Droplet-based microfluidics for the genotype-phenotype mapping of model enzymes

Dany Chauvin

Thèse de doctorat
de l'Université Sorbonne Paris Cité
Préparée à l'Université Paris Diderot

Ecole doctorale Frontières du Vivant (ED 474)

Laboratoire de Biochimie, ESPCI Paristech

Droplet-based microfluidics for the genotype-phenotype mapping of model enzymes

Par Dany Chauvin

Thèse de doctorat de Biophysique
Dirigée par Clément Nizak, Olivier Rivoire et Andrew Griffiths

Présentée et soutenue publiquement à Paris le 29 Septembre 2017

Président du jury :	Soumillion, Patrice, Professeur, Université catholique de Louvain
Rapporteurs :	Hollfelder, Florian, Professeur, University de Cambridge
	DeMello, Andrew, Professeur, ETH Zurich
Examinateur :	Urvoas-Cisse, Agathe, Maître de conférence, Université Paris-Sud
Directeur de thèse :	Nizak, Clément, Chargé de recherche CNRS, ESPCI Paristech
Co-directeurs de thèse :	Rivoire, Olivier, Chargé de recherche CNRS, Collège de France
	Griffiths, Andrew, Professeur, ESPCI Paristech

Abstract

Titre:

Microfluidique en gouttelettes pour la cartographie génotype-phénotype d'enzymes modèles.

Résumé:

Je présente ici le développement d'un système expérimental permettant d'effectuer, quantitativement et à haute fréquence, la cartographie génotype-phénotype d'enzymes modèles. Cette cartographie est réalisée en combinant la microfluidique en gouttelettes et les technologies de séquençage nouvelle génération.

La relation qui lie la séquence d'une protéine à sa fonction nous échappe toujours en grande partie, pourtant elle est essentielle à la compréhension de l'évolution moléculaire. Des études théoriques basées sur des modèles physiques, des informations structurales ou des données de séquençage, tentent de prédire les propriétés de la relation génotype-phénotype ainsi que les effets des mutations. Ces prédictions doivent être testées expérimentalement et affinées par comparaison avec des données expérimentales de cartographie génotype-phénotype à grande échelle.

Les expériences de "Deep Mutational Scanning" profitent des technologies de séquençage nouvelle génération afin de séquencer des bibliothèques de mutants de protéines, générées par mutagenèse saturée, sous différentes pressions de sélection. Ces expériences mettent en lumière comment l'"épistasie", la "robustesse" et les compromis entre les propriétés des protéines (e.g stabilité et/ou différentes activités) sont structurellement et statistiquement distribués. Mais ces études souffrent de certaines limitations: mesurer le phénotype d'enzymes in vitro requiert de (i) coupler physiquement le génotype et le produit de la réaction enzymatique, (ii) découpler l'activité totale du niveau d'expression du mutant étudié. Ces deux éléments sont rarement rencontrés dans la littérature, et à notre connaissance, jamais associé à un système d'essai enzymatique in vitro et haut-débit.

La microfluidique permet de remplacer les traditionnels tubes à essais par des micro-gouttelettes afin de tester séparément des mutants d'enzyme à des fréquences de l'ordre du kilohertz. Cette technique fournit un moyen de coupler le génotype et le produit de l'activité enzymatique. Sélectionner et récupérer les gouttelettes sur demande et séquencer leur contenu permet d'effectuer la cartographie génotype-phénotype de millions de mutants d'enzymes en une seule expérience.

Au cours de cette thèse, j'ai tout d'abord développé un système microfluidique basé sur l'expression de protéines in vitro afin de pouvoir réaliser la cartographie génotype-phénotype de Streptomyces griseus aminopeptidase (SGAP). Des gènes mutants de l'enzyme SGAP sont encapsulés (un

par gouttelette au maximum) amplifiés, exprimés et testés contre un substrat fluorogénique. Des incompatibilités entre les étapes d'amplification, d'expression et d'essai enzymatique en gouttelettes obligent à réaliser chacune de ces étapes séparément et successivement, afin de diluer le produit de chaque réaction par l'électro-coalescence des gouttelettes. Je montre qu'un work-flow microfluidique dans lequel (i) les gènes sont encapsulés et amplifiés dans des gouttes de 0.2 pL, (ii) exprimés in vitro, (iii) testés contre un substrat fluorogenique dans des gouttelettes de 20 pL, permet de mesurer l'activité de variants de SGAP avec un contraste important. Afin d'optimiser l'essai enzymatique en gouttelettes de SGAP, j'ai aussi développé, en collaboration avec Dr. Johan Fenneteau (Laboratoire de Chimie Organique, ESPCI Paristech), un nouveau substrat fluorogénique basé sur une rhodamine hydrophile. Cette sonde est caractérisée par un échange limité de la rhodamine entre les gouttelettes.

J'ai ensuite développé un work-flow microfluidique in vivo, pour Ratus norvegicus trypsin (la trypsine du rat), dans lequel les capacité de sécrétion de Bacillus subtilis sont utilisées afin de simplifier les expériences. Des cellules uniques de B. subtilis sont encapsulées dans des gouttelettes de 20 pL où elles sécrètent des mutants de la trypsine en protéine de fusion avec un rapporteur permettant de mesurer le niveau d'expression. Les mutants sont testés par électro-coalescence avec des gouttelettes de 2 pL contenant un substrat fluorogénique de la trypsine. En normalisant l'activité totale détectée par la fluorescence du rapporteur du niveau d'expression, l'efficacité catalytique peut être directement mesurée en gouttelettes. C'est la première fois qu'un système expérimental d'essai enzymatique haut-débit fournit l'opportunité de mesurer directement l'efficacité catalytique de mutants d'une enzyme à une fréquence de l'ordre du kilo Hertz. Une méthode afin de réaliser la mutagenèse saturée (tous les simples mutants) du gène de la trypsine du rat a aussi été développée. Combinée au séquençage nouvelle génération, la méthode microfluidique développée permettra de réaliser la première cartographie génotype-phénotype de tous les simples mutants de la trypsine du rat.

Mots-clés:

Cartographie génotype-phénotype, enzyme, trypsine, SGAP, microfluidique en gouttelettes, deep mutational scanning.

Title:

Droplet-based microfluidics for the genotype-phenotype mapping of model enzymes.

Abstract:

In this thesis manuscript, I present the development of an experimental set-up to perform high-throughput quantitative genotype-phenotype mapping of model enzymes, combining droplet-based microfluidics and next generation sequencing.

The question of how sequence encodes proteins' function is essential to understand molecular evolution but still remains elusive. Theoretical studies based on physical model, structural information or protein sequence, predict mutational effects and properties of the highly dimensional genotype-phenotype landscape. However such predictions need to be empirically tested and refined using large scale high-throughput genotype-phenotype mapping data.

Deep mutational scanning experiments take advantage of next generation sequencing technologies to sequence protein libraries generated by saturated mutagenesis and tested under varying selection pressures. Such studies shed a new light on how epistasis, robustness, promiscuity and trade-offs between proteins' properties (e.g stability and activity), are structurally and statistically distributed. But these studies suffer from limitations in measuring enzymes' phenotype in vitro, as it is necessary to (i) physically couple the genotype and the product of the enzymatic activity and (ii) decouple the total activity measured from the enzyme expression level.

Droplet-based microfluidics allows to use micro-metric droplets as reaction vessels to separately assay enzyme variants at the kHz frequency. It also provides an elegant solution to couple the genotype with the product of the catalytic activity of enzymes. Sorting droplets on demand and sequencing their content enables to map the genotype of millions of enzyme variants to their phenotype in a single experiment.

First, I developed a cell-free microfluidic work-flow to carry out genotype-phenotype mapping of *Streptomyces griseus* aminopeptidase (SGAP). Single enzyme variant genes are encapsulated and amplified in droplets, expressed, and assayed against a fluorogenic substrate. Incompatibilities between gene amplification, expression and assay reactions, constrain to execute each one of those steps successively and to dilute the product of each reaction by droplet electro-coalescence. I show that a work-flow in which (i) genes are encapsulated and amplified into 0.2 pL droplets, (ii) expressed using cell-free expression reagents in 2 pL droplets and (iii) assayed with a fluorogenic substrate in 20 pL droplets, allows to measure SGAP variants activity with high contrast. To optimize the SGAP droplet assay, I also developed in collaboration with Dr. Johan Fenneteau (Laboratory of

Organic Chemistry, ESPCI Paristech), a hydrophilic rhodamine based substrate, characterized by limited exchange of the released fluorophore between droplets.

Second, I developed an in vivo microfluidic work-flow on *Ratus norvegicus* trypsin (rat trypsin), in which *Bacillus subtilis* secretion abilities are used to simplify the microfluidic work-flow. Single *B. subtilis* cells are encapsulated in 20 pL droplets where they secrete trypsin variants as fusion proteins with a fluorescent expression-level reporter. The variants are assayed by droplet electro-coalescence with 2 pL droplets containing a trypsin fluorogenic substrate. Trypsin variants catalytic efficiency can be directly measured in droplets, by normalizing the total trypsin activity by the expression-level reporter fluorescence. This is the first time a high-throughput protein assay work-flow gives the opportunity to directly measure the catalytic efficiency of enzyme variants at the kHz frequency. A method to carry out saturated mutagenesis on the rat trypsin gene was also developed. Together with deep sequencing, the developed experimental work-flow will allow to perform the first quantitative genotype-phenotype mapping of all single point mutants of the rat trypsin protein.

Keywords:

Genotype-phenotype mapping, enzyme, trypsin, SGAP, droplet-based microfluidics, deep mutational scanning.

Remerciements

Je tiens à remercier en premier lieu mes directeurs Andrew, Olivier et Clément, pour leur accompagnement bienveillant et leur confiance. Ils m'ont permis de progresser dans une atmosphère informelle, mais rigoureuse et de haut niveau scientifique. Je remercie également l'IPGG, qui a financé mes trois premières années de thèse et plus particulièrement l'équipe du plateau technologique grâce à qui j'ai pu travailler dans des conditions très confortables.

Je remercie chaleureusement Aude Bernheim (pour les transfos), Marie Leman (pour la litho mais pas seulement), Johan Feneteau, Elodie, Alexei, Gab, Thomas, Guansheng, Baptiste, Klaus, Yacine, Pablo et Sandeep (my "favourite" post-doc!) pour leur nombreux conseils, leur aide et leur bonne humeur permanente face à mes questions et/ou à mon humour douteux. J'espère pouvoir retrouver dans le futur une ambiance digne de celle du Laboratoire de Biochimie ! Merci à Isa et Hélène qui maintiennent le navire à flot et jettent les bouées de sauvetage avant même que nous ne tombions à l'eau. Merci à Amandine, Raph et Antoine avec qui j'ai partagé quatre ans de joyeuse vie commune. Sont nominés aux victoires de la musique, dans la catégorie "Pop-Rock-Blues-Musique-du-Monde", Marco "Big Droplet Daddy" Ribezzi et Marina "F" Theodorou, mes camarade du groupe « The Droplets ». Mes amis, si l'on perce un jour dans la musique, ça sera à titre post-posthume. Toujours pour les victoires de la musique, dans la catégorie "révélations" sont nominés : Angga, le soliste du silane ainsi que Matt. Merci Matt de n'avoir jamais refusé que je massacre avec toi, a capella, un classique de Bowie lors de nos moments de solitude à la paillasse. La politesse canadienne est fidèle à sa réputation. Merci à Sophie et Simon « Little Freak", le ying et le yang, pour leur soutien et leur « non-soutien » respectif. Je remercie aussi Stéphanie, notre "managueuse" de nous avoir laissé siffler du whisky dans les bureaux sans jamais prévenir les autorités compétentes. Merci aux stagiaires, qui ont beaucoup apporté à ce projet de thèse: Ryma, Antonio, Judith, Guillaume, Elisa, Jérémie-Luc et Sebastian. Merci à nos compagnons du bout du couloir (Hifibio), toujours prompts à me faire l'aumône de quelques réactifs et à Marcel en particulier pour ses nombreux conseils techniques en microfluidique.

Merci à Bill, Rama, David et Victor pour leur aide sur ce projet et pour leur accueil chaleureux et les nombreuses découvertes gastronomiques à Dallas !

Merci à l'école doctorale Frontières du Vivant et plus particulièrement à Sophie Leon, David Tareste et Elodie Kaslikowski pour leur réactivité et leur bienveillance tout au long de la thèse. L'école fut l'occasion de rencontrer de nombreuses personnes qui ont considérablement élargis mon champs d'horizon. Je pense en particulier à Adrien le « spleen doctor » et Léa, Iryna, Ameline (et Polo!), Marion et Bé, avec qui j'ai eu maintes fois l'occasion de décompresser autour d'un verre.

Merci au Club des 5, excellent exutoire dans les moments « difficiles »: Charles G.C, Olivier (un type gentil comme pas deux), Pierre, Clément et Alex. Alex, ta presque victoire à « Ma thèse en 180s », qui restera dans les

annales, m'a beaucoup inspiré !

Merci aux nombreux amis dont la curiosité et l'enthousiasme (re)donnent de l'énergie : les Sébastiennnais et autres Nantais toujours aussi festifs et les Raideurs et associés. Le catalyseur se reconnaîtra. Merci à Julien, Pit, Tanguy et Jack Sparrow : si seulement votre amour pour la biologie nous vous avait pas rendu aveugle...

J'ai la chance d'avoir une famille qui m'a gâté pendant toutes ces (longues) années d'études. Mes pensées vont à eux en terminant cette page de remerciements: merci à Papa, Maman, à Flo et Nini ainsi qu'à Amandine, pour votre soutien inconditionnel.

Contents

Abstract iii

Remerciements viii

1 Introduction 1
1.1 The genotype-phenotype relationship . . . 2
1.1.1 Genotype, phenotype and the landscape metaphor . 2
1.1.2 The high dimensionality of genotype space or sequence space . . . 3
1.1.3 Questioning the properties of the genotype-phenotype landscape . . . 5
1.1.4 Knowing the genotype-phenotype landscape: opportunities and potential applications . . . 5
1.1.5 Approaches to understand the genotype-phenotype landscape . . . 8
Numerical approaches from the first principle of physics and protein structural information . . . 8
Statistical inference on protein sequence data . . . 8
Experimental approaches: the need for large scale genotype-phenotype mapping experiments . . . 11
1.2 Current high-throughput genotype-phenotype mapping approaches: deep mutational scanning experiments . . . 11
1.2.1 Generate sequence variants by mutagenesis . . . 12
1.2.2 Opportunities and challenges of next generation sequencing . . . 13
1.2.3 Mapping the genotype to the phenotype in deep mutational scanning experiments . . . 13
General considerations about deep mutational scanning 14
Limitations of those approaches . . . 16
Overcoming those limitations for a quantitative genotype-phenotype mapping . . . 17
1.2.4 Properties of the genotype-phenotype landscape as seen through deep mutational scanning . . . 18
Distribution of robustness and agreement with the phylogeny . . . 18
Distribution of beneficial mutations . . . 22
Distribution of epistatic effects . . . 22
Influence of robustness on molecular evolution . . . 24
Influence of epistasis on molecular evolution . . . 25
The roles of promiscuity and evolvability in molecular evolution . . . 25
Trade-offs and couplings between proteins' properties 26

1.2.5 Performing genotype-phenotype mappings of enzymes is limited with current high-throughput approaches . 28
1.3 Droplet-based microfluidics to perform high-throughput enzymatic assays . . . 28
1.3.1 Droplet-based microfluidics: high-throughput manipulation of micro-metric reaction vessels . . . 29
The opportunities offered by the picoliter droplet format in biology . . . 29
The development of bio-compatible materials dedicated to microfluidics . . . 29
Microfluidic devices to perform high-throughput operations . . . 30
1.3.2 Constraints of the droplet format on performing quantitative enzymatic assays . . . 32
General droplet-based microfluidics constraints . . . 34
Constraints specific to enzymatic assays . . . 34
1.3.3 Current droplet-based microfluidic work-flows . . . 38
1.3.4 Conclusion: requirements for a quantitative genotype-phenotype mapping of model enzymes in droplets . 40
1.4 Studying model enzymes to better understand the link between genotype and phenotype . . . 40
1.4.1 *Streptomyces griseus* Aminopeptidase (SGAP) to study allostery and promiscuity . . . 42
1.4.2 *Ratus norvegicus* trypsin (rat trypsin). Are trypsin sectors independent functional units? What is the relation between the trypsin sectors and epistasis? . . . 44

2 Material and methods: droplet-based microfluidics and fluorogenic substrates 49
2.1 Droplet-based microfluidics . . . 49
2.1.1 Microfluidic devices fabrication . . . 49
2.1.2 Fabrication and preparation of disposable devices to handle aqueous phases and emulsions . . . 51
2.1.3 Microfluidic devices operation . . . 53
Optical and electrical setup . . . 53
Preparing the microfluidic devices . . . 53
Making and manipulating droplets . . . 54
Measuring droplet fluorescence . . . 54
2.2 Fluorogenic substrates . . . 58
2.2.1 SGAP fluorogenic substrates . . . 58
2.2.2 Trypsin fluorogenic substrates . . . 58

3 Development of a cell-free microfluidic work-flow for the genotype-phenotype mapping of *Streptomyces griseus* Aminopeptidase (SGAP) 59
3.1 Previous work on a cell-free microfluidic work-flow to assay SGAP in droplets . . . 60
3.2 Development of the in vitro workflow 1: pico-injecting the substrate in droplet containing the enzyme . . . 63
3.2.1 The PCR reagents inhibit cell-free expression in bulk 63
3.2.2 Cell-free expression is successful in bulk and droplets 63

3.2.3 Cell-free expression reagents inhibit SGAP enzymatic activity in bulk: picoinjection is incompatible with activity detection . . . 64
3.2.4 Cell-free expression droplets have to be diluted in the assay droplets . . . 68
3.3 Development of the in vitro workflow 2: diluting the enzyme containing droplet into the substrate containing droplet . . . 70
3.3.1 Fusing 2 pL expression and 20 pL assay droplets allows aminopeptidase activity detection with high contrast . . . 70
3.3.2 Synthesizing a new non leaky substrate to improve SGAP enzymatic assay in droplets . . . 72
3.3.3 SGAP PCR amplification is successful in 0.2 pL droplets 78
3.3.4 0.2-2pL droplet electro-coalescence development . . 80
3.4 Discussion, Conclusion and Perspectives . . . 80

4 Development of an in vivo microfluidic workflow for the genotype-phenotype mapping of *Ratus norvegicus* trypsin 83
4.1 The in vivo workflow 1: rat trypsin periplasmic expression in *E. coli* . . . 84
4.1.1 Osmotic shock of E.coli cells in hypotonic buffer allows trypsin activity detection in bulk . . . 86
4.1.2 MUGB inhibits trypsin activity but cannot be used as a reporter for its concentration . . . 86
4.1.3 Normalizing rat trypsin activity using mCherry as a reporter of its expression level . . . 91
4.1.4 Difficulties with the *E. coli* expression system . . . 92
4.2 The in vivo workflow 2: trypsin secretion by *B. subtilis* . . . 95
4.2.1 The rat trypsin - mCherry fusion protein is secreted by WB800N as a full protein in the supernatant . . . 97
4.2.2 The rat trypsin - mCherry fusion protein is fluorescent and enzymatically active in the culture medium 99
4.2.3 The mCherry fluorescence can be used as a reporter of the trypsin expression level . . . 100
4.2.4 Optimizing incubation time for the trypsin-mCherry expression in droplets . . . 102
4.2.5 Shaking the emulsion during incubation for expression reduces emulsion size polydispersity . . . 102
4.2.6 Measuring the catalytic efficiency of trypsin variants in bulk . . . 104
4.2.7 Measuring the catalytic efficiency of trypsin variants in droplets . . . 107
4.3 Towards a library of all rat trypsin single point mutants . . . 108
4.3.1 "Around the horn" site directed mutagenesis principle 111
4.3.2 Designing the mutagenic primers . . . 113
4.3.3 Performing saturated mutagenesis on the rat trypsin protein . . . 114
4.3.4 Analyzing the first rat trypsin library by deep sequencing . . . 114
First deep sequencing run results . . . 116

Second deep sequencing run results, with fraction of the mutagenic primers redesigned. 117
Coverage of all single point mutants 119
4.4 Discussion, Conclusion and Perspectives 123

5 Conclusion **125**
5.1 Conclusion and further work 125
5.2 Perspectives 126

A Appendix: Development of a cell-free microfluidic work-flow for the genotype-phenotype mapping of *Streptomyces griseus* Aminopeptidase (SGAP) **129**
A.1 Experimental details 129
A.1.1 Plasmid maps and primer sequences 129
A.1.2 General comments about the experiments 129
A.1.3 SGAP PCR protocol with Phusion polymerase 131
A.1.4 The PCR reagents inhibit cell-free expression in bulk 3.2.1 131
A.1.5 Cell-free expression is successful in bulk and droplets 3.2.2 131
A.1.6 Cell-free expression reagents inhibit SGAP enzymatic activity in bulk: pico-injection is incompatible with activity detection 3.2.3 132
A.1.7 Fusing 2 pL expression and 20 pL assay droplets allows aminopeptidase activity detection with high contrast 3.3.1 132
A.1.8 SGAP PCR amplification is successful in 0.2 pL droplets 3.3.3 133
A.1.9 Random mutagenesis on the SGAP gene 133
A.2 Synthesis of New Hydrophilic Rhodamine Based Enzymatic Substrates Compatible with Droplet-Based Microfluidic Assays 3.3.2 134

B Appendix: Development of an in-vivo microfluidic work-flow for the genotype-phenotype mapping of *Ratus norvegicus* trypsin **159**
B.1 *Ratus norvegicus* trypsin expression in *E. coli* 159
B.1.1 Plasmid maps 159
B.1.2 General comments about the experiments 159
B.1.3 Osmotic shock of E.coli cells in hypotonic buffer allows trypsin activity detection in bulk 4.1.1, 4.1.3 . . 161
B.1.4 Lysozyme and sucrose only marginally improve trypsin activity detection 4.1.1 161
B.1.5 MUGB inhibits trypsin activity but cannot be used as a reporter for its concentration 4.1.2 162
B.1.6 Development of a non leaking substrate based on the fret pair EDANS-Dabcyl 4.1.4 162
B.1.7 Rat trypsin in *B. subtilis*: plasmid maps 163
B.2 *Ratus norvegicus* trypsin expression in *B. subtilis* 163
B.2.1 General comments about the experiments 163
B.2.2 *B. subtilis* transformation protocol 164
B.2.3 Induction protocol 164

B.2.4 Droplet induction protocol . . . 165
B.2.5 Bulk assay protocol . . . 165
B.2.6 Droplet assay protocol . . . 165
B.2.7 The rat trypsin - mCherry fusion protein is secreted by WB800N as a full protein in the supernatant 4.2.1 166
B.2.8 Measuring the catalytic efficiency of trypsin variants in droplets 4.2.1 . . . 166
B.3 Towards a library of all rat trypsin single point mutants . . . 167
B.3.1 Designing the mutagenic primers 4.3.2 . . . 167
B.3.2 Performing saturated mutagenesis on the rat trypsin protein 4.3.3 . . . 168
B.3.3 Library preparation 4.3.3 . . . 169
B.3.4 Redesigning part of the mutagenic primers . . . 170
B.3.5 Sequencing data analysis . . . 172

Bibliography **175**

List of Figures

1.1 Two different representation of the myoglobin molecule . . . 2
1.2 How does the sequence determine the protein function? . . 2
1.3 The genotype-phenotype landscape metaphor 4
1.4 Strategies to evolve proteins in laboratory experiments . . . 7
1.5 Conserved and coevolving residues within an artificial multiple sequence alignment 9
1.6 PDZ specificity switch via class-bridging mechanism or class-switching mechanism 10
1.7 Typical results of deep mutational scanning experiments . . 15
1.8 Distribution of fitness of single amino-acid point mutation in TEM-1 20
1.9 Fitness distribution of the DNA methyl transferase M.HaeIII 21
1.10 Distribution of epistasis in the IgG1 binding domain of the G protein (56 residues) 23
1.11 Mutational effects single point mutants of TEM-1 against ampicilline and cefotaxime 24
1.12 Network properties of the functional variants of ParD 26
1.13 A selected set of operations than can be performed on droplets 31
1.14 Scheme of a high resolution genotype-phenotype mapping using droplet-based microfluidics 33
1.15 Limiting rhodamine leakage using a hydrophilic tag 38
1.16 Visualization and sequencing of the whole distribution of phenotypes 41
1.17 SGAP structure and specificity 43
1.18 Statistical Coupling Analysis (SCA) reveals the existence of three independent groups of co-evolving residues 45
1.19 Mutational effects of a few single point and double point mutations on rat trypsin stability and catalytic efficiency. 46

2.1 Typical soft-lithography protocols 50
2.2 Soft-lithography and PDMS chip fabrication 51
2.3 Microfluidic station optical setup 53
2.4 Microfluidic station additional references 53
2.5 2-5 pL droplet maker 55
2.6 20-30 pL droplet maker 55
2.7 20-30 pL co-flow droplet maker 56
2.8 20-30 droplet reinjector 56
2.9 2-20 pL droplet electro-coalescence device (integrated with delay lines and sorting device) 57
2.10 Structure of the two commercial fluorogenic substrates used in the SGAP experiments 58
2.11 Structure of the two commercial fluorogenic substrates used in the trypsin experiments 58

3.1 In vitro microfluidic workflow 1 as in [197]. 62
3.2 Aminopeptidase activity of final expression products prepared with varying SGAP PCR product volume fractions 64
3.3 Aminopeptidase activity for different expression incubation times in bulk. 65
3.4 Droplet expression product aminopeptidase activity and comparison with bulk controls. 66
3.5 Aminopeptidase activity of different expression product volume fractions in the assay mix. 67
3.6 In vitro microfluidic workflow 2 69
3.7 Validation of the enzymatic assay in droplets by fusing 2 pL expression droplets with 20 pL substrate droplets. 71
3.8 0.2 pL droplet production and its application to droplet PCR. 79

4.1 In vitro microfluidic work-flow envisioned for the *E. coli* expression system. 85
4.2 Trypsin activity for different periplasmic extraction protocols 87
4.3 Trypsin activity for different periplasmic extraction protocols involving lyzozyme 88
4.4 Trypsin activity before and after adding MUGB 89
4.5 MugB cannot be used as a reporter for trypsin expression level 90
4.6 Activity of a subset of rat trypsin variants (hypotonic conditions) 91
4.7 MCherry and trypsin mCherry expression in *E. coli* 92
4.8 EDANS-Dabcyl substrate structure 93
4.9 The EDANS Dabcyl FRET substrate is cleaved by trypsin activity and not leaking between droplets. 94
4.10 Partial map of the trypsin mCherry *B. subtilis* expression vector 95
4.11 In vitro microfluidic workflow envisioned for the *B. subtilis* expression system. 96
4.12 Anti-mCherry western blot performed on the supernatants of induced culture of *B. subtilis* 98
4.13 Trypsin activity and Mcherry fluorescence detected in WB800N cultures 99
4.14 Effect of the temperature on *B. subtilis* secretion levels 100
4.15 Using mCherry fluorescence to normalize trypsin activity . . 101
4.16 Kinetic of trypsin mCherry expression in droplets 103
4.17 Effect of the agitation on induction in droplets and emulsion polydispersity. 104
4.18 Bulk ranking of a subset of trypsin mutants 105
4.19 Expected ranking of a subset of trypsin variants. 106
4.20 Trypsin activity versus mCherry fluorescence in droplets . . 109
4.21 Normalizing trypsin activity in droplets to measure catalytic efficiencies 110
4.22 "Around the horn" mutagenesis principle [211] 112
4.23 Melting temperature distribution of the designed mutagenic primers 113
4.24 Mutagenic PCRs (band size and amplification) according to different PCR protocols. 115
4.25 Characterization of the library of single point mutants (first deep sequencing). 118

4.26 Characterization of sequencing errors (first deep sequencing). 119
4.27 Characterization of the second trypsin library of single point mutants by deep sequencing 120
4.28 Characterization of the third trypsin library of single point mutants. 120
4.29 Frequency matrix of single point mutations along the trypsin polypeptide chain. 121
4.30 Total frequency of occurrence of mutations at each position of the trypsin peptide chain. 122
4.31 Coverage of all single point mutants as a function of the number of recovered colonies for the third library. 123

A.1 pIVEX SGAP plasmid map 130
A.2 Structure of the two main substrates used in the experiments of this chapter. 130

B.1 Rat trypsin sequence in pET28, plasmid map 160
B.2 Rat trypsin mCherry sequence in pET28, plasmid map . . . 160
B.3 pHT43c rat trypsin mCherry, plasmid map 163
B.4 pET28(b) rat trypsin mutagenesis partial plasmid map. . . . 164

Chapter 1

Introduction

Wilhelm Johannsen introduced the words "gene", "genotype" and "phenotype" in 1911 [1] to respectively define the unit of heredity, "the sum total of all the genes" and the "forms, structures, sizes, colors and other characters of living organisms".

A little more than a hundred of years later, these definitions have barely changed. But in the mean time, our understanding of the mechanisms by which the units of heredity are expressed in observable traits greatly improved. Thomas Hunt Morgan discovered that heredity is mostly supported by the chromosomes [2], Oswald Avery discovered that the latter are made of deoxyribonucleotides (DNA) sequences [3–5] and Watson and Crick elucidated the double helix structure of DNA [6]. By 1965, Nirenberg, Khorana and Holley had cracked the genetic code [7], by which DNA sequences, the genes, are translated into proteins, the active molecules which support the phenotype.

Just a few years before, in 1958, John Kendrew had already published the first crystallographic structure of a protein, the myoglobin, a blood protein in charge of transporting the oxygen throughout the body [8]. Kendrew and many others thought the mechanisms of proteins' function would naturally emerge from their three dimensional structure, i.e the way the chains of amino-acids fold on themselves to adopt the conformation that confers their function. But the complexity of Kendrew's finding was unexpected (see figure 1.1), as he declared:

"Perhaps, the most remarkable features of the molecule are its complexity and its lack of symmetry. (...) it is more complicated than has been predicated by any theory of protein structure".[9]

Fifty years later, the two questions that Kendrew underlined are still unresolved: how does the protein sequence (genotype) determine the protein fold and how does the protein fold determine the protein's function (phenotype)?

The work I present in this manuscript aims at setting up a new method to experimentally and quantitatively map the genotype of millions of enzyme variants to their phenotype ("genotype-phenotype mapping").

This method consists in combining droplet-based microfluidics and next-generation sequencing to perform millions of enzymatic assays on as many

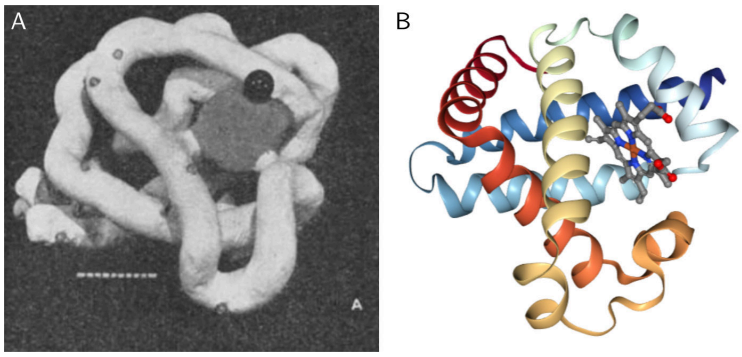


FIGURE 1.1: **Models of the myoglobin molecule**. **A**: Photographs of the first model of the myoglobin molecule as published by Kendrew et al. in 1958. Polypeptide are white. The grey disk is the haem group that binds to oxygen. **B**: Crystallographic structure of the same molecule as deposited on the Protein Data Bank in 2011 (PDB: 3RGK).

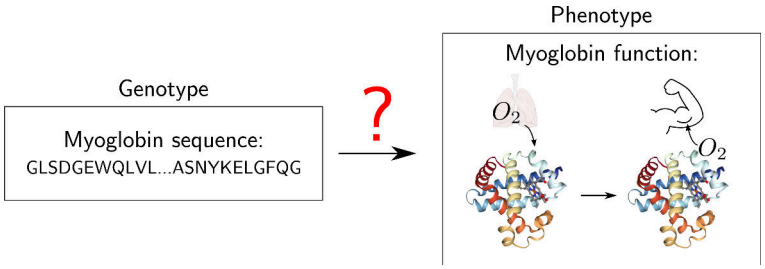


FIGURE 1.2: How does the myoglobin sequence encode the myoglobin function: transporting the oxygen from the lungs to the muscles?

enzyme variants and to sequence their genotype. By analyzing the relationship of millions of genotype-phenotype couples, we wish to determine the statistical invariants that determine the relationship between the genotype and the phenotype and more generally to better understand molecular evolution (see figure 1.2).

In this introductory chapter, I will first discuss in what aspects knowledge of the genotype-phenotype relationship is fundamental and how in-silico, statistical and experimental approaches tackle this problem. Then, I will discuss how high-throughput experiments are shedding a new light on the genotype-phenotype relationship, via deep mutational scanning experiments. I will discuss the current limitations of these experimental approaches when it comes to study enzymes and how droplet-based microfluidics can overcome the challenges. Finally, I will present the model enzymes for which we intend to develop droplet-based microfluidic work-flow.

1.1 The genotype-phenotype relationship

1.1.1 Genotype, phenotype and the landscape metaphor

In the following paragraphs, I will adopt the molecular perspective. The genotype is here defined as a particular DNA sequence and the phenotype is defined as any observable biochemical property presented by the protein (stability, catalytic properties, etc.). The term function can either refer to the

phenotype or to a sum of biochemical properties which confer a role to the protein within the cell (e.g solubility combined with catalytic activity).

Molecular darwinian evolution via natural selection consists in repeated cycles of variation and selection. A parental genotype is replicated in offspring genotypes bearing variations compared to the parental genotype. Each genotype is expressed as a protein with a phenotype that confers a replication rate to the genotype. If the phenotype is adapted to the environment, the replication rate is high and vice versa. Generation after generation, the frequency of genotypes that give the most adapted phenotypes, increases in the pool of genotypes. In an environment where the resources for genotype replication are limited, the least adapted genotype-phenotype couples will eventually disappear from the population.

For evolution to occur, a physical linkage must exist between the genotype and the phenotype so that the selection of the phenotype implies the co-selection of the genotype. In living organisms, such linkage is carried out by the cell membrane.

The first visualization of a possible relationship between the genotype and the phenotype, was depicted by Wright [10]. Wright imagined this relationship as a 3D "landscape", where the horizontal axes represent the genotype space (or "sequence space"), and the vertical axis represents the phenotype[1]1.3: "low" and "high" phenotypes being respectively less adapted and more adapted phenotypes. Evolution by natural selection can be seen as a process by which a population explores the sequence space and is attracted towards the regions of more adapted phenotypes through genotype variation and phenotype selection.

This "landscape metaphor" is highly questionable. First, the landscape is not static but dynamic: the fluctuating environment might select for different functions overtime, in magnitude (e.g different binding affinities) and in nature (e.g binding another ligand). Second, the sequence space is represented as a 2 dimensional continuous space whereas it is highly dimensional and discrete: particular topological properties might emerge from its high dimensionality[11]. Nevertheless, this metaphor is widely used to represent evolution and mutational effects [12].

1.1.2 The high dimensionality of genotype space or sequence space

The diversity of sequences offered by the sequence space is virtually infinite. The dimension of the sequence space is equal to the number of residues composing the protein, L. With $L = 300$ (close to the average length of proteins in *E. coli* [13]) and 20 natural amino-acids, the sequence space is made of $20^{300} \approx 10^{390}$ possible combinations.

[1]To be precise, Wright intended to depict the relationship between the genotype and the reproductive success of organisms, sometimes called "fitness". The "fitness landscape" metaphor was then extended to the molecular perspective via the concepts of "genotype-phenotype landscape", "function landscape" or "mutational landscape".

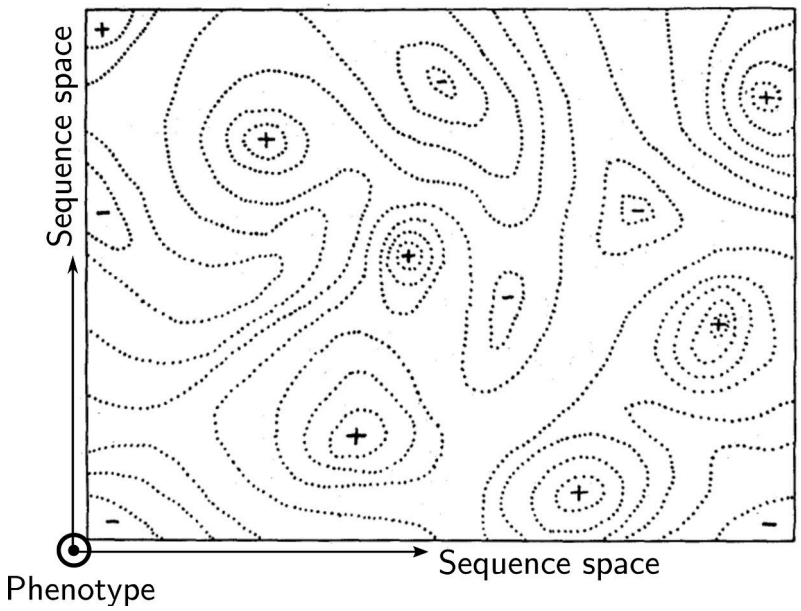


FIGURE 1.3: **The genotype-phenotype landscape metaphor.** In the plan of the page (2 dimensions) is represented the sequence space. In the direction orthogonal to the page is represented the protein's phenotype. Dotted lines represent lines of equivalent phenotype. The closer they are from each other, the steeper is the slope towards maxima or minima of phenotypes, which are respectively signaled with '+' and '-'. Adapted from [10]

Even from a "local" point of view, the diversity if overwhelming as mutations accumulate. Let us compute the number of neighbors that are one mutation away from a reference sequence of 300 amino-acids: $(20-1) \times 300 = 5700$ neighbors. Let us now consider the two mutations away neighbors: $((20-1)^2 \times 300 \times 299)/2 \approx 10^7$ neighbors. In general, the number n of neighbors, k mutations away from the reference is:

$$n = \binom{300}{k} \times 19^k$$

Within one family of homologous proteins, amino-acids homology between two different sequences can be less than 25% [14]. A 20% sequence homology accounts for $0.8\% \times 300 = 240$ mutations. The number of neighbors that are 240 mutations away from the reference is then $\approx 10^{64}$.

Thanks to this available sequence diversity and owing to the diverse physical and chemical properties of the amino-acids, proteins display a multitude of functions: they catalyze reactions (polymerization, condensation) or the folding of other proteins (chaperonins), bind other objects (DNA, proteins, atoms, ions, lipids, polymers), play the role of molecular motors, arrange in large structures to constitute the cytoplasmic scaffold of cells (actine filaments, micro-tubules), etc [15]. How is protein function distributed in the vast and highly dimensional genotype-phenotype landscape?

Given the size of the sequence space, only small fractions of it can be

probed experimentally. It is known that mutations can have very different effects on the phenotype of proteins: some are highly deleterious while some others are neutral, even beneficial or conferring new functions to the proteins. It also has been shown that the effect of mutations depends on the presence of other mutations. Are those mutational effects reflecting global properties of genotype-phenotype landscapes?

1.1.3 Questioning the properties of the genotype-phenotype landscape

Given the individual mutational effects already measured, concepts have emerged to qualify proteins' properties upon mutation:

- Robustness: the capacity for a protein to accumulate mutations while keeping a constant function.
- Epistasis: the context dependence of mutational effects (i.e mutational effects are different in the presence of other mutations).
- Evolvability: the protein potential to evolve towards an improved function or a new function.

The genotype-phenotype landscape is the complex product of at least millions (if only two bodies interactions are considered) of physical and chemical interactions. Those concepts need to be questioned statistically, with high-throughput experimental genotype-phenotype mapping approach:

- What are the parameters of protein "robustness"? How is it distributed into the local sequence space?
- What is the distribution of epistasis? How is it influencing the shape of the genotype-phenotype landscape? If its influence is important, is considering epistatic effects between pairs or residues enough to model the landscape? In other words, which epistatic order is the most relevant to model the landscape?
- What are the parameters of evolvability?

Are those concepts even relevant to describe the genotype-phenotype landscape? If yes, how are they related? If no, what parameters should we use to reduce the high-dimensionality of the landscape? Answering those questions should provide insights to understand how protein sequence encodes the proteins' fold and function and how proteins evolve in the sequence space.

1.1.4 Knowing the genotype-phenotype landscape: opportunities and potential applications

Knowing the properties of the genotype-phenotype relationship thanks to genotype-phenotype mapping would considerably improve our understanding of molecular evolution. Population and human genetics, protein engineering and protein design would highly benefit from this knowledge.

Population genetics and human genetic variation The emergence and rise of antibiotic resistances among bacteria is of great concern [16]. Knowing the genotype-phenotype relationship of proteins involved in the resistance against antibiotics would reveal the mechanisms of resistance at the molecular level. It would also highlight the evolutionary paths available towards improved or new resistance. Such insights could be used to push their evolution towards evolutionary dead-ends. The same ideas could be applied to prevent virus propagation and predict their evolution in order to conceive effective vaccines against emerging strains [17].

In the field of human diseases, understanding the relationship between protein sequences and functions would allow to predict the impact of rare mutations in the frame of complex diseases [18].

Protein engineering and design Most of the proteins used in the industry (from biotechnology to textile manufacturing) are engineered proteins [19, 20]. Many were evolved using directed evolution, a laboratory technique which mimics evolution by natural selection, to enhance the properties of natural or artificial proteins, such as their stability, activity or solubility. Directed evolution was notably applied on:

- Antibodies evolved towards higher affinity or different specificities to bind factors involved in inflammatory reactions [20].
- Enzymes evolved towards higher stability (to be processive at high temperatures), or higher catalytic capacities [19].

Directed evolution consists in generating genotype variation from a natural protein, assaying the variants for a given function and selecting the best of them for a new variation selection cycle. Protein can also be designed computationally, but this technique leads to proteins characterized with low activities compared to natural proteins, and directed evolution has to be used to optimize them [21, 22] (see figure 1.4).

In some cases, directed evolution is limited: if the starting point of the directed evolution process is already "trapped" in a phenotype peak and the diversity generated is not large enough to escape it, the improvements will be limited. Also, proteins properties are often coupled: mutations that increase the activity might be destabilizing, mutations that increase the solubility might impair the protein's function [23, 24]. This phenomenon, termed pleiotropy, is not well understood and limits the optimization of enzymes. As a result, high-throughput methods to measure the phenotype of millions or proteins are required to screen the largest possible libraries in order to increases the chances of discovering suitable genotypes. Also, a better knowledge of the genotype-phenotype landscape would allow to design better strategies to improve proteins' properties without impairing their function.

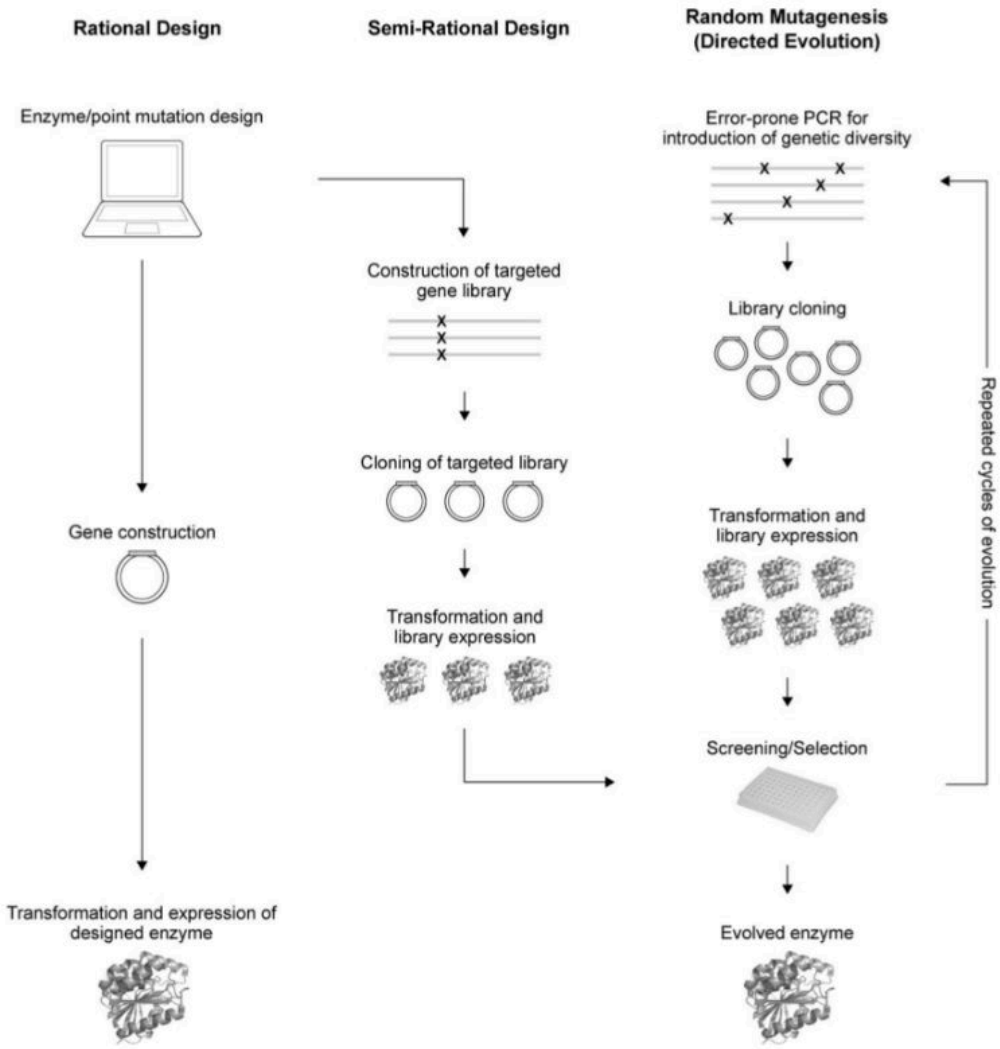


FIGURE 1.4: **Strategies to evolve proteins in laboratory experiments. Adapted from [19].**

1.1.5 Approaches to understand the genotype-phenotype landscape

In order to map the genotype to the phenotype and better understand mutational effects, three different approaches were developed:

Numerical approaches from the first principle of physics and protein structural information

The first approach consists in modeling the folding of known protein sequences using molecular dynamics [25, 26] to predict their 3D conformation. Molecular dynamics simulates the 3D motion of each residue in interaction with the others and within an energy potential. This approach is also used to predict protein's functions. As mentioned in [9], the success of those approaches has improved but is limited: structures cannot be predicted consistently and the binding affinity of small molecules to protein structures cannot be determined with enough accuracy, notably because of the computational cost of these simulations.

Other approaches such as Popmusic [27] or foldX [28] use structural information to compute proteins' free energy. They also predict the free energy destabilization upon mutations. However, predictions are moderately accurate ($\approx 60\%$) as mentioned in [29].

Statistical inference on protein sequence data

The function of folded proteins is dictated by a very large number of molecular interactions. Statistical methods are used to study their global properties and infer the parameters that govern their behaviour. The inference of those parameters leads to hypotheses regarding the distribution mutational effects and their influence on protein evolution. Those hypothesis can be tested by experimental genotype-phenotype mapping.

Large amount of homologous protein sequences are now available owing to the sequencing of full genomes via next generation sequencing [30]. Statistical inference methods can be performed on alignments of homologous sequences [31] to build mutational landscape models that predict mutational effects.

Figliuzzi et al. used the multiple sequence alignment of the β-lactamase enzyme family to measure the residue conservation and the conservation of correlation between pairs of residues (to capture second order epistatic effects which had been experimentally highlighted by [32, 33]). They use that information to infer a predictive model of mutational effects on β-lactamase enzyme TEM-1. The validation of such models requires the availability of quantitative genotype phenotype mapping data and is thus a motivation for my work.

The conservation of correlation between pairs of residues in the phylogeny of protein families is also used to identify networks of coevolving residues (see figure 1.5), forming groups of residues physically connected in the fold of the studied proteins, the sectors [34–37].

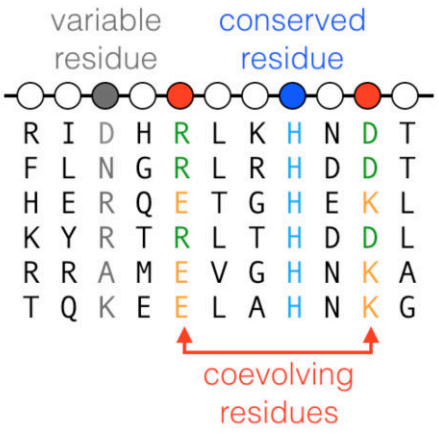


FIGURE 1.5: **Conserved and coevolving residues within an artificial multiple sequence alignment**. Adapted from [38].

Russ et al. and Socolich et al. [39, 40] showed that in the case of the WW domain family, the conserved pairs of residues were encoding the fold and the function. They identified the set of conserved residues and the set of conserved pairs of residues from the sequence alignment of the WW domain family. They generated and synthesized artificial sequences that were either random (i), respecting the uncoupled conservation pattern (ii) or respecting the coupled conservation pattern (iii). They tested the artificial proteins capacity to fold and to bind proline rich peptides which are bound by natural WW domains. None of the sequences from the groups (i) and (ii) folded. 28% of the sequences from the group (iii) folded as natural WW domains. 10 of them were tested and showed specifity and affinity comparable to natural WW domains. This suggests that pair-wise conservation contains sufficient information to specify the fold and function of proteins. These results also suggest that protein fold is only controlled by a sparse set of interacting residues.

Sectors were shown to mediate allostery in the case of HSP70 [41] and DHFR [35] domain families. Smock et al. identified a sector connecting the ATP binding site to the substrate binding site in the HSP70 protein family. The same residues were shown to disrupt allostery between the two domains upon mutation. In the case of the DHFR family, a sector connects the DHFR catalytic site to distant surface residues. A domain that changes its conformation upon photo-activation was fused to each residue at the surface of a member of the DHFR protein family. It was shown that its photo-activation was modulating DHFR's catalytic activity, preferentially when the photo-activable domain was fused to surface sector residues.

These results suggest that sector can be involved in the mediation of allostery.

In the case of the S1A serine protease family, sectors were hypothesized to be functional units [34]. In this family, three independent sectors were identified. The residues involved in each sector and preliminary mutagenesis data suggest that each sector is mutationally independent from the others (there is no epistasis between them, only within them) and controlling a particular feature of the enzyme: its stability, its specificity or its catalytic

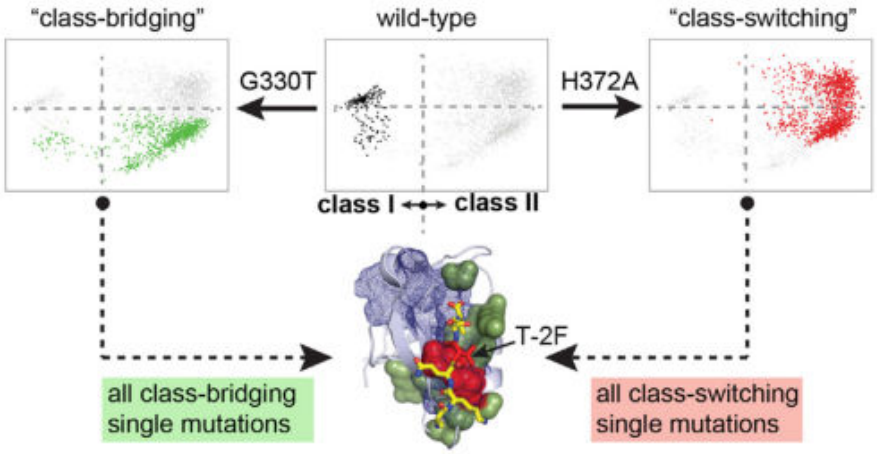


FIGURE 1.6: **PDZ specificity switch via class-bridging mechanism or class-switching mechanism.** The ligand sequence space of the PDZ domain can be projected according to two modes. The first mode (x axis) separates two different specificities: class I and class II. Evolution from class I to class II can be carried out by mutating red positions in the binding site ("direct switching") or by mutating green positions surrounding the binding site ("class bridging"). Class-bridging mechanism connects the two classes through a third specificity projected on the second mode of the ligand sequence space (y axis).The sector positions are depicted in blue. Adapted from [36].

efficiency.

Finally, in the case of the PDZ domain, sector positions are shown to be involved in the evolution towards new specifities. They were shown to be less tolerant to mutations than non sector positions [37] and two sector residues are shown to mediate a specificity switch of the PDZ between two different classes of ligands. Raman et al. [36] investigated this specificity switch and experimentally showed that a set of physically connected residues, connecting near active site residues to distal residues in an allostery-like pattern, controls this switching through two different mechanisms: a direct-switching or a class-bridging mechanism involving a third specificity that connects both previous classes in the ligand sequence space (see figure 1.6). This suggests that allostery is linked to protein adaptation.

Altogether, the statistical signature of the co-evolution between residues, raises testable hypotheses:

- Pair-wise conservation of correlation encodes fold and function in natural proteins.
- Sectors is a common signature of the coupling or uncoupling between proteins' properties. Sectors are structures which reflect a general mechanism to evolve allostery or towards new functions.

These hypothesis lead to the opportunity to considerably reduce the high dimensionality of the genotype-phenotype problem. But testing them requires high-throughput quantitative experimental approaches to systematically map mutational effects of single mutations and pairs of mutations.

Experimental approaches: the need for large scale genotype-phenotype mapping experiments

The computational approaches described above were performed using natural sequence data, which are the product of unknown walks through the sequence space, under unknown pressures of selection. Testing the predictions issued from these approaches requires large scale data, covering various regions of the sequence space, to reconstruct landscapes. The motivation for my thesis is precisely to develop high-throughput methods to generate these data and test these predictions. In turn, the data generated will also feed the computational approaches to refine their predictions.

High-throughput protein assay methods and now, DNA synthesis as well as Next Generation Sequencing technologies (NGS), allow to perform high-throughput genotype-phenotype mapping experiments termed "deep mutational scanning experiments".

1.2 Current high-throughput genotype-phenotype mapping approaches: deep mutational scanning experiments

Carrying out an experimental genotype-phenotype mapping consists in:

- (i) Generating a library of variant sequences where a chosen set of residues of a protein of interest are mutated.
- (ii) Measuring the phenotype of each variant through a protein assay.
- (iii) Sequencing its genotype.

The scope of genotype-phenotype mapping experiments can be limited by the capacity to mutate sequences, the throughput of the protein assay and the throughput of sequencing. A classical genotype-phenotype mapping experiment consists in measuring the mutational effect of each residue of a protein being independently mutated to alanine. In 1989, Cunningham et al. [42] measured the effect of 62 such mutations on the human growth hormone to map the binding epitope of the human growth hormone receptor. The scope of this experiment was limited to 62 residues mutated to alanine because each mutant had to be expressed, assayed and sequenced in separate experiments. Therefore, it was not possible to study subtle and heterogeneous perturbations offered by the diversity of available mutations towards amino-acids other than alanine, on all the residues constituting the human growth factor protein.

This is now reachable. Thanks to mutagenesis and sequencing developments, that can be interfaced with high-throughput protein assay, hundred of thousands to millions of protein variants can be studied simultaneously in what has been termed "deep-mutational scanning" experiments.

1.2.1 Generate sequence variants by mutagenesis

To produce a library of variant sequences, two possible methods that are based on Polymerase Chain Reaction (PCR) are available: random mutagenesis and site-directed mutagenesis.

Random mutagenesis As in a classical PCR reactions, primers flanking the 5′ and 3′ regions are designed so as to amplify the sequence of interest in a PCR reaction. But contrary to a faithful PCR reaction, the polymerase incorporates errors into the newly synthesized sequences of DNA. This can be achieved using a modified polymerase, adding destabilizing ions to the buffer of the reaction (manganese instead of magnesium for instance) or using an unbalanced concentration ratio of the four deoxyribonucleotides phosphate [22].

- **Advantages**: Many mutants can be generated in one single reaction. All of them are cloned at once, simultaneously.
- **Disadvantages**: It is a random process following a Poisson distribution: the average number of mutations can be controlled by varying the amount of template DNA, the number of cycles, the polymerase or the buffer conditions. But the number of mutations cannot be controlled at the level of the individual sequence. Mutations requiring two or three mutations on the same codons are very unlikely to occur. Hence, this technique is limited to amino-acid mutations characterized by single nucleotide mutations. The process is biased towards certain nucleotide mutations. This approach also generates non-sense codons which can be selected and enriched in the library if the protein is toxic for the host used for cloning.

Site-directed mutagenesis In site-directed mutagenesis, primers which contain the desired mutations are used to amplify the sequence of interest. The mutated amplicons are cloned in the desired vector.

- **Avantages**: The mutations are designed. Non-sense codons should not appear if not designed. The final library contains an arbitrary ratio of mutations, independently of their requirements in terms of nucleotide mutations.
- **Disadvantages**: This method requires at least, as many pairs of primers, as many PCR reactions as mutated positions. The different variant PCR products have to be mixed to achieve the final library.

Large-scale site-directed mutagenesis demands more work (more PCR) and is more expensive (more primers to order) than random mutagenesis methods. But new technical developments have made site-directed mutagenesis more competitive. Thanks to the increasing yield and efficiency of column-based oligo nucleotide synthesis methods. The price of DNA synthesis and its error-rate have dropped over the years to respectively reach 0.05-0.15 $ per nucleotide and 1 in 200 nucleotides [43, 44]. Besides, new methods to reduce the number of primers (costly for large genes) and to

make this process as efficient as possible [45, 46], have been developed. As a result, site-directed mutagenesis is now broadly used to perform uniform saturated mutagenesis of whole genes.

1.2.2 Opportunities and challenges of next generation sequencing

The development of next-generation sequencing was driven by the Human Genome Project [47]. In this paragraph, I purposely limit the scope of the discussion to the description of the performances of the Illumina sequencing technology which accounts for the most widely used short read sequencing platforms. A complete review of the next generation sequencing technologies that are now available has been carried out in [47].

Illumina Miseq v3 can deliver up to approximately 20 millions of forward and reverse 300 base pairs reads in one sequencing run. This is too short to sequence average sized genes in one single read (approximately 1 kilo base pair). Comparable technologies that could sequence millions of longer reads (such as Roche 454 GS) are characterized with 50-fold less reads and a 90-fold higher price per giga base [47].

Next-generation sequencing technologies are prone to sequencing errors. The Illumina Miseq v3 is characterized with an error rate of 0.1 %, which represents one sequencing error every three 300 base pairs reads. Sequencing strategies have to be adapted to be robust to sequencing errors. Different methods were already described. The same sequence can be read multiple times (using pair-end reads [48]), or short DNA tag can be added to the long variant sequences of interest, in the expression vector. Once the tag sequence is known for each variant, only sequencing the tag is required. As the tags are much smaller and very diverse, the analysis is very robust to sequencing errors [49]. Others have also added to the sequenced samples, wild type sequences that are not supposed to be mutated [50], in order to directly measure the error-rate in each sequencing run. Finally, "unique molecular identifiers" (UMI), short random DNA tags, can be used to identify different reads accounting for an identical sequence in order to produce a error-free consensus sequence [51].

Overall, as long as the sequencing process is designed to be robust to errors, next-generation sequencing gives the opportunity to sequence millions of variants genotypes in a few days.

1.2.3 Mapping the genotype to the phenotype in deep mutational scanning experiments

How can next-generation sequencing and mutagenesis be interfaced to high-throughput protein assays in deep mutational scanning experiments? What are the limitations of those techniques? What do they tell us about the structure of the genotype-phenotype landscape?

General considerations about deep mutational scanning

Once the desired library of variant sequences has been generated, the sequences are expressed into protein variants to be assayed. A lot of steps that cannot be easily automated using "high-throughput plate format" are required: expression, purification, assay and measurement. As a result, expressing and assaying each single mutant separately, is not scalable to large libraries (typically more than 10^3).

In deep mutational scanning experiments, all variant proteins are expressed and assayed simultaneously. The variant proteins displaying a phenotype higher than a given threshold are selected and extracted from the population. The genotype and the phenotype have to physically linked in order to be co-selected. Both unselected and selected libraries are sequenced separately using next generation sequencing. The comparison of the distribution of genotypes between the selected and the unselected libraries, allows to map the genotype to the phenotype. In practice, the frequency of each genotype is computed in the selected and in the unselected library and the ratio of those two frequencies, termed enrichment is used as a proxy to assess the genotype's phenotype for the assayed property (see 1.7).

Two different assay approaches are used to carry out high-throughput protein assay while keeping the necessary linkage between genotype and phenotype: the **in vitro approach** and the **in vivo approach**.

In vitro assay approach The in vitro approach comprises all the methods in which the protein is assayed outside of the cellular environment. The phenotype of the protein can be directly measured. It encompasses display methods such as: ribosome display [53, 54], phage display [48, 55–59], yeast display [50, 60–62] or mammalian cell display [63]. The protein of interest is expressed and kept at the surface of an object that contains the genotype information. This approach is mostly used to assay the stability and the binding affinity of proteins such as antibodies.

In the case of cell display, the cells are incubated within a solution containing the binding target labeled with a fluorescent compound. The cells are washed and selection is carried out by Fluroescence Activated Cell Sorting (FACS). The cell content is then sequenced. In the case of phage display or ribosome display, beads or a flat surface are coated with the binding target. After washing the fraction that bound is eluted and sequenced.

In vitro display methods are limited when it comes to study enzymes, as the genotype and the product of the activity have to be physically linked. With the methods presented above, the product of the activity would diffuse in the environment and phenotype measurement would be impossible. We will see in a next section that droplet-based microfluidics allow to overcome this limitation.

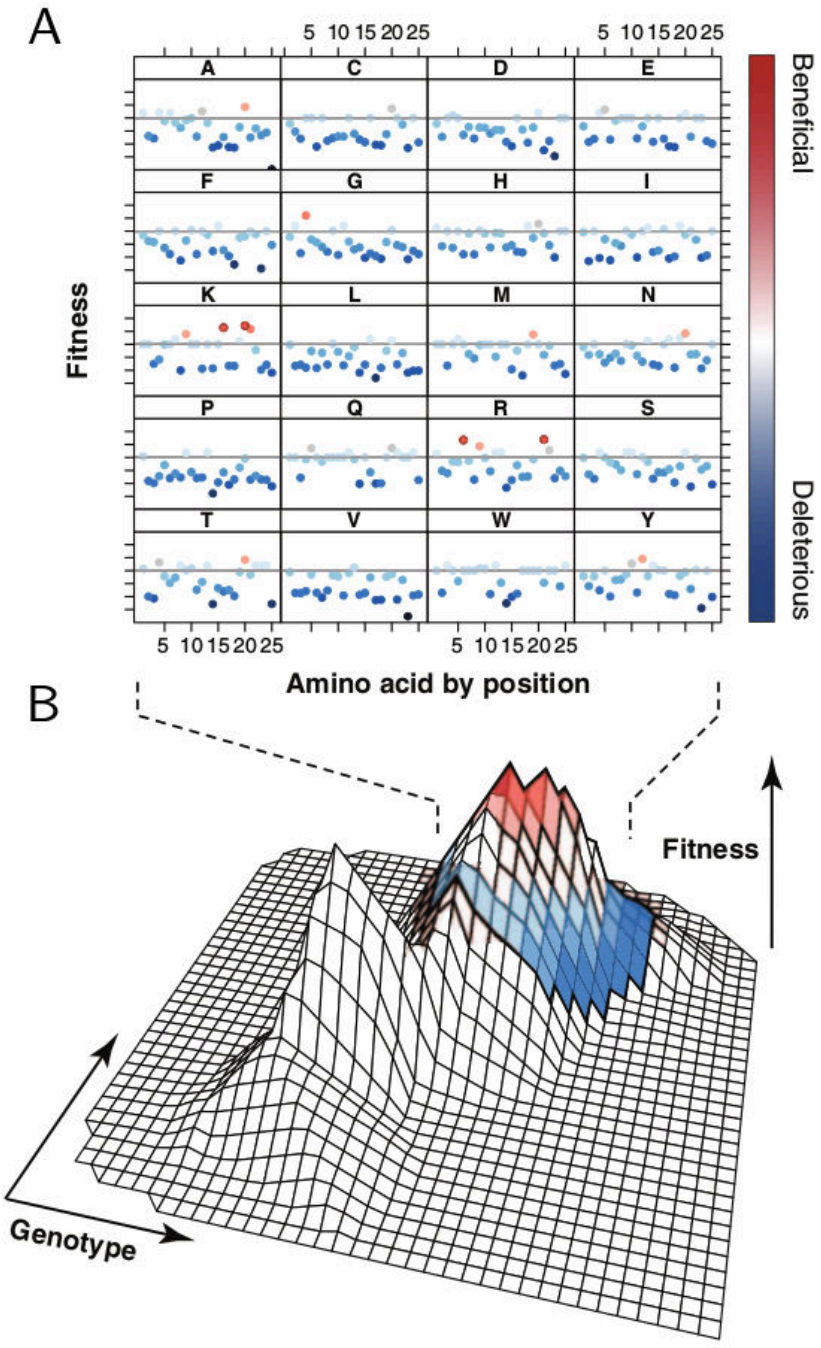


FIGURE 1.7: **Typical results of deep mutational scanning experiments. A** Typical results of a deep mutational scanning experiment where each one of 25 residues were mutated towards every possible amino-acids. Each square represents an amino-acid (from A to Y). In each square, each dot represents the ratio of enrichment between the selected and the unselected library at one position, for a mutation towards this amino-acid. Here 25 positions are represented. Red dots notify enrichment while blue dots represents depletion. Gray dots represents no enrichment.**B** The fitness landscape landscape is depicted according to the enrichment values illustrated in A. Adapted from [52].

In vivo assay approach The in vivo approach comprises all the methods in which the protein is assayed within the cellular environment. The in vivo approach is notably used in the case of enzymes as two physical linkages are required. The genotype has to be linked to the phenotype and the phenotype has to be linked to the product of its activity which is usually a hydrolyzed substrate whose concentration is measured by fluorescence or absorbance. Cells or phages and their target cells are providing a means to express the enzyme and to ensure both necessary physical linkages.

In this case, the fitness of the organism is used as a proxy for the variants phenotype. This method has been used to a large extent to study the β-lactamase enzyme TEM-1 [32, 33, 62, 64, 65]: a periplasmic protein responsible for the bacterial resistance against β-lactam antibiotics such as penicillin or ampicillin. In those experiments, cells transformed with the library of TEM-1 variants are incubated in presence of different concentrations of antibiotics. Resistant growing cells are sequenced and the TEM-1 genotypes are mapped to a range of minimum inhibitory concentrations (MIC).

In a method termed EMPIRIC [66], Hietpas et al. describe deep mutational experiments where yeast cells are transformed with a library of variant enzymes for which they are deficient. The cells are grown in liquid culture and the cells are sequenced at different time points to determine the growth-rate of each genotype [67–69].

Finally, in the case of phages, enzymes that are necessary for their replication and infection can be studied [70, 71]. In the extra-cellular medium is sequenced to assess the replication rate of the phages genotype.

In vivo approach are also used for recognition assay. McLaughlin et al., Raman et al. have engineered a quantitative two-hybrid system which triggers eGFP intracellular expression (sorted using FACS) or chloramphenicol resistance that can be selected for [36, 37]. Sarkisyan et al. [72] used *E. coli* to express and measure the fluorescence of GFP variants.

Limitations of those approaches

1. **How to relate frequency enrichment and thermodynamic properties?**: In general, the described mutational approaches do not give access to the absolute thermodynamic constants of the studied protein, such as the dissociation constant (K_d) or the catalytic efficiency (k_{cat} and K_M). The relative enrichment can give access to relative values, but the relationship between enrichment and the constants have to be determined. This has been studied in a few cases [37, 61]. This is of tremendous importance because the absence of absolute constant values hinders the comparison between different studies.

2. **The resolution on low phenotypes is poor**: In most cases, a single selection threshold is used. Selection is performed to select the top phenotypes providing a good resolution on the genotype-phenotype landscape for phenotypes which are close to the wild type in terms

of magnitude. But deleterious mutations are considerably depleted or even wiped out in the selected library. As an example, Roscoe et al. [68] failed to discriminate mutants with phenotype below 40 % of the wild type. As a result, if deleterious mutations can indeed be identified, their relative impact on the phenotype is not known. To quote Firnberg et al. [33] we know where the valley is, but we do not know what it looks like.

3. **Fitness is not the phenotype**: In the case of the in vivo methods, the link between fitness and proteins' phenotype is not trivial. The fitness is in general not a linear function of the phenotype [64] or even a monotonous function of the phenotype [70]. The relationship between the phenotype and the fitness has to be characterized in order to map the genotype to the phenotype.

4. **Mutations are more likely to appear neutral at high expression level**: In the case of in vivo methods, the expression levels are usually higher than in the native environment of the studied proteins. Mutations that appear to be neutral at high concentration reveal to be deleterious at lower protein concentration, because high total activity can vouch for the partial loss of function [33].

5. **Total activity is not the phenotype**: Except in a few cases [50, 61, 63, 72], the phenotype of the protein variants are not normalized by the expression yield. As a result, the phenotype is convoluted by the expression rate, which leads to false positive (negative) if a particular mutant expression level is high (low).

Overcoming those limitations for a quantitative genotype-phenotype mapping

1. **Droplet-based microfluidics to perform deep mutational scanning on enzymes**:

   Some of the constrains imposed by enzymes are now overcome thanks to droplet-based microfluidics (presented in the next section 1.3).

   Romero et al. [73] implemented, for the first time to our knowledge, a droplet-based microfluidics work-flow to perform high-throughput genotype-phenotype mapping of an enzyme. They encapsulated a library of β-glucosidase variants in micro-metric aqueous droplets (one variant per droplet) and performed the enzymatic assay within the droplets to measure the phenotype of the variants. Briefly, cells expressing the variant proteins within their cytoplasm are encapsulated in droplets (one bacteria per droplet at most) and lysed inside the droplets in the presence of a fluorogenic substrate. The fluorogenic substrate is hydrolyzed by the enzyme variants and the droplets are sorted according to their fluorescence in a way similar to FACS. The content of the sorted droplets is then sequenced to map the genotype to the phenotype.

   This solution allows to physically link genotype, phenotype and the

product of the phenotype. It is to be noted that one selection threshold was used to sort even low activities: a mutant displaying 17% of the WT activity was enriched 2-fold compared to a null mutant. The wild type was enriched approximately 30-fold compared to the null mutant.

2. **Selecting in multiple bins to improve the resolution of the lowest variants phenotype**

   Most of these works are depicting a bimodal distribution of fitness with two peaks [67]: a nearly neutral peak and a deleterious peak. As pointed out in the previous paragraph, deep mutational scanning methods usually fail to discriminate low phenotype mutants. More quantitative data on the TEM-1 $beta$-lactamase protein were generated by Firnberg et al. [33]. They engineered a "band pass filter" selection scheme which selects *E. coli* cells displaying a certain range of TEM-1 activity [74]. As a result, they can selectively sequence a narrow range of phenotype and discriminate between low phenotype mutants. Their methods even enabled to determine and discuss phenotypic differences between synonymous codons.

3. **Normalizing measurements by the expression yield**: To normalize the intracellular GFP variants fluorescence, Sarkisyan et al. [72] expressed it as a fusion protein with the fluorescent reporter mKate. Measuring simultaneously the mKate and GFP fluorescences, they normalize the measurements to obtain the phenotype of each variant. Such normalization was also performed in [61], by binding a fluorescent compound to a tag on the assayed antibody.

4. **Repeating the selection process to determine absolute thermodynamic constants**: In [61], Adams et al. repeated the selection process at different concentrations of ligand in order to generate binding curves. Doing so, they determined the binding affinity of a thousand of antibody variants simultaneously.

In spite of these limitations, which can be overcome as demonstrated by the few quoted examples, deep-mutational approaches are shedding a new light on the properties of the genotype-phenotype landscape.

1.2.4 Properties of the genotype-phenotype landscape as seen through deep mutational scanning

In this section, I will detail some deep-mutational scanning results. They are difficult to compare: phenotype or fitness are defined differently and the methods to determine the mutational effects are also diverse. Nonetheless, here is an overview of the insights that they provide on molecular evolution, through the study of robustness, epistasis and evolvability, in terms of statistical and structural distribution.

Distribution of robustness and agreement with the phylogeny

Are proteins robust to mutations? The results of those deep mutational scanning experiments show that the very large majority of mutations are

Protein and reference	Frequency of deleterious amino-acid mutations
WW domain [48]	97% (over 600 000 mutants mostly composed of single point mutants)
PhoP-PhoQ system [75]	More than 99% of the tested mutants not functional (all combinations of all possible mutations at four residues).
Methyl-transferase [76]	Between 61% and 83% of deleterious single point nucleotide mutations (averaged over different sequence backgrounds).
GFP [72]	75%
HIV protease NS3 [71]	73%
TEM-1 [33]	64% of all single point nucleotide mutants had lost 50% of the TEM-1 activity while 30% had lost 90% of the wild type activity (see 1.8).

TABLE 1.1: **Frequency of deleterious mutations according to deep mutational scanning experiments.**

deleterious and that robustness is not uniformly distributed in the 3D structure of the proteins. Studies also show that phylogeny conservation and even more, conservation of correlation between pairs of residues if a good predictor of the mutational sensitivity to mutations.

Robustness to mutations Table 1.2.4 summarizes an example of deep mutational scanning results regarding the share of deleterious mutations.

Mutational robustness is not uniformly distributed in the protein 3D structure. Buried residues are usually more sensitive to mutations than solvent exposed residues (except for the active site) [67, 77]. This property is used in the case of the bacterial toxin Ccdb [77] to infer a model predicting the degree to which a residue is buried as a function of its mutational tolerance. This model is then used to perform structure discrimination using mutagenesis experimental data.

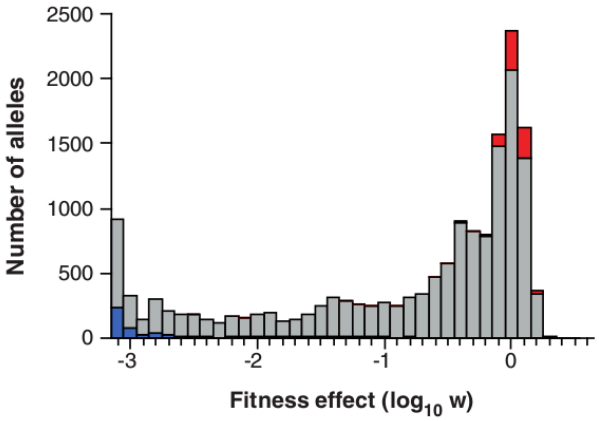


FIGURE 1.8: **Distribution of fitness of single amino-acid point mutation in TEM-1.** In red: synonymous codons. In blue: nonsense codon. In gray: missense codons. Here, fitness is a function of bacterial growth. Cells transformed with the TEM-1 library are selected for ranges of TEM-1 activity thanks to an engineered "band pass filter". This method enables to measure fitness over 3 orders of magnitude. Adapted from [33].

Residues within the active site or binding site are mostly intolerant to mutations and tolerance to mutations increases as residues are located further away from the active site. This distribution of mutational tolerance is found in other proteins such as the cetuximab (a therapeutic antibody) [63], ubiquitin [68], Aminoglycoside-3'-phosphotransferase[78], or TEM-1 [65].

The resolution for mutants with low phenotype is usually very poor in deep mutational scanning experiments. The example presented in figure 1.8, where fitness can be precisely measured over three orders of magnitude is not representative of most deep mutational scanning experiments. Most of the time, the relationship between the enzyme activity and the organismal fitness is not studied and the phenotype of the enzyme cannot be measured. There is a need for methods to directly measure protein's phenotype with high sensitivity, in order to understand the genotype-phenotype landscape of at low phenotypes.

Agreement with the phylogeny Is conservation correlated with mutational sensitivity? Even though some highly conserved positions are showing tolerance to mutations, conservation is in general a good predictor of the fitness or phenotype. This demonstrates that residues involved in proteins' function are conserved and that this conservation is not only due to inheritance, but also to an ongoing pressure of selection for proteins' properties.

Fowler et al. [48] showed that 10 of the 11 residues that are shared between the assayed member of the WW family and the family consensus sequence are intolerant to mutations and that mutations towards consensus amino-acid is enriched during selection for binding affinity.

In the case of HSP90 [79], indole-3-glycerol phosphate synthase (IGPS, TIM barrel) [67], the CH3-CH2 domain of humain IgG1 [50], and the GFP

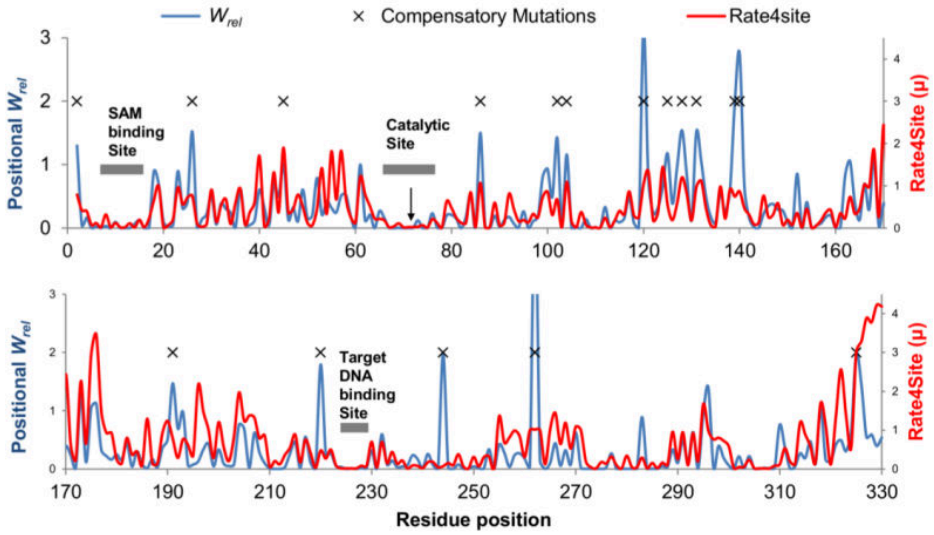


FIGURE 1.9: **Fitness distribution of the DNA methyl transferase M.HaeIII.** The fitness of a mutation is defined as the mutation enrichment between the beginning of the neutral drift and the end. In blue: experimental fitness of each position as the mean fitness over all mutations. In red: evolutionary rate of each position as predicted from the multiple sequence alignment of the enzyme orthologs (using Rate4site). Experimental fitness and evolutionary rate are in strong agreement. Adapted from [76].

[72], conservation correlates with low mutational tolerance.

It has been reported that positional entropy is more important in experiments than in the multiple sequence alignments [79]. But it was suggested, that this was due to the fact that misense mutations were considered as probably equal, whereas in nature single base pair misense mutations are more likely to occur, which biases mutations towards particular amino-acids at different positions [76]. To take into account this bias, Rockah-Shmuel et al. performed a systematic mapping of all single nucleotide mutations of the DNA methyl transferase M.HaeIII by random mutagenesis and neutral drift. They showed that experimental mutational effects at each position are in agreement with their rate of evolution and mutational effects as predicted by multiple sequence alignment tools (Rate4site and PROVEAN) (see figure 1.9). Agreement is improved by restricting the phylogeny to sequences more similar to the assayed "wild type" (more than 50% homology) [76]. In this work, the mutation sampling was indeed probably more "natural" than in other systematic studies. Unfortunately, the different regions of the methyl transferase were sequenced independently and the authors could only consider the total frequency of appearance of each mutation with no access to epistasic effects.

Finally, sector analysis of the PDZ domain and genotype-phenotype mapping of all its single point mutants showed that the most sensitive residues are better predicted by the sector composition than by conservation, distance to the core or by their direct interaction with the ligand [37]. This suggests that conservation of correlation between pairs of residues

might be a better predictor of mutational tolerance than conservation in general.

Distribution of beneficial mutations

How are distributed beneficial mutations? Mutations that improve the functions are rare and comprise residues located at the periphery of the active site.

By generating all possible amino-acid single point mutants in the variable regions of an antibody fragment, Fujino et al. identified the most enriched mutations upon selection for binding (approximately 10% of the mutations assayed): most of them occurred at the periphery of the binding site [54]. A combination of those mutations improved the binding affinity of the antibody fragment by 2000-fold.

Whitehead et al. generated de novo designed inhibitors (50 amino-acids) of the influenza hemagglutinin [60]. Deep mutational scanning identified beneficial mutations at the periphery of the binding site. Those are thought to be involved in long range electrostatic interactions which were not modeled in the de novo algorithm.

Interestingly, even on TEM-1 which is supposed to be a "perfect enzyme" under strong pressure of selection and functioning close the diffusion limit, Firnberg et al. identified beneficial mutations which accounts for approximately 7% of the single nucleotide misense mutations [33].

These results imply that evolving proteins might reveal more successful by mutating distal residues rather than mutating active site residues, which results in the impairment of function, as the robustness distribution suggests.

Distribution of epistatic effects

To what extent are mutations sequence dependent? As shown by the following examples, epistatic effects are rare among all possible pairs of residues, and mostly negative (the double mutant phenotype is lower than what could be expected from the two single point mutant phenotypes). This is mainly due to the fact that natural protein are marginally stable: two mutations which would be nearly neutral independently can together destabilize the protein and impair its function, leading to a large negative epistatic effect. In the following examples, epistasis is measured differently from one work to the other, making comparison difficults.

In the case of the IgG1 binding domain of the G protein (all single point and double point mutations, ribosome display), Olson et al. [53] showed that strong epistasis ($\epsilon > 1$ on figure 1.10) accounted for $\approx 4\%$ of all 500 000 double point mutants (which is consistent with [57] and [80]). Even if most single point mutations are deleterious, each one of them can be beneficial in the presence of one other mutation. A clusters of 12 positions interacting

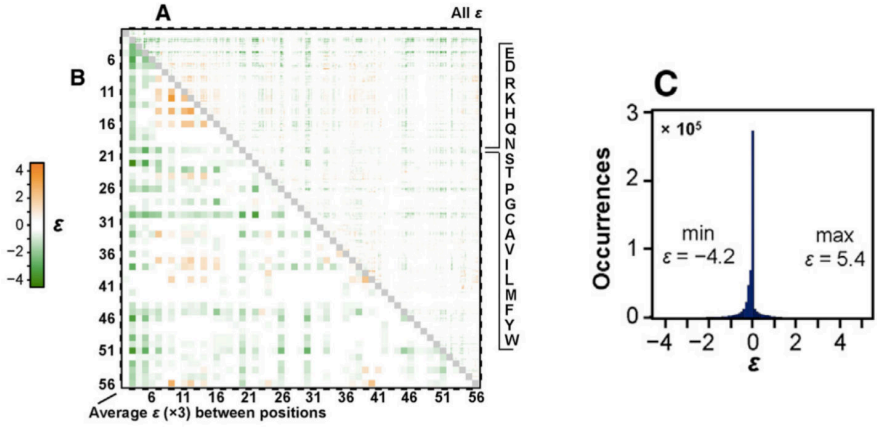


FIGURE 1.10: **Distribution of epistasis in the IgG1 binding domain of the G protein (56 residues).** ϵ represents the epistatic effet between each pair of positions. Negative epistasis is in green. Positive epistasis is in orange. **A**: The epistasis between each single point mutant is represented. **B**: The mean epistasis between each pair of residues is represented. Most of the epistasis is negative, except for the region surrounding position 11 where a cluster of residues interacting through positive epistasis has been identified. **C**: Distribution of epistatic effects. Most of the pair-wise mutations effect are additives. Strong epistatic effects are rare (defined here as ϵ>1). Adapted from [53]

through positive epistasis (orange cluster on figure 1.10) were also identified in a dynamic region which contains half of all positive epistatic interactions. Interestingly, the combination of two highly deleterious mutations at two residues that are co-evolving, leads to a near neutral double mutant.

But epistatis is not limited to pair-wise effects as epistasis between 3 (third order epistasis), 4 (fourth order epistasis) or more residues can be defined [81] and measured. Which epistatic order is the most relevant to study the genotype phenotype landscape?

In the case of the PhoP-PhoQ system, all combinations of mutations of 4 different residues leads to 1659 functional variants which represents 1% of all combinations [75]. This result is striking compared to the previous example as 99% of the combinations display strongly negative epistasis. But here, 4^{th} order epistasis is considered whereas in the previous example, only 2^{nd} order epistasis were considered. Also the phenotype is limited to "functional" or "not functional", depending on a single threshold set on a fluorescent reporter expression level. The assay is not quantitative and might hide a more complex distribution of epistasis. Also 4^{th} order epistasis is not compared to 2^{th} epistasis. More work has to be conducted in order to question the relevance of second order epistasis compared to higher orders. Obviously systematic studies are for now, limited by the high number of variants that would have to be tested in order to measure high order epistatic effects systematically.

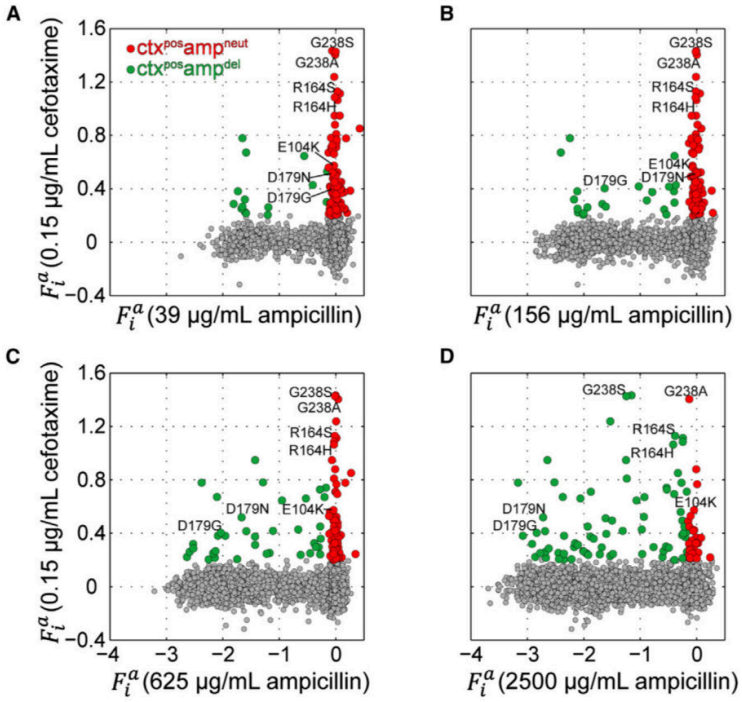


FIGURE 1.11: **Mutation effect of the single point mutants of TEM-1.** In each figure, x-axis: mutational effect under the mentioned ampicillin concentration, y-axis: mutational effect under the mentioned cefotaxime concentration. As ampicilin concentration increases, mutations that are beneficial against cefotaxime and neutral against ampiciline become deleterious against ampiciline. The evolvability reservoir towards cefotaxime resistance (in red) is depleted. Adapted from [64].

Influence of robustness on molecular evolution

Robustness is often depicted as an intrinsic property of the proteins: proteins would have a "stability reservoir", exhausted through the accumulation of mutations. Deep mutational experiments are showing that higher robustness allow to accumulate conditionally neutral mutations which act as a reservoir to evolve towards new function[53]. They also show that robustness does also depend on the pressure of selection under which proteins are evolving [64].

Stiffler et al. [64] have studied the environment dependence of robustness and its relationship with evolvability. They studied the effect of each single point mutations on TEM-1, the enzyme responsible for the bacterial antibiotic resistance against ampiciline and cefotaxime. Wild type TEM-1 degrades ampiciline but is weak against cefotaxime. Nevertheless, single point mutations that are beneficial to the resistance against cefotaxime were identified. At low ampicilline concentration, those beneficial mutations are neutral to the fitness under moderate ampicilline selection and could play a role in an evolution towards cefotaxime resistance. But under strong ampiciline selection, they are deleterious and could not accumulate in a drifting population (see figure 1.11).

Overall, this work emphasizes the role of neutral mutations, which can

appear and accumulate thanks to robustness, in the potential to evolve towards new functions. It also highlight the fact that higher pressures of selection lead to lower robustness which might lead to lower evolvability. Experimental evolution with transient selection pressures, is required to confirm those insights.

Influence of epistasis on molecular evolution

Systematic studies where the phenotype of all possible intermediary mutants between two different functions or specificities were measured, have shown that only a small portions of all possible evolutionary paths are characterized by an increasing fitness or phenotype [82–85]. The order of appearance of the mutations is not neutral, as in presence of epistasis, the effect of mutation a in the presence of mutation b is different from the effect of mutation b in the presence of mutation a.

In the case of the toxin antitoxin system ParE-ParD, 92% of the two substitutions evolutionary paths that enable to evolve between two specificities are epistatic [86].

In the case of PhoQ [75], 100 functional double mutants were identified. For 23% of them, both single mutants were functional, for 46% of them only one single mutant was functional and in the remaining 31%, none were functional. As a result, at least 77% of the two substitution path were identified as epistatic.

Epistasis is considerably constraining evolution in a reduced number of available evolutionary paths. Yet, controlled experimental evolution experiments such as the work presented in [87] highlights how irreproducibility can arise from epistasis and fluctuations due to limited population size.

The roles of promiscuity and evolvability in molecular evolution

Promiscuity is defined as the fact of bearing multiple specificities or functions. Promiscuity has been reported multiple times and in most cases the different functions are mechanistically close (sesquiterpene synthases [78, 88]. Cases of orthogonal promiscuity are more rare [89]. If a protein has to evolve from a function a to a function b, is it more favorable to switch directly, or is it necessary to go through a promiscuous intermediary step which can accommodate both functions? Deep mutational scanning experiments results suggest that natural evolution might favor promiscuous intermediates and that single nucleotide misense mutations are enriched in neutral mutations, favoring the accumulation of potentially adaptive mutations.

In [86], where the antitoxin ParD controls the deleterious effects of two orthogonal toxins, the network of ParD mutants separated by single point

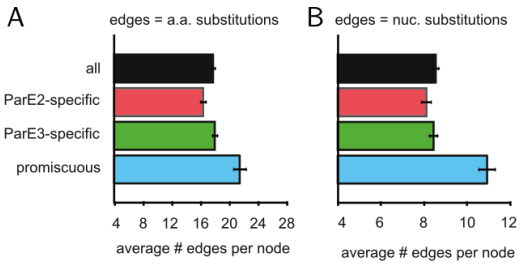


FIGURE 1.12: **Network properties of the functional variants of ParD**. The network of ParD functional variant was computed. Two variant (two nodes) are linked by an edge if they are different by **A** one amino-acid substitution or by **B** one nucleotide substitution. In black: all functional variants. In red: functional variants against the first toxin ParE2. In green: functional variants against the second toxin ParE3. In blue: functional variants against both toxins. Promiscuous variants are significantly more connected than "specialist" variants. Adapted from [86]

nucleotide mutations has been computed. On average, promiscuous mutants that can accomodate both toxins are 31% more connected than random mutants, suggesting that the genotype-phenotype landscape is favoring evolution through promiscuous variants (see 1.12).

Some results suggest that the genetic code might favor robustness and adaptive mutations. In the case of TEM-1 [33], it has been reported that 1 base pair misense mutations are on average less deleterious than 2 base pairs and 3 base pairs misense mutations. Hietpas et al. also showed that for the heat shock protein HSP90, single base misense mutations are enriched in functional mutants compared to random mutations [79].

There is a current need to test if completely artificially evolved but functional sequences, present the same distributions that seems to locally promote evolvability in natural sequences.

Trade-offs and couplings between proteins' properties

Proteins display different properties which are prone to be selected by natural selection: stability, processivity, specificity, catalytic efficiency in the case of enzymes, solubility, allostery, aggregation propensity, etc. Are those different properties coupled or uncoupled? Is it possible to modify those properties independently through single mutations, or are there trade-offs? Deep mutational scanning experiments probing those different properties sequentially or simultaneously have been performed. There seems to be a general trade-off between stability or solubility and the activity. Nevertheless, systematic approaches allow to find, in each case, mutations that only influence one property and mutations that are beneficial to several traits simultaneously.

Activity and solubility Active site residues are solvent exposed residues which are not optimized towards solubility [23, 90]. This underlies a possible trade-off between activity and solubility. Knowing that beneficial mutations that increase the catalytic activity are rare, Kelsmith et al. [62],

questioned the frequency of mutations that were increasing both solubility and catalytic activity. They used published data on the fitness conferred by TEM-1 against antibiotics and a deep mutational scanning approach to map the genotype of TEM-1 variants on their solubility (defined as the fraction of protein displayed at the surface of expressing cells). They showed the following:

- Negative coupling: 55 % of the mutations that are beneficial for the solubility are deleterious to the enzymatic activity.
- Positive coupling: They assessed that $\approx 30\%$ of the beneficial mutations confer an increased solubility.
- Those experimental data and the knowledge of the TEM-1 structure allowed Klesmith et al. to build a predictor of mutational neutrality. They showed that 90% of the mutations that (i) were extracted from the multiple sequence alignment of TEM-1, (ii) displayed a low number of contacts with neighboring residues (≤ 16) and (iii) were distant from the active site (≥ 15 Å), were neutral.

This definition of solubility integrates protein expression, folding and aggregation. Besides, in this example, the protein was expressed as a fusion partner which might have a behaviour different from TEM-1 alone. Finally the fitness is considered and not the phenotype. Nevertheless, this work confirms the trade-off between activity and solubility and highlights the existence and the distribution of residues which can promote both properties simultaneously.

Affinity and stability Koenig et al. [55] performed the saturated mutagenesis of the variable domain of a high affinity anti-VEGF antibody. They carried out deep mutational scanning experiments to measure both stability and affinity of the variants by phage display. They could decouple binding mutational effects from stability mutational effects: many inactive mutants were indeed strongly destabilized. Among others, they identified a residue distal from the binding region (25 Å away) that increased both stability ($T_m + 5C$) and affinity ($3 - fold$) upon mutation. Interestingly, this particular residue is frequently mutated in human antibodies. Structural investigations suggest this increase of affinity and stability is respectively due to a conformational change of the antibody light chain and a change of orientation of both antibody light and heavy chains.

The S1A serine protease family: are sectors independent functional units? Halabi et al. [34] performed statistical analysis ons the S1A serine protease family of sequences [34]. Three independent sectors were identified. These sectors are comprising residues involved in three different features of the protein: its catalytic efficiency, its specificity and its stability. Preliminary mutagenesis experiments performed on Rattus norvegicus trypsin (rat trypsin), suggest that each one of them is mutationally independent from the others. Thus, it was hypothesized that these sectors were functional units, i.e that the trade-off between the serine proteases stability, specificity

and catalytic efficiency would be limited.

Are sectors functional units? Rattus norvegicus trypsin (rat trypsin) is one of the two model systems for which we would like to develop a high-throughput genotype-phenotype mapping experimental approach. This model system will give me the opportunity to test this question by quantitatively measuring the mutational effects of all single points mutations on its catalytic efficiency, its specificity and its stability. I come back, in details, on the work of Halabi et al. in the last section of this chapter 1.4.2.

1.2.5 Performing genotype-phenotype mappings of enzymes is limited with current high-throughput approaches

Deep mutational scanning approaches yield interesting results regarding the properties of the genotype-phenotype landscape. But we are currently lacking high-throughput methods to quantitatively measure absolute thermodynamic constant values. Some efforts have been made in this direction for binding proteins [61], but not for enzymes.

Also, there is a need for methods to perform in vitro enzymatic assay to decouple fitness from catalytic efficiency. To my knowledge, Romero et al. [73] were the firsts to couple an in vitro enzymatic assay with a deep mutational scanning approach to perform a systematic genotype phenotype mapping of an enzyme. Thanks to droplet-based microfluidics, they could physically couple the genotype, the phenotype and the product of the enzymatic reaction. But, the product of the activity was not normalized by the expression level, which prevents quantitative measurements.

Thanks to its high-throughput and the possibility it offers to couple the genotype to the product of the activity, droplet-based microfluidics is very well suited to perform genotype-phenotype mapping of enzymes. I started developing high-throughput and quantitative droplet-based work-flows to measure the catalytic efficiency of large number of enzyme variants by simultaneously measuring their expression level and the product of their activity. The next section is dedicated to discussing the state of the art in terms of enzymatic assays in droplet-based microfluidics.

1.3 Droplet-based microfluidics to perform high-throughput enzymatic assays

In this chapter I will present techniques based on droplet-based microfluidics to perform high-throughput enzymatic assays. I will focus on the specificities of the droplet format, its advantages and the constraints it imposes. I will discuss those specificities in the light of the recently published microfluidics literature.

1.3.1 Droplet-based microfluidics: high-throughput manipulation of micro-metric reaction vessels

The opportunities offered by the picoliter droplet format in biology

Micro-metric emulsions are made of micro-metric aqueous droplets dispersed in an oil continuous phase. The study of the generation, the properties and the manipulation of such droplets belongs to the domain of microfluidics (also defined as the study of flows in micrometric systems [91]).

In biology, the use of those droplets as reaction vessels is of great interest, as it presents the following opportunities:

- It reduces the volume of samples and reagents by 10^6-fold (from milliliters to picoliters) and massively increases the number of experiments performed in parallel.
- Generating micro-metric size droplets out of a solution of genes, cells or organisms, allows to compartmentalize single objects. It is then possible to perform individual assays on each "monoclonal" droplet and to probe the phenotypic diversity of the population. In theory droplet size is only limited by the size of the object to encapsulate.
- In the case of enzymatic assays, droplets offer the opportunity to co-compartmentalize the enzyme genetic sequence, the enzyme itself and the product of enzymatic activity by preventing the diffusion of the product of the enzymatic activity.

The development of bio-compatible materials dedicated to microfluidics

Those opportunities and the development of miniaturization methods related to MEMS, such as soft lithography [92], led to the elaboration of techniques to easily prototype microfluidic devices (or "microfluidic chips") made of blocks of solid PDMS (poly(dimethylsiloxane)), bearing micrometric channels, stuck on glass slides. Immiscible oil and water can be injected in the channels of such devices and droplets can be generated by the shearing of the aqueous phase by the oil phase [93].

Picoliter aqueous droplets are metastable in the oil phase. To ensure that they do not coalesce (i.e fuse when in contact), it is necessary to add surfactants to the continuous oil phase. Surfactants are amphiphile compounds that gather at the oil water interface. They lower its surface tension and prevent the coalescence of droplets in contact by steric stabilization and by opposing the drainage of the layer of oil between them [94–96].

Applying droplet-based microfluidics techniques to biological questions requires a combination of polymer (e.g PDMS), oil and surfactant that are inert to biological reactions, or in other words bio-compatible. In fact, PDMS is bio-compatible. Perfluorinated oils are now widely used as continuous phases: they are hydrophobic, lipophobic, non miscible with water, they are inert to biochemical reactions and are soluble to gases which enables

cellular respiration in droplets [97]. Finally, fluorophile perfluoroblocks attached to hydrophylic PEG head groups were developed [98]. This class of bio-compatible surfactant stabilizes emulsion to the point where it is possible to thermocycle emulsions with limited coalescence [99].

Microfluidic devices to perform high-throughput operations

PDMS microfluidics chips, together with the right combination of fluorinated oil and surfactant, enable the generation of droplets highly monodisperse in size as well as their precise manipulation (see 1.13 for an illustration of a representative set of microfluidic devices to manipulate droplets). The following operations are described in the litterature (I limit the scope of the discussion to the techniques which are not involving valves or thermomechanical actuation (e.g [100]) of the PDMS):

- Droplets whose volumes span from femtoliters [101–103] to nanoliters [104] can be generated. The interested reader will find valuable information about the influence of the geometry of the generation devices on the size of the final droplets in the following litterature: [93, 105–109].
- It is possible to mix two different aqueous phases or more upon droplet generation [110].
- A small volume of reagents can be added to droplets (a fraction of their size) through injection (also called picoinjection [111]).
- Different populations of droplets can be fused one-to-one on demand using a triggered electric field (electro-coalescence) [112–115] or passively [116, 117].
- Droplets can be divided in several smaller droplets [118].
- Droplets can be trapped or anchored to rails [119].
- Droplets can be incubated on chip in delay lines [120] or static chambers[121].
- Droplets can be incubated off-chip [122], thermocycled [123] and reinjected into another microfluidic device.
- Droplet fluorescence [124], size, hydrodynamic resistance, [125–127], absorbance [128] and capacitance [129] can be measured. Droplets can be sorted according to those physical properties using dielectric forces [130, 131]. They can also be sorted using acoustic waves [132].
- Water droplets in oil can be encapsulated again, forming water in oil in water droplets ("double emulsion") which can be analyzed and sorted by Fluorescence Activated Cell Sorting technology (FACS). [133, 134].

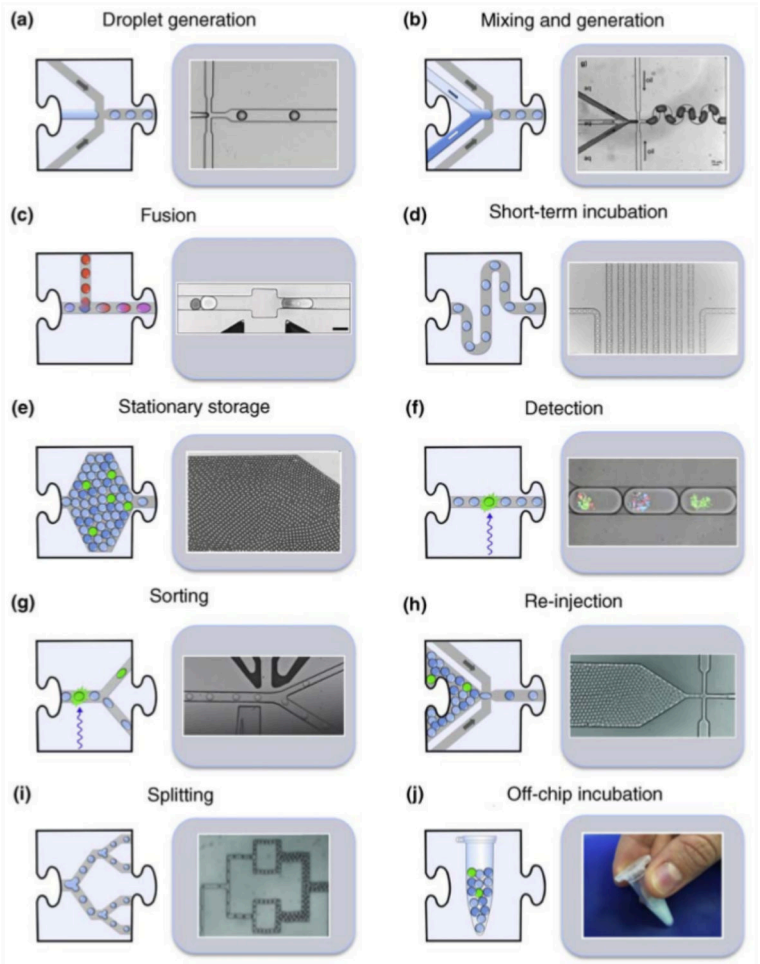


FIGURE 1.13: **A selected set of operations than can be performed on droplets. (a)**: Droplet generation, **(b)**: mixing of two different aqueous phases upon droplet production, **(c)**: one to one droplet electrocoalescence, **(d)**: short term incubation (seconds) in a single file without the loss of the droplet order, **(e)**: long term incubation in delay lines (minutes) or stationary storage (in both cases droplet order is lost), **(f)**: measurement of the droplet fluorescence, **(g)**: droplet sorting (here conditioned by the droplet fluorescence), **(h)**: droplet re-injection (after incubation off chip), **(i)**: droplet splitting, **(j)**: off-chip incubation or storage. Reprinted from [135].

Fluorogenic and chromogenic substrates are available for a large number of enzymatic reactions and can be used in microfluidic work-flows. The mentioned microfluidic techniques make it possible to, encapsulate enzymes in picoliter droplets, add a cocktail of substrates to each droplet and measure the enzymatic activity as a fluorescence or absorbance read-out (see figure 1.14). Given that droplets can be operated at kHz frequencies, the experimenter has access to the phenotypic distribution of millions of enzyme variants in a single day. Sorting devices and the FACS technology enable to extract selected droplets belonging to an arbitrary region of the distribution. If the enzyme genetic sequences can be recovered from within the droplets and sequenced, high-resolution genotype-phenotype mappings can be carried out.

To be able to establish such a mapping, droplet-based enzymatic assays have to be quantitative. Let me now define what I define as "quantitative" and discuss the constraints of the droplet-format.

1.3.2 Constraints of the droplet format on performing quantitative enzymatic assays

My experiments should allow me to measure the properties of millions of enzyme variants: their catalytic efficiency (ideally the Michaelis Menten constants k_{cat} and K_M should be measured), their stability (melting temperature) and their specificity (detection of more than one type of activity simultaneously using two or more substrates). Sorting droplets and sequencing their content should allow me to relate the phenotype to the genotype.

Fluorescence Activated Droplet Sorting Devices (FADS) [136] are operated under a microscope bearing a single objective: the experimenter can only focus and measure the droplets characteristics at one particular location of the microfluidic device at a time (at the sorting fork in this case). Droplets pass only once in front of this detection point. As a result and until now, high-throughput droplet-based enzymatic assays have been limited to one point measurement in the course of an enzymatic assay.

The fluorescence of each droplet at this detection point should provide enough information to extract at a high frequency the properties of the variants. The amount of information given depends on the three following quantities:

- The noise of the experimental measurement: how variable is the result of the measurement on one particular enzyme variant?
- The resolution of the assay: what is the smallest difference of activity that can be distinguished?
- The dynamic range: what is the amplitude between the lowest and the highest activities that can be measured?

Those quantities are influenced by the following:

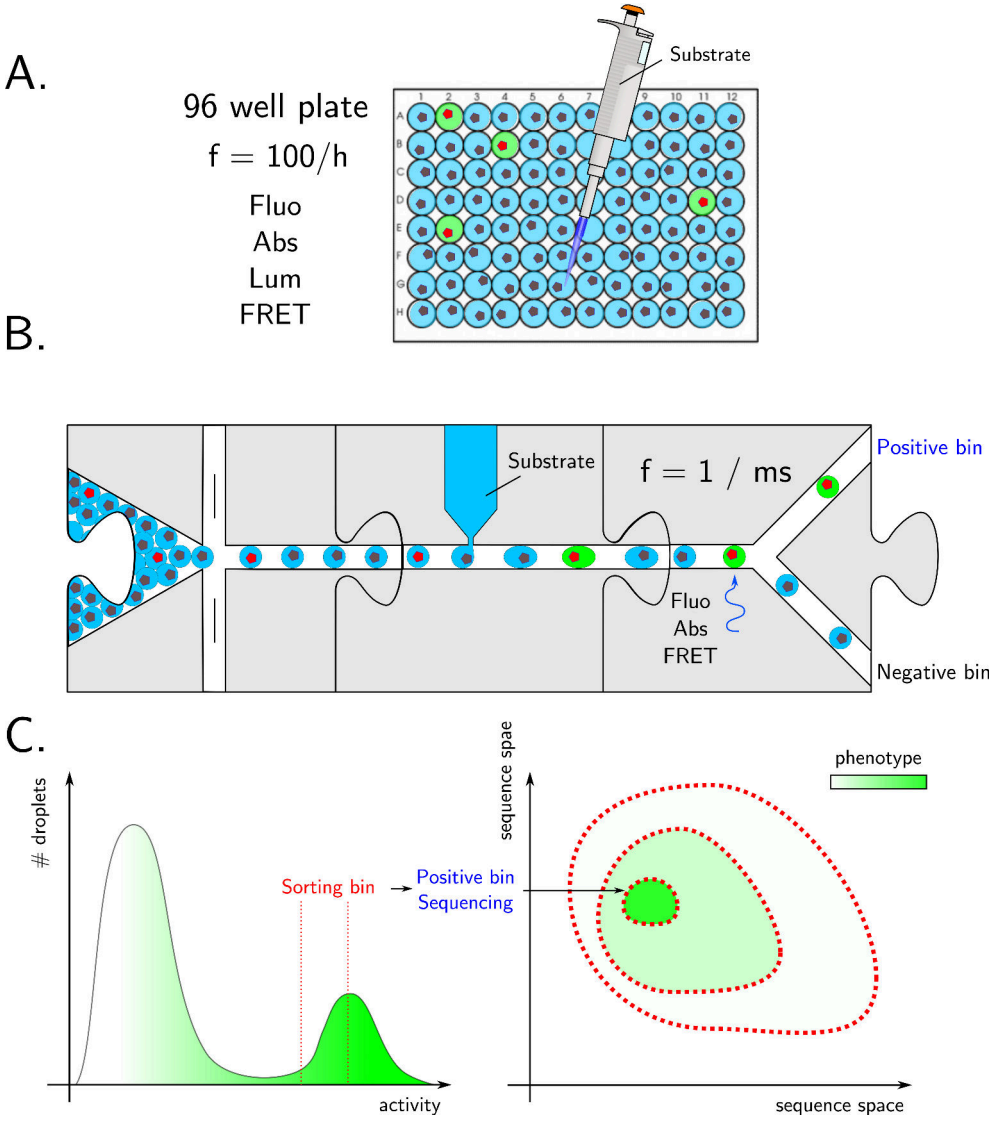


FIGURE 1.14: **Scheme of a high resolution genotype-phenotype mapping using droplet-based microfluidics**. **A**: Classical 96 well plate format, compatible with fluorescence, absorbance, luminescence and FRET measurement. Enzymes are represented as pentagons within the aqueous droplets. Active enzymes are red. Inactive enzymes are grey. Substrate is pipetted into the wells. **B**: Droplet-based microfluidc work-flow to perform an enzymatic assay in droplets. A substrate is added to each droplets by picoinjection. Droplets containing the active enzymes are sorted in the positive bin. **C**: Activity distribution of the droplets represented in B. All the droplets displaying a fluorescence contained in the sorting gate defined between the two red dashed thresholds are sorted in the positive bin. Their content is sequenced to establish the genotype-phenotype mapping (right).

- The microfluidic chip format and the droplet format impose specific constraints. Droplets are not solid reaction vessels: they can break, coalescence and have their content leaking in the continuous phase or the other droplets.
- The enzyme, the expression system and the substrate.

Let us discuss those issues and see how they are managed in the droplet-based microfluidic work-flows depicted in the recent literature.

General droplet-based microfluidics constraints

As suggested on figure 1.13, droplet manipulation techniques can be seen as modules. Those modules can be combined together in an integrated work-flow to encapsulate, assay, measure and sort biological objects. The safest and fastest route to manipulate emulsions is to integrate different modules on the same microfluidic device. Precisely manipulating the droplets (synchronize two populations of droplets to be paired one-to-one for instance), requires squeezing them in channels which have a size comparable to the droplets' diameter. The hydrodynamic resistance of a rectangular microfluidic channel is inversely proportional to the product of its length, width and height [91]. Thus, the hydrodynamic resistance of a channel increases quadratically with a decreasing droplet diameter. The number of modules that can be integrated on the same device is limited by the increasing hydrodynamic resistance of the channels. Also, controlling multiple microfluidic steps simultaneously can make a device difficult to control.

As a result, multiple steps are usually performed sequentially on different microfluidic chips. But emulsions are fragile and each operation generates partial loss of size mono-dispersity. Osmotic pressure, Ostwald ripening [94, 137], evaporation during incubation, splitting and coalescence due to collision with debris within the microfluidic channels are sources of poly-dispersity. This is a problem because mono-dispersity is critical to precisely control the droplets, but also to be able to compare them as identical reaction vessel. Thus the number of different operations that can be performed on one emulsion with different devices is limited.

Constraints specific to enzymatic assays

Each droplet has to be monoclonal[2] (1) so as to ensure genotype-phenotype linkage. As a result, the enzyme is either physically linked to its gene (via a cell, an organism, a microbead, etc.) or expressed in the droplet where its genetic sequence has been encapsulated (2). Finally a substrate (3) must be added to the droplets so that the product of the enzymatic activity can be measured (4).

[2]Even though multiple variants could be pooled in the same droplet to increase the throughput of the experiment, but I did not explore that possibility.

(1) The droplets should be monoclonal The gene to encapsulate is either in a plasmid, in a cell or in an organism. If the objects to encapsulate are not interacting, the encapsulation process follows a Poisson distribution. As a result, encapsulating at most one gene per droplet requires diluting the samples so that the encapsulation of two or more genes is limited in frequency (the average number of objects per droplet is typically $\lambda = 0.3$) [138–140]. As a result, most of the droplets are empty (if $\lambda = 0.3$, 74% of them are empty), which reduces the throughput of droplet-based experiments. Some devices were developed to overcome the Poisson distribution [141, 142]. They rely on sorting out empty droplets or ordering the objects to encapsulate in the flow.

(2) Choice of the expression system The choice of the expression system is of paramount importance in droplet-based enzymatic assays, as it conditions the nature and the number of microfluidic steps that will have to be performed. They are two different approaches: the in vitro and the in vivo approaches. In the in vitro approach, the enzyme expression is performed with purified components (RNA polymerase, rybosomes, tRNA, etc.) or cells extracts. In the in vivo approach a living organism that expresses the enzyme of interest is encapsulated. The table 1.2 gathers the advantages and disadvantages of each approach, according to the following criteria: microfluidic complexity, high surface to volume ratio of the droplet format, yield of expression, variability of expression level, preparation of variants library, toxicity, potential background activity, substrate access and gene recovery.

In a few words, about the choice of the expression system:

- **In vitro approach main advantages**. The yield of expression is expected to be relatively uniform between droplets and the experimenter might not need to use a reporter for the gene expression to decouple activity and expression [143]. Not limited by the toxicity towards the host. No need to transform a library of variants into the host. The enzyme is available to the substrate.
- **In vitro approach main disadvantages**. The gene might need amplification prior to expression. The expression system is less robust than in vivo expression systems and can be incompatible with the assay. These impose a microfluidic work-flows characterized with more steps which can be difficult to handle and detrimental to the emulsion mono-dispersity. The yield of expression is low with cell-free expression systems based on purified components, which might decrease the dynamic range of the experiment.
- **In vivo approach main advantages**. The gene and the product of its expression are encapsulated all at once. Expression yield are in general higher than with the in vivo approach. The microfluidic workflow is simplified and bears fewer steps.
- **In vivo approach main disadvantages**. Expression is variable, need for a reporter of the expression level to decouple the activity and the

Criteria	In vitro expression	In vivo expression
Microfluidic complexity	Genes are encapsulated. If the cell-free expression mix is not sensitive enough to yield detectable amount of enzyme from one single copy of the gene per droplet, the gene has to be amplified in the droplets, and the cell-free expression mix has to be added afterwards (amplification and expression cannot be performed simultaneously) [122, 143, 144]. If the cell-free expression mix is sensitive enough the work flow is simplified [121]. In some cases the substrate can even be added to the droplets at the same time as the cell-free expression mix ([143]). In other cases, the substrate has to be added later on the work-flow (see laccase enzyme [145]).	Gene encapsulation and expression can be performed in one step, simplifying the work-flow: expression is either intracellular and done before encapsulation ([73, 146–148], or extracellular and then expression is performed in the droplets (by surface display [149] or secretion [150, 151].
Influence of the high surface to volume ratio of the droplet format	The DNA and the components (RNA polymerase, ribosomes, etc.) of the transcription translation mix adsorb to water oil interface [152]. Additives such as yeast RNA, BSA, or pluronic have to be added to the water phase to limit this effect [101, 153]	Not a limitation.
Yield of expression	Very high yield ($\approx$ mg/ml) were reported from *E. coli* cytoplasmic extracts [154]. Cell-free expression systems based on purified components give lower yields (10 to 200 µg/mL^3) [155, 156].	Organisms such as *B. subtilis* were reported to express more than 10% of cell weight [157] ($\approx$ 10 µg/mL/cell in 10 pL droplets). It is possible to grow organism in the droplets while inducing for expression to obtain even better yields [158].
Post-translation modifications	Not possible.	Possible in eukaryotic expression systems such as yeasts.
Variability of expression level	In some cases, expression is uniform from one droplet to another and therefore activity products can be directly compared between droplets (Fallah Araghi et al. [143] reported a activity distribution coefficient of variation of only 5% for a monoclonal population).	Expression is variable from one droplet to another [140, 159, 160]. Organisms can show different plasmid copy numbers, phenotypic variability and history from one droplet to another can vary which might have an impact on microorganisms that are sensitive to oxygenation. To decouple expression and activity, the detected activity has to be normalized by a reporter for the expression level. Shim et al. [161] reported the use of a cistronic construct to co-express the alkaline phosphatase and a red fluorescent protein reporter in *E. coli*. The fusion protein was also tested but it was characterized by a phosphatase activity ten-fold lower than the regular protein.
Preparation of variant libraries	Libraries of enzyme variants do not need to be transformed.	If the gene of interest is not endogenous, it is necessary to transform the library of variants into a host, the competency of the host has to be taken into account (high for *E. coli*, lower for others such as yeats).
Toxicity	Not limited by the enzyme potential toxicity.	Limited by the enzyme toxicity towards the host. The least toxic variants are likely to be selected.
Potential background activity	More likely with cell-free expression systems based on cells extracts than with cell-free expression systems based on purified components [162].	More likely if the organism has to be lysed to release the product of the expression, as the cytoplasm might have a background activity. Enzyme deficient strains are sometimes available.
Substrate access	The expressed enzyme is directly available to the substrate.	If the enzyme is expressed within the organisms, the latter might have to be lysed [146]. In some cases, lysis is not compulsory [158]. If the enzyme is displayed at the surface (e.g horse radish peroxydase displayed on yeast cells [163]), secreted (e.g α-amylase secreted by yeast cells [151]), substrate access should be guaranteed.
Gene recovery	DNA can be directly recovered, amplified and sequenced or tested in a bulk transcription translation mix to validate the corresponding phenotype.	If the host is not killed in the process, it can be recovered, grown. Otherwise, the sorted DNA has to be transformed again [146].

TABLE 1.2: **Advantages and disadvantages of the in vitro and in vivo expression approaches according to the available literature.**

expression level. Need to transform the library of variants. Potential toxicity might be an impediment and introduce biases in the library. The substrate might not have access to the enzyme if it is not secreted or displayed at the surface of the host.

(3) Substrate properties and addition Small and hydrophobic [164] molecules can diffuse through the fluorinated oil or be exchanged by micellar transport mechanisms from one droplet to another [165, 166]. As a result, available fluorophores, which are generally small hydrophobic molecules, might not be suitable to perform quantitative measurement in droplets [167]. It is to be noted that neither proteins nor DNA are exchanged [168].

Several solutions or alternative are available to use fluorescence as a read out of droplet-based enzymatic assays. Modifications to make classical fluorophores more hydrophilic have shown to greatly improve their retention in droplets [167, 169, 170](see figure 1.15). FRET pairs to probe proteases activity have been designed[4] [171]: their size prevents leakage. Also, it has been shown that adding compounds such as BSA could improve retention in the droplets [172]. Alternatively, if the product of the activity exchanges slowly between droplets, short incubation times can be considered. In that case, measuring the activity of low activity enzyme variants might be compromised.

(4) Conditions on the signal detection

- The relation between the intensity of the detected signal at the detection point and the catalytic properties of the variants have to be characterized.
- In the case of the fluorescence, different signals can be detected in different channels characterized by different spectral windows (for instance to detect several activities simultaneously). As each droplet fluorescence profile is measured once only at the detection point, all the fluorescences are measured simultaneously. Fluorophores have emission spectra which stretch towards the red, which causes "cross-talk", i.e detection of a fluorophores in non specific channels. This can lead to false positives or hinder the detection of weak signals. Compensations coefficients or synchronous detection can be used to correct this effect.

Those constraints must be considered when designing a droplet-based microfluidic work-flow dedicated to enzymatic assays: to reduce the noise and increase the dynamic range and the resolution of the assay. I will finish this section by discussing how those constraints were handled in published droplet-based microfluidic work-flows dedicated to directed evolution, mutational scanning, strain or enzyme discovery and functional metagenomics.

[4] Those are made of a fluorophore/quencher pair linked by a peptide chain cleaved by the assayed protease activity. When the peptide is intact the fluorophore fluorescence is quenched by the presence of the quencher. When the peptide is cleaved, the pair is separated and the fluorescence of the fluorophore can be detected.

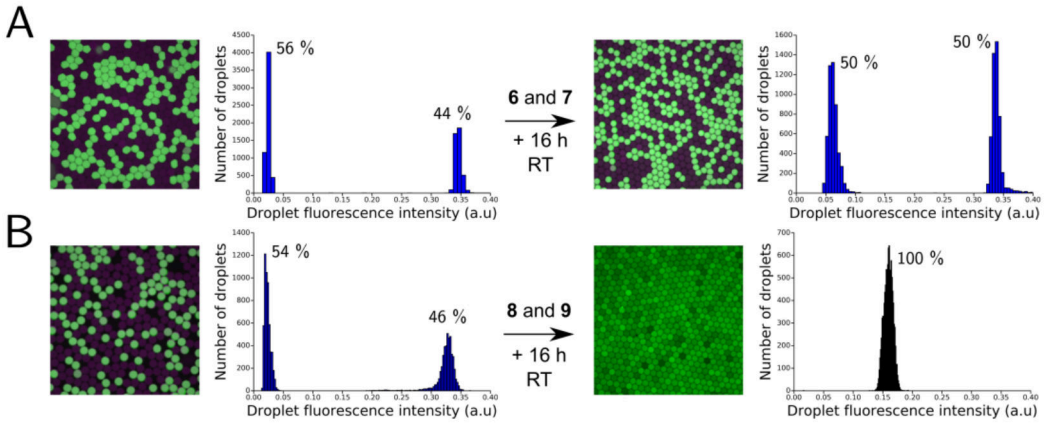


FIGURE 1.15: **Limiting rhodamine leakage using a hydrophilic tag.** Rhodamine is a typical example of a fluorophore that can be exchanged between droplets, at a speed which depends on the nature of the oil phase, the nature of the surfactant and the content of the aqueous phase. In each case, an emulsion containing the rhodamine substrate totally hydrolyzed is mixed with an emulsion containing the intact rhodamine substrate. The resulting emulsion is imaged on a fluorescence microscope and the fluorescence distribution is analyzed upon mixing and after 16h of incubation. **A**: Rhodmaine substrate with the hydrophilic tag. **B**: Rhodamine substrate wihtout the hydrophilic tag. Leakage is controlled thanks to the hydrophilic tag. Reprinted from [167].

1.3.3 Current droplet-based microfluidic work-flows

Operational picoliter droplet-based work-flows, as published in the microfluidics literature, are reported in the table 1.3.I limited the scope of the state of the art to works where the enzymatic assay dynamic range, resolution or noise could be at least, roughly assessed: the quantitative aspects of those work-flows are not always available.

Most of the work-flows presented here are dedicated to enzyme discovery or directed evolution of strains or enzymes. In those cases, the priority is to extract from the population the content of the most active droplets 1.16: be that a mutant strain characterized by high expression yield or a particularly active or thermostable enzyme. Even if the expression and the activity are not decoupled, it is expected to enrich the population with the most active mutants selection after selection. As we saw, this type of work-flow can also be applied to deep mutational scanning experiments [73], by sequencing sorted and unsorted libraries of variants.

Dynamic ranges and characteristic noise vary a lot from one work to the other. Enzymes, expression systems, measurement methods are different. Dynamic ranges span from a few fold to 50-fold, and noise is complicated to assess, mostly because in none of those works, the activity is decoupled from the expression. Nevertheless, noise is generally less important when the assay and the read out or integrated on the same microfluidic chip.

Expression approach	Ref.	Enzyme	Work-flow	Quantitative aspects
In vitro	[173]	β-galactosidase	Single gene encapsulation Gene amplification in the droplets off-chip. electro-coalescence with droplets containing the cell-free expression mix and the fluorogenic substrate. Expression and assay simultaneously off-chip. Reinjection and fluorescence measurement.	Monoclonal population: low noise and important dynamic range. Active droplets are 20-fold more fluorescent than inactive or empty ones.
	[143]	Idem	Idem	Wild type β-galactosidase versus inactive mutant: the most active droplets have a fluorescence intensity 8-fold the fluorescence intensity of the inactive or empty ones. Noise is important: one third of the active droplets show fluorescent intensity spanning from the non active ones to the most active ones.
	[144]	Ribozyme (X-motif)	The cell-free expression mix is replaced by a cell-free transcription mix.	No data available on a monoclonal population to assess the experimental noise dynamic range or resolution.
In vivo ***E. coli* (intracellular)**	[146]	Pseudomonas aeruginosa sulfatase	Intracellular expression. Single cells encapsulation with a fluorogenic substrate and a lysis buffer. Incubation off-chip. Reinjection and fluorescence measurement.	Wild type sulfatase versus 8-fold less active mutant: droplets containing the wild type sulfatase are on average 5-fold more fluorescent than the mutant after 7 min of incubation in delay lines (on-chip). The two populations are not overlapping. A third population could be distinguished between them.
	[147]	Sulfate monoester and phosphotriester hydrolases	Idem	No data available on a monoclonal population to assess the experimental noise dynamic range or resolution.
	[73]	Glycosidase	Idem	No data are available to assess the experimental noise on one monoclonal population. The phenotype of a subset of glycosidase mutants is assessed by determining their enrichment in the library after sorting with an arbitrary activity threshold to sort even low activities. A mutant displaying 17% of the WT activity was enriched approximately 30-fold compared to the null mutant.
	[148]	Amino-acid dehydrogenases	Intracellular expression. Single cells encapsulation with a substrate and a lysis buffer. The substrate is indirectly reduced by the catalytic activity to give the absorbent WST-1 formazan. Incubation off-chip. Reinjection and absorbance measurement.	The activity distribution of a monoclonal population is not available. But most active mutants display a fluorescence intensity more than 4-fold higher than the inactive variants (or empty droplets).
	[158]	CotA laccase	Single cells encapsulation. Growth and expression in droplets. picoinjection of a fluorogenic substrate. Incubation off-chip. Reinjection and fluorescence measurement.	Wild type laccase versus inactive laccase mutant: after 90 min of incubation active laccase droplets are on average four times more fluorescent than inactive ones. Active droplets fluorescence distribution is flat and spreads from 1 to 50-fold the inactive droplets distribution.
	[174]	Artificial aldolase	Single cells encapsulation with the substrate. Incubation on-chip. Fluorescence measurement and sorting.	No data available to assess the experimental noise or dynamic range of the assay.
In vivo **Yeast cell secretion**	[151]	α-amylase	Cells are washed. Single cells encapsulation in presence of a fluorogenic substrate. Incubation off-chip. Reinjection and fluorescence measurement.	Wild type α-amylase secreting cells versus non secreting cells: non secreting cells are displaying background activity. The two fluorescence distributions are overlapping and their average is different by a few dozen of percent.
	[150]	Heterologous enzymes Xylanase, cellobiase, protease.	Single cells are encapsulated. Growth and expression in droplets. Picoinjection of a fluorogenic substrate. Incubation on chip in delay lines (10 min of incubation). Fluorescence measurement.	The activity signal variability is low (CV from 3 to 15% depending on the activity) for a monoclonal population. The dynamic range is high: droplets containing the cells are on average 18 times more fluorescent than empty ones.
In vivo **Yeast cell display**	[149]	Heterologous enzymes (among them butyryl-cholinesterase)	Expression in liquid culture. Single cells encapsulation with a fluorogenic substrate in water in oil in water droplets. Incubation off-chip (3 h). Reinjection and fluorescence measurement (FACS).	Mutants corresponding to three different regions of the activity distribution are extracted from a population of butyrylcholinesterase mutants. When reintroduced into the microfluidic work-flow the three populations activitiy distributions are overlapping.
In vivo **Cyanobacteria**	[175]	L-lactate dehydrogenase	Single cells encapsulation. Growth and expression in droplets. Substrate picoinjection. Off-chip incubation. Reinjection and fluorescence measurement.	Three different cyanobacteria populations can be clearly distinguished: wild type and two others strains expression characterized with more important lactate expression levels. The distribution of the droplet fluorescence for one population span over the entire dynamic range. This is probably due to the expression level variability from one droplet to another.

TABLE 1.3: **Quantitative aspects of already published work-flow dedicated to enzymatic assays in picoliter droplets.**

My aim is to perform high resolution quantitative genotype-phenotype mapping experiments to eventually measure the catalytic efficiency of a large number of variants, in the spirit of [61]. In this prospect and given the presented literature, let me develop the requirement of the work-flow I developed during this PhD.

1.3.4 Conclusion: requirements for a quantitative genotype-phenotype mapping of model enzymes in droplets

In the prospect of a high-throughput quantitative genotype-phenotype mapping of enzymes in droplets (see figure 1.16) the following has to be considered:

- The droplets in which the assay is performed should be as small as possible to increase the sensitivity.
- The substrate should be added to the droplets once expression is at its maximum (and constant), to decouple expression and activity and to optimize sensitivity and sorting throughput of the assay.
- The substrate used in the assay should not leak, or the time-scale of the assay should be negligible compared to the time-scale of the molecular exchange.
- The expression system should display a background activity negligeable compared to the lowest activity level that is to be detected.
- If seconds (minutes) of incubations are needed, droplets should be incubated in a single file (delay lines) to limit the time dispersion of the assay between different droplets.
- Activity and a reporter for the expression level should be measured simultaneously at the detection point to normalize the activity by the expression level (in the spirit of [61, 72, 161]). If expression level is uniform, a reporter might not be necessary.
- If a reporter is used, it should be inert towards the variants' activity.
- I should be able to recover the content of the sorted droplets to interface our work-flow with next generation sequencing technologies, in order to establish the genotype of the sorted variants.
- Since I am limited to one detection point, determining the k_{cat} and K_M of the enzyme will require performing several experiments at different concentrations of substrate, in the spirit of [61].

In the last section of this introductory chapter, I will introduce the model enzymes and the fundamental questions we would like to address.

1.4 Studying model enzymes to better understand the link between genotype and phenotype

We chose to study two different enzymes: *Streptomyces griseus* Aminopeptidase (SGAP) and the *Ratus norvegicus* trypsin (rat trypsin).

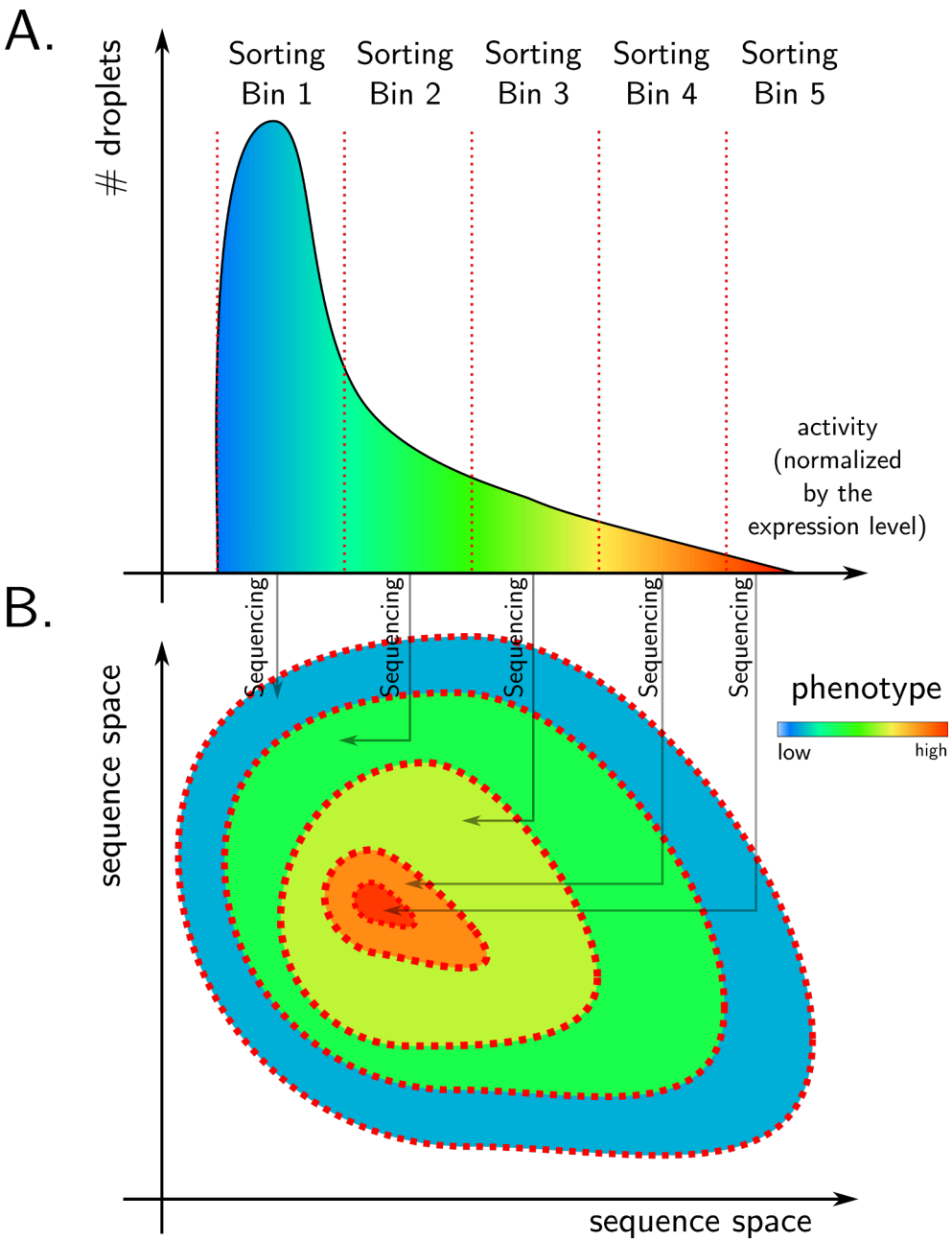


FIGURE 1.16: **Visualization and sequencing of the whole distribution of phenotypes. A**: In a directed evolution experiment, the red part of the distribution would be extracted, sequenced. Variation would be generated from the extracted variants to perform a new selection experiment. In a deep mutational scanning experiment, droplets would be sorted according to their fluorescence. The product of the sorting would be sequenced and the sequencing results would be compared to the sequencing results of the unsorted library. Here I propose to extract different parts of the distribution of activities normalized by the expression level in each droplet so that each sorting bin will be representative of a range of catalytic efficiency at a given concentration of substrate. Each bin will be separately sequenced and the genotype will be mapped on the phenotype.(**B**) Representation of the genotype-phenotype mapping.

1.4.1 *Streptomyces griseus* Aminopeptidase (SGAP) to study allostery and promiscuity

Streptomyces griseus Aminopeptidase (SGAP) is an ideal model system to apply genotype-phenotype mapping so as to study allostery and the mechanisms of enzyme promiscuity.

SGAP is a monomeric metalloenzyme (284 amino-acids) that cleaves the last amino-acid at the N terminus of polypeptides. Its structure is known (see figure 1.17) and its catalytic mechanism has been experimentally studied [176–190]:

- SGAP catalytic pocket natively binds two cations Zn^{2+}. They are involved in the catalytic mechanism as Lewis acids and activate the deprotonation of an embedded water molecule into an hydroxyde anion. This nucleophile hydroxyde anion attacks the carbonyl group of the last amino-acid to hydrolyze the peptide bond.
- SGAP preferably cleaves large hydrophobic residues such as leucine, but displays detectable activity towards other amino-acids (see figure 1.17).
- SGAP possesses a Ca^{2+} binding site away from its catalytic site (2.5 nm). Calcium binding modulates its activity, its specificities and increases its stability 1.17. This allosteric modulation has been studied but not systematically. [184, 185, 191].
- SGAP displays two other orthogonal activities: it catalyzes the hydrolysis of phosphoester and catechol substrates. [186, 188–190]
- The zinc cations can be replaced by other metallic cations (Co^{2+}, Cu^{2+}, etc.) to modulate those different activities.
- SGAP is a member of the M28 peptidase family (PFAM: PF04389, MEROPS: MER0002161) which contains more than 10^4 sequences. Those sequence are a valuable source of data to perform statistical analysis such as Statistical Coupling Analysis (SCA, [34]).

A systematic genotype-phenotype mapping of SGAP, for its different activities, could provide valuable insights about the mechanisms of enzymatic promiscuity and its role in the evolution towards new specificities or orthogonal activities. Besides, experiments at different concentration of calcium would reveal the residues involved in the allostery mechanism which modulates SGAP activities. Those data could be interpreted in the light of the phylogeny of the M28 peptidase family, through statistical analysis such as the already mentioned Statistical Coupling Analysis.

Dr. Alexei Godina had already started developing a droplet-based microfluidic work-flow in our laboratory, based on a cell-free expression system. I took over his work in order to complete a functional cell-free quantitative work-flow. I discuss my work on this topic in the Chapter III (3).

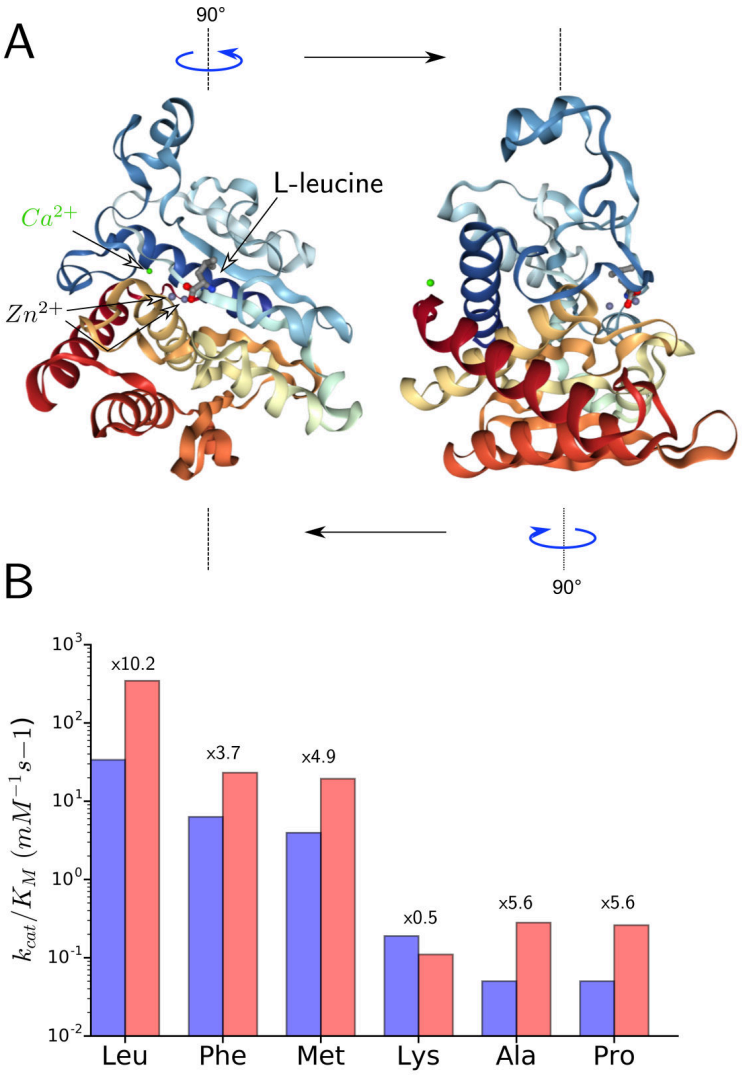


FIGURE 1.17: **SGAP structure and specificity. A**: SGAP crystallographic structure in complex with L-Leucine (PDB 1F2O). In gray: Zn^{2+}, in green: Ca^{2+}. **B**: SGAP catalytic efficiency on different L-Xaa-*p*NA substrates. In blue: without Ca^{2+}, in red: with 10 mM Ca^{2+}. SGAP displays aminopeptidase activities, spanning over 3 orders of magnitude, depending on the N terminus amino-acid. Calcium increases the activity with different ratio, except in the case of lysine where catalytic efficiency is divided by two. Data from [185].

1.4.2 *Ratus norvegicus* trypsin (rat trypsin). Are trypsin sectors independent functional units? What is the relation between the trypsin sectors and epistasis?

Ratus norvegicus trypsin is a member of the S1A serine protease family which has already been studied in [34] where it was chosen as a model system to study the sector hypothesis: are trypsin sectors mutational independent function units? We chose this model system to address this question systematically.

The serine proteases Serine proteases are characterized by a catalytic triad Asp-His-Ser where the serine is used as a nucleophilic center to attack carbonyl groups in the polypeptide chain [192]. This mechanism is thought to have evolved independently several times, as it is found in different families of proteases that do not share the same fold [193].

Among the serine proteases, trypsin is a classical biochemistry model system. It is well characterized, abundant (usually purified from bovine or porcine pancreas) and many substrates are commercially available.

Ratus norvegicus trypsin is a member of the S1A serine protease subfamily (chymotrypsin A like serine proteases), which is the largest subfamily of serine proteases (40596 sequences according to the MEROPS database[5]). Inside this subfamily, chymotrypins and trypsins display different specificities in spite of similar structures. Chymotrypsins cleave polypeptide on the carbonyl side of large hydrophobic residues whereas trypsins prefer lysines or arginines residues. Switching from trypsin to chymotrypsin specificities is not trivial: it has been shown that it not only requires mutating residues in the binding pocket but also in surface loops which are not directly involved in catalysis [194–196]. But the systematic measurement of the specificity of all possible mutants separating two representative trypsin and chymotrypsin has not been performed yet. The mechanisms of molecular adaption from trypsin to chymotrypsin (and vice versa) are therefore unknwown.

Consequently, the S1A serine proteases and more particularly the rat trypsin, are models of interest to study the mechanisms of molecular adaptation to new specificity. Besides, statistical analysis performed on the sequences of this family by Halabi et al. [34] raised hypotheses regarding the structure of their genotype-phenotype landscape. I present those hypotheses in the next paragraph.

***Ratus norvegicus* trypsin: a model system to study the independence of biochemical properties through the "sector hypothesis"** Halabi et al. [34] built a multiple sequence alignment of the S1A serine proteases (trypsin) family and measured the conservation of correlation between each pair of residues in the alignment. This statistical analysis (termed SCA for "Statistical Coupling Analysis"), revealed a decomposition of the proteins into

[5] http://merops.sanger.ac.uk/cgi-bin/famsum?family=s01c

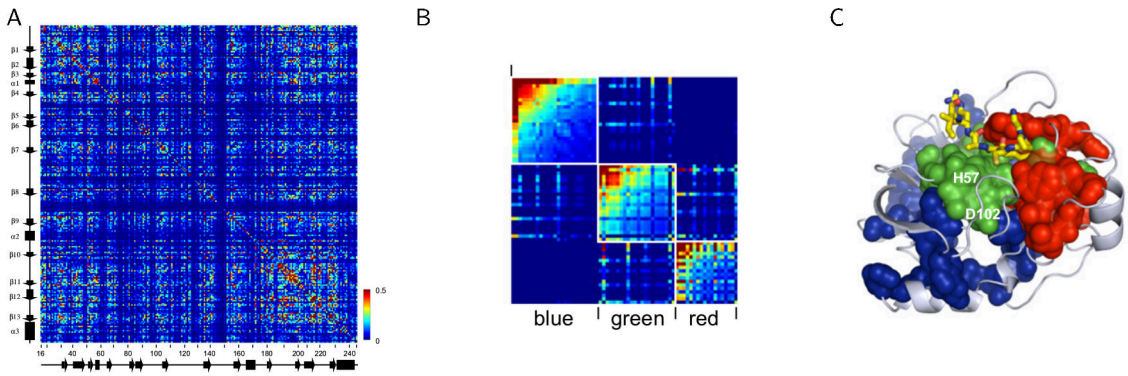


FIGURE 1.18: **Statistical Coupling Analysis (SCA) reveals the existence of three independent groups of co-evolving residues. A:** SCA matrix of the trypsin family (1470 sequences). Each coefficient of the SCA matrix is a measurement of the conserved correlation between pairs of positions in the rat trypsin polypeptide sequence. The subset of pairs of position displaying significant contributions ($\approx 10\%$ of all positions) can be clustered in three independent groups (see **B**), each one of them defines a sector (blue, green, red). **C:** Sector positions are represented in the 3D structure of the rat trypsin protein (PDB 3TGI). Each sector is made of contiguous residues involved in different biochemical properties of the protein. The red sector comprises residues flanking the serine catalytic pocket and involved in trypsin specificity. The green sector comprises residues directly involved in the catalysis mechanism such as the catalytic triad. The blue sector comprises residues distributed within the protein core, extending to the catalytic pocket. Hence the hypothesis that the red, the green and the blue sectors are mutationally independent functional units respectively mediating the specificity, catalytic efficiency and stability in the S1A serine protease family.

three quasi-independent groups of correlated residues that they called sectors. Further investigations showed that the residues within each sector were physically connected in the tertiary structure of a representative member of the trypsin family: the rat trypsin protein. Each of the sectors embeds residues involved in one particular feature of the trypsin protein: its stability, its substrate specificity or its catalytic efficiency. Halabi et al. hypothesized that the sectors might not only be a decomposition in co-evolving subsets of residues but also a decomposition in mutationally independent functional units 1.18.

To begin to test this hypothesis, the activity and the stability of a few single and double points mutants were measured for each one of the sectors 1.19. In agreement with their hypothesis, each of these sectors displays mutational independence (i.e absence of epistasis between two residues belonging to two different sectors) and bears a distinct property (e.g a mutations in the stability sector lowers the stability of the protein but has no effect on its specificity or its catalytic efficiency). Testing this hypothesis systematically requires measuring the phenotype (activity, substrate specificity and stability) of all single point mutants of the rat trypsin protein ($223 \times 19 = 4237$ mutants).

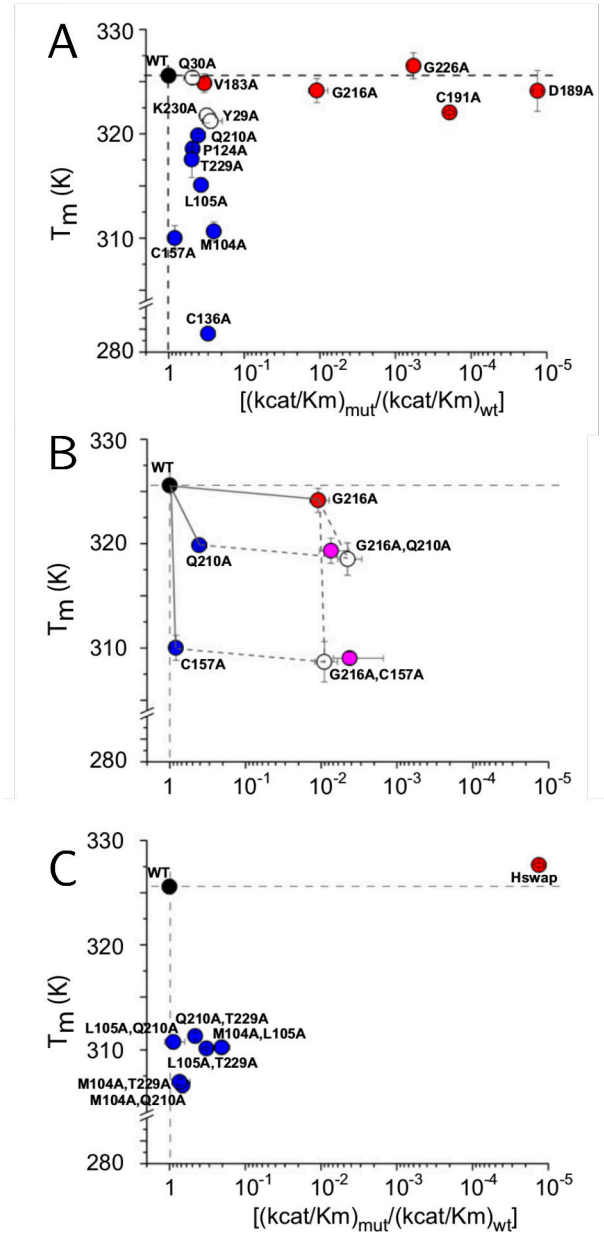


FIGURE 1.19: **Mutational effects of a few single point and double point mutations on rat trypsin stability and catalytic efficiency. A**: Mutational effects of single point mutations. Mutating the red sector mostly impairs the trypsin protease catalytic efficiency. Mutating the blue sector mostly impairs the melting temperature. **B** Mutational effect of a few double point mutants. For two mutations in two different sectors, the expected phenotypes (in absence of epistasis, white dots) are close to the measured phenotypes (purple), which suggest the sectors are mutationally independent. **C:** On the other hand, double mutants (or multiple, Hswap) within one sector display specific, non additive effects. Those preliminary results suggest trypsin sectors behave as mutationally independent functional units.

The rat trypsin is therefore an ideal model enzyme to develop a quantitative high-throughput genotype-phenotype mapping work-flow, that will bring the data to test this sector hypothesis. Our system could even be used to measure the phenotype of all double point mutants of the rat trypsin protein to reconstruct networks of positions interacting by epistasis and question its relationship to the sector structure. This second step would require measuring the phenotype of $(223 \times 19) \times (222 \times 19) \times 1/2 \approx 9.10^6$ mutants.

We started a collaboration with the Green Center for Systems Biology (Ranganathan et al. UT Southwestern Medical Center) to develop a microfluidic work-flow dedicated to the genotype-phenotype mapping of the rat trypsin protein. They had already set up an in vivo expression system in *E. coli* I could make use of. The development of this in vivo microfluidic work-flow is discussed in Chapter IV (4).

Chapter 2

Material and methods: droplet-based microfluidics and fluorogenic substrates

This chapter is dedicated to the material and methods related to droplet-based microfluidics. In a first section, I will present the tools and methods I used to produce and manipulate picoliter droplets, as well as the optical setup used to excitate and detect fluorescent compounds in the droplets. In a second section, I will briefly describe the fluorogenic substrates used to carry out enzymatic assays on *Streptomyces griseus* Aminopeptidase and *Ratus norvegicus* trypsin. The details of the experiments presented in further chapters are systematically detailed in the associated appendices.

2.1 Droplet-based microfluidics

2.1.1 Microfluidic devices fabrication

Here I present the protocols that I followed to fabricate the chips used in the microfluidic experiments, from the mask design to the preparation of the microfluidic setup.

Mask design All microfluidic devices presented in this manuscript were designed using the industrial design software Autocad (Autodesk, inc). The mask were printed at Selba S.A (Switzerland), at a resolution of 25400 dpi. If the design were presenting features smaller than 7 µm, resolution was improved to 50800 dpi. In the printed mask the inside (ouside) of the channels were set to be transparent (black).

Soft-lithography protocol to produce master wafers (figure 2.2) A photoresist SU8 (MicroChem Corp.) resin is spin-coated on a silicon wafer, forming a homogeneous layer whose thickness depends on the speed of sspin-coating and on the viscosity of the resin. The resin is then partially fixed on the wafer by heating according to the MicroChem Corp. protocol. The printed mask are used to perform soft lithography on silicon wafers using a MJB4 mask aligner (SUSS MicroTec) set with a UV lamp ($\approx 12mW$), a UV filter (365 nm) and an appropriate illumination time: the resin is illuminated through the mask where the channels should be located. After more heating steps following the manufacturer protocol, the wafer is

SU8 thickness (µm)	Resin	Spin 1 (RPM)	Time 1 (s)	Acceleration 1 (RPM/s)	Spin 2 (RPM)	Time 2 (s)	Acceleration 2 (RPM/s)	Softbake 65°C	Softbake 95°C	Exposure time (s)	Postbake 65°C (min)	Postbake 95°C (min)	Final bake
10	Su8 2010	500	10	200	2600	30	300	1	3	15	1	4	2 min 200°C
20	Su8 2015	500	10	200	1750	30	300	1	3	18	1	5	2 min 200°C
70	Su8 2075	500	10	200	3000	30	300	3	8	50	3	8	2 min 200°C

FIGURE 2.1: **Typical soft-lithography protocols**.

washed with an organic solution (Propylene glycol monomethyl ether acetate) to wash away the resin which was not illuminated, leaving only the negatives of the microfluidic channels on the wafer. The quality of the wafer is controlled by optical or mechanical profilometry. If the result is satisfying, the wafer is washed with isopropanol and set for one last heating step in order to fix the resin on the wafer.
For wafers characterized with multiple layers, this protocol is repeated several times on the same wafer, layer after layer. References (crosses or other motifs) are designed on the masks in order to allow alignment of the wafer with the mask corresponding to the next layer. Before spin-coating the wafer, the references already on the wafer are protected with band tape in order to keep the references visible during the alignment.

Protocols for typical resin thickness are available in figure 2.1

Fabricating PDMS chips Sylgard 184 silicone elastomer is mixed to sylgard curing agent in 10:1 w/w proportions. The mixture is degassed by centrifugation and in a vacuum chamber to get rid of the air bubbles. The degassed mixture is poured on a clean wafer (45 g per wafer to obtain an appropriate PDMS thickness) bearing the microfluidic design required. The wafer is set to incubate at 70°C for at least 2 h. After 2 h, the solid PDMS is gently peeled off the wafer. Inlets and outlets are pierced using a 0.5 mm biopsy (Uni-core), attacking the PDMS from the printed side. Isopropanol is flushed, from the non-printed side, through the holes in order to get rid of trapped PDMS particles. The inside of the holes are then dried using dry compressed air, from the non-printed side. In parallel, a glass slide[1] is washed with water and dish-washing soap. The slide is rinsed with water and isopropanol, and put to dry at 70°C for 5 min. Both punched PDMS slab and glass slide are treated by oxygen plasma exposition (Diener electronic) and the printed side of the PDMS slab is put in contact with the glass slide to seal the microfluidic device. Finally, a fluoro-silane solution compound (perfluorododecyl- 1H, 1H, 2H, 2H-triethoxy-silane, SigmaAldrich) is diluted 100 times in fluorinated oil (HFE 7100, 3M) and the resulting solution injected into the oil inlet of the chip in order to coat the inside of the channels to make them hydrophobic. Channels dedicated to electrodes (if any) are filled with 51In/32.5Bi/16.5Sn (Indium Corporation) after heating the chip at 120°C on a heating plate. Electric wires have their ends stripped (typically half of the thickness of the microfluidic chip) and are inserted in the electrodes inlets while the chip is still on the heating plate.

[1] In case the microfluidic device should bear electrodes, an indium tin oxide coated glass slide (Delta Technologies) is used instead. The coated side should face outwards the microfluidic chip.

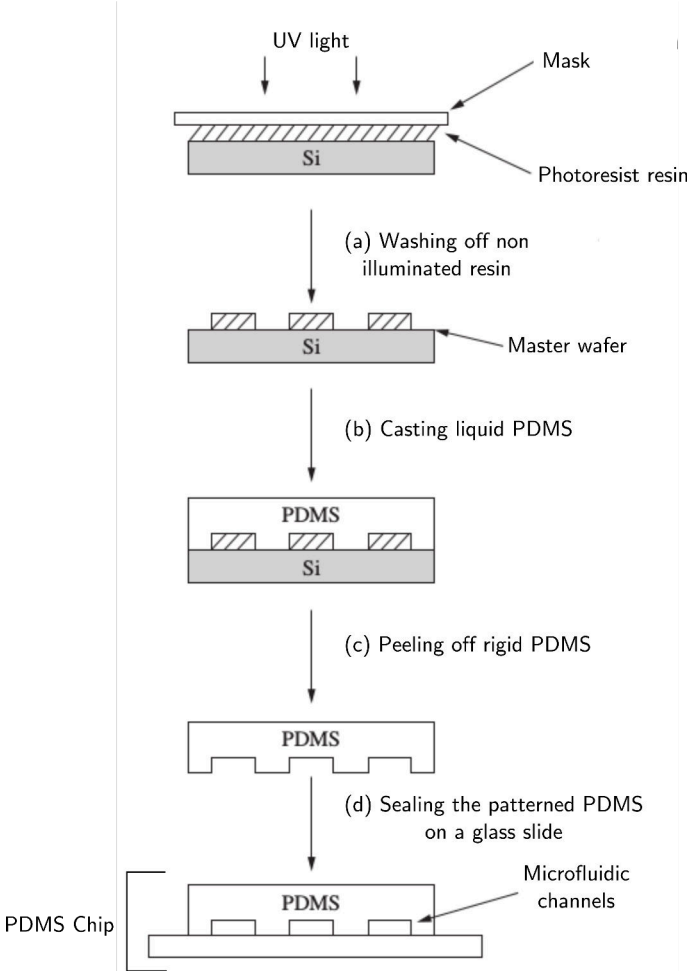


FIGURE 2.2: **Soft-lithography and PDMS chip fabrication**. **(a)**: Soft-lithography process. The master wafer is patterned with resin forming negatives of the microfluidic channels. **(b)**: Liquid PDMS is mixed with curing agent are poured on the master wafer and heated at 70°C for 2 h. **(c)**: Rigid PDMS is peeled off the master wafer. **(d)**: After plasma treatment, the PDMS slab is sealed on a glass slide. Adapted from [92].

Soft-lithography and the PDMS chips fabrication process are summarized in figure 2.2

2.1.2 Fabrication and preparation of disposable devices to handle aqueous phases and emulsions

Here I present the protocols that I followed to fabricate disposable devices that I used to inject aqueous solutions in the microfluidic chips and collect and reinject droplets.

Fabricating and preparing 1 mL aqueous phase injectors Rigid PDMS plugs (cylinders) are cut out of PDMS remnants using a 6 mm diameter biopsy (height of the plug is approximately 5 mm). A hole is punched in the plug of PDMS (0.5 mm). The hole is cleaned using isopropanol and then dried using dry compressed air. A piece of tubing is inserted throughout

the hole (approximately 10 cm long, at least half a centimeter should be going through). The plug is then inserted into a 1 mL plastic syring (latex free, Omnifix, Braun). The piston is evacuated from the body of the syringe and the plug is inserted inside, with the long part of the tubing facing outwards. Needle (27G, Terumo) and tubing (PTFE, 0.3mm ID 0.76 mm OD, Adtech) are adapted to the body of the plastic syringe. The latter is filled with fluorinated oil from its bottom, using another syringe full of oil.

Once the free tubing of the injector is filled with oil, its tip is put inside the aqueous solution. The aqueous phase is gently sucked inside the injector, by pulling the piston of the syringe that was used to fill the injector with oil.

Fabricating and preparing 0.2 mL collection devices Rigid PDMS plugs (cylinders) are cut out of PDMS remnants using a 6 mm diameter biopsy (height of the plug is approximately 5 mm). Two holes are punched in the plug of PDMS (0.5 mm) and the holes are cleaned using isopropanol and then dried using dry compressed air. A piece of tubing is inserted in each hole (approximately 10 cm long): one is inserted through the plug to reach the bottom of the collection tube, the other one is inserted through the PDMS so that its end reaches the bottom of the PDMS plug. The plug is then inserted into a 0.2 mL PCR tube.

Oil (HFE7500, 3M) supplemented with 2% surfactant (008-FluoroSurfactant, Ranbiotechnologies) is injected into the collection tube through the tube reaching the bottom of the collection tube. The latter is entirely filled until all air bubbles disappeared and until oil is dripping from the tip of the free tubing.

If the droplets are to be collected, the tip this tubing is plugged into the outlet of the microfluidic device and the tubing which connects the collection device to the filling syringe is cut immediately after.

Once collection is over, the tubing connecting the device outlet to the collection device is cut, and the filling syringe is introduced back into the tubing that reached the bottom of the tube. No air should be introduced into the collector at this step.

Fabricating and preparation of droplet collection device to grow *B. subtilis* in droplets The piston of a 5 mL syringe (Omnifix latex-free, Braun) is evacuated. A needle and a long tubing (20 cm) are adapted to the syringe (now called collection device). The latter is attached upside down to the microfluidic station, on top of the microfluidic chip from which the droplets are collected. Using another syringe, the collection device is filled with 1 mL of oil supplemented with 2% surfactant. Tissue soaked with water is added to the top of the collection device to prevent droplet evaporation.

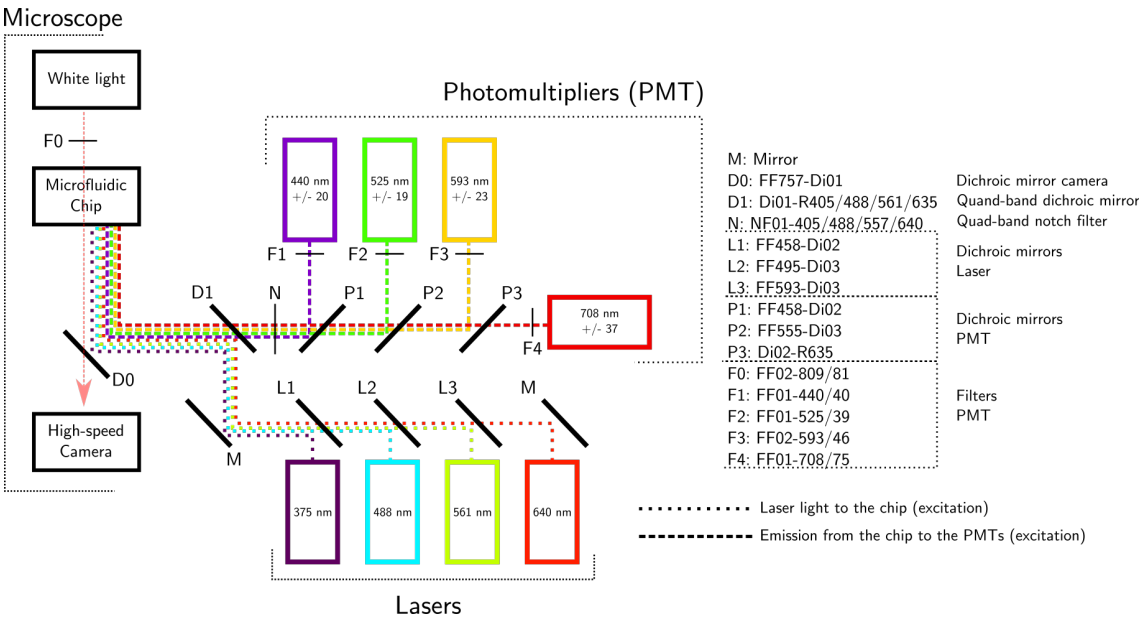


FIGURE 2.3: **Microfluidic station optical setup**.

	Provider	Reference
Microscope	Nikon	Eclips Ti
High-speed camera	Vision Research	Phantom v4.2 High speed camera
Laser 375 nm	Obis	Obis 375LX (50 mW)
Lasers 488 nm	CrystalLaser	DL-488-050
Laser 561 nm	Obis	Obis 561LS (50 mW)
Laser 640 nm	CrystalLaser	DL-640-050-O
Photomultipliers	Hamamatsu	H10723-20
Mirrors, filters and dichroic mirrors	Semrock Inc.	See associated figure
Function generator	Agilent	33220 A
High voltage amplifier	TREK, Inc.	623B

FIGURE 2.4: **Microfluidic station additional references**.

2.1.3 Microfluidic devices operation

Optical and electrical setup

The optical setup of the microfluidic station is presented in figure2.3. The optical setup allow to simultaneously excite fluorescence at λ=375, 488, 561 and 640 nm. Fluorescence can be detected simultaneously in 4 different wave-length windows, set by the optical filters upstream the photomultipliers: 440 nm +/- 20, 525 nm +/- 19, 593 +/- 23, 708 nm +/- 37.

When electrodes are used to generate electric fields within the chips, a function generator and a high voltage amplifier are used. Their reference, along with the references of the microscope, high-speed camera, lasers and photomultipliers are available in figure 2.4.

Preparing the microfluidic devices

All phases (aqueous and fluorinated oil) are operated using Fluigent Microfluidic Flow Control System with Fluiwell-4C reservoirs (connected to 6.5 bar of compressed air via a manometer and a air filter). All reservoirs

are filled with oil if connected to aqueous phase injectors or oil supplemented with surfactant otherwise.

Peek tubing (JR-T-5664, OD = 1/32", ID = 64µm, CIL Cluzeaux), are fitted into the reservoirs. If their pressure path is to be used to inject a solution or an emulsion, their free end is connected to a high pressure peek connectors (P-771 high pressure peek union, F-126S, P-553, CIL Cluzeaux). The peek tubings are filled by turning pressures on. Once oil is dripping from their free end or from the free end of the peek union, pressures are turned. The free end of the peek union is connected to the bottom tubing of the aqueous phase injector device, or the peek tubing is directly plugged into the chip.

If no air bubbles are visible within the different devices, the different solutions are injected into the microfluidic chip for microfluidic operations, using the setting that are described in the next section.

Making and manipulating droplets

In this section, I will detail how droplets volume is determined and detail the microfluidic that were used in this work.

Determining droplet volume The diameter D of a droplet of volume V, squeezed between two parallel planes, as seen from the perspective of a direction orthogonal to those planes, is given by the following relation:

$$D = \frac{2(V + (h/2)^3(\pi^2 + 1.25\pi))}{\pi(h/2)^2(2 + \pi)}$$

According to the desired droplet volume V and the height of the microfluidic channels h, the pressures were adjusted in the droplet maker to achieve the desired diameter.

Microfluidic devices The microfluidic devices used during this PhD project and mentioned in this manuscript are depicted in figure 2.5, 2.6, 2.7, 2.8 and 2.9.

Measuring droplet fluorescence

As the droplets crossed the laser beam at the detection point, the PMTs signal are acquired via an FPGA card (National Instruments) connected to a computer. The trace of the signal is live processed by a custom-developed software. The user chooses a reference fluorescence channel among the four already described. The software detects the time events corresponding to the beginning, the end and the maximum fluorescence of each droplet.

Droplet fluorescence is either defined as:

- The fluorescence in all four channels at the time of the maximum in the reference channel.

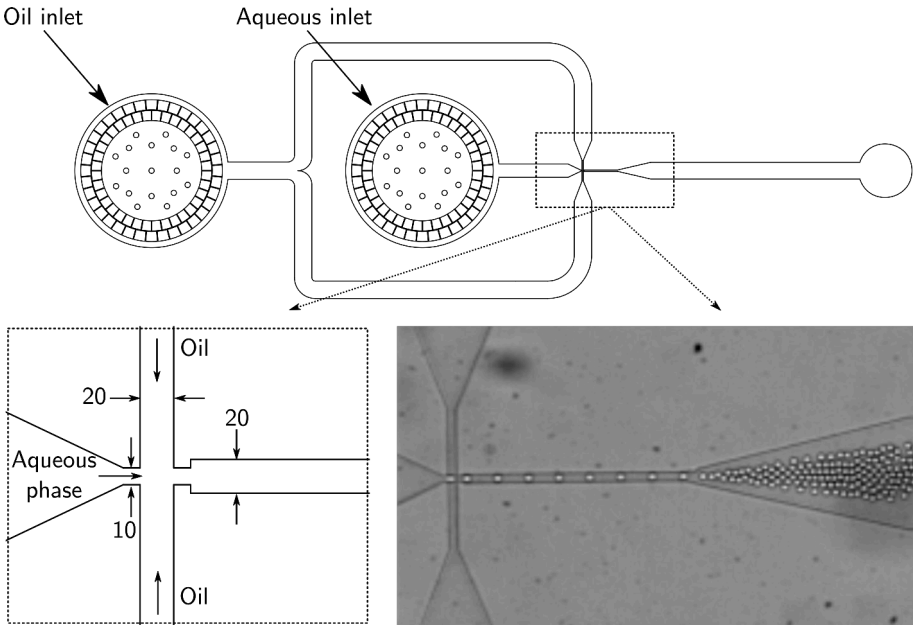


FIGURE 2.5: **2-5 pL droplet maker** All dimensions are in µm. Channels height is 10 µm. Pressure settings: aqueous phase: 1100 mBar, oil phase: 1000 mBar.

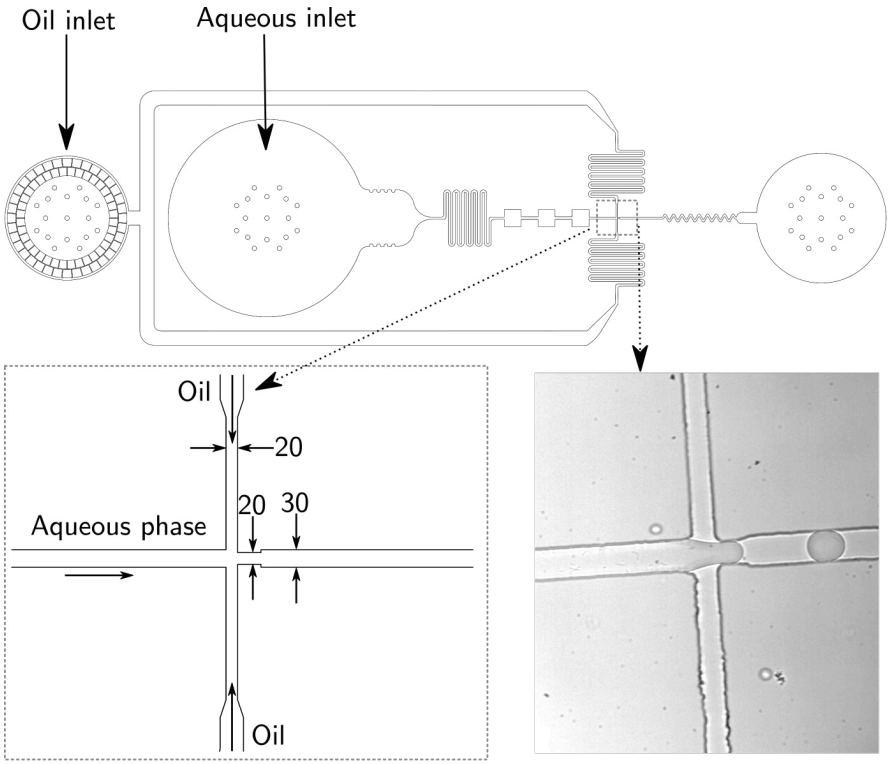


FIGURE 2.6: **20-30 pL droplet maker** All dimensions are in µm. Channels height is 20 µm. Pressure settings: aqueous phase: 1100 mBar, oil phase: 1000 mBar.

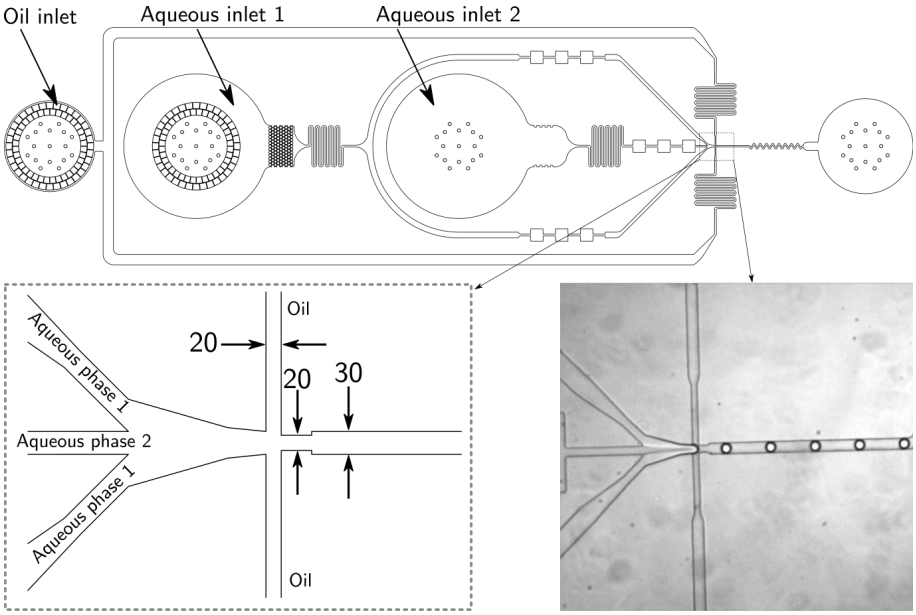


FIGURE 2.7: **20-30 pL co-flow droplet maker** All dimensions are in µm. Channels height is 20 µm. Pressure settings: both aqueous phases: 1100 mBar, oil phase: 1000 mBar.

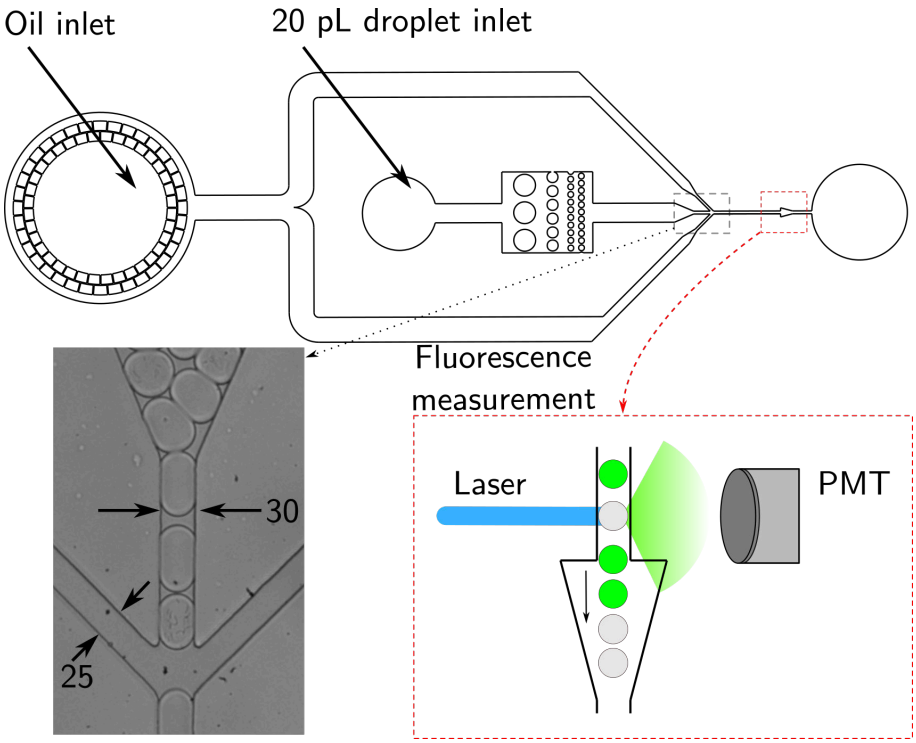


FIGURE 2.8: **20-30 droplet reinjector** All dimensions are in µm. Channels height is 20 µm. Pressure settings: Emulsion: 600 mBar, oil phase: 600 mBar.

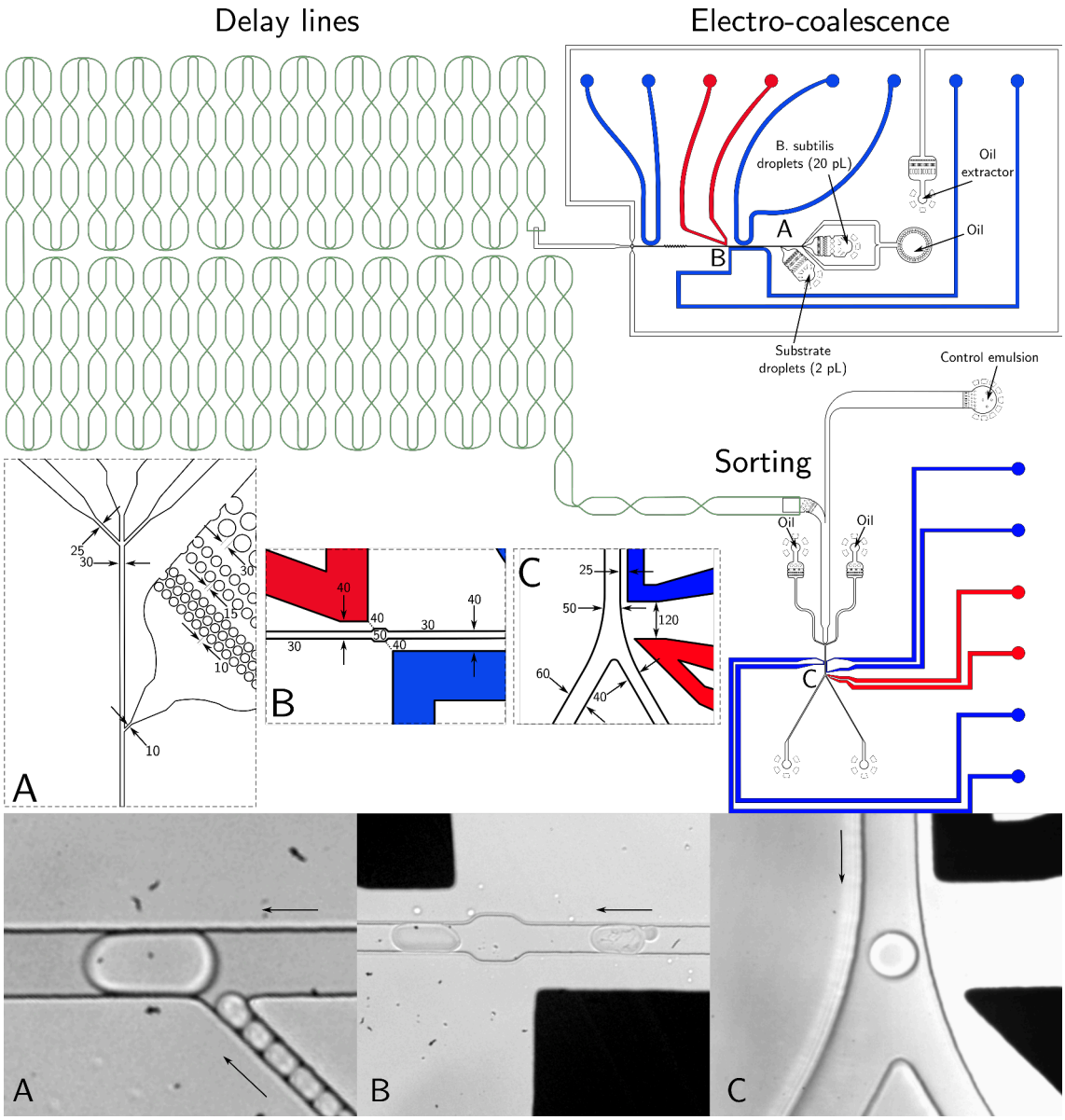


FIGURE 2.9: **2-20 pL droplet electro-coalescence device (integrated with delay lines and sorting device)**. All dimensions are in µm. Channels height is 20 µm except for the delay lines where channels height is 75 µm. Pressure settings: Emulsions: 600 mBar, oil phase: 600 mBar. The extractor is used to extract oil and increase the incubation time in the delay lines, or filled with oil and blocked with a piece of electrical wire. **A** Droplets are paired one to one by adjusting the pressures and **B** droplets are fused by applying a voltage (600 V, 30 kHz) between the positive electrode (red) and the negative electrodes (connected to the electrical ground), using a function generator connected to a x1000 high voltage amplfier. **C**: Droplet fluorescence is measured between the two sorting electrodes (similar to 2.8).

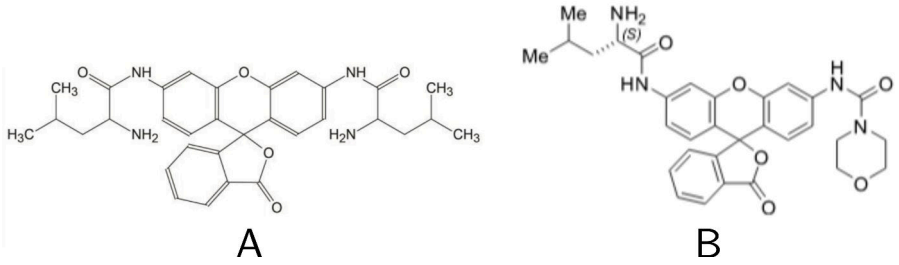


FIGURE 2.10: Structure of the two main substrates used in the SGAP experiments. **A**: Leu_2Rho. **B**: LeuRhoMS. $\lambda ex = 488nm, \lambda em = 535nm$

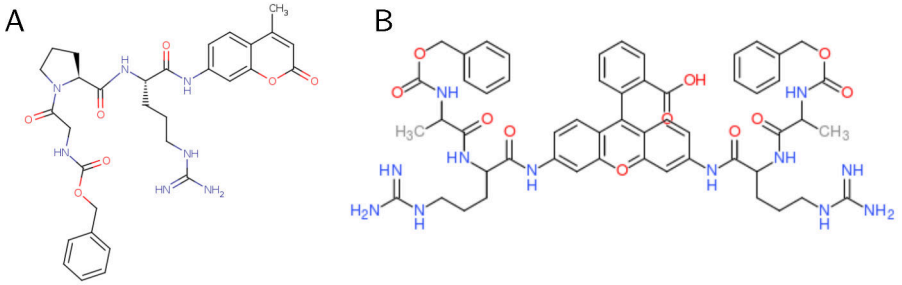


FIGURE 2.11: Structure of the two main substrates used in the trypsin experiments. **A**: Coumarin based trypsin substrate $\lambda_{ex} = 372nm$,$\lambda_{em} = 445nm$. **B**: Rhodamine 110 based trypsin substrate $\lambda_{ex} = 488nm$, $\lambda_{em} = 535nm$

- The mean of the fluorescence in each channel between the beginning and the end of the droplet fluorescent time trace.

In this manuscript, if not stated otherwise, all experiments were performed with the first measurement method.

2.2 Fluorogenic substrates

2.2.1 SGAP fluorogenic substrates

Two different substrates were used in this work to assay SGAP: Bis-(l-leucinyl)-rhodamine 110 (Promega, Charbonniere, France) and Leu-Rho110-MC (NewChem Technologies, UK). They are respectively denoted as Leu_2Rho and LeuRhoMS in the following chapters. Their structure is available in figure 2.10).

2.2.2 Trypsin fluorogenic substrates

In the case of trypsin, two substrates were used: Benzyloxycarbonyl-Gly-Pro-Arg-7-amino-4-methylcoumarin (Enzo Life Science) or (Benzyloxycarbonyl-Ala-Arg)2-Rhodamine110 2HCl (Anaspec). The rhodamine based substrate possesses two sites cleavable by trypsin activity. The structures are depicted in figure 2.11.

Chapter 3

Development of a cell-free microfluidic work-flow for the genotype-phenotype mapping of *Streptomyces griseus* Aminopeptidase (SGAP)

This chapter is dedicated to the development of a cell-free droplet-based microfluidic work-flow to perform the genotype-phenotype mapping of the enzyme *Streptomyces griseus* Aminopeptidase (SGAP). Dr. Alexei Godina had previously started developing such a work-flow. The latter consisted in performing droplet enzymatic assay starting from one single SGAP plasmid per droplet. In the droplets, the SGAP gene was amplified, and expressed in a functional enzyme via a cell-free expression system. The assay was supposed to be carried out by pico-injecting a fluorogenic substrate into the droplets containing the enzyme (see section 3.1).

Unfortunately this work-flow suffers from limitations. I performed bulk control experiments which show that the amplification reagents inhibit the cell-free expression reaction which in turns inhibit SGAP catalytic activity. Both those inhibitions can be lifted if the amplification and the cell-free reaction products are diluted (see section 3.2).

Therefore, I started developing a new work-flow so as to: dilute the amplification product in the cell-free expression mixture and dilute the cell-free expression product in the assay mixture containing the substrate. Bulk and droplets control experiments show that, in principle, two successive dilutions allow the detection of SGAP activity with high contrast, starting from one single SGAP gene per droplet. I developed new microfluidic devices and a new class of hydrophilic fluorogenic substrate (in collaboration with Dr. Fenneteau, Laboratory of Biochemistry ESPCI Paristech) to transpose this scheme to a functional multi-scale microfluidic work-flow, manipulating droplets from 0.2 pL to 20 pL (see section 3.3).

3.1 Previous work on a cell-free microfluidic work-flow to assay SGAP in droplets

In this section, I briefly discuss the work already carried out by Alexei Godina to assay SGAP in droplets.

During his PhD, Alexei Godina developed an in vitro expression microfluidic work-flow [197] based on the Pure system in vitro transcription translation kit (PURExpress, Cosmobio) to screen SGAP libraries for aminopeptidase and phosphodiesterase activities simultaneously. This expression kit is made out of purified components (RNA polymerase, ribosomes, tRNA, etc.)[155, 156] and showed no detectable background aminopeptidase or phosphodiesterase activities contrary to other cell-free commercialized protein synthesis kits based on cells extracts. The work-flow included the following steps:

1. Encapsulating at most one SGAP mutant plasmid per droplet to obtain at most one genotype per droplet.
2. Adding the expression mix to the droplets containing the SGAP mutants.
3. Expressing the SGAP mutant proteins in droplets.
4. Adding a cocktail of fluorogenic substrates to each one of the droplets by picoinjection.
5. Letting the aminopeptidase and phosphodiesterase enzymatic reactions occur in droplets.
6. Measuring the fluorescence of the hydrolyzed substrates.

In his PhD thesis, Alexei Godina reported the following:

1. The expression kit is not sensitive enough to show detectable aminopeptidase activity starting with a SGAP gene concentration of one per 2 pL droplets. It is necessary to perform DNA amplification in the droplets prior to expression. To do so, Alexei Godina opted for an unspecific, Hyperbranched Rolling Circle Amplification (HRCA) using the phi29 polymerase. This isothermal amplification is performed at 30°C which maintains emulsion size monodispersity. Taq polymerase also displayed good results in droplets but at the cost of the emulsion integrity: thermocycling at high temperatures increases size poly-dispersity.
2. The use of Pluronic F-127 and yeast RNA (concentrations available in the appendix A) enables amplification and expression in droplets, possibly because they saturate the interfaces (oil/water, water/plastic).
3. Alexei Godina did not manage to find conditions to perform isothermal unspecific amplification and SGAP expression simultaneously.
4. He also reported that it is possible to perform expression and enzymatic assay simultaneously but at the cost of considerably reducing protein expression yield. This solution also does not allow to decouple expression and enzymatic assay.

5. The optimal incubation time for bulk protein expression lies in between 4 and 6 hours at 37°C.

Because of 3 and 4, the steps of this work-flow have to be performed consecutively. Besides, as the maximum sorting frequency scales as the inverse of the droplet diameter [101], it was chosen to start with 2 pL droplets, the smallest droplet size mastered in our laboratory at that time, and to screen for SGAP activity in 20 pL droplets. The resulting work-flow is depicted in figure 3.1. My first attempts at performing the entire work-flow failed: no SGAP activity could be detected in droplets. I performed bulk control experiments to identify the issues at each step.

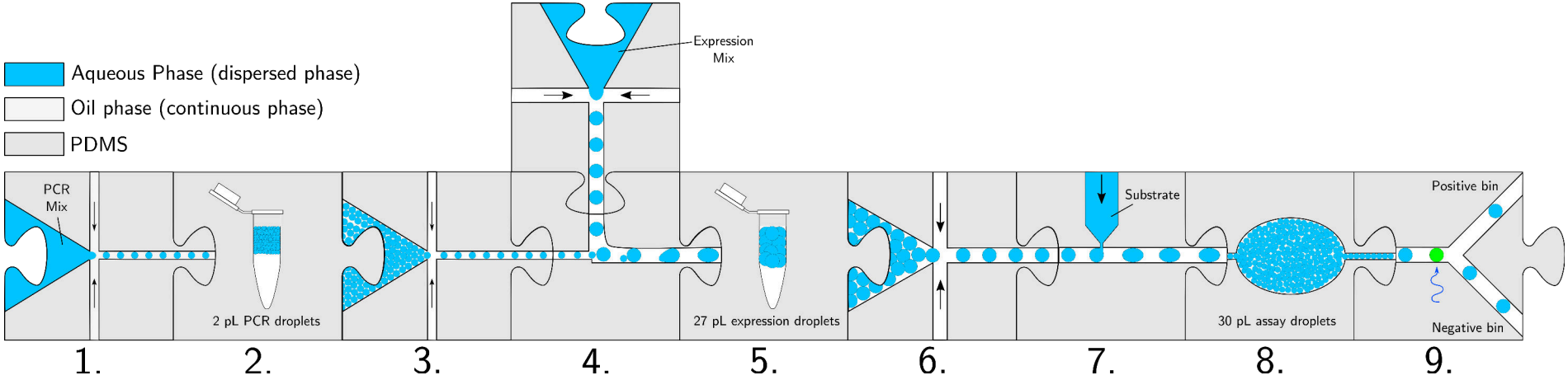


FIGURE 3.1: In vitro microfluidic workflow 1 as in [197].

Step	Description
1.	DNA amplification mix encapsulation (at most, one SGAP plasmid per droplet) in 2 pL droplets and collection off chip in a PCR tube.
2.	DNA amplification in droplets.
3.	Reinjection of 2 pL droplets containing the amplification product in an electro-coalescence device.
4.	Electro-coalescence of the 2 pL droplets with 25 pL droplets containing the expression mix and collection off-chip.
5.	Incubation 37°C for 3 h.
6.	Reinjection of the 27 pL droplets containing the expression product into a picoinjection device.
7.	picoinjection of the 27 pL droplets with substrates cocktail solution.
8.	Incubation for the enzymatic activity to occur in delay lines.
9.	Fluorescence measurement and droplet sorting.

3.2 Development of the in vitro workflow 1: pico-injecting the substrate in droplet containing the enzyme

Bulk control experiments performed at each step of the work-flow (expression and enzymatic assay), show that picoinjection is not suitable to detect SGAP activity in droplets. Those control experiments are detailed in this section.

3.2.1 The PCR reagents inhibit cell-free expression in bulk

The use of a new fluorosurfactant (008-FluoroSurfactant, Ranbiotechnologies) maintains emulsion stability throughout PCR cycling. With a 5% concentration of fluorosurfactant, coalescence is kept between 1% and 5% [99]. Thus, I switched from non specific hyperbranched rolling circle amplification to a classical PCR protocol using the high-fidelity phusion polymerase in order to amplify the SGAP gene in the first step of the work-flow.

First, I determined what volume fraction of PCR product in the expression mix led to the highest final aminopeptidase activity (see A.1.4). It appears that the more diluted the PCR product in the expression mix is, the more active the final product is (figure 3.2): a 20-fold dilution leads to a 6-fold more active final product than a 2-fold dilution.

As 10-fold dilution leads to decent level of aminopeptidase activity. Thus, mixing the PCR reaction and the cell-free expression mix via the electro-coalescence of 2 pL PCR droplets with 25 pL expression droplets is compatible with the droplet enzymatic assay.

3.2.2 Cell-free expression is successful in bulk and droplets

I determined how the incubation time influences the in vitro expression in bulk and the aminopeptidase activity of the expression final product. I also tested expression in droplets to compare it to the bulk expression level. The results are depicted in the figures 3.3 (positive control) and 3.4 (droplet and negative control).

The bulk optimal incubation time in those conditions lies between 6 and 10 h of incubation. Regarding expression in droplets, the activity of the final product is ten times lower than the bulk control for the same incubation time (see figure 3.3 and 3.4). This might be explained by the partial adsorption of the components of the expression kit on the surface of the droplets and of the tubes, despite the use of pluronic and yeast RNA to saturate the interfaces. Nevertheless, the activity from the expression in droplets is still 200-fold higher than the negative control (- DNA, 0.1 RFU/min), validating the SGAP expression step in 27 pL droplets after 2 pL - 25 pL droplet electro-coalescence (step **4** on the figure 3.1). The protocol details of this experiment are available in paragraph A.1.5.

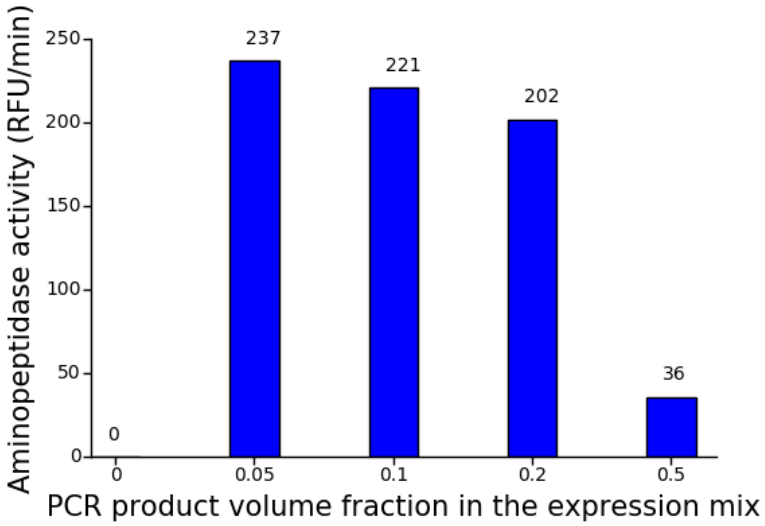


FIGURE 3.2: **Aminopeptidase activity of final expression products prepared with varying SGAP PCR product volume fractions**. An SGAP PCR product was mixed with an expression mix according to the following volume fractions: 0, 0.05, 0.1, 0.2 and 0.5. After 3 h of incubation (sub-optimal incubation time, but enough to obtain a decent aminopeptidase activity) at 37°C for the expression to occur, an enzymatic assay was performed on each final product by diluting the latter ten times in the Leu_2Rho substrate solution. The more diluted the SGAP PCR product is, the more active is the final product.

3.2.3 Cell-free expression reagents inhibit SGAP enzymatic activity in bulk: picoinjection is incompatible with activity detection

Following previous results, I investigated the expression product compatibility with the aminopeptidase detection assay on the Leu_2Rho substrate: what should be the dilution ratio between the expression mix and the aminopeptidase assay mix? I first addressed those questions in bulk, before addressing them in droplets. The results of this experiment and the details of the associated protocol are available in figure 3.5 and paragraph A.1.6.

Unexpectedly, the most important activity is detected when the expression product is diluted 10 times in the assay mix (86 RFU/min). On the contrary, diluting the assay mix 10 times in the expression product results in an activity more than 10-fold lower (7 RFU/min). This is unexpected as, the latter sample is characterized by a final SGAP concentration 10 times more important. My interpretation is that the expression kit reagents inhibit the SGAP aminopeptidase activity against the substrate.

So far I have shown that the amount of aminopeptidase activity detected in the droplets is lower than the activity detected in bulk. Besides, the expression product needs to be diluted in the assay mix to allow an optimal detection of the aminopeptidase activity. Those two facts might explain the difficulties I encountered in detecting any aminopeptidase activity with the workflow 1 as seen in figure 3.1. The picoinjection step (step **7**), which dilutes the assay mix in the expression product droplet has to be changed

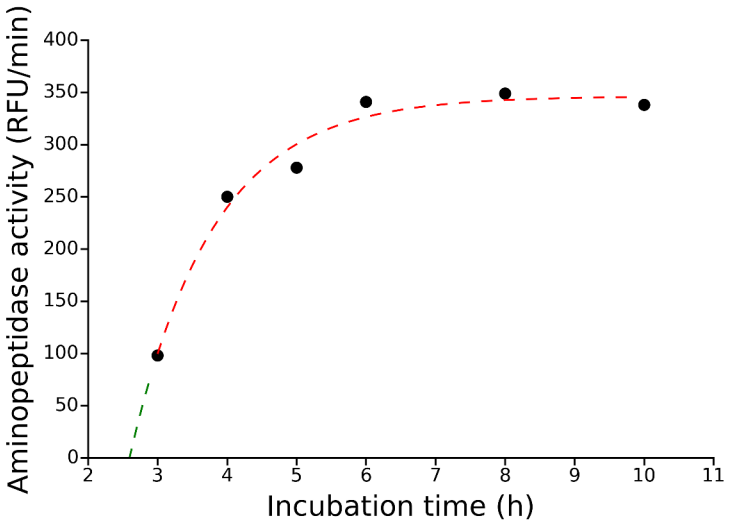


FIGURE 3.3: **Aminopeptidase activity for different expression incubation times in bulk.** Expression products that were incubated 3, 4, 5, 6, 8 and 10 h for expression at 37°C, were assayed against the Leu_2Rho substrate to measure their aminopeptidase activity. The negative control assay (without DNA) was performed after 4 h of incubation. The activity seemed to saturate after 6 h of incubation at 346 RFU/min (exponential fit). The extrapolation of the exponential fit to low level of activity (in gray on the figure) gives a time offset of 2.6 h compared to 0 h, time of mixing. This is certainly not representative of the evolution of the activity at short incubation times, but it suggests that during the first 3 h of incubation, aminopeptidase activity has increased slowly. This might be due to the fact that the mRNA concentration needs to reach a certain level before translation can start and aminopeptidase activity can be detected.

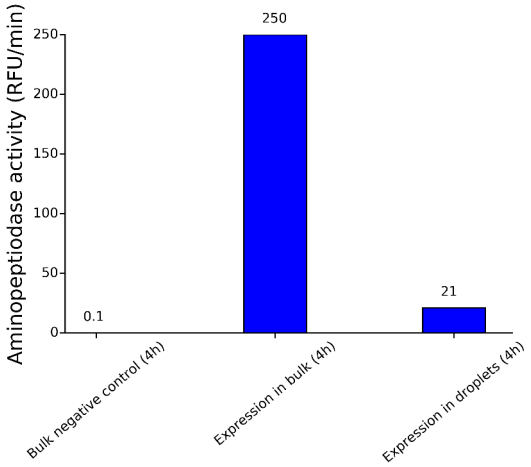


FIGURE 3.4: **Droplet expression product aminopeptidase activity and comparison with bulk controls.** A solution of SGAP plasmid was encapsulated in 2 pL droplets. The SGAP plasmid concentration (approximately 2600 plasmids per droplet) was chosen to yield gene concentration comparable to PCR products in droplets as reported by Alexei Godina. An expression mix was prepared without DNA. Half of the latter solution was injected into a 2 pL - 25 pL electro-coalescence to produce 25 pL expression mix droplets which were fused with the 2 pL DNA droplets produced earlier. The resulting emulsion was collected into a syringe and kept at 37°C during 4 h for the expression to occur inside the droplets. The emulsion was then recovered into a PCR tube and broken to recover the aqueous phase. Its aminopeptidase activity against the Leu_2Rho fluorogenic substrate was then measured. Expression in droplets leads to an activity 100 times higher than the bulk negative control but 10 times lower than the positive bulk control.

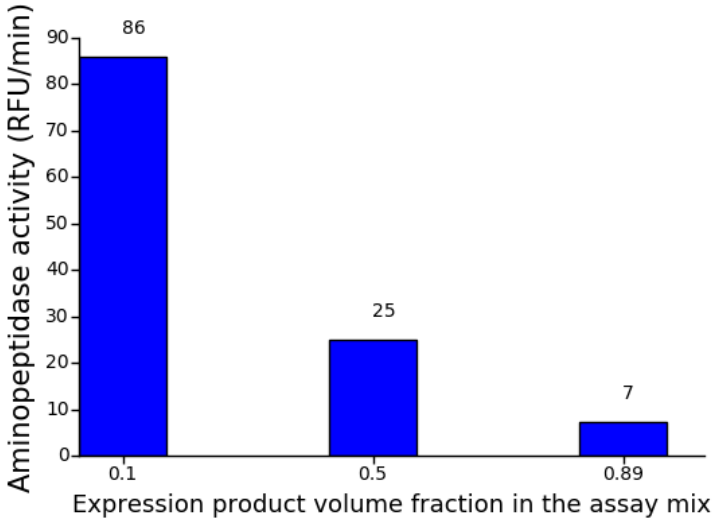


FIGURE 3.5: **Aminopeptidase activity of different expression product volume fractions in the assay mix.** A SGAP PCR product was diluted 20 times in an expression mix. The latter was then incubated for 3 h at 37°for expression (sub-optimal, but enough to see a decent aminopeptidase activity). Next, I added 10 µL, 50 µL and 89.4 µL of this expression product to solutions of Leu_2Rho (final concentration 20 µM in all cases), topped up each sample to 100 µL with water and performed fluorescence measurements to measure aminopeptidase activity. Unexpectedly, diluting the expression product ten times in the assay mix (volume fraction of 0.1) leads to the highest activity (86 RFU/min). On the contrary, diluting the assay mix ten times in the expression product (volume fraction of 0.89) leads to low activity level (7 RFU/min).

for another step to dilute the expression product droplet in the assay mix droplet.

3.2.4 Cell-free expression droplets have to be diluted in the assay droplets

I chose to replace the picoinjection step by a second electro-coalescence step to dilute the expression droplet into the assay droplet. Diluting a 27 pL expression droplet 10 times would have resulted in a final droplet volume of 270 pL which would have required the development of a new sorting device with a 10 times reduced frequency, as the maximum sorting frequency is inversely proportional to the droplet volume [198]. An alternative was to keep the final droplet size unchanged and to start the work-flow with smaller droplets (0.2 pL). Those would be diluted 10 times twice, from 0.2 to 2 pL and from 2 to 20 pL at the expression and at the enzymatic assay steps respectively. It was an opportunity to start a collaboration with Dr. Marie Leman at the Microfluidics, MEMS and Nanostructures laboratory at ESPCI Paristech, who already mastered femtoliter droplets generation, to develop a multi-scale high-throughput microfluidic work-flow. Besides, I could benefit from the fast production frequency of the femtoliter droplets. The imagined in vitro microfluidic work-flow 2 is depicted in figure 3.6.

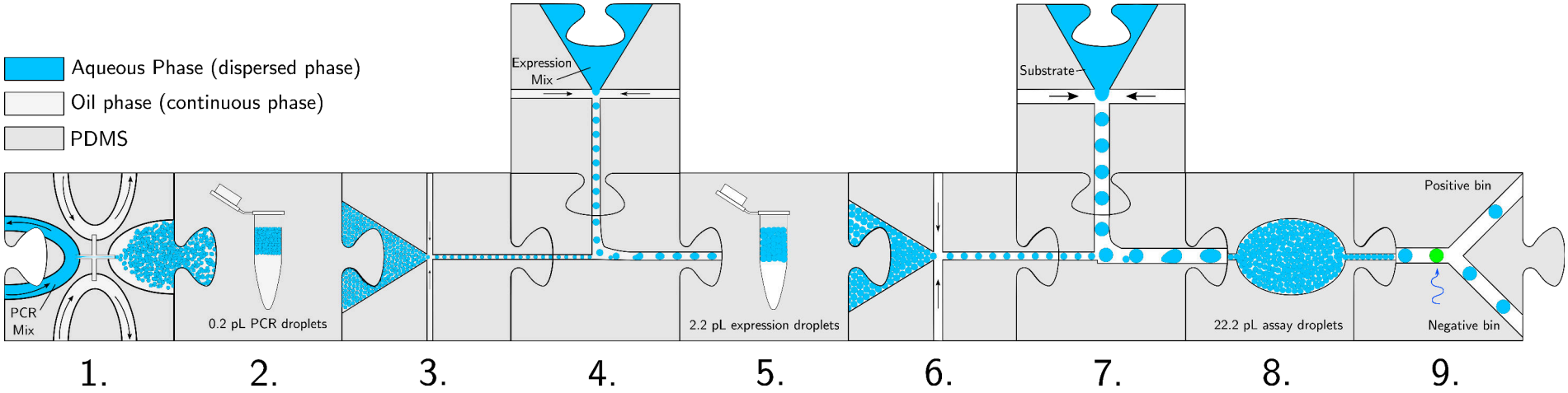


FIGURE 3.6: In vitro microfluidic workflow 2

Step	Description
1.	PCR mix encapsulation (at most, one plasmid per droplet) in 0.2 pL droplets and collection off chip in a PCR tube.
2.	PCR in droplets.
3.	Reinjection of 0.2 pL droplets containing the PCR product in an electro-coalescence device.
4.	Electro-coalescence of the 0.2 pL droplets with 2 pL droplets containing the expression mix and collection off-chip.
5.	Incubation 37°C for 8 h.
6.	Reinjection of the 2.2 pL droplets containing the expression product into an electro-coalescence device.
7.	Electro-coalescence of the 2.2 pL droplets with 20 pL droplets containing the substrate.
8.	Incubation for the enzymatic activity to occur in delay lines.
9.	Fluorescence measurement and droplet sorting.

3.3 Development of the in vitro workflow 2: diluting the enzyme containing droplet into the substrate containing droplet

I first performed bulk and droplet control experiments to validate that two successive 10-fold dilution steps allow the detection of SGAP activity in droplets. Then I developed a new class of rhodamine substrates compatible with SGAP droplet-based enzymatic assay in collaboration with Dr. Johan Feneteau (this work was published, see [167]). I started transposing the two dilution steps scheme to microfluidics by developing a 0.2 pL droplets maker device and a 0.2-2 pL electro-coalescence device. Here I show that the SGAP gene can be amplified successfully in 0.2 pL droplets.

3.3.1 Fusing 2 pL expression and 20 pL assay droplets allows aminopeptidase activity detection with high contrast

I validated that the aminopeptidase enzymatic activity could be detected via the electro-coalescence of 2 pL expression droplets with 20 pL assay droplets. Bulk expression products with and without DNA ("expression mix + DNA" and "expression mix - DNA" respectively) were emulsified in 2 pL droplets in presence of two different concentrations of a red fluorescent dye ("coding dye"). The mix of those two emulsions was then fused with 20 pL assay droplets, and the final emulsion was reinjected on the microfluidic station to measure its fluorescence distributions. The results are depicted in 3.7. The protocol details are available in paragraph A.1.7.

The 2 pL droplets showed identical basal substrate fluorescence during their production and all 20 pL droplets displayed identical level of fluorescence before electro-coalescence. Therefore, the high fluorescence displayed by the positive emulsion is interpreted as the outcome of the SGAP aminopeptidase activity. The contrast of aminopeptidase activity fluorescence signals between the positive and negative emulsions enables to potentially screen for up to 7 different levels of activities (difference of the means of the green and gray distributions, normalized by the sum of their standard deviation).

Some droplets displayed low red fluorescence and no SGAP activity. Those droplets are interpreted as substrate droplets that were not fused with droplets containing the expression product. Also, it can be seen on the 2D histogram that the activity distribution tends to homogenize at the average activity fluorescence. This is caused by two main factors: (i) exchange of the green rhodamine 110 compound released after processing of Leu_2Rho by SGAP, between the positive and negative emulsions (this phenomenon starts being visible after 1 h of incubation) and (ii) partial coalescence of the expression droplets before or during electro-coalescence. Besides, the results of this experiment have to be nuanced for two main reasons. First, a bulk PCR was used as a DNA starting material whose yield is better in bulk than in a droplets (see figure 3.8): activity might not be as important with the full droplet work-flow and 1 h of incubation might not be enough.

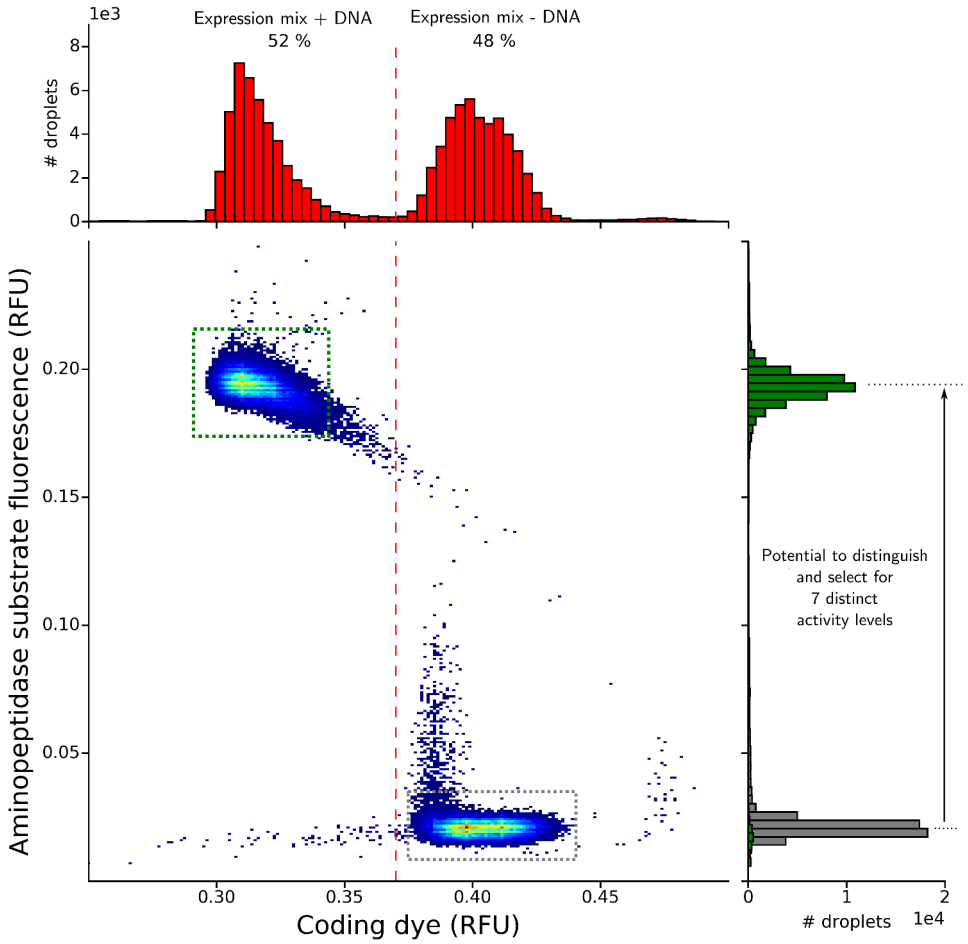


FIGURE 3.7: **Validation of the enzymatic assay in droplets by fusing 2 pL expression droplets with 20 pL substrate droplets.** I prepared two expression mixes with (expression mix + DNA) and without DNA (expression mix without DNA), with respectively 2.5 µM and 25 µM of color coding dye. After incubation for bulk expression at 37°C for 3 h, the two mixes were encapsulated separately but simultaneously in 2 pL droplets and collected in the same PCR tube. I prepared an enzymatic assay mixture containing the fluorogenic substrate LeuRhoMS which was subsequently emulsified in 25 pL droplets and fused with the 2 pL droplet emulsion prepared previously, using a 2 - 25 pL electro-coalescence device. The collected droplets were incubated 1 h at room temperature, off chip, to let the enzymatic reaction occur. The emulsion was then reinjected to measure the fluorescence distribution in the substrate and the coding dye fluorescence channels. **Top histogram**: Color coding dye fluorescence channel. Low fluorescence accounts for the expression mix + DNA emulsion, high fluorescence accounts for the expression mix - DNA emulsion. **Right histogram**: Aminopeptidase activity fluorescence channel. The expression mix + DNA emulsion (green square) shows high activity compared to the negative control (grey square). Up to 7 different populations can be discriminated (difference of the means of the green and gray distributions, normalized by the sum of their standard deviation), as seen on the aminopeptidase substrate histogram.

Second, a better negative control could have been chosen with an inactive SGAP mutant.

Still, this result shows the possibility to detect aminopeptidase activity through electro-coalescence of 2 pL expression droplet with 20 pL assay droplets. This assay could be improved by the development of new non exchangeable fluorogenic substrates. This development was uptaken through a collaboration with the Laboratoire de Chimie Organique at ESPCI Paristech (LCO) and detailed through the publication [167] presented in the following paragraph 3.3.2.

3.3.2 Synthesizing a new non leaky substrate to improve SGAP enzymatic assay in droplets

As mentioned in a previous section 3.3.1, fluorophores such as rhodamine 110 are "leaky", i.e, they are exchanged between the water in oil droplets during incubation either because they are partially soluble in the carrier oil, or because of micellar exchanges. In the course of a genotype phenotype mapping experiment on an enzyme, droplets bearing very active enzyme mutant will be characterized with a high fluorophore concentration while droplets bearing enzyme mutant with low activity will be characterized with low fluorophore concentration. Because of the exchange, the fluorescent contrast between active and non active mutants will be lowered or even canceled, lowering the fluorescence resolution between variants displaying different activities. Such a transport phenomenon can be seen in figure 3.7. It has been shown that the exchange between droplets could be controlled using substrates based on hydrophilic fluorescent compounds which are exchanged less quickly [169, 170]. In this spirit, we started a collaboration with Dr. Fenneteau and Prof. Cossy from the Laboratoire de Chimie Organique (LCO, ESPCI Paristech).

Briefly, Dr. Johan Feneteau synthesized a new hydrophilic sulfonylated rhodamine110-based substrate dedicated to the detection of peptidase activity. I characterized the properties of this new substrate, in bulk and in droplets, and showed it was compatible with droplet-based microfluidic enzymatic assays:

- Whereas the fluorescence between droplets containing either hydrolyzed or non hydrolyzed rhodamine 110 substrate equilibrates within 10 h, the hydrophilic rhodamine 110 substrate contrast stabilizes within 3h and retains 84% of its initial value over more than 16h.
- The hydrophilic rhodamine 110 substrate substrate is hydrolyzed in presence of the SGAP enzyme in bulk and in droplets. Droplet-based experiments show that, after 16 h on incubation, it allows to quantify aminopeptidase activity into more than 20 different levels of activity between the fluorescence of droplets with no enzyme and droplets bearing SGAP in which substrate hydrolysis was complete.

Our work is summarized in the following manuscript: "Synthesis of New Hydrophilic Rhodamine Based Enzymatic Substrates Compatible With

Droplet-Based Microfluidic Assays", published in the Chemical Communication journal (RSC, 2017) [167], which is integrated to this thesis manuscript. The supplementary information associated with the submitted manuscript is availaible in the appendix (A.2).

ChemComm

COMMUNICATION

Cite this: *Chem. Commun.*, 2017, 53, 5437

Received 25th February 2017,
Accepted 25th April 2017

DOI: 10.1039/c7cc01506b

rsc.li/chemcomm

Synthesis of new hydrophilic rhodamine based enzymatic substrates compatible with droplet-based microfluidic assays†

Johan Fenneteau, ‡[a] Dany Chauvin,‡[b] Andrew D. Griffiths,[b] Clément Nizak *[b] and Janine Cossy*[a]

Here we report the conception, synthesis and evaluation of new hydrophilic rhodamine-based enzymatic substrates for detection of peptidase activity compatible with high-throughput screening using droplet-based microfluidics.

Droplet-based microfluidic techniques allow the production of highly monodisperse picoliter droplets that can be considered as independent miniaturized reactors.[1] They provide a powerful ultrahigh-throughput screening platform that can be used, in particular, in the field of directed evolution of enzymes.[2] For this purpose, millions of enzyme mutants are assayed individually in separate droplets, and the catalytic activity of each mutant is measured using a fluorescent read-out allowing the sorting of droplets containing active mutants. After the selection of active mutants, mutagenesis is applied to generate new mutants, and the selection–mutation cycle is repeated until the most active mutants reach a desired target activity level. This process has been used for fundamental studies of enzyme evolution,[3] and can be applied in the industrial sector to develop new enzymes (*e.g.* for applications in detergent formulations, the food industry, and biotechnology).[4] In such high-throughput screening, each droplet contains the gene for a different enzyme variant that is expressed *in situ*.[2,5] Droplet-based enzymatic assays typically rely on fluorogenic enzymatic substrates present in each droplet that release a fluorescent moiety when processed by an active enzyme (Fig. 1A). Droplets can then be individually sorted according to their fluorescence signal to allow collection of droplets containing active and non-active enzymes and their encoding genes (Fig. 1B). However, this sorting step is possible only if transport of the chromophore between the droplets is limited. Although exchange between droplets is greatly reduced by using fluorinated carrier oils in which non-fluorinated molecules have very poor solubility,[6] transport can also be mediated by micellar exchange.[7] More hydrophilic molecules however, tend to exchange less quickly.[8] As a consequence, one key property of these fluorogenic substrates is that the released fluorescent products must be hydrophilic enough to stay confined inside the droplets until assay readout. Thus, no significant exchange between droplets should occur on the timescale of the enzymatic assay in droplets (Fig. 1C).

In this study, we focused on the evolution and adaptation of a gene encoding a well-characterised enzyme, *Streptomyces*

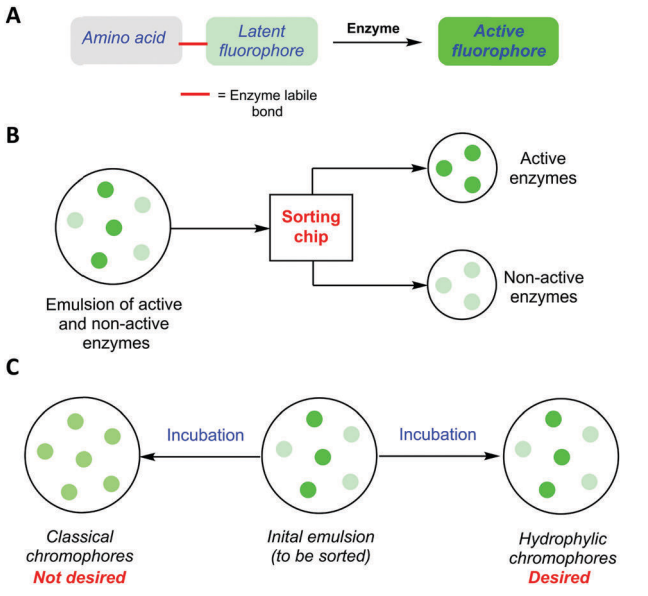


Fig. 1 Fluorescence based enzymatic assay in droplets: (A) typical fluorogenic enzymatic substrate used for these assays. (B) Fluorescence based sorting device. (C) Evolution of droplet fluorescence over time during incubation.

[a] *Laboratoire de Chimie Organique, Institute of Chemistry, Biology and Innovation (CBI), UMR 8231, ESPCI Paris/CNRS, PSL Research University, 10 rue Vauquelin, 75231-Paris Cedex 05, France. E-mail: janine.cossy@espci.fr*

[b] *Laboratoire de Biochimie, Institute of Chemistry, Biology and Innovation (CBI), UMR 8231, ESPCI Paris/CNRS, PSL Research University, 10 rue Vauquelin, 75231-Paris Cedex 05, France. E-mail: clement.nizak@espci.fr*

† Electronic supplementary information (ESI) available: Experimental procedures, characterization data, copy of ^{1}H and ^{13}C NMR, device fabrication. See DOI: 10.1039/c7cc01506b

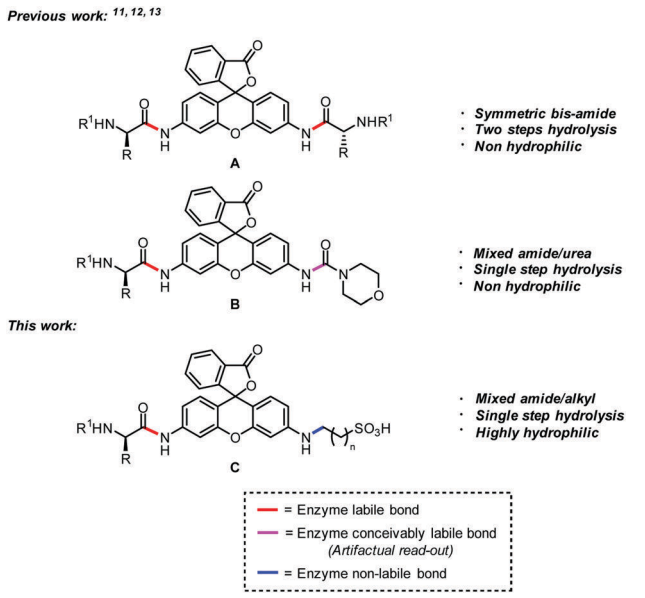
‡ These authors contributed equally.

griseus Amino-Peptidase (SGAP). SGAP is a 284 amino acid enzyme that catalyses the hydrolysis of *N*-terminal amino acid residues from polypeptides, with a strong preference for leucine residues and detectable activity (10 to 1000-fold lower) towards other residues such as methionine and alanine.[9]

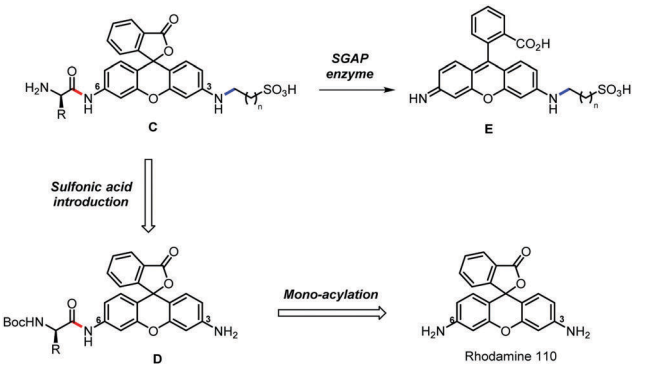
Our long-term goal is to evolve SGAP to modify its activity profile towards other *N*-terminal amino-acids. For this purpose fluorogenic substrates releasing a non-exchanging fluorophore (*i.e.* highly hydrophilic) have to be prepared. Due to its great photostability and photochemical properties, rhodamine has been widely used as fluorescent moiety to monitor various enzymatic activities.[10] We thus focused on the modification of the rhodamine core to obtain fluorogenic substrates releasing a non-exchanging dye. Rhodamine bis-amides **A** were previously developed to assay serine protease activity,[11] then extended to assay other enzymatic activities (Scheme 1).[12] However, due to the two hydrolysis steps required to unmask the bulk fluorescence of rhodamine 110, these substrates show complex kinetic profiles.[11,12a] Mixed amide/urea rhodamine substrates **B** were therefore developed to monitor a single step hydrolysis that releases the moderately fluorescent rhodamine urea.[13]

In the context of directed evolution, we would like to exploit this single step hydrolysis principle but with a non-hydrolysable linker, to avoid possible artifactual read-out of urease activity. We thus decided to aim for rhodamine of type **C** with an amide bond and a non-hydrolysable linker bearing a sulfonic acid group to avoid the exchange of the released dye **E**, thus extending a concept that has been demonstrated previously in the case of coumarin fluorophores (Scheme 2).[8]

The targeted substrate **C** would be synthesized by introduction of a sulfonic acid moiety on intermediate **D**, possessing the required amino-acid, which would be obtained from the commercially available rhodamine 110 by mono-acylation of one of



Scheme 1 Different type of fluorogenic rhodamine-based substrates developed for detection of enzymatic activity.

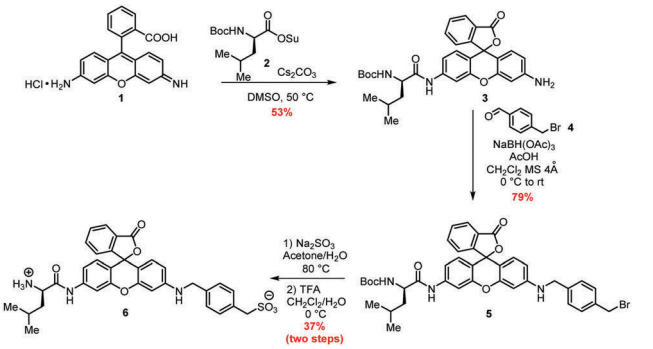


Scheme 2 Retrosynthesis of sulfonylated rhodamine-based fluorogenic substrates for SGAP activity detection.

the amino groups (Scheme 2). It is expected that the late stage introduction of the sulfonic acid function would simplify the purification steps of highly polar sulfonates and avoid the use of sulfonate protecting groups.[14]

To establish the proof of concept, we focused on the synthesis of compound **6** that comprises a leucine residue (Scheme 3). After extensive optimisation of the reaction conditions,[15] compound **1** was condensed with *N*-hydroxysuccinimidyl activated Boc-leucine ester **2** under basic conditions (Cs_2CO_3, DMSO, 50 °C) to produce the desired mono-acylated product **3** in 53% yield.[16] At this stage, introduction of the sulfonate group or a precursor was investigated. When **3** was treated with different electrophiles such as 1,3-propane sultone,[17] bis-(2-chloroethyl)amine,[18] 3-bromopropan-1-ol,[19] or 1,4-bis-(bromomethyl) benzene, the remaining aniline group was not alkylated, even under thermal activation. In contrast, when **3** was reacted with 4-bromomethyl benzaldehyde (**4**) under reductive amination conditions ($NaBH(OAc)_3$, AcOH, MS 4 Å, CH_2Cl_2) the benzylated product **5** was isolated (79% yield). To the best of our knowledge, this is the first example of modification of the rhodamine core with the use of simple aldehydes as electrophiles. This method opens up access to various dissymmetric rhodamines that are still a challenge for synthetic chemists. Treatment of **5** with sodium sulfite followed by TFA-mediated *N*-Boc-cleavage afforded the desired product **6** in 37% yield over two steps.

The absorption and emission spectra of this substrate were measured. Compound **6** displays a λ_{max} at 505 nm and green



Scheme 3 Synthesis of fluorogenic SGAP substrate **6**.

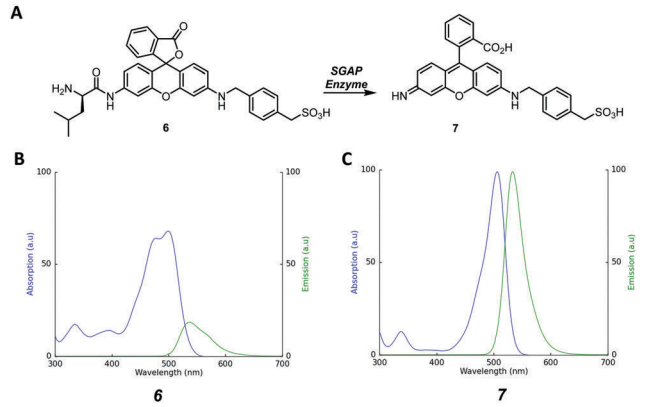


Fig. 2 Enzymatic evaluation of substrate **6**: (A) enzymatic release of sulfonylated rhodamine **7**. (B) Absorption and emission spectra of compound **6** (10 μM) in pH = 8.0 SGAP buffer. (C) Absorption and emission spectra of compound **7** (10 μM) in pH = 8.0 SGAP buffer.

fluorescence emission (λ_{em} = 534 nm) (Fig. 2). Incubation of **6** with SGAP enzyme in pH = 8.0 buffer releases the highly fluorescent sulfonylated rhodamine **7** (λ_{em} = 538 nm, Φ = 0.76),[20] by cleavage of the amide bond, resulting in up to 11-fold increase of the fluorescence signal at 538 nm,[21] consistent with previous observations on rhodamine dyes.[13g]

Exchange experiments were performed to confirm the beneficial effect of the sulfonyl group on the retention time of rhodamine derivatives in aqueous droplets. Substrates **6** and **8** were selected to study the respective behaviour of the released sulfonylated chromophore **7** and non-sulfonylated chromophore **9** (Fig. 3A).[22] Substrates **6** and **8** were incubated in the presence and in the absence of SGAP at 37 °C for 30 min to obtain full conversion of **6** and **8** into respectively **7** and **9** as determined by fluorescence spectrophotometry. Stability of **6** and **8** in the absence of SGAP was also confirmed (see Fig. S1, ESI†). For each substrate, the solutions containing either the released or the non-released fluorophore were separately compartmentalized along with a non-exchanging control dye into 20 pL droplets and collected simultaneously in the same tube.[23] The resulting mixed emulsion was incubated at room temperature and droplet fluorescence was monitored over time (Fig. 3B, see also Fig. S2, ESI†). In droplets containing the non-sulfonylated compounds **8** and **9**, droplet fluorescence equilibrated completely after 10 h to the average fluorescence level of the initial emulsion, relatively to the control dye. (Fig. 3C, see also Fig. S4 and S5 for detailed kinetic analysis, ESI†). In contrast, the two populations of droplets containing **6** and **7** showed well-separated fluorescence distributions over 16 h (Fig. 3C, see also Fig. S3, ESI†), again in the absence of exchange for a control dye. The difference between the average fluorescence of both populations remained constant at 84% of its initial value for 16 h (see Fig. S3 and S5 for detailed kinetic analysis, ESI†).

Finally, we confirmed the activity of SGAP towards **6** in droplets. Two populations of 20 pL droplets were produced simultaneously on the same chip and collected in the same tube (Fig. 4A, see also Fig. S6, ESI†).

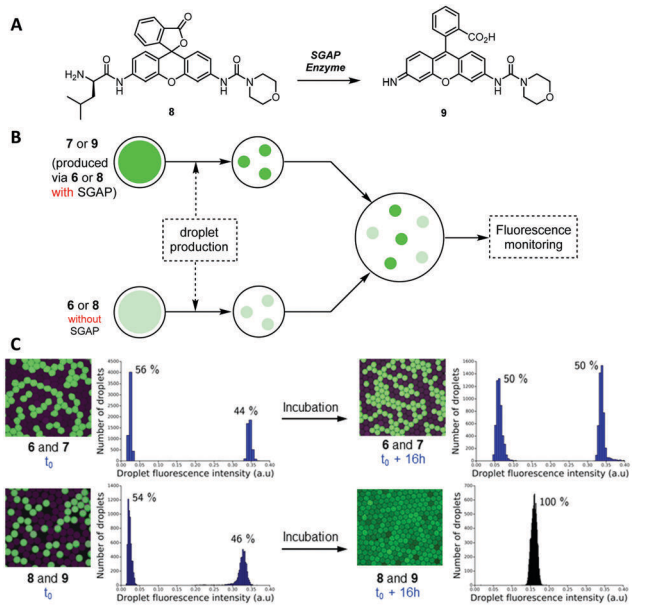


Fig. 3 Exchange test experiment: (A) enzymatic release of non-sulfonylated rhodamine **9**. (B) Principle of formation of the heterogeneous emulsion (see Fig. S2 for more details, ESI†). (C) Fluorescence signal evolution of the emulsions containing **6** and **7**, but also **8** and **9**. An image of a 2D-confined sample of the emulsion (20 pL droplets) and the corresponding droplet fluorescence histograms are shown for **7** and **9** at t_0 and t_0 + 16 h (see Fig. S3–S5 for detailed kinetics, ESI†). For each subpopulation the ratio of droplet counts are indicated.

The first population contained **6** without SGAP enzyme (negative population), the second population contained **6** and SGAP (positive population). SGAP and **6** (or buffer and **6**) were co-flowed immediately prior to droplet formation, ensuring that there was no enzymatic action before droplet formation. The two droplet populations showed the same fluorescence profile during droplet production (t_0). After 16 h of incubation at room temperature, the average fluorescence of SGAP-containing droplets increased by 7.7-fold (Fig. 4B), demonstrating SGAP activity towards **6** in droplets. In turn, the average fluorescence of negative control droplets increased by 2.7-fold (Fig. 4B).[24] The gap between the fluorescence distributions of the SGAP-containing population and the negative population is more than 48 times the typical standard deviation of the two populations after 16 h of incubation. The dynamic range of our droplet enzymatic assay allows thus to distinguish 24 activity levels between background noise (*e.g.* null mutant) and wild-type SGAP. We have further shown that our assay allows discriminating droplets containing different levels of enzymatic activity using different concentrations of SGAP to mimic different k_{cat} values (Fig. S8, ESI†).

In conclusion, we have demonstrated that the sulfonylated rhodamine-based enzymatic substrate **6**, produced through our new reductive amination approach, is compatible with microfluidic high-throughput screening and can be used to assay enzymatic activity in droplets. The exchange of the released sulfonylated chromophore **8** between droplets was considerably reduced in comparison to non-sulfonylated rhodamine,[7] and

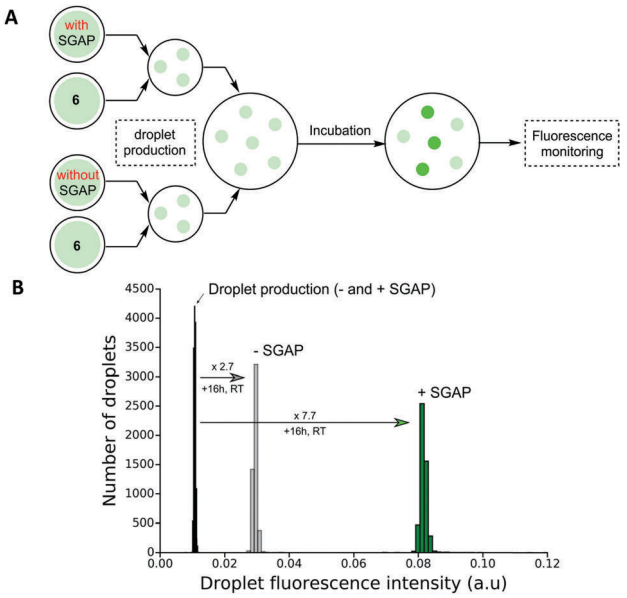


Fig. 4 Detection of SGAP activity in droplets: (A) two emulsions were produced, one with droplets containing **6** and SGAP, the other with droplets containing **6** without SGAP. (B) Evolution of the fluorescence distribution of positive and negative droplet populations over 16 h of incubation. In the presence of SGAP, the average droplet fluorescence increased 7.7-fold, while it increased only 2.7-fold in the absence of SGAP.

the dynamic range of our droplet enzymatic assay allows to quantify activity into >20 levels between background noise and wild-type SGAP. This new class of sulfonylated rhodamine-based substrates opens up droplet microfluidics enzymatic assays in the green fluorescence channel, thus complementing already available sulfocoumarin-based substrates in the blue fluorescence channel.[8]

This project was supported by a grant from Agence Nationale de la Recherche (ANR) under the SIMBAD project (ANR-14-CE16-0020), and a PhD fellowship to D. C. from the Institut Pierre-Gilles de Gennes. We thank Alexei Godina for advice on enzymatic assays in droplets and fluorogenic substrates at the onset of our project.

Notes and references

1 A. B. Theberge, F. Courtois, Y. Schaerli, M. Fischlechner, C. Abell, F. Hollfelder and W. T. S. Huck, *Angew. Chem., Int. Ed.*, 2010, **49**, 5846.
2 J. J. Agresti, E. Antipov, A. R. Abate, K. Ahn, A. C. Rowat, J.-C. Baret, M. Marquez, A. M. Klibanov, A. D. Griffiths and D. A. Weitz, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 4004.
3 (*a*) R. Obexer, M. Pott, C. Zeymer, A. D. Griffiths and D. Hilvert, *Protein Eng., Des. Sel.*, 2016, **29**, 355; (*b*) R. Obexer, A. Godina, X. Garrabou, P. R. E. Mittl, D. Baker, A. D. Griffiths and D. Hilvert, *Nat. Chem.*, 2017, **9**, 50.
4 T. Beneyton, I. P. M. Wijaya, P. Postros, M. Najah, P. Leblond, A. Couvent, E. Mayot, A. D. Griffiths and A. Drevelle, *Sci. Rep.*, 2016, **6**, 27223.
5 A. Fallah-Araghi, J.-C. Baret, M. Ryckelynck and A. D. Griffiths, *Lab Chip*, 2012, **12**, 882.
6 (*a*) R. L. Scott, *J. Am. Chem. Soc.*, 1948, **70**, 4090; (*b*) J. H. Simons and M. J. Linevsky, *J. Am. Chem. Soc.*, 1952, **74**, 4750.
7 P. Gruner, B. Riechers, B. Semin, J. Lim, A. Johnston, K. Short and J.-C. Baret, *Nat. Commun.*, 2016, **7**, 10392.
8 (*a*) G. Woronoff, A. El Harrak, E. Mayot, O. Schicke, O. J. Miller, P. Soumillion, A. D. Griffiths and M. Ryckelynck, *Anal. Chem.*, 2011, **83**, 2852; (*b*) M. Najah, E. Mayot, I. P. Mahendra-Wijaya, A. D. Griffiths, S. Ladame and A. Drevelle, *Anal. Chem.*, 2013, **85**, 9807.
9 D. Ben-Meir, A. Spungin, R. Ashkenazi and S. Blumberg, *Eur. J. Biochem.*, 1993, **212**, 107.
10 (*a*) L. D. Lavis and R. T. Raines, *ACS Chem. Biol.*, 2008, **3**, 142; (*b*) M. Beija, C. A. M. Afonso and J. M. G. Martinho, *Chem. Soc. Rev.*, 2009, **38**, 2410.
11 (*a*) S. P. Leytus, L. L. Melhado and W. F. Mangel, *Biochem. J.*, 1983, **209**, 299; (*b*) S. P. Leytus, W. L. Patterson and W. F. Mangel, *Biochem. J.*, 1983, **215**, 253.
12 (*a*) L. Jixiang, M. Bhalgat, C. Zhang, D. Zhenjun, B. Hoyland and D. H. Klaubert, *Biorg. Med. Chem. Lett.*, 1999, **9**, 3231; (*b*) S. S. Chandran, K. A. Dickson and R. T. Raines, *J. Am. Chem. Soc.*, 2005, **127**, 1652; (*c*) S.-T. Huang and Y.-L. Lin, *Org. Lett.*, 2006, **8**, 265.
13 (*a*) A. P. Guzikowski, J. J. Naleway, C. T. Shipp and R. C. Schutte, *Tetrahedron Lett.*, 2000, **41**, 4733; (*b*) S. X. Cai, H.-Z. Zhang, J. Guastella, J. Drewe, W. Yang and E. Weber, *Biorg. Med. Chem. Lett.*, 2001, **11**, 39; (*c*) L. Jixiang, M. Bhalgat, C. Zhang, D. Zhenjun, B. Hoyland and D. H. Klaubert, *Biorg. Med. Chem. Lett.*, 2001, **9**, 3231; (*d*) S. Lorey, J. Faust, C. Mrestani-Klaus, T. Kähne, S. Ansorge, K. Neubert and F. Bühling, *J. Biol. Chem.*, 2002, **277**, 33170; (*e*) H.-Z. Zhang, S. Kasibhatla, J. Guastella, B. Tseng, J. Drewe and S. X. Cai, *Bioconjugate Chem.*, 2003, **14**, 458; (*f*) Z.-Q. Wang, J. Liao and Z. Diwu, *Bioorg. Med. Chem. Lett.*, 2005, **15**, 2335; (*g*) L. D. Lavis, T.-Y. Chao and R. T. Raines, *ACS Chem. Biol.*, 2006, **1**, 252.
14 (*a*) T. W. Greene and P. G. M. Wuts, *Protective Groups in Organic Synthesis*, John Wiley & Sons, New York, NY, 3rd edn, 1999; (*b*) S. C. Miller, *J. Org. Chem.*, 2010, **75**, 4632.
15 See ESI† for details.
16 J. A. B. Ferreira, V. V. Serra, A. Sanchez-Coronilla, S. M. G. Pires, M. A. F. Faustino, A. M. S. Silva, M. G. P. M. S. Neves, J. A. S. Cavaleiro and S. M. B. Costa, *Chem. Commun.*, 2013, **49**, 8809.
17 (*a*) N. H. Ho, R. Weissleder and C. H. Tung, *Tetrahedron*, 2006, **62**, 578; (*b*) J. Jose, Y. Ueno and K. Burgess, *Chem. – Eur. J.*, 2009, **15**, 418.
18 (*a*) M. A. Clark, S. A. Hilderbrand and S. J. Lippard, *Tetrahedron Lett.*, 2004, **45**, 7129; (*b*) K. G. Liu and A. J. Robichaud, *Tetrahedron Lett.*, 2005, **46**, 7921.
19 V. H. J. Frade, P. J. G. Coutinho, J. C. V. P. Moura and M. S. T. Gonçalves, *Tetrahedron*, 2007, **63**, 1654.
20 Quantum yield of 7 (Φ = 0.76 in SGAP buffer) was estimated by comparison with fluorescein in NaOH (0.1 M) at 488 nm (Φ = 0.92) using the method described by Williams *et al.*: A. T. R. Williams, S. A. Winfield and J. N. Miller, *Analyst*, 1983, **108**, 1067. Absolute quantum yield of rhodamine 110 (Φ = 0.88 in HEPES buffer) was measured by Lavis *et al.*: J. B. Grimm, A. J. Sung, W. R. Legant, P. Hulamm, S. M. Matlosz, E. Betzig and L. D. Lavis, *ACS Chem. Biol.*, 2013, **8**, 1303.
21 This factor depends on substrate concentration and is optimal in the range 0.1–1 μM. See Fig. S7 (ESI†) for details.
22 Compound **8** was purchased from NewChem Technologies, Holly lodge, Whitesmocks, Durham, DH1 4LH, United Kingdom, Tel: +44 (0) 191 375 7294.
23 See ESI† for detailed droplet production setup.
24 This 2.7-fold increase results from initial transient exchange as detailed in Fig. S3 and S5 (ESI†).

3.3.3 SGAP PCR amplification is successful in 0.2 pL droplets

The device I developed and used to produce 0.2 pL droplets is described in figure 3.8, **A** and **B**. First attempts failed to show any SGAP gene amplification by performing PCR in 0.2 pL droplets. The fluorinated oil (HFE 7500, 3M) is characterized by a thermal diffusivity of 4 times lower than the water thermal diffusivity at 30°C. I simply increased by four the time of each thermocycling step in the PCR protocol. This new cycle is entitled "Slow PCR" by comparison with the classical PCR protocol. The results are illustrated in the figure 3.8 and protocol is detailed in A.1.8. Intensity of the band suggests an approximately 800-fold specific amplification in droplets.

It is worth noting that this 0.2 pL droplet maker, while functional, is complicated to use. First, the dimensions of the inner cross makes it prone to clogging, which destabilizes the emulsion and result in important polydispersity (even in the presence of filters). During the chip fabrication process, carefully washing the inside of the PMDS slab holes and their surroundings with isopropanol and drying them considerably reduce the occurrence of clogging events as most of the PDMS particles comes from hole punching. Still, silane precipitation brings in a lot of micrometric particles and gas phase silanisation, which brings in less particles, is not efficient enough for our purposes (the aqueous phase wets the walls of the channels). To get rid of the silanisation step and take advantage of the fact that PDMS is naturally hydrophobic, I tried to use PDMS coated glass slides, but as the PDMS is too soft and the height of the channel small, the latter would collapse. It would be worth repeating this with additives hardening the PDMS.

Besides, even though the frequency of generation is high, 4 h are necessary to produce emulsion volume that can be easily handled. During that amount of time, resistance in the channels is continuously increasing as the filters get clogged. The water phase "tongue" necessary to perform step emulsification retracts in the channel, stopping the emulsion production. While using this chip, it is necessary to manually increase the pressure to start the step emulsification again. This can be solved by using pressure pumps controlled via a feed back loop targeted on constant flow-rates values.

So far the following has been shown:

- SGAP DNA can be amplified $\approx 10^3$-fold in 0.2 PCR droplet.
- SGAP is expressed in droplets by fusing 2 pL PCR droplets with 25 pL expression droplets.
- SGAP enzymatic assay can be performed by fusing 2 pL expression droplets with 20 pL substrate droplets.
- We developed a functional SGAP substrate characterized with limited leaking that can be improved with a newly developed rhodamine based hydrophilic substrate.

The last microfluidic device I needed to develop to complete the workflow was a 0.2-2pL droplet fuser device.

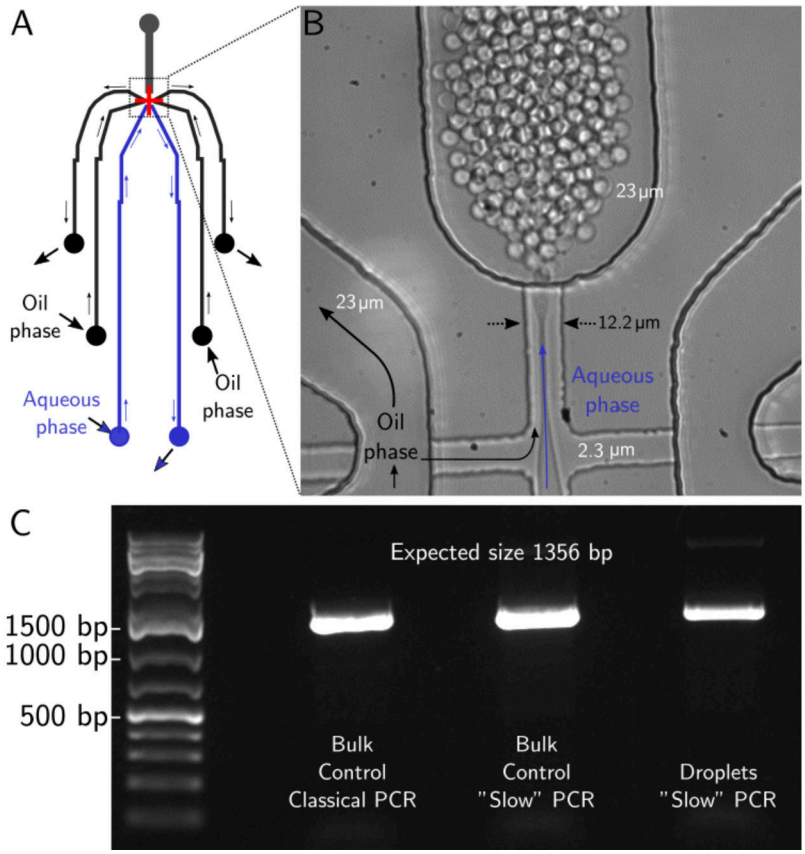


FIGURE 3.8: **0.2 pL droplet production and its application to droplet PCR. A**: 0.2 pL droplet production device design used to produce 0.2 pL droplets. Each phase is plugged to an open channel to prevent particles from entering the inner cross (in red) where the step emulsification takes place. **B**: Zoom in on the inner cross channel during 0.2 pL droplet production. **In white**: Height of the channels. I adapted the microfluidic design from [103] by setting the diameter and the height of the inner cross channels respectively to 12.2 µm and 2.3 µm, to produce 0.2 pL droplets. **In black**: Oil phase. **In blue** water phases are put under pressure using Fluigent pressure pumps with the following respective pressures 1500 mBar and 600 mBar. **C**: Electrophoresis gel after PCR in 0.2 pL droplets. A high concentration of plasmid was used ($\lambda = 5$) to make sure each droplet was containing at least one SGAP gene: the result of the PCR in droplets is representative of the PCR efficiency, as no PCR mix should have been lost in empty droplets. After thermocycling, the emulsion was broken, the aqueous phase was recovered and loaded on a gel (agarose, 1%) to assess the yield compared to bulk "slow PCR" and classical PCR. Band intensities show that amplification is approximately 800-fold and 2000-fold the amount of template in the initial PCR mix respectively in droplets and bulk.

3.3.4 0.2-2pL droplet electro-coalescence development

Achieving proper electro-coalescence requires pairing one to one droplets from the two emulsions that are to be fused. If at least one of the emulsions has to be reinjected, pairing requires reinjecting the droplets at a stable frequency by constraining ("squeezing") them into their reinjecting channel in a single file. As a result, the typical dimensions of the reinjection channels of an electro-coalescence device should be smaller than the diameter of the droplets. The diameter of spherical 0.2 pL and 2 pL droplets are respectively 7.2 µm and 15.6 µm. The latter dimension is small enough to make the devices prone to clogging. Also, the electro-coalescence channel should be small enough to slow down the 2 pl droplets by friction with the channel's walls and long enough to enable the faster smaller 0.2 pL droplets to catch up and come in contact with the 2 pL droplets, as contact is necessary for the two droplets to fuse. As it is possible to produce channels with different heights within one device and to produce or reinject the droplets containing the expression mix, I envisioned and tested the following solutions:

1. All channels at the same height. Produce 2 pL expression droplets in the electro-coalescence device.

2. All channels at the same height. Produce 2 pL expression droplets off chip and reinject them into the electro-coalescence device.

3. Low height channels for the reinjection of the 0.2 pL droplets. Producing 2 pL expression droplets in the electro-coalescence device.

4. Low height channels for the reinjection of the 0.2 pL droplets. Producing 2 pL expression droplets off chip and reinjecting them into the electro-coalescence device.

1. and 3. were not functional as the length of the narrow channel necessary for the pairing of the 0.2 pL droplets with the 2 pL droplets was increasing the resistance to a point where 2 pL droplet production was impaired.

2. and 4. gave better results. The limiting factor proved to be the small dimensions of the 0.2 pl reinjection channel, which was prone to clogging. A new version with better filters should be designed at the inlets to prevent the particles from entering further into the chip.

3.4 Discussion, Conclusion and Perspectives

In this chapter dedicated to the development of a cell-free expression work-flow to perform genotype phenotype mapping of the enzyme SGAP, I have shown the following:

- The PCR reagents are inhibiting cell-free expression and the cell-free expression reagents are inhibiting SGAP aminopeptidase activity. Those incompatibilities led me to develop a new microfluidic work-flow where the droplet of each step is diluted into the droplet of the next step: the PCR droplet is diluted into the cell-free expression droplet

which is diluted into the assay droplet. Bulk experiments results suggest that a ten fold dilution is enough to detect the SGAP activity.

- I started developing a new work-flow with femtoliter droplets and successfully designed a chip to produce 0.2 pL droplets in collaboration with Dr. Leman at the MEMS laboratory (ESPCI Paristech).
- I confirmed that the SGAP gene can be PCR amplified in droplets, that SGAP can be expressed using the cell-free expression system in droplets and that diluting 2 pL cell-free expression droplets into 20 pL assay droplets enabled the detection of aminopeptidase activity with a high contrast compared to a negative control.
- I developed in collaboration with Dr. Fenneteau a new non leaky hydrophilic substrate compatible with SGAP assay in droplets, that could prove very useful to probe SGAP variants displaying weak aminopeptidase activities and for which long time of incubation would be required.
- I prepared a first library of SGAP mutants (not mentioned in the body of this chapter) by random mutagenesis and determined the conditions to obtain between 1 and 3 amino-acid substitutions per SGAP gene (see appendix A.1.9).

The in vitro work-flow revealed more difficult to set up than expected, mainly because of the miniaturization. Retrospectively, starting the work-flow with bigger droplets would have been a better solution. Starting with 1 pL droplets and assaying 100 pL droplets might have required designing new production, electro-coalescence and sorting chips (with a maximum sorting frequency divided by 5 compared to 20 pL droplets, according to [198]), but it might have been more robust and easier to set up.

Nevertheless, further development on the 0.2-2 pL electro-coalescence device should allow to set up, to my knowledge, the first high-throughput microfluidic work-flow to assay SGAP using femtoliter droplets.

I present in Chapter IV an in vivo work-flow which was developed in parallel and which revealed highly promising and more robust than this cell-free work-flow. Also, preliminary results I present in the Chapter IV show that it could be used to perform high resolution genotype-phenotype mapping on SGAP. Therefore, the development of the cell-free microfluidic work-flow was put on hold to focus on the in vivo work-flow presented in the next chapter.

Chapter 4

Development of an in vivo microfluidic workflow for the genotype-phenotype mapping of *Ratus norvegicus* trypsin

This chapter is dedicated to the development of an in vivo droplet-based microfluidic work-flow to perform a genotype-phenotype mapping of the Rattus norvegicus trypsin or rat trypsin (the motivation to study this model enzyme are detailed in the section 1.4.2).

Briefly, Halabi et al.[34] hypothesized that each of the three rat trypsin sectors might be an independent functional unit controlling the catalytic efficiency, the specificity or the stability of the protein. Preliminary mutagenesis data on a few residues comprised in those sectors are comforting this hypothesis. Those data show that mutating a particular sector only impairs the biochemical property associated with the sector. They also show that epistasis is weak between residues belonging to different sectors and that it is high within a sector. Developing a high-throughput quantitative phenotype-genotype mapping approach for the rat trypsin protein will allow to systematically test this hypothesis, by measuring the enzymatic activity, the specificity and the stability of each single point mutant ($\approx 4000 mutants$) and all double point mutants (≈ 9 millions mutants).

Dr. Halabi, Dr. Russ and Dr. Bonventre (Green Center for Systems Biology, UT Southwestern, Dallas, Texas) developed a periplasmic rat trypsin expression system based on *E. coli*. My attempt to adapt it to a droplet-based microfluidic work-flow was unsuccessful. The associated results are nevertheless presented in this manuscript (see section 4.1).

Facing the difficulties imposed by the *E. coli* expression system, I changed for another bacterial host: *Bacillus subtilis*. This bacteria successfully secretes the rat trypsin protein as a fusion protein with a fluorescent reporter, mCherry. The mCherry reporter can be used to measure trypsin expression level and normalize trypsin activity so as to decouple the activity from the expression level. I developed a microfluidic work-flow compatible with this expression system and show that it is possible to detect both the trypsin activity and the reporter fluorescence in droplets. Droplet-based enzymatic assays performed with this work-flow on a subset of rat trypsin mutants, shows that it is possible to select trypsin variants based on their catalytic

efficiency, independently of the heterogeneous expression level displayed in droplets (see section 4.2).

In parallel, I developed a method to generate all single point mutants of the rat trypsin protein, using saturated mutagenesis by inverse PCR ("Around the horn site-directed mutagenesis"). The first library generated should allow me to measure the catalytic efficiency of $\approx 80\%$ of all the single point mutants in a few droplet-based microfluidic experiments (see section 4.3).

4.1 The in vivo workflow 1: rat trypsin periplasmic expression in *E. coli*

Dr. Halabi, Dr. Russ and Dr. Bonventre (Green Center for Systems Biology, UT Southwestern, Dallas, Texas) had already developed an expression system for the rat trypsin protein, based on periplasmic expression in *E. coli*. A periplasm export signal peptide and the rat trypsin gene had been cloned as a fusion protein downstream of the lac and T7 promoters in an expression vector (pET 28 plasmid). I transformed this plasmid in an *E. coli* strain which expresses the T7 polymerase under the control of arabinose. Upon induction with arabinose and IPTG, the rat trypsin protein is expressed and exported, with the signal peptide, to the periplasm. The signal peptide is cleaved during the transmembrane transport resulting in a mature and active rat trypsin protein.

My first efforts aimed at adapting this expression system to droplet based enzymatic assays according to the following work-flow (4.1):

- Transforming *E. coli* cells with a library of trypsin variants.
- Inducing a fresh culture of transformed *E. coli* cells by adding arabinose and IPTG in the culture medium. At that point, the OD should guarantee the encapsulation of at most one cell per droplet.
- Encapsulating the cells in 2 pL droplet.
- Incubating the emulsion to let the cells divide and express the rat trypsin protein.
- Fusing the 2 pL droplets with 20 pL droplets containing a cocktail of substrates in an osmotic shock buffer, to release the trypsin protein out of the periplasm or to make it accessible to the substrate.
- Incubating to let the enzymatic reaction occur.
- Measuring the substrate fluorescence intensity and normalize it by a proxy for the trypsin concentration to obtain the catalytic efficiency at a given concentration of substrate. Sorting the droplets and their content based on the catalytic efficiency. Sequencing the content of the sorted droplets to establish the relationship between catalytic efficiency and genotype.

The microfluidic work-flow I envisioned is presented in figure 4.1.

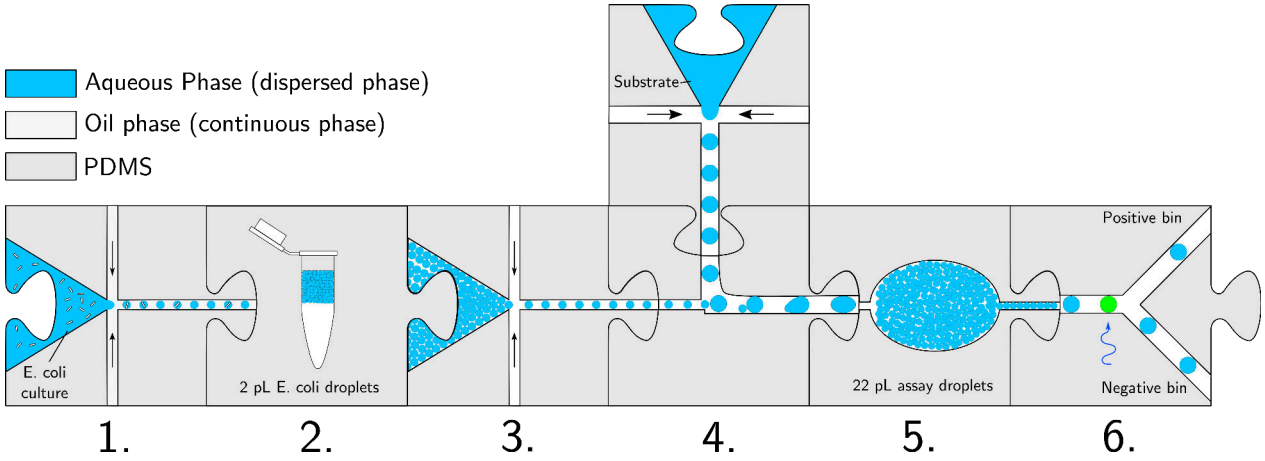


FIGURE 4.1: In vitro microfluidic work-flow envisioned for the *E. coli* expression system.

Step	Description
1.	*E. coli* cells encapsulation (at most, one cell per droplet, in presence of inducer) in 2 pL droplets and collection off chip.
2.	Rat trypsin expression in the periplasm of the encapsulated cells.
3.	Reinjection of 2 pL droplets containing the cells and the trypsin in an electro-coalescence device.
4.	Electro-coalescence of the 2 pL droplets with 20 pL droplets containing the substrate in a periplasmic extraction buffer.
5.	Incubation for the enzymatic activity to occur in delay lines.
6.	Fluorescence measurement and droplet sorting.

All the following results regarding trypsin expression in E.coli were performed in bulk.

4.1.1 Osmotic shock of E.coli cells in hypotonic buffer allows trypsin activity detection in bulk

Following the protocol suggested by Dr. Eric Bonventre (see B.1.3), I induced cells transformed with the expression vector containing the wild type trypsin gene. After induction, the trypsin activities of the supernatant and of the resuspended pellet were measured by diluting them twice in an assay mixture containing a fluorogenic trypsin substrate based on coumarin. Fluorescence measurements were then performed. The results are depicted in figure 4.2. The details of this protocol are available in the paragraph B.1.3.

The culture supernatant shows little activity compared to the other conditions. The highest trypsin activity level is detected when the cells are flash frozen in isotonic buffer or resuspended in hypotonic buffer. This result was nevertheless promising as diluting cells in a hypotonic buffer can be easily adapted to the droplet format.

Further experiments using harsher conditions (lysozyme and sucrose in hypotonic conditions) led to marginal improvements regarding the activity level (see figure 4.3), indicating that the lysozyme activity does not improve substrate capacity to diffuse into the periplasm where it is processed by trypsin. Protocol details are available in the paragraph B.1.4.

With those successful protocols to measure trypsin activity from *E. coli* cells, I undertook to develop a suitable method to measure trypsin concentration in the culture medium, in order to decouple trypsin activity product fluorescence from its expression level. I first tested a compound (MUGB) to titrate the concentration of trypsin.

4.1.2 MUGB inhibits trypsin activity but cannot be used as a reporter for its concentration

Quantitatively measuring catalytic efficiencies in droplets requires normalizing the product of the trypsin activity with a proxy for its concentration. 4-methylumbelliferyl p-guanidobenzoate (MUGB) is a trypsin inhibitor releasing a fluorescent methyl coumarin compound upon binding with the trypsin active site. Halabi et al. [34] successfully used that compound to stop trypsin activity and to determine its concentration, when they measured the mutational effects of sector mutations on the rat trypsin. MUGB is not an ideal compound as it does not bind all trypsin variants with the same affinity [34]. Nevertheless, it is easy to use and offer the opportunity to better control the timing of the assay by stopping trypsin activity. I tested its compatibility with the *E. coli* expression system.

Using a protocol already detailed in the previous paragraph, the cells were induced and used to perform enzymatic assays in presence of a trypsin substrate. In the course of the enzymatic reaction, MUGB was added to the assay mixtures to different final concentrations. The activity was measured

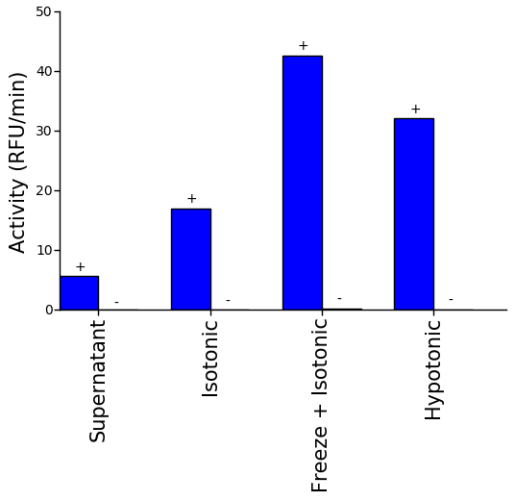


FIGURE 4.2: **Trypsin activity for different periplasmic extraction protocols.** Overnight cultures of *E. coli* cells transformed (+) or not transformed (-) with the expression plasmid containing the wild type trypsin gene were diluted 100 times in fresh culture medium. Once the OD reached 0.5, the cells were induced by adding arabinose and IPTG in the culture medium. After 1.5 h of incubation at 37°C the following conditions were applied: **Supernatant**: Supernatant is diluted twice in a solution of substrate in isotonic buffer. **Isotonic**: Cells are resuspended in isotonic buffer and diluted twice in a solution of substrate in isotonic buffer. **Freeze+Isotonic**: Cells are resuspended in isotonic buffer, flash frozen and diluted twice in a solution of substrate in isotonic buffer. **Hypotonic**: Cells are resuspended in hypotonic buffer and diluted twice in a solution of substrate in hypotonic buffer. Best results are obtained when the cells are lysed via flash freezing or with an osmotic chock in hypotonic buffer. This shows that when the cells are not treated, most of the trypsin remains in the periplasm.

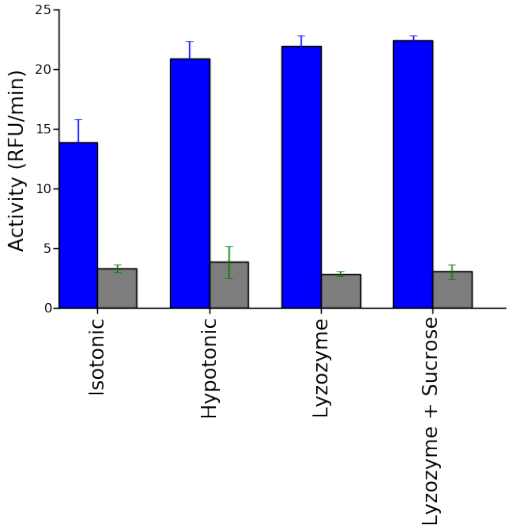


FIGURE 4.3: **Trypsin activity for different periplasmic extraction protocols.** Overnight cultures of *E. coli* cells transformed with the expression plasmid containing the wild type trypsin gene were diluted 100 times in fresh culture medium. Once the OD reached 0.5, the cells were induced (blue bars) or not induced (grey bars). After 1.5 h of incubation at 37°C the cells were resuspended in isotonic buffer to an OD of 4. They were diluted ten times in the following assay mixtures: **Isotonic**: Substrate in isotonic buffer. **Hypotonic**: Substrate in hypotonic buffer. **Lyzozyme**: Substrate in hypotonic buffer containing lysozyme. **Lysozyme + Sucrose**: Substrate in hypotonic buffer containing lysozyme and sucrose. Lyzozyme and lyzozyme + sucrose conditions only marginally increase the magnitude of the detected trypsin activity compared to hypotonic conditions.

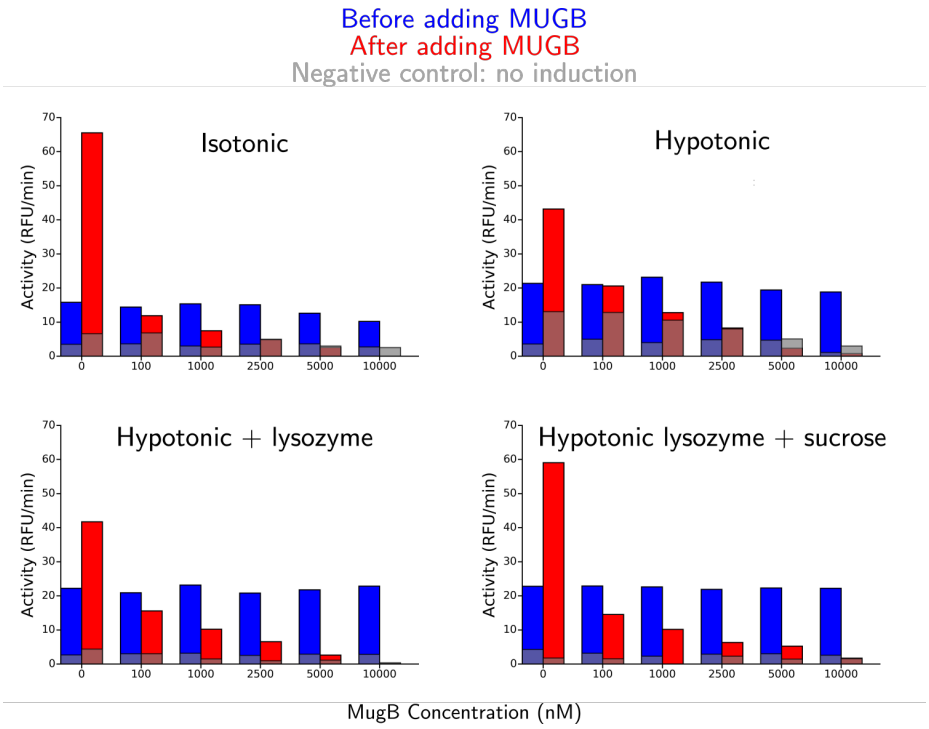


FIGURE 4.4: **Trypsin activity before and after adding MUGB**. The conditions are identical to the conditions described in 4.3. After 60 min of enzymatic assay, MUGB was added to the assay mixtures to a final concentration of 0, 100, 1000, 2500, 5000 and 10000 nM. The activity was measured before (blue bars) and immediately after MUGB addition (red bars). The activity of the negative controls (not induced, grey bars) was also measured. Addition of MugB clearly inhibits the trypsin activity in every conditions, to the point where the trypsin activity is no longer detectable with 10 µM of MUGB). After MUGB addition, the activity of the negative controls increase (difference between red and blue bars at 0 nM MUGB). Between the last activity measurement before MUGB addition and the first measurement following MUGB addition, a few minutes passed. During those few minutes the slope of the fluorescence evolution over time increased and therefore the activity (in RFU/min) increased. The non linearity of the fluorescence profile is due to the fact that the substrate used has two trypsin reactive sites and that the substrates hydrolyzed once and twice both display fluorescence.

before and after MUGB addition to check if trypsin was indeed inhbited (see figure 4.4). The evolution of the coumarin fluorescence was also observed before and after MUGB addition to test if it could be used to measure the concentration of trypsin in each condition (see figure 4.5). The detailed protocol is available in paragraph B.1.5.

Upon MUGB addition the trypsin enzymatic activity is indeed stopped compared to the negative control (no MUGB added on figure 4.4). But whether the cells are induced or not, the evolution of the coumarin fluorescence remain the same (figure 4.5). The use of lysozyme or lysozyme and sucrose does not improve the results. Thus, MUGB is not suitable compound to normalize trypsin activity. Given those results, I turned myself to another solution: assaying trypsin as a fusion protein with mCherry as a expression level reporter.

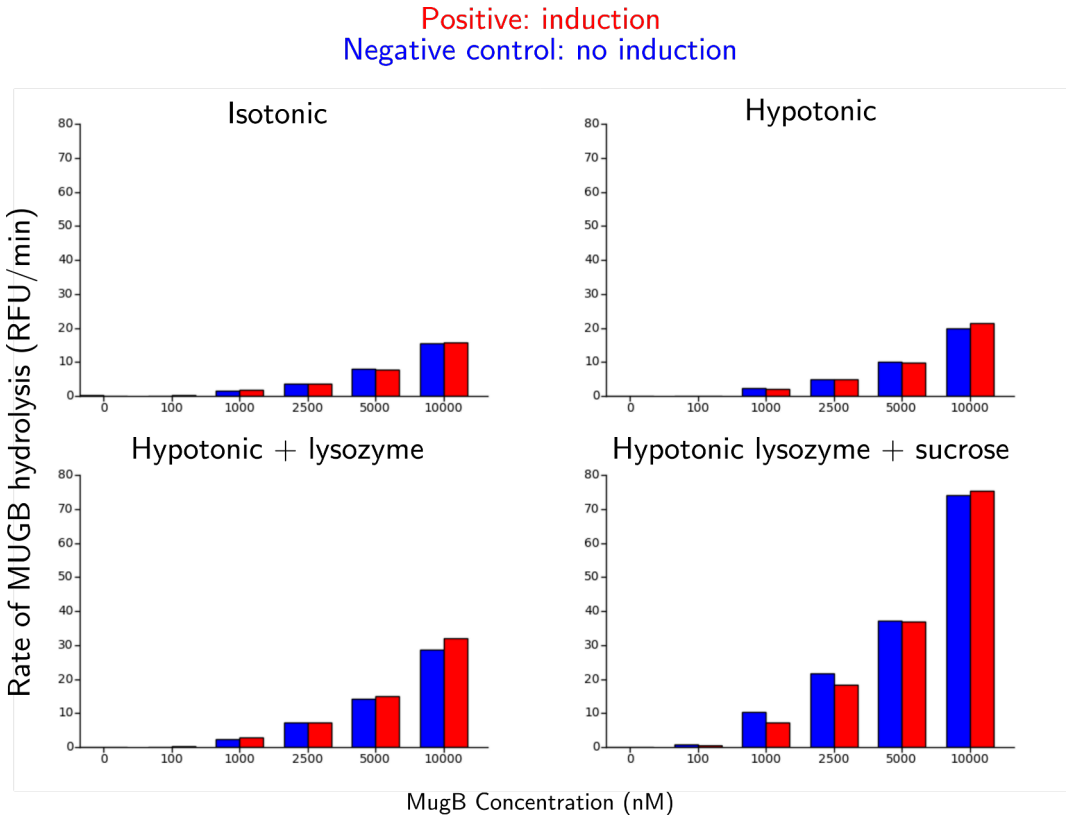


FIGURE 4.5: **MugB cannot be used as a reporter for trypsin expression level.** After 60 min of enzymatic assay, MUGB was added to the assay mixtures to a final concentration of 0, 100, 1000, 2500, 5000 and 10000 nM. The evolution of the fluorescence of the compound released by MUGB hydrolization was measured in the positive (red bars) and the negative controls (not induced, blue bars)). Unfortunately, negative and positive experiments show the same evolution of fluorescence. Therefore the fluorescence in the coumarin fluorescence channel and the detected activity are not related, preventing the use of MUGB as a method to normalize the trypsin activity.

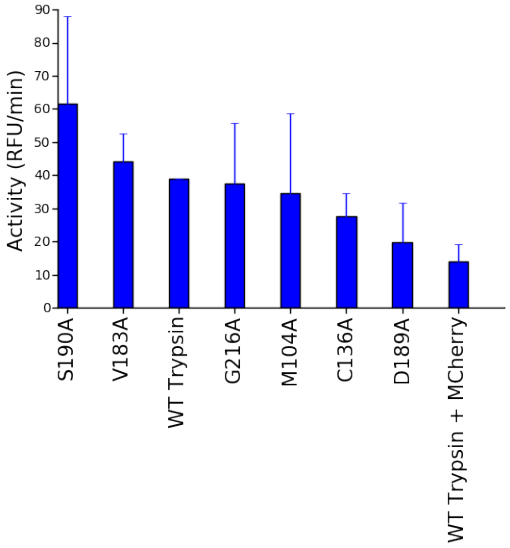


FIGURE 4.6: **Activity of a subset of rat trypsin variants (hypotonic conditions).** Hypotonic conditions were used to assess the activity of a subset of trypsin mutants and wild-type trypsin mCherry expressed as a fusion protein. All activities were normalized relatively to the wild type trypsin activity and multiplied by its mean to compare this experiment with the previous ones. The ranking between the different mutants is highly variable from one experiment to the other and the fusion protein has, on average, an activity more than 2-fold lower than the wild type trypsin alone. MCherry fluorescence of the trypsin mCherry fusion protein could not be detected in the culture medium.

4.1.3 Normalizing rat trypsin activity using mCherry as a reporter of its expression level

Normalizing the expression level using a fused fluorescent protein as a reporter has already been described in the literature. Sarkisyan et al. used mKate to normalize GFP fluorescence, [72], Tokuriki et al used GFP to normalize PON1 activity [23]. We chose mCherry for the following reasons: it is a monomeric protein fluorescing at 610 nm which is compatible with our microfluidic optical setup, it is less prone to photo-bleaching than other fluorescent proteins and it is characterized by a short maturation time ($t_{0.5} = 15$ min at 37°C) [199]. Besides, it has been shown that fusion proteins with mCherry were successfully exported to the periplasm in *E. coli* [200].

I cloned a glycine serine flexible linker and the mCherry gene downstream of the rat trypsin gene. In parallel, I received a set of trypsin mutants in the expression vector from Dr. Eric Bonventre. I performed enzymatic assays on those along with the trypsin mCherry fusion protein using the hypotonic conditions protocol (see results in figure 4.6, B.1.4 for the detailed protocol).

The trypsin activity can be detected in presence of mCherry but it is more than twice lower than the activity detected from trypsin alone. This suggests that mCherry is either inhibiting the trypsin activity, hindering substrate access to the enzyme or lowering the expression yield. Also, mCherry fluorescence cannot be detected in E.coli culture expressing the

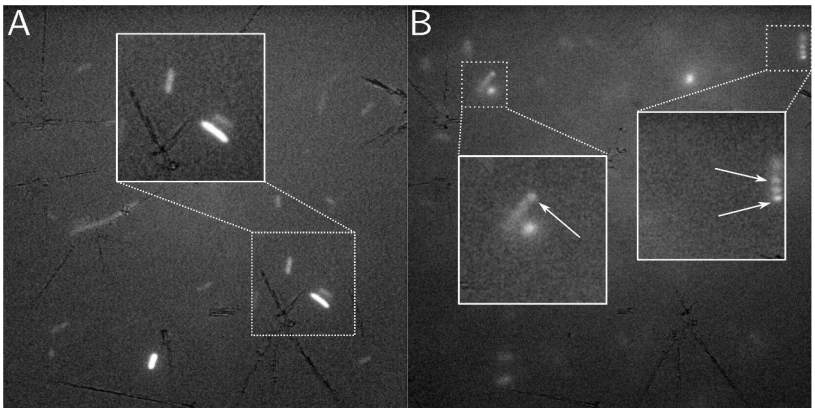


FIGURE 4.7: **MCherry and trypsin mCherry expression in *E. coli*.** Fluorescence microscope pictures of *E. coli* cells expressing either (**A**) the mCherry protein or (**B**) the trypsin mCherry fusion protein. A few microliter of induced *E. coli* cells culture were dropped on a microscope glass-slide and imaged on a fluorescence microscope (100X, λ_{ex}: 510-560 nm, λ_{em}: > 590 nm, 1 ms of exposure). Cells expressing the fusion protein display a lower fluorescence and do not show uniform fluorescence over their surface (white arrows in the insets), contrary to the cells expression mCherry. I interpret those as inclusion bodies, which might explain the reduced activity detected compared to trypsin alone.

trypsin mCherry fusion protein. Besides, the relative activity of the assayed mutants is highly variable from one experiment to the other.

Microscopy experiments confirmed that *E. coli* is not a suitable host to express the fusion protein. We also developed a non leaky trypsin FRET substrate which revealed non suitable to detect trypsin activity as expressed by *E. coli* cells. Those difficulties are discussed below. They convinced us to change for another expression system based on B.subtilis.

4.1.4 Difficulties with the *E. coli* expression system

The presence of inclusion bodies upon trypsin mCherry expression To understand the low level of activity displayed by trypsin mCherry, I performed fluorescence microscopy experiments to visualize how the trypsin mCherry fusion protein is distributed within *E. coli*. Induced *E. coli* expressing either mCherry alone or trypsin mCherry were imaged on a fluorescence microscope set on the mCherry fluorescence channel (see figure 4.7). In the case of the fusion protein, mCherry fluorescence is lower in intensity and localized inside inclusion bodies which are absent when mCherry is expressed alone (white arrows in the pictures). I did not investigate further to test if the inclusion bodies are also present in the case of trypsin alone. My interpretation is that lower expression level and the presence of the inclusion bodies explain the low trypsin mCherry activity as part of the enzyme is not in the cell periplasm.

MCherry fluorescence cannot be easily measured with our optical setup in droplets It is not convenient to measure the mCherry fluorescence with our optical setup. The mCherry fluorescence is sequestrated within the

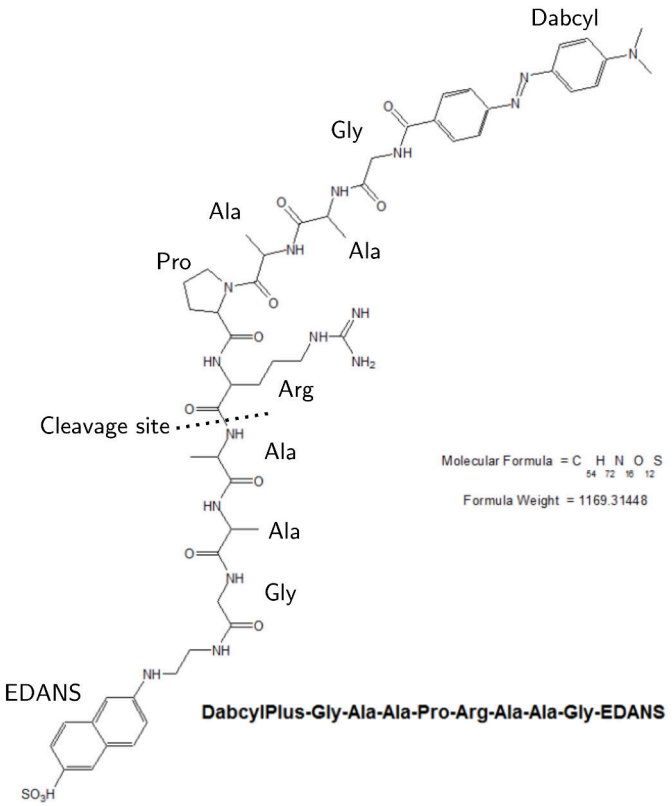


FIGURE 4.8: **EDANS-Dabcyl substrate structure.** When the peptide chain is intact, the Dabcyl quenches the EDANS fluorescence. When the trypsin protein catalyzes the cleavage of the peptide chain at the carboxyl side of the arginine residue, the quenching is suppressed and the EDANS fluorescence can be detected ($\lambda_{ex} = 352nm$ $\lambda_{em} = 510nm$).

cells, which implies that they would have to cross the detection point for mCherry fluorescence to be measured. But *E. coli* cells are at least 14 times smaller than 2 pL droplets in terms of diameter, which makes this method not suitable. I tried more rigorous conditions to extract trypsin mCherry from the cells, but I was unsuccessful to recover more activity or mChery fluorescence than with the hypotonic protocol[1].

***E. coli* culture is not reactive against a non leaky trypsin FRET substrate**
The only commercialized fluorogenic substrates to specifically detect trypsin activity are based on leaky fluorophores such as coumarin or rhodamine. Dr. Sandeep Ameta, a post-doctoral researcher in our laboratory, synthesized a new substrate based on a trypsin cleavable peptide chain separating the EDANS DABCYL FRET pair. This type of substrate is known not to leak between droplets and is therefore suitable for trypsin enzymatic assay in droplets [171]. The sequence and structure of this substrate are detailed in figure 4.8.

I tested both the stability and the reactivity of this substrate against trypsin activity. I performed a simple enzymatic assay experiment by adding

[1]Sucrose, EDTA, lysozyme, polymixin B and osmotic shock in presence high concentration of magnesium.

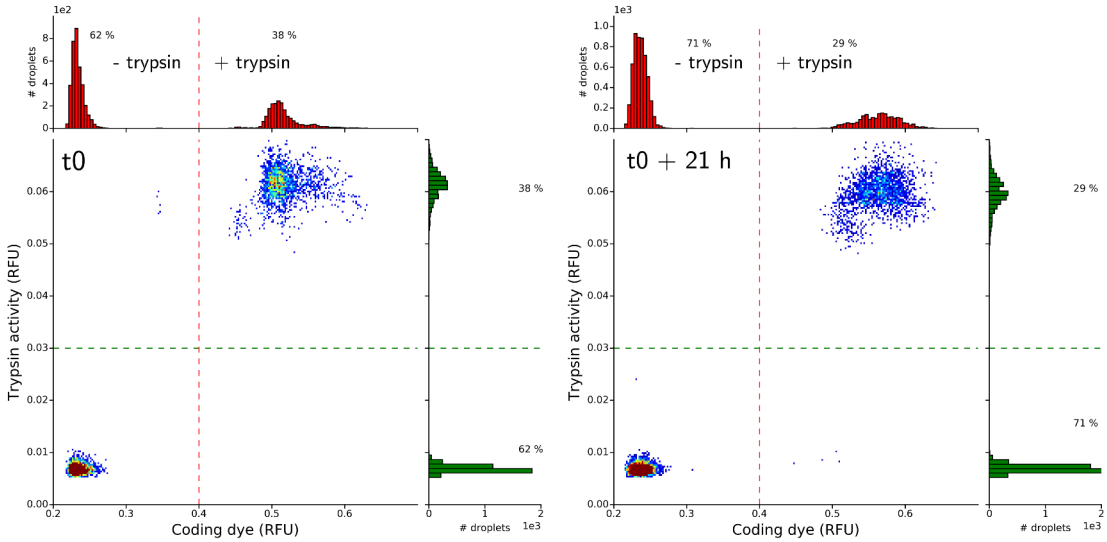


FIGURE 4.9: **The EDANS Dabcyl FRET substrate is cleaved by trypsin activity and not leaky between droplets.** Assay mixtures containing the EDANS-Dabcyl substrate in presence (+ trypsin) and in absence (- trypsin) of purified trypsin were prepared. Once the substrate was completely hydrolyzed and fluorescence no longer increasing, coding dye was added to the two solutions at low and high concentrations for the "- trypsin" and the "+ trypsin" solutions respectively. Solutions were emulsified simultaneously and collected in the same tube. Once collection was over, the resulting emulsion was immediately reinjected and fluorescence distribution was measured in the coding dye and EDANS-Dabcyl fluorescence channels. The emulsion was also reinjected after 21 h of incubation. As the fluorescence distributions did not change between t_0 and $t_0 + 21h$, I conclude that no detectable exchange occurred between droplets. I note a change in the relative fractions of the two emulsions (positive emulsion: from 38% to 29%) which, I believe, is due to local fluctuations in the composition of the stacked emulsion in the collection tube. Those data show that the substrate is indeed cleaved by the trypsin activity and that it not exchanged between droplets.

purified trypsin in a FRET substrate solution. As expected, an increase of fluorescence intensity was detected whereas none could be detected in the absence of the enzyme. I confirmed that the substrate was not exchanged in emulsions during incubation. Both with and without trypsin assay mixtures were emulsified separately and mixed together. Their fluorescence distribution was measured immediately after encapsulation and after 21 h of incubation (see figure 4.9). The protocol is detailed in paragraph B.1.6.

Unfortunately, I could not detect any trypsin activity using induced *E. coli* cells expressing the rat trypsin, with this substrate. I was equally unsuccessful with another protease substrate based on bodipy and casein (EnzCheck E6638, Bodipy-FL), which suggests that the substrate size might prevent it from diffusing into the periplasm of the cells expressing the trypsin enzyme.

Facing those difficulties with *E. coli*, I switched to an engineered *B. subtilis* strain to take advantage of *B. subtilis* secretion abilities.

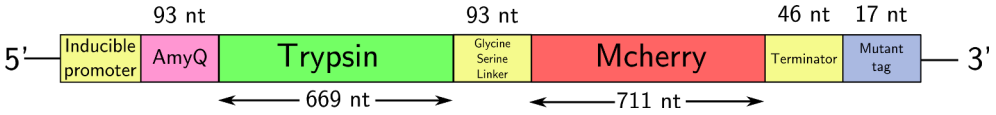


FIGURE 4.10: **Partial map of the trypsin mCherry *B. subtilis* expression vector.** The promoter (Pgrac) is ITPG inducible. The amyQ sequence codes for a signal peptide that targets expression by secretion. Trypsin and mCherry sequences are linked by a glycine serine flexible linker. The mutant tag cassette is meant to comprise a short sequence coding for the trypsin genotype.

4.2 The in vivo workflow 2: trypsin secretion by *B. subtilis*

Bacillus species are well known for their capacity to secrete proteins. They are widely used in the industry to produce enzymes [201]. As mentioned in the introduction, a good expression system should not display unspecific activity towards the assayed substrate. Wu et al. and Nguyen et al. built WB800N, a *B. subtilis* strain knocked out for a total of 8 extracellular proteases [202, 203] (δnprE δnprB δaprE δepr δmpr δbpr δvpr δwprA). It has been shown that it presented a low background activity towards a subtilisin substrate [204]. In the following, I compare the expression capacity of this engineered strain to the behaviour of a classical domesticated strain, *B. subtilis* D168.

Stable expression vector dedicated to *B. subtilis* have been previously designed [205, 206]. We chose to work with pHT43, a 8 kb plasmid comprising an expression cassette to induce the secretion of recombinant proteins using the P_{grac} inducible promoter and the amyQ secretion signal peptide sequence. This plasmid is also compatible with *E. coli* ("shuttle vector") which is used as a host for cloning.

I cloned the trypsin - glycine serine linker - mCherry fusion protein inside pHT43, downstream of the P_{grac} promoter and juxtaposed to the 3' end of the amyQ secretion signal peptide sequence. Upon induction with IPTG, the transformed *B. subtilis* cells are supposed to secrete the fusion protein in the culture medium. A partial plasmid map is depicted in 4.10. The complete plasmid map and the transformation protocol are available in appendices B.1.7 and B.2.2.

The workflow I envisioned for the *B. subtilis* expression system is depicted in figure 4.11.

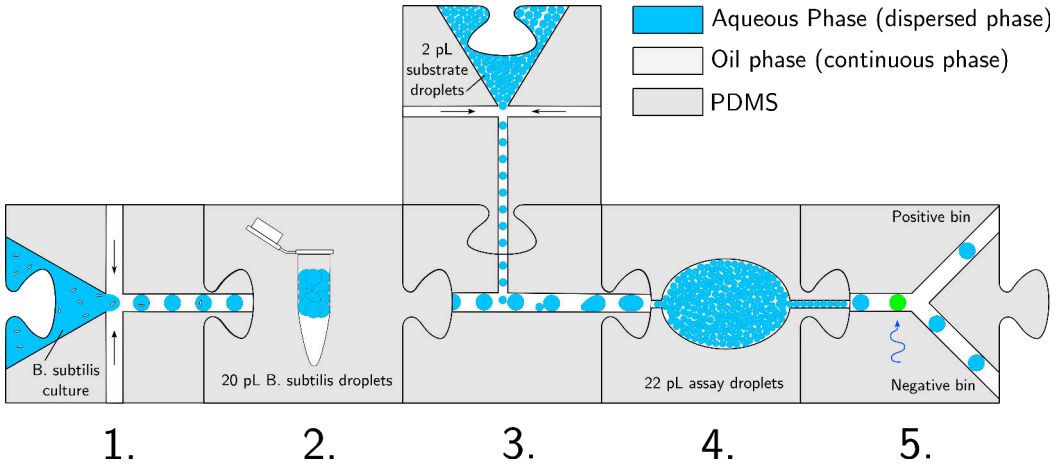


FIGURE 4.11: In vitro microfluidic workflow envisioned for the *B. subtilis* expression system.

Step	Description
1.	*B. subtilis* cells encapsulation (at most, one cell per droplet, in presence of inducer) in 20 pL droplets and collection off chip.
2.	Induction off chip: rat trypsin secretion in the droplets.
3.	Reinjection of 20 pL droplets containing the cells and electro-coalescence of the 20 pL droplets with 2 pL droplets containing the substrate
4.	Incubation for the enzymatic activity to occur in delay lines.
6.	Fluorescence measurement and droplet sorting.

4.2.1 The rat trypsin - mCherry fusion protein is secreted by WB800N as a full protein in the supernatant

The rat trypsin protein possesses an autolysis activity [207][2]. To validate that the trypsin mCherry fusion is correctly secreted by WB800N, I mutated the catalytic serine residue to an alanine (inactive trypsin S195A) and cloned both active and inactive trypsin sequences in the expression vector. The latters were transformed into two different strains of *B. subtilis*: D168 and WB800N (see transformation protocol in appendix B.2.2). I carried out western blots on the supernatants of D168 and WB800N cultures using anti-mCherry antibodies (see 4.12).

As seen on figure 4.12, **A**: "WB800N trypsin S195A", part of the detected mCherry is linked to the full rat trypsin protein (band at $\approx 60kDa$), but most of the mCherry is either alone, or linked to a small C terminal trypsin fragment. I noticed the presence of an arginine at the C terminus of the flexible linker between the trypsin and the mCherry. Moreover, the wild type mCherry sequence displays two lysines in an unfolded N terminal region. I got rid of the arginine in the linker and mutated those two lysines to alanine residues (mCherry K3A K13A).

I performed new western blots according to the same protocol (see figure 4.12, **B**). In the case of S195A in WB800N, most of the mCherry detected is linked to the rat trypsin, as it presents the expected full protein size ($\approx 60kDa$). In the case of the active trypsin in WB800N, the band at 60 kDa is replaced by two smaller bands: mCherry is linked to two smaller trypsin fragments. This is consistent with the fact that the rat trypsin possesses two autolysis sites [207]. The lowest bands (around 35kDa) are interpreted as the result of residual WB800N activiy which partially hydrolyzes the C terminal region of the trypsin or the linker ("Mcherry", "Trypsin fragment 3"). In the D168 case, no trypsin fragments could be detected confirming the presence of a protease activity towards the secreted trypsin. This strain is therefore not suitable for our purpose.

mCherry is robust to trypsin activity, D168 and WB800N residual activities as no bands lower than the mCherry band can be seen on the gels.

Following those results, I kept the linker with no arginine residues, K3A K13A mCherry and the WB800N *B. subtilis* strain for further expression experiments.

[2]Trypsin proteases resistant to autolysis are available. I failed to clone one of them (Streptomyces erythraeus trypsin [208]) in the expression vector. Only clones displaying nonsense or misense mutations could be isolated on plate. This is certainly due to its high toxicity: it shows no autolysis and is 10 times more active than the rat trypsin protein.

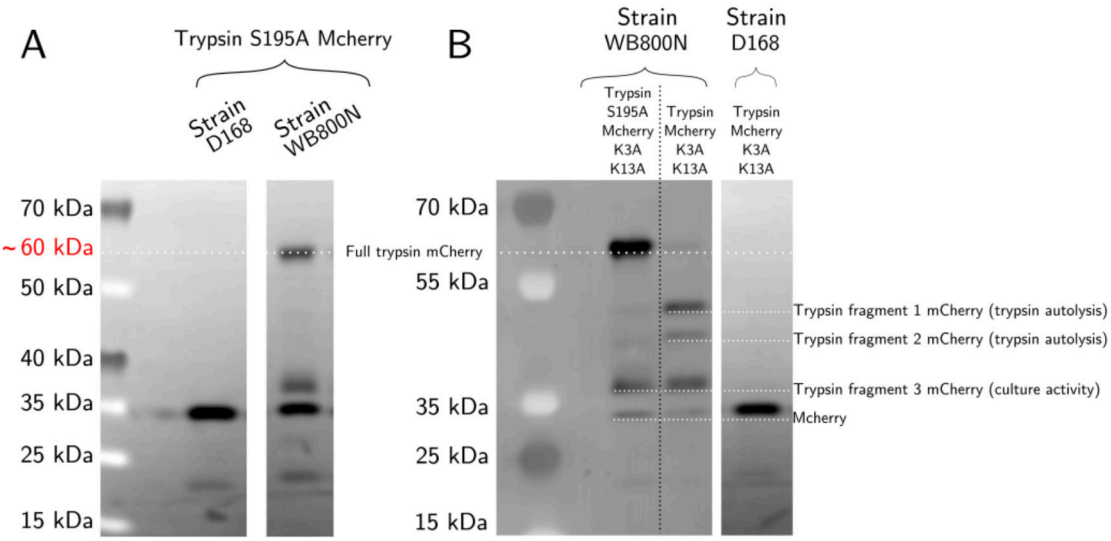


FIGURE 4.12: **Anti-mCherry western blot performed on the supernatants of induced culture of *B. subtilis***. **A**: Using an anti-mCherry antibody, I performed western blots on the supernatants of induced cultures of *B. subtilis* (see protocol in appendix B.2.7). D168 supernatant mostly displays a bright band at 35kDa. On the other hand, WB800N shows three bands: a band at approximately 60 kDa, close to the expected full trypsin mCherry size (54 kDa), a band above 35 kDa and the brighest band at 35 kDa. Other western blots performed on WB800N and D168 only secreting mCherry showed only one bright band at 35kDa (data not shown), suggesting that this 35 kDa bright band accounts for mCherry alone (molecular weight: 26 kDa). I hypothesize that the shallow band above 35 kDa is the result of the activity of WB800N in the C terminal region of the trypsin sequence. I interpreted the highest band ($\approx 60kDa$) as accounting for the full trypsin Mcherry fusion protein. D168 does not display such band, suggesting at the least, hydrolysis at the Cter region of trypsin flexible linker, upon expression or in the extracellular medium of D168 culture. **B**: An arginine residue in the linker and the first two lysines of the mCherry protein were mutated. In the case of the inactive trypsin in WB800N, the brighest band ("Full trypsin mCherry") displays the expected size: in absence of trypsin autolysis, most of the mCherry is physically linked to the full trypsin. In the case of the active trypsin in WB800N, two smaller bands replaced the band at 60 kDa. This is consistent with the fact that the rat trypsin protein possesses two autolysis sites [207], leaving two C terminal fragments (Trypsin fragment 1: 169 amino-acids, Trypsin fragment 2: 113 amino-acids) linked with mCherry. None of those bands are visible in the case of D168, demonstrating that this strain is not suitable for our purpose. Again, bands accounting for mCherry alone (35 kDa, "Mcherry") and accounting for WB800N activity in the C terminal region of trypsin are visible ("Trypsin fragment 3 mCherry (culture activity)"). The combination of WB800N and the mutated expression vector display suitable band pattern on the western blot and were selected for further investigations.

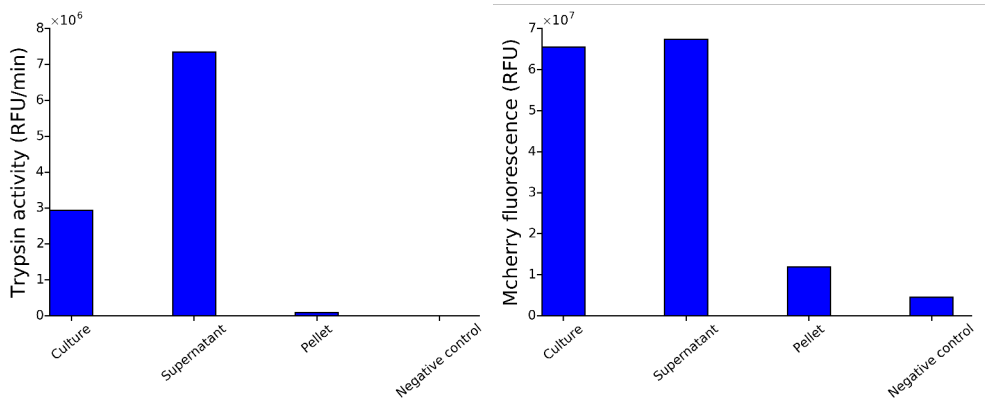


FIGURE 4.13: *B. subtilis* cells were induced overnight for trypsin mCherry secretion in presence of IPTG, in 2xYT medium. In the morning, a trypsin substrate solution was diluted ten times in either the culture, the supernatant or cells resuspended in 2xYT medium. The fluorescence in the substrate and the mCherry channels were measured simultaneously. The results show that the vast majority of the trypsin activity (**A**) and mCherry fluorescence are in the supernatant. Interestingly, the activity of the supernatant is almost three times higher than the culture activity.

4.2.2 The rat trypsin - mCherry fusion protein is fluorescent and enzymatically active in the culture medium

I then confirmed that the rat trypsin - mCherry fusion protein was secreted, active and fluorescent in the culture medium. Following the protocol detailed in appendix B.2.3, I induced overnight *B. subtilis* cells transformed with the new expression vector containing the rat trypsin mCherry fusion protein sequence. I measured the trypsin activity and the mCherry fluorescences of the culture (whole culture, supernatant and pellet). Results are depicted in figure 4.13 and the protocol for the enzymatic assay is available in appendices B.2.5.

The vast majority of the trypsin activity and mCherry fluorescence are detected in the supernatant. Interestingly, trypsin activity detected in the supernatant ($\approx 7.10^6$ RFU/s) is more than twice the trypsin activity detected in the full culture ($\approx 3.10^6$ RFU/s). Peptides at the surface of the cells might compete with the substrate for the enzymatic activity.

Lowering the induction temperature to 30°C leads to a 2-fold decrease in both trypsin activity and mCherry fluorescence, as seen in figure 4.14. Lowering the temperature cannot be used to increase the expression yield as it is sometimes done with *E. coli*.

I also assayed the trypsin activity with the previously synthesized EDANS-Dabcyl substrate.

Trypsin activity using the EDANS-Dabcyl FRET substrate with the WB800N expression system Unfortunately, the EDANS-Dabcyl showed limited signal to noise ratio compared to the rhodamine based trypsin substrate. Bulk enzymatic assay on induced cells, either with the EDANS-Dabcyl substrate or the rhodamine110 based substrate showed that those two substrates exhibited a respective contrast (difference of fluorescence intensity between

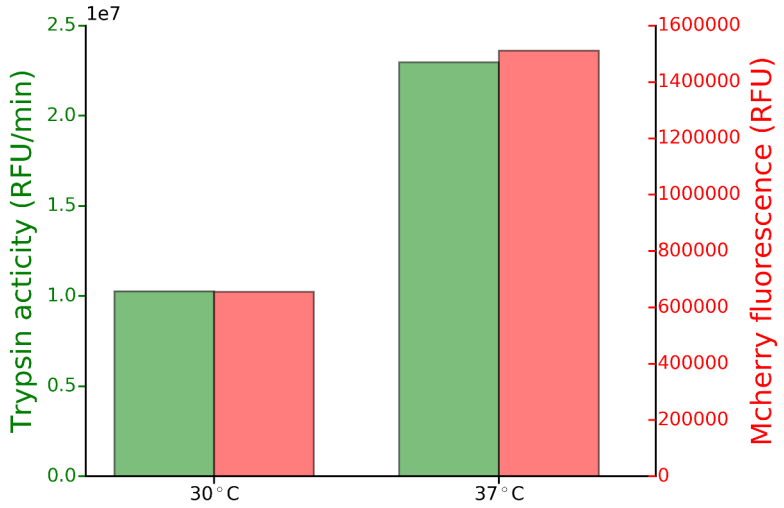


FIGURE 4.14: **Effect of the temperature on *B. subtilis* secretion levels.** *B. subtilis* cells were induced overnight for trypsin mCherry secretion in presence of IPTG 1 mM at 30 or 37°C. The cultures were assayed according to the protocol available in the appendix B.2.5. Inducing the cells at 30°C instead of 37°C reduces the amount of trypsin activity detected and the amount of mCherry fluorescence by 2 folds.

non and fully hydrolyzed substrate) of 4.5×10^8 and 2.2×10^7, corresponding to respective signal to noise ratios (ratio of fluorescence intensity between non and fully hydrolyzed substrate) of 224 and only 4.0. Besides, we noticed that the assays performed with the inactive trypsin mutant (S195A) exhibited twice as much activity in presence of EDANS-Dabcyl than with the rhodamine 110 substrate, which reduces even more the available dynamic range of the assay.

To sum up, the low dynamic range of the EDANS-Dabcyl substrate in the culture medium considerably limits the dynamic range of activities that can be detected. Therefore, I used the substrate based on the rhodamine 110 in further experiments, which hinders enzymatic assays with long incubation times.

4.2.3 The mCherry fluorescence can be used as a reporter of the trypsin expression level

In the prospect of normalizing the product of the trypsin activity in droplets, I assessed the correlation between trypsin activity and mCherry fluorescence. *B. subtilis* cells were induced overnight with different IPTG concentrations to vary the induction level. The trypsin activity of the cultures and their mCherry fluorescence were measured according to the protocol detailed in appendix B.2.5. Results are depicted in figure 4.15.

The western blot displays trypsin fragments mCherry bands whose intensity is increasing with increasing concentration of IPTG (4.15, A). A constant shallow band is visible at 60 kDa, even in the absence of IPTG. This band is certainly the cause of expression leakage, which also explains the non null activity detected in the absence of IPTG (≈ 0.2 RFU/min). I hypothesize that in absence of IPTG, smaller bands are not visible because the

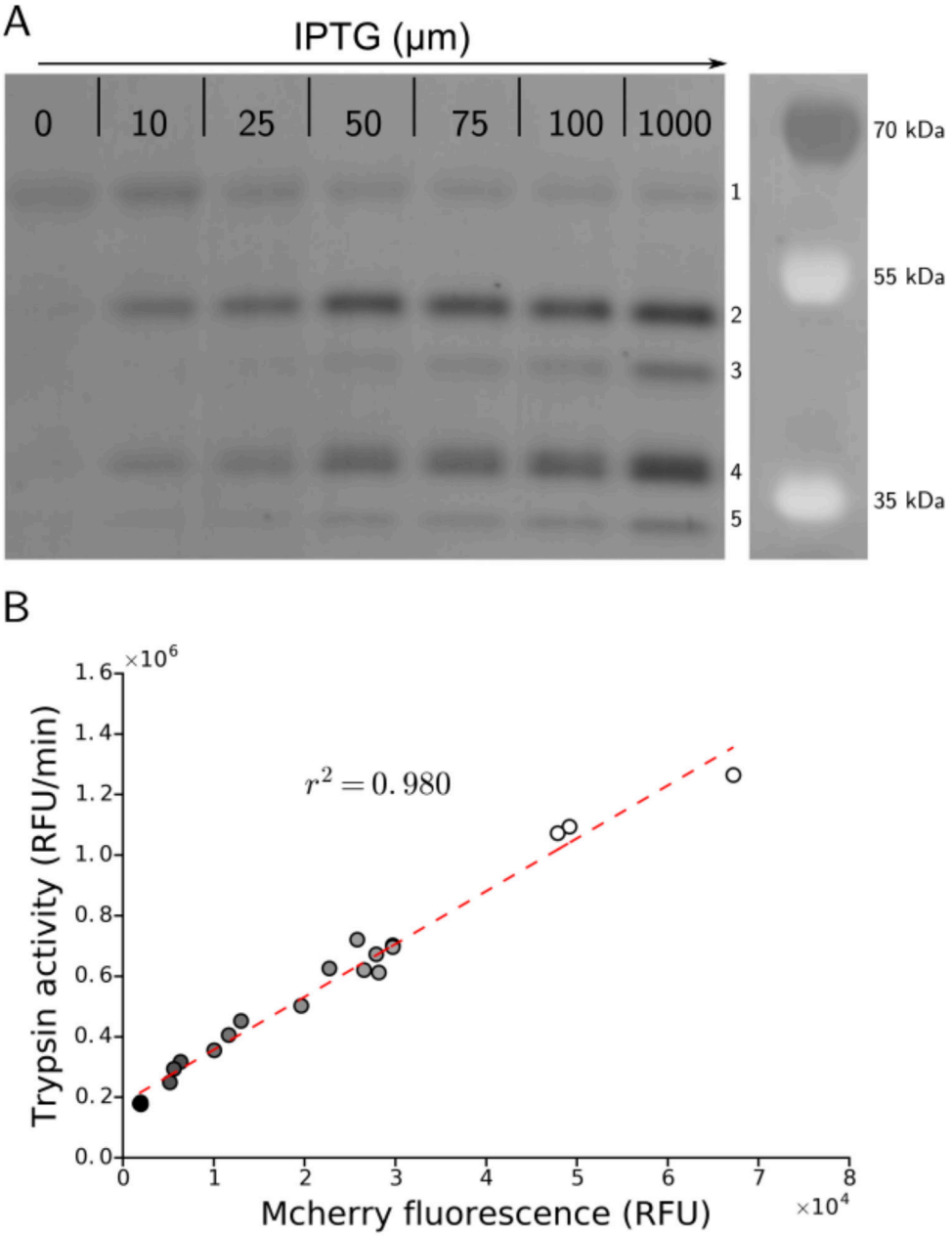


FIGURE 4.15: **Using mCherry fluorescence to normalize trypsin activity.** Effect of IPTG concentration on the secretion level **A** WB800N cells previously transformed with the wild type trypsin (WT) were induced overnight in presence of different concentrations of IPTG to vary the secretion levels. A western blot using anti-mCherry antibodies was then performed on the supernatants of the induced cultures. The intensity of all the bands accounting for the different C terminal trypsin fragments linked with mCherry (2,3,4 and 5) correlate with the IPTG concentrations, the mCherry fluorescence and the trypsin activity level (see B). Surprisingly, the intensity of upper band (1), that I interpret as the full trypsin mCherry fusion protein (approximately 60 kDa), is roughly constant, even in the absence of IPTG. I hypothesize that this is the product of the expression leakage, and that in absence of IPTG, the concentration of the trypsin-mCherry protein is not high enough to detect the effects of intermolecular autolysis. **B:** Trypsin activity versus fluorescence intensity in the mCherry fluorescence channel. The dots grey level indicates the final IPTG concentration that was used: 0 µM (black), 10 µM, 25 µM, 50 µM, 75 µM, 100 µM, 1000 µM (white). The strong linear correlation between the activity and the mCherry fluorescence allow to use mCherry as a proxy for the trypsin concentration using a simple affine equation

concentration of trypsin-Mcherry protein is not high enough to detect intermolecular autolysis products.

As expected, the relation between trypsin activity and mcherry fluorescence is linear over a ten-fold range of trypsin activity (4.15, B). This is coherent with Michaelis Menten kinetics where the speed of the reaction is linear with the enzyme concentration at a constant concentration of substrate. MCherry fluorescence can therefore be used as a reporter for the trypsin concentration in the culture medium, using a simple affine equation.

4.2.4 Optimizing incubation time for the trypsin-mCherry expression in droplets

Measuring catalytic efficiency requires being able to detect a level of mCherry fluorescence above the auto-fluorescence of the culture medium. To determine induction times leading to suitable contrasts between the mCherry fluorescence level and the fluorescence of the culture medium, I induced *B. subtilis* cultures secreting the rat trypsin mCherry fusion protein in droplets (see protocol in appendix B.2.4). A sample of the emulsion was imaged at different induction times on a fluorescence microscope to compare the fluorescence of empty droplets to the fluorescence of droplets containing cells (see 4.16).

Results show that before 6 hours of induction, no mcherry fluorescence can be distinguished from the background fluorescence. After 6 h of induction mcherry fluorescence keeps increasing gradually up until 22 hours of induction (see figure 4.16). The contrast does not improve with longer incubation times. Thus, I used overnight induction for further droplet experiments. I also noticed that the mCherry fluorescence was variable between droplets containing cells (from 1-fold to 4-fold typically). Although the average number of cells encapsulated upon droplets production was not precisely controlled, this suggests a heterogeneous level of expression in droplets. Which is confirmed by the results I present in the following sections.

4.2.5 Shaking the emulsion during incubation for expression reduces emulsion size polydispersity

In the course of the experiments in droplets, I noticed an increase in the droplet size heterogeneity of the emulsion during the induction in the tube where the emulsion was stored. During the incubation, droplets form a tightly packed emulsion in which they are not diffusing. Droplets on the top of the emulsion, in contact with air, are likely to evaporate and shrink faster than the droplets at the bottom of the emulsion. Their cells are also more aerated. To limit this source of heterogeneity between droplets,

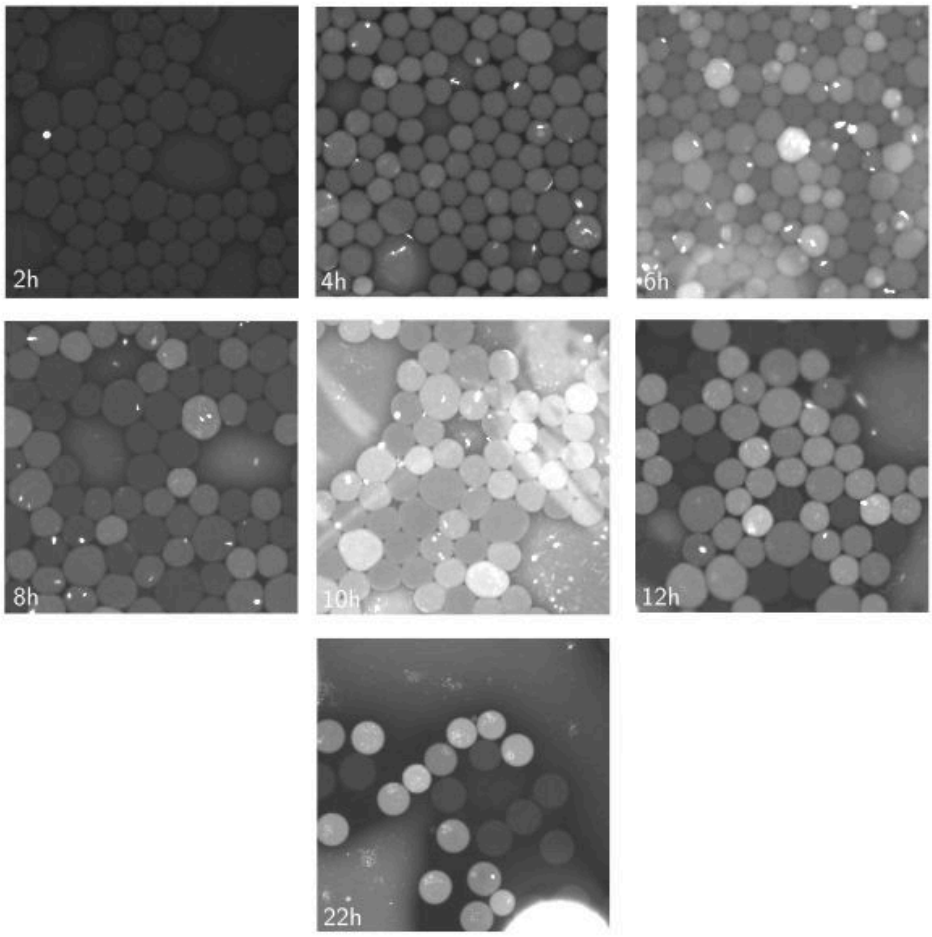


FIGURE 4.16: **Kinetic of trypsin mCherry expression in droplets.** *B. subtilis* cells transformed with the expression vector containing the rat trypsin mCherry sequence were encapsulated in 20 pL droplets to compare empty droplets to droplets containing cells. After 2h, 4h, 6h, 8h, 10h, 12h and 22h of induction, a sample of the emulsion was dropped on a microscope glass slide and imaged on a fluorescence microscope (mCherry fluorescence channel, 20 X, 1 s of exposure). MCherry fluorescence cannot be distinguished from the medium fluorescence before 6 h of induction.

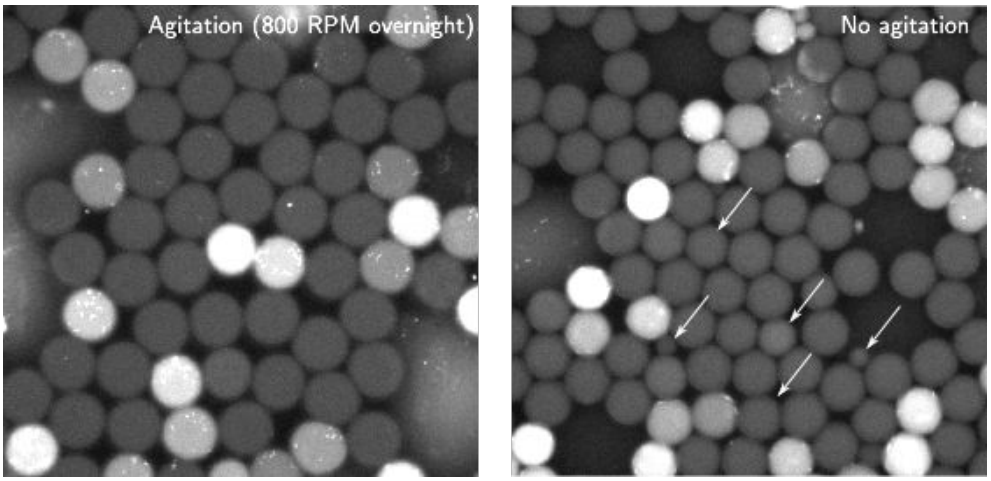


FIGURE 4.17: **Effect of the agitation on induction in droplets and emulsion polydispersity.** *B. subtilis* cells transformed with the expression vector containing the rat trypsin mCherry sequence were encapsulated in 20 pL droplets in two different emulsions. The first was shaken during overnight incubation for induction, the other was not. In the morning, a sample of each emulsion was dropped on different microscope glass slides. The state of the emulsion was imaged on a fluorescence microscope (mCherry fluorescence channel, 200X, 1 s of exposure) in the mCherry fluorescence channel. **Left**: The collector containing the emulsion was shaken at 800 RPM overnight. **Right**: The collector was not shaken, small droplets are indicated with white arrows. The size distribution seemed more uniformed in the shaken case, which might be due to the fact that droplet evaporation is more homogeneous.

I tested shaking the emulsion during the overnight induction[3]. The results are depicted in figure 4.17.

I did not notice any clear difference between the shaken and unshaken emulsions regarding the mCherry fluorescence distributions. Shaken emulsions are less polydisperse than the unshaken emulsions after an overnight incubation. Therefore the emulsions containing the induced *B. subtilis* cells were shaken in further experiments.

4.2.6 Measuring the catalytic efficiency of trypsin variants in bulk

Given the encouraging results regarding the expression of the wild type rat trypsin mCherry fusion protein in droplets, I cloned a set of selected trypsin mutants in the expression vector and measured their activity in bulk using the *B. subtilis* expression system. As mentioned in 1.4.2, those mutants were studied by Halabi et al. [34] who purified them and measured their Michaelis Menten constants. I assayed those mutants using the protocols detailed in appendices B.2.3 and B.2.5. Results are depicted in figure 4.18.

Results show the robustness of the assay using *B. subtilis* as an expression system: the results are reproducible, especially when the activity is

[3]It has been shown that emulsion could be shaken in humidity saturated environment to improve oxygen availability, with limited coalescence [209].

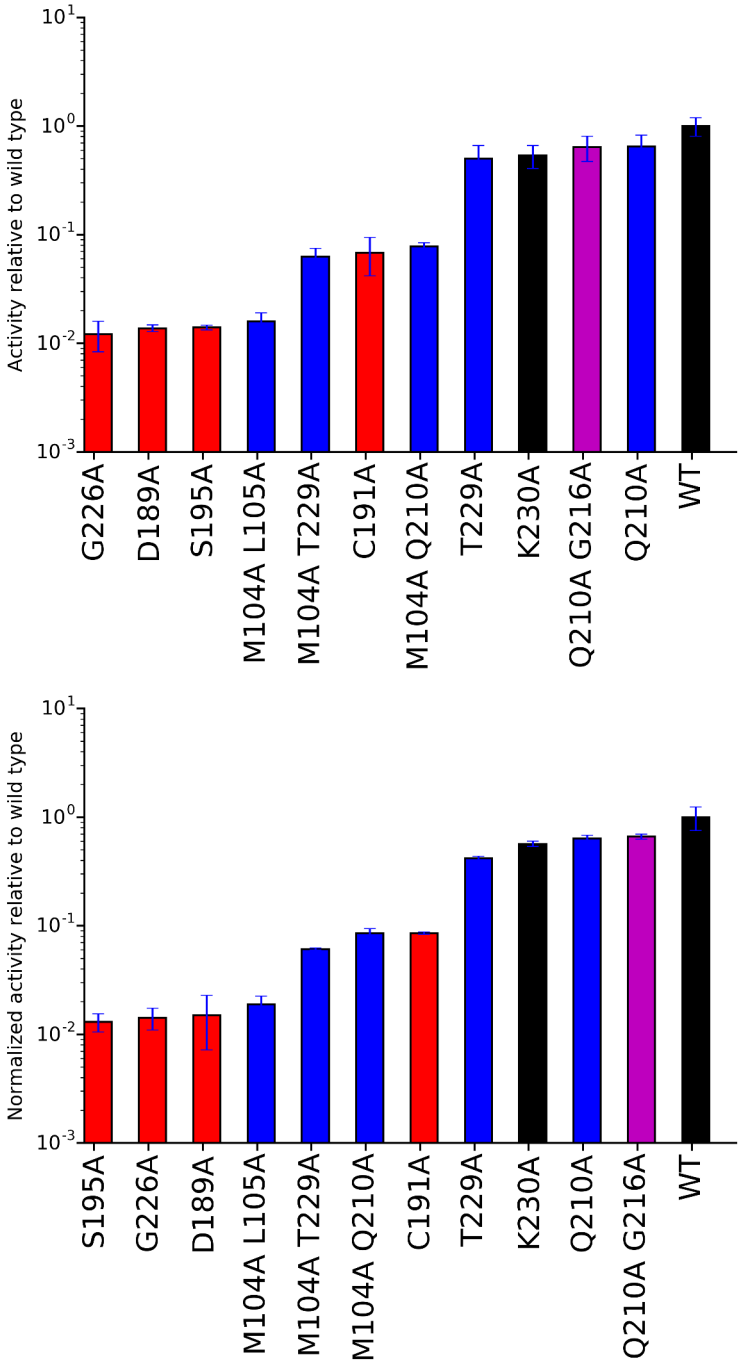


FIGURE 4.18: **Bulk ranking of a subset of trypsin mutants**. *B. subtilis* cells expressing different rat trypsin variants were induced overnight for secretion. For each trypsin variant, induction was performed in triplicates, starting from three different colonies. Assays were performed according to the protocol detailed previously (B.2.3 and B.2.5). The activity of the variants were normalized by the wild type activity. Each bar represents the average of three measurements. **A**: Activity not normalized by mCherry fluorescence. **B**: Activity normalized by the mCherry fluorescence. Compared to expression experiments in *E. coli* (see 4.6), results are more reproducible, especially when the activity is normalized (except for the mutant D189A).

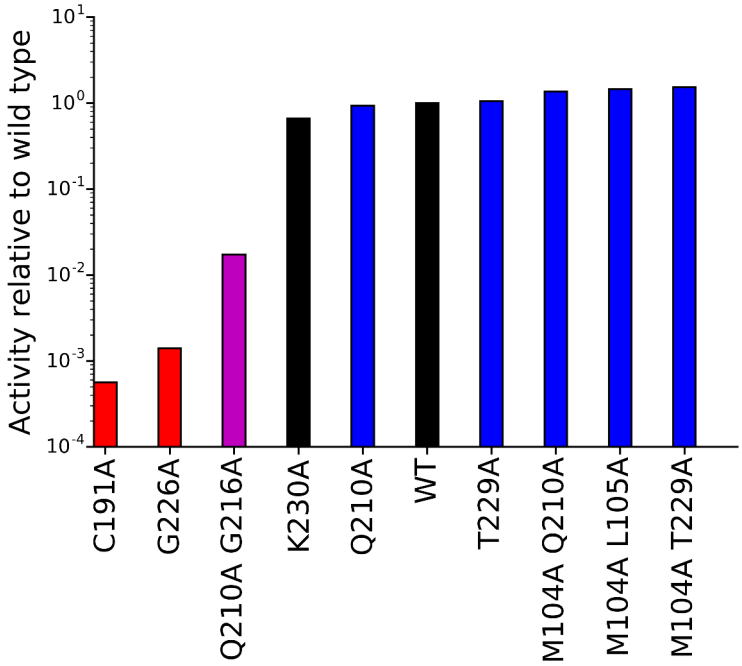


FIGURE 4.19: **Expected ranking of a subset of trypsin variants.** The expected ranking was determined using the formula: $k_{cat}S/(K_m + S)$ with $S = 10^{-5}M$ the substrate concentration used to obtain the results depicted in figure 4.18. k_{cat} and K_m values were taken from [34] (Supplementary information). Those constants were determined at room temperature after secretion by S. cerevisiae and purification. S195A and D189A do not appear here as no activity could be detected in [34].

normalized by the mCherry fluorescence intensity. Interestingly, the mCherry normalization does not seem to change the ranking between mutants that show distinct ranges of activities, but it changes the ranking of mutants that have comparable activities. This might be explained by the fact that the yield of secretion is comparable from one variant to the other in bulk conditions. As reported previously, the supposedly inactive S195A, shows detectable activity in our assay (approximately 100-fold lower than the wild type). Further investigation showed that its activity level was similar to the residual activity presented by *B. subtilis* cells with no rat trypsin gene.

Are the results comparable to the data from Halabi et al. [34]? The ranking obtained is quite different from an expected ranking based on Halabi et al. Michaelis-Menten data. A theoretical expected ranking, based on the K_M and k_{cat} they determined, is depicted in figure 4.19.

The elements that might explain such differences are the following:

- I used different substrates. Halabi et al. used an absorbance substrate Suc-Ala-Ala-Pro-Lys-PNA (Bachem) with only one trypsin hydrolysis site. I used a fluorogenic substrate with two moieties cleaved by trypsin activity, and both of the activity products are fluorescent, with different quantum yields.

- Trypsin expression and assays were performed at 37°C, whereas [34] data were obtained at room temperature. Stability of some mutants might be compromised at 37°C.
- The mCherry protein might have an influence on the trypsin activity itself, or its fluorescence might be impacted by the trypsin activity.

To compare Halabi et al. data and our own expression system, it would be necessary to perform Michaelis-Menten kinetics to determine the kinetic constants of the mutants. This could be achieved by purifying the fusion protein with a tag (His tag for instance), assessing its concentration and performing the enzymatic assays at different concentrations of substrate. But because of trypsin autolysis, I would only recover fragments of the trypsin.

A simpler solution is to characterize the relation between mCherry fluorescence and its concentration (mCherry alone and mCherry as a fusion protein with trypsin, using a tag again or other methods) in the culture medium and to perform enzymatic assays at different concentrations of a substrate such as Suc-Ala-Ala-Pro-Lys-PNA while measuring the mCherry fluorescence.

These experiments are mostly meant to identify mutants with activities distinct from the wild type trypsin in order to test the microfluidic work-flow that I developed in parallel 4.11. I applied this work-flow to three trypsin variants displaying 3 distinct activities: the wild-type, C191A and S195A. The results are detailed in the next paragraph.

4.2.7 Measuring the catalytic efficiency of trypsin variants in droplets

I tested the possibility to reproduce the bulk ranking of trypsin variants[4] in droplets with a subset of those mutants using the work-flow depicted in 4.11.

To speed up the rate of the enzymatic reaction and be able to detect substantial signal before leakage was noticeable in droplets, the rhodamine 110 substrate concentration was increased to 100 µM (final concentration in the fused droplets). Also, to reduce noise in the measurements, the phenotype of each droplet was no longer defined as the maximum of the droplet fluorescence, as in the previous experiments, but as the integral of the fluorescence signal over each droplet. Also, the oil extractor depicted in 2.9, is not used in order to limit the substrate leakage, as it has been shown that increasing the distance between droplets reduced this effect [210].

Following the results of the bulk experiments, I chose the following variants: the wild type trypsin (high activity), C191A (medium activity: 8.5% of the wild type activity) and S195A (low activity: 1.3% of the wild type activity). The three variants were secreted by *B. subtilis* cells in three different emulsions generated separately and incubated in the same tube. From there, the microfluidic work-flow depicted in 4.11 was followed. Trypsin

[4] In this paragraph, trypsin residues' numbering is in agreement with the bovine chymotrypsin sequence.

activity as a function of mCherry fluorescence for each droplet, is depicted in the figure 4.20. Protocol details are available in the appendix B.2.8.

The three variants display a strong linear correlation between trypsin activity and mCherry fluorescence (Correlation coefficient r^2 for wild type, C191A and S195A respectively: 0.70, 0.93, 0.71). The activity displayed by the two mutants is higher (S195A: 4.9-fold higher, C191A: 2.4-fold) which might be due to the rhodamine substrate partial leakage from very active droplets to low activity droplets. In spite of this effect, resolution of the assay is sufficient to selectively extract each population based on its catalytic efficiency, even when the mCherry fluoresence channel is not compensated for the rhodamine cross-talk. The level of mCherry fluorescence displayed by the three variants is similar, spanning from 5 RFU to 40 RFU, which suggests that the variant catalytic property has no influence on trypsin mCherry expression in droplets.

Using the linear relation between mCherry and trypsin activity, normalization can be carried out (see figure 4.21).

In the absence of normalization (4.21, **B**), substrate fluorescence distributions of the three variants are overlapping. This is due to heterogeneous trypsin expression level between the droplets, which prevents quantitative measurement of the distribution of catalytic efficiencies of the variants. Normalizing by the mCherry fluorescence allows to visualize the distribution of catalytic efficiency at a given substrate concentration and to select for an arbitrary range of catalytic efficiency.

In order to perform genotype phenotype mapping experiments with this system, I started building a first trypsin library composed of all single point mutants of the rat trypsin protein.

4.3 Towards a library of all rat trypsin single point mutants

A first application of the developed high-throughput microfluidic workflow is to measure the phenotype of each single point mutants of the rat trypsin protein (see paragraph 1.4.2), in order to test the sector hypothesis. I used a site-directed mutagenesis approach based on inverse PCR [211] ("Around the horn site directed mutagenesis" [5]) to generate all possible amino-acid single point mutants of the rat trypsin polypeptide sequence.

I designed a first set of mutagenic primers for which I optimized the PCR conditions leading to acceptable results in terms of PCR yield and specificity. The resulting library was analyzed by deep sequencing. The data revealed that for some positions the mutagenesis failed. I redesigned those positions, carried out the corresponding mutagenic PCR reactions

[5] http://www.openwetware.org/wiki/'Round-the-horn_site-directed_mutagenesis

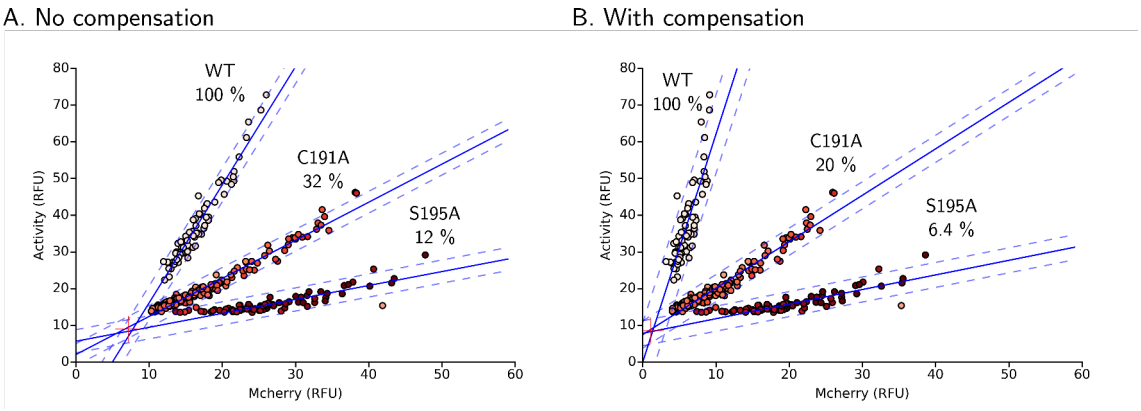


FIGURE 4.20: **Trypsin activity versus mCherry fluorescence in droplets.** The three different *B. subtilis* strains bearing either the wild type trypsin gene or the variants S195A or C191A, were encapsulated separately in presence of IPTG and coding dye. The coding dye concentration was set to be specific to each variant: high concentration for S195A (red), medium concentration for C191A (pink), low concentration for wild type (white). The three emulsions were collected in the same collection tube. After *B. subtilis* overnight induction in droplets, enzymatic assay was performed according to B.2.6. Only active droplets are depicted here (21% of all droplets, S195A: 78 , C191A: 176 and WT: 72 droplets). After the experiment, the cross-talk of the rhodamine fluorescence into the mCherry channel was determined to substract the rhodamine fluorescence signal from the mCherry channel (compensation). Not compensated data are presented in **A**. Compensated data are presented in **B**. For each variant population, linear regression was performed (blue line). The slope relative to the wild type was computed (black percentage). In the case of the compensated data, the null mutant S195A displays 6.4% of the wild type activity and C191A displays 20% of the wild type activity. 3 standard deviations from the linear fit are represented in dashed blue lines. In each case, more than 95% of the population is comprised in the interval. As a result, identical sorting gates would selectively recover more than 95% of each population. The barycenter of the three intersections between each pair of linear regression is represented (red cross). It defines the background fluorescence of the experiment in the substrate channel and in the mCherry channel. Populations closer to the barycenter are harder to discriminate for low expression and low activity (low mCherry intensity). With the barycentre as a center of rotation, different sorting gates or sorting threshold with different angles can be determined to extract populations displaying variable catalytic activities. When compensation is not carried out, the mCherry signal and expression level are over-estimated, the dynamic range and resolution of the assay are lower.

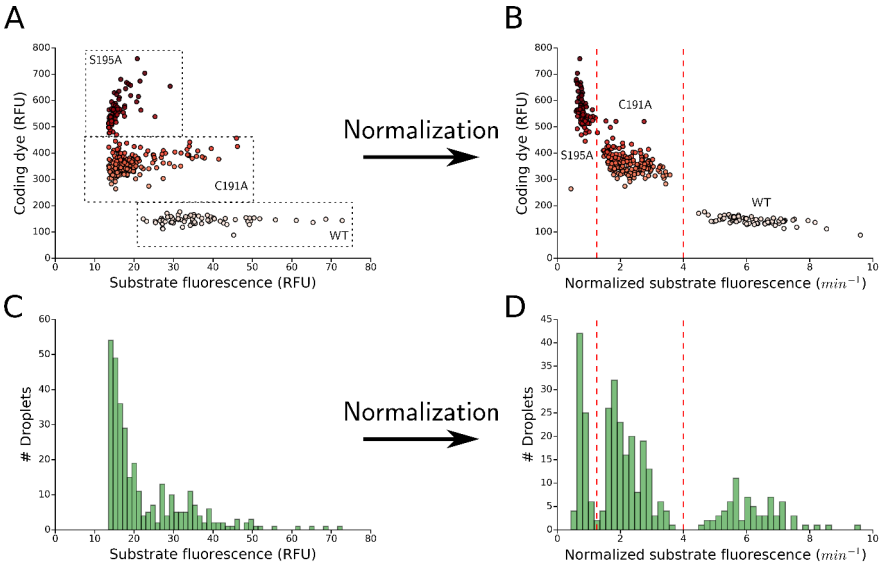


FIGURE 4.21: **Normalizing trypsin activity in droplets to measure catalytic efficiencies.** **A**: Coding dye versus the substrate fluorescence. Thanks to the coding dye it is possible to distinguish the three different emulsions. Their substrate fluorescence distributions overlap because of the heterogeneous expression level between the droplets. **B**: Substrate fluorescence distribution. Because of the overlap, it is impossible to measure catalytic efficiency and to select for it. **C**: Substrate fluorescence normalized by the droplet fluorescence in the mCherry channel. The equation of the linear fit computed in in 4.20 was used to normalize trypsin activity. The different substrate fluorescences of the variants are no longer overlapping. **D**: Normalized substrate fluorescence distribution. It is now possible to measure and select for an arbitrary range of catalytic efficiencies, at a given concentration of substrate.

and sequenced this second library. Those two libraries should allow to perform a genotype-phenotype mapping of 80% of all single point mutants of the rat trypsin protein.

4.3.1 "Around the horn" site directed mutagenesis principle

Differences between site-directed mutagenesis and random mutagenesis are already detailed in 1.2.1. The main reason to choose site-directed mutagenesis is that amino-acid mutations requiring more than one nucleotide substitution are not likely to be seen in a library generated by random mutagenesis. In the case of the wild type rat trypsin, only 32% (1340 out of 4237) of all single point amino-acids mutations are reachable via single nucleotide substitution.

Besides, the rat trypsin gene is toxic in *E. coli*. Library preparation supposes to clone the library in the expression vector, to transform the cloning product in *E. coli*, to recover the cells that were transformed to extract the library and transform it into WB800N. Preliminary experiments showed that with a random mutagenesis protocol yielding on average two residue mutations per trypsin sequence, 50% of the recovered sequences displayed nonsense codons or frame shifts after selection on plates containing the proper antibiotics. By using random mutagenesis I would have selected for inactive trypsin mutants, especially at the positions that are easily mutated to nonsense codons (52 codons out of 223 in the wild type trypsin, can be mutated to nonsense codons with only one nucleotide substitution).

Site-directed mutagenesis requires designing primers that bear the desired mutations and that are used to PCR amplify the mutant sequence from the wild type sequence. The "around the horn" site directed mutagenesis [211] consists in amplifying by PCR the whole plasmid containing the gene of interest using a forward primer whose first three nucleotides are coding for the desired amino acid mutation and a reverse primer that is faithful to the wild type sequence (see figure 4.22, **A** and **B**). In theory, it is necessary to perform one PCR reaction for each position to mutate. But it is possible to use a mix of 19 forward primers, each one of them bearing one of the 19 possible mutations, in each PCR reaction. As a result, single point saturated mutagenesis can be done performing as many PCR reactions as positions to mutate the sequence of interest[6].

Generating a library of all possible single point mutants of the 223 residues of the rat trypsin protein requires:

- Designing and ordering $19 \times 223 = 4237$ forward primers and 223 reverse primers.
- Performing 223 mutagenic PCR reactions, purifying the correct size PCR product on gel, measuring their concentration, mixing and circularizing them.

[6] By repeating the same operationss, using a library of single point mutants as a template, it is possible to generate a library comprising more than 90% of all possible double point mutants.

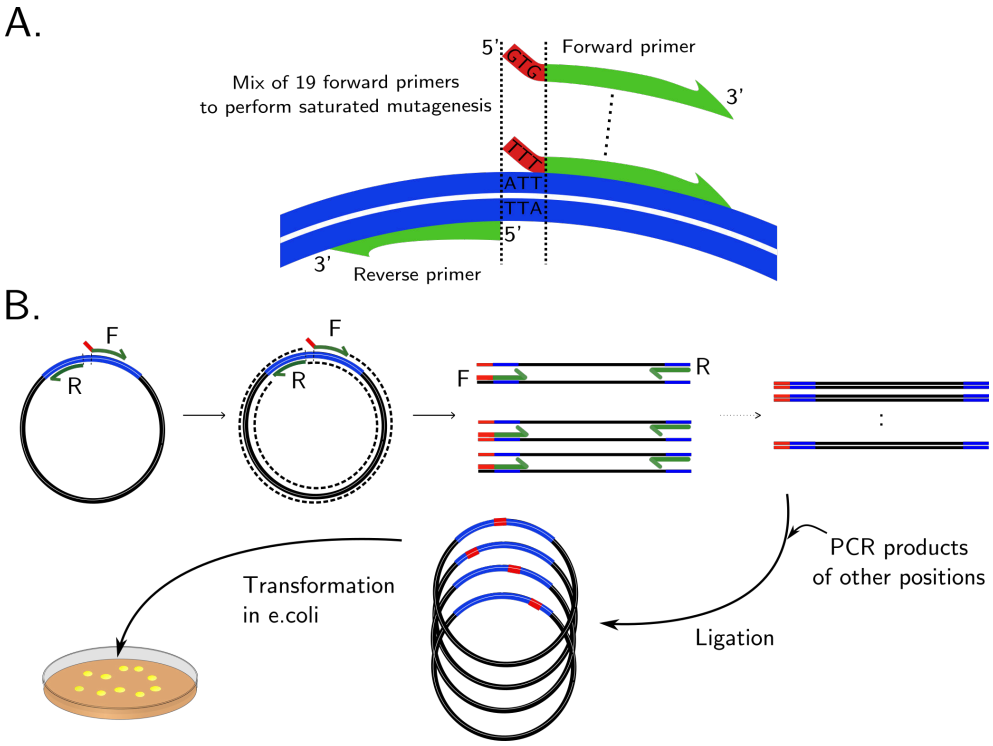


FIGURE 4.22: "Around the horn" mutagenesis principle. **A**: For each amino acid to mutate, a set of 19 mutant codons is designed (forward primer variable part, in red). They are added a constant part (forward primer, in green) faithful to the wild type sequence. The reverse primer is also designed faithful to the wild type sequence. **B**: For each position to mutate in the rat trypsin sequence, the whole plasmid is amplified using the designed reverse primer and the mix of the 19 forward primers. For each position, the PCR product is purified on a gel and its concentration is assessed. All the PCR products are mixed in equimolar ratio and circularized in one single batch. The ligation product is transformed in highly competent cells and the colonies are recovered to extract the library DNA, or to store the transformants in glycerol stock.

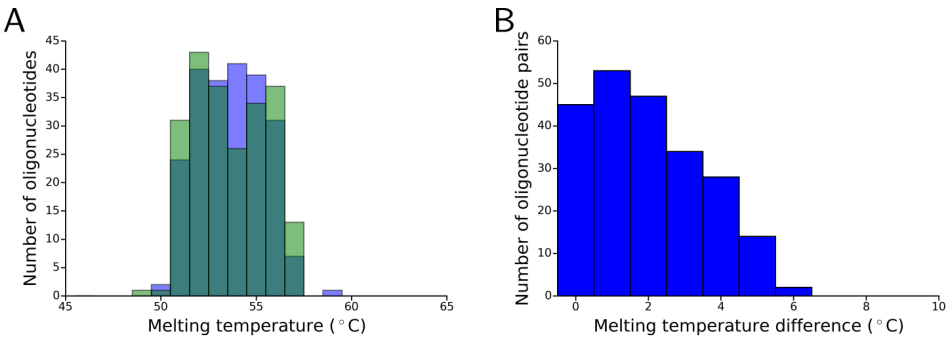


FIGURE 4.23: **Melting temperature distribution of the designed mutagenic primersA**: **Blue**: Melting temperature distribution of the forward primers (constant parts). **Green**: Melting temperature distribution of the reverse primers. **B**: Melting temperature difference distribution $|T_{m,forward} - T_{m,reverse}|$. The distribution of melting temperatures is kept in a narrow range around the target temperature: 54°C. Outliers at 50 and 60°C are characterized by either very low or very high GC content. Melting temperature difference between forward and reverse primers is kept within 6°C.

- Transforming the circularized product in highly competent *E. coli* cells.
- Extracting the library from the recovered transformed cells and deep sequencing the library.
- Analyzing the sequencing data to assess the success of mutagenesis.

4.3.2 Designing the mutagenic primers

Sets of mutant codons were designed for each wild type position to mutate, using degenerate nucleotides (e.g, N coding for A, T, C or G). This strategy allowed me to only design 3 to 5 mutant codons per position, instead of 19. An example and the full table of amino acids and their set of associated degenerate codons are available in B.3.1.

The constant part of the forward primers and the reverse primers were designed using a script whose principle is detailed in the appendix B.3.1. I designed primers with similar melting temperatures to be able to perform a maximum number of PCR reactions in parallel in 96 well plates. In brief, for each position to mutate, the constant part of the forward primer and the reverse primer were set to 10 nucleotides. They were then extended using the wild type trypsin sequence as a template to reach a target melting temperature of 54°C (according the NEB, "Q5 polymerase T_m calculator"). The melting temperature distributions of the designed primers are depicted in figure 4.23. As the distributions show, the designed primers melting temperatures are contained in a narrow range, spanning over less than 10°C. The difference of melting temperature between forward and reverse primers is also kept within 6°C. Outliers are primers corresponding to sequence regions displaying low or high GC content.

Final forward primers were obtained by concatenating the mutant codons with their respective forward primer constant part. The primers were synthesized (IDTDNA) and I performed all mutagenic PCRs on a modified

version of the pET28 plasmid (used in the Chapter III to perform trypsin expression in *E. coli*, see B.4). This plasmid (pET28, 6kb) is smaller than the *B. subtilis* expression vector (pHT43, 9.6 kb) and leads to higher amplification yields. The PCR results are detailed in the next section.

4.3.3 Performing saturated mutagenesis on the rat trypsin protein

The Q5 polymerase was used to perform those PCRs as it is characterized with a low error-rate ($\approx 6.10^{-6}$ substitution per base pair and $\approx 2.2.10^{-7}$ deletions and insertions per base pair). It also proved to be suitable for full-plasmid amplification in preliminary experiments [212].

The manufacturer Q5 polymerase PCR protocol using different annealing temperatures, led to satisfying results in terms of yield and product size for 55% of the 223 pairs of primers. The others primer pairs either failed at producing amplicons or produced amplicons with unexpected sizes. Therefore I tested different protocol to improve specific amplification. First, I increased the amount of template by ten folds in each PCR reactions. Second, to favor specific amplification, I performed "Touch down" PCR [213]. That protocol is characterized by a high initial annealing temperature which decreases at each cycle. This strategy enabled to obtain acceptable results for 33% of the primer pairs. Unfortunately, I did not manage to see specific amplification for 11% of the positions[7] (26 primer pairs, see figure 4.24). The details of the PCR protocol are given in appendix B.3.2.

As explained in the previous sections, the PCR products showing the expected amplicon size were purified on gel (4.24, **B**) and circularized. The final circular product was transformed in highly competent *E. coli* cells which were plated on selective medium. After transformation, the recovered colonies were used to extract the library which was subsequently deep sequenced. All the details regarding the preparation of the library and the deep sequencing are available in the appendix B.3.3. In the next section, I present the deep sequencing results for this first library.

4.3.4 Analyzing the first rat trypsin library by deep sequencing

Frequency of expected and unexpected mutations, diversity of mutations at each position and errors To assess the success of the saturated mutagenesis, three observables were quantified for each position of the trypsin polypeptide chain: the frequency of expected mutant codons $f_{expected}$ (i.e mutations encoded by the site-directed mutagenesis primers), the total mutation frequency f_{total} and the diversity of expected mutant codons, D (i.e the number of different expected codons that are read). Here I define those observables:

The total frequency of mutations at the position i:

$$f_{i,total} = \frac{M_i}{N_i}$$

[7] In this paragraph, trypsin residues' numbering is sequential, from 1 to 223.

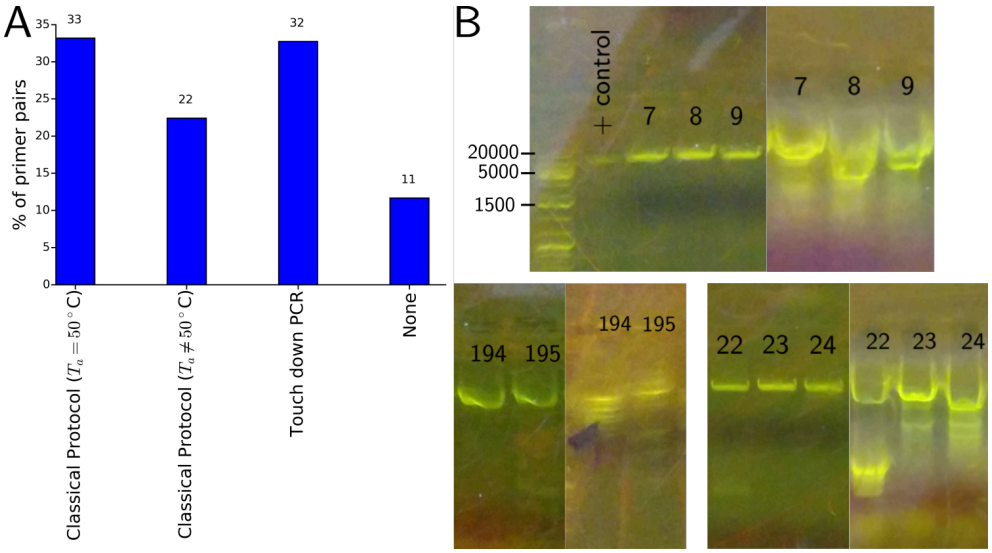


FIGURE 4.24: **Mutagenic PCRs (band size and amplification) according to different PCR protocols. A**: For the saturated mutagenesis using each primers pair, three successive protocols were tested: classical PCR protocol with an annealing temperature of 50°C, classical PCR protocol with another annealing temperature (59 °C for most of them) and "Touch down" PCR starting with a high initial annealing temperature (61 °C), decreasing at each cycle (-0.5°C by cycle for the 20 first cycles and then 50°C). "None" designates the primers pairs for which none of those PCR protocols enabled proper amplification. Overall, mutagenic PCRs gave acceptable results, in terms of size and yield of the product, for 89 % of the positions. **B**: Mutagenic PCR products ran in an agarose gel (0.8%, low melting point agarose) before gel purification. For the subsets of PCR reactions presented here, the left and right pictures represent respectively PCR products after "touch down" PCR and classical PCR reactions. The "touch down" protocol greatly improved the results in terms of specificity and yield, of the most difficult mutagenic PCR reactions.

Where N_i is the total number of times position i is read and M_i the total number of times codon at position i is read and different from the wild type (mutation).

For the same position i, I define the frequency of expected mutations as the following:

$$f_{i,expected} = \frac{M_{i,expected}}{N_i}$$

Where $M_{i,expected}$ is the total number of times the codon at position i is read as one of the 19 expected mutant codons.

Finally, I define the diversity of expected mutant codons at position i, D_i, as the number of different expected mutations read at position i.

If the mutagenesis is successful at position i, sequencing data should display either the wild type codon (no mutation) or one of the 19 mutant codons designed previously. As a result, $f_{i,expected}/f_{i,total}$ should equal one. Also, sequencing data should display each one of the 19 mutant codons and therefore $D_i = 19$. Besides, the frequency of occurrence of the 19 mutant codons should also be distributed uniformly.

The 26 pairs of primers that failed to yield any correct mutagenic PCR product were not included in the library. The associated positions were analyzed to measure the level of errors, in absence of mutagenesis, along the trypsin sequence. Ideally, I should have had sequenced the wild type trypsin sequence with the library, in order to directly measure the sequencing error rate and its distribution along the sequence. I could also added "unique molecular identifiers" (random sequences). By comparing reads displaying the same UMI, it should have been possible to establish consensus sequences.

First deep sequencing run results

Sequencing results are summarized on figure 4.25. Positions that were not included in the library are notified in red in this figure. They represent the errors from sequencing, but also potential errors from the mutagenesis itself (mutagenesis at a given position generating mutations at another). Those errors are characterized by both low D and high $f_{expected}/f_{total}$ values. I defined arbitrary thresholds to assess the success of the saturated mutagenesis: $D_i > 15$ and $f_{expected}/f_{total} > 0.3$ which includes 73% of the positions.

Further analysis of the noise of the experiment are presented in figure 4.26. The mean and standard deviation of the f_{total} and $f_{expected}$ distributions for the 26 positions that were not included in the library were computed. A threshold error level was defined as the sum of the mean plus two standard deviations for all 26 positions not included in the library: $f_{total,error} = 3.10^{-3}$ and $f_{expected,error} = 9.10^{-4}$. Respectively 65%

and 74% of the positions are displaying mutation rates above the threshold error level.

Those results show that the level of noise in the sequencing data is important. Multiple sources of errors can be considered. This particular miseq run was characterized by an important sequencing error-rate (typically > 1% as determined by Dr. Russ who sequenced samples in the same run). I transformed the library in *E. coli* before sequencing, which might have introduced some biases. Finally, as resolution on electrophoresis gels is poor for large PCR products, I might have interpreted unspecific PCR products as specific.

I used the outcome of this analysis to redesign part of the primers corresponding to positions which did not display expected mutagenesis profile.

Second deep sequencing run results, with fraction of the mutagenic primers redesigned.

Most of the primers corresponding to the unsuccessful positions (27%) were very short (less than 13 bp) and characterized by high GC content (higher than 70% on average). I redesigned 32% of the positions (71 pairs of primers including the 26 unsuccessful ones and the positions that did not meet the two criteria on $f_{expected}/f_{total}$ and D (see appendix B.3.4). In brief, the minimal length was set to 16 nucleotides with no target melting temperature. The few pairs of primers for which the melting temperature exceeded 72°were designed manually. The PCR reactions were performed for those 71 positions, their product were mixed with the leftovers of the already successful positions PCR reactions and the already mentioned protocol was followed. The analyzed deep sequencing results are depicted in figure 4.27.

This new library is characterized with a success rate of 86% according to the criteria already mentioned. Unfortunately, there is a bias between the positions for which the primers were redesigned and the positions for which the primers pairs stayed the same. The former positions had benefited from fresh PCR products whereas the latter positions where added to the library using PCR products old from the first library. Those old PCR products might have partially decayed between the makings of the two libraries. Given those results, I decided to mix the first and the second library (50% and 50%) to limit the bias and produce a more uniform library. The foreseen characteristics of this library, that is currently being made, are depicted in figure 4.28. 94% of the positions should meet the criteria already mentioned.

More detailed visualizations of the composition of this library in terms of single point mutants are available in figure 4.29 and 4.30.

Among all detected mutations, only 7% of them are nonsense mutations, which is very low compared to what could be performed with random mutagenesis: preliminary results show that approximately half of the

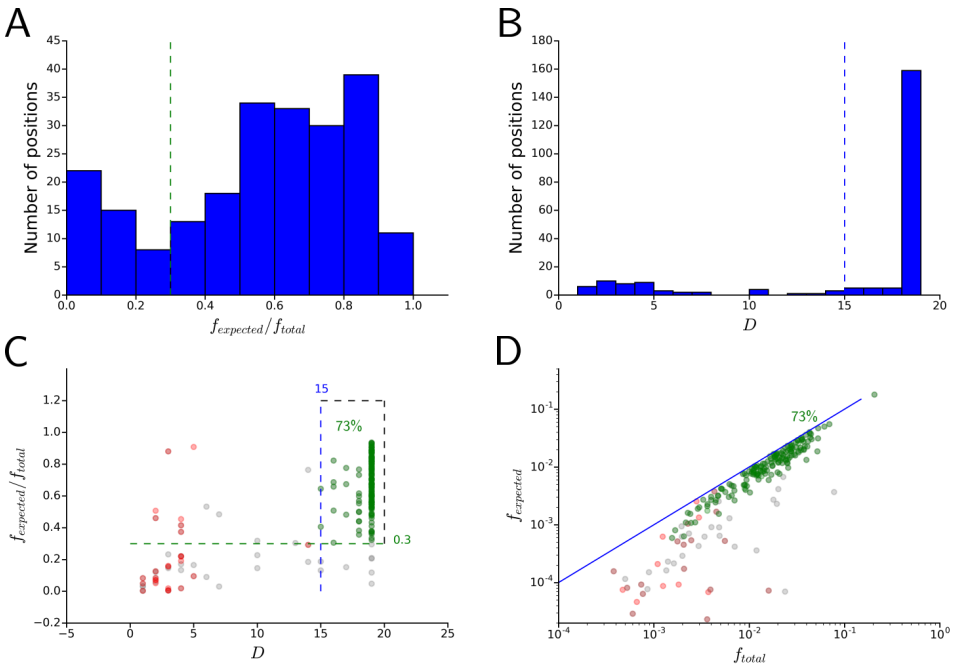


FIGURE 4.25: **Characterization of the first trypsin library of single point mutants (first deep sequencing). A**: $f_{expected}/f_{total}$ distribution. The distribution is bimodal with a main mode at 0.7. An arbitrary threshold is set at $f_{expected}/f_{total}$=0.3 (green dashed line). **B**: D distribution. 160 positions over 223 show each expected mutant codon at least once. A threshold is set at $D = 15$. **C**: $f_{expected}/f_{total}$ versus D. Each dot represents a position. Positions for which no mutagenic PCR products could be obtained are represented in red (26 positions out of 223): they are used to assess the level of sequencing errors noise in the sequencing data (see figure 4.26). They are characterized by a low D and a low $f_{expected}/f_{total}$. 73% of the positions showed satisfying results (successful positions in green, $D > 15$ and $f_{expected}/f_{total} > 0.3$). As expected, the green and red distributions do not overlap. In grey are represented the positions for which the mutagenic PCR reaction seemed to yield a correct product, but that do not display satisfying D or $f_{expected}/f_{total}$ successful results. **D**: $f_{expected}$ versus f_{total}. The colors of the dots is the same as in C. This representation allow to visualize the distributions of $f_{expected}$ and f_{total} which span over three orders of magnitude from 10^{-3} to 10^{-1}: in practice, this means that some mutations are 1000 folds more likely to be seen in a genotype phenotype mapping experiments than others. The distribution is not linear and deviates towards very low values $f_{expected}$ as f_{total} decreases. I interpret this part of the distribution as positions for which the mutagenic PCR product concentration was low in the library sent for sequencing: they are dominated by the sequencing errors ($f_{total} >> f_{expected}$).

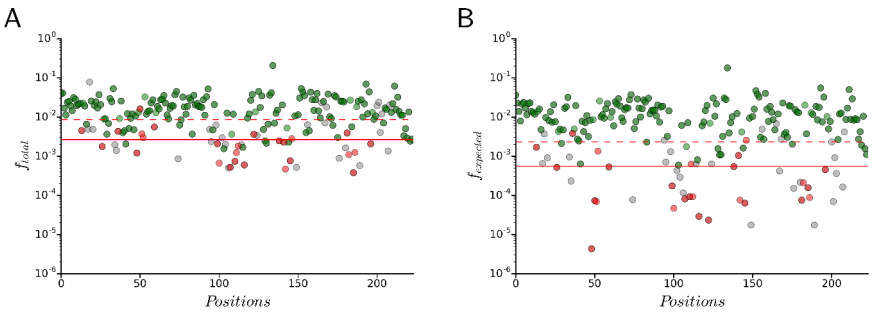


FIGURE 4.26: **Characterization of sequencing errors (first deep sequencing).** Colors of the dots are defined as in 4.25. Continuous red line: mean value for the 26 positions not included in the library sent for sequencing. Dashed red line defines the error threshold: mean value plus two standard deviations computed from all 26 positions not included in the library. **A**: f_{total} as a function of the amino-acid position in the trypsin sequence (1 to 223). Positions above the error threshold account for 65% of the positions included in the mix. **B**: $f_{expected}$ as a function of the amino-acid position in the trypsin sequence (1 to 223). Positions above the error threshold account for 74% of the positions included in the mix.

mutations that were recovered from clones were either nonsense codons or frame-shifts due to deletion and insertion. Although a better understanding of the level of noise is necessary, this shows that our approach manages to contain the amount of nonsense mutations that are generated.

The distribution of mutations is not uniform within a position and between positions, as some particular mutations are over or under represented in the population. Again, several explanations are possible to explain non-uniformity of the distribution: biased pool of mutagenic primers, unspecific PCR products, bias due to the transformation, incorporation of errors during the preparation of the library via PCR and errors in the sequencing process. This new library will be sequenced with a proper negative control (trypsin wild-type sequence), in order to rigorously characterize sequencing errors.

Coverage of all single point mutants

The last library will be cloned in the *B. subtilis* expression vector and transformed in *B. subtilis*. If the distribution depicted here is correct, and if the clonings and the transformation do not change the distribution presented in 4.29, the proportion of single point mutants (or coverage) whose phenotype is going to be measured can be assessed. This coverage depends on the number of *B. subtilis* colonies that are recovered. The coverage has been computed and is presented in figure 4.31.

B. subtilis presents a natural but low competency: the transformation of one µg of *B. subtilis* (WB800N) leads to approximately 10^3 colonies (results I obtained with the transformation protocol detailed in the appendix B.2.2). This range can be improved using unspecific rolling circle amplification to achieve 10^5 to 10^6 colonies [204]. Therefore obtaining between 10^4 and 10^5 colony should be reachable, leading to a 80% coverage of all single point

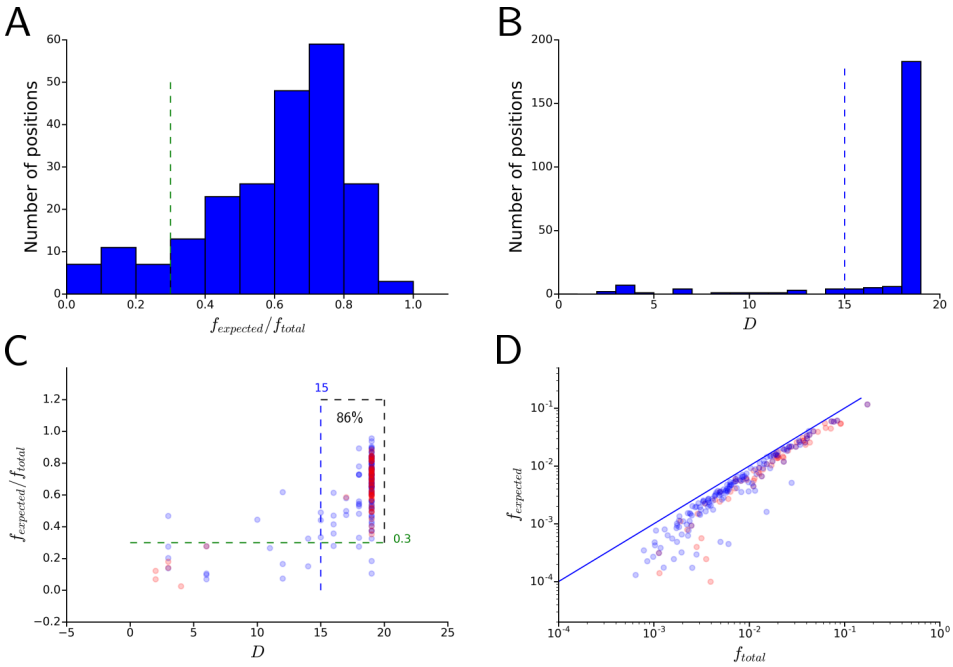


FIGURE 4.27: **Characterization of the second trypsin library of single point mutants.** **A**: $f_{expected}/f_{total}$ distribution. Distribution seemed to present a main mode at 0.7, as for the first library. We kept the arbitrary threshold at 0.3 (green dashed line). **B**: D distribution. More than 180 positions over 223 showed each expected mutant codon at least once. The threshold was kept at $D = 15$. **C**: $f_{expected}/f_{total}$ versus D. Positions for which the pair of primers were redesigned are represented in red (71 positions out of 223). Positions for which the pair of primers were identical as in the first library are represented in blue. Saturated mutagenesis seems to be successful for 86% of the positions (to compare to 73% in the first library, according to the same criteria). **D**: $f_{expected}$ versus f_{total}. The colors of the dots is the same as in **C**. Even though the results seems to be better as 86% of the positions gave successful results (73% in the first library), the distribution of frequencies is bimodal between positions for which the primers were redesigned (red) and the other positions (blue). This might be due to the fact that the PCR products used for the blue positions were older: their concentration might have decayed.

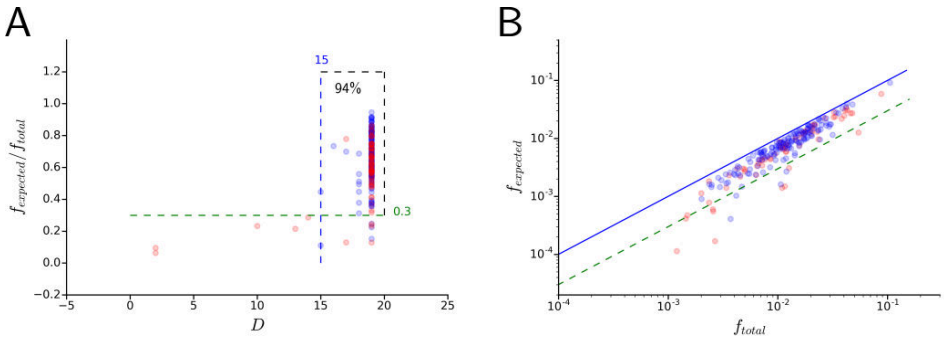


FIGURE 4.28: **Characterization of the third trypsin library of single point mutants.** **A**: $f_{expected}/f_{total}$ versus D. Colors are the same than in figure 4.27. Mixing the two libraries should enable us to bring the rate of successful positions to 94%. **D**: $f_{expected}$ versus f_{total}. The colors of the dots is the same as in **A**. Mixing the two libraries should enable to maintain the frequency distribution within 100 folds.

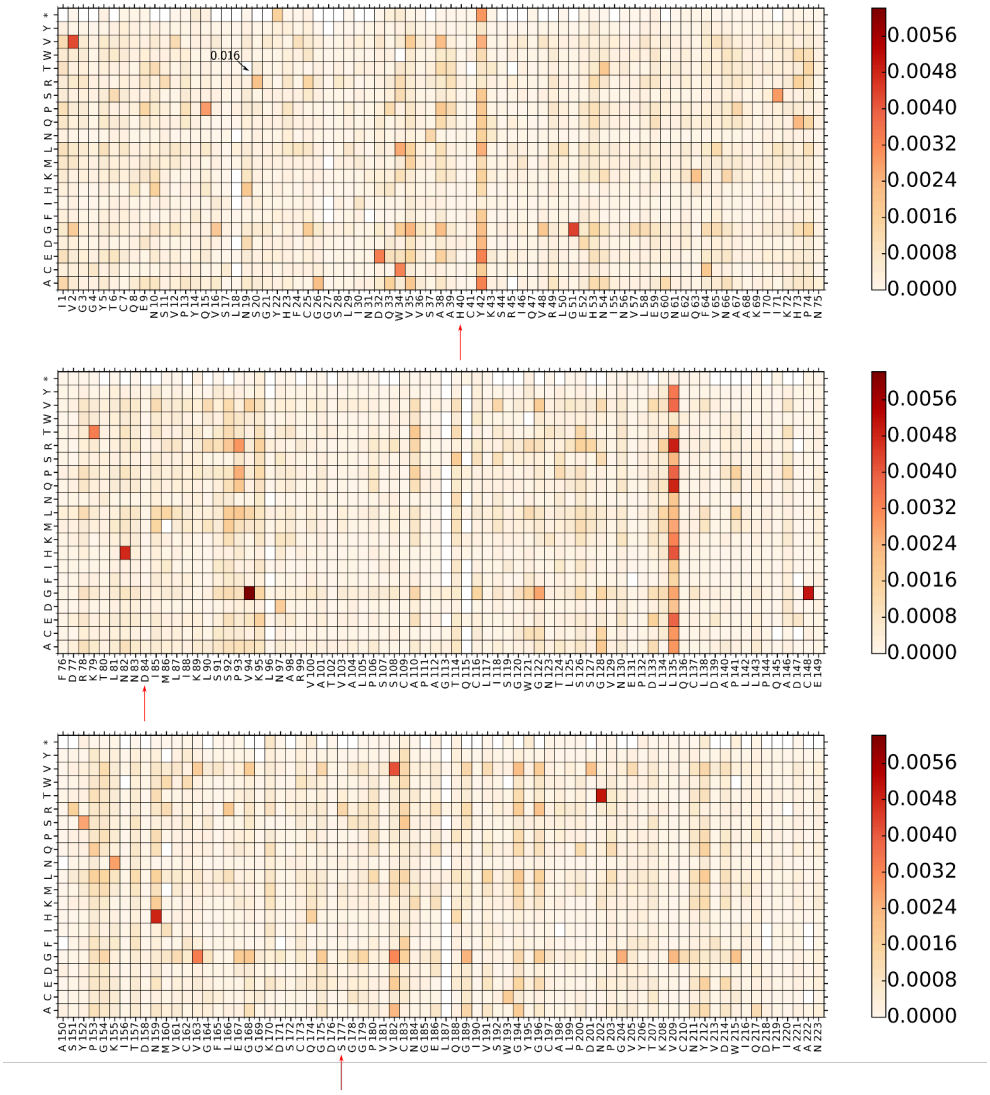


FIGURE 4.29: **Frequency matrix of single point mutations along the trypsin polypeptide chain (50% of the first library and 50% of the second library).. X axis**: trypsin peptide chain, each position number is preceded with the wild type amino acid. **Y axis**: Mutations. Each box represents the frequency of occurrence of one particular mutation at one particular position. Red arrows designate the catalytic triad of the trypsin. I did not notice any particular enrichment towards mutations in this catalytic triad that would specifically inactivating the trypsin activity. In fact, Y42 and L135 mutations were enriched either in sequencing data coming from library 1 or library 2, but not both, suggesting their enrichment is not reproducible. Within some positions, the distribution of mutations is highly biased: e.g N19 and N19T for which the particular not expected codon 'ACC' is invading the population in the first library. I interpret that type of results as positions for which the mutagenic PCR product is unspecific, or as the consequence of biases in the pool of primers used to perform mutagenesis. The redesigned primers to mutate N19 leads to much better results on the second library (not shown here). As expected, the frequency of stop codons is low: 7%.

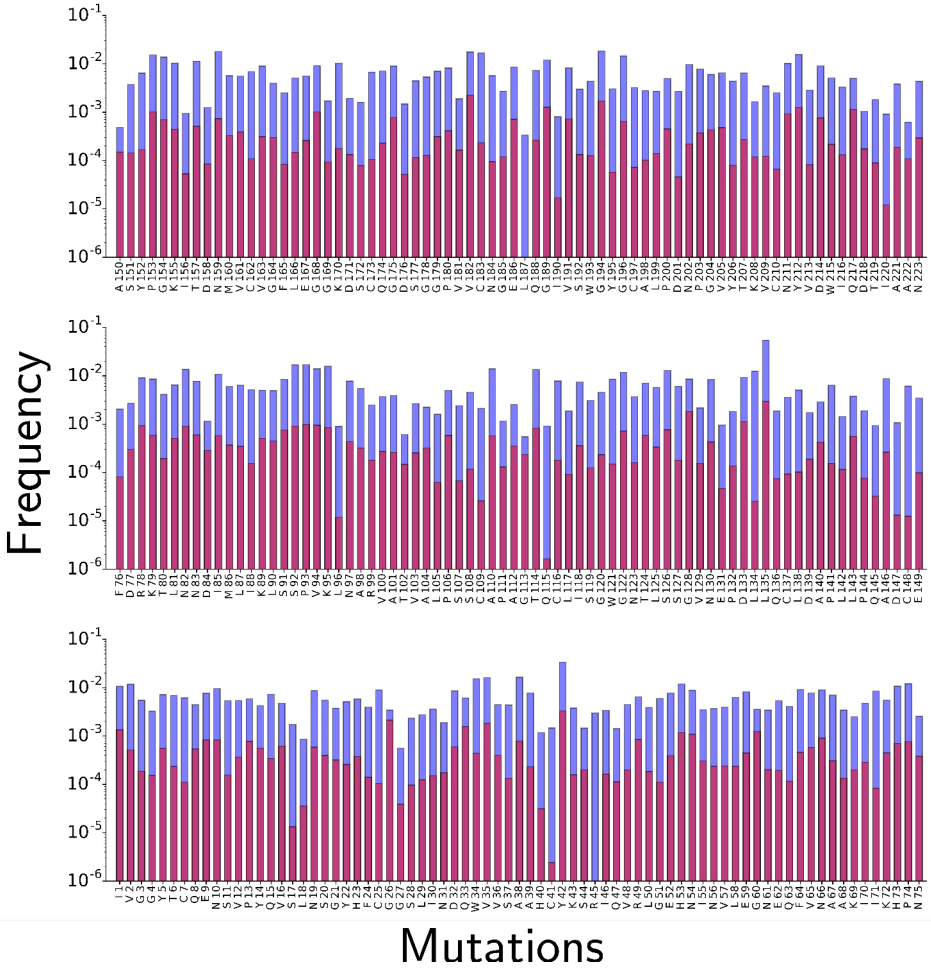


FIGURE 4.30: **Total frequency of occurrence of mutations at each position of the trypsin peptide chain.** In blue: total frequency of misense mutations, in red: total frequency of nonsense mutations (stop codons). High value positions are already highlighted in figure 4.29: N 19, Y 42, L 135. The coefficient of variation (ratio of standard deviation / mean) is 0.90, which highlights the fact that the distribution is not uniform. Positions with the lowest frequencies are typically seen 100 folds less than high frequency positions, as already seen in figure 4.28.

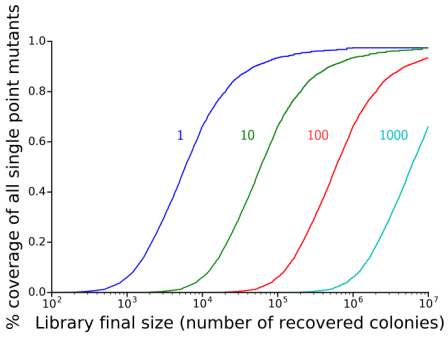


FIGURE 4.31: **Coverage of all single point mutants as a function of the number of recovered colonies for the third library.** The frequency matrix presented in 4.29 was multiplied by the final library size (number of recovered *B. subtilis* colonies). The number of single point mutants that are seen more than 1, 10, 100 and 1000 times was computed. As an example, recovering 10^4 colonies enables to measure the phenotype of 80% of all possible single point mutants at least once, whereas recovering 10^5 colonies enables to measure the phenotype of 80% of all single point colonies at least 10 times.

mutants, each one of them represented by at least one colony.

It is possible to measure the fluorescence of 10^7 droplets in one single experiment. Cells have to be diluted to prevent double encapsulations of two different trypsin mutants in the same compartment. Therefore, it is possible to measure the phenotype of $\approx 10^6$ mutants per experiment. As a result, without considering sampling effects, our setup allows to cover 80% of all single point mutants, with the phenotype of each one of them measured at least 100 times. This should enable us to begin to gather data to establish the genotype-phenotype landscape of the rat trypsin protein, and question the sector hypothesis.

4.4 Discussion, Conclusion and Perspectives

In this chapter dedicated to the development of an in vivo expression work-flow to perform genotype phenotype mapping of the rat trypsin enzyme, I have shown the following:

Regarding the *E. coli* expression system:

- Soft osmotic conditions (low salt) on *E. coli* cells expressing rat trypsin allow to detect trypsin activity in the culture medium using small fluorogenic substrates based on coumarin or rhodamine 110.
- Rat trypsin periplasmic expression with *E. coli* was difficult to adapt into droplets mainly, I think, because the expressed protein remained within the cells. As a result, bigger substrate such as a FRET substrate based on EDANS-Dabcyl were not cleaved in the culture medium of induced cells.

- My attempts at developping a reporter for trypsin concentration failed with E.coli expression system. The MUGB compound inhibits trypsin activity in our experiment, but it has proven not suitable to measure trypsin concentration. MCherry cannot neither be used as a reporter for the trypsin concentration as its fluorescence is not detected within the supernatant. Besides, *E. coli* cells showed inclusion bodies upon induction to express the rat trypsin mCherry fusion protein and low trypsin activity.

Regarding the *B. subtilis* expression system:

- The rat trypsin mCherry fusion protein is secreted by the engineered strain WB800N as a whole in the culture medium. The trypsin activity and the mCherry fluorescence can be both detected and measured after overnight induction.
- The mCherry fluorescence intensity is linearly correlated with the trypsin activity detected in the culture medium in bulk. The trypsin activity can be normalized thanks to a simple linear relationship.
- It is possible to assay trypsin mutants in droplets using WB800N expression system and to directly measure their catalytic efficiency, at a given concentration of substrate, by simultaneously measuring the fluorescence of the activity product of the trypsin activity and the fluorescence of the mCherry reporter. This opens the possibility to quantitatively measure the phenotype of large number of trypsin mutants in droplets.

Towards a first library of all single point mutants of the rat trypsin protein In parallel, I started generating a first library of rat trypsin mutants via site-directed mutagenesis by inverse PCR. I now have at my disposal a library with which I should be able to measure the catalytic efficiency of $\approx 80\%$ of all single point mutants ($\approx 3400 mutants$), each one of them being measured on average, at least 100 times, in a single experiment.

Chapter 5

Conclusion

5.1 Conclusion and further work

I developed two different approaches to perform the genotype-phenotype mapping of two model enzymes: SGAP and the rat trypsin protein.

In the case of the work-flow dedicated to SGAP, I have shown that the incompatibilities between gene amplification/gene expression and expression/enzymatic assay were constraining the microfluidic work-flow. Both amplification and expression products have to be diluted in order to respectively perform expression and enzymatic assay. The following microfluidic work-flow is in principle functional and should allow to measure SGAP variants activity with high contrast: (i) encapsulating and amplifying SGAP gene in 0.2 pL droplets, (ii) expressing SGAP gene in 2 pL droplets using cell-free expression reagents (iii) assaying the enzyme with a fluorogenic substrate in 20 pL droplets. The development of a new hydrophilic rhodamine-based substrate characterized by limited leakage should even improve the dynamic range of the assay, especially for variants displaying low catalytic efficiency.

Unfortunately, this cell-free work-flow is limited by the miniaturization imposed by femtoliter droplets and the number of steps involving non trivial droplet operations. Yet, the cell-free approach has advantages that motivate further developments: less biases due to toxicity, no need for cloning, no transformation required and less background activity leading to a higher dynamic range. Highly active but toxic enzymes such as Streptomyces erythraeus trypsin (mentioned in the Chapter IV) would benefit from this type of approach that could be applied not only to genotype-phenotype mapping but also to laboratory artificial evolution.

Using the same cell-free expression system, two options are available to complete this work-flow. The first option is to finish developing the 0.2-2 pL droplets electro-coalescence step and to improve the handling of femtoliter droplets. A second option is to work with bigger droplets, at the expense of the throughput of the experiments. Alternative cell-free expression systems could also be tested or developed so as to find conditions for which expression can be carried out in a single step after gene encapsulation. Such cell-free expression systems would considerably simplify the cell-free microfluidic work-flow. Once a solution has been found, fluctuations of the expression yield will have to be assessed so as to determine if normalization for the expression level is necessary.

In the case of the rat trypsin, the use of *B. subtilis* secretion abilities simplifies the microfluidic work-flow, as DNA does not have to be amplified in droplets prior to expression. I have shown that it is possible to encapsulate and grow *B. subtilis* cells secreting trypsin as a fusion protein with a fluorescent reporter and to normalize trypsin activity using the reporter fluorescence. This allows to select droplets based on their displayed catalytic efficiency at the kHz frequency. I also generated by saturated mutagenesis a library of trypsin variants, which is ready to be screened. Further work, underway, are nevertheless required before performing the first genotype-phenotype mapping experiments. There is a need to understand why the few tested trypsin variants display catalytic activities that are different from what is expected from the literature. While the results presented in this manuscript are promising, I have to demonstrate that a culture of *B. subtilis* cells transformed with a subset of trypsin variants can be encapsulated, assayed, sorted and that the content of the droplets can be recovered and sequenced. If successful, such an experiment would be a full demonstration that the setup is operational. Once this has been shown, the library that I prepared will be transformed in *B. subtilis*, to perform the first genotype-phenotype mapping experiments.

Besides, preliminary experiments have shown that SGAP could also be expressed as a fusion protein with mCherry using WB800N (data not shown in the manuscript). Therefore, SGAP properties could also be studied using the in vivo work-flow in synergy with the newly developed rhodamine substrate. This suggests that the in vivo microfluidic work-flow is versatile and can be adapted to several proteins, provided microfluidics compatible fluorogenic substrates are available.

5.2 Perspectives

The two model systems were chosen to tackle fundamental questions about the properties of the genotype-phenotype landscape: SGAP to study the mechanisms of allostery and promiscuity and the rat trypsin to systematically test the sectors' properties in terms of functions and epistasis. How should those questions be approached with the developed microfluidic work-flows?

The optical setup allows to measure four different fluorescence channels simultaneously. Two channels being used for the coding dye and the reporter fluorescence, two channels are left to measure the fluorescence of two different substrates. It is therefore possible to assay SGAP with a cocktail of two substrates characterized with different fluorophores (e.g hydrophilic rhodamine and non leaky sulfo-coumarin) and different amino-acids in order to carry out specificity genotype-phenotype mapping, with specificity defined as the ratio of the two substrates fluorescence at the detection point. This experiment could be performed at different substrate and calcium concentrations in order to systematically test allosteric effects on the SGAP specificity profile. Such an experiment would also allow to

systematically determine the set of residues which are involved in the calcium allosteric effect.

The in vivo work-flow depicted in this manuscript only allows to test the variants for their catalytic activity. Testing the rat trypsin sectors's properties requires setting up experiments to also test the variants specificity and stability. Different probes with different amino-acid and different fluorophores will have to be designed, in order to test trypsin variants specificity in a way similar to SGAP. Testing the stability could simply be performed by comparing catalytic efficiency genotype-phenotype mapping data with and without heat shocks at different temperature: destabilized variants should display loss of activity. Nevertheless, this could be limited by the presence of B. subtilis cells in the droplets.

Finally, the same experiments will be repeated on libraries made of double point mutants in order to determine second order epistasis networks of residues and compare them with the sectors. At this point, other members of the S1A serine protease family will have to be interrogated by such genotype-phenotype mapping experiments and the results compared. Indeed, determining what is shared from what is idiosyncratic is fundamental so as to identify the statistical invariants of the protein sequence-function relationship.

Appendix A

Appendix: Development of a cell-free microfluidic work-flow for the genotype-phenotype mapping of *Streptomyces griseus* Aminopeptidase (SGAP)

A.1 Experimental details

A.1.1 Plasmid maps and primer sequences

SGAP gene had been cloned previously in the pIVEX vector downstream of a T7 promoter and upstream of a T7 terminator (see plasmid map in figure A.1).

The primers used for the PCR reactions are given in the paragraph A.1.3.

A.1.2 General comments about the experiments

Regarding the SGAP enzymatic assay experiments:

- The SGAP buffer is made of the following : 2 mM $CaCl_2$, 200 µM $ZnCl_2$, 40 mM Tricine (pH = 7.8).
- The fluorogenic substrates Bis-(l-leucinyl)-rhodamine 110 (Promega, Charbonniere, France) and Leu-Rho110-MC (NewChem Technologies, UK) are respectively denoted as Leu_2Rho and LeuRhoMS in the following paragraphs (see figure A.2). If not stated otherwise, they are used at a final concentration of 20 µM in SGAP buffer in all SGAP enzymatic assay experiments.
- If not stated otherwise, all enzymatic assay experiments are performed by adding 10 µL of sample to 90 µL of substrate in SGAP Buffer, in a 96 well plate (flat black bottom). Fluorescence kinetics measurements are performed for 1 h at 37°C in a spectrophotometer. The activity (in RFU/min) is defined as the linear slope of the fluorescence vs time graph over the last 20 minutes of kinetics.

Regarding the microfluidic experiments:

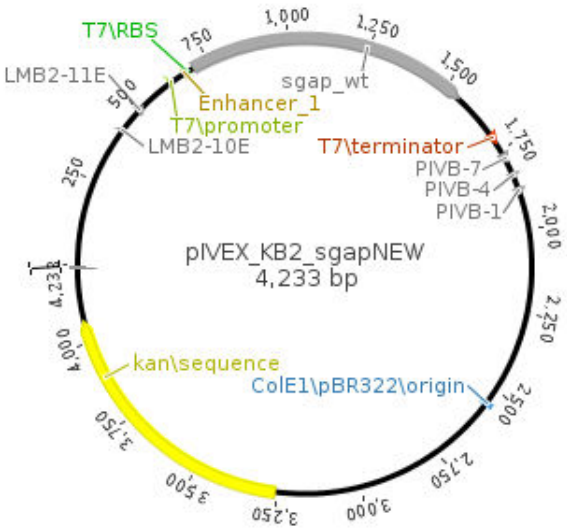


FIGURE A.1: **pIVEX SGAP plasmid map.** SGAP gene had been cloned previously downstream of a T7 promoter and upstream of a T7 terminator to enable in vitro transcription and translation with the "Pure" expression system.

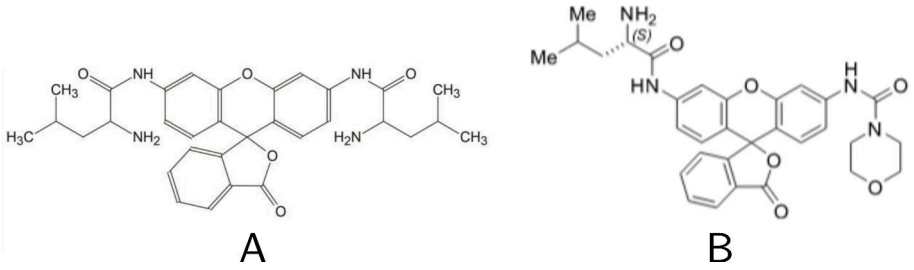


FIGURE A.2: **Structure of the two main substrates used in the experiments of this chapter. A**:Leu$_2$Rho. **B**:LeuRhoMS. $\lambda ex = 488nm, \lambda em = 535nm$

- The coding dye is an inert red fluorescent dye (dy647P1, Dyomics) used to determine emulsions identity, control Z focus on the microfluidic station and/or to control events distribution (such as pico-injection, fusion).
- In the microfluidics experiments, Pluronic F-127 yeast RNA were kept respectively 0.4 % w/v and 1 ng/µL in the water phase at all time.
- All emulsions are broken using 1H, 1H, 2H, 2H Perfluoro-1-octanol, 97% (Sigma Aldrich) to a final concentration of 40% v/v.

A.1.3 SGAP PCR protocol with Phusion polymerase

.

All SGAP bulk PCR reactions were performed with the following primers, mix and cycles:

PIVB-7seq	ACCACACCCGTCCTGTG
LMB2-10seq	ATGGCGCCCAACAGTCC

dNTP's	0.2 mM
DMSO	3% v/v
Primers	0.5 µM each
DNA template	1 ng
Polymerase	1 unit

Step	Temperature °C	Time
1	98	30 s
2	98	5 s
3	67	20 s
4	72	25 s
5	72	5 min

A.1.4 The PCR reagents inhibit cell-free expression in bulk 3.2.1

An IVTT mix was prepared according to the commercial protocol and mixed with the product of a bulk PCR (see A.1.3) on the SGAP gene. The following PCR dilution ratios were tested in a 100 µL final volume: 0 (no PCR product), 0.05, 0.1, 0.2, 0.5. After 3h of incubation at 37°C, we added 10 µL of the IVTT product to 90 µL of a solution of Leu_2Rho (20 µM). Fluorescence kinetics measurement were then performed.

A.1.5 Cell-free expression is successful in bulk and droplets 3.2.2

A SGAP plasmid solution ($\lambda = 2600$, 7 ng/µL) was emulsified in 2 pL droplets in presence of dy647 (80 µM). The emulsion was collected into a PCR tube. An IVTT mix with no DNA (dy647 1 mM) was prepared and injected into an 2 pL - 25 pL electrofusion device to produce 25 pL droplets which were fused with the 2 pL emulsion produced earlier, using syringe pumps and the following flow-rates (Nemesys):

Phase	Fluxes
Oil	250 µL/h
IVTT Mix	25 µL/h
2 pL droplets	25µL/h

The resulting emulsion was collected into a syringe and kept at 37°C during 4 h for the expression to occur inside the droplets. The emulsion was then recovered into a PCR tube and broken. After centrifugation, the aqueous phase was recovered and kept on ice. 10 µL of the aqueous phase were added to 90 µL of Leu_2Rho solution (final concentration 150 µM) in a 96 wells plate. Fluorescence kinetics measurement were then performed.

The other half of the IVTT solution without DNA was used to perform bulk positive and negative controls. DNA was added to the IVTT solution to 7 ng/µL final concentration. In the case of the negative control, water was added instead of the DNA solution. The resulting IVTT solutions were kept at 37°C. After 3, 4, 5, 6, 8 and 10 h of incubation, 10 µL of the positive control solution was used to perform enzymatic assays. Negative control assay was performed after 4 h of incubation.

A.1.6 Cell-free expression reagents inhibit SGAP enzymatic activity in bulk: pico-injection is incompatible with activity detection 3.2.3

A bulk PCR was performed and mixed to an IVTT mixture to a volume ratio PCR/IVTT of 5/95 (final volume: 150 µL). After 3h of incubation at 37°C, we added different volumes of IVTT product to 10 µL of a solution containing the fluorogenic substrate Leu_2Rho (20 µM final concentration): 10 µL, 50 µL and 89.4 µL. All mixes were topped up with water to a final volume of 100 µL and fluorescence kinetics measurement were then performed.

A.1.7 Fusing 2 pL expression and 20 pL assay droplets allows aminopeptidase activity detection with high contrast 3.3.1

Two IVTT mixtures were prepared with either 2.5 µM (+) or 25 µM (-) of emulsion identifier. In 100 µL of the + solution was added 5 µL of SGAP PCR product. In 100 µL of - was added 5 µL of water. Both mixtures were incubated at 37°C for 3 h for the in vitro expression to occur. The two mixtures were encapsulated in 2 pL droplets in two different droplet maker devices linked by their outlet to the same PCR tube. The fact that the two emulsions were showing identical basal fluorescence was confirmed. The mixture was kept on ice during the whole encapsulation process. An enzymatic assay mixture containing the LeuRhoMS mixture (final concentration 20 µM) and the emulsion identifier (5 µM) was used to produce 25 pL droplet into a 2 pL - 25 pL electrofusion device where they were fused with the 2 pL emulsion previously prepared. The resulting emulsion was collected into a PCR tube. The following flow rates were used:

Phase	Pressure
Oil	3400 mBar
Substrate solution	2000 mBar
2 pL emulsion	1630 mBar

The emulsion was incubated at room temperature for 1 h, to let the enzymatic reaction occur. The emulsion was then reinjected to measure the fluorescence distribution in the substrate and the emulsion identifier fluorescence channels. We used the following pressures:

Phase	Pressure
Oil	600 mBar
Emulsion	600 mBar

The emulsion identifier fluorescence distribution was analysed to determine the minimum of the red fluorescence distribution: 0.37 RFU. This minimum was used to discriminate the substrate droplets that where fused with the positive expression product (expression mix + DNA, 2.5 µM identifier, 52%) or the negative one (expression mix - DNA, 25 µM identifier, 48%). Aminopeptidase activity fluorescence distribution was then computed. Knowing that the 2 pL droplets showed identical basal activity fluorescence during their production and that all 25 pL droplets were identical before the fusion, the high fluorescence displayed by the positive emulsion was interpreted as the outcome of the SGAP aminopeptidase activity.

A.1.8 SGAP PCR amplification is successful in 0.2 pL droplets 3.3.3

The thermal diffusivity of the oil and the water at 30°C are the following: 1.4×10^{-7} and 0.35×10^{-7} m^2/s.
For this experiment, a SGAP PCR mix ($\lambda = 5$, 99 % occupancy according to the Poisson distribution) was encapsulated in 0.2 pL droplets using the following pressure settings (MFCS 7 bar, Fluigent):

Phase	Pressure
Oil (2 channels)	600 mBar
Water	1500 mBar

Thermocycling was performed according to the following cycle:

Step	Temperature °C	Time
1	98	2 min s
2	98	20 s
3	67	1 min 20 s s
4	72	1 min 40 s
5	72	20 min

After thermocycling, we broke the emulsion, recovered and loaded the aqueous phase on a gel (1% agarose) to assess the PCR yield in comparison with bulk "slow PCR" and classical PCR.

A.1.9 Random mutagenesis on the SGAP gene

Mutagenic PCR was performed according to manufacturer's protocol (Genemorph II, Agilent), with varying concentrations of target DNA in the PCR mix. The product of the mutagenic PCR was subcloned into the pIVEX plasmid using notI and ncoI restriction enzymes. E.coli chemically competent cells (top10, invitrogen) were transformed with the ligation product

and spread to grow on agar plates containing kanamycin. In each PCR condition, ten clones were isolated, plasmid DNA was purified and sequenced (GATC, Konstanz, Germany). Mutations were deduced by aligning sequencing results with the SGAP sequence using ClustalW2 alignment algorithm.

A.2 Synthesis of New Hydrophilic Rhodamine Based Enzymatic Substrates Compatible with Droplet-Based Microfluidic Assays 3.3.2

Synthesis of New Hydrophilic Rhodamine Based Enzymatic Probes Compatible With Droplet-Based Microfluidic Assays.

Johan Fenneteau,‡[a] Dany Chauvin,‡[b] Andrew D. Griffiths,[b] Clément Nizak[b] and Janine Cossy.[a]

[a]Laboratoire de Chimie Organique, Institute of Chemistry, Biology and Innovation (CBI), UMR 8231, ESPCI Paris/CNRS, PSL Research University, 10 rue Vauquelin, 75231-Paris Cedex 05, France

[b]Laboratoire de Biochimie, Institute of Chemistry, Biology and Innovation (CBI), UMR 8231, ESPCI Paris/CNRS, PSL Research University, 10 rue Vauquelin, 75231-Paris Cedex 05, France

Supplementary Information

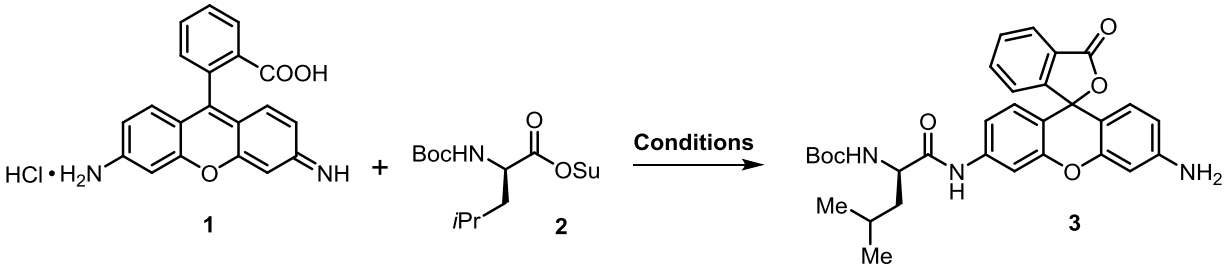
Table of Contents

General information S2
Optimization of conditions for compound **3**: S3
Experimental procedure S4
Microfabrication: S8
Enzymatic assays and substrate exchange experiments: S8
Fig S1: Bulk kinetics assay: S9
Fig S2: Workflow of the exchange test experiments in droplets: S10
Fig. S3: Exchange test experiment in droplets for compound **6**: S12
Fig. S4: Exchange test experiment in droplets for compound **8** : S13
Fig. S5: Quantifying the exchange of probes **6** and **8**: S14
Fig. S6: Testing SGAP activity towards **6** in droplets - microfluidic chip design: S15
Fig. S7: Signal to noise ratio at different concentrations of **6**: S16
Fig. S8: Discrimination of different enzymatic activity levels: S17
Copy of ^{1}H and ^{13}C spectra : S18

General information

All reactions were carried out under an argon atmosphere unless otherwise specified. Flasks were oven-dried at 120 °C and cooled under argon prior to use. CH_2Cl_2 was distilled over calcium hydride. All others commercially available chemicals were purchased from Aldrich and used directly without purification. TLCs were performed on Merck $60F_{254}$ silica gel plates and visualized with UV lamp (254 or 360 nm), and by treatment with a solution of $KMnO_4$, K_2CO_3, NaOH (5%) in H_2O followed by heating. Reverse phase TLC were performed on Merck 60 RP-18 F_{254} silica gel plates and visualized with UV lamp (254 or 360 nm). Flash column chromatographies were performed with Merck Geduran Si 60 silica gel (40-63 μm). Reverse phase chromatography were performed using a Biotage Isolera One flash purification system using Biotage SNAP cartridge KP-C18-HS 12g and eluted with HPLC grade MeCN/H_2O. The fractions were collected using UV detection at 254 and 360 nm and concentrated under reduced pressure at 40 °C, residual water was removed *via* lyophilisation.^{1}H NMR spectra were recorded on a Bruker Avance 400 at 400 MHz. The chemical shifts δ are reported in ppm relative to tetramethylsilane (TMS) or residual protonated solvents (TMS: δ_H = 0.00 ppm, $CHCl_3$: δ_H = 7.26 ppm, MeOD: δ_H = 3.31 ppm) The multiplicity and shape of signals are designated by the following abbreviations: s = singlet, d = doublet, t = triplet, q = quartet, quin = quintet, m = multiplet, br = broad. Coupling constants *J* are reported in Hertz (Hz).^{13}C NMR spectra were recorded on a Bruker Avance 400 at 100 MHz. The chemical shifts δ are reported in ppm relative to the solvent as an internal indicator ($CDCl_3$ δ 77.16 ppm, CD_3OD δ 49.00 ppm). IR-spectra were recorded on a Bruker TENSOR^{TM} 27 (IRTF). The samples were prepared as neat films or as fine powders. Only selected absorbances (ν_{max}) are reported, and wave numbers are reported in cm^{-1}. High resolution mass spectra (HRMS) were performed by the Centre Regional de Microanalyse (Université Pierre et Marie Curie VI, Paris, France).

Optimization of conditions for compound 3:

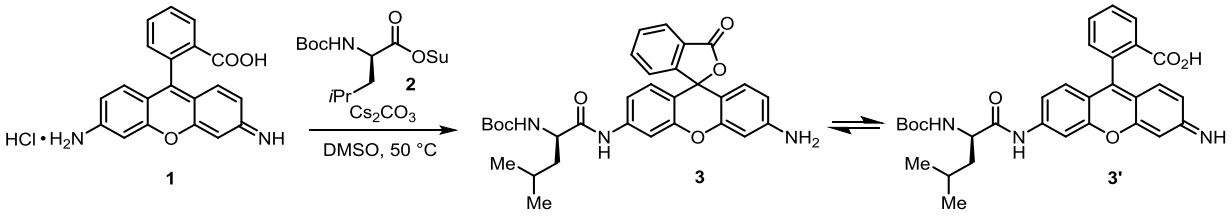


Reagent	Base	Solvent	Temperature	Yield [a]
2, (1.1 eq.)	K_2CO_3 (5 eq.)	DMF/H_2O	23 °C	12% [b]
2, (1.5 eq.)	Pyridine (excess)	DMF	23 °C	-
2, (1.5 eq.)	Cs_2CO_3 (1 eq.)	DMF/Pyridine	23 °C	18% [c]
2, (1.1 eq.)	Et_3N (3 eq.)	1,4-dioxane	50 °C	29%
2, (1.5 eq.)	Et_3N (2 eq.)	DMSO	23 °C	27%
2, (1.5 eq.)	Cs_2CO_3 (3 eq.)	DMSO/H_2O	23 °C	< 10% [d]
2, (1.5 eq.)	Cs_2CO_3 (3 eq.)	DMSO	23 °C	37%
2, (1.5 eq.)	Cs_2CO_3 (3 eq.)	DMSO	50 °C	53%
2, (1.5 eq.)	-	DMSO	23 °C	-

[a] Isolated yield after purification. [b] DMF/H_2O (9/1) mixture was used, [c] DMF/Pyridine (3/1) mixture was used, [d] DMSO/H_2O (9/1) mixture was used. All the reactions were performed on a 0.1 mmol scale at 0.1 M concentration.

Experimental procedure

***tert*-Butyl ((2*R*)-1-((3'-amino-3-oxo-3*H*-spiro[isobenzofuran-1,9'-xanthen]-6'-yl)amino)-4-methyl-1-oxopentan-2-yl)carbamate(3):**



To a solution of rhodamine 110 hydrochloride (40 mg, 0.109 mmol, 1 eq) in DMSO (1 mL) was added Cs_2CO_3 (106 mg, 0.327 mmol, 3 eq) followed by *N*-Boc-L-leucine *N*-hydroxysuccinimide ester (53 mg, 0163 mmol, 1.5 eq) and heated at 50 °C for 12 h. The mixture was then dilute with EtOAc (40 mL), washed with H_2O (3 x 10 mL), brine (3 x 10 mL), dried over $MgSO_4$, filtered and concentrated under reduced pressure. The resulting orange oil was purified by flash chromatography on silica gel 10:90 to 30:70 (EtOAc/$CHCl_3$) to afford compound **3** as an amorphous orange solid (31 mg, 0.057 mmol, 53%).

This compound exists as an equilibrium mixture of **3** and its open form **3'** in $CDCl_3$.

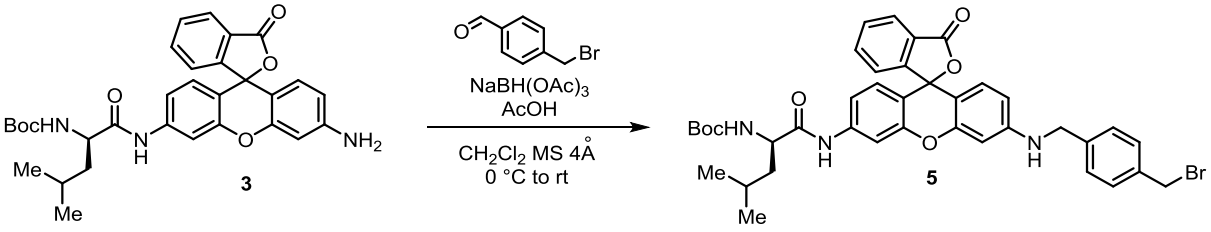
^{1}H NMR (400 MHz, $CDCl_3$) *Described as 1/0.9 mixture of tautomers: Major tautomer* δ 8.74 (brs, 1H), 8.00 (s, 1H), 7.69 (d, *J* = 2.1 Hz, 1H), 7.65 – 7.55 (m, 2H), 7.07 (d, *J* = 7.7 Hz, 1H), 6.99 (dd, *J* = 8.6, 2.2 Hz, 1H), 6.67 (d, *J* = 2.5 Hz, 1H), 6.51 (s, 1H), 6.49 (d, *J* = 2.3 Hz, 1H), 6.33 (t, *J* = 2.0 Hz, 1H), 5.05 (d, *J* = 8.1 Hz, 1H), 4.26 (brs, 1H), 3.91 (brs, 2H),1.80 – 1.70 (m, 2H), 1.62 – 1.52 (m, 1H), 1.44 (s, 9H), 0.96 (t, *J* = 6.7 Hz, 6H) ppm. *Minor tautomer* δ 8.74 (brs, 1H), 7.98 (brs, 1H), 7.66 (d, *J* = 2.1 Hz, 1H), 7.65 – 7.55 (m, 2H), 7.10 (d, *J* = 7.7 Hz, 1H), 6.95 (dd, *J* = 8.6, 2.2 Hz, 1H), 6.65 (d, *J* = 2.5 Hz, 1H), 6.53 (s, 1H), 6.45 (brs, 1H), 6.31 (t, *J* = 2.0 Hz, 1H), 5.05 (d, *J* = 8.1 Hz, 1H), 4.26 (brs, 1H), 3.91 (brs, 2H), 1.80 – 1.70 (m, 2H), 1.62 – 1.52 (m, 1H), 1.44 (s, 9H), 0.96 (t, *J* = 6.7 Hz, 6H) ppm.

^{13}C NMR (100 MHz, $CDCl_3$) *Described as 1/0.9 mixture of tautomers* δ [171.89, 171.83], 169.97, 156.56, 153.29, [152.55, 152.51], 151.86, [149.11, 149.07], 139.92, [135.16, 135.12], [129.70, 129.67], [129.00, 128.96], 128.50, [126.87, 126.83], 124.94, [124.13, 124,13], [115.32, 115.13], [114.61, 114.46], [111.74, 11.69], [108.24, 108.21], [107.93, 107,81], [101.63, 101, 58], 83.95, 80.69, 54.11, 40.91, [28.47, 28.45], 24.88, [23.14, 23.12], 21.89 ppm.

IR (neat): 3306, 2957, 2926, 2854, 1744, 1672, 1613, 1534, 1508, 1366, 1286, 1249, 1208 1168, 1113, 870 cm^{-1}.

ESI-HRMS: *m*/*z* calculated for $C_{31}H_{34}N_3O_6$ $[M+H]^+$: 544.2442, found: 544.2436.

***tert*-Butyl ((2*R*)-1-((3'-((4-(bromomethyl)benzyl)amino)-3-oxo-3*H*-spiro[isobenzofuran-1,9'-xanthen]-6'-yl)amino)-4-methyl-1-oxopentan-2-yl)carbamate (5)**:



To a solution of **3** (170 mg, 0.324 mmol, 1 eq) and 4-bromomethyl benzaldehyde[1] (77 mg, 0.389 mmol, 1.2 eq) in CH_2Cl_2 (5 mL) was added activated 4 Å MS (200 mg), $NaBH(OAc)_3$ (109 mg, 0.518 mmol, 1.6 eq) followed by the addition of AcOH (100 µL). The solution turned from orange to red and the flask was covered from the light with aluminium foil. After 3 h, an aqueous K_2CO_3 solution (10%) was added and the reaction mixture was extracted with CH_2Cl_2 (3 x 10 mL). The combined organic phases were dried over $MgSO_4$, filtered, concentrated under reduced pressure and purified by flash chromatography (protected from light with aluminium foil) on silica gel 10:90 to 20:80 (EtOAc/$CHCl_3$) to afford the unstable and light sensitive compound **5** (189 mg, 0.260 mmol, 79%) as a red oil.

This compound is unstable and light sensitive and has to be kept in the dark and engaged in the next step within 2 h, otherwise extensive decomposition was observed.

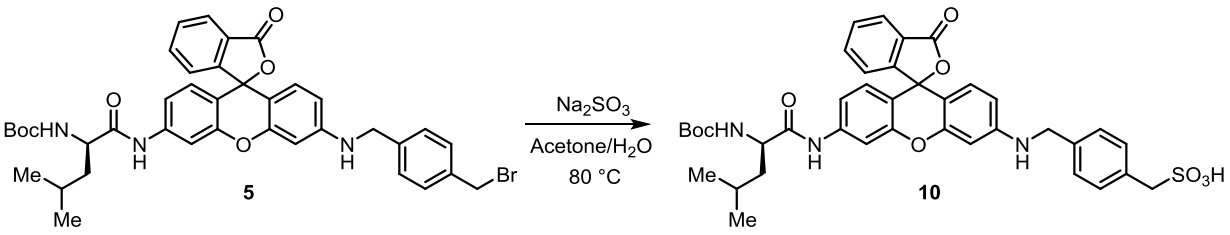
^{1}H NMR (400 MHz, $CDCl_3$) δ 8.94 (brd, 1H), 7.98 (m, 1H), 7.71 – 7.52 (m, 3H), 7.36 – 7.26 (m, 4H), 7.02 (m, 2H), 6.62 (d, J = 8.6 Hz, 1H), 6.47 (d, J = 8.6 Hz, 1H), 6.37 – 6.23 (m, 2H), 5.18 (brs, 1H), 4.47 (s, 2H), 4.36 – 4.23 (m, 3H), 4.28 (brs, 1H), 1.77 – 1.66 (m, 2H), 1.62 – 1.53 (m, 1H), 1.43 (s, 9H), 0.98 – 0.81 (m, 6H) ppm.

IR (neat): 3307, 2958, 2929, 2854, 1747, 1674, 1613, 1510, 1413,1366, 1335,1285, 1250, 1166, 1124, 1224, 1084, 1020, 870 cm^{-1}.

ESI-HRMS: *m*/*z* calculated for $C_{39}H_{41}BrN_3O_6$ $[M+H]^+$: 726.2173, found: 726.2173.

[1]Prepared according procedure described by : L. Wen, M. Li and J. B. Schlenoff, *J. Am. Chem. Soc*. 1997, **119**, 7726.

Sodium (4-(((3'-((*R*)-2-((*tert*-butoxycarbonyl)amino)-4-methylpentanamido)-3-oxo-3*H*-spiro[isobenzofuran-1,9'-xanthen]-6'-yl)amino)methyl)phenyl)methanesulfonate (10):



To a solution of bromide **5** (152 mg, 0.209 mmol, 1 eq) in acetone (4 mL) was added a solution of Na_2SO_3 (131 mg, 1.04 mmol, 5 eq) in H_2O (2 mL). The mixture was then heated at 80 °C for 1 h under vigorous stirring. The mixture was then cooled to room temperature, concentrated and engaged in the next step without further purification.

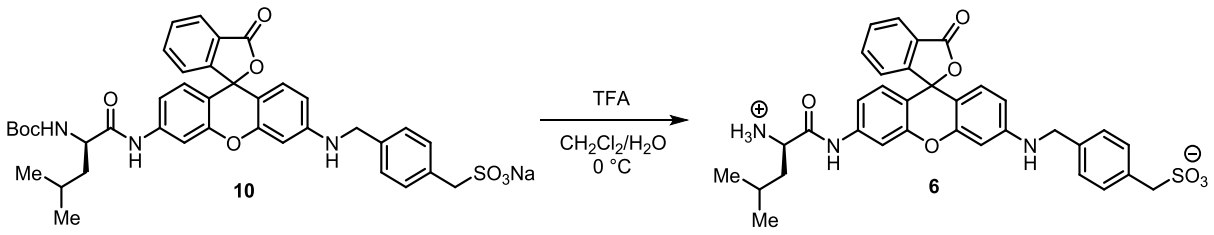
An aliquot (15 mg) was purified with reverse phase silica gel chromatography (40:60 to 60:40 MeCN/H_2O + 0.1% TFA) to afford, after lyophilisation, pure sulfonate **10** (4 mg) for characterisation purpose.

^{1}H NMR (400 MHz, MeOD) δ 8.32 (m, 2H), 7.85 (m, 2H), 7.48 – 7.37 (m, 4H), 7.37 – 7.30 (m, 2H), 7.20 (m, 1H), 7.11 (m, 1H), 7.03 – 6.86 (m, 2H), 4.63 (s, 2H), 4.28 (m, 1H), 4.04 (s, 2H), 1.75 (m, 1H), 1.68 – 1.50 (m, 2H), 1.45 (s, 9H), 0.99 (s, 3H), 0.97 (s, 3H) ppm. *-SO_3H and -NH protons(x3) were not observed in these conditions*.

IR (neat): 3326, 2963, 2926, 2854, 2362, 1744, 1684, 1614, 1517, 1414, 1366, 1286, 1249, 1168, 1120, 873 cm^{-1}.

ESI-HRMS: *m*/*z* calculated for $C_{39}H_{42}N_3O_9S$ $[M+H]^+$: 728.2636, found: 728.2639.

(4-(((3'-((*R*)-2-ammonio-4-methylpentanamido)-3-oxo-3*H*-spiro[isobenzofuran-1,9'-xanthen]-6'-yl)amino)methyl)phenyl)methanesulfonate (6):



The crude sulfonate **10** was dissolved in CH_2Cl_2/H_2O (10:1 - 2.2 mL) at 0 °C and TFA (2 mL) was added dropwise to the mixture. After 0.5 h the mixture was carefully concentrated under reduced pressure. The residue was then adsorbed on C18 SiO_2, using MeCN, and twice purified on reverse phase silica gel chromatography (20:80 to 50:50 $MeCN/H_2O$ + 0.1% TFA) to obtain pure sulfonate **6** (49 mg, 78.1 μmol, 37% over two steps after lyophilisation).

The purity of **6** was measured to be >95% by UV (at 254 nm) and ^{1}H NMR.

^{1}H NMR (400 MHz, MeOD) δ 8.33 (m, 2H), 7.91 – 7.74 (m, 2H), 7.52 – 7.46 (m, 1H), 7.44 (d, *J* = 8.1 Hz, 1H), 7.39 (d, *J* = 7.3 Hz, 2H), 7.34 (d, *J* = 8.1 Hz, 2H), 7.22 (d, *J* = 9.1 Hz, 1H), 7.12 (d, *J* = 9.3 Hz, 1H), 6.98 (d, *J* = 9.3 Hz, 1H), 6.90 (s, 1H), 4.64 (s, 2H), 4.12 (t, *J* = 6.8 Hz, 1H), 4.05 (s, 2H), 1.87 – 1.72 (m, 3H), 1.05 (d, *J* = 5.5 Hz, 3H), 1.04 (d, *J* = 5.5 Hz, 3H) ppm. *–NH protons (x 5) were not observed in these conditions*.

^{13}C NMR (100 MHz, MeOD) δ 170.5, 168.3, 161.0, 160.8, 156.0, 146.6, 136.7, 134.4 (2C), 132.8, 132.2, 132.0, 131.8, 131.7, 131.1, 130.7, 128.5, 120.7, 119.3, 119.1, 117.9, 117.7, 107.8, 96.8, 58.1, 54.0, 48.1, 41.4, 38.2, 23.3, 21.8 ppm. *One C^{IV} is not observed due to relaxation*.

IR (neat): 2965, 2854, 2360, 1680, 1599, 1557, 1512, 1445, 1340, 1259, 1051, 1014, 1202, 1136, 1033 cm^{-1}.

ESI-HRMS: *m*/*z* calculated for $C_{34}H_{34}N_3O_7S$ $[M+H]^+$: 628.2112, found: 628.2107

Microfabrication:

All microfluidic chips were fabricated in PDMS using soft-lithography techniques (J. Cooper McDonald *et al.*[3]) according to the protocol already described by Mazutis *et al.*[2]. If not stated otherwise, microfluidic channels are 20 µm deep.

Enzymatic assays and substrate exchange experiments:

Purified Aminopeptidase I from SGAP was purchased from Sigma Aldrich (A9934 SIGMA) and dissolved in SGAP buffer (2 mM $CaCl_2$, 200 µM $ZnCl_2$, 40 mM Tricine, pH = 8.0) to a concentration of 1 mg/mL. Solid substrate was dissolved in SGAP buffer to a concentration of 10 µM. For each substrate (**6** or **8**), a positive assay and a negative assay (no enzyme) were performed at 37 °C. In each well, 1 µL of SGAP solution or 1 µL of enzymatic buffer was added to 99 µL of substrate in enzymatic buffer (10 µM final substrate concentration). Evolution of the fluorescence was recorded in a fluorescence plate reader (SpectraMax i3x, Molecular Devices) with the following spectral settings: λ_{ex}= 488 nm, λ_{em}= 535 nm.

[2] L. Mazutis, J. Gilbert, W. L. Ung, D. A. Weitz, A. D. Griffiths and J. A. Heyman, Nat. Protoc. 2013, 8, 870.

[3] J. C. McDonald, D. C. Duffy, J. R. Anderson, D. T. Chiu, H. Wu, O. J. A. Schueller and G. M. Whitesides, *Electrophoresis* 2000, **21**, 27.

Fig S1: Bulk kinetics assay:

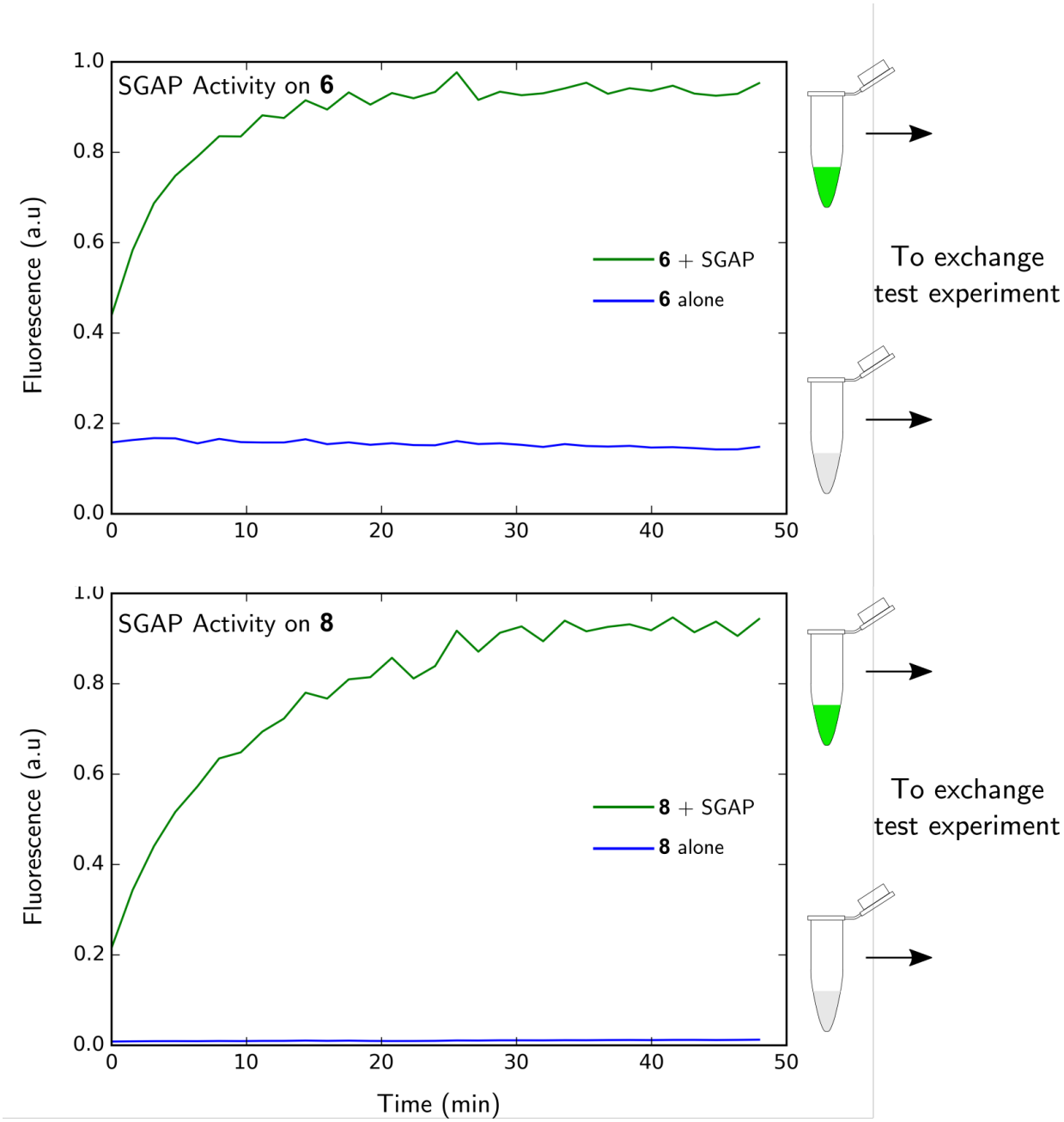


Fig. S1: Fluorescence evolution of **6** and **8** in the presence (green) or absence (blue) of SGAP enzyme.

For each substrate (**6** or **8**), once saturation of the fluorescence of the positive assay was achieved, well contents from positive and negative assays were retrieved into Eppendorf tubes. A control fluorescent dye (dy647, dyomics, ex: 653 em: 672 nm) that does not get exchanged between droplets was added to the positive and the negative solutions at a final concentration of respectively 500 nM and 1 µM. Once mixed, the solutions were used to perform exchange test experiments in droplets (Fig. S2).

Fig S2: Workflow of the exchange test experiments in droplets:

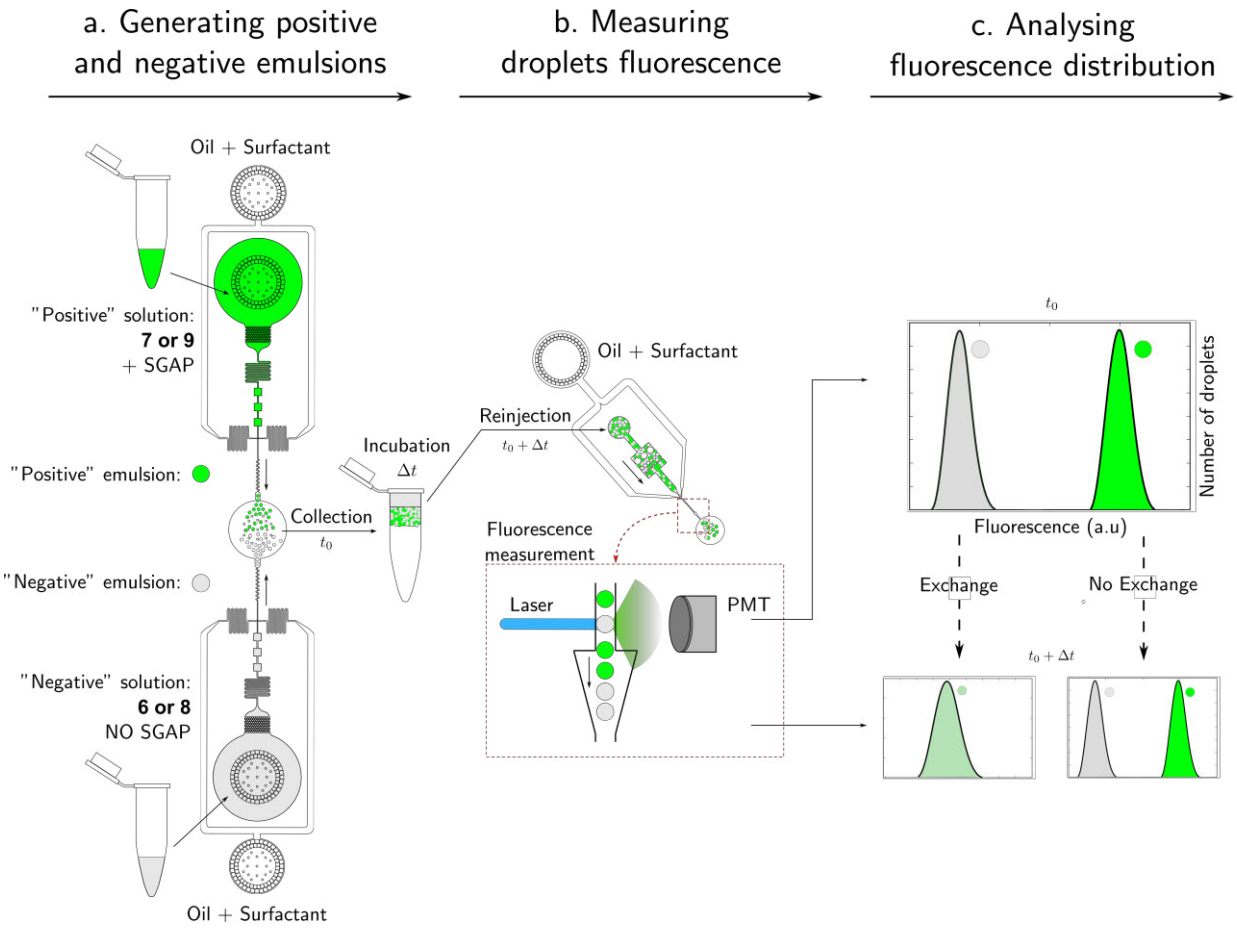


Fig. S2: Workflow of the exchange test experiment in droplets

a. Both positive (green, with SGAP) and negative (grey, without SGAP) solutions were compartmentalized separately in 20 pL droplets at a generation frequency of approximately 1kHz. They were collected simultaneously using a double droplet maker design (two droplet makers with a common outlet). We used HFE 7500 (3M) fluorinated oil containing "008-fluorosurfactant" (RAN biotechnologies) as surfactant (2% wt.). In each droplet maker, the oil phase was injected at 1 bar into the oil inlet of the chip, while aqueous phase was injected at 1.2 bar. Solutions were pressurised using Fluigent pressure-controlled pumps.

Collection was performed in an Eppendorf tube sealed with a plug of PDMS. The resulting emulsion was left to incubate at room temperature (23°C), in the dark.

b. The reinjection microfluidic chip was placed on the stage of an inverted fluorescence microscope (Nikon Eclipse Ti). Two monochromatic lasers (488 nm Coherent and 640 nm Crystalaser), and two photomultipliers (PMT, Hamamatsu) were directed to the microscope objective using a dichroic mirror and a dedicated set of filters to hit the microfluidic channel in which droplets were circulating and to measure their fluorescence. As the droplets crossed the laser beam, the PMT signal was acquired and recorded via an FPGA card (National Instruments) connected to a computer. The signal was then analysed by custom-developed software which measures and records the fluorescence intensity of each droplet, in two spectral windows: 506-545 (rhodamine fluorescence) and 671-746 (control fluorescence).

To assess exchange progression over time of the hydrolysed substrate, part of the emulsion was reinjected using pressure pumps (Fluigent) immediately after the end of collection (t_0) and after different incubation time points ($t_0 + \Delta t$). Oil and aqueous phases were pressurised at 0.5 mBar, 1 kHz reinjection frequency. During reinjection the fluorescence of each droplet was recorded. The fluorescence of the control dye is used to ensure a constant Z-focus of the microscope throughout the experiment. Fluorescence distributions in both spectral windows were then compared with t_0 fluorescence distributions. Reinjection and fluorescence measurements were performed at t_0 + 0.5 h, t_0 + 1 h, t_0 + 2 h, t_0 + 3 h, t_0 + 4 h and t_0 + 16 h (fluorescence distributions over time are depicted in Supporting Figure 3).

c. The evolution of the fluorescence intensity distribution can be analysed to assess the exchange rate of hydrolysed substrate between the emulsions. The two possible extreme cases are illustrated here. “No exchange”: negative + positive emulsion fluorescence distribution is bimodal and identical to t_0 distribution. “Exchange”: the distribution collapsed into one single mode at an intermediate intensity with respect to initial positive and negative emulsions. In the latter case, we can no longer distinguish positive and negative populations after the incubation time Δt. Therefore, in the case of a variant screening, it would be impossible to discriminate between low and high activity mutants of the SGAP enzyme.

Fig. S3: Exchange test experiment in droplets for compound 6:

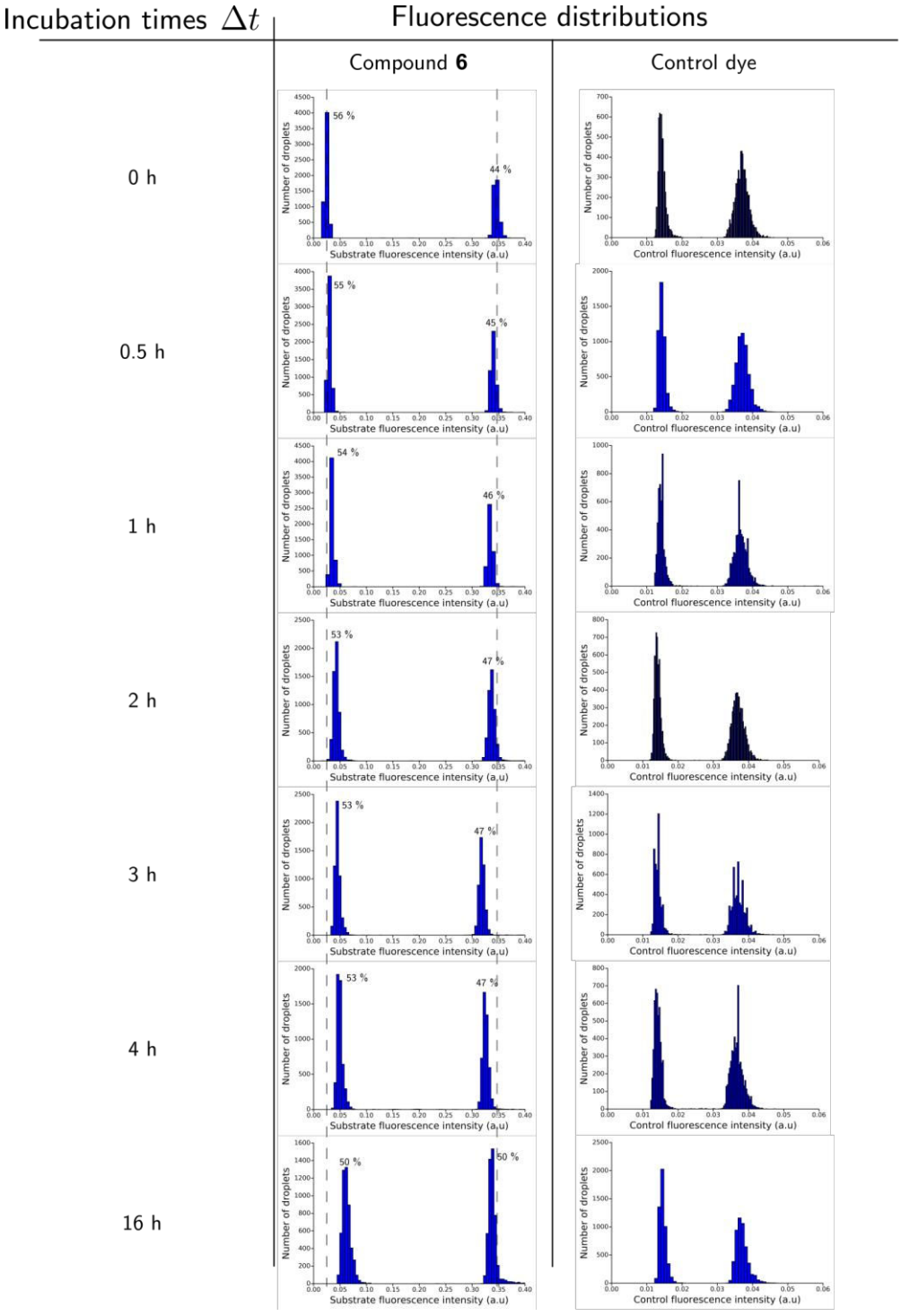


Fig. S3: Exchange test experiment for compound **6** and control dye.

The fluorescence profile (substrate fluorescence and control) of 10 000 droplets was measured at each time point. Minor changes in the droplet count ratio of each subpopulation are explained by the sampling fluctuations and possible local heterogeneities in the collection tube.

Fig. S4: Exchange test experiment in droplets for compound 8 :

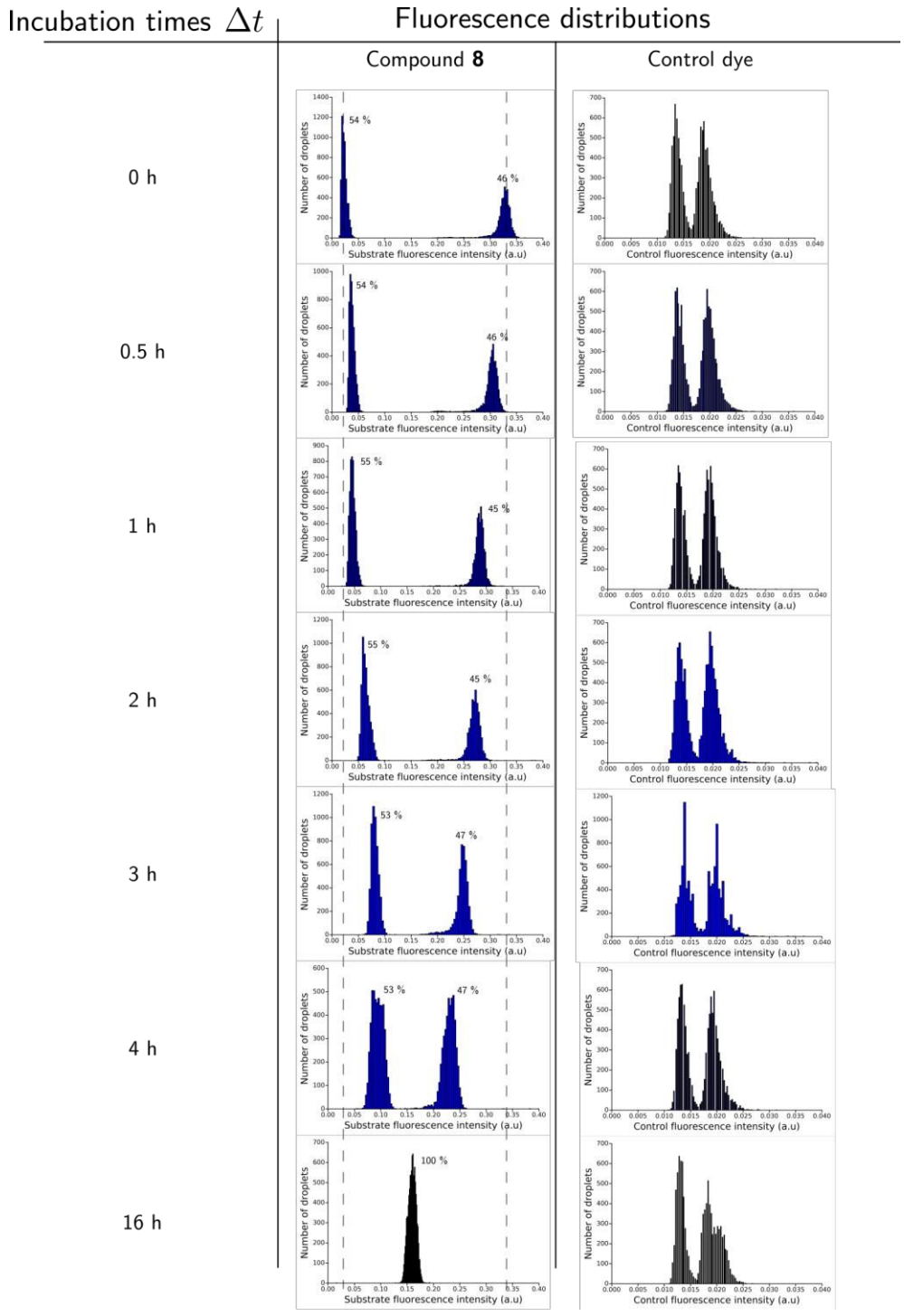


Fig. S4: Exchange test experiment for compound **8** and control dye.

The fluorescence profile (substrate fluorescence and control) of 10 000 droplets was measured at each time point. Minor changes in droplet counts ratio of each subpopulation are explained by the sampling fluctuations.

Fig. S5: Quantifying the exchange of probes 6 and 8:

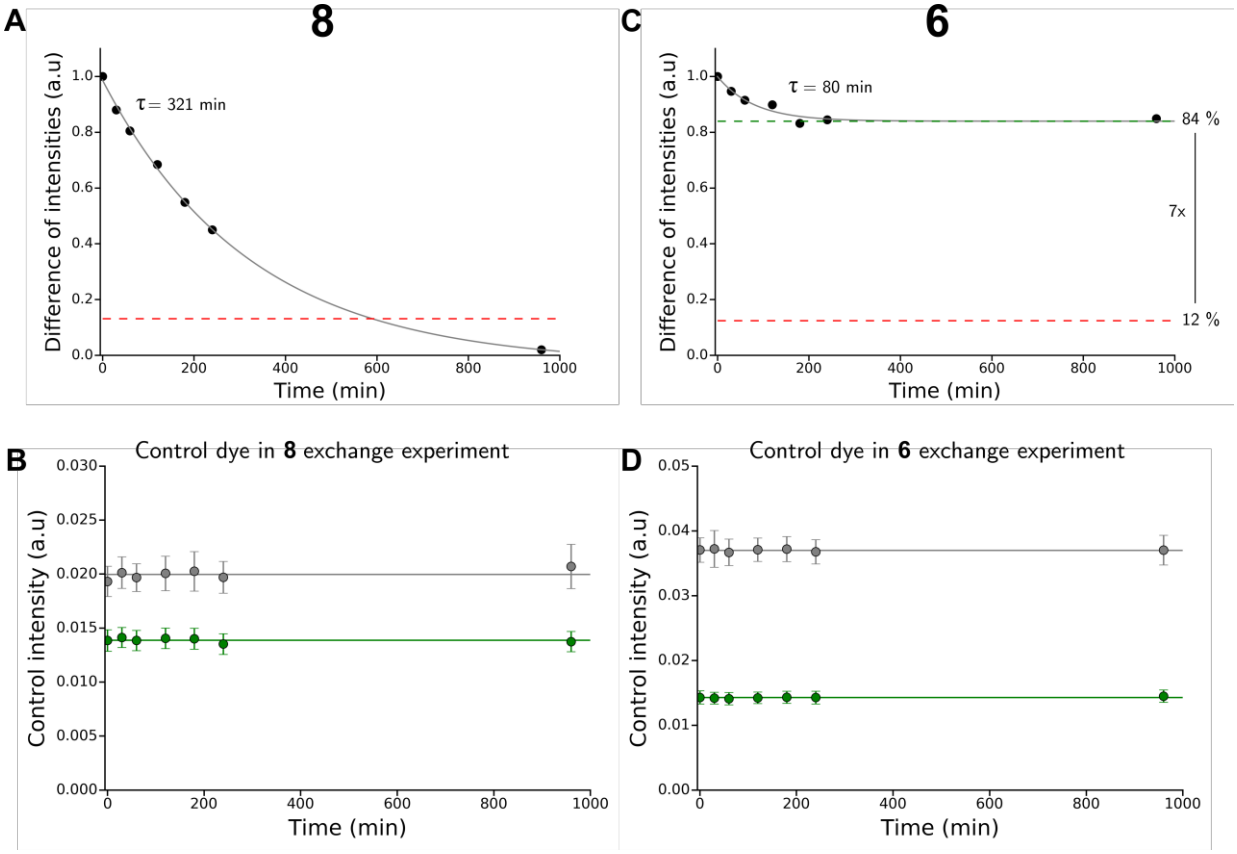


Fig. S5: Quantifying the exchange of **6** and **8**

A and **C**: For the compounds **6** and **8** respectively, the difference of the average fluorescence of positive and negative populations is computed at each time point and normalized to the initial difference. The red dotted line represents the sum of the standard deviations: $\sigma_{positive} + \sigma_{negative}$. **A**: For **6,** it is no longer possible to distinguish between the two populations after 600 min. Data were fitted according to exponential decay with a time constant of 320min+/-15min. **C**: For **8**, following an initial transient exchange (exponential decay fitted time constant of 80 min+/-25min), the difference between the average fluorescence of positive and negative emulsions remains constant at 84% of the initial contrast. The mechanism of the initial transient exchange remains to be investigated.

B and **D**: A control dye that emits in infrared and that does not get exchanged between droplets, dy647, is present in the same droplets as **6** or **8**. Its fluorescence intensity is monitored simultaneously to that of **6** or **8** to ensure constant optical settings throughout the experiment. **B:** Control intensity during the **8** exchange experiment. **D:** Control intensity during the **6** exchange experiment.

Fig. S6: Testing SGAP activity towards 6 in droplets - microfluidic chip design:

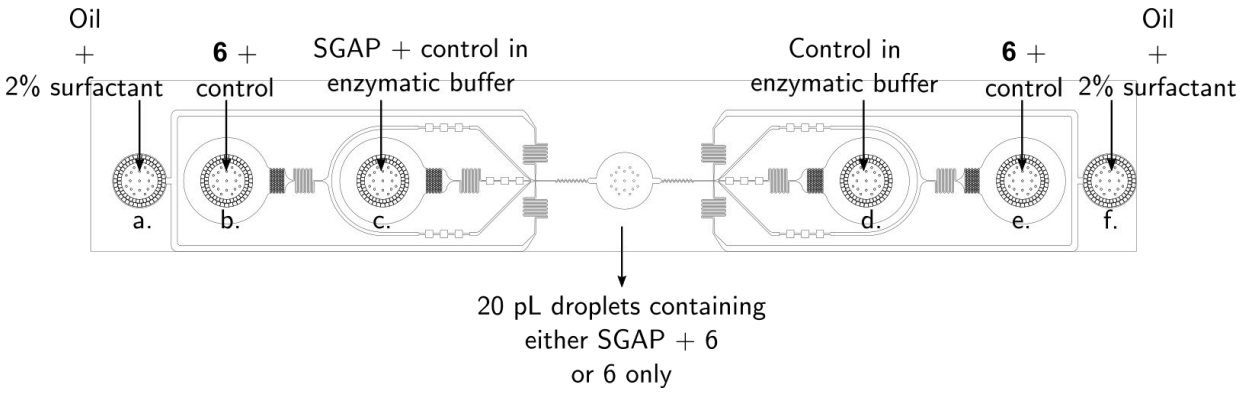


Fig. S6: Testing SGAP activity towards 6 in droplets

For the purpose of this experiment, pluronic F-127 (Sigma Aldrich) was added to the enzymatic buffer (0.4% w/v). A solution of **6** and control dye "dy647" was prepared in enzymatic buffer to the respective final concentrations 100 µM and 10 µM and connected to the b. and e. inlet. A solution of SGAP enzyme and control dye "dy647" was prepared in enzymatic buffer to the respective final concentrations 0.2 mg/mL and 10 µM and connected to the c. inlet. Finally a solution of control "dye647" was prepared in enzymatic buffer to 10 µM final concentration and connected to the d. inlet. Except for the presence of two aqueous inlets in this experiment, everything was performed according to the exchange test experiments protocol. The droplets fluorescence was measured in the same spectral windows, during production (t_0) and after 16h of incubation at room temperature (23°C) in the dark during reinjection (same device as in Fig. S2). The control fluorescence was used to ensure comparable measurements between t_0 and t_0 + 16h.

Fig. S7: Signal to noise ratio at different concentrations of 6:

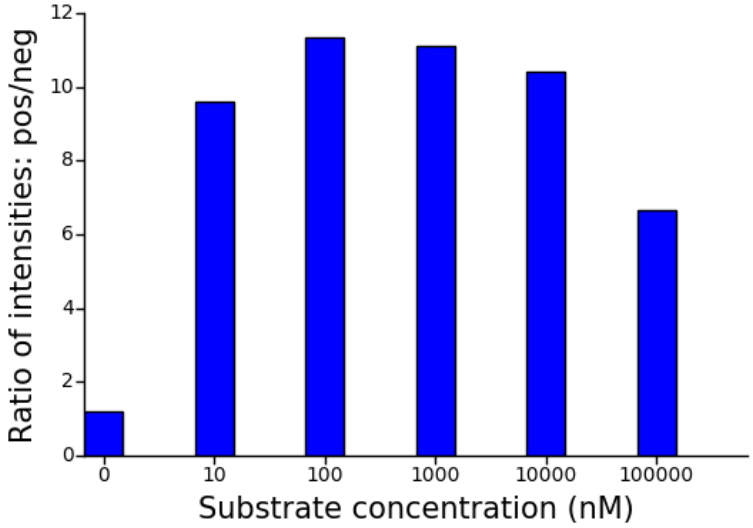


Fig. S7: Signal to noise ratio as different concentrations of 6

To investigate the influence of substrate concentration on signal to noise ratio, enzymatic assays were performed with the following concentrations in enzymatic buffer supplemented with pluronic F-127, 0.4%, in a spectrophotometer (100 µL per assay):

Assay	SGAP final concentration (mg/mL)	**6** concentration (nM)
1	0.01	0
2	0.01	10
3	0.01	100
4	0.01	10^3
5	0.01	10^4
6	0.1	10^5

In each case, a negative control (no enzyme) was also performed. For each concentration, once hydrolysis was complete, the ratio between the final positive assay fluorescence and the final negative assay fluorescence was computed. An optimum is found between 100 nM and 1 µM where the positive assay is 11 times more fluorescent than the negative control. Between 10 µM and 100 µM this ratio is located between 10 and 7, which is consistent with the results of our experiments in droplets.

Fig. S8: Discrimination of different enzymatic activity levels:

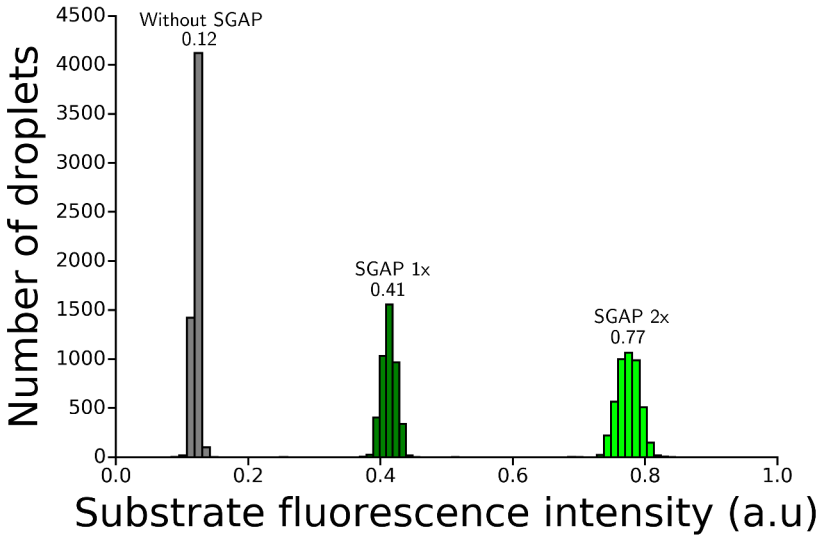


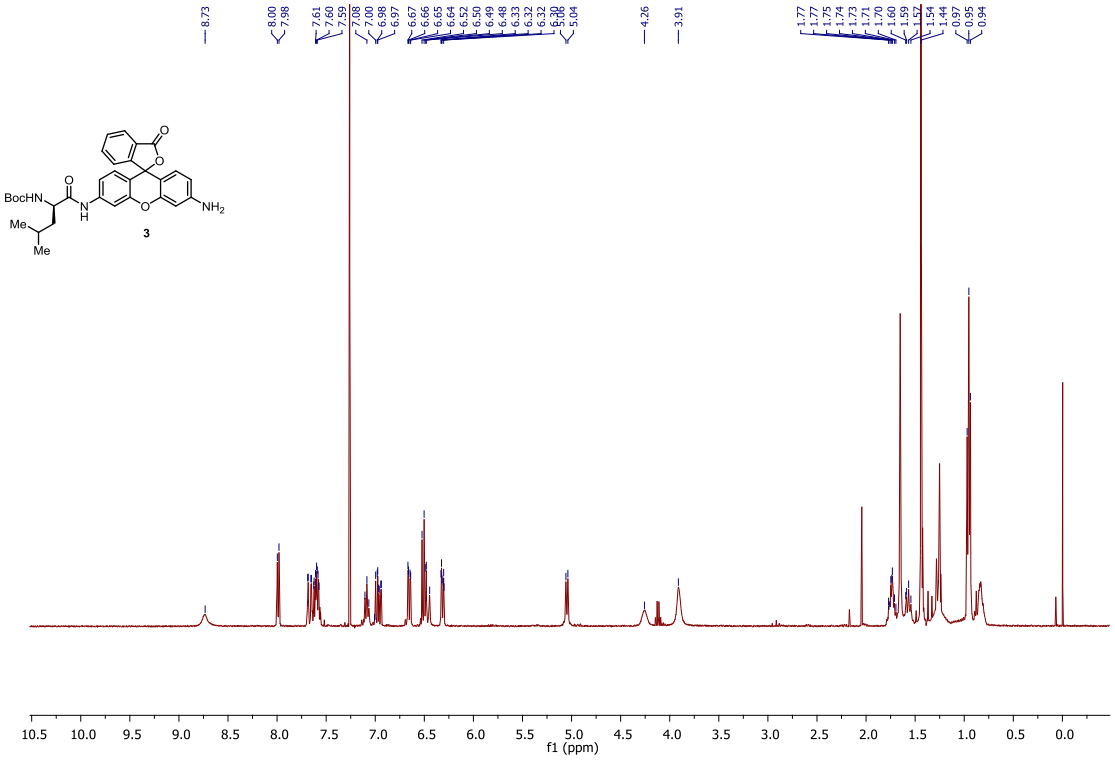
Fig S8: Discrimination of different enzymatic activity levels using our droplet microfluidics assay.

We have applied our assay to droplets containing SGAP at three different concentrations, mimicking enzymes with different k_{cat} Michaelis Menten parameters. We produced three emulsions containing probe **6** and either no SGAP, SGAP at 0.1 mg/mL (1x), or SGAP at 0.2 mg/mL (2x). The fluorescence distributions of the three droplet populations after 30 s of incubation are plotted here. The three populations are very well discriminated, which would make sorting for either activity level straightforward. The average fluorescence of the 0.2 mg/mL SGAP emulsion (0.77 A.U.) is approximately twice as high as the fluorescence of the 0.1 mg/mL SGAP emulsion (0.41 A.U.), confirming that our droplet assay based on probe **6** allows to quantitatively measure relative reaction rates in droplets.

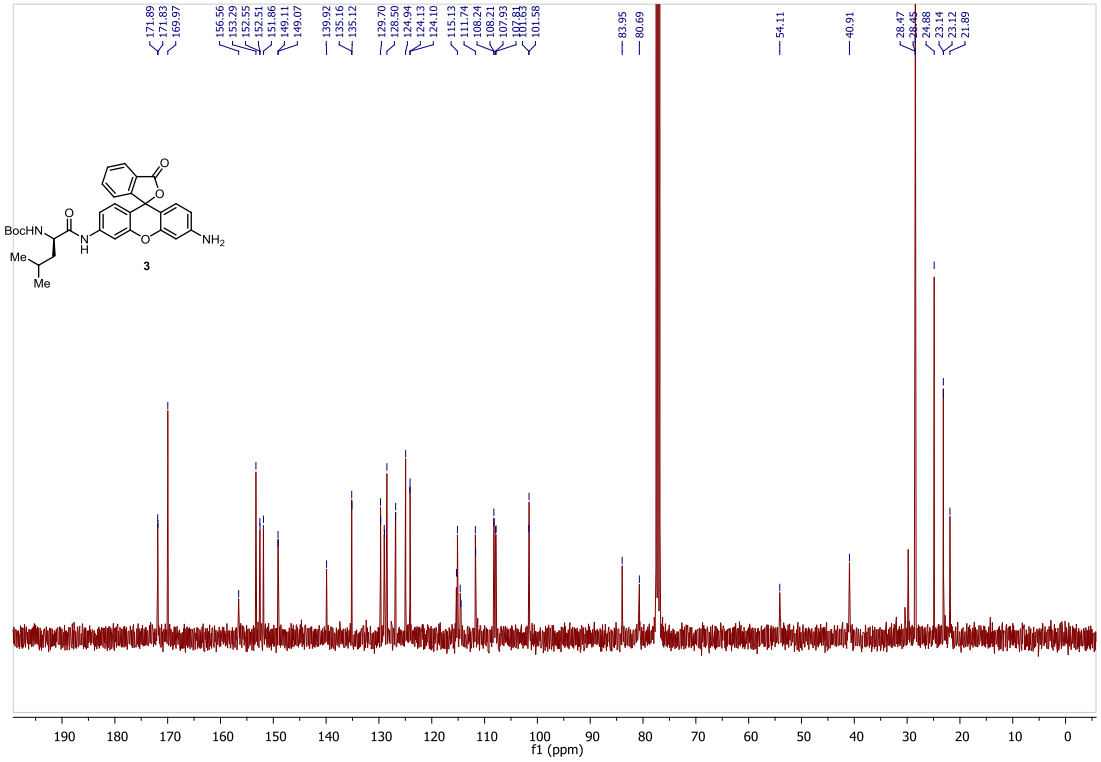
Protocol : This experiment was performed in a co-flow device similar to the design depicted in Fig. S6, but with three connected droplet makers instead of two and a second outlet to split the flow at the end of the device. Three SGAP solutions (0, 0.1 and 0.2 mg/mL) in enzymatic buffer supplemented with pluronic F-127 (final concentration, 0.4 % w/v) were encapsulated with a solution of probe **6** (100 µM in enzymatic buffer) using syringe pumps (Nemesys, aqueous phases: 50 µL/h, oil phases: 100 µL/h). Once the production was stable, the first outlet was directly connected to a reinjection chip using a piece of PTFE tubing (length: 5 cm, internal diameter: 0.33 mm). The time for the droplets to cross the tubing (time of incubation) was assessed to be approximately 30 s. During a mutant enzyme screening experiment, the incubation time should be optimized to cover the desired range of enzymatic activity levels. The substrate fluorescence of the droplets was recorded in the reinjection chip. A control dye was used to monitor the substrate concentration in all three populations (not shown here).

Copy of ^{1}H and ^{13}C spectra :

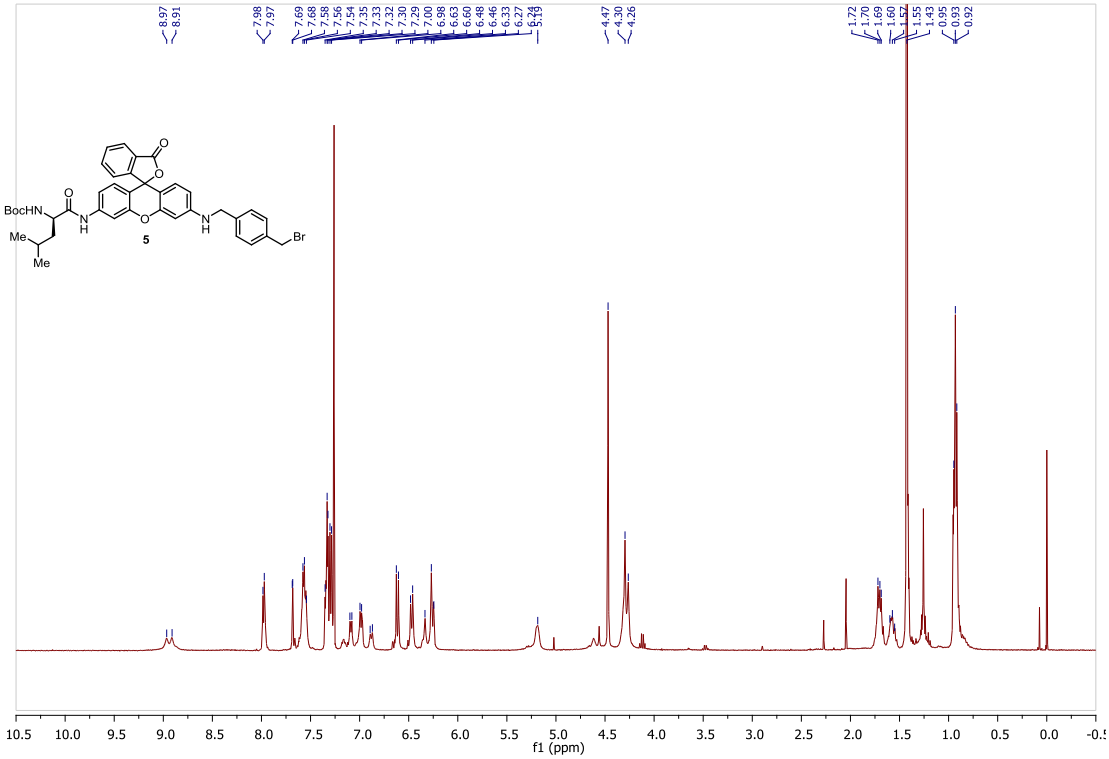
Copy of ^{1}H NMR spectra of **3**:



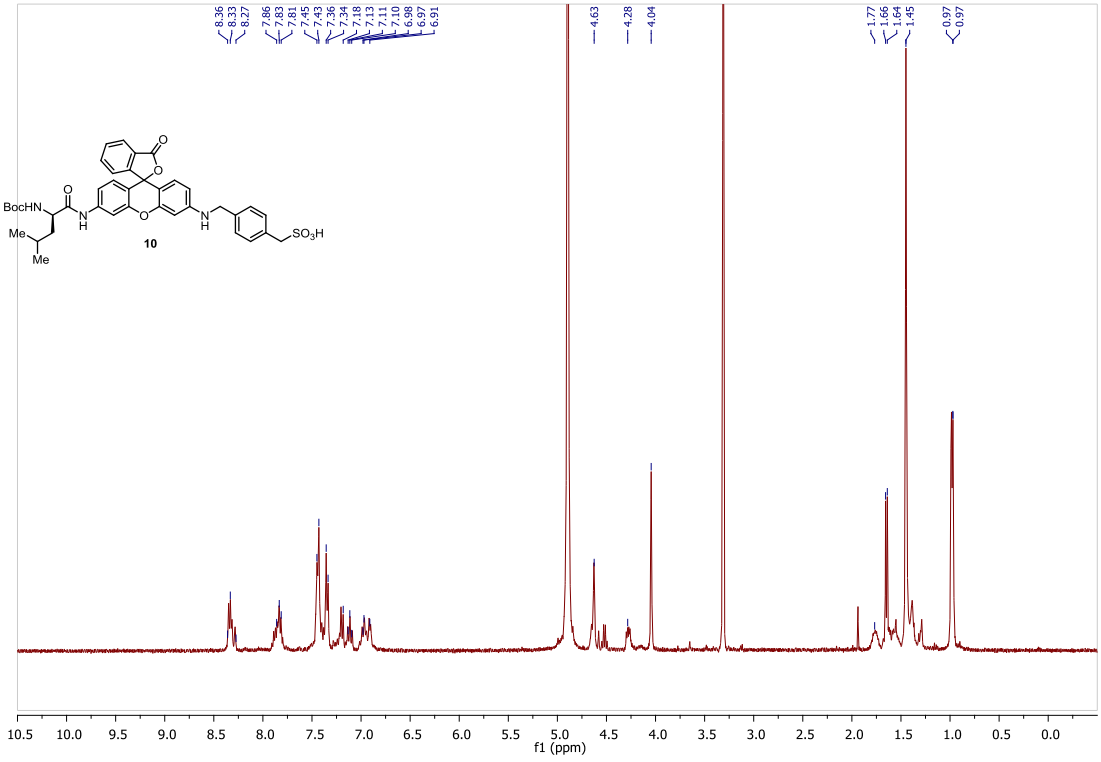
Copy of ^{13}C NMR spectra of **3**:



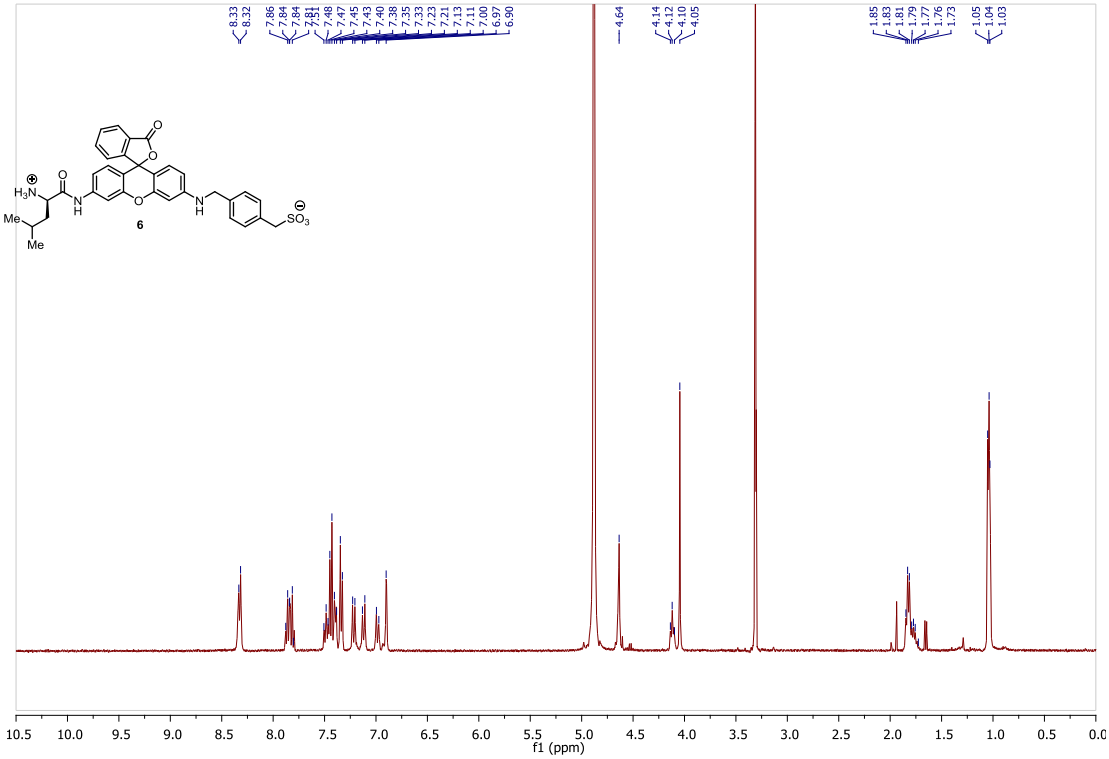
Copy of ^{1}H NMR spectra of **5**:



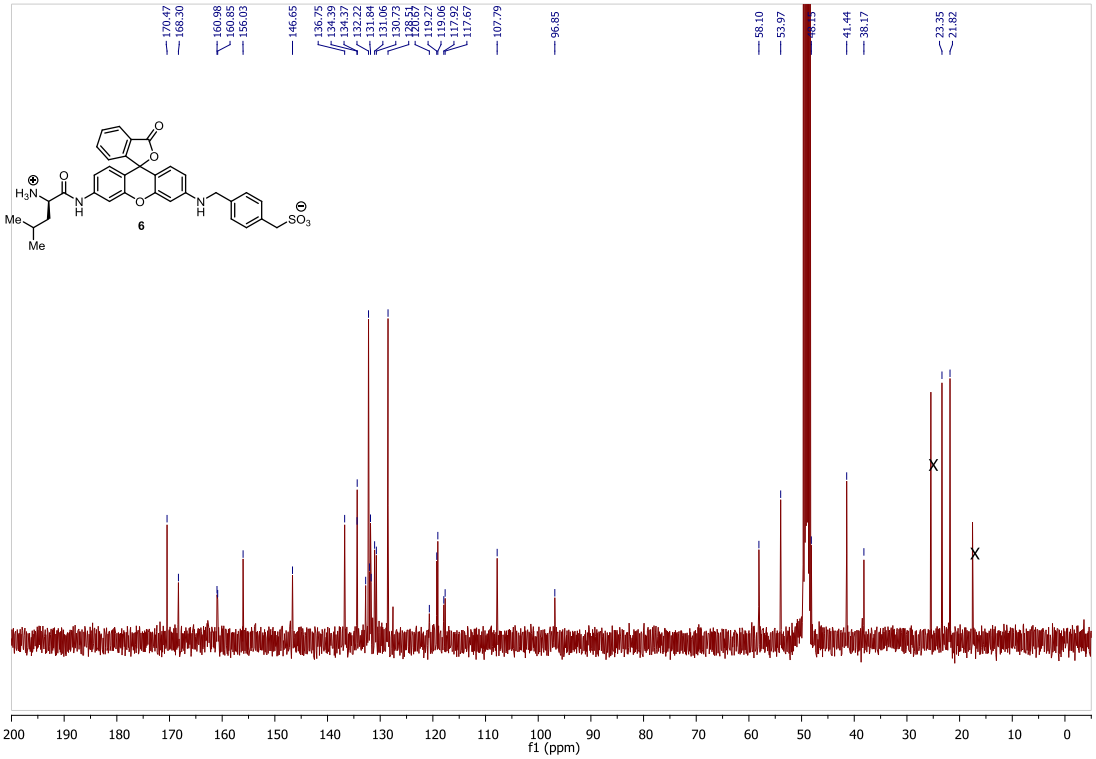
Copy of ^{1}H NMR spectra of **10**:



Copy of ^{1}H NMR spectra of **6**:



Copy of ^{13}C NMR spectra of **6**:



Appendix B

Appendix: Development of an in-vivo microfluidic work-flow for the genotype-phenotype mapping of *Ratus norvegicus* trypsin

B.1 *Ratus norvegicus* trypsin expression in *E. coli*

The different expression vectors used to perform periplasmic expression of either rat trypsin or rat trypsin mCherry fusion protein are presented in figures B.1 and B.2.

B.1.1 Plasmid maps

B.1.2 General comments about the experiments

- The hypotonic buffer is made of the following: 50 mM HEPES, 10 mM $CaCl_2$, pH=8.0.
- The isotonic buffer is made of the following: 50 mM HEPES, 10 mM $CaCl_2$, 100 mM NaCl, p=8.0.
- Kanamycin is used at all time at 50 µg/mL final concentration in *E. coli* cultures transformed with the pet28 plasmid.
- Cells were spinned down at 4000 g during 4 min.
- Glucose is used at 1 % w/v final concentration.
- IPTG is used to 1 µg/mL final concentration.
- Arabinose is used to 0.2 % w/v final concentration.
- For expression purposes, the *E. coli* strain used is BL21-AI with the following genotype: F-ompT hsdSB (rB- mB-) gal dcm araB::T7RNAP-tetA. The expression vector is pet28. The signal peptide is hisJ.
- If not stated otherwise, all enzymatic assay experiments are performed by adding 10 µL of active sample to 90 µL of substrate, in a 96 well plate (flat black bottom). Fluorescence kinetics measurements are performed for 1 h at 37°C in a spectrophotometer. The activity (in RFU/min)

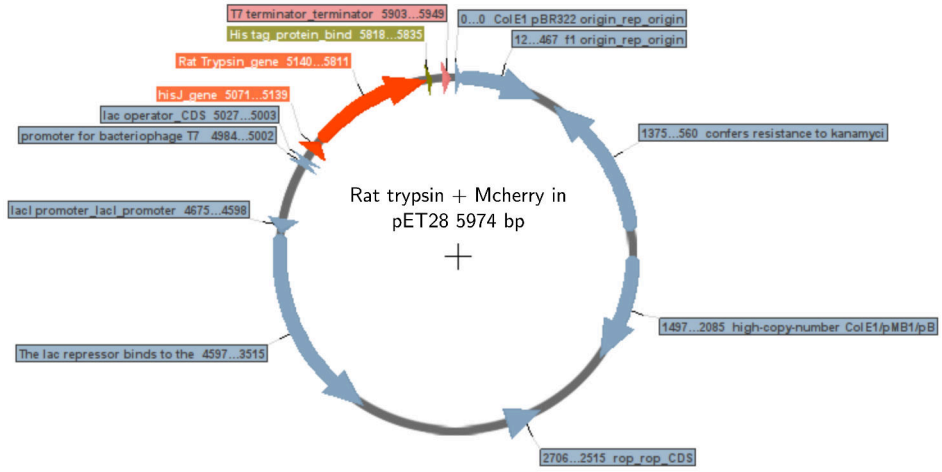


FIGURE B.1: **Rat trypsin sequence in pET28, plasmid map.**

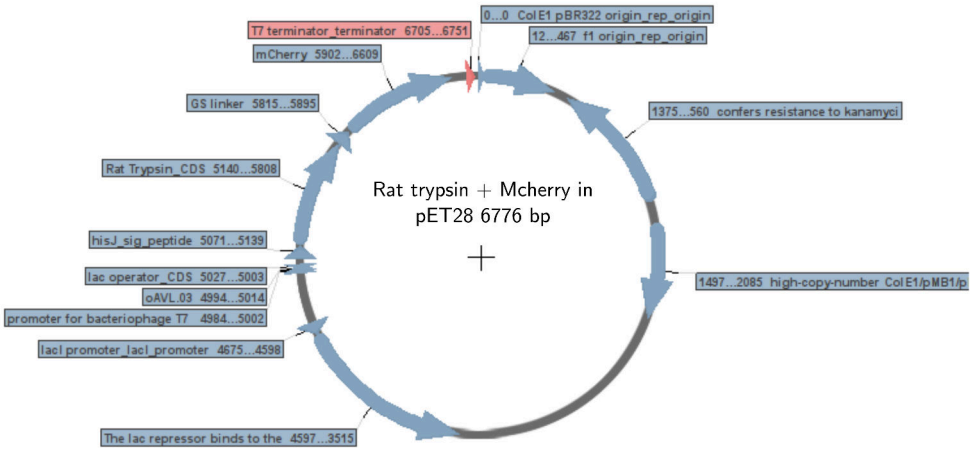


FIGURE B.2: **Rat trypsin mCherry sequence in pET28, plasmid map.**

is defined as the linear slope of the fluorescence vs time graph over the last 20 minutes of kinetics.

B.1.3 Osmotic shock of E.coli cells in hypotonic buffer allows trypsin activity detection in bulk 4.1.1, 4.1.3

Fresh LB medium containing the proper antibiotics and glucose were inoculated from *E. coli* (transformed or not transformed) glycerol stocks overnight and incubated at 37°C, 180 RPM overnight. In the morning the cultures were diluted 100 folds in LB medium containing kanamycin. Once the OD reached 0.5, the cells were induced by adding arabinose and iptg in the culture medium for wild type trypsin expression. After 1.5h of incubation at 37°C, the cultures were treated according to the following protocols.

Conditions	Protocol
Supernatant	50 µL of culture added to 50 µL substrate in isotonic buffer
Isotonic	50 µL of culture spinned down, resuspended in isotonic buffer added to 50 µL of substrate in isotonic buffer
Hypotonic	50 µL culture spinned down, resuspended in hypotonic buffer added to 50 µL of substrate in hypotonic buffer
Flash freezing and Isotonic	50 µL of culture spinned down, resuspended in isotonic buffer, flash frozen in liquid nitrogen and added to 50 µL of substrate in isotonic buffer

The substrate used in this experiment was Benzyloxycarbonyl-Gly-Pro-Arg-7-amino-4-methylcoumarin (Enzo Life Science) to a final concentration of 10 µM ($\lambda_{ex} = 372nm$,$\lambda_{em} = 445nm$). When the rat trypsin was expressed as a fusion protein with mCherry, the mCherry fluorescence was measured with the following wavelengths: ($\lambda_{ex} = 585nm$,$\lambda_{em} = 610nm$). Fluorescence kinetics experiments were then performed in a spectrophotometer.

B.1.4 Lysozyme and sucrose only marginally improve trypsin activity detection 4.1.1

Fresh LB containing the proper antibiotics and glucose were inoculated from *E. coli* glycerols stocks overnight and placed in the incubator at 37°C, 180 RPM overnight. In the morning they were diluted 100 folds in fresh LB medium containing kanamycin. Once OD hit 0.5, the cells where either not induced or induced by adding arabinose and iptg in the culture medium. After 1.5 h of induction, OD was measured and the cells were resuspended in isotonic buffer containing 2.5 µM of coding dye to an OD of 4. 10 µL of the resuspended cells were added to 90 µL of a substrate solution in a 96 wells plate containing the dy647 250 nM and the substrate. The substrate used in this experiment was (Benzyloxycarbonyl-Ala-Arg)2-Rhodamine110 2HCl (Anaspec) to a final concentration of 10 µM ($\lambda_{ex} = 488nm$,$\lambda_{em} = 535$). The tested buffers were the following:

Buffer	Composition
Isotonic	Isotonic buffer
Hypotonic	Hypotonic buffer
Hypotonic + Lyzozyme	2 mg/mL lysozyme in hypotonic buffer
Hypotonic + Lyzozyme + Sucrose	2 mg/mL lysozyme and 500 mM sucrose in hypotonic buffer

Fluorescence kinetics experiments were then performed in a spectrophotometer.

B.1.5 MUGB inhibits trypsin activity but cannot be used as a reporter for its concentration 4.1.2

Fresh LB containing kanamycin and glucose were inoculated from *E. coli* glycerols stocks overnight and placed in the incubator at 37°C, 180 RPM overnight. In the morning they were diluted 100 folds in fresh LB medium containing kanamycin. Once OD reached 0.5, the cells were not induced or induced by adding arabinose and iptg in the culture medium. After 1.5 h of induction, OD was measured and the cells were resuspended in isotonic buffer to an OD of 4. 10 µL of the resuspended cells were added to 90 µL of a substrate solution in a 96 wells plate containing the substrate (Benzyloxycarbonyl-Ala-Arg)2-Rhodamine110 2HCl (Anaspec) ($\lambda_{ex} = 488nm$, $\lambda_{em} = 535$) to a final concentration of 10 µM. After 1 h of enzymatic reaction, 4-methylumbelliferyl p-guanidobenzoate (MUGB) was added into the wells to the following final concentrations: 0 (negative control), 100, 1000, 2500, 5000 and 10000 nM. Immediately after adding the MUGB to the assay mixtures, the fluorescence kinetics was measured in the coumarin ($\lambda_{ex} =$,$\lambda_{em} =$) and the rhodamine 110 channels. The trypsin activity (in RFU/min) and the rate of MUGB hydrolysis are defined as the linear slope of the fluorescences vs time graph over the 10 minutes following MUGB addition.

B.1.6 Development of a non leaking substrate based on the fret pair EDANS-Dabcyl 4.1.4

In 90 µL of fresh M9 medium were added 5 µL of EDANS DABCYL substrate (approximate final concentration: 50 µM) and 5 µL of purified trypsin enzyme (trypsin from bovine pancreas, Sigma Aldrich, 30 µg/mL final concentration). In the case of the negative control sample, 5 µL of water were added instead of the enzyme. Fluorescence kinetics measurements were performed at the following wavelengths: $\lambda_{\text{ex}} = 352nm$ and $\lambda_{\text{em}} = 510nm$. Once the positive sample fluorescence achieved saturation, the positive and negative samples were recovered in eppendorf tubes and added coding dye to the respective final concentrations of 3 µM and 1 µM. Both samples were vortexed and emulsified into 20 pL droplets separately but simultaneously (MFCS, Fluigent) with the following pressures:

Phase	Pressure
Oil	1200 mBar
Water	450 mBar

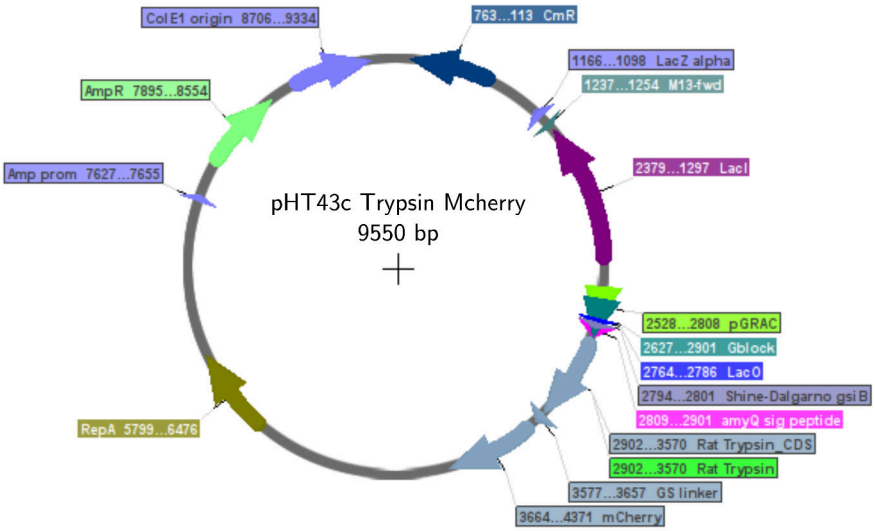


FIGURE B.3: **pHT43c rat trypsin mCherry, plasmid map.**

Immediately after collection and after 21 h of incubation at room temperature, the emulsion was reinjected to measure its fluorescence distribution in the activity and coding dye fluorescence channels using the UV (375 nm) and red (648 nm) lasers. We used the following pressures:

Phase	Pressure
Oil	600 mBar
Water	600 mBar

B.1.7 Rat trypsin in *B. subtilis*: plasmid maps

B.2 *Ratus norvegicus* trypsin expression in *B. subtilis*

B.2.1 General comments about the experiments

- The strain used in all those experiments is WB8000N, purchased from Mobitech.
- The expression vector is pHT43, also purchased from Mobitech.
- In all experiments, if mentionned, the final concentrations of the following compounds are: chloramphenicol : 50 µg/mL, glucose: 1% w/v, IPTG: 1mM.
- The substrate used in these experiments was (Benzyloxycarbonyl-Ala-Arg)2-Rhodamine110 2HCl (Anaspec) to a final concentration of 10 µM ($\lambda_{ex} = 488nm$, $\lambda_{em} = 535$).
- The mCherry fluorescence channel is ($\lambda_{ex} = 585nm$, $\lambda_{em} = 610$)

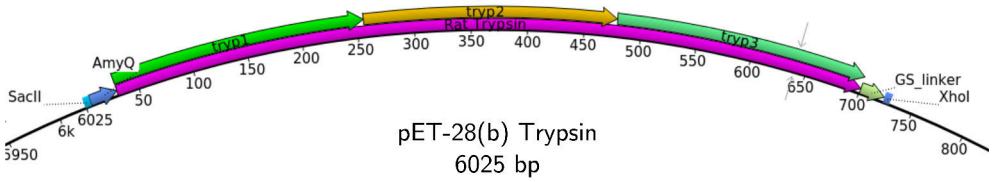


FIGURE B.4: **pET28(b) rat trypsin mutagenesis partial plasmid map**. The trypsin sequence has been cloned inside the pET28 plasmid along with the end of the amyQ signal peptide sequence and the beginning of the GS flexible linker. As a result, the designed mutagenic primers are compatible with both expression and mutagenic plasmids. Also, it is possible to subclone the mutagenic product directly into the expression vector without amplification to add compatible restriction sites.

B.2.2 *B. subtilis* transformation protocol

Before transformation, 10xPC solution was prepared (K_2HPO_4: 112 g/L, KH_2PO_4: 48 g/L, Sodium citrate $2H_2O$: 6.8 g/L) and autoclaved. The day before the transformation, *B. subtilis* glycerol stocks were streaked on a 2xYT medium agar plate. The plate was incubated at 37°C overnight. The day of the transformation, MD medium was prepared (1xPC, 2% w/v glucose, 0.05 mg/mL L-tryptophan, 0.011 mg/mL ferric ammonium citrate, 2.5 mg/mL DL aspartic acid potassium salt, 3 µM $MgSO_4$), its pH is corrected to 7.0 with KOH and filtered (0.2 µm). At the same time, casa amino acid was prepared (20% w/v) and filter sterilized (0.2 µm).

In an empty 100 mL erlenmeyer, 5 mL and 25 µL of MD and casa amino acid solutions were added. A *B. subtilis* colony was added to the erlenmyer which was set to incubate at 37°C, 200 RPM. The rest of the MD medium was kept at 37°C. When the OD reached 1, 5 mL of warm MD medium were added to the erlenmeyer which was incubated again at 37°C. 10 mL culture tubes were warmed in the incubator. After 1 hour of incubation, 400 µL of culture were added to each of the prewarmed culture tube. Plasmid DNA was added to each one of the tube to a final concentration of 1 µg/mL. The tubes were incubated again for 20 min and then 12.5 µL of 20% casa amino acid were added to each tube. The tubes were incubated for 1 h more at 37°C and then spread on plates containing glucose 1 % and chloramphenicol (50 µg/mL) (100 µL per plate maximum). This protocol led to approximately 1000 colonies per µg of plasmid DNA (pHT43 plasmid).

B.2.3 Induction protocol

: 2xYT agar plates supplemented with chloramphenicol and glucose are streaked and left to incubate overnight at 37°C. In the morning, one colony is used to inoculate 5 mL of fresh 2xYT supplemented with chloramphenicol and glucose, in a 50 mL falcon. After 4 h of growth (OD should be 3 to 4) at 37°C, 180 RPM, the fresh culture is diluted to an OD of 0.15 in 2xYT supplemented with chloramphenicol. Once the OD hits 0.7, IPTG is added to the medium and the culture is incubated overnight at 37°C, 180 RPM.

B.2.4 Droplet induction protocol

: 2xYT agar plates supplemented with chloramphenicol and glucose are streaked and left to incubate overnight at 37°C. In the morning, one colony is used to inoculate 5 mL of fresh 2xYT supplemented with chloramphenicol and glucose, in a 50 mL falcon. After 4 h of growth (OD should be 3 to 4) at 37°C, 180 RPM, the fresh culture is diluted to an OD of 0.15 in 2xYT supplemented with chloramphenicol. Once the OD hits 0.7, IPTG is added to the medium. Coding dye is added to the culture which is briefly vortexed. The culture is then encapsulated in 20 pL droplets, using the pressure pumps with the following settings:

Phase	Pressure
Oil + 2% surfactant	1000 mBar
Culture	1100 mBar

The emulsion is collected into an open plastic syringe. Once generation is over, the syringe containing the emulsion is incubated at 37°overnight. If the emulsion is to be shaken, the syringe is placed into a 50 mL falcon tube containing tissue soaked with water. The falcon is incubated at 37°C, 800 RPM in a thermomixer overnight.

B.2.5 Bulk assay protocol

: Following induction, as detailed in B.2.3, 90 µL of culture (or supernatant, or pellet) are added to the wells of a 96 well plates. 10 µL of trypsin substrate (final concentration: 10 µM) are added to each well. Fluorescence kinetics is measured at 37°C in a spectrophotometer in the rhodamine and the mCherry fluorescence channels. Activity is defined as the initial slope of the fluorescence vs time graph (RFU/min).

B.2.6 Droplet assay protocol

: Following induction in droplets B.2.4, the emulsion is transfered into a collector for reinjection. A 2 pL substrate (trypsin substrate, 1 mM) is generated and collected. Both emulsion are reinjected into a 2-20 pL electrofusion device with delay lines using the following pressure settings:

Phase	Pressure
Oil + 2% surfactant	600 mBar
2 pL droplets (substrate)	600 mBar
20 pL droplets (cells)	600 mBar
Sorting oil 1 and 2	300 mBar

The pressure settings are corrected to achieve one to one pairing between the droplets containing the cells and the droplets containing the substrate. A electrical function generator is connected to the electrodes of the chip through a voltage amplifier (x 1000) with the following settings: 5 kHz, 500 mVpp. The voltage settings are corrected to achieve fusion without micelles. At the end of the delay lines, the droplets fluorescence is measured using the optical setup already described in the material and methods (2.1.3).

B.2.7 The rat trypsin - mCherry fusion protein is secreted by WB800N as a full protein in the supernatant 4.2.1

B. subtilis glycerol stocks of cells (WB800N and D168) transformed with either the wild type or S195A rat trypsin were used to perform induction experiments as described in B.2.3. After induction, the cultures were spinned down. 15 µL of each supernatants were denatured in Laemmli denaturing buffer supplemented with β-mercaptoethanol (95°C, 5 min). The samples, along with a protein weight marker (Ladder PageRuler 26619) are loaded and ran on a SDS-Page gel with the following current, voltage and buffer: 25 mA, 100 V, 20X MOPS SDS. The gel was fixed in fixing solution (40% ethanol v/v, 10% acetic acid v/v) for 15 min. The gel was then washed in milliQ water and stained in Coomassie blue overnight. In the morning, the gel was washed in mQ water 3 times. Western blotting was performed on a nitrocellulose membrase according to the XCell II Blot Module documentation (Biorad). The membrane was immerged in a blocking solution (2% powder milk, 0.1% Tween-20 in PBS 1X) for 1 h. An anti-mCherry antibody (polyclonal, rabbit, ab167453, AbCam) and a conjugated fluorescent anti-Rabbit (Anti-Rabbit IgG, 711-545-152, Jackson Immuno Research) antibodies were diluted 1000 times in blocking solutions. The membrane was incubated 1 h in the anti-Mcherry solution, rinsed in 0.1 % Tween-20 in PBS and incubated 1 h in the conjugated antibody solution. The membrane was rinsed again and imaged in a fluorescence chamber in the Alexa 488 fluorescence channel (exposure time: 5 s).

B.2.8 Measuring the catalytic efficiency of trypsin variants in droplets 4.2.1

Three emulsions with either cells containing C191A, S195A or WT trypsin subtilis were prepared simultaneously but separately for induction in droplets according to B.2.4 (the emulsions were mixed upon production). In the morning, the resulting emulsion was displaced from the open collector used for induction to a 0.2 mL PCR tube in order to make it ready for reinjection. In parallel, a 2 pL substrate emulsion was prepared from a 1 mM rodamine 110 based trypsin substrate in water, using Fluigent pressure pumps (pressures: oil: 1000 mbar, aqueous phase: 1100 mBar). The emulsion was collected into a 0.2 mL PCR tube, ready for the reinjection. Both emulsions, as well as oil phase were reinjected into the chip and pressures turned on (300 mBar each). The 2 and 20 pL emulsions were paired and electrodes were connected to deliver 600 Vpp at 30 kHz in the fusion chamber. After 15 min of incubation on chip in delay lines, the droplets reached the sorting device (pressures at 200 mbar for each oil inputs) where their fluorescence was measured in the rhodamine 110 ($\lambda_{ex} = 488nm, \lambda_{em} = [505 : 545]$) and the mCherry ($\lambda_{ex} = 561nm, \lambda_{em} = [570, 616]$) fluorescence channels.

B.3 Towards a library of all rat trypsin single point mutants

B.3.1 Designing the mutagenic primers 4.3.2

As an example, let us consider the following "wild type" sequence:

ATTGTGGGTGGTTATACGTG

To mutate the isoleucine residue coded by **ATT**, primers were designed and ordered to begin with the following codons (using the IUPAC nucleotide code):

- **TTT**: 1 codon (F)
- **NRT**: (A or T or C or G) and (A or G) and T (8 codons coding for N, S, Y, C, H, R, D, G).
- **TGG**: 1 codon (W)
- **VHG**: (A or C or G) or (A or C or T) and G: (9 codons coding for K,T, M, Q, P, L, E, A, V)

The complete list of residues and their associated ambiguous codons is the following:

Wild type residue	Ambiguous mutant codons to perform saturated mutagenesis
A	WKK, MMG, GDA, NAT
L	RTG, NRT, TGG, VMA, WTT
R	KGG, WKT, VHG, NAT
K	SAW, BGG, VYG, WDC
N	BGG, BAT, WKT, VHG
M	NDT, TGG, VMA
D	KGG, VHG, HRT, WTT
F	BRT, TGG, VHG, ADC
C	TGG, SVT, VWG, WHC
P	RMA, WKG, CAS, DDT
Q	NHT, RRA, TGS, ATG
S	WWC, TGS, SRT, VHG
E	GBC, WKK, MMG, NAT
T	TGG, SVT, VWG, WDC
G	WGS, VHG, NAT, WTT
W	NRT, VHG, WTT
H	VMA, WKG, DDT
Y	TGS, VRT, VHG, WTT
I	TTT, NRT, TGG, VHG
V	NRT, TGG, VMA, WTK

My script processed the rest of the sequence (GTGGGTGGTTATACGTG in the example) to generate primers with the following characteristics:

- The primers should finish by 'C' or 'G'
- The melting temperature of the primers should be as close as possible to a target value (here 54°C).
- The primers should have a length greater than X nucleotides.

The detailed procedure of the script was the following:

- Variables: l: starting length of the oligonucleotides, T_t: target melting temperature of the nucleotides.
- Determine the melting temperature of the oligos made of the l next nucleotides starting after the codon to mutate.
- Calculate $\Delta T = T_t - T_{oligo}$
- By adding or deleting nucleotides on 3', determine the sequence of:
  - The highest T_m oligonucleotide among those showing $\Delta T > 0$ and finishing by 'C' or 'G'.
  - The lowest T_m oligonucleotide among those showing $\Delta T < 0$ and finishing by 'C' or 'G'.
  - Compare the two of them and keep in a separate variable the sequence showing the lowest $\|\Delta T\|$.

The same process was repeated to design the reverse oligonucleotides, on the indirect strand. Finally, the melting temperature of the primers (constant part of the forward and reverse) was determined on the NEB website to ensure the sequences had the desired properties. For each position, the constant part of the forward primer was concatenated to each of the 3 to 5 ambiguous codons that we had designed, leaving us with as many forward primers to order. The primers were then ordered at Integrated DNA Technologies which mixed the nucleotides during synthesis according to the ambiguous nucleotides incorporated in our primers sequences. For each position in the trypsin peptide chain, I mixed the 3 to 5 primers I received so that all the codons were represented in equimolar ratio.

B.3.2 Performing saturated mutagenesis on the rat trypsin protein 4.3.3

The Q5 PCR reaction mix (25 µL) were prepared according to the NEB protocol with GC enhancer with the following final concentration: 0.2 mM of dNTP's, 0.5 µM of each primers (reverse or mix of forward primers) and 0.5 ng of template DNA (see B.4 for the plasmid map). The following thermocycling programs were used:

Classical PCR:

Step	Temperature °C	Time
1	98	30 s
2	98	10 s
3	Tm	30 s
4	72	4 min s
5	72	5 min
5	4	For ever min

Steps 2 to 4 are repeated 30 times.

"Touch down" PCR: In this case, the amount of template is 5 ng per 25 µL PCR mix.

Step	Temperature °C	Time
1	98	30 s
2	98	10 s
3	61 - 0.5 at each cycle for the 20 first cycle, then 50 for the last 10 cycles	30 s
4	72	4 min s
5	72	5 min
5	4	For ever min

Steps 2 to 4 are repeated 30 times.

B.3.3 Library preparation 4.3.3

The following protocol was followed for all the positions except for the positions for which the primers pairs did not lead to satisfying PCR product:

14, 27, 37, 49, 51, 52, 53, 60, 100, 101, 108, 111, 112, 113, 117, 123, 139, 142, 143, 146, 147, 182, 183, 186, 187 and 197.

After the PCR, each product was added 0.5 µL of dpnI restriction enzyme to digest the template. Each reaction was incubated at 37°C for 1 h. DpnI was inactivated at 80°for 20 min. Samples were cooled down and loaded on a low melting point agarose gel where the correct size bands were purified using Macherey Nagel PCR clean up columns and eluted in water. Except for 26 positions for which mutagenic PCR failed, the concentration of each product was quantified on nanodrop and three groups of PCR products were mixed in equimolar ratio to a final concentration of 1 ng/µL: positions 1 to 74 included, 75 to 149 included, 150 to 223 included. Each sublibrary was concentrated 5 times under vaccum to an approximate concentration of 5 ng/µL. 50 ng of DNA were then taken, phosphorylated using T4 polynucleotide kinase (PNK, NEB) following the manufacturer's instructions. The PNK was inactivated at 72°C during 10 min. 5 µL (approximately 12 ng) of this sample was circularized using electroligase (NEB). The ligase was inactivated (65°C, 15 min) and dyalised on a membrane for 30 min (MF-Millipore Membrane, VSWP02500), in a closed petri dish filled

with milliQ water. The ligation reaction was used to transform highly competent *E. coli* cells (MegaX DH10B T1R Electrocomp Cells, 1 µL per transformation) according to the manufacturer's protocol. The cells were spread on selective medium (kanamycin, glucose 1%) and scraped from the plate the following day, approximately 2 millions of colonies were recovered each time per sublibrary and stored in a glycerol stock whose OD was previously measured.

To prepare for deep sequencing, DNA was first extracted from the glycerol stocks using miniprep kits (1 mL at OD 6) and the following primers were used to perform a PCR (1 ng of template, 18 cycles with Q5 polymerase, T_a=58°C).

Sublibrary 1	forMseqv1tryp1for	ACACTCTTTCCCTACACGACGCTCTTCC-GATCTNNNNNGTTTAGCAGCGCGAC
Sublibrary 1	revMseqv1tryp1rev	GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTNNNNNGCTCAGCTTAATCAGCATG
Sublibrary 2	forMseqv1tryp2for	ACACTCTTTCCCTACACGACGCTCTTCC-GATCTNNNNNAACGAACAGTTTGTGAACG
Sublibrary 2	revMseqv1tryp2rev	GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTNNNNNGAAAGCCCACACACAC
Sublibrary 3	forMseqv1tryp3for	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNT-GCAGTGCCTGGATG
Sublibrary 3	revMseqv1tryp3rev	GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTNNNNNTCAGTGGTGGTGGTG

For each sublibrary, the product size was checked on a gel, and then diluted 50 folds. The new sample was used to perform a new PCR (1 µL of template, 25 cycles with Q5 polymerase, T_a=65°C).

Sublibrary 1	D501	AATGATACGGCGACCACCGAGATCTACACtatagcctACACTCTTTCCCTACACGAC
Sublibrary 1	D701	CAAGCAGAAGACGGCATACGAGATcgagtaatGTGACTGGAGTTCAGACGTG
Sublibrary 2	D502	AATGATACGGCGACCACCGAGATCTACACatagaggcACACTCTTTCCCTACACGAC
Sublibrary 2	D702	CAAGCAGAAGACGGCATACGAGATtctccggaGTGACTGGAGTTCAGACGTG
Sublibrary 3	D503	AATGATACGGCGACCACCGAGATCTACACcctatcctACACTCTTTCCCTACACGAC
Sublibrary 3	D503	CAAGCAGAAGACGGCATACGAGATaatgagcgGTGACTGGAGTTCAGACGTG

B.3.4 Redesigning part of the mutagenic primers

Oligos were designed again for the following trypsin residues (each residue being referenced by a number from 1 to 223):

14, 19, 21, 27, 30, 35, 36, 37, 38, 42, 49, 51, 52, 53, 92, 93, 95, 96, 97, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 123, 124, 125, 130, 139, 140, 142, 143, 146, 147, 148, 150, 157, 168, 169, 173, 177, 181, 182, 183, 184, 186, 187, 190, 197, 202, 203, 204, 205, 208, 223.

The first three nucleotides of the forward oligos were designed as in the previous part B.3.1.

My script processed the rest of the sequence (GTGGGTGGTTATACGTG) to generate primers with the following characteristics:

- The primers should finish by 'C' or 'G'
- The primers should have a length greater than 14 nucleotides.

The detailed procedure of the script was the following:

- Variables: l: minimum length of the oligonucleotides.
- Initialize the forward and reverse oligos with the sequence made of the $l + i$ nucleotides after the codon to mutate, i being the length of the extension in the 3' direction so that the oligos finish with 'G' or 'C' (respectively on the direct and the indirect strand).
- Determine the T_m of the oligos and determine which oligo has the lowest T_m.
- By adding nucleotides on 3' of the lowest T_m oligo, determine the sequence of:
  - The highest T_m oligo with a T_m lower than the T_m of the other oligo and finishing by 'C' or 'G'.
  - The lowest T_m oligo with a T_m higher than the T_m of the other oligo and finishing by 'C' or 'G'.
  - Compute the two differences of $abs(T_{m,forward} - T_{m,reverse})$ between. Keep the oligos showing the lowest of them two.

Finally, the melting temperature of the oligos was tested on the NEB website to ensure the sequences had the desired properties.

The library was prepared according to the protocol detailed in B.3.3. For the preparation to deep sequencing, we used different primers. DNA was first extracted from the glycerol stocks using miniprep kits (1 mL at OD 6) and the following primers (forward and reverser) were used to perform a PCR (1 ng of template, 18 cycles with Q5 polymerase, T_a=71°C).

Sublibrary 1	forMseqv1tryp1forv2	ACACTCTTTCCCTACACGACGCTCTTCC-GATCTNNNNNGCGGTTTGCCGATTACAAAAAC-TAGTGCT
Sublibrary 1	revMseqv1tryp1revv2	GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTNNNNNGTTGTTCAGGGTCTTGCGATCGAA
Sublibrary 2	forMseqv1tryp2forv2	ACACTCTTTCCCTACACGACGCTCTTCC-GATCTNNNNNGCGGCGAAAATTATCAAGCATC-CGAAC
Sublibrary 2	revMseqv1tryp2revv2	GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTNNNNNGGTAATTTTGCCCGGATAGCTGGC
Sublibrary 3	forMseqv1tryp3forv2	ACACTCTTTCCCTACACGACGCTCTTCC-GATCTNNNNNCCGCAGGCGGATTGCGAA
Sublibrary 3	revMseqv1tryp3revv2	GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTNNNNNACCGGTCACCACTTGAGGATCC

For each sublibrary, the product size was checked on a gel, and then diluted 50 folds. The new sample was used to perform a new PCR (1 µL of template, 25 cycles with Q5 polymerase, T_a=65°C).

Sublibrary 1	D504	AATGATACGGCGACCACCGAGATCTACACggctctgaACACTCTTTCCCTACACGAC
Sublibrary 1	D704	CAAGCAGAAGACGGCATACGAGATggaatctcGTGACTGGAGTTCAGACGTG
Sublibrary 2	D505	AATGATACGGCGACCACCGAGATCTACACaggcgaagACACTCTTTCCCTACACGAC
Sublibrary 2	D705	CAAGCAGAAGACGGCATACGAGATttctgaatGTGACTGGAGTTCAGACGTG
Sublibrary 3	D506	AATGATACGGCGACCACCGAGATCTACACtaatcttaACACTCTTTCCCTACACGAC
Sublibrary 3	D706	CAAGCAGAAGACGGCATACGAGATacgaattcGTGACTGGAGTTCAGACGTG

B.3.5 Sequencing data analysis

The forward and reverse sequencing data were aligned to the trypsin sequence using the software Bowtie2-2.2.9:

```
bowtie2 -x trypsinindex -1 D1_S1_L001_R1_001.fastq.gz
-2 D1_S1_L001_R2_001.fastq.gz
--maxins 600 --minins 450 -S tryp30D1P.sam
--rdg 10,6 --rfg 10,6

bowtie2-build trypsin.fa trypsinindex
```

The reads were indexed using samtools in order to visualize the global alignment using the software Tablet (The James Hutton Institute).

Then, the reads were aligned by name and recorded as a binary file (BAM). Using a python script I wrote, the reads where filtered in order to only keep paired, mapped reads whose cigarstring only bear matches or indels. Pair of reads not covering the full region of trypsin under investigation were also discarded.

Following this, each pair of reads was processed in order to build a sparse vector of all possible mutants codons at all positions (size $N \times 4^4$ where N is the number of positions considered). A mutation is called if its Phred quality score is higher than 30. If the forward and the reverse reads are in agreement about a mutation it is keps. Otherwise, it is discarded. A sparse matrix is built from the concatenation of those different vectors. It is used to perform statistical analysis on the sequencing data.

First deep sequencing run:

- Positions 1 to 74 included, 232372 reads, 58 % of pair of reads displaying one mutation.
- Positions 75 to 149 included, 172130 reads, 64 % of pair reads displaying one mutation.
- Positions 150 to 223 included, 171638 reads, 62 % of pair of reads displaying one mutation.

Second deep sequencing run:

- Positions 1 to 74 included, 114865 reads, 57% of pair of reads displaying one mutation.

- Positions 75 to 149 included, 216120 reads, 64 % of pair of reads displaying one mutation.
- Positions 150 to 223 included, 171638 reads, 62 % of pair of reads displaying one mutation.

Bibliography

[1] W. Johannsen. “The Genotype Conception of Heredity”. In: *The American Naturalist* 45.531 (1911), pp. 129–159. ISSN: 0003-0147.

[2] T. H. Morgan. “The Theory of the Gene”. In: *The American Naturalist* 51.609 (1917), pp. 513–544. ISSN: 0003-0147.

[3] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES”. In: *The Journal of Experimental Medicine* 79.2 (Feb. 1944), pp. 137–158. ISSN: 0022-1007.

[4] Maclyn McCarty and Oswald T. Avery. “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES”. In: *The Journal of Experimental Medicine* 83.2 (Jan. 1946), pp. 89–96. ISSN: 0022-1007.

[5] Maclyn McCarty and Oswald T. Avery. “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES”. In: *The Journal of Experimental Medicine* 83.2 (Jan. 1946), pp. 97–104. ISSN: 0022-1007.

[6] J. D. Watson and F. H. Crick. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid”. eng. In: *Nature* 171.4356 (Apr. 1953), pp. 737–738. ISSN: 0028-0836.

[7] Marshall Nirenberg. “Historical Review: Deciphering the Genetic Code – a Personal Account”. In: *Trends in Biochemical Sciences* 29.1 (Jan. 2004), pp. 46–54. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2003.11.009.

[8] J. C. Kendrew et al. “A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis”. en. In: *Nature* 181.4610 (Mar. 1958), pp. 662–666. ISSN: 0028-0836. DOI: 10.1038/181662a0.

[9] Ken A. Dill and Justin L. MacCallum. “The Protein-Folding Problem, 50 Years On”. en. In: *Science* 338.6110 (Nov. 2012), pp. 1042–1046. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1219021.

[10] S Wright. “The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution.” In: ((1932)), pp. 1, 356–366.

[11] Sergey Gavrilets. “Evolution and Speciation on Holey Adaptive Landscapes”. In: *Trends in Ecology & Evolution* 12.8 (Aug. 1997), pp. 307–312. ISSN: 0169-5347. DOI: 10.1016/S0169-5347(97)01098-7.

[12] J. Arjan G. M. de Visser and Joachim Krug. “Empirical Fitness Landscapes and the Predictability of Evolution”. en. In: *Nature Reviews Genetics* 15.7 (July 2014), pp. 480–490. ISSN: 1471-0056. DOI: 10.1038/nrg3744.

[13] Jianzhi Zhang. "Protein-Length Distributions for the Three Domains of Life". In: *Trends in Genetics* 16.3 (Mar. 2000), pp. 107–109. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(99)01922-8.

[14] William R Pearson and Michael L Sierk. "The Limits of Protein Sequence Comparison?" In: *Current opinion in structural biology* 15.3 (June 2005), pp. 254–260. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2005.05.005.

[15] Joram PIATIGORSKY and Joram Piatigorsky. *Gene Sharing and Evolution: The Diversity of Protein Functions*. en. Google-Books-ID: _4YaeCobdakC. Harvard University Press, June 2009. ISBN: 978-0-674-04212-4.

[16] Anil Kumar and Nikita Chordia. "Bacterial Resistance Against Antibiotics". en. In: *Drug Resistance in Bacteria, Fungi, Malaria, and Cancer*. Ed. by Gunjan Arora, Andaleeb Sajid, and Vipin Chandra Kalia. Springer International Publishing, 2017, pp. 171–192. ISBN: 978-3-319-48682-6 978-3-319-48683-3. DOI: 10.1007/978-3-319-48683-3_7.

[17] X. Li and J. C. Phillips. "Prediction (Early Recognition) of Emerging Flu Strain Clusters". In: *Physica A: Statistical Mechanics and its Applications* 479 (Aug. 2017), pp. 371–378. ISSN: 0378-4371. DOI: 10.1016/j.physa.2017.02.073.

[18] Elizabeth T. Cirulli and David B. Goldstein. "Uncovering the Roles of Rare Variants in Common Disease through Whole-Genome Sequencing". en. In: *Nature Reviews Genetics* 11.6 (Jan. 2010), pp. 415–425. ISSN: 1471-0056. DOI: 10.1038/nrg2779.

[19] Joanne L. Porter, Rukhairul A. Rusli, and David L. Ollis. "Directed Evolution of Enzymes for Industrial Biocatalysis". en. In: *ChemBioChem* 17.3 (Feb. 2016), pp. 197–203. ISSN: 1439-7633. DOI: 10.1002/cbic.201500280.

[20] Richard A. Lerner. "Combinatorial Antibody Libraries: New Advances, New Immunological Insights". en. In: *Nature Reviews Immunology* 16.8 (Aug. 2016), pp. 498–508. ISSN: 1474-1733. DOI: 10.1038/nri.2016.67.

[21] Hajo Kries, Rebecca Blomberg, and Donald Hilvert. "De Novo Enzymes by Computational Design". In: *Current Opinion in Chemical Biology*. Bioinorganic Chemistry • Biocatalysis and Biotransformation 17.2 (Apr. 2013), pp. 221–228. ISSN: 1367-5931. DOI: 10.1016/j.cbpa.2013.02.012.

[22] Michael S. Packer and David R. Liu. "Methods for the Directed Evolution of Proteins". en. In: *Nature Reviews Genetics* 16.7 (July 2015), pp. 379–394. ISSN: 1471-0056. DOI: 10.1038/nrg3927.

[23] Nobuhiko Tokuriki et al. "How Protein Stability and New Functions Trade Off". In: *PLOS Computational Biology* 4.2 (Feb. 2008), e1000002. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000002.

[24] Xiaojun Wang, George Minasov, and Brian K. Shoichet. "Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-Offs". In: *Journal of Molecular Biology* 320.1 (June 2002), pp. 85–95. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(02)00400-X.

[25] Joseph D. Bryngelson et al. "Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis". en. In: *Proteins: Structure, Function, and Bioinformatics* 21.3 (Mar. 1995), pp. 167–195. ISSN: 1097-0134. DOI: 10.1002/prot.340210302.

[26] Martin Karplus and J. Andrew McCammon. "Molecular Dynamics Simulations of Biomolecules". en. In: *Nature Structural & Molecular Biology* 9.9 (Sept. 2002), pp. 646–652. ISSN: 1072-8368. DOI: 10.1038/nsb0902-646.

[27] Yves Dehouck et al. "PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality". In: *BMC Bioinformatics* 12 (2011), p. 151. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-151.

[28] Joost Schymkowitz et al. "The FoldX Web Server: An Online Force Field". In: *Nucleic Acids Research* 33.suppl_2 (July 2005), W382–W388. ISSN: 0305-1048. DOI: 10.1093/nar/gki387.

[29] Sofia Khan and Mauno Vihinen. "Performance of Protein Stability Predictors". en. In: *Human Mutation* 31.6 (June 2010), pp. 675–684. ISSN: 1098-1004. DOI: 10.1002/humu.21242.

[30] Supratim Mukherjee et al. "Genomes OnLine Database (GOLD) v.6: Data Updates and Feature Enhancements". In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D446–D456. ISSN: 0305-1048. DOI: 10.1093/nar/gkw992.

[31] Simona Cocco et al. "Inverse Statistical Physics of Protein Sequences: A Key Issues Review". In: *arXiv:1703.01222 [cond-mat, q-bio]* (Mar. 2017). arXiv: 1703.01222 [cond-mat, q-bio].

[32] Hervé Jacquier et al. "Capturing the Mutational Landscape of the Beta-Lactamase TEM-1". en. In: *Proceedings of the National Academy of Sciences* 110.32 (June 2013), pp. 13067–13072. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1215206110.

[33] Elad Firnberg et al. "A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape". In: *Molecular Biology and Evolution* 31.6 (June 2014), pp. 1581–1592. ISSN: 0737-4038. DOI: 10.1093/molbev/msu081.

[34] Najeeb Halabi et al. "Protein Sectors: Evolutionary Units of Three-Dimensional Structure". en. In: *Cell* 138.4 (Aug. 2009), pp. 774–786. ISSN: 00928674. DOI: 10.1016/j.cell.2009.07.038.

[35] Kimberly A. Reynolds, Richard N. McLaughlin, and Rama Ranganathan. "Hot Spots for Allosteric Regulation on Protein Surfaces". English. In: *Cell* 147.7 (Dec. 2011), pp. 1564–1575. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2011.10.049.

[36] Arjun S. Raman, K. Ian White, and Rama Ranganathan. "Origins of Allostery and Evolvability in Proteins: A Case Study". In: *Cell* 166.2 (July 2016), pp. 468–480. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.05.047.

[37] Richard N. McLaughlin Jr et al. "The Spatial Architecture of Protein Function and Adaptation". In: *Nature* 491.7422 (Oct. 2012), pp. 138–142. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11500.

[38] Simona Cocco et al. "Inverse Statistical Physics of Protein Sequences: A Key Issues Review". In: *arXiv:1703.01222 [cond-mat, q-bio]* (Mar. 2017). arXiv: 1703.01222 [cond-mat, q-bio].

[39] William P. Russ et al. "Natural-like Function in Artificial WW Domains". en. In: *Nature* 437.7058 (Sept. 2005), pp. 579–583. ISSN: 0028-0836. DOI: 10.1038/nature03990.

[40] Michael Socolich et al. "Evolutionary Information for Specifying a Protein Fold". en. In: *Nature* 437.7058 (Sept. 2005), pp. 512–518. ISSN: 0028-0836. DOI: 10.1038/nature03991.

[41] Robert G Smock et al. "An Interdomain Sector Mediating Allostery in Hsp70 Molecular Chaperones". In: *Molecular Systems Biology* 6 (Sept. 2010), p. 414. ISSN: 1744-4292. DOI: 10.1038/msb.2010.65.

[42] B. C. Cunningham and J. A. Wells. "High-Resolution Epitope Mapping of hGH-Receptor Interactions by Alanine-Scanning Mutagenesis". en. In: *Science* 244.4908 (June 1989), pp. 1081–1085. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.2471267.

[43] Sriram Kosuri and George M. Church. "Large-Scale de Novo DNA Synthesis: Technologies and Applications". en. In: *Nature Methods* 11.5 (May 2014), pp. 499–507. ISSN: 1548-7091. DOI: 10.1038/nmeth.2918.

[44] Randall A. Hughes and Andrew D. Ellington. "Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology". en. In: *Cold Spring Harbor Perspectives in Biology* 9.1 (Jan. 2017), a023812. ISSN: , 1943-0264. DOI: 10.1101/cshperspect.a023812.

[45] Jacob O. Kitzman et al. "Massively Parallel Single Amino Acid Mutagenesis". In: *Nature methods* 12.3 (Mar. 2015), pp. 203–206. ISSN: 1548-7091. DOI: 10.1038/nmeth.3223.

[46] Emily E. Wrenbeck et al. "Plasmid-Based One-Pot Saturation Mutagenesis". en. In: *Nature Methods* 13.11 (Nov. 2016), pp. 928–930. ISSN: 1548-7091. DOI: 10.1038/nmeth.4029.

[47] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of Age: Ten Years of next-Generation Sequencing Technologies". en. In: *Nature Reviews Genetics* 17.6 (June 2016), pp. 333–351. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.49.

[48] Douglas M. Fowler et al. "High-Resolution Mapping of Protein Sequence-Function Relationships". en. In: *Nature Methods* 7.9 (Sept. 2010), pp. 741–746. ISSN: 1548-7091. DOI: 10.1038/nmeth.1492.

[49] Jacob O. Kitzman et al. "Massively Parallel Single Amino Acid Mutagenesis". In: *Nature methods* 12.3 (Mar. 2015), pp. 203–206. ISSN: 1548-7091. DOI: 10.1038/nmeth.3223.

[50] Michael W. Traxlmayr et al. "Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by Combining Directed Evolution with High Throughput Sequencing". In: *Journal of Molecular Biology* 423.3 (Oct. 2012), pp. 397–412. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2012.07.017.

[51] Teemu Kivioja et al. "Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers". en. In: *Nature Methods* 9.1 (Jan. 2012), pp. 72–74. ISSN: 1548-7091. DOI: 10.1038/nmeth.1778.

[52] Carlos L. Araya and Douglas M. Fowler. "Deep Mutational Scanning: Assessing Protein Function on a Massive Scale". In: *Trends in Biotechnology* 29.9 (Sept. 2011), pp. 435–442. ISSN: 0167-7799. DOI: 10.1016/j.tibtech.2011.04.003.

[53] C. Anders Olson, Nicholas C. Wu, and Ren Sun. "A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain". English. In: *Current Biology* 24.22 (Nov. 2014), pp. 2643–2651. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.09.072.

[54] Yasuhiro Fujino et al. "Robust in Vitro Affinity Maturation Strategy Based on Interface-Focused High-Throughput Mutational Scanning". In: *Biochemical and Biophysical Research Communications* 428.3 (Nov. 2012), pp. 395–400. ISSN: 0006-291X. DOI: 10.1016/j.bbrc.2012.10.066.

[55] Patrick Koenig et al. "Mutational Landscape of Antibody Variable Domains Reveals a Switch Modulating the Interdomain Conformational Dynamics and Antigen Binding". en. In: *Proceedings of the National Academy of Sciences* 114.4 (Jan. 2017), E486–E495. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1613231114.

[56] Gábor Pál et al. "Comprehensive and Quantitative Mapping of Energy Landscapes for Protein-Protein Interactions by Rapid Combinatorial Scanning". en. In: *Journal of Biological Chemistry* 281.31 (Apr. 2006), pp. 22378–22385. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M603826200.

[57] Carlos L. Araya et al. "A Fundamental Protein Property, Thermodynamic Stability, Revealed Solely from Large-Scale Measurements of Protein Function". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.42 (Oct. 2012), pp. 16858–16863. ISSN: 0027-8424. DOI: 10.1073/pnas.1209751109.

[58] Andreas Ernst et al. "Coevolution of PDZ Domain–ligand Interactions Analyzed by High-Throughput Phage Display and Deep Sequencing". en. In: *Molecular BioSystems* 6.10 (Oct. 2010), pp. 1782–1790. ISSN: 1742-2051. DOI: 10.1039/C0MB00061B.

[59] Lea M. Starita et al. "Activity-Enhancing Mutations in an E3 Ubiquitin Ligase Identified by High-Throughput Mutagenesis". en. In: *Proceedings of the National Academy of Sciences* 110.14 (Feb. 2013), E1263–E1272. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1303309110.

[60] Timothy A. Whitehead et al. "Optimization of Affinity, Specificity and Function of Designed Influenza Inhibitors Using Deep Sequencing". en. In: *Nature Biotechnology* 30.6 (June 2012), pp. 543–548. ISSN: 1087-0156. DOI: 10.1038/nbt.2214.

[61] Rhys M Adams et al. "Measuring the Sequence-Affinity Landscape of Antibodies with Massively Parallel Titration Curves". In: *eLife* 5 (2016). ISSN: 2050-084X. DOI: 10.7554/eLife.23156.

[62] Justin R. Klesmith et al. "Trade-Offs between Enzyme Fitness and Solubility Illuminated by Deep Mutational Scanning". en. In: *Proceedings of the National Academy of Sciences* 114.9 (Feb. 2017), pp. 2265–2270. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1614437114.

[63] Charles M. Forsyth et al. "Deep Mutational Scanning of an Antibody against Epidermal Growth Factor Receptor Using Mammalian Cell Display and Massively Parallel Pyrosequencing". In: *mAbs* 5.4 (July 2013), pp. 523–532. ISSN: 1942-0862. DOI: 10.4161/mabs.24979.

[64] Michael A. Stiffler, Doeke R. Hekstra, and Rama Ranganathan. "Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase". In: *Cell* 160.5 (Feb. 2015), pp. 882–892. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.01.035.

[65] Zhifeng Deng et al. "Deep Sequencing of Systematic Combinatorial Libraries Reveals β-Lactamase Sequence Constraints at High Resolution". In: *Journal of Molecular Biology* 424.3–4 (Dec. 2012), pp. 150–167. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2012.09.014.

[66] Ryan Hietpas et al. "Fitness Analyses of All Possible Point Mutations for Regions of Genes in Yeast". en. In: *Nature Protocols* 7.7 (July 2012), pp. 1382–1396. ISSN: 1754-2189. DOI: 10.1038/nprot.2012.069.

[67] Yvonne H. Chan et al. "Correlation of Fitness Landscapes from Three Orthologous TIM Barrels Originates from Sequence and Structure Constraints". en. In: *Nature Communications* 8 (Mar. 2017), p. 14614. ISSN: 2041-1723. DOI: 10.1038/ncomms14614.

[68] Benjamin P. Roscoe et al. "Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate". In: *Journal of molecular biology* 425.8 (Apr. 2013), pp. 1363–1377. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2013.01.032.

[69] Claudia Bank et al. "A Systematic Survey of an Intragenic Epistatic Landscape". In: *Molecular Biology and Evolution* 32.1 (Jan. 2015), pp. 229–238. ISSN: 0737-4038. DOI: 10.1093/molbev/msu301.

[70] Nicholas C. Wu et al. "Systematic Identification of H274Y Compensatory Mutations in Influenza A Virus Neuraminidase by High-Throughput Screening". In: *Journal of Virology* 87.2 (Jan. 2013), pp. 1193–1199. ISSN: 0022-538X. DOI: 10.1128/JVI.01658-12.

[71] Mariona Parera and Miguel Angel Martinez. "Strong Epistatic Interactions within a Single Protein". In: *Molecular Biology and Evolution* 31.6 (June 2014), pp. 1546–1553. ISSN: 0737-4038. DOI: 10.1093/molbev/msu113.

[72] Karen S. Sarkisyan et al. "Local Fitness Landscape of the Green Fluorescent Protein". en. In: *Nature* 533.7603 (May 2016), pp. 397–401. ISSN: 0028-0836. DOI: 10.1038/nature17995.

[73] Philip A. Romero, Tuan M. Tran, and Adam R. Abate. "Dissecting Enzyme Function with Microfluidic-Based Deep Mutational Scanning". en. In: *Proceedings of the National Academy of Sciences* 112.23 (Sept. 2015), pp. 7159–7164. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1422285112.

[74] Takayuki Sohka et al. "An Externally Tunable Bacterial Band-Pass Filter". en. In: *Proceedings of the National Academy of Sciences* 106.25 (June 2009), pp. 10135–10140. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0901246106.

[75] Anna I. Podgornaia and Michael T. Laub. "Pervasive Degeneracy and Epistasis in a Protein-Protein Interface". en. In: *Science* 347.6222 (Feb. 2015), pp. 673–677. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1257360.

[76] Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, and Dan S. Tawfik. "Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations". In: *PLoS Computational Biology* 11.8 (Aug. 2015). ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1004421.

[77] Bharat V. Adkar et al. "Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing". In: *Structure* 20.2 (Feb. 2012), pp. 371–381. ISSN: 0969-2126. DOI: 10.1016/j.str.2011.11.021.

[78] Alexandre Melnikov et al. "Comprehensive Mutational Scanning of a Kinase in Vivo Reveals Substrate-Dependent Fitness Landscapes". In: *Nucleic Acids Research* 42.14 (Aug. 2014), e112–e112. ISSN: 0305-1048. DOI: 10.1093/nar/gku511.

[79] Ryan T. Hietpas, Jeffrey D. Jensen, and Daniel N. A. Bolon. "Experimental Illumination of a Fitness Landscape". en. In: *Proceedings of the National Academy of Sciences* 108.19 (Oct. 2011), pp. 7896–7901. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1016024108.

[80] Daniel Melamed et al. "Deep Mutational Scanning of an RRM Domain of the Saccharomyces Cerevisiae Poly(A)-Binding Protein". In: *RNA* 19.11 (Nov. 2013), pp. 1537–1551. ISSN: 1355-8382. DOI: 10.1261/rna.040709.113.

[81] Frank J. Poelwijk, Vinod Krishna, and Rama Ranganathan. "The Context-Dependence of Mutations: A Linkage of Formalisms". In: *arXiv:1502.00726 [q-bio]* (Feb. 2015). arXiv: 1502.00726 [q-bio].

[82] Jamie T. Bridgham, Eric A. Ortlund, and Joseph W. Thornton. "An Epistatic Ratchet Constrains the Direction of Glucocorticoid Receptor Evolution". In: *Nature* 461.7263 (Sept. 2009), pp. 515–519. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08249.

[83] Jack da Silva et al. "Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region". In: *Genetics* 185.1 (May 2010), pp. 293–303. ISSN: 0016-6731. DOI: 10.1534/genetics.109.112458.

[84] Elena R. Lozovsky et al. "Stepwise Acquisition of Pyrimethamine Resistance in the Malaria Parasite". en. In: *Proceedings of the National Academy of Sciences* 106.29 (July 2009), pp. 12025–12030. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0905922106.

[85] Kyle M. Brown et al. "Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase". In: *Molecular Biology and Evolution* 27.12 (Dec. 2010), pp. 2682–2690. ISSN: 0737-4038. DOI: 10.1093/molbev/msq160.

[86] Christopher D. Aakre et al. "Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates". English. In: *Cell* 163.3 (Oct. 2015), pp. 594–606. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.09.055.

[87] Bryan C. Dickinson et al. "Experimental Interrogation of the Path Dependence and Stochasticity of Protein Evolution Using Phage-Assisted Continuous Evolution". en. In: *Proceedings of the National Academy of Sciences* 110.22 (May 2013), pp. 9007–9012. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1220670110.

[88] P.E. O'Maille et al. "Quantitative Exploration of the Catalytic Landscape Separating Divergent Plant Sesquiterpene Synthases". In: *Nature chemical biology* 4.10 (Oct. 2008), pp. 617–623. ISSN: 1552-4450. DOI: 10.1038/nchembio.113.

[89] Altan Ercan, Hyun Ik Park, and Li-June Ming. "A "Moonlighting" Dizinc Aminopeptidase from *Streptomyces Griseus* : Mechanisms for Peptide Hydrolysis and the 4×10^{10} -Fold Acceleration of the Alternative Phosphodiester Hydrolysis †". en. In: *Biochemistry* 45.46 (Nov. 2006), pp. 13779–13793. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/bi061086x.

[90] Xiaojun Wang, George Minasov, and Brian K. Shoichet. "Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-Offs". In: *Journal of Molecular Biology* 320.1 (June 2002), pp. 85–95. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(02)00400-X.

[91] Patrick Tabeling. "Introduction to Microfluidics". en. In: *Angewandte Chemie International Edition* 45.47 (Dec. 2006), pp. 7875–7875. ISSN: 1521-3773. DOI: 10.1002/anie.200585389.

[92] Yeshaiahu Fainman, Demetri Psaltis, and Changhuei Yang. "Basic Microfluidic and Soft Lithographic Techniques". In: *Optofluidics: Fundamentals, Devices and Applications*, 2010th ed. McGraw Hill Professional, 2010.

[93] Shelley L. Anna. "Formation of Dispersions Using "Flow Focusing" in Microchannels". In: *Applied Physics Letters* 82.3 (Jan. 2003), pp. 364–366. ISSN: 0003-6951. DOI: 10.1063/1.1537519.

[94] J. Bibette, F. Leal Calderon, and P. Poulin. "Emulsions: Basic Principles". en. In: *Reports on Progress in Physics* 62.6 (1999), p. 969. ISSN: 0034-4885. DOI: 10.1088/0034-4885/62/6/203.

[95] Jean-Christophe Baret. "Surfactants in Droplet-Based Microfluidics". eng. In: *Lab on a Chip* 12.3 (Feb. 2012), pp. 422–433. ISSN: 1473-0189. DOI: 10.1039/c1lc20582j.

[96] Emilia Nowak et al. "Effect of Surfactant Concentration and Viscosity of Outer Phase during the Coalescence of a Surfactant-Laden Drop with a Surfactant-Free Drop". In: *Colloids and Surfaces A: Physicochemical and Engineering Aspects*. 6th International Workshop on Bubble and Drop Interfaces 505 (Sept. 2016), pp. 124–131. ISSN: 0927-7757. DOI: 10.1016/j.colsurfa.2016.02.016.

[97] Maciej Pilarek, Julia Glazyrina, and Peter Neubauer. "Enhanced Growth and Recombinant Protein Production of Escherichia Coli by a Perfluorinated Oxygen Carrier in Miniaturized Fed-Batch Cultures". In: *Microbial Cell Factories* 10 (2011), p. 50. ISSN: 1475-2859. DOI: 10.1186/1475-2859-10-50.

[98] C. Holtze et al. "Biocompatible Surfactants for Water-in-Fluorocarbon Emulsions". en. In: *Lab on a Chip* 8.10 (Sept. 2008), pp. 1632–1639. ISSN: 1473-0189. DOI: 10.1039/B806706F.

[99] Amandine Trouchet. "PCR Digitale Pour La Détection et La Caractérisation de Micro-Organismes Pathogènes Au Niveau de La Cellule Unique." In: (2016).

[100] Vincent Miralles et al. "A Versatile Technology for Droplet-Based Microfluidics: Thermomechanical Actuation". en. In: *Lab on a Chip* 15.9 (Apr. 2015), pp. 2133–2139. ISSN: 1473-0189. DOI: 10.1039/C5LC00110B.

[101] Marie Leman et al. "Droplet-Based Microfluidics at the Femtolitre Scale". en. In: 15.3 (Jan. 2015), pp. 753–765. ISSN: 1473-0189. DOI: 10.1039/C4LC01122H.

[102] Rerngchai Arayanarakool et al. "Single-Enzyme Analysis in a Droplet-Based Micro- and Nanofluidic System". en. In: *Lab on a Chip* 13.10 (Apr. 2013), pp. 1955–1962. ISSN: 1473-0189. DOI: 10.1039/C3LC41100A.

[103] Z. Li et al. "Step-Emulsification in a Microfluidic Device". en. In: 15.4 (Feb. 2015), pp. 1023–1031. ISSN: 1473-0189. DOI: 10.1039/C4LC01289E.

[104] Alim Dewan et al. "Growth Kinetics of Microalgae in Microfluidic Static Droplet Arrays". en. In: *Biotechnology and Bioengineering* 109.12 (Dec. 2012), pp. 2987–2996. ISSN: 1097-0290. DOI: 10.1002/bit.24568.

[105] Claudiu A. Stan, Sindy K. Y. Tang, and George M. Whitesides. "Independent Control of Drop Size and Velocity in Microfluidic Flow-Focusing Generators Using Variable Temperature and Flow Rate". In: *Analytical Chemistry* 81.6 (Mar. 2009), pp. 2399–2402. ISSN: 0003-2700. DOI: 10.1021/ac8026542.

[106] Benjamin Dollet et al. "Role of the Channel Geometry on the Bubble Pinch-Off in Flow-Focusing Devices". In: *Physical Review Letters* 100.3 (Jan. 2008), p. 034504. DOI: 10.1103/PhysRevLett.100.034504.

[107] P. B. Umbanhowar, V. Prasad, and D. A. Weitz. "Monodisperse Emulsion Generation via Drop Break Off in a Coflowing Stream". In: *Langmuir* 16.2 (Jan. 2000), pp. 347–351. ISSN: 0743-7463. DOI: 10.1021/la990101e.

[108] Charles N. Baroud, Francois Gallaire, and Rémi Dangla. "Dynamics of Microfluidic Droplets". en. In: 10.16 (July 2010), pp. 2032–2045. ISSN: 1473-0189. DOI: 10.1039/C001191F.

[109] A. R. Abate et al. "Impact of Inlet Channel Geometry on Microfluidic Drop Formation". In: *Physical Review E* 80.2 (Aug. 2009), p. 026310. DOI: 10.1103/PhysRevE.80.026310.

[110] Bo Zheng, Joshua D. Tice, and Rustem F. Ismagilov. "Formation of Droplets of Alternating Composition in Microfluidic Channels and Applications to Indexing of Concentrations in Droplet-Based Assays". In: *Analytical Chemistry* 76.17 (Sept. 2004), pp. 4977–4982. ISSN: 0003-2700. DOI: 10.1021/ac0495743.

[111] Adam R. Abate et al. "High-Throughput Injection with Microfluidics Using Picoinjectors". en. In: *Proceedings of the National Academy of Sciences* 107.45 (Sept. 2010), pp. 19163–19166. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1006888107.

[112] Xize Niu et al. "Electro-Coalescence of Digitally Controlled Droplets". In: *Analytical Chemistry* 81.17 (Sept. 2009), pp. 7321–7325. ISSN: 0003-2700. DOI: 10.1021/ac901188n.

[113] Michele Zagnoni, Charles N. Baroud, and Jonathan M. Cooper. "Electrically Initiated Upstream Coalescence Cascade of Droplets in a Microfluidic Flow". In: *Physical Review E* 80.4 (Oct. 2009), p. 046303. DOI: 10.1103/PhysRevE.80.046303.

[114] Michele Zagnoni and Jonathan M. Cooper. "On-Chip Electrocoalescence of Microdroplets as a Function of Voltage, Frequency and Droplet Size". en. In: *Lab on a Chip* 9.18 (Sept. 2009), pp. 2652–2658. ISSN: 1473-0189. DOI: 10.1039/B906298J.

[115] Linas Mazutis, Jean-Christophe Baret, and Andrew D. Griffiths. "A Fast and Efficient Microfluidic System for Highly Selective One-to-One Droplet Fusion". en. In: *Lab on a Chip* 9.18 (Sept. 2009), pp. 2665–2672. ISSN: 1473-0189. DOI: 10.1039/B903608C.

[116] Yung-Chieh Tan et al. "Design of Microfluidic Channel Geometries for the Control of Droplet Volume, Chemical Concentration, and Sorting". en. In: *Lab on a Chip* 4.4 (July 2004), pp. 292–298. ISSN: 1473-0189. DOI: 10.1039/B403280M.

[117] Xize Niu et al. "Pillar-Induced Droplet Merging in Microfluidic Circuits". en. In: *Lab on a Chip* 8.11 (Nov. 2008), pp. 1837–1841. ISSN: 1473-0189. DOI: 10.1039/B813325E.

[118] Adam R. Abate and David A. Weitz. "Faster Multiple Emulsification with Drop Splitting". en. In: *Lab on a Chip* 11.11 (June 2011), pp. 1911–1915. ISSN: 1473-0189. DOI: 10.1039/C0LC00706D.

[119] Paul Abbyad et al. "Rails and Anchors: Guiding and Trapping Droplet Microreactors in Two Dimensions". en. In: *Lab on a Chip* 11.5 (Mar. 2011), pp. 813–821. ISSN: 1473-0189. DOI: 10.1039/C0LC00104J.

[120] Lucas Frenz et al. "Reliable Microfluidic On-Chip Incubation of Droplets in Delay-Lines". eng. In: *Lab on a Chip* 9.10 (May 2009), pp. 1344–1348. ISSN: 1473-0197. DOI: 10.1039/b816049j.

[121] Fabienne Courtois et al. "An Integrated Device for Monitoring Time-Dependent in Vitro Expression From Single Genes in Picolitre Droplets". en. In: *ChemBioChem* 9.3 (Feb. 2008), pp. 439–446. ISSN: 1439-7633. DOI: 10.1002/cbic.200700536.

[122] Linas Mazutis et al. "Droplet-Based Microfluidic Systems for High-Throughput Single DNA Molecule Isothermal Amplification and Analysis". en. In: *Analytical Chemistry* 81.12 (June 2009), pp. 4813–4821. ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac900403z.

[123] Eleonora Zonta et al. "Multiplex Detection of Rare Mutations by Picoliter Droplet Based Digital PCR: Sensitivity and Specificity Considerations". In: *PLOS ONE* 11.7 (July 2016), e0159094. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0159094.

[124] Jean-Christophe Baret et al. "Fluorescence-Activated Droplet Sorting (FADS): Efficient Microfluidic Cell Sorting Based on Enzymatic Activity". en. In: *Lab on a Chip* 9.13 (July 2009), pp. 1850–1858. ISSN: 1473-0189. DOI: 10.1039/B902504A.

[125] Hirosuke Maenaka et al. "Continuous and Size-Dependent Sorting of Emulsion Droplets Using Hydrodynamics in Pinched Microchannels". In: *Langmuir* 24.8 (Apr. 2008), pp. 4405–4410. ISSN: 0743-7463. DOI: 10.1021/la703581j.

[126] Yung-Chieh Tan, Yao Li Ho, and Abraham Phillip Lee. "Microfluidic Sorting of Droplets by Size". en. In: *Microfluidics and Nanofluidics* 4.4 (Apr. 2008), p. 343. ISSN: 1613-4982, 1613-4990. DOI: 10.1007/s10404-007-0184-1.

[127] P. Sajeesh et al. "A Microfluidic Device with Focusing and Spacing Control for Resistance-Based Sorting of Droplets and Cells". en. In: *Lab on a Chip* 15.18 (Aug. 2015), pp. 3738–3748. ISSN: 1473-0189. DOI: 10.1039/C5LC00598A.

[128] Fabrice Gielen et al. "Ultrahigh-Throughput–directed Enzyme Evolution by Absorbance-Activated Droplet Sorting (AADS)". en. In: *Proceedings of the National Academy of Sciences* 113.47 (Nov. 2016), E7383–E7389. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1606927113.

[129] Arjen M. Pit. "High-Throughput Sorting of Drops in Microfluidic Chips Using Electric Capacitance". In: *Biomicrofluidics* 9.4 (July 2015), p. 044116. DOI: 10.1063/1.4928452.

[130] Ahn. "Dielectrophoretic Manipulation of Drops for High-Speed Microfluidic Sorting Devices". In: *Applied Physics Letters* 88.2 (Jan. 2006), p. 024104. ISSN: 0003-6951. DOI: 10.1063/1.2164911.

[131] Adam Sciambi and Adam R. Abate. "Accurate Microfluidic Sorting of Droplets at 30kHz". In: *Lab on a chip* 15.1 (Jan. 2015), pp. 47–51. ISSN: 1473-0197. DOI: 10.1039/c4lc01194e.

[132] Thomas Franke et al. "Surface Acoustic Wave (SAW) Directed Droplet Flow in Microfluidics for PDMS Devices". en. In: *Lab on a Chip* 9.18 (Sept. 2009), pp. 2625–2627. ISSN: 1473-0189. DOI: 10.1039/B906819H.

[133] Kalia Bernath et al. "In Vitro Compartmentalization by Double Emulsions: Sorting and Gene Enrichment by Fluorescence Activated Cell Sorting". In: *Analytical Biochemistry* 325.1 (Feb. 2004), pp. 151–157. ISSN: 0003-2697. DOI: 10.1016/j.ab.2003.10.005.

[134] Anastasia Zinchenko et al. "One in a Million: Flow Cytometric Sorting of Single Cell-Lysate Assays in Monodisperse Picolitre Double Emulsion Droplets for Directed Evolution". In: *Analytical Chemistry* 86.5 (Mar. 2014), pp. 2526–2533. ISSN: 0003-2700. DOI: 10.1021/ac403585p.

[135] Balint Kintses et al. "Microfluidic Droplets: New Integrated Workflows for Biological Experiments". In: *Current Opinion in Chemical Biology*. Nanotechnology and Miniaturization/Mechanisms 14.5 (Oct. 2010), pp. 548–555. ISSN: 1367-5931. DOI: 10.1016/j.cbpa.2010.08.013.

[136] Jean-Christophe Baret et al. "Fluorescence-Activated Droplet Sorting (FADS): Efficient Microfluidic Cell Sorting Based on Enzymatic Activity". en. In: *Lab on a Chip* 9.13 (July 2009), pp. 1850–1858. ISSN: 1473-0189. DOI: 10.1039/B902504A.

[137] J. Bibette et al. "Stability Criteria for Emulsions". In: *Physical Review Letters* 69.16 (Oct. 1992), pp. 2439–2442. DOI: 10.1103/PhysRevLett.69.2439.

[138] Sarah Köster et al. "Drop-Based Microfluidic Devices for Encapsulation of Single Cells". en. In: *Lab on a Chip* 8.7 (June 2008), pp. 1110–1115. ISSN: 1473-0189. DOI: 10.1039/B802941E.

[139] Jenifer Clausell-Tormos et al. "Droplet-Based Microfluidic Platforms for the Encapsulation and Screening of Mammalian Cells and Multicellular Organisms". In: *Chemistry & Biology* 15.5 (May 2008), pp. 427–437. ISSN: 1074-5521. DOI: 10.1016/j.chembiol.2008.04.004.

[140] Ansgar Huebner et al. "Development of Quantitative Cell-Based Enzyme Assays in Microdroplets". In: *Analytical Chemistry* 80.10 (May 2008), pp. 3890–3896. ISSN: 0003-2700. DOI: 10.1021/ac800338z.

[141] David J. Collins et al. "The Poisson Distribution and beyond: Methods for Microfluidic Droplet Production and Single Cell Encapsulation". en. In: *Lab on a Chip* 15.17 (Aug. 2015), pp. 3439–3459. ISSN: 1473-0189. DOI: 10.1039/C5LC00614G.

[142] Jon F. Edd et al. "Controlled Encapsulation of Single-Cells into Monodisperse Picolitre Drops". en. In: *Lab on a Chip* 8.8 (July 2008), pp. 1262–1264. ISSN: 1473-0189. DOI: 10.1039/B805456H.

[143] Ali Fallah-Araghi et al. "A Completely in Vitro Ultrahigh-Throughput Droplet-Based Microfluidic Screening System for Protein Engineering and Directed Evolution". en. In: *Lab on a Chip* 12.5 (Feb. 2012), pp. 882–891. ISSN: 1473-0189. DOI: 10.1039/C2LC21035E.

[144] Michael Ryckelynck et al. "Using Droplet-Based Microfluidics to Improve the Catalytic Properties of RNA under Multiple-Turnover Conditions". en. In: *RNA* 21.3 (Jan. 2015), pp. 458–469. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.048033.114.

[145] Linas Mazutis et al. "Multi-Step Microfluidic Droplet Processing: Kinetic Analysis of an in Vitro Translated Enzyme". en. In: *Lab on a Chip* 9.20 (Oct. 2009), pp. 2902–2908. ISSN: 1473-0189. DOI: 10.1039/B907753G.

[146] Balint Kintses et al. "Picoliter Cell Lysate Assays in Microfluidic Droplet Compartments for Directed Enzyme Evolution". In: *Chemistry & Biology* 19.8 (Aug. 2012), pp. 1001–1009. ISSN: 1074-5521. DOI: 10.1016/j.chembiol.2012.06.009.

[147] Pierre-Yves Colin et al. "Ultrahigh-Throughput Discovery of Promiscuous Enzymes by Picodroplet Functional Metagenomics". en. In: *Nature Communications* 6 (Dec. 2015), p. 10008. ISSN: 2041-1723. DOI: 10.1038/ncomms10008.

[148] Fabrice Gielen et al. "Ultrahigh-Throughput–directed Enzyme Evolution by Absorbance-Activated Droplet Sorting (AADS)". en. In: *Proceedings of the National Academy of Sciences* 113.47 (Nov. 2016), E7383–E7389. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1606927113.

[149] Stanislav S. Terekhov et al. "Microfluidic Droplet Platform for Ultrahigh-Throughput Single-Cell Screening of Biodiversity". en. In: *Proceedings of the National Academy of Sciences* 114.10 (July 2017), pp. 2550–2555. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1621226114.

[150] Thomas Beneyton et al. "Droplet-Based Microfluidic High-Throughput Screening of Heterologous Enzymes Secreted by the Yeast Yarrowia Lipolytica". In: *Microbial Cell Factories* 16 (2017), p. 18. ISSN: 1475-2859. DOI: 10.1186/s12934-017-0629-5.

[151] Staffan L. Sjostrom et al. "High-Throughput Screening for Industrial Enzyme Production Hosts by Droplet Microfluidics". en. In: *Lab on a Chip* 14.4 (Jan. 2014), pp. 806–813. ISSN: 1473-0189. DOI: 10.1039/C3LC51202A.

[152] L. Spencer Roach, Helen Song, and Rustem F. Ismagilov. "Controlling Nonspecific Protein Adsorption in a Plug-Based Microfluidic System by Controlling Interfacial Chemistry Using Fluorous-Phase Surfactants". In: *Analytical Chemistry* 77.3 (Feb. 2005), pp. 785–796. ISSN: 0003-2700. DOI: 10.1021/ac049061w.

[153] Shigeyoshi Matsumura et al. "Transient Compartmentalization of RNA Replicators Prevents Extinction Due to Parasites". en. In: *Science* 354.6317 (Dec. 2016), pp. 1293–1296. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aag1582.

[154] Jonathan Garamella et al. "The All E. Coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology". In: *ACS Synthetic Biology* 5.4 (Apr. 2016), pp. 344–355. DOI: 10.1021/acssynbio.5b00296.

[155] Yoshihiro Shimizu et al. "Cell-Free Translation Reconstituted with Purified Components". en. In: *Nature Biotechnology* 19.8 (Jan. 2001), pp. 751–755. ISSN: 1087-0156. DOI: 10.1038/90802.

[156] Yoshihiro Shimizu, Takashi Kanamori, and Takuya Ueda. "Protein Synthesis by Pure Translation Systems". In: *Methods*. Engineering Translation 36.3 (July 2005), pp. 299–304. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2005.04.006.

[157] Marian Wenzel et al. "Self-Inducible Bacillus Subtilis Expression System for Reliable and Inexpensive Protein Production by High-Cell-Density Fermentation ▽". In: *Applied and Environmental Microbiology* 77.18 (Sept. 2011), pp. 6419–6425. ISSN: 0099-2240. DOI: 10.1128/AEM.05219-11.

[158] Thomas Beneyton et al. "CotA Laccase: High-Throughput Manipulation and Analysis of Recombinant Enzyme Libraries Expressed in E. Coli Using Droplet-Based Microfluidics". en. In: *Analyst* 139.13 (2014), pp. 3314–3323. DOI: 10.1039/C4AN00228H.

[159] A. Huebner et al. "Quantitative Detection of Protein Expression in Single Cells Using Droplet Microfluidics". en. In: *Chemical Communications* 12 (Mar. 2007), pp. 1218–1220. ISSN: 1364-548X. DOI: 10.1039/B618570C.

[160] Pierre R. Marcoux et al. "Micro-Confinement of Bacteria into w/o Emulsion Droplets for Rapid Detection and Enumeration". In: *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 377.1–3 (Mar. 2011), pp. 54–62. ISSN: 0927-7757. DOI: 10.1016/j.colsurfa.2010.12.013.

[161] Jung-uk Shim et al. "Simultaneous Determination of Gene Expression and Enzymatic Activity in Individual Bacterial Cells in Microdroplet Compartments". In: *Journal of the American Chemical Society* 131.42 (Oct. 2009), pp. 15251–15256. ISSN: 0002-7863. DOI: 10.1021/ja904823z.

[162] Alexei Godina. *In Vivo and in Vitro Directed Evolution of Enzymes Using Droplet-Based Microfluidics*. Strasbourg, Feb. 2013.

[163] Jeremy J. Agresti et al. "Ultrahigh-Throughput Screening in Drop-Based Microfluidics for Directed Evolution". en. In: *Proceedings of the National Academy of Sciences* 107.9 (Feb. 2010), pp. 4004–4009. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0910781107.

[164] Yong Nam Ahn et al. "Molecular Transport through Surfactant-Covered Oil-Water Interfaces: Role of Physical Properties of Solutes and Surfactants in Creating Energy Barriers for Transport". In: *Langmuir* 27.6 (Mar. 2011), pp. 2420–2436. ISSN: 0743-7463. DOI: 10.1021/la103550v.

[165] Yousr Skhiri et al. "Dynamics of Molecular Transport by Surfactants in Emulsions". en. In: *Soft Matter* 8.41 (Oct. 2012), pp. 10618–10627. ISSN: 1744-6848. DOI: 10.1039/C2SM25934F.

[166] Philipp Gruner et al. "Controlling Molecular Transport in Minimal Emulsions". en. In: *Nature Communications* 7 (Jan. 2016), p. 10392. ISSN: 2041-1723. DOI: 10.1038/ncomms10392.

[167] Johan Fenneteau et al. "Synthesis of New Hydrophilic Rhodamine Based Enzymatic Substrates Compatible with Droplet-Based Microfluidic Assays". en. In: *Chemical Communications* (Apr. 2017). ISSN: 1364-548X. DOI: 10.1039/C7CC01506B.

[168] Farid J. Ghadessy, Jennifer L. Ong, and Philipp Holliger. "Directed Evolution of Polymerase Function by Compartmentalized Self-Replication". en. In: *Proceedings of the National Academy of Sciences* 98.8 (Oct. 2001), pp. 4552–4557. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.071052198.

[169] Gabrielle Woronoff et al. "New Generation of Amino Coumarin Methyl Sulfonate-Based Fluorogenic Substrates for Amidase Assays in Droplet-Based Microfluidic Applications". In: *Analytical Chemistry* 83.8 (Apr. 2011), pp. 2852–2857. ISSN: 0003-2700. DOI: 10.1021/ac200373n.

[170] Majdi Najah et al. "New Glycosidase Substrates for Droplet-Based Microfluidic Screening". In: *Analytical Chemistry* 85.20 (Oct. 2013), pp. 9807–9814. ISSN: 0003-2700. DOI: 10.1021/ac4022709.

[171] Ee Xien Ng et al. "Single Cell Multiplexed Assay for Proteolytic Activity Using Droplet Microfluidics". In: *Biosensors and Bioelectronics* 81 (July 2016), pp. 408–414. ISSN: 0956-5663. DOI: 10.1016/j.bios.2016.03.002.

[172] Fabienne Courtois et al. "Controlling the Retention of Small Molecules in Emulsion Microdroplets for Use in Cell-Based Assays". In: *Analytical Chemistry* 81.8 (Apr. 2009), pp. 3008–3016. ISSN: 0003-2700. DOI: 10.1021/ac802658n.

[173] Linas Mazutis et al. "Droplet-Based Microfluidic Systems for High-Throughput Single DNA Molecule Isothermal Amplification and Analysis". In: *Analytical Chemistry* 81.12 (June 2009), pp. 4813–4821. ISSN: 0003-2700. DOI: 10.1021/ac900403z.

[174] Richard Obexer et al. "Emergence of a Catalytic Tetrad during Evolution of a Highly Active Artificial Aldolase". en. In: *Nature Chemistry* 9.1 (Jan. 2017), pp. 50–56. ISSN: 1755-4330. DOI: 10.1038/nchem.2596.

[175] Petter Hammar et al. "Single-Cell Screening of Photosynthetic Growth and Lactate Production by Cyanobacteria". In: *Biotechnology for Biofuels* 8 (2015), p. 193. ISSN: 1754-6834. DOI: 10.1186/s13068-015-0380-2.

[176] A. Spungin and S. Blumberg. "Streptomyces Griseus Aminopeptidase Is a Calcium-Activated Zinc Metalloprotein. Purification and Properties of the Enzyme". eng. In: *European Journal of Biochemistry* 183.2 (Aug. 1989), pp. 471–477. ISSN: 0014-2956.

[177] Daniella Ben-Meir et al. "Specificity of Streptomyces Griseus Aminopeptidase and Modulation of Activity by Divalent Metal Ion Binding and Substitution". en. In: *European Journal of Biochemistry* 212.1 (Feb. 1993), pp. 107–112. ISSN: 1432-1033. DOI: 10.1111/j.1432-1033.1993.tb17639.x.

[178] Bruno Maras et al. "Aminopeptidase from Streptomyces Griseus". en. In: *European Journal of Biochemistry* 236.3 (Mar. 1996), pp. 843–846. ISSN: 1432-1033. DOI: 10.1111/j.1432-1033.1996.00843.x.

[179] Orna Almog et al. "Crystallization and Preliminary Crystallographic Analysis of Streptomyces Griseus Aminopeptidase". In: *Journal of Molecular Biology* 230.1 (Mar. 1993), pp. 342–344. ISSN: 0022-2836. DOI: 10.1006/jmbi.1993.1146.

[180] H. M. Greenblatt et al. "Streptomyces Griseus Aminopeptidase: X-Ray Crystallographic Structure at 1.75 \AA Resolution". In: *Journal of molecular biology* 265.5 (1997), pp. 620–636.

[181] Michael N Harris and Li-June Ming. "Different Phosphate Binding Modes of Streptomyces Griseus Aminopeptidase between Crystal and Solution States and the Status of Zinc-Bound Water". In: *FEBS Letters* 455.3 (July 1999), pp. 321–324. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(99)00879-0.

[182] R. Gilboa et al. "Interactions of Streptomyces Griseus Aminopeptidase with a Methionine Product Analogue: A Structural Study at 1.53 \AA Resolution". In: *Acta Crystallographica Section D: Biological Crystallography* 56.5 (2000), pp. 551–558.

[183] Yifat Fundoiano-Hershcovitz et al. "Identification of the Catalytic Residues in the Double-Zinc Aminopeptidase from Streptomyces Griseus". en. In: *FEBS Letters* 571.1-3 (July 2004), pp. 192–196. ISSN: 00145793. DOI: 10.1016/j.febslet.2004.07.001.

[184] J. Arima. "Modulation of Streptomyces Leucine Aminopeptidase by Calcium: IDENTIFICATION AND FUNCTIONAL ANALYSIS OF KEY RESIDUES IN ACTIVATION AND STABILIZATION BY CALCIUM". en. In: *Journal of Biological Chemistry* 281.9 (Dec. 2005), pp. 5885–5894. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M509025200.

[185] Jiro Arima et al. "The Role of Glu196 in the Environment around the Substrate Binding Site of Leucine Aminopeptidase from Streptomyces Griseus". en. In: *FEBS Letters* 580.3 (Feb. 2006), pp. 912–917. ISSN: 1873-3468. DOI: 10.1016/j.febslet.2006.01.014.

[186] Giordano F. Z. da Silva and Li-June Ming. "Catechol Oxidase Activity of Di-Cu2+-Substituted Aminopeptidase from Streptomyces Griseus". In: *Journal of the American Chemical Society* 127.47 (Nov. 2005), pp. 16380–16381. ISSN: 0002-7863. DOI: 10.1021/ja056034u.

[187] Altan Ercan, Hyun Ik Park, and Li-June Ming. "A "Moonlighting" Dizinc Aminopeptidase from Streptomyces Griseus: Mechanisms for Peptide Hydrolysis and the 4 × 1010-Fold Acceleration of the Alternative Phosphodiester Hydrolysis". In: *Biochemistry* 45.46 (Nov. 2006), pp. 13779–13793. ISSN: 0006-2960. DOI: 10.1021/bi061086x.

[188] Altan Ercan et al. "Mechanistic Role of Each Metal Ion in Streptomyces Dinuclear Aminopeptidase: Peptide Hydrolysis and 7 × 1010-Fold Rate Enhancement of Phosphodiester Hydrolysis". In: *Journal of Inorganic Biochemistry* 104.1 (Jan. 2010), pp. 19–29. ISSN: 0162-0134. DOI: 10.1016/j.jinorgbio.2009.09.019.

[189] Hyun Ik Park and Li-June Ming. "A 1010 Rate Enhancement of Phosphodiester Hydrolysis by a Dinuclear Aminopeptidase—Transition-State Analogues as Substrates?" en. In: *Angewandte Chemie International Edition* 38.19 (Oct. 1999), pp. 2914–2916. ISSN: 1521-3773. DOI: 10.1002/(SICI)1521-3773(19991004)38:19<2914::AID-ANIE2914>3.0.CO;2-P.

[190] Altan Ercan, Hyun Ik Park, and Li-June Ming. "Remarkable Enhancement of the Hydrolyses of Phosphoesters by Dinuclear Centers: Streptomyces Aminopeptidase as a 'Natural Model System'". en. In: *Chemical Communications* 24 (Jan. 2000), pp. 2501–2502. ISSN: 1364-548X. DOI: 10.1039/B004544F.

[191] Jiro Arima, Masaki Iwabuchi, and Tadashi Hatanaka. "Gene Cloning and Overproduction of an Aminopeptidase from Streptomyces Septatus TH-2, and Comparison with a Calcium-Activated Enzyme from Streptomyces Griseus". en. In: *Biochemical and Biophysical Research Communications* 317.2 (Apr. 2004), pp. 531–538. ISSN: 0006291X. DOI: 10.1016/j.bbrc.2004.03.082.

[192] Lizbeth Hedstrom. "Serine Protease Mechanism and Specificity". In: *Chemical Reviews* 102.12 (Dec. 2002), pp. 4501–4524. ISSN: 0009-2665. DOI: 10.1021/cr000033x.

[193] Guy Dodson and Alexander Wlodawer. "Catalytic Triads and Their Relatives". In: *Trends in Biochemical Sciences* 23.9 (Sept. 1998), pp. 347–352. ISSN: 0968-0004. DOI: 10.1016/S0968-0004(98)01254-7.

[194] L. Hedstrom, L. Szilagyi, and W. J. Rutter. "Converting Trypsin to Chymotrypsin: The Role of Surface Loops". en. In: *Science* 255.5049 (Mar. 1992), pp. 1249–1253. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1546324.

[195] L. Hedstrom et al. "Converting Trypsin to Chymotrypsin: Ground-State Binding Does Not Determine Substrate Specificity". eng. In: *Biochemistry* 33.29 (July 1994), pp. 8764–8769. ISSN: 0006-2960.

[196] L. Hedstrom, J. J. Perona, and W. J. Rutter. "Converting Trypsin to Chymotrypsin: Residue 172 Is a Substrate Specificity Determinant". eng. In: *Biochemistry* 33.29 (July 1994), pp. 8757–8763. ISSN: 0006-2960.

[197] Alexei Godina. "In Vivo and in Vitro Directed Evolution of Enzymes Using Droplet-Based Microfluidics". PhD thesis. Université de Strasbourg, 2013.

[198] Thomas Beneyton et al. "High-Throughput Screening of Filamentous Fungi Using Nanoliter-Range Droplet-Based Microfluidics". eng. In: *Scientific Reports* 6 (June 2016), p. 27223. ISSN: 2045-2322. DOI: 10.1038/srep27223.

[199] Nathan C. Shaner et al. "Improved Monomeric Red, Orange and Yellow Fluorescent Proteins Derived from Discosoma Sp. Red Fluorescent Protein". en. In: *Nature Biotechnology* 22.12 (Dec. 2004), pp. 1567–1572. ISSN: 1087-0156. DOI: 10.1038/nbt1037.

[200] Thuy Dinh and Thomas G. Bernhardt. "Using Superfolder Green Fluorescent Protein for Periplasmic Protein Localization Studies ▽". In: *Journal of Bacteriology* 193.18 (Sept. 2011), pp. 4984–4987. ISSN: 0021-9193. DOI: 10.1128/JB.00315-11.

[201] Lidia Westers, Helga Westers, and Wim J. Quax. "Bacillus Subtilis as Cell Factory for Pharmaceutical Proteins: A Biotechnological Approach to Optimize the Host Organism". In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*. Protein Export/Secretion in Bacteria 1694.1–3 (Nov. 2004), pp. 299–310. ISSN: 0167-4889. DOI: 10.1016/j.bbamcr.2004.02.011.

[202] Sau-Ching Wu et al. "Functional Production and Characterization of a Fibrin-Specific Single-Chain Antibody Fragment from Bacillus Subtilis: Effects of Molecular Chaperones and a Wall-Bound Protease on Antibody Fragment Production". en. In: *Applied and Environmental Microbiology* 68.7 (Jan. 2002), pp. 3261–3269. ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.68.7.3261-3269.2002.

[203] Duc Nguyen Hoang. "Construction of Plasmid-Based Expression and Secretion Vectors and Study of the Immobilization of Proteins on the Surface of Bacillus Subtilis Cells". In: (2006).

[204] Ran Tu et al. "A Flow Cytometry–Based Screening System for Directed Evolution of Proteases". en. In: *Journal of Biomolecular Screening* 16.3 (Mar. 2011), pp. 285–294. ISSN: 1087-0571. DOI: 10.1177/1087057110396361.

[205] Hoang Duc Nguyen et al. "Construction of Plasmid-Based Expression Vectors for Bacillus Subtilis Exhibiting Full Structural Stability". In: *Plasmid* 54.3 (Nov. 2005), pp. 241–248. ISSN: 0147-619X. DOI: 10.1016/j.plasmid.2005.05.001.

[206] Trang Thi Phuong Phan, Hoang Duc Nguyen, and Wolfgang Schumann. "Novel Plasmid-Based Expression Vectors for Intra- and Extracellular Production of Recombinant Proteins in Bacillus Subtilis". In: *Protein Expression and Purification* 46.2 (Apr. 2006), pp. 189–195. ISSN: 1046-5928. DOI: 10.1016/j.pep.2005.07.005.

[207] Éva Várallyay et al. "Two Mutations in Rat Trypsin Confer Resistance against Autolysis". In: *Biochemical and Biophysical Research Communications* 243.1 (Feb. 1998), pp. 56–60. ISSN: 0006-291X. DOI: 10.1006/bbrc.1997.8058.

[208] Jianying Z. Kiser et al. "Streptomyces Erythraeus Trypsin for Proteomics Applications". In: *Journal of Proteome Research* 8.4 (Apr. 2009), pp. 1810–1817. ISSN: 1535-3893. DOI: 10.1021/pr8004919.

[209] Lisa Mahler et al. "Enhanced and Homogeneous Oxygen Availability during Incubation of Microfluidic Droplets". en. In: *RSC Advances* 5.123 (2015), pp. 101871–101878. DOI: 10.1039/C5RA20118G.

[210] Philipp Gruner et al. "Controlling Molecular Transport in Minimal Emulsions". en. In: *Nature Communications* 7 (Jan. 2016), p. 10392. ISSN: 2041-1723. DOI: 10.1038/ncomms10392.

[211] Pankaj C. Jain and Raghavan Varadarajan. "A Rapid, Efficient, and Economical Inverse Polymerase Chain Reaction-Based Method for Generating a Site Saturation Mutant Library". In: *Analytical Biochemistry* 449 (Mar. 2014), pp. 90–98. ISSN: 0003-2697. DOI: 10.1016/j.ab.2013.12.002.

[212] David F. Lee et al. "Mapping DNA Polymerase Errors by Single-Molecule Sequencing". In: *Nucleic Acids Research* 44.13 (July 2016), e118–e118. ISSN: 0305-1048. DOI: 10.1093/nar/gkw436.

[213] R H Don et al. "'Touchdown' PCR to Circumvent Spurious Priming during Gene Amplification." In: *Nucleic Acids Research* 19.14 (July 1991), p. 4008. ISSN: 0305-1048.