Characterising fitness landscapes in protein evolution by next-generation sequencing

Maya Petek

Supervisor: Prof. F. Hollfelder

Department of Biochemistry
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Pembroke College

April 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation contains fewer than 60,000 words excluding tables, bibliography, footnotes and appendices. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

Parts of this work have been published:

Stephane Emond, Maya Petek, Emily J. Kay, Brennen Heames, Sean R. A. Devenish, Nobuhiko Tokuriki and Florian Hollfelder. Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nature Communications* **11**, 3469 (2020). DOI: 10.1038/s41467-020-17061-3.

Parts of this work are in preparation for publication:

Remkes A. Scheele*, Laurens H. Lindenburg*, Maya Petek*, Markus Schober, Kevin N. Dalby and Florian Hollfelder. Sequence-function mapping of MKK1-ERK2 interactions reveals positive epistasis in the MKK1 D-domain. *In preparation. * Shared first authorship.*

Maya Petek
April 2020

Abstract

A protein's amino acid sequence determines its structural, chemical and physical properties, yet how sequence variation influences protein function is still incompletely understood. Protein fitness landscapes powerfully describe the sequence-function relationship by dividing sequence space into functional hills and valleys. This representation is often invoked yet lacks experimental evidence; the immense vastness of possible sequence space makes comprehensive high-quality datasets difficult to obtain. Laboratory directed evolution has focused on optimal utilisation of substitution libraries, however examination of functional innovation in Nature shows that short insertions and deletions (InDels) also play a key role. Beyond rare targeted studies of specific InDels, high-throughput data on fitness landscape for mutations other than substitutions are lacking entirely.

In my PhD, I worked towards experimentally describing the fitness landscapes of InDels and substitutions in three systems: GFP, phosphotriesterase (PTE) and the kinase MKK1 docking domain. Towards this goal, I established two experimental assays (GFP, PTE) for deep mutational scanning and a new software toolkit, InDelScanner, for interpreting resulting data that contain InDels.

With GFP, I sorted the deletions and substitution libraries into three activity fractions using FACS, then deep sequenced them with Illumina MiSeq to obtain a pilot dataset. The comparison of deletion effects between different lengths of deletions (-3, -6 and -9 bp) indicates that deletions are partially tolerated in eGFP, with tolerance improved for short deletions and in the stabilised starting point GFP8. Further interpretation of data was complicated by limited resolution in the sequencing dataset stemming from poor FACS separation, so I optimised the conditions for better sorting resolution using the mKate2 fluorescent protein as an expression reporter. In the second iteration of the activity sorting I additionally included UMIs in the plasmid design to improve the utilisation of NGS capacity.

In the case of PTE, I performed proof-of-concept experiments for microfluidic droplet sorting in an integrated device with an in-line incubation line and a fluorescent sorting design. In parallel, testing of solubility and activity of random InDel variants showed that functional

InDels do not necessarily suffer from a stability handicap, making InDel mutagenesis a viable strategy for gene randomisation in directed evolution.

One challenge of InDel library data analysis is that InDels are not compatible with existing, substitution-focused software. Using the GFP deletions dataset, I developed the InDelScanner scripts which accurately detect, aggregate and filter insertions, deletions and substitutions. Using the scripts for composition analysis of TRIAD libraries in PTE showed these libraries are well balanced and highly diverse.

Finally, I used the InDelScanner scripts to interpret a deep mutational scanning dataset that recorded the sequence preferences in the MKK1 docking domain, acting to activate ERK2. This experiment showed that the fitness landscape in this kinase pair is shaped by the activating effect of hydrophobic residues in the docking groove, as well as widespread positive epistasis.

Together, the projects in this thesis demonstrate that deep mutational scanning experiments are a powerful method for exploring the sequence-function relationship in proteins, which can extend into comparison of different types of mutations as well as probing their (epistatic) interactions.

Table of contents

List of figures

List of tables

List of Abbreviations

Roman Symbols

E. coli *Escherichia coli* bacteria: it may refer to any commonly used laboratory strain unless otherwise specified

AADS absorbance-activated droplet sorting

aTc Anhydrotetraycline

avGFP A variant of GFP as originally isolated from *Aequorea victoria* jellyfish, sometimes referred to as wild type GFP

bp base pair in DNA or RNA

CAPS N-Cyclohexyl-3-aminopropanesulfonic acid

CHES N-Cyclohexyl-2-aminoethanesulfonic acid

eGFP enhanced green fluorescent protein

esGFP extra-superfolder green fluorescent protein, with 12 mutations in addition to those in sfGFP

FADS fluorescence-activated droplet sorting

GFP The green fluorescent protein; this may denote any of its variants and derivatives that still fluoresce in the green spectrum

GFP8 a green fluorescent protein variant with six additional stabilizing mutations compared to eGFP and fewer total mutations than superfolder GFP

HEPES 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid, a zwitterionic buffer with pKa_{a1} 3 for the sulfonic acid and pKa_{a2} 7.5 for the amine mL Millilitre

NADPH Dihydronicotinamide adenine dinucleotide phosphate, i.e. reduced form of NADP+

Ni-NTA Nickel-nitrilotriacetic acid

NW Needleman-Wunsch

OD_{600nm} Optical density of cell culture, measured as absorbance at 600 nm

PBS Phosphate buffered saline

PDB The Protein Data Bank, available at https://www.wwpdb.org/

rcf Relative centrifugal force

rpm Revolutions per minute

SDS-PAGE Sodium dodecyl sulphate - polyacrylamide gel electrophoresis; the denaturing gel used to determine protein purity and molecular weight

sfGFP superfolder green fluorescent protein, combining eGFP, cycle-3 and six additional stabilizing or neutral mutations

SW Smith-Watermann

Chapter 1

Introduction

1.1 Protein evolution in Nature

Changes in living organisms are shaped by natural selection, where adaptations can occur at multiple levels; ranging from behaviour, to developmental patterns, down to changes in protein interaction networks or the function of a single protein. While selection in natural evolution always occurs at the level of the whole organism, investigation of fitness changes in essential components allows us to learn how new functions are acquired at a molecular level.

The diversity of the protein sequences observed in Nature is immense, and is still increasing through selection for new functions. The early evolution of protein domains must have occurred through *de novo* acquisition of function via modification of non-coding sequences, and it may still be ongoing, but it now happens more rarely due to excellent error-correction in DNA replication and the low probability of finding functional sequences in the vast protein space. For most of evolutionary time including the present, the main drivers of expansion of protein sequences have been gene duplication, recombination and gene divergence through introduction of mutations (Chothia et al., 2003).

1.1.1 Comparing the occurrence of InDels and substitutions in natural evolution

Sequence modification through introduction of mutations is the most frequent process in evolution and comprises three classes of mutations (see Figure 1.1):

- synonymous mutations, which keep the protein sequence unchanged. They are the most conservative class, which can affect the protein function through changes in mRNA stability and translational changes due to different codon usage.

- non-synonymous substitutions, which exchange one amino acid for another. These span a large range of effects, from no impact to complete loss of function due to effects on stability, to various possible functional effects on folding, substrate binding or enzyme catalysis. Their effects are amplified when they appear in the carefully organised enzyme active sites.
- insertions or deletions (InDels), where a short sequence is added or removed in the starting DNA sequence.

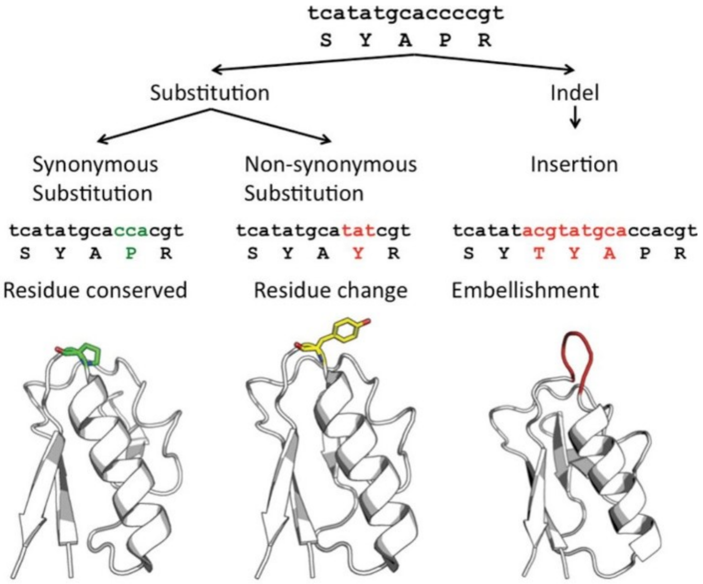


Figure 1.1 An illustration of the possible effect of the three classes of mutation, showing the effect of substitutions compared to small InDels on protein structure. Adapted from Studer et al. (2013).

Amino acid substitutions maintain the length of the protein backbone and only change the structure of the side chains, which is a small local change (although the effect can propagate further in the protein through successive conformation changes). In contrast, InDels alter both the protein backbone and side chain composition, so they can cause more drastic changes to the protein structure. Furthermore, a random change in coding DNA length has a 2 out of 3 chance of causing a change in protein reading frame, which generally result in non-functional proteins and are therefore purged from the genome.

While the functional impact of substitutions is well studied, the role of small InDels on protein structure and function is less well understood, despite InDels making up 15-20% of human genome polymorphisms (Mullaney et al., 2010). It has been shown that both deletions and insertions typically occur during DNA replication due to replication fork slippage, and additionally insertions have been traced back to recombination events (Kvikstad et al., 2007).

InDels are also rarer than substitutions, occurring at rates between 1:5 InDels to substitutions in mammalian genomes (including human) and 1:20 across tens of bacterial genomes. This variation is mostly due to much higher substitution rates in coding sequences, where the ratio can be as high as 1:60, and is lower but still clearly in favour of substitutions in non-coding regions (Chen et al., 2009).

An analysis of 35 proteomes from fungi, plants and metazoans showed that simple InDels are mostly small (1-5 amino acids long) and the most frequent length is a single amino acid (Ajawatanawong and Baldauf, 2013), corroborating earlier observations that small InDels predominate in available protein crystal structures (Benner et al., 1993; Pascarella and Argos, 1992). Interestingly, insertions were observed at higher frequency than deletions, which may indicate a different impact on protein function or a difference in molecular mechanisms generating insertions compared to deletions (Ajawatanawong and Baldauf, 2013). The higher frequency of insertions over deletions could also be a result of different experimental design: an exploration of proteomes, which filters out small proteins by design, may facilitate the detection of insertions over deletions. Other studies that compared genomic sequences found that deletions outnumbered insertions among non-frameshifting InDels (Lin et al., 2017; Taylor et al., 2004), so the debate about the frequency of InDels is on-going.

Curiously, it has been shown that the frequency of substitution is sharply increased in close proximity to InDels, then decreases to background substitution rate over several hundred bases (Tian et al., 2008). It appears the InDel event either has mutagenic effect on the surrounding sequence, or the change started by the InDel creates a selective pressure for adaptation of surrounding sequence to evolve and better accommodate the InDel. Alternatively, a comparison of protein orthologs in fungi suggests that the substitutions in a protein sequence accumulate first, followed by the first InDel event at a later point in divergent evolution (Tóth-Petróczy and Tawfik, 2013). The 'substitution-first' view points towards certain sequences being intrinsically more tolerant of mutations, both of the substitution and InDel variety, such that the InDels do not require further remodelling of the local sequence.

In combination, these observations suggest that short InDels are relatively abundant in natural protein evolution, and may act as drivers for structural and functional innovation (Grishin, 2001). Furthermore, there is an open question regarding the interplay between the roles of InDels and substitutions in protein evolution. Given the lack of comprehensive data on the effect of InDels, I started the projects described in this thesis in order to contribute towards filling this gap in experimental data.

1.1.2 Epistasis and the influence of starting point stability

It has been previously shown that increased protein stability promotes evolvability in proteins (Bloom et al., 2006). This observation follows from the notion that most proteins are only marginally stable , which is good enough for the proteins to perform their function under normal conditions. However, during the process of evolution, the new mutations often push the protein below the 'stability threshold', such that mutations conferring a functional advantage are not tolerated because of their impact on overall protein stability. This argument was originally made with respect to *thermodynamic* stability, but has since also been commonly used with respect to *kinetic* stability - that is, postulating that mutations can disrupt fragile protein folding funnels if the original folding pathways is not sufficiently robust.

Consequently, the fitness effect of a specific mutation often depends on the wider sequence background into which it is introduced. Mechanistically, this sequence dependence can be due to:

- Specific interactions: these occur when two residues are either in direct physical contact or otherwise specifically affect the function of each other. For example, if there is a salt bridge present in the protein, changing either residue first will be strongly deleterious - because that stabilizing interaction is removed - but mutating the second residue will have a lesser effect.
- General interactions: since many functionally beneficial mutations are also detrimental to overall protein stability, some epistasis occurs because of a build-up of destabilizing mutations. Overall destabilization can reach the point where any additional mutation (even only mildly destabilizing) will register as strongly deleterious, because the protein now falls under the overall stability cut-off.

The introduction of InDels into a protein structure is often deleterious if it disrupts the protein structure, which is one mechanistic explanation for why InDels much more commonly appear outside of defined structural elements in protein. The InDels may also be less tolerated if they cause a global stability issue for the protein.

At minimum, InDels require small local repositioning of the protein backbone, and potentially introduce large structural changes into the tertiary/quaternary structure. Therefore, the functional effect of these mutations depends on the protein scaffold. Some factors that may affect the tolerance of a particular protein scaffold to InDels are:

- Rigidity/flexibility of the starting structure in general and the region where the InDel is introduced specifically; flexible regions are more likely to be tolerant to InDels.

- Thermodynamic stability (ΔG of folding), typically measured with T_m as a proxy parameter (but see Chapter 5 for additional discussion). Generally, the more thermodynamically stable the fold of the protein is, the more mutations that introduce high energy, strained areas into the protein structure can be tolerated.
- Robustness of the folding pathway and the susceptibility of the protein to aggregation. If this becomes an issue, there may be a thermodynamically stable fold available for the protein, but the nascent peptide chain has little chance of getting there.

  For example, this effect has been shown in a lipoyl domain, where a two-residue deletion in a long surface loop caused misfolding of the protein. However, incubating the poorly folded protein at increased temperature overcame the kinetic barrier and restored the native form (Stott et al., 2009).
- The secondary structure elements: while β sheets and α helices are generally more sensitive to structural changes than unstructured regions, β sheets may be particularly susceptible to InDels of an odd length. These change the registry of the sheet and can have a large impact, especially if one side is buried in the core of the protein and the other side is exposed to the solvent. Such repositioning was demonstrated when a single residue was deleted in the B1 domain in protein G (O'Neil et al., 2000).

Protein can be stabilised through evolution or by computational design. Evolutionarily stabilised variants typically originate from thermophilic organisms, or by using variants that have been improved with directed evolution. Alternatively, starting point stability can be improved by computational re-design which introduces thermodynamically stabilising substitutions, although the effect on catalytic activity in enzymes can be unpredictable (Broom et al., 2017; Goldenzweig et al., 2016; Wijma et al., 2018).

Given this important role of protein stability in directed evolution, I also specifically probed the effect of some InDels on protein stability. When planning the high throughput experiments testing the effect of InDels, I planned to perform these in two protein backgrounds (i.e. starting points) with different stability. Then, the results can be compared between starting points to identify mutations that show different effects, and thus explore whether the effects are mediated through local or global interactions.

1.1.3 Laboratory campaigns showing adaptability of InDels

Despite the relative scarcity of methods for generating libraries of InDel, several examples of InDels substantially altering the position of the backbone and side chains have been demonstrated.

Short InDels have been shown to have a key role in controlling dimerization of homodimer proteins, indicating they are involved in determining the stability and specificity in formation of protein complexes (Hashimoto and Panchenko, 2010). Furthermore, it has been shown that the switch from tetrameric to monomeric structure, in members of the rapidly evolving and structurally diverse *o*-succinylbenzoate synthase family, was associated with increased divergence and accumulation of large InDels (Odokonyero et al., 2014).

Several case studies have shown the involvement of InDel in the change of enzymatic specificity and activity: examples are the reconstruction of the active site in an engineered metallo-β-lactamase (Park et al., 2006), the specificity switch in tRNA nucleotidyltransferases from a CCA-adding enzyme to CC-adding activity through deletion of a flexible loop (Neuenfeldt et al., 2008), and a contribution to functional divergence in the lactate *vs* malate dehydrogenases (Boucher et al., 2014). Similarly, the gradual introduction of InDels over time has contributed to functional divergence in the FMN-dependent nitroreductase enzyme family, where the presence or absence of medium-length insertions supports the clustering of the family subgroups (Akiva et al., 2017).

TEM-1 β-lactamase. The TEM-1 β-lactamase is a very well studied model enzyme, primarily because of the clinical significance of understanding the mechanisms of microbial resistance to penicillin and cephalosporin antibiotics (Palzkill, 2018). In the test tube, the activity of TEM-1 can be assayed with a simple growth assay at a range of antibiotic concentrations, and the activity of variants is typically described with the minimum inhibitory concentration of the antibiotic.

An early study of Ala insertions (1-3 residues) showed that insertions into an α-helix are accommodated either through looping out or by replacing other residues in the helix, while the interface between the target helix and the rest of the protein is maintained (Heinz et al., 1993). Another investigation tested the tolerance of shorts insertions in combination with randomization of two neighbouring residues in three TEM-1 loops, and found that insertion tolerance was not related to substitution mutational tolerance (Mathonet et al., 2006). One loop tolerated substitutions but not insertions, while the other two loops accepted a variety of inserted residues with preference for wild type residues at the ends.

One deep sequencing study reported on the individual tolerance of 98% of single amino acid deletions and of 78% random amino acid insertion variants, in a library constructed with inverse PCR without any adjacent substitutions. They observed that some InDels were tolerated (but not beneficial) in the N-terminal signal sequence and in the loops of the protein, yet no functionally beneficial variants were identified, corroborating that the majority of InDels are deleterious (Gonzalez et al., 2019).

1.1.4 Methods for generating InDel libraries for directed evolution

This section draws on ideas developed with Dr. Stéphane Emond and published in Emond et al. (2020), especially Supplementary Table S16 and discussion of it.

It has long been speculated that InDels could be a valuable tool for protein engineering, giving access to an untapped pool of structural variation (Shortle and Sondek, 1995). Given that the impact of an individual InDel is largely unpredictable, the use of random libraries of InDels at multiple positions, with different composition and length seems advisable. The reliability of gene randomization methods is essential for success in directed evolution experiments; the resulting InDel libraries could then be used in directed evolution campaigns as usual. A high-yielding library generation protocol should create a large number of variants, avoid bias in gene composition or type of variant introduced, and be technically straightforward. When it comes to amino acid substitutions, several approaches (e.g. error-prone PCR, site-saturation mutagenesis starting with synthetic oligonucleotides) have been developed that partially or fully meet these criteria and are widely used. However, until recently the methods for InDel library generation have been limited, and consequently the use of InDel libraries in protein engineering has been sparse.

The first protocol on record to create InDel libraries, random insertion-deletion (RID; Murakami et al. (2002)) succeeds in creating random InDels along the gene of interest. However, it relies on a complex protocol involving random cleavage of single stranded DNA, so that random substitutions are introduced unintentionally alongside the target mutations. Two other early methods, segmental mutagenesis (Pikkemaat and Janssen, 2002) and RAISE (Fujii et al., 2006), do not control for the length of the InDel and consequently produce libraries that primarily contain frame-shifted variants. In contrast, a codon-based protocol dubbed COBARDE (Osuna et al., 2004) generates a pool of multiple codon-based deletions with <5% frameshifts. In principle, this is an attractive method which could also be used to combine InDels with substitutions, but it has the practical disadvantage that it requires custom reprogramming of an oligonucleotide synthesizer to create mutagenic oligonucleotides.

Transposon-based methods

Multiple transposon-based protocols have been established for generation of random substitutions, deletions of various sizes and to a limited extent, insertions.

Deletions. The methodology for transposon-mediated mutagenesis with a mini-Mu transposon was first established by Jones (2005). The engineered transposon inserts into the target gene with weak sequence preference, which integrates the transposon and duplicates the five nucleotide target sequence. The transposon is excised with Type IIS restriction enzymes, which remove the transposon and four nucleotides in the gene on each side, which results in a -3 bp deletion library.

The Jones method has been extended to generate up to five codon deletions, by re-engineering the Mu transposon and using inverse PCR to control the deletion length (Liu et al., 2016). A transposon-based protocol that generates gene truncation variants has also been described (Morelli et al., 2017), and a commercial kit that introduces stop codons via transposon insertions is available (STOP Kit by ThermoFisher).

Substitutions. While substitution libraries are typically generated with error-prone PCR or with randomised oligonucleotides, transposon-mediated methods for introducing randomised codons (NNN, NNK or NDT) that build on the Jones deletion method have been described; the protocol for generating trinucleotide exchange libraries (TriNEx) is particularly noteworthy (Baldwin et al., 2008).

The TriNEx method has also been combined with an intein system for reading frame selection, which purges any cross-codon transposon insertions from the library, and extended to replace up to three codons at once (Daggett et al., 2009; Gerth et al., 2004; Liu and Cropp, 2012). This combination of random codon substitution(s) and intein selection creates an attractive platform for creation of balanced substitution libraries, but unfortunately suffers from a high proportion of off-target and frameshifted variants in the library (60% or more). Other more complex protocols have been proposed which can create longer substitutions (Kim et al., 2009).

Insertions. In contrast with substitutions and deletions, the available methods for introducing insertions are limited to PCR-based methods that target specific protein sites. The only reported transposon-mediated protocol, pentapeptide-scanning mutagenesis, creates insertions in random positions along the gene but only with fixed insertion size and sequence (Hallet et al., 1997; Hayes and Hallet, 2000). In fact, transposon-mediated insertion of a fixed pentapeptide sequence at random positions in the gene of interest has been developed into a commercial kit (Mutation Generation System Kit by ThermoFisher).

TRIAD

Building on the single triplet deletion protocol by Jones (2005) and the TriNEx protocol by Baldwin et al. (2008), Transposon-mediated Random Insertion And Deletion (TRIAD) mutagenesis was developed in the Hollfelder research group (Emond et al., 2020). The TRIAD protocol gives easy access to large, diverse InDel libraries, where each variant contains a single short InDels of one, two or three nucleotide triplets (± 3, 6 or 9 bp). The main advantage of this protocol is that it creates both insertions and deletions, with randomized insertions sequence, and all within a streamlined general protocol. The variants may be located between codon boundaries or span two adjacent codons, which therefore may create an InDel with an adjacent amino acid substitution.

The TRIAD protocol starts with a transposition reaction, which inserts an engineered mini-Mu transposon (TransDel or TransIns) into random positions of the target plasmid. These transposon insertion libraries are transferred into a fresh plasmid backbone (neither the target gene nor the plasmid backbone should contain any restriction sites used in the process). Next, the transposons are excised with restriction enzymes and new DNA cassettes introduced, which allow the creation of InDels of 3, 6 or 9 bp length. Thus, the protocol is technically straightforward, reliable and gives easy access to libraries of in-frame InDels at random positions and in the case of insertions, random composition.

The process for generating TRIAD libraries has similar practical steps as the TriNEx protocol for the creation of NNN substitution libraries, only the DNA cassette sequences differ. The TriNEx libraries are easily prepared in parallel with the TRIAD deletion libraries, by ligation of an NNN substitution casette instead of Del2/Del3 (Figure 1.2A).

The protocol has also been adapted for creation of libraries targeted to a smaller target region, and could in principle be extended to introduce longer insertions (+12 bp and more), although this would lead to library sizes above the typical screening capacity. Alternatively, if one wishes to focus on analysis of insertions without adjacent substitutions (between-codon insertions only), the insertions libraries could be created with NNK or NDT codons.

Exploring the diversity of TRIAD libraries. At the start of the projects presented in this thesis, the TRIAD method was available and the libraries had been prepared in two model systems, the green fluorescent protein (GFP) and phosphotriesterase (PTE). Low throughput screening performed by Dr. Stéphane Emond had shown that the libraries contain functional or possibly improved variants, but the full diversity of these libraries was still unexplored. Therefore, I started the project with the aim to couple high throughput screening

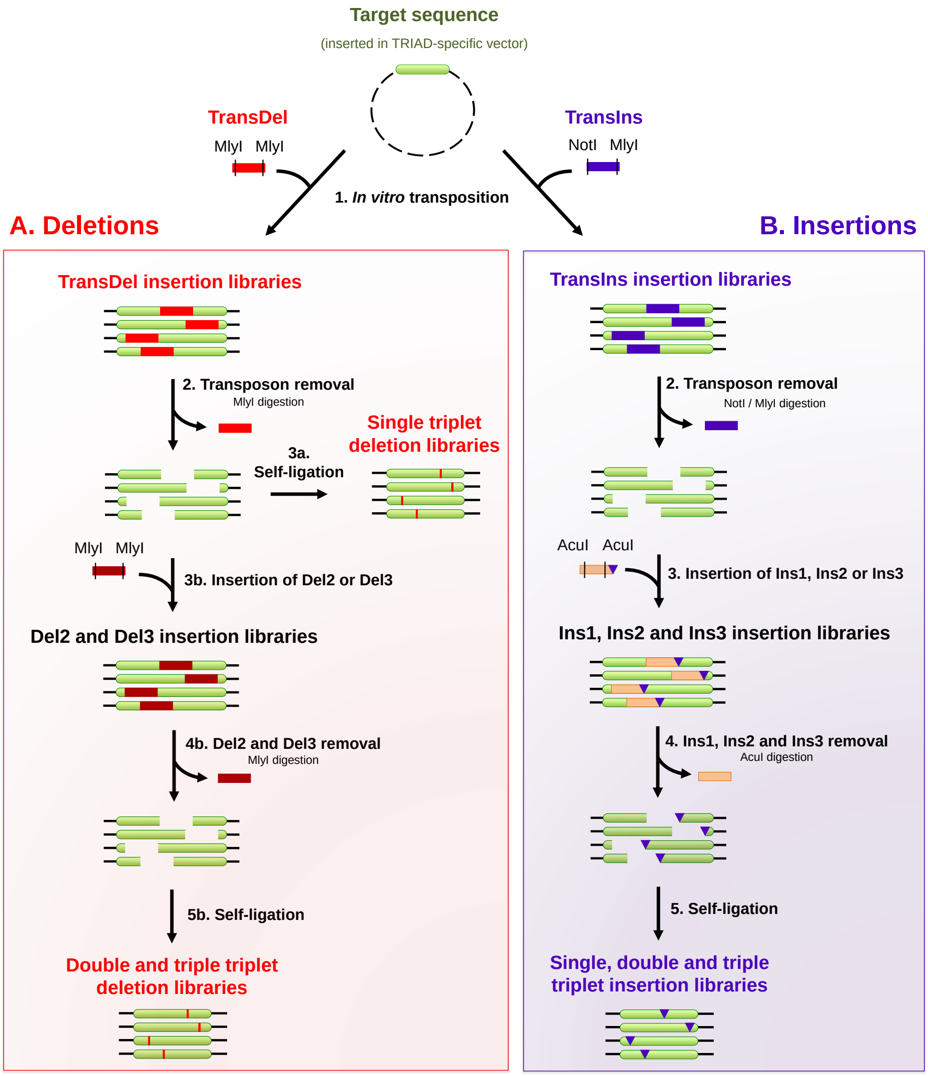

Target sequence
(inserted in TRIAD-specific vector)
TransDel
MlyI MlyI
TransIns
NotI MlyI
1. In vitro transposition
A. Deletions
B. Insertions
TransDel insertion libraries
2. Transposon removal
MlyI digestion
Single triplet
deletion libraries
3a.
Self-ligation
MlyI MlyI
3b. Insertion of Del2 or Del3
Del2 and Del3 insertion libraries
4b. Del2 and Del3 removal
MlyI digestion
5b. Self-ligation
Double and triple triplet
deletion libraries
TransIns insertion libraries
2. Transposon removal
NotI / MlyI digestion
AcuI AcuI
3. Insertion of Ins1, Ins2 or Ins3
Ins1, Ins2 and Ins3 insertion libraries
4. Ins1, Ins2 and Ins3 removal
AcuI digestion
5. Self-ligation
Single, double and triple
triplet insertion libraries

Figure 1.2 The TRIAD protocol work flow. Both insertion and deletion protocols require the target gene and the plasmid backbone to be free of MlyI, NotI and AcuI restriction sites. A) Deletion libraries. The TransDel transposon is inserted into the target sequence on a circular plasmid with an in-vitro transposition reaction, followed cloning of the gene with the transposon into a fresh plasmid backbone. TransDel is excised through MlyI digestion and re-ligation, with generates the -3 bp library (Jones, 2005). Alternatively, after TransDel removal the Del2 or Del3 casettes are ligated into the library. Finally, the cassettes are removed with MlyI digestion and the plasmids re-ligated, which gives -6 and -9 bp deletion libraries. B) Insertion libraries. After an analogous first steps with the TransIns transposon, the transposon sequence is removed with NotI and MlyI double digestion. Next, DNA casettes Ins1, Ins2 or Ins3, containing randomised NNN triplets, are ligated into the libraries, and NotI sticky-ends facilitate efficient ligation. In the last cloning step, AcuI digestion removes the cassette shuttle sequence (leaving behind randomised codons) and short digestion with the Klenow fragment removes 3'-overhangs. Finally, blunt-ended self-ligation yields libraries of single insertion variants with the randomised nucleotides in random positions. Reproduced from Emond et al. (2020).

and next-generation sequencing to deepen our insight into the effects of InDels on these two model proteins.

1.2 Recording protein fitness landscapes through deep mutational scanning (DMS)

1.2.1 The concept of fitness landscapes in protein evolution

The systematic relationship between sequence and function is captured by the idea of fitness landscapes, which was first introduced in the context of population genetics (Wright, 1932). It links protein sequence space, the collection of all possible gene or protein sequences, with a functional fitness score. In graphical representation of fitness landscapes (Figure 1.3), regions with higher elevation represent adaptive variants, and evolution moves from sequence to sequence in an uphill fashion. In natural evolution, the ’fitness’ reflects the ability of the organism to survive and reproduce in a given selective environment; in directed evolution, fitness is a function selected by the scientist. Even in directed evolution, protein fitness is a multi-trait property, defined as the product of multiple traits: protein stability and soluble expression concentration, a long lifetime in cell lysate, a high binding affinity to the target, good catalytic performance against the target substrate (which may or may not be a good model for the natural reaction), or perhaps high discrimination between multiple substrate or reaction pathways. Some of these component properties can be measured (k_{cat}/K_M, enzyme concentration, % of soluble protein compared to total amount).

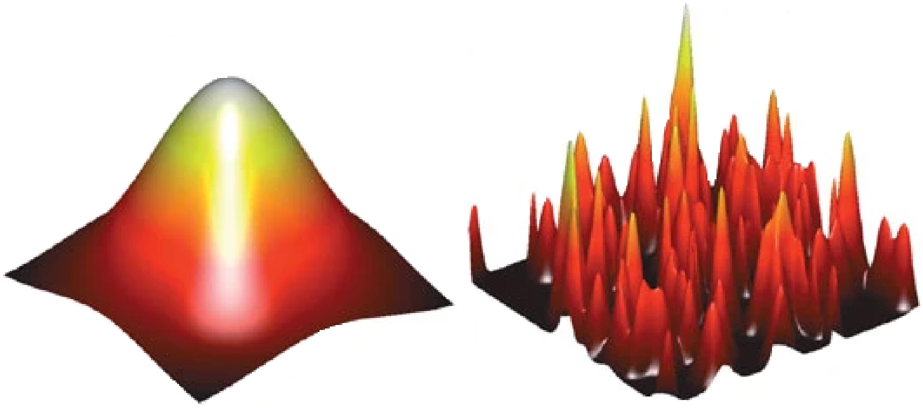


Figure 1.3 Smooth and rough fitness landscapes. In this representation, the horizontal plane represents available sequence space and the vertical dimension shows increasing fitness. A smooth landscape enables an easy transition towards improved variants, analogous to climbing a hill. A rough, rugged landscape has multiple fitness peaks (locally optimal sequences), but moving from one region of sequence space to another is difficult because of intervening valleys of non-functional sequences. Figure adapted from Romero and Arnold (2009).

It follows that with a folded, functional protein as a starting point, a search for interesting variants is therefore a walk in sequence space, constrained by the requirement that all intermediates are functional (Maynard Smith, 1970). Some rare areas of sequence space are home to a cluster of highly functional variants, which through a set of mutations may turn to inactive variants. This concept is often represented as metaphorical hills and valleys in sequence space (as in Figure 1.3). It should be noted that although fitness landscapes are often shown as connected hills and valleys, the true mutational network is formed in discrete spaces (since a single position cannot be half one and half another amino acid), and is highly multi-dimensional.

Multiple approaches to studying fitness landscapes have been described. One approach derives from the analysis of natural sequence variation between protein homologs in different species, either by examining co-variation between co-evolving positions (Teşileanu et al., 2015) or through 'looking back in time' with ancestral sequence reconstruction (Bridgham, 2006; Harms and Thornton, 2010; Zakas et al., 2016). The evolution trajectories thus identified show broad (i.e. far from the evolutionary starting point) coverage of the fitness landscape, since the protein had thousands to millions of years to roam in sequence space. However, the coverage is simultaneously quite shallow: perhaps evolution could have taken multiple different trajectories, but all we observe are those that survived to give extant proteins.

Examination of experimental fitness landscapes through directed evolution allows the possibility of addressing some of these questions (Poelwijk et al., 2007; Romero and Arnold, 2009). Is there a multitude of possible evolutionary paths across the same landscape, signifying the adaptability of the protein, or is the landscape full of functional 'cliffs' leading to non-functional proteins? Are the landscapes rugged, containing a multitude of medium-fitness peaks that represent evolutionary dead-ends, or are they smooth with easy paths towards improved function? How epistatic are they, and how important is the order of mutations in functional transitions? Are the landscapes of related protein similar, such that identical key mutations can jump-start functional innovation, or is each protein an island in sequence space with unique properties?

1.2.2 Recording local fitness landscapes with DMS

Deep mutational scanning describes a class of studies that experimentally describe the *local* fitness landscape of the protein of interest. In contrast with study of evolutionary pathways, the coverage of the fitness landscape is *narrow* - the datasets typically sample variants with one to three mutations compared to parental gene sequence - and *deep*, such that a substantial proportion of all possible variants is experimentally tested (50% or more).

A set of detailed protocols for DMS experiments have been published (Starita and Fields, 2015a,b,c). Here I briefly summarise the steps involved these experiments and some considerations for experimental design.

Library randomisation

First, the target gene, which may encode the whole protein or a specific region of interest, is randomised to introduce the desired number of substitutions into the gene. The use of a variety of library constructions methods is possible; whole gene randomisation is often performed with error-prone PCR, while small regions can be randomised in a more controlled way (i.e. creating a balanced mix of different substitutions) with methods utilising mutagenic oligonucleotides.

The size of the region that should be randomised is constrained by size of the protein - small proteins can be randomised in their entirety while keeping the final library size moderate - and the maximum length of DNA that can be sequenced. Currently, the longest read length achieved by Illumina MiSeq is 2×300 bp, so the maximum length that can be sequenced with a single read per base is 550-600 bp. Depending on experimental design, some of the read length might be used on technical sequence such as UMIs (barcodes), which increase the accuracy of sequence but take up a proportion of available sequencing length.

The choice of the randomisation method also controls the scope of the recorded fitness landscape. Ideally, the mutations should be spread equally across the randomised region and no mutations should occur outside the randomised region, to avoid confounding effects. The size of the libraries needs to be adjusted to the number of variants that can be experimentally screened, as well as sequenced within the planned type and number of sequencing runs. The library size is controlled either through library design or with a restrictive DNA transformation that reduces the number of variants in the library.

Fitness screening

The second step is an experimental screen of the library of variants for the desired protein function. This can take the form of a growth competition assay, a display experiment to enrich for binding, or an assay for catalytic function. The experimenter has a free choice of the experimental platform, as long as it meets three conditions: high throughput, maintenance of a genotype-phenotype linkage, and the ability to recover variant DNA after the screening.

A high throughput is essential, since good coverage of even a small DMS library requires testing upwards of 10^4 variants.

The majority of DMS studies utilise growth selection experiments (Deng et al., 2012; Hietpas et al., 2011; Melnikov et al., 2014; Roscoe et al., 2013) and selections based on detecting binding using phage display (Fowler et al., 2010; Starita et al., 2013). More recently, a fitness landscape of a TIM-barrel glycosidase enzyme has been recorded using a microfluidic assay (Romero et al., 2015). The authors successfully quantified sequence preferences for 31% possible amino acid substitutions and showed that the β-sheet core of the protein was easily disrupted, while mutations were largely tolerated in the outer helices, especially on the outside surfaces. While this experiment had drawbacks in library design - it sorted an error-prone PCR library with an average of 3.8 mutations per gene, yet only quantified single site preferences - it was a notable achievement from an experimental point of view.

Sequencing

The DNA encoding the randomised region is recovered from both the input library and the sorted fractions, and sequenced with a NGS protocol. Illumina is the predominant current technology for sequencing in DMS experiments because it allows the sequencing of $> 10^5$ variants per sequencing run at single base resolution, which is not quite achievable with long read PacBio or Nanopore sequencing (yet). The technical requirements for Illumina sequencing are typically the controlling factor in DMS design; the sequencing generates millions of reads with information about the variants, which allows for the depth of the experiment, while being limited by the length of Illumina sequencing reads.

From the start, DMS experiment scope has been guided by the length of the gene that can be sequenced in a single sequencing run (single read or paired-end reads), and much technical improvement in the use of DMS has derived from improved strategies to optimally exploit Illumina sequencing capacity. The pioneering deep mutational scanning experiment used phage display to describe the fitness landscape of a 50-amino acid WW domain binding to its peptide ligand (Fowler et al., 2010). They randomised a 33-residue region in the middle of the domain and sequenced 25 residues within that region, which was the maximum available sequencing length at the time (76 bp). Currently, the maximum length of Illumina sequencing is 2×300 bp in paired sequencing, and soon the length limitation may be further lifted by the use of Nanopore or PacBio long-read sequencing.

NGS data analysis

The exact manner of analysis depends on the fitness assay, and on whether only the active variants are sequenced (as is inherent in a growth selection) or whether the dataset comprises both active and inactive variants. The resulting description of the fitness landscape typically includes a heat map, which described the tolerated amino acid substitutions at each randomised position, an examination of epistasis between key positions (if the sequenced region is long enough), and comparison of tolerated mutations to variants that occur in natural evolution.

Since DMS is still a relatively new approach, the data analysis approach is not yet fully standardised. Furthermore, now that new datasets are published increasingly often, cross-analysis is becoming possible. One cross-comparison of 16 DMS datasets, collected on 14 proteins that were tested for a variety of functions, examines the average tolerance of substitutions in general, and in different structural contexts (Gray et al., 2017); the selected datasets were curated to examine only single amino acid substitutions, and re-scaled to express fitness changes on a common magnitude scale. The authors observe that of the 20 possible amino acid substitutions, the introduction of Pro was overwhelmingly more disruptive and other residues, while Met, Val, Ala, Ser, Thr and Cys were considerably better tolerated than the median effect of all mutations. In line with expectations, they observed that substitution in turns/loops more likely to be accepted than in α-helices or β-sheets.

1.2.3 Functional assays in deep mutational scanning

Since the first reported studies, the local substitution fitness landscapes of a variety of proteins have been recorded with DMS, using a range of functional assays. The tested fitness functions span GFP fluorescence (Sarkisyan et al. (2016), discussed in Section 1.3.2), chaperone activity, antibiotic resistance (Firnberg et al., 2014; Melnikov et al., 2014), transcription factor activity (Firnberg et al., 2014), selectivity in ligand binding (Raman et al., 2016), kinase activity (Brenan et al., 2016) and others (see Gray et al. (2017) and references therein for a selection).

Overview of current literature on DMS shows that success is possible with a number of functional assays, as long as the chosen format reflects the desired fitness function and maintains a genotype-phenotype linkage. While sophisticated statistical analysis can increase the understanding of a given dataset, ultimately the results of a DMS experiment are only as good as the resolution afforded by the experimental assay (Starita and Fields, 2015c).

Experimental formats for ultra-high-throughput screening of enzymatic activity

Successful DMS studies require the use of high-throughput platforms for fitness screening, typically recording data on $10^4 - 10^5$ individual variants and screening up to 10× more. Because of the requirement for this throughput level, classic screening technologies such as agar plate colony screen or microtitre-plate screening (even with the use of robotics) are not suitable. Instead, the available ultra-high-throughput methodologies fall into two group: screening and selection. Both maintain a link between the variants' genotype and phenotype in some fashion, but they differ in the fate of non-selected variants.

Selection methods create an environment of fitness competition between variants, typically based on growth or biophysical properties, such that variants below the fitness threshold are eliminated from the system. Selection methods include survival assays / growth competition assays (e.g. selection for growth in liquid culture medium at increasing concentration of an antibiotic; MacBeath et al. (1998); Reetz et al. (2008)) and protein display methods: phage display (Pande et al., 2010), yeast display (Cherf and Cochran, 2015), ribosome display (Yanagida et al., 2010; Zahnd et al., 2007) and bead-surface display (Diamante et al., 2013; Mankowska et al., 2016) are popular options in protein engineering. The advantage of selection methods lies in the extremely large number of variants that can be tested (up to 10^{11} in ribosome display). However, they are only applicable to proteins encoding a selectable trait (e.g. essential metabolic enzymes) or proteins binders amenable to surface display.

In contrast with selections, screening methods test all variants, which reduces the throughput but they have the advantage of enabling the search for more complex protein activities and the discovery of improved catalytic efficiency. Still, modern ultra-high-throughput screening methods achieve a good total throughput per experimental day ($< 10^7$ variants in a well-optimised system). There are two main screening methods in use at the moment: fluorescence-activated cell sorting (FACS) and in-vitro compartmentalisation through droplet microfluidics.

FACS. During a FACS experiment, live single cells are surrounded with sheath fluid (PBS buffer) and injected through a small nozzle at high speed, which enabled controlled separation of the sheath fluid into regularly spaced droplets. Each droplet is illuminated with up to 4 lasers and fluorescence signals are detected in up to 16 channels. Depending on the signal, the target droplets are charged with an appropriate level of positive or negative electric charge. The droplets (and the cells inside) pass magnetic plates with fixed field strength and are separated with high precision into 2-4 collection tubes. FACS instruments are very well engineered, have the advantage of multi-colour sorting and good throughput (Yang and

Withers, 2009). However, the use of FACS is restricted to compatible biochemical systems; this primarily encompasses live cells where the fluorogenic moiety is retained within the cell, or incorporation of elements of protein display on the cell surface.

Microfluidic droplet sorting. Droplet microfluidics refers to a screening technology that creates millions of separate reaction compartments in less than 100 μL combined sample volume, by controlled mixing an aqueous solution with oil in the presence of a stabilising surfactant. In this manner, each individual cell or bead expressing a single library variant is separated into its own droplet. Once inside a droplet, the cells can be lysed to bring the expressed enzyme in contact with the chemical substrate present in the solution, while the droplet maintains the genotype-phenotype linkage.

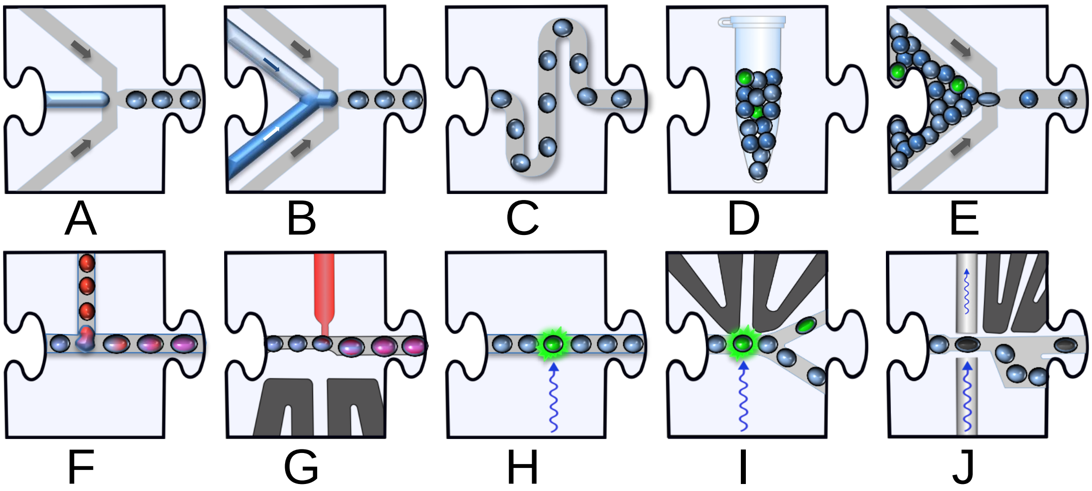


Figure 1.4 A selection of available microfluidic modular devices. A) Single-inlet droplet generation, B) double-inlet generation with mixing of two aqueous solutions, C) on-chip incubation in a delay line, D) off-chip incubation, E) droplet re-injection and spacing, F) droplet fusion, G) pico-injection, H) fluorescence detection, I) fluorescence-activated droplet sorting (FADS), J) absorbance-activated droplet sorting (AADS). Figure adapted from Kintses et al. (2012) with additional panels by Dr. Liisa van Vliet.

Microfluidic droplet technology is a modular platform that uses separate patterned ’chip’ devices for individual steps in an assay workflow: cell-encapsulation / droplet creation, incubation, pico-injection of additional reagents, droplet merging, droplet splitting and signal detection (Figure 1.4). While the use of microfluidics requires some specialised equipment (access to a clean room for master fabrication, fluid pumps, sorting rigs) and knowledge (device design, operation of sensitive chips), the field is growing in popularity as the individual components become more reliable for everyday operation. The modularity is a clear advantage of the technology, allowing fast adaptation of existing expertise to new assays and applications.

A breakthrough in high-throughput screening with microfluidic droplets came with the design of a fluorescence-activated droplet sorting (FADS) chip (Baret et al., 2009; Sciambi and Abate, 2014). Detection of a fluorescent read-out benefits from high sensitivity and low background in negative droplets, so it is not surprising that FADS has since been widely adopted in the microfluidic community (Colin et al., 2015; Guo et al., 2012; Kintses et al., 2012; Shembekar et al., 2016; Zinchenko et al., 2014).

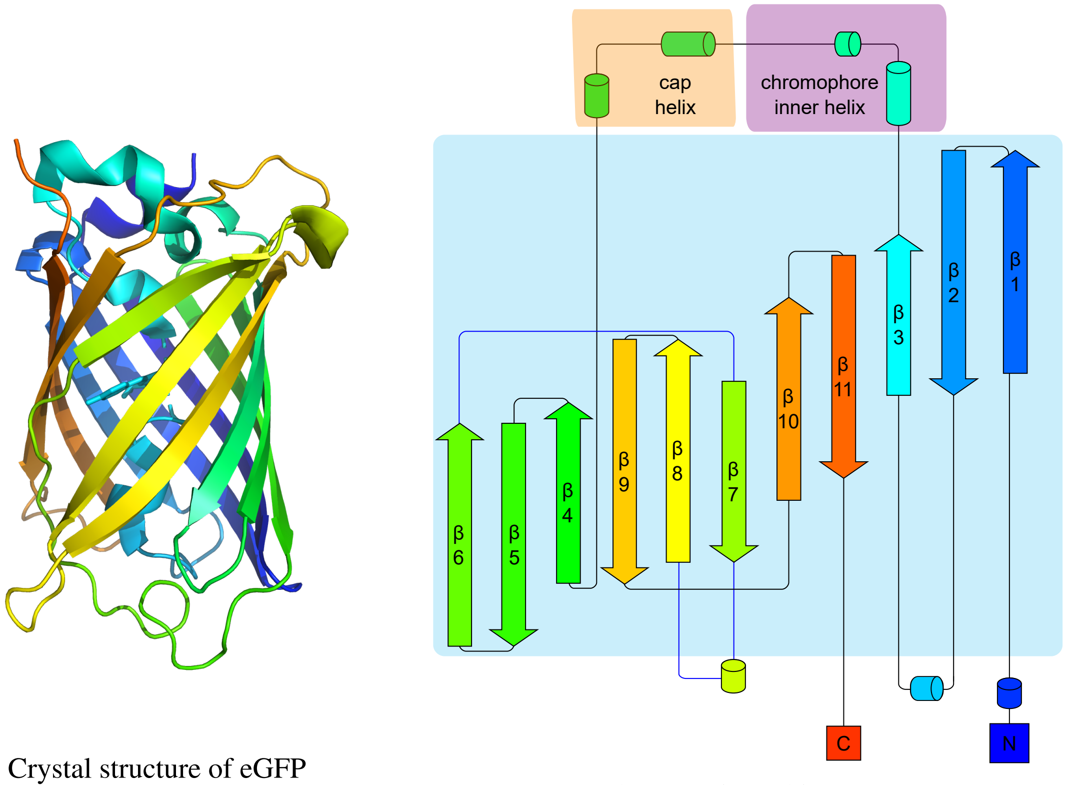
The main drawback of FADS - the requirement for a fluorogenic substrate, which may bias the outcome of the selection compared to a natural substrate - can be avoided with other sorting methods. The most useful is absorbance-activated droplet sorting (AADS), which measures the change in absorbance of a chromogenic substrate, co-factor (such as NADH) or coupled reaction product (Gielen et al., 2016).

1.3 Exploring the effect on InDels on structure: GFP

Looking forward to implementing DMS with InDel libraries, I first wished to set-up the DMS protocol on a protein with an established experimental assay. I chose the green fluorescent protein (GFP), which is a popular model system for testing new mutagenesis methods. Assaying the fitness of GFP is experimentally simple: the fitness is defined as the fluorescence intensity under certain experimental conditions, and the genotype-phenotype linkage is maintained by the cells - they carry both the plasmid DNA and the protein expressed in the cytosol.

GFP was first discovered in *Aequorea* jellyfish as a side product during purification of the chemiluminescent protein aequorin (for a comprehensive review, see Tsien (1998). It contains 238 amino acid residues and folds into a so-called "β-can": 11 anti-parallel strands form a barrel and an α-helix threads through the centre (see Figure 1.5). The fluorophore forms auto-catalytically from three amino acids at positions 65-67, which have the sequence SYG in the native *A. victoria* jellyfish, or TYG in commonly used engineered variants.

GFP has found widespread use as an expression marker in bacteria, yeast and mammalian cells, because it does not need an external cofactor or enzymatic partner to achieve fluorescence. It does, however, have some limitations: without additional mutations, expression of the properly folded protein is reduced at temperatures above 20°C, which is not surprising given that source jellyfish live in cold water. However, once folded, the protein exhibits



(a) Crystal structure of eGFP

(b) Domain architecture

Figure 1.5 a) A ribbon diagram showing the crystal structure of eGFP (PDB structure 4EUL). The N-terminus is shown in blue and the C-terminus in red. The central helix that threads through the β-barrel is shown in light blue and is followed by an additional helical structure on top of the helix 'cap'. b) A diagram of domain architecture in GFP, showing connectivity and relative alignment of β-sheets in the barrel. The central helix and the cap helix are highlighted in at the top of the diagram. (The exact number of helices in such representations depends on the algorithm used to extract secondary structure elements from crystallographic coordinates.)

excellent stability and fluorescence has been observed even when precipitated in inclusion bodies in *E. coli* (García-Fruitós et al., 2005).

Due to its popularity, numerous GFP variants have been developed that show improved folding, a different spectrum profile or both. One particularly popular variant is enhanced GFP (eGFP), which was the first sufficiently stable GFP variant that enabled use as marker for protein expression in mammalian cell culture (Heim et al., 1995).

1.3.1 A model for mutagenesis studies

The first advantage of GFP is that it does not require any cofactors to achieve fluorescence, hence it is amenable to various experimental conditions and host species. Generally it shows strong and robust levels of protein expression. Still, it should be noted that the maturation of the chromophore includes an air oxidation step that generates one equivalent of H_2O_2, which does place the host organism under some oxidative stress. It appears one of the intermediates formed during GFP maturation catalyses the reduction of atmospheric oxygen to generate superoxide radical anion $O_2^{\cdot-}$ with stoichiometric use of NADPH (Ganini et al., 2017). Nevertheless, eGFP is regularly well expressed in bacteria, yeast, mammalian cell lines and higher organisms.

Second, the excitation maximum of eGFP is at 488 nm, which corresponds exactly with the readily available 488 nm blue laser. Consequently, GFP fluorescence can be observed by naked eye under a blue light / orange filter combination, in a plate fluorimeter, under a basic fluorescent microscope and in every flow cytometry device (which typically always use the 488 nm laser for forward scatter / side scatter measurements).

Finally, the use of GFP as a model system for mutagenesis methods is relatively straightforward: in this case, "protein fitness" is defined as "fluorescence intensity at specified excitation/emission wavelengths" and is comparable to a growth assay in term of experimental complexity. In contrast, to measure enzymatic activity, specialised assay development is often needed: one must identify a suitable substrate, detection method, coupled reagents (to produce a fluorescent, luminescent or highly absorbent product), as well as find suitable experimental conditions that give a good signal. In GFP, bacteria brightness is a good proxy for GFP "fitness", although the exact brightness depends on:

- Total expression level,
- Proportion of well folded and matured fluorescent protein,
- Efficiency of excitation, energy transfer and fluorescent emission,
- Absence of fluorophore quenching,

- Resistance to photobleaching,
- Susceptibility of the protein to degradation.

Thus, in GFP fitness is a composite measure - amongst these processes, efficient protein folding is the most variable component.

1.3.2 A high-throughput substitution fitness landscape in GFP

Advances in next-generation sequencing read length and increasing maturity of deep mutational scanning have led to recent publication of a comprehensive local fitness landscape in *av*GFP by the Kondrashov lab (Sarkisyan et al., 2016). This study screened a library of 5.0×10^4 variants generated through error-prone PCR with a relatively high mutational load (average 3.7 mutations per genes). The high theoretical combinatorial complexity resulted in relatively good coverage of single point mutations, where 90% of protein sequences that are accessible through single nucleotide substitutions were characterised, but coverage fell precipitously for double and higher order mutants. As is typically observed with GFP mutagenesis studies, the observed distribution of fluorescence was bimodal, with most variants exhibiting either wild-type-like fluorescence or none at all.

The effect of up to three random amino acid substitutions was generally mild, with only 9.4% of single mutants exhibiting more than a five-fold decrease in fluorescence. Point mutations on the surface of GFP were almost all neutral, while sites with the side chains oriented towards the chromophore were largely intolerant of substitutions (Figure 1.6). These intolerant sites are also more conserved between GFP orthologs and likely to not accept amino acid deletions.

The study design allowed the authors to obtain full-length gene sequences through an UMI-guided sequencing approach that incorporated several key endonuclease restriction sites in the starting gene, thereby allowing investigation of epistatic interactions between distant sites. The avGFP fitness landscape shows a strong presence of epistasis; the tolerance of mutations at a given position is strongly correlated with evolutionary conservation, yet mutations to residues present in orthologs were often non-fluorescent. Across the whole dataset, negative epistasis (where a double mutant exhibits reduced fluorescence compared to the sum of the individual effects) was common while positive epistasis (the double mutant appearing more fluorescent than expected) was rare. Epistatic effects were observed between 96% of variants with weak individual effect, supporting the conclusion that epistasis is common in the compact, marginally stable structure of *av*GFP.

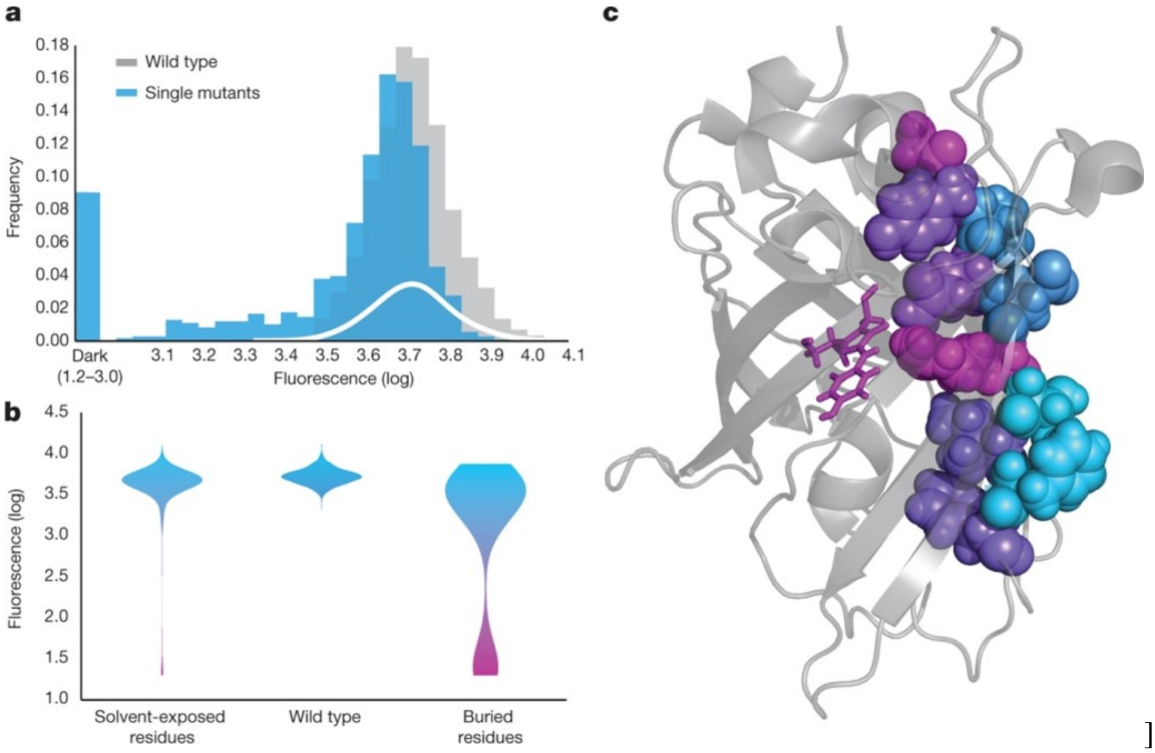


Figure 1.6 The distribution of fluorescence brightness in variants with a single amino acid substitution. a) The fluorescence brightness for 1,114 single missense mutations (blue), compared to 2,442 independently measured wild-type sequences (grey). Most single mutants retain strong fluorescence with similar distribution and mildly decreased brightness compared to wild type. b) The impact of single substitutions differs depending on solvation status of the affected residue, with buried residues showing lower average fluorescence. Almost all non-fluorescent single mutants carry mutations affecting buried residues. c) Illustration of the effect of residue orientation on mutation susceptibility on one β-strand. Figure reproduced from Sarkisyan et al. (2016).

1.3.3 Low-throughput studies on the effect of InDels in GFP

In contrast to substitutions, the effect of InDels on GFP has not been comprehensively explored, but smaller scale studies show that the compact β-can structure shows some tolerance towards InDels. An early observation was that GFP retains fluorescence with an up to 9 amino acid C-terminal truncation, while N-terminal, internal or longer C-terminal deletions abolish fluorescence (Kim and Kaang, 1998). A medium scale study of single triplet deletions in eGFP was demonstrated during transposon-mediated single triplet deletion mutagenesis (from which TRIAD derives, (Jones, 2005)) by Jones and others (Arpino et al., 2014a). They performed a colony screen in eGFP, where they showed that 10% of the colonies displayed green fluorescence when grown at room temperature and 2.5% when grown at 37°C. Sequence analysis of active variants showed that tolerated trinucleotide deletions were primarily clustered in loop (61% of variants) compared to helical (19%) and β-sheet (21%) regions. Surprisingly, they identified the variant G4Δ, which exhibited improved cellular fluorescence while the fluorescence properties of the protein remained similar. The improved folding and re-folding properties of this variant were attributed to a shift in the hydrogen bonding network at the N-terminal 3-10 cap helix, which may improve the transition from the folding intermediate to mature protein.

Two further well-tolerated trinucleotide deletions have also been characterised in detail (Arpino et al., 2014b): the buried residue A227Δ in the C-terminus of β-strand S11, and solvent-exposed D190Δ in the loop between S9 and S10. D190Δ is located in a 10-residue loop and the effect of the deletion is restricted to neighbouring residues, such that the positioning of β-strands before and after the loop is essentially unaffected. In contrast, the D190Δ deletion creates only small changes in the strand S11 in which it is located, but introduces alternative side chain conformations in nearby strands S7, S8 and S10 and also greatly increases their flexibility.

A similar screening campaign in a UV-optimised uvGFP variant, using trinucleotide deletion mutagenesis followed by colony screening, identified thirteen unique variants that retain fluorescence (Liu et al., 2015), eight of which were located in loops. The tolerated deletions showed impaired cellular fluorescence with increasing temperature of expression, often surpassing uvGFP fluorescence at 20°C but falling sharply at 30°C and 37°C. The temperature sensitivity was correlated to a decrease in soluble protein expression, which indicated impaired protein folding as well as impaired chromophore maturation. The introduction of folding-enhancing mutation F64L (one of two mutations present in eGFP) partially rescued fluorescence in all five internal single deletions at 37°C, as well as rescuing double deletion mutants at 20°C.

1.4 Examining the effect of InDels on catalysis: phosphotriesterase

Phosphotriesterase as a model enzyme

In this part of the project, I aimed to examine how short random InDels affect the catalytic activity of a model enzyme. There are multiple model enzymes that have been used in literature for fitness landscape studies, with the choice guided by the experimenter's familiarity with the enzyme and the availability of suitable activity assays. I chose to work with the *Brevidomonas diminuta* enzyme phosphotriesterase (PTE), which is a recently evolved enzyme in the amidohydrolase superfamily. It shows a very high phosphotriesterase activity towards paraoxon - approaching the diffusion limit (Briseño-Roa et al., 2011; Caldwell et al., 1991) - and additionally a promiscuous arylesterase hydrolase activity towards various substrates (see Figure 1.7). It has been hypothesised that the enzyme provides a selective advantage to host organisms in environments contaminated with the parathion and paraoxon pesticides, providing them with phosphorus through their degradation. Thanks to PTE's evolutionary novelty, the high catalytic efficiency and promiscuity both in choice of substrate and mechanism, PTE has become an established model enzyme in the field of directed evolution (Griffiths and Tawfik, 2003).

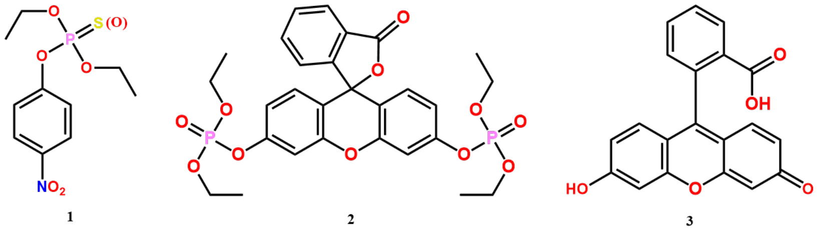


Figure 1.7 The structure of **1** - parathion and paraoxon, **2** - fluorescein diethyl phosphotriester and **3** - fluorescein (shown as the protonated structure present at acidic pH)

Structurally, PTE is a dimeric α-helical protein in the TIM-barrel fold that binds two Zn^{2+} ions per subunit in the form of a binuclear centre in the active site (Figure 1.8). Denoting the two Zn^{2+} ions as α and β, the former is more buried and coordinated with His55, His57 and Asp301, while the β Zn^{2+} is more solvent-exposed and coordinates His201, His230 and when present, the substrate through the P=O double bond. A molecule of water is coordinated between α and β, and acts as the nucleophile in the attack on the coordinated and polarised substrate, to give a pentavalent coordinated intermediate in an associative nucleophilic substitution mechanism. In the next step, the enzyme facilitates a proton chain that eliminates a leaving group and leads to dissociation of all reaction products, thus regenerating the enzyme (Aubert et al., 2004).

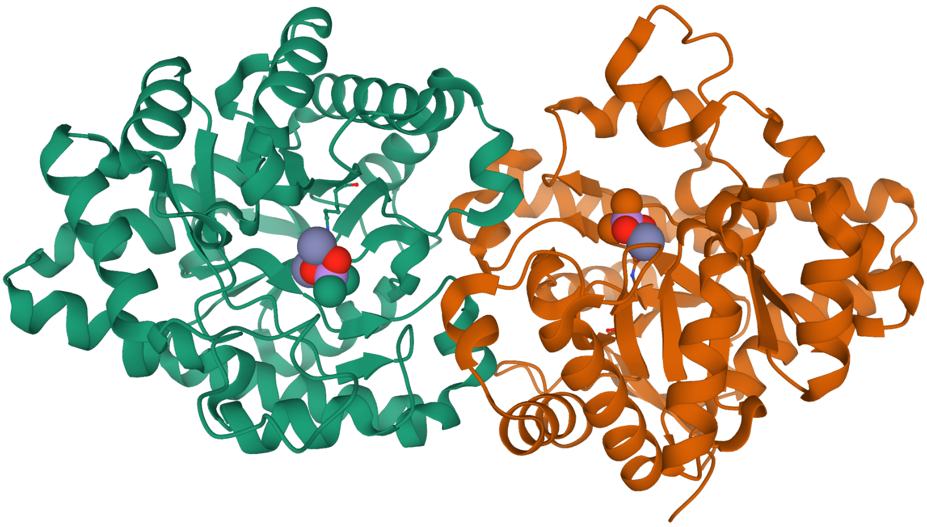


Figure 1.8 A cartoon illustration of the dimeric phosphotriesterase crystal structure. Active Zn^{2+} residues are shown as grey spheres, the coordinated cacodylate ions as coloured spheres and the amino acid residues that coordinate Zn^{2+} ions are depicted as ball and stick models. Drawn from PDB structure with accession code 4PCP.

Functionally, PTE essentially utilises metal-mediated nucleophilic activation, substrate activation and proximity based rate acceleration. These mechanism features are similar between ester bond hydrolysis and phosphotriester bond hydrolysis, so it is perhaps not surprising that PTE also exhibits promiscuous esterase activity towards arylester substrates, and some lactonase activity (Roodveldt and Tawfik, 2005). These substrates show similar structural features as paraoxon through a flat, aromatic core and the polar carbonyl bond which is susceptible to nucleophilic attack. The promiscuous arylesterase activity which has been explored in a directed evolution campaign which achieved a full functional switch to a highly effective arylesterase (Tokuriki et al., 2012) over 22 rounds of directed evolution, which utilised error-prone PCR and DNA shuffling. In a later campaign (Kaltenbach et al., 2015), the functional evolution was reversed back to a highly active paraoxonase. However, comparison of these two evolution pathways reveals that the fitness landscape of PTE is epistatic and the effect of individual mutations is highly contingent on the background.

Remodelling of phosphotriesterases. Though the PTE ezymes isolated from different bacterial species are very closely related, their sequence identity with their closest relatives, the PTE-like lactonases (PLLs), is only 30%, which is too low to signify a direct evolutionary ancestry between the two families. However, the PTEs do show some lactonase activity and some PLLs exhibit low level promiscuous phosphotriesterase catalytic activity, which supports the premise that PTEs evolved from some form of lactonases (Afriat et al., 2006). One PLL has been successfully converted into a phosphotriesterase with incremental expansion of the specificity-controlling loop 7 with randomised residues(Hoque et al., 2017). Additional evidence for this hypothesis is the fact that remodelling of PTE loop 7 into a PLL-like state, which used deletions of five and nine amino acids, and one substitution, restored this presumed ancestral-like catalytic activity (Afriat-Jurnou et al., 2012).

In summary, PTE is a recently evolved, highly active enzyme that has been near "evolutionarily optimised" for degradation of paraoxon, and also exhibits multiple promiscuous activities on ester and lactone substrates. These features make it a popular model system in the field of protein evolution and underpin the choice of PTE as a first model enzyme when exploring the effects of short InDels on enzymatic structure and function.

1.4.1 The effect of InDels on PTE catalytic activity

Following this choice of PTE as a model system, six TRIAD libraries were prepared in *wt*PTE and evaluated for their native paraoxonase and promiscuous arylesterase activity (Emond et al., 2020). The TRIAD mutagenesis methods introduces random short in-frame InDels at random positions along the target gene, with maximum diversity of hundreds of deletion variants and between 5×10^4 and $> 10^5$ insertion variants per library. While the large diversity of insertion variants necessitates a high-throughput screening method, insight into the effect of InDels is still available using medium throughput methods.

In that study, the main screening campaign was conducted by Dr. Stéphane Emond and done in two arms, one examining the native paraoxonase activity and the other searching for improved promiscuous arylesterase activity. All six TRIAD libraries (3, 6 or 9 bp deletions or insertion) were screened, as well as trinucleotide substitution libraries prepared according to TriNEx procol to serve as a control. A plate-based screen on >800 variants for phosphotriesterase activity revealed that on average, InDels were more deleterious than comparable substitutions, while also revealing some variants with >1.5-fold improved activity - compared to none in 342 screened substitution. In a screen for arylesterase activity, the InDel libraries again showed a greater proportion of hits with and the adaptive variants showed different functional profile; the top hits achieve the functional improvement through k_{cat} rather than K_M, which suggests a novel adaptive pathway.

1.5 Core project outline: applying deep mutational scanning to insertions and deletions

1.5.1 Experimentally recording protein fitness landscapes

The previous sections describe how fitness landscapes have been discussed in the context of protein evolution, but actual experimental data are still very limited in enzyme evolution in general, and with respect to the function of InDels specifically. In this project, I therefore set out to experimentally record fitness landscapes of InDels in two model proteins and generate datasets that can be used to relate protein sequences to functions. Specifically, I wished to synthesise the principles developed in deep mutational scanning of substitutions, the still under-explored potential of InDels in protein evolution, and high throughput screening technologies.

Library composition. The starting point for DMS were the already available TRIAD libraries in the green fluorescent protein and phosphotriesterase, both established model proteins. The libraries had already been screened to some extent to find variants that are wild-type like or show improved fitness, but had not been fully explored. The first question I address in this thesis is the open-ended query of library composition: *what variants are found in these libraries? Where in the target genes are the mutations located and how does the distribution depend on the target sequence? How good is the quality of the libraries?*

Establishing the experimental assays. Compared to substitutions, InDels are generally more deleterious to protein stability and activity, yet multiple adaptive InDels have also been identified (Hoque et al., 2017; Simm et al., 2007). In order to fully explore the potential of TRIAD InDel libraries, high-througput screening is therefore essential - and it is also integral to generating a local fitness landscape with good coverage. GFP is relatively readily amenable to screening by sorting the bacteria with FACS, followed by an analysis the output of the sorting with flow cytometry and testing the fluorescence of randomly selected variants.

Phosphotriesterase cannot be screened with FACS in the same way, but requires a different experimental approach. A DMS experiment on an enzyme has been demonstrated by using a microfluidic assay (Romero et al., 2015), which was an inspiration for using a similar approach with PTEthe However, PTE is a very fast enzyme, so the first open question was whether it is in fact compatible with a microfluidic assay, and assay development.

Here, the enzyme fitness is defined as the product concentration in each microfluidic droplet, which is measured through the intensity of fluorescence when excited with a 488nm laser. The amount of turned-over substrate is a reflection of soluble enzyme concentration, protein folding and catalytic efficiency.

Deep sequencing. Overlapping with development of the assay, I worked on sorting the libraries and improving the deep sequencing strategy. Because of the large library size, I decided to primarily use Illumina sequencing, which has the highest available throughput and moderate maximum read length (up to 2×300 bp). The efficiency of sequencing projects can be improved with the use of unique molecular identifiers (UMIs; also dubbed barcodes), which improves the quality of sequencing and reduces the amount (and thereby cost) required. Therefore, from the beginning of the project I explored ways in which UMIs could be incorporated into the project design. These considerations were addressed in parallel with choosing the deep sequencing strategy.

Variant detection with InDelScanner. Deep mutational scanning is a fairly new method, so data analysis generally requires familiarity with command-line tools and scripting languages even for substitution libraries. InDel libraries present some additional challenges: because the length of the variants is different, each variant must be aligned to the reference sequence and the alignment used to identify the variants. Thus, during these projects I tested multiple methods for analysing DMS data, both for InDel and for substitution libraries. From the start, this was an open-ended challenge that might find an easy solution early on or present substantial obstacles.

Statistical modelling and validation. The DMS datasets are a gold mine of information, but also contain complexity: there is noise, artefacts from sequencing errors, stochastic effects due to random sampling of variants during sequencing, and varying frequency of variants between the libraries. Therefore, inferring the 'true' activity of variants is not obvious. I set out to test different way of modelling variant activity from deep sequencing data, both statistical methods and machine learning models.

Chapter 2

Tolerance of deletions in GFP

In this chapter I describe the deep mutation scanning of GFP with substitutions and short deletions (1, 2 and 3 residues). I worked with the libraries in two GFP starting points with different stabilities: eGFP and the stabilised variant GFP8. The libraries were sorted with FACS into three activity fractions, deep sequenced with Illumina 2×75 *paired end sequencing and the fitness of individual variants was inferred with InDelScanner scripts. While primarily a pilot experiment, this dataset also provides some insight into how stability affects the tolerance of InDels compared to substitutions.*

2.1 Introduction

2.1.1 DMS with deletions and substitutions: experimental design

While high-throughput deep mutational scanning in the green fluorescent protein has been done for substitution libraries (Sarkisyan et al., 2016) and some screening has been done in single triplet deletion (-3 bp) libraries (Arpino et al., 2014a), there is a gap in understanding of longer deletions. Nor are there any published studies that systematically examine the effect of short insertions of random sequence. Here, I wished to fill this gap in knowledge, and I set out to adapt the established protocols for deep mutational scanning to the TRIAD InDel libraries, and use the results to build fitness landscapes of all three types of small mutations.

In terms of complexity, the steps required for achieving fluorescence of a protein are fewer than the number of factors involved in enzymatic catalysis, which makes GFP a simpler test system than an enzyme. To fluoresce, the fluorophore in the GFP core must form in an autocatalytic process and the mature fluorophore must be located in a suitable microenvironment in the protein core, protected from bulk solvent exposure. The autocatalytic fluorphore formation requires the correct positioning of the key residues and the water network in the core, but there is no added difficulty of maintaining both substrate binding and catalysis of the chemical reaction (as there would be in an enzyme).

In addition to aiming for a fitness landscape of InDels in a fluorescent protein, I also wished to explore the effect of InDels on protein stability. The starting hypothesis was that a stabilised starting point would generally show better tolerance of stability-impacting mutations; however, the magnitude of the effect and the structure- and position-specific effects were unpredictable. The extent to which the importance of starting point stability, which has been demonstrated for evolution of catalysts (see 1.1.2), is relevant to the function of a fluorescent protein is also unclear. In order to address the question of stability affecting fluorescence, I chose to screen the InDel libraries in two backgrounds with different starting stability to better understand these effects.

While numerous adaptive mutations in GFP and related fluorescent proteins have been identified, the stabilised variants typically contain multiple mutations. This feature creates difficulty in distinguishing adaptive mutations from neutral ones unless comprehensive combinatorial screening is performed. Building on previous work, I set out to test the effect of small InDels on the following two model proteins:

- eGFP: a well-established variant with two mutations (F64L/S65T) relative to the source jellyfish variant *av*GFP, engineered to have a single peak in the absorption spectrum and 35-fold improved folding at 37°C compared to avGFP (Cormack et al., 1996). It is a well-folded, well-expressed protein and as such a good baseline reference point for testing the tolerance of insertions or deletions;

- GFP8: a new variant with further six mutations in addition to two in eGFP. It combines the cycle 3 mutations F99S, M153T and V163A (Crameri et al., 1996) and three out of the six additional mutations in sfGFP, S30R, N105T and A206V (Pédelacq et al., 2006).

GFP8 was designed by Dr. Nobuhiko Tokuriki, by examining the properties of variants that incorporated various combinations of mutations present in sfGFP. The aim of those combinatorial experiments was to find a GFP variant that was stable, very soluble and functional as a folding reporter when placed in fusion with other proteins (see Gupta and Tawfik (2008) for background). GFP8 was a variant that exhibited the best such properties: this is achieved by keeping the mutations in sfGFP that individually increase thermodynamic stability, while discarding the mutations affecting the folding rate. The relationship between mutations in eGFP, GFP8 and related constructs is shown in Figure 2.1 and the relevant mutations are listed in Table 2.1. Finally, compared to sfGFP, analysis of the mutational landscape is easier in GFP8, because fewer mutations in GFP8 compared to eGFP mean that there are fewer confounding effects. Thus, GFP8 is a 'cleaner' construct that differs from eGFP in final stability and solubility, but is similar in the rate of folding and maturation.

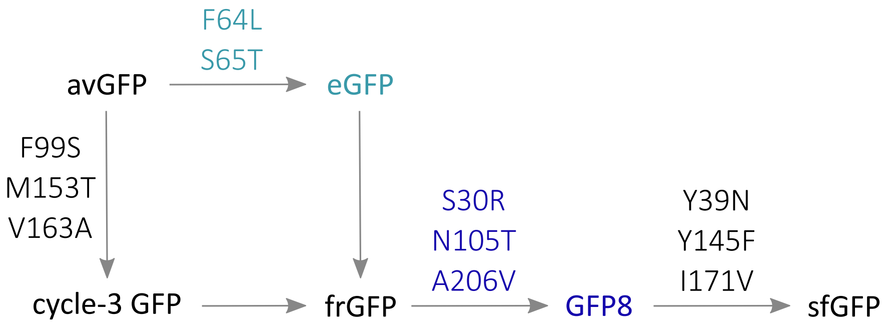


Figure 2.1 The relationship of early stabilised GFP variants to each other and to GFP8. References: *av*GFP is the ancestral *A. victoria* sequence containing the Q80R mutation, introduced in the original isolation of GFP when it was first discovered (Tsien, 1998); eGFP was discovered by Cormack et al. (1996) and popularised by Yang et al. (1996); the cycle-3 mutations were identified by Crameri et al. (1996); frGFP and sfGFP were described by Pédelacq et al. (2006).

Prior to the start of this work, TRIAD libraries (-3, -6, -9 bp deletions and +3, +6, +9 bp insertions) had been prepared by Dr. Stéphane Emond in both eGFP and GFP8, following the procedure outlined in section 1.1.4. Additionally, trinucleotide substitution libraries had also been prepared using the TriNEx protocol (Baldwin et al., 2008), which were used as control libraries for the DMS. Both TriNEx and TRIAD libraries are made with a combination of Mu transposon insertion and cloning steps, so they have comparable design and diversity.

Practically, performing DMS on InDel libraries is challenging because the libraries have different theoretical diversity (~10^3 for each deletion length, 5×10^4 for substitutions and +3 bp insertions, and $> 10^5$ for +6 and +9 bp insertions). In the initial design, I was concerned about the amount of costly Illumina sequencing that would be required for DMS, since both input and sorted libraries need to be sequenced at sufficiently high coverage. I decided to add UMIs to existing libraries, so that we could sequence only the starting libraries with (more expensive) longer-read Illumina sequencing. The hope was that the sorted fractions could be sequenced at higher throughput and at a lower cost by just sequencing the UMIs.

Overall, the planned experimental side of the project comprised:

1. the cloning of UMIs into InDel and TriNEx libraries;
2. GFP expression and separation into different activity fractions with FACS;
3. extraction and sequencing of DNA from starting and sorted fractions;

4. and the fluorescence screening of some individual variants for validation of the high-throughput dataset.

On the data analysis side, the raw sequencing data had to be parsed to extract the mutations, followed by quality control and filtering, and exploration of models to map from NGS counts to estimated variant fluorescence.

2.1.2 Stabilised GFP variants

This section gives a brief overview of known adaptive substitutions in GFP variants and how they guided the choice of GFP variants to use for deep mutational scanning. Numerous GFP constructs have been described to date, many of which are in some way stabilized or optimized for expression, so the choice of a stabilized variant is not immediately obvious.

Since its discovery, GFP has been a popular marker for biochemical studies as well as a target of extensive protein engineering (Tsien, 1998). The two key mutations that give rise to eGFP, F64L and S65T, were discovered more than two decades ago (Heim et al., 1995; Yang et al., 1996). These mutations improve expression at 37°C and modify the protonation state of the chromophore to O^{-} at cellular pH, thereby increasing absorption of blue light and the brightness relative to avGFP, when excited at 488 nm.

Beyond eGFP, numerous variants have been engineered that show a different absorption or emission profile, increased thermodynamic stability, improved folding (kinetic stability) and expression in different organisms or reduced propensity to aggregate. Wild-type *av*GFP and eGFP have a weak to moderate propensity to aggregate, depending on protein concentration, which can confound results when GFP is used as reporter of cellular properties.

Two prominent variants that are especially relevant to this work are the folding reporter GFP (frGFP) and superfolder GFP (sfGFP) (Pédelacq et al., 2006). The former is constructed by combining mutations in eGFP with three additional mutations identified by shuffling of *av*GFP substitution libraries (Crameri et al., 1996). Superfolder GFP was identified after four further rounds of mutagenesis starting from frGFP through error-prone PCR and DNA shuffling, which screened for the brightest variants when in fusion with a poorly-folded protein. After back-shuffling with starting sequence to remove some neutral mutations, additional six adaptive mutations were identified (Table 2.1). As a result, the sfGFP shows in increase in thermodynamic stability ($\Delta\Delta G_{WT-sf} = -2.3kcal/mol$), a faster folding and maturation rate (Pédelacq et al., 2006) as well as a reduced propensity for aggregation. While *av*GFP and eGFP retain a weak to moderate tendency to dimerise, dependent on protein concentration, sfGFP remains a monomer even at high concentration (Costantini et al., 2012).

Other adaptive mutations have been engineered more recently. One important residue is Glu222, which is highly conserved in the family of fluorescent proteins and is thought to stabilise the hydrogen bond network facing the chromophore (Royant and Noirclerc-Savoye, 2011). However, it is also prone to photodecarboxylation, which results in bleaching of the fluorescence even at low temperature. A substitution to similarly acidic residue histidine (E222H) protects GFP from photobleaching and expands the pH range over which it exhibits fluorescence (Auerbach et al., 2014). This mutation was not included in the GFP8 construct, but is expected to appear in the substitution library as a tolerated or beneficial substitution. Second, an extra superfolder GFP (esGFP) variant with further 12 mutations was developed during exploration of tolerance of multiple simultaneous non-canonical amino acid substitutions (Nagasundarapandian et al., 2010). Inspection of the crystal structure of this heavily engineered variant assigned the stabilizing effect to five surface mutations that form highly structured salt bridges or hydrogen bond networks (Choi et al., 2017), suggesting that surface mutations to charged residues are adaptive in part because they increase solubility of the protein. This observation forms a hypothesis that these classes of mutations could be better tolerated: deletions of surface hydrophobic residues, as well as substitutions and insertions that introduce charged residues on the surface.

Mutation	Location	Functional improvement	Construct	In GFP8
S30R	Surface	Increased thermodynamic stability, faster refolding rate	sfGFP	Yes
Y39N	Surface	Improved refolding rate	sfGFP	No
F64L	Core	Improved fluorescence at 37°C	eGFP	Yes
S65T	Chromophore	Promotes chromophore deprotonation	eGFP	Yes
F99S	Surface	Improved stability	cycle 3	Yes
N105T	Surface	Faster refolding rate in presence of urea	sfGFP	Yes
Y145F	Core	Increased fluorescence, decreased refolding rate	sfGFP	No
M153T	Surface	Improved stability	cycle 3	Yes
V163A	Core	Improved thermostability	cycle 3	Yes
I171V	Surface	Improved thermostability	sfGFP	No
A206V	Surface	Improved urea tolerance	sfGFP	Yes

Table 2.1 A summary of mutations present in eGFP, sfGFP and intermediate constructs, the mutation location and their published effect on GFP brightness, folding and thermostability.

2.2 Results

2.2.1 GFP8 crystal structure

Mutations can affect GFP fluorescence by altering the protein structure, flexibility or by altering factors involved in protein folding and solubility (for specifics, see section 1.3). While GFP8 shares mutations with multiple variants that have been crystallised, there are differences between structures that make it harder to model the structure of GFP8 in detail. Therefore I set out to obtain a crystal structure, which would allow distinguishing between the effects of mutations and the effects of the starting point, and facilitate later interpretation of screening results. The crystallography work was done in collaboration with Dr. Mariana Rangel Pereira.

Many crystal structures of GFP variants are available on the Protein Data Bank (PDB), where I examined the published protocols to discover common crystallization conditions. I searched the PDB for existing GFP crystal structures and compared crystallization conditions in structures at the following accession codes: 4EUL, 5BT0, 5BTT, 4ZF3, 5HZO, 5DY6, 4KA9, 4OGS, 4P1Q, 3P28 and 2YOG. There was no clear theme in published conditions: different variants have been crystallised under both acidic and basic conditions, in a variety of buffers and with several different precipitants. I therefore decided to start the crystallization screen from the beginning, with commercial crystallization plates.

The 6×His tagged GFP8 in the pID-Tet plasmid was expressed in *E. coli*, the cells lysed and the protein purified on a nickel affinity column, followed by size exclusion chromatography and the final combined step of desalting and buffer exchange. During initial experiments, the concentrated protein was frozen between experiments, but that sometimes triggered precipitation. Since GFP is typically not prone to aggregation, it was later stored at 4°C between expression and the start of crystallization experiments.

Each batch of purified GFP8 was assessed by SDS-PAGE after nickel affinity purification and after size exclusion chromatography. Initially, I observed two bands close the expected size (28.4 kDa) as shown in Figure 2.2. This was likely due to incomplete denaturation during sample preparation, because only one band was observed when the samples were heated for >10 min at 95°C and the protein stocks re-analyzed.

The purified protein was screened against sparse matrix conditions in Wizard I&2, Classics and JCSG+ crystallisation plates. The three initial hits were:

- 0.1 M HEPES pH 7.5, 200 mM NaCl, 10% v/v IPA
- 0.1M CHES pH 9.5, 200 mM NaCl, 10% w/v PEG 8000
- 0.1M CAPS pH 10.5, 200 mM NaCl, 20% w/v PEG 8000

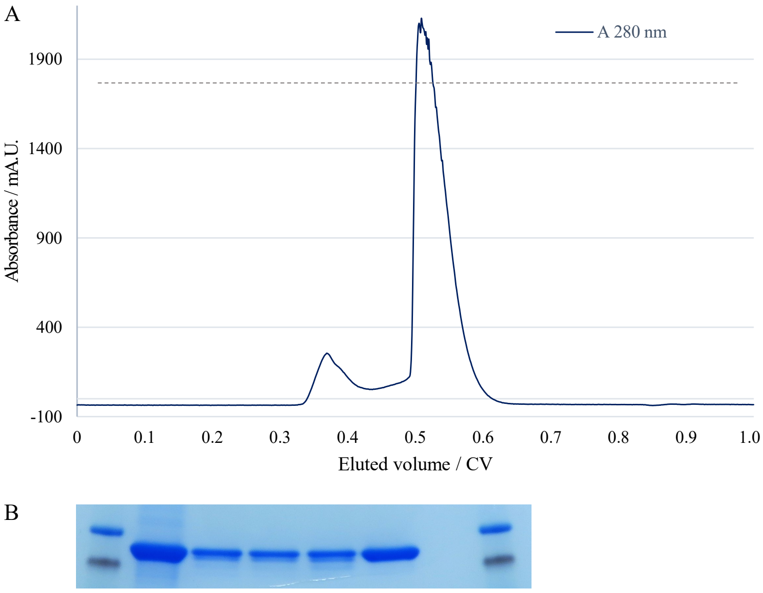


Figure 2.2 Purification of GFP through immobilised metal-affinity chromatography and gel filtration. A) During purification of the protein with size exclusion chromatography, the protein content of the elute was monitored by UV-VIS absorption at 280 nm. A minor and a major peak were observed. The major peak was fluorescent and was collected. B) Denaturing polyacrylamide gel of GFP8-His6 after gel filtration shows a strong main band and a faint second peak at an apparent lower molecular weight.

The pH and salt concentration were further optimised with manual screening by hanging drop technique, which gave crystals in needle, plate and cubic shape in multiple conditions. The promising crystals were used for data collection at the Diamond Light Source synchotron, where 7 out of 13 crystals gave a good diffraction pattern. The best dataset was used to solve the crystal structure using automatic molecular replacement with a set of sfGFP/W57A coordinates (PDB code 4LQT). The chromophore structure was manually linked with the backbone, then the structure was refined with Phenix (Liebschner et al., 2019) until all parameters were satisfactory. The final structure with 1.29Å resolution is available at the Protein Data Bank at accession code 6HUT and the structure parameters are listed in Table 2.2.

Data Reduction Statistics	**GFP8**
Beamline	I04-1
Wavelength (Å)	0.9282
Space group	P 21 21 21
a, b, c (Å)	52.16, 63.37, 69.21
Resolution range (Å)	31.69-1.29
Total reflections measured	325,473
Unique reflections	58,248
Completeness (%) (inner shell)	98.8 (100.0)
I/σ (inner shell)	16.9
B(iso) from Wilson (Å^2)	15.6
Refinement Statistics	
Protein atoms excluding H	1886
Solvent atoms	211
R-factor (%)	18.6
R-free (%)	20.3
Ramachandran Plot Statistics	
Rmsd angles (Å)	1.23
Core region (%)	99.2
Allowed region (%)	0.8
Additionally allowed region (%)	0
Disallowed region (%)	0
PDB code	**6HUT**

Table 2.2 Data collection, reduction, refinement and Ramachandran plot statistics of GFP8 crystal structure with PDB code 6HUT.

Both for the purpose of screening and crystallization, GFP8 was expressed as a C-terminally 6×His-tagged construct. GFP8 crystallised with a single protein molecule per crystallographic unit cell, where the protein C-terminus formed an additional α-helix (Figure 2.3A). The backbone of the structure very closely follows the backbone of eGFP, with root mean square difference between backbone atoms only 0.231Å- that is, they are essentially the same structure. The alignment with the structure of sfGFP is similarly close, except for the loop between β9-β10, which follows a different position between sfGFP and eGFP, and where GFP8 matches the positions in eGFP (Figure 2.3B, top right). The following paragraphs describe the similarities and differences between these three structures, to attempt to understand the properties of the DMS starting point.

The chromophore in GFP is unusual in that it is located in a polar, water-accessible environment, and its brightness and spectral properties are closely linked to the structure of the hydrogen bonding network in the vicinity. The aromatic region of the chromophore is located close to four water molecules, two of which form hydrogen bonds with it (Figure 2.3C). The phenolic OH/O^- group is coordinated with a water molecules and the side chains of His148 and Thr203. The carbonyl group in the imidazole ring coordinates with Glu94 side chain, Thr62 carbonyl group and the positively charged Arg96, which polarises the imidazole ring and contributes to stabilisation of the negatively charged chromophore (and so increases brightness when excited at 488 nm). The Arg96 residue is itself stabilised by a hydrogen bond network with Thr206 backbone carbonyl group, a water molecule and Glu183. On the other face of the imidazole ring, the nitrogen of the ring forms a hydrogen bond with a coordinated water molecule. This water molecule is part of a larger water network (consisting of 6 crystal waters in eGFP, and 4 in GFP8 and sfGFP) that is organised by Glu222. Given how close the water molecules are to the chromophore, their organisation is essential to maintenance of fluorescence: while ePISA analysis shows that all residues in GFP are accessible to water, that water is not the bulk solvent - the chromophore must nevertheless be shielded from bulk solvent to prevent fluorescence quenching.

A comparison of three structures. Specifically examining the six positions where GFP8 carries an additional mutation compared to eGFP, the largest difference appears in the orientation of residue Ser30Arg (Figure 2.4). Residue 30 is located on the surface of GFP, in the middle of the strand $\beta 2$. In eGFP, the Ser30 occupies two rotamers, one of which form an internal hydrogen bond with the backbonce carbonyl group. In GFP8 and sfGFP, this position is occupied by Arg30, but the side chain positions are different: in GFP8, the side chain forms a polar contact with a co-ordinated water molecule (which is present in all three structures), while in sfGFP the side chain NH_2 forms a salt bridge with Asp16. Here we see that when the stabilising Ser30Arg mutation - identified in sfGFP - is transplanted into GFP8, the structural effect is different. However, the change from a polar to a charged residue is in line with the previous observation that charged residues on GFP surface are stabilising because of increased solubility (Choi et al., 2017).

Of the three mutations originating from the cycle-3 stabilisation campaign, Val163Ala is located in strand $\beta 8$ and is the only residue of the six mutations facing towards the barrel

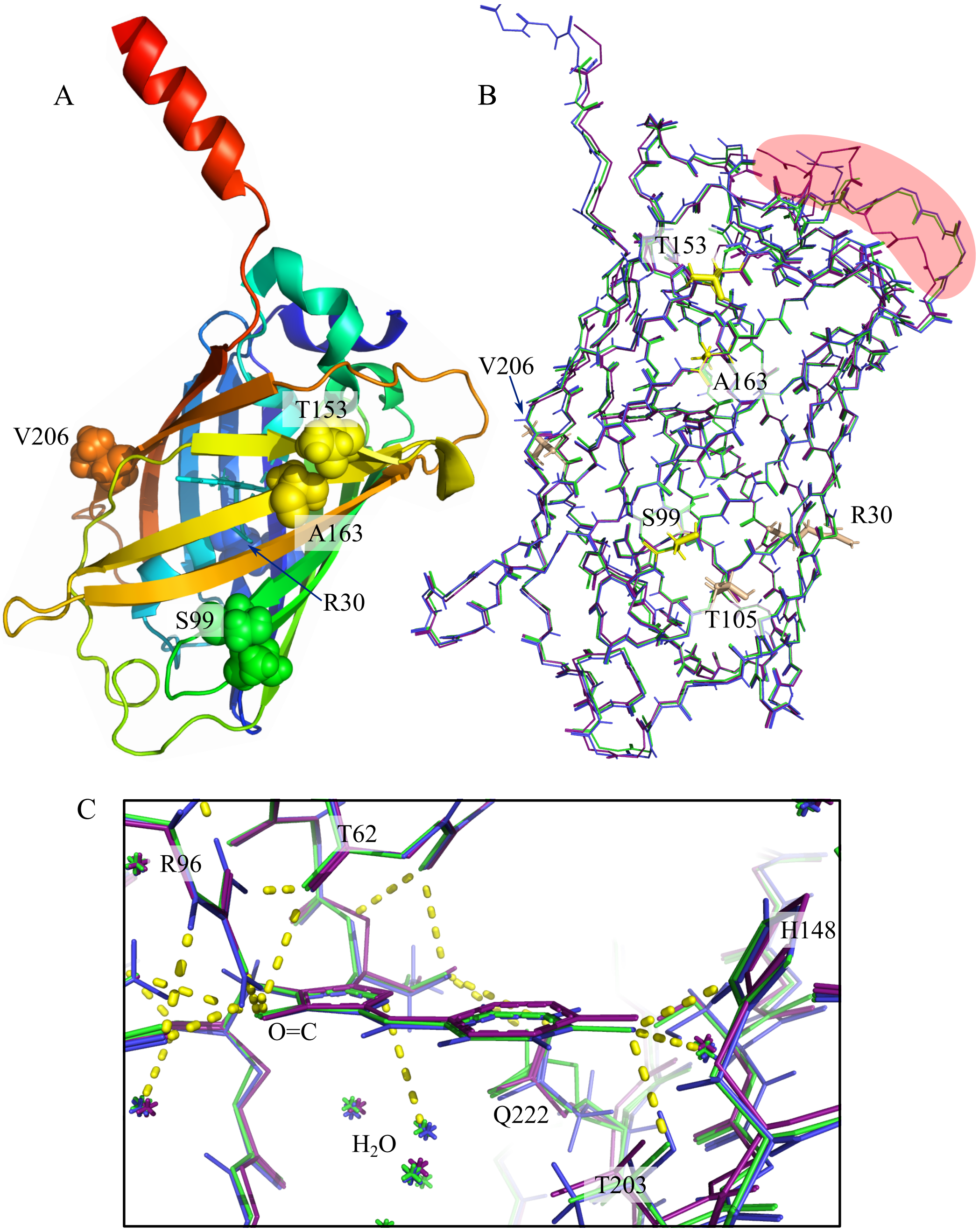


Figure 2.3 The crystal structure of GFP8-His (accession code 6HUT) compared to eGFP and sfGFP. A) Representation of the secondary structure in the GFP8-His construct, highlighting the position of six additional mutations present in GFP8 compared to eGFP. The colour scheme follows the protein sequence from the N to C terminus, from blue to red. B) Alignment of eGFP (green), GFP8 (blue) and sfGFP (magenta) backbone positions. The positions highlighted in yellow originate from frGFP and those in beige from sfGFP. All three backbones are in excellent alignment, with the exception of the protein termini and the loop $\beta 9 - \beta 10$ (shaded red), which takes a different conformation in sfGFP. The position of this loop in GFP8 aligns well with eGFP. C) An illustration of key residues that form hydrogen bonds or polar contacts with the aromatic, flat chromophore moiety. The organising residue Glu222 appears behind the chromophore in this perspective and organises three water molecules in front.

core. In eGFP it is occupied by Val, which packs closely against nearby Gln183. In sfGFP and GFP8 it is replaced by Ala, leaving open the space taken up by the larger Val side chain without changing the position of Gln183 (significant, because this residue is involved in coordinating the hydrogen bond network around the chromophore). The second cycle-3 position, Met153Thr, is on the surface towards the end of strand $\beta 7$ and shows two alternative rotamers in eGFP, indicating a large degree of flexibility. The smaller Thr residue in GFP8 and sfGFP maintains this flexibility, and the residue occupies different rotamers between the two structures. The third mutation, Phe99Ser, is on the surface in strand $\beta 4$ and in close proximity to residue 105. Both Phe99Ser and Met153Thr increase the polarity of the GFP surface, which may have a solubilising effect and thereby improve GFP folding kinetics.

The mutation Asn105Thr is curious in that it eliminates the hydrogen bond between the side chain and backbone that Asn105 exhibits in eGFP, while the Thr residues in GFP8 and sfGFP show flexibility: they do not form defined polar contacts and occupy different rotamers between the structures. Again, the contribution of this mutation to the GFP8's increase tolerance of other mutations may be primarily due to solubilisation. However, Ala206Val is also a mutation of a surface residue that, in contrast, increases hydrophobicity of the surface, so the physical causes for these effects are less clear. Here, the residue Val206 occupies the same position in GFP8 as it does in sfGFP.

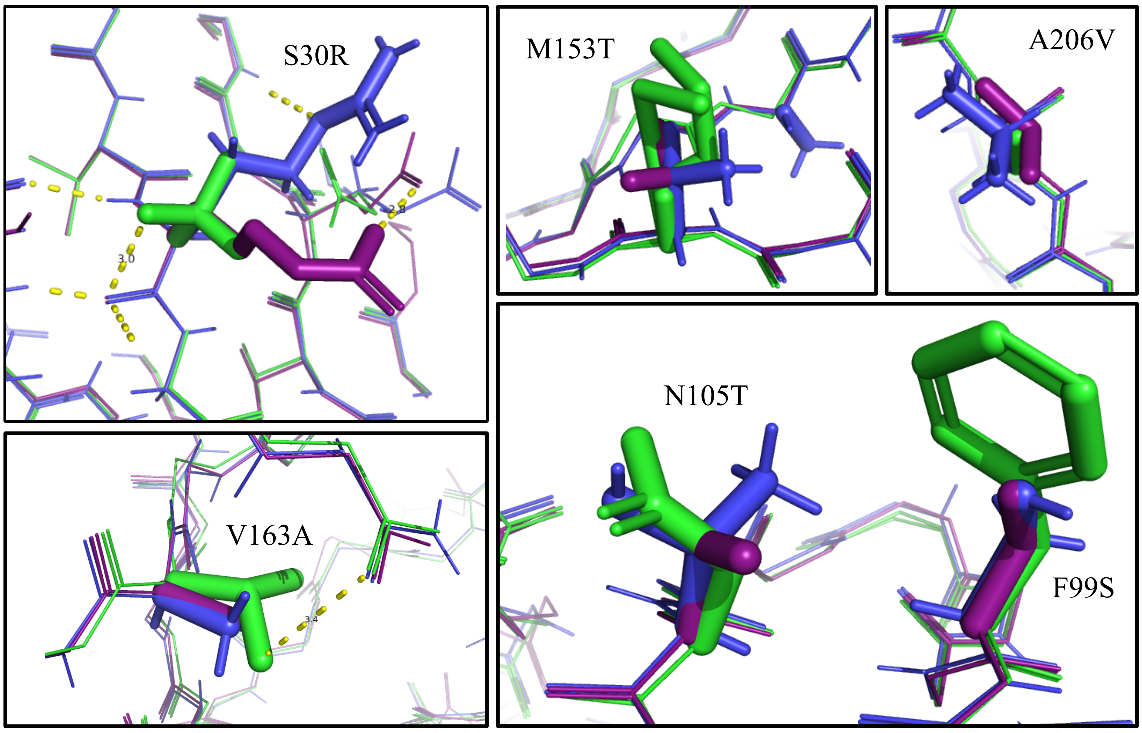


Figure 2.4 Close up representation of side chain positions in GFP8 (blue), compared to eGFP (green) and sfGFP (magenta).

Flexibility. Finally, because stability is also influenced by rigidity of different parts of the protein, I examined the flexibility in the different parts of the protein by looking at B-factors in four related structure: eGFP (PDB code 4EUL) , folding reporter GFP (PDB code 2B3Q), GFP8 (PDB code 6HUT) and superfolder GFP (PDB code 2B2P), shown in Figure 2.5. It should be noted that B-factors in crystal structure depend on the resolution and so the absolute values cannot be compared between structures, and the comparison here is on a qualitative level. In these four structures, frGFP was solved to a lower resolution (2.3Å) compared to the other three variants (1.3-1.4Å), so the comparison focuses on the latter three structures.

In all four structures, the B factors are much lower in the parts of the protein with a defined secondary structure, the β strands and the central α helix, but much higher in the loops on both ends of the barrel. The N- and C-terminal ends of the protein are also very flexible, which is a common observation for protein crystallography, and aligns with the observation that multiple C-terminal residues can be deleted in GFP without loss of fluorescence. There are three regions where there is an apparent change in relative flexibility, which are highlighted in orange and purple in Figure 2.5.

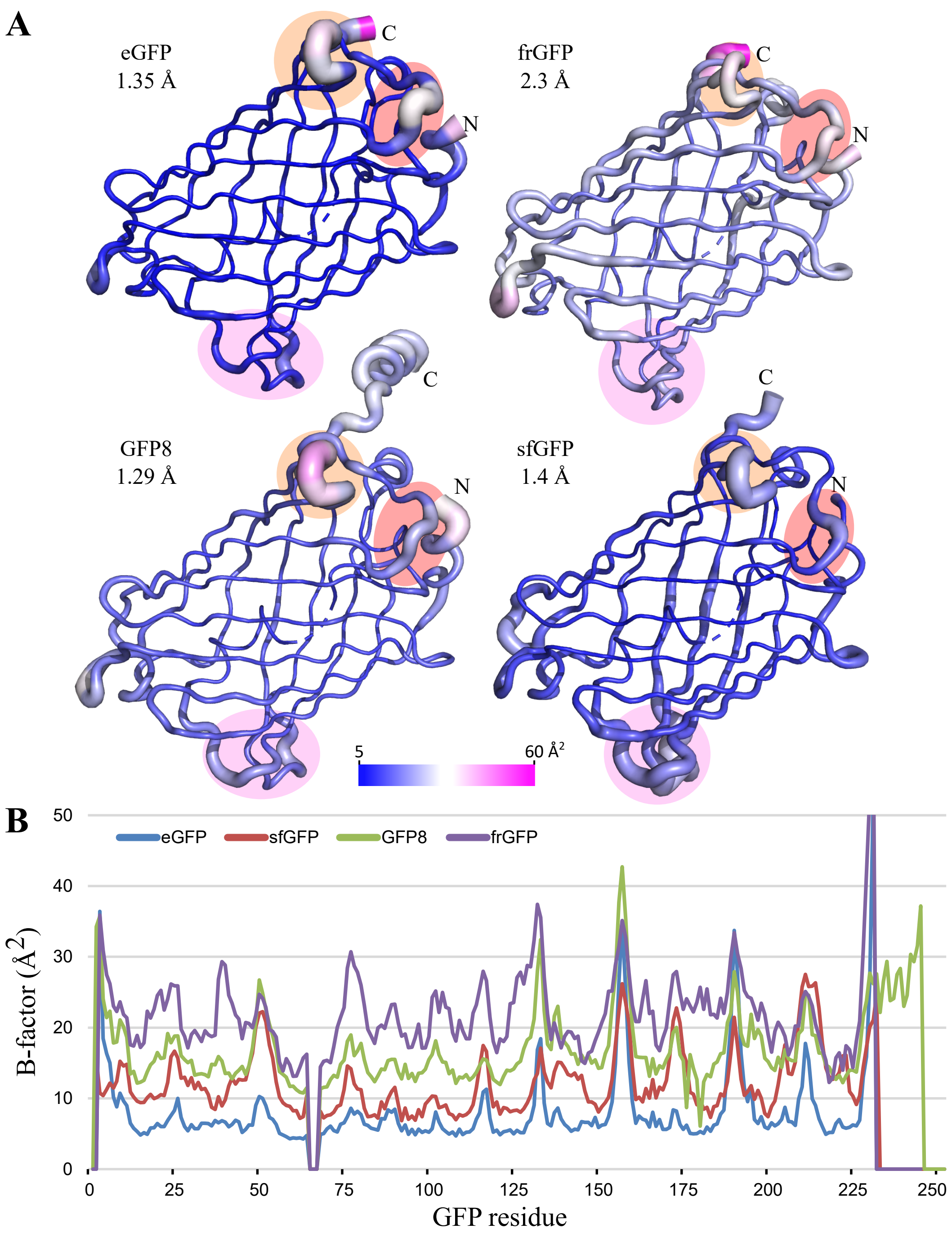


Figure 2.5 A comparison of B-factors in crystal structures of four related GFP variants. Both size and colour of protein loops indicate the crystal temperature factors; the structures are coloured from blue (low atomic motion) to magenta (high atomic motion) and the size of each loop is proportional to the B-factor value. Regions of difference are highlighted with shaded background: orange circle - loop between strands $\beta 7$ and $\beta 8$; red circle - loop $\beta 9$-$\beta 10$; purple circle - front loop is between $\beta 3$ and the central helix, back loop is between strands $\beta 10$ and $\beta 11$.

First, the loop β9-β10 (red shading in Figure 2.5) is very flexible in eGFP, shifting to reduced flexibility in frGFP and GFP8 while keeping the same backbone position. In sfGFP, the conformation of this loop changes compared to eGFP, frGFP and GFP8 (see also Figure 2.3B), but the trend towards a decrease in flexibility in sfGFP remains. Second, three residues in the loop β7-β8 (orange circles, residues 155-159 inclusive) forms a small cluster of local high flexibility in-between a highly organised region, and the B-factors remain high across all four structures. Finally, on the other end of the β-can there are two loops (shown in purple) that counter-intuitively appear relatively static in eGFP and frGFP, but show a mild (GFP8) and strong (sfGFP) increase in flexibility in sfGFP. The front loop is the transition region between the β3 strand and the central helix (residues 49-54) and the back loop is the $\beta 10 - \beta 11$ loop (residues 211-214).

These data suggest that the introduction of additional mutations in the progression from eGFP to sfGFP changes the dynamics of the protein, and the higher stability at least of sfGFP may originate in greater rigidity of the structure as evidenced by changes in B-factors, rather than in obvious changes amino acid backbone or side chain positions. On the other hand, the differences both in kinetic and thermodynamic stability could be due to changes to residual structure in the unfolded states of these proteins, which are entirely invisible when examining the crystal structures of folded proteins.

Taken together, at this point the main conclusion regarding the difference between the eGFP and GFP8 crystal structures is the protein chain flexibility in the loops highlighted in Figure 2.5. Beyond the difference in flexibility, there does not appear to be a large difference in either backbone position or the mutated side chains. The protein backbone in GFP8 aligns very closely with eGFP placement, including in the loop which has different placement in sfGFP. The GFP8 structure, much like the sfGFP structure, shows less atom movement in the loops at the end of the β-barrel. However, the meaning of changes in temperature factors is controversial, because it can be interpreted in two different ways:

- Increased fluctuations in atomic coordinates indicate disruptive destabilisation in the protein structure, which in the extreme leads to protein unfolding or aggregation.
- Increased fluctuations indicate adaptive flexibility, which increases the protein's ability to accept new mutations since the more flexible protein regions can change conformation without ill effect on overall protein structure (see Campbell et al. (2016) for an example of an adaptive role in conformational changes in protein evolution).

At this point I formed the following hypotheses:

- Solubilising mutations on the surface of the protein will be well tolerated, even in the middle of β-strands. Therefore, special attention should be directed to deletions that introduce adjacent substitutions and classify them by type of substitution.,
- because the β-strands are very rigid, InDels will be much better tolerated in the protein loops,
- while accepting InDels in the core of the protein is unlikely, some could be accepted if the total size of the residues is similar.

2.2.2 The strategy for deep mutational scanning

Introducing UMIs into InDel libraries

At the start of this project, I envisaged the functional screening of GFP to be relatively straightforward with FACS, however it was likely there would be challenges with processing of NGS results. I therefore decided to initially focus the DMS data collection on the smaller deletion and substitution libraries, and leave the much larger insertion libraries for the second stage of the work. Thus, I could be reasonably sure to obtain some results in the first stage, with smaller libraries and fewer samples, and later achieve better outcomes with the larger insertion libraries that require more optimization.

The deletion libraries have a maximal theoretical diversity of 720 DNA variants per library; this diversity is equal to the length of the gene, since a fixed length deletion can at most occur once at each position in the gene. The actual DNA and protein diversity are lower because of repetition in DNA sequence and codon redundancy. If I were only interested in screening the effect of deletions, these libraries could be mostly screened in 96-well plates and the results analysed manually. However, I was interested both in a fitness landscape of short deletions and in developing a scalable methodology, so I used deletion libraries as a test system on which to develop the InDel DMS tools (Chapter 4.

Looking ahead to using UMIs for the larger libraries, I first wanted to introduce 20 nt long UMIs with random sequence at the end of each DNA sequence. I designed these UMIs to be placed as close to the end of the gene as possible, to maximise the gene length that gets sequenced in the same Illumina read: if the UMI is placed 100 nt after the gene, the sequencing read must start at the UMI and run for another 100 cycles before it reaches the gene, so that sequencing read length is effectively wasted. Given that the maximum Illumina read length is 300 nt per read, clearly every nucleotide counts.

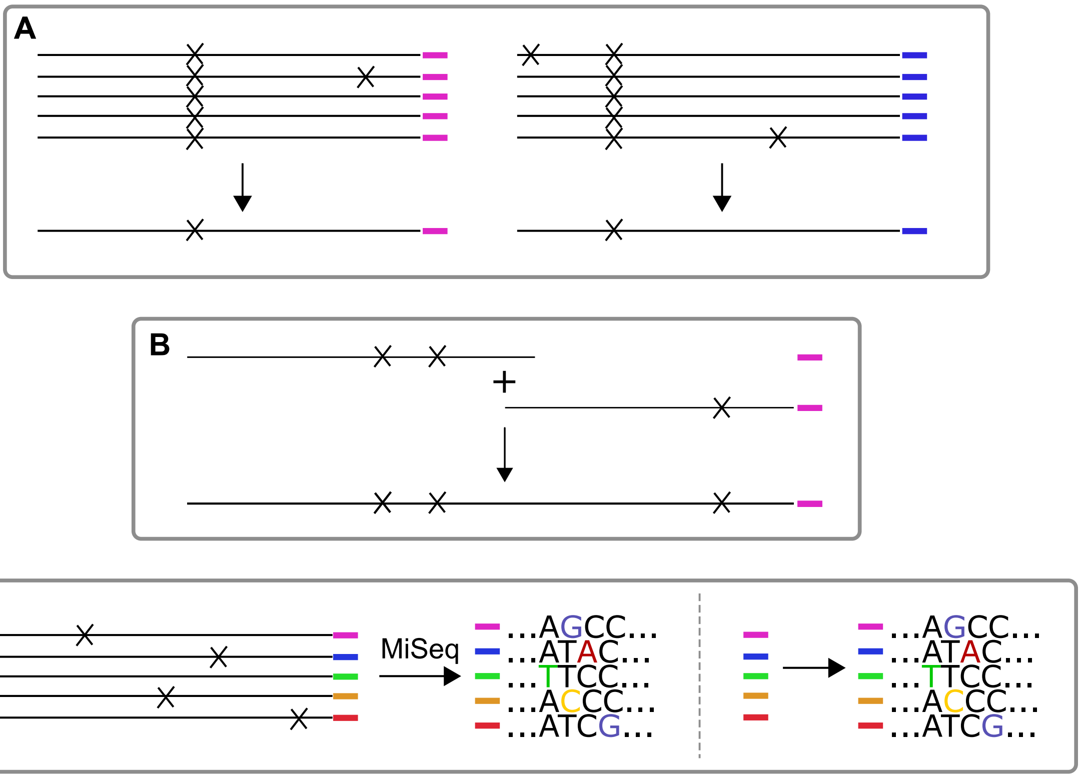


Figure 2.6 The uses of UMIs in amplicon deep sequencing. A) Error correction. Each variant is tagged with a unique UMI, such that reads belonging to the same variant can be grouped. Only mutations present in the majority of reads are accepted, while rare mutations in the same UMI cluster are safely discarded as sequencing errors. B) Extending the available sequencing length through subassembly of distal parts of the sequence. If the target gene is longer than available read sequencing, then two distal parts can be linked to the same UMI and later re-joined computationally. C) A dictionary of UMIs reduces the amount of sequencing amount required for sorted libraries. If all variants in the starting library are identified with a longer-read sequencing run, then the meaning of every UMI in this library is known. After the selection, only the UMIs need to be sequences, which reduces the cost. Additionally, the same library can undergo multiple selection conditions with only a moderate increase in sequencing cost. The dictionary application is easily combined with error correction and/or subassembly.

My original experimental plans explored the use of UMIs in combination with a custom Illumina amplicon sequencing protocol, in order to enable effective deep sequencing of both deletion and insertion libraries. However, after exploring the possible protocols I felt there were too many unknowns in that plan to use it as the first experimental strategy. I was particularly concerned about the issues of read diversity and achieving sufficient sequencing depth.

Variant coverage. If the libraries are 'ideal', meaning they contain only the desired variants in equal proportion, calculation of required sequencing depth to achieve the desired number of reads per variants is straightforward. For example, if the ideal library contains 50,000 variants (such as TriNEx or the +3 bp insertion libraries), the sequenced fragment is 800 bp long, each read covers 200 bp and the aim is to observe each variant 20 times, it takes $50,000\ variants \times \frac{800\ bp\ in\ the\ gene}{200\ bp\ read\ lenght} \times 20\ observations = 4.0$ million reads to cover this library. However, calculation of the required amount of sequencing for a real library requires an assumption about the distribution of variant frequencies: if some variants are present at 10% of the average, that requires approximately $10\times$ more total sequencing depth to detect them as well. Consequently, the sequencing depth needed for detection of less abundant variants is very sensitive to any bias in variant distribution.

Diversity. Illumina sequencing that starts at the same starting position in the gene has a higher rate of failure, especially if the amplicon sequence has very low diversity. The issue is that the Illumina software uses the first 4 sequencing cycles to identify individual DNA clusters; essentially each cycle takes a high resolution image of the flow cells and analyses the images to choose the boundary of each cluster, by finding borders between differently coloured spots in the image. The optical focus does not change in later cycles or between reads (if using multiplexed and/or paired-end sequencing), thus creating the information of which reads belong in a pair. However, if the sequence is low diversity and most clusters fluoresce with the same colour, the process of cluster identification will fail.

The issue of low diversity amplicons can be addressed by mixing in a proportion of random DNA (PhiX viral genome), up to 20% of total loaded DNA. Additionally, if the amplicons are sequenced from the same starting position (that is, not starting at random points in the gene), the outcomes are improved if the first sequenced nucleotides are very diverse. Therefore, UMIs are best positioned such that they are sequenced in the first cycles of the sequencing run. However, taking advantage of UMIs to improve sequencing quality in this way requires precise engineering of the library plasmid flanking sequence or the use of custom sequencing primers, which complicates the experimental protocol.

Pilot experiment design

The experiments described in this chapter were the the first attempt at collecting a large dataset describing the effect of InDels through deep sequencing. At the time, I had not yet developed data analysis pipelines, so I decided that tackling the complexity of a custom sequencing protocol (see above) at the same time would be over-ambitious.

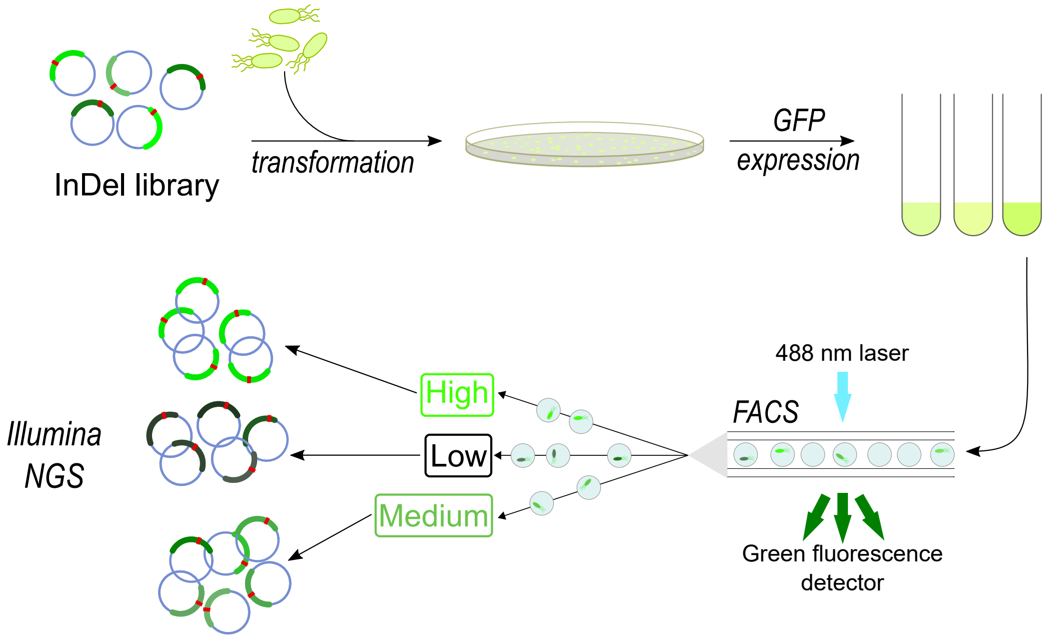


Figure 2.7 An outline of the pilot DMS experiment with GFP libraries. The starting InDel libraries are separately transformed into the appropriate highly competent *E. coli* strain and grown overnight on large agar plates or in liquid culture. After GFP expression, the individual cells are sorted with based on green fluorescence excited at 488 nm, into three activity fractions. The plasmid DNA in each fraction is recovered and the GFP gene variants submitted for Illumina MiSeq sequencing.

For this first data collection, I chose a simplified protocol that was sure to generate *some* good data, even if it was not optimised to generate the maximum amount of insight per sequencing run, or take advantage of UMIs. In this way, I could generate a ’good-enough’ experimental dataset on which to optimise the pipeline, and still learn about the functional effect of InDels on GFP. I chose to compare eight libraries: -3 bp, -6 bp, -9 bp deletions and 3 bp substitutions (TriNEx libraries), each prepared both in eGFP background and in the stabilised GFP8. These libraries have lower theoretical diversity than insertion libraries, so any NGS dataset should be large enough to provide good information on their activity.

The steps in this pilot experiment (Figure 2.7) were:

- Use FACS to separate each library into three or more activity fractions (see Figure 2.8 for gating strategy),
- Verify the efficacy of the sorting and repeat sorting as needed, aiming to minimise the overlap in fluorescence between the fractions.
- Extract the GFP gene DNA and obtain an Illumina NGS dataset.
- Measure the fluorescence of randomly selected variants and identify the mutations with Sanger sequencing, and use these variants to validate the analysis of the NGS dataset.
- Use the NGS dataset to develop a bioinformatic pipeline for analysis.

2.2.3 Fitness measurements

Fluorescence-Activated Cell Sorting

In GFP, fitness is experimentally defined as the fluorescence brightness of single cells expressing the protein. Biochemically, brightness is a composite measure of the amount of protein expression per bacterial cell, protein folding, rate of chromophore maturation, resistance to photobleaching and the inherent brightness of the chromophore in each variant. Of these contributing factors, efficient protein folding is likely the strongest source of true variation, and variability in protein expression the largest source of experimental noise.

The libraries were transformed into an *E. coli* strain, grown overnight in liquid culture, then diluted into fresh medium to induce protein expression with anhydrotetracyclin. After expression for 2 hours at 30°C, the cultures were diluted in phosphate-buffered saline to approximate cell density 2×10^7 and kept on ice until sorting.

The input distribution of fluorescence between libraries is shown in Figure 2.8. The samples were gated on forward and side scatter to select only singlet bacterial cells, and then sorted into three gates according to GFP fluorescence. The boundaries of the *high* gate were chosen based on the same-day positive control sample that was grown under the same conditions; the *low* gate was set on the lower side of the peak in the negative control; and the *medium* gate was chosen arbitrarily in-between with some distance from both low and high gates. Each library was sorted for 10^5 events per bin, which ensured oversampling of the library >1000-fold for deletion libraries and >10-fold for the substitution libraries. The sorted cells were collected in tubes with LB-ampicillin medium, grown overnight at 37°C to increase the amount of DNA present, and harvested to recover the DNA.

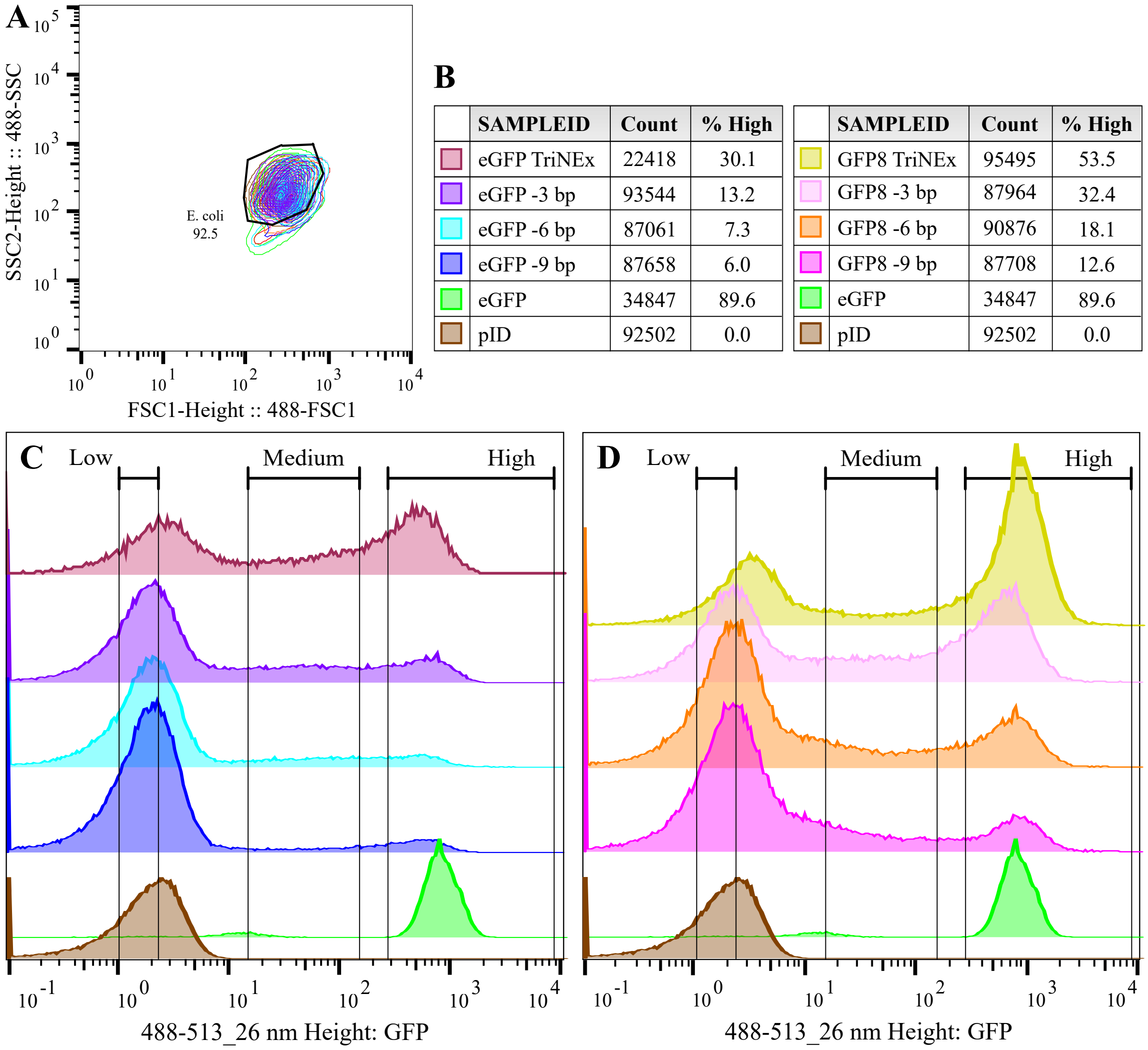
The distribution in Figure 2.8 show that there is a substantial proportion of fluorescent variants in all libraries, which decreases in the expected order across four libraries in the same GFP background (TriNEx > -3 bp > -6 bp > -9 bp). Similarly, the increased starting robustness of GFP8 shifts the entire distribution to the high fluorescence region of the plot, compared to the fluorescence distribution in eGFP.

Sorting efficiency validation

After sorting, the *E. coli* bacteria expressing GFP were re-grown in liquid culture and the DNA recovered as plasmid mini-preps, separate for each library and fraction. Since no sorting technology - either FACS or microfluidic droplet sorting - is 100% precise and efficient, the outcome of the sorting was validated with flow cytometry. The plasmids were re-transformed into fresh *E. coli* cells, GFP re-expressed and the distribution of the fluorescence recorded in comparison with same-day positive and negative controls (Figure 2.9).

Comparing the fluorescence histograms for the three fractions (high, medium and low), it was apparent that the distribution was good for the high and low fractions. However, the medium fluorescence fraction exhibited a very broad distribution after re-expression, so that it overlapped the control peaks to a significant extent. This broadening is undesirable yet not unusual and is caused by stochastic changes in protein expression levels at the level of individual cells.

In order to improve the separation between activity fractions, the medium fractions were re-sorted to eliminate variants in the extreme of the fluorescence distribution, in the same manner as before. This reduced the width of the medium fraction histograms to substantially reduce the overlap with low and high histograms, so the sorting was concluded at this point.



	SAMPLEID	Count	% High
🟥	eGFP TriNEx	22418	30.1
🟪	eGFP -3 bp	93544	13.2
🟦	eGFP -6 bp	87061	7.3
🟦	eGFP -9 bp	87658	6.0
🟩	eGFP	34847	89.6
🟫	pID	92502	0.0

	SAMPLEID	Count	% High
🟨	GFP8 TriNEx	95495	53.5
🟪	GFP8 -3 bp	87964	32.4
🟧	GFP8 -6 bp	90876	18.1
🟪	GFP8 -9 bp	87708	12.6
🟩	eGFP	34847	89.6
🟫	pID	92502	0.0

Figure 2.8 The gating strategy during sorting of starting TRIAD libraries in eGFP and GFP8. A) The events detected by the sorter are first gated on forward and side scatter to isolate *E. coli* cells from random electronic noise. The position of the cells in SSC vs. FSC plot changes between days, so this gate was adjusted to capture most of the cells at every sort. After gating on scatter, the cells were gated based on GFP fluorescence. B) Colour legend for panels A, C and D, also showing the number of sorted cells for that library and the percentage of variants in the high gate . C) Histograms of eGFP libraries (see legend panel), compared to the negative control (pID) and the positive control (pID plasmid expressing eGFP). The sorting gates were set relative to fluorescence of same-day control cultures. The high gate captured the entire positive control and higher, the low gate was set below the main peak of the negative control. The medium gate was arbitrarily set in-between, with some separation from the other two gates to reduce the number of variants that are sorted into two gates because of stochastic variation in protein expression. D) Histograms of GFP8 libraries with the same control cultures and gating strategy.

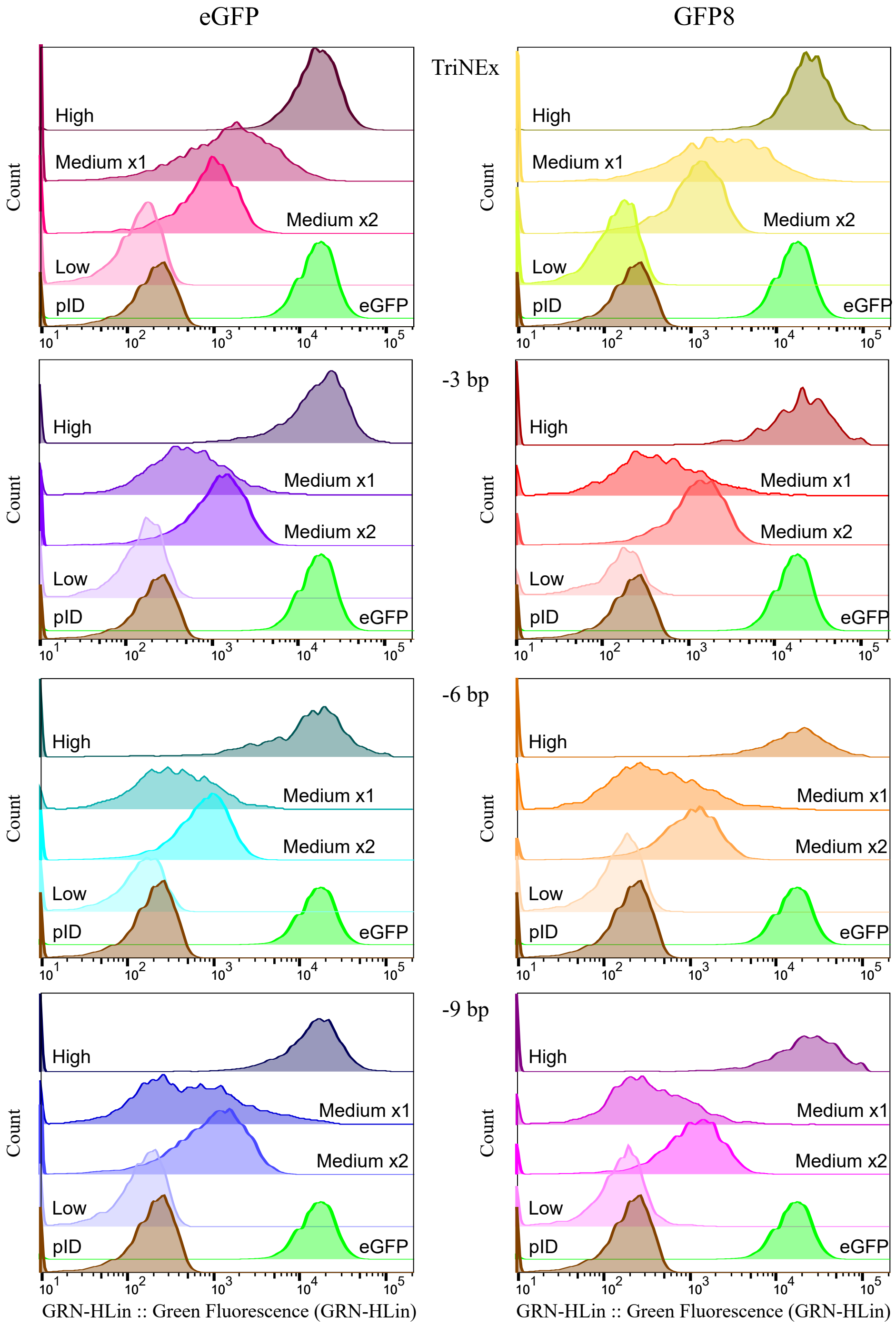


Figure 2.9 Post-sorting fluorescence histograms in eight sorted libraries, comparing the distribution in high, medium (sorted once and re-sorted) and low gates after sorting. All variants are compared to positive and negative controls - direct comparison to sorting gates is not possible because fluorescence was measured on a flow cytometry analyser, not on the sorter instrument.

Nextera library preparation and short paired-end sequencing. The pool of GFP genes was extracted from the sorted plasmid DNA with restriction digestion, using restriction sites located at the start of the GFP gene (NdeI) and after UMI in the plasmid (PstI, 25 bp after the end of the GFP gene). The restriction digests were carefully separated by agarose gel electrophoresis and purified. Finally, because accurate quantification of input DNA amounts is essential for Illumina sequencing, the DNA concentration was adjusted on the basis of NanoDrop measurements and verified with the Qubit assay that is specific for double-stranded DNA. In order to accurately dilute the DNA for sequencing, all DNA samples needed to be handled in dedicated DNA low-binding tubes.

The 738-bp long fragments from all eight libraries were submitted to the Department of Biochemistry Sequencing Facility, where sequencing libraries were prepared with the Nextera XT DNA Library Preparation Kit: the fragments are tagmented with random insertion of the Illumina proprietary transposase enzyme, which simultaneously breaks the DNA into smaller pieces and adds the first part of sequencing adapters. In the second step, five cycles of PCR are used to resolve the transposition junction, add the remaining adapter sequence and mildly amplify the amount of DNA. After sequencing library preparation, all libraries were sequenced in the same Illumina MiSeq 2×75 bp paired end run and de-multiplexed to give a pair of FASTQ files containing the raw sequence information for every activity fraction.

2.2.4 TRIAD library composition

The FASTQ files were the starting point of the analysis with the InDelScanner scripts; thus, this section is analogous to the description of composition of the TRIAD libraries in *wt*PTE and is a useful starting point for later discussion of variant enrichment.

Sequencing depth. Figure 2.10 shows the sequencing coverage across the GFP genes for all sequencing libraries, derived from the initial mapping of reads that generates SAM files with all aligned reads (Table 2.3). In all libraries, the sequencing depth is similar across fractions in the core part of the gene, but drops off at the ends of the fragment, which is typical for amplicons sequenced with Nextera library preparation - the transposase has a lower insertion efficiency close to the ends of linear DNA. The horizontal line shows a depth of $>$50,000 reads per base and the shaded area shows the part of the gene where depth is at or above this depth in the sequencing library with lowest coverage (low activity for eGFP and high for GFP8).

A consequence of limited sequencing depth at the ends of the gene is that the following analysis focuses on the parts of the gene with $\geq$50,000 reads per base, which is from

eGFP	**Naïve**	**High**	**Medium**	**Low**
Total reads pairs	1.45×10^6	1.6×10^6	1.68×10^6	1.31×10^6
Assembled read pairs	9.73×10^5	1.08×10^6	1.16×10^6	8.74×10^5
Assembled alignment rate	95.9%	96.1%	95.3%	95.1%
Unassembled read pairs	4.77×10^5	5.11×10^5	5.2×10^5	4.34×10^5
Unassembled alignment rate	96.8%	97.2%	96.8%	96.5%
% Assembled reads	66.9%	67.7%	68.8%	66.6%
Total aligned reads	1.86×10^6	2.03×10^6	2.11×10^6	1.67×10^6
Mean coverage per base	2.12×10^5	2.32×10^5	2.41×10^5	1.91×10^5
Coverage standard deviation	8.12×10^4	9.08×10^4	9.14×10^4	7.22×10^4

GFP8	**Naïve**	**High**	**Medium**	**Low**
Total reads	1.38×10^6	1.21×10^6	1.44×10^6	1.41×10^6
Assembled read pairs	8.89×10^5	7.82×10^5	8.92×10^5	9.63×10^5
Assembled alignment rate	96.0%	95.4%	94.8%	94.8%
Unassembled read pairs	4.83×10^5	4.2×10^5	5.42×10^5	44.69×10^4
Unassembled alignment rate	97.0%	96.8%	96.6%	96.0%
% Assembled reads	64.6%	64.9%	62.0%	68.1%
Total aligned reads	1.79×10^6	1.56×10^6	1.89×10^6	1.77×10^6
Mean coverage per base	2.04×10^5	1.78×10^5	2.16×10^5	2.02×10^5
Coverage standard deviation	7.8×10^4	6.86×10^4	8.54×10^4	7.53×10^4

Table 2.3 The Illumina sequencing statistics for eGFP and GFP8 libraries (Nextera library preparation, 2 × 75 bp paired-end sequencing). All eight libraries are a mix of deletion libraries and the TriNEx library, and were sequenced with equal loading on the flowcell. The assembly statistics are the output of PEAR read assembler and the coverage per depth is calculated with the Samtools depth feature.

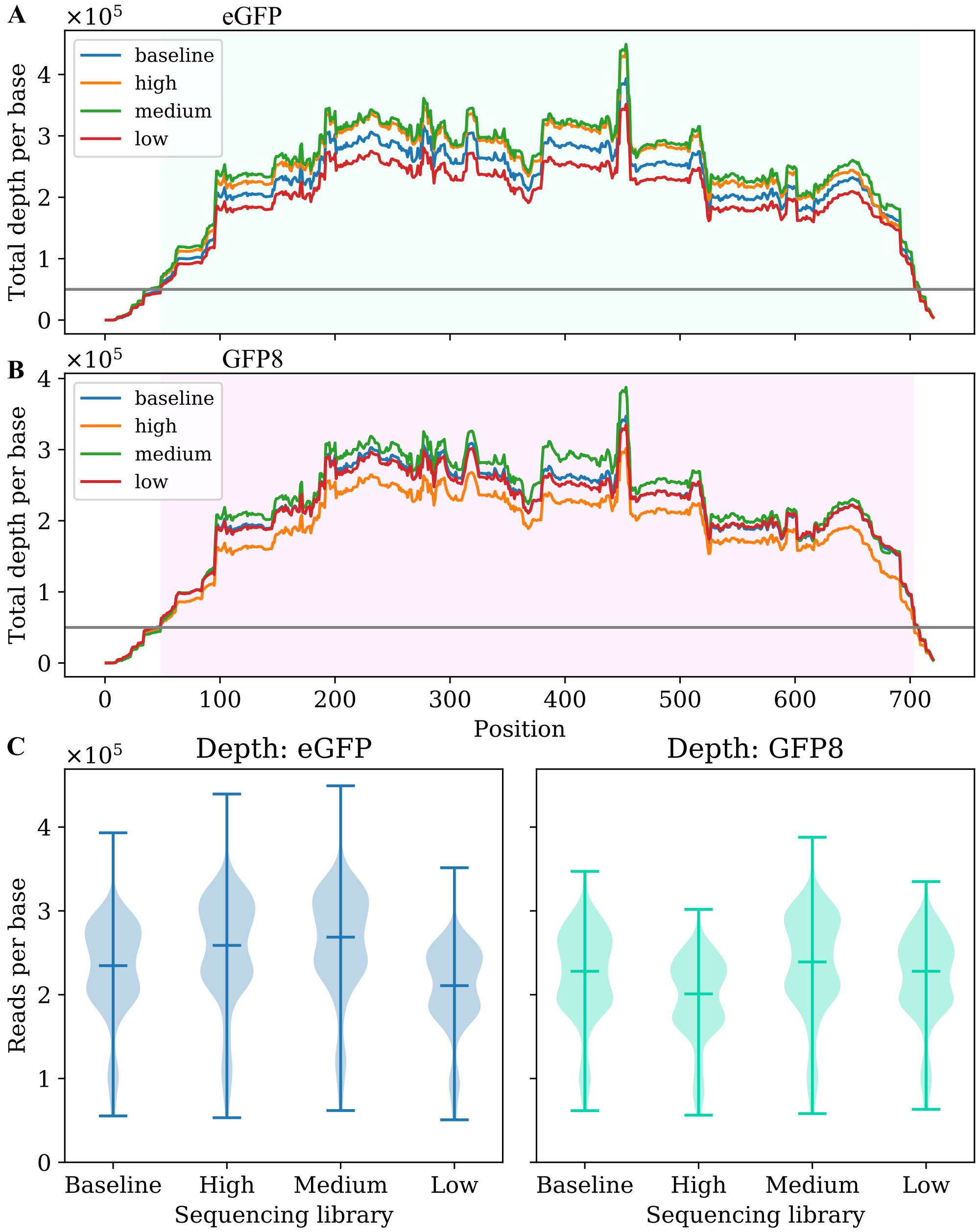


Figure 2.10 The combined sequencing coverage of eGFP and GFP8 sequencing libraries, calculated on the basis of all aligned reads. Each sequencing library pools the DNA from -3, -6, -9 bp deletion libraries and the TriNEx substitution library, according to the FACS activity fraction. A) The coverage across eGFP gene, the parts of the gene with >50,000 reads per base are highlighted in cyan. There is systematic variation in coverage across the gene, likely originating from Nextera transposase sequence preference. B) Coverage in GFP8 libraries. C) Violin plots of sequencing depth per base GFP libraries, from the shaded regions in panels A and B. The median depth and interquartile range are depicted with horizontal lines.

nucleotide 49 (Glu17) to nucleotide 703 (eGFP; Asp234 as the last complete codon) cf. 706 (GFP8; Glu235 the last complete codon). Thus, this sequencing dataset does not have the power to identify the known adaptive variant Gly4Δ, although it is present in the -3 bp deletion library and has been identified in low-throughput screening of eGFP TRIAD libraries. Similarly, not all C-terminal deletions can be identified with these cut-offs, although C-terminal deletions up to 9 amino acids long have been shown as tolerable or even beneficial to GFP fluorescence (Kim and Kaang, 1998).

Diversity of starting libraries

The deletion libraries are comparatively small and well sequenced, so that the majority of variants are observed $\geq 10\times$ in the starting libraries (Figure 2.12). Additionally, the errors in Illumina sequencing are not typically InDels; DNA polymerase error either in library preparation or during the sequencing run generate single nucleotide substitution, or (rarely) single nucleotide InDels from polymerase slippage. Therefore, the 3/6/9 nt long deletions can be reliably identified even at low sequencing coverage.

The deletions are detected across most of the gene length in both eGFP and GFP8 libraries before sorting (Figure 2.11). As expected, there is a decrease in the number of detected variants in the first 100 bp of the gene, which is caused by a lack of sequencing depth in that region.

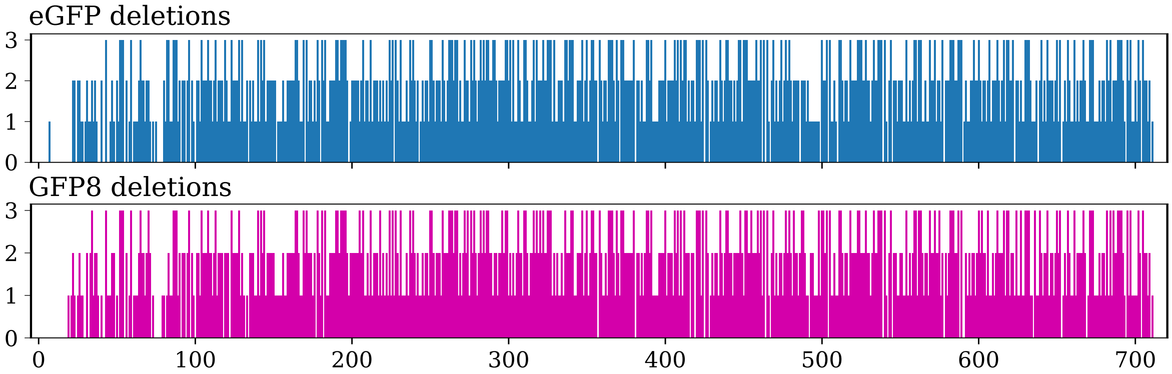


Figure 2.11 Number of detected on-target deletions per gene position in GFP. Each deletion is assigned to the first nucleotide position that is deleted, although of course the each deletion spans multiple nucleotides.

However, the detection of substitution variants is less straightforward; a rare variant could be genuine or a sequencing artefact; a variant with two distant mutations could indicate true variant with an accidental distant mutation or a sequencing error. The reality of these falsely detected variants is clear in Figure 2.12, which shows that >12,000 variants are only

	Deletions			**Substitutions**
eGFP	**-3 bp**	**-6 bp**	**-9 bp**	**TriNEx**
Accessible DNA diversity	527	500	511	33029
Observed unique on-target DNA variants	443	436	394	?
Observed DNA diversity (% accessible)	*84%*	*87%*	*77%*	
Accessible protein diversity	404	389	407	8698
Observed unique in-frame protein variants	349	337	329	5286
Observed protein diversity (% accessible)	*86%*	*87%*	*81%*	*61%*
InDels with no adjacent aa substitution	208	205	200	3108
InDels with adjacent aa substitution	141	132	129	2178
InDels resulting in truncated variants	25 (combined)			386

	Deletions			**Substitutions**
GFP8	**-3 bp**	**-6 bp**	**-9 bp**	**TriNEx**
Accessible DNA diversity	530	504	513	33024
Observed unique on-target DNA variants	434	439	399	?
Observed DNA diversity (% accessible)	*82%*	*87%*	*78%*	
Accessible protein diversity	405	389	408	8659
Observed unique on-target protein variants	339	345	335	5605
Observed protein diversity (% accessible)	*84%*	*89%*	*82%*	*65%*
InDels with no adjacent aa substitution	206	210	203	3208
InDels with adjacent aa substitution	133	135	132	2397
Truncated variants	25 (combined)			383

Table 2.4 Observed diversity of detected variants in GFP deletion and TriNEx libraries. The accessible diversity was calculated by running the InDelScanner pipeline on the set of computationally-generated reads, which covered all possible variants of the desired type (ie. deletions or NNN substitutions of the correct length). The observed diversity of detected deletions can be stated even without filtering, so this table shows *all* detected variants (count = 1 or more). However, the number of observed substitutions is much more dependent on the chosen cut-off value separating real variants from sequencing artefacts, so the total number of detected DNA variants is fairly meaningless - denoted by "?". Once the variants are translated into protein sequence, they can be filtered to only consider those that also occur in the ’ideal’ dataset, so the proportion of observed protein diversity is included for substitution libraries.

observed 5 or fewer times. These rare observation are very likely to be sequencing errors - indeed, manual inspection of a selection of these variants shows that they typically appear in tandem with more frequent variants without distant ’passenger’ mutations.

The majority of artefacts can be filtered from the dataset by only considering variants that appear more than a set number of times. While there is some choice in the designated minimum number of supporting observations, a choice between 5 and 20 reads is quite typical in DMS experiments. I chose the cut-off of 10 observations as a middle option: it is stringent enough to eliminate most (if not all) sequencing artefacts while still detecting 2.5×10^3 DNA variants in each TriNEx library, which is large enough to generate insight into variant tolerance (Figure 2.4).

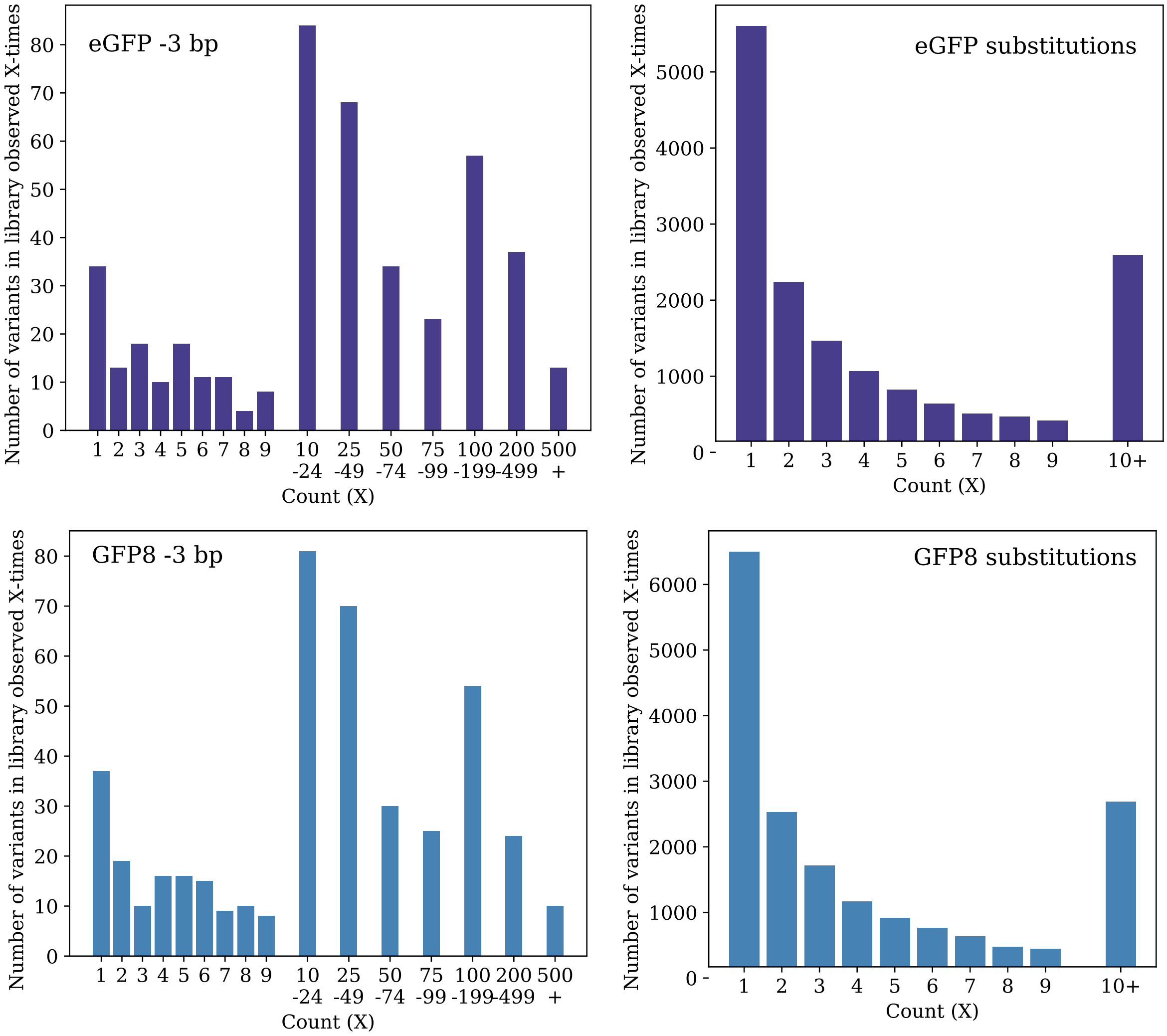


Figure 2.12 Histograms of the number of sequencing observation supporting each detected DNA variant. The X-axes show how many reads support a given variants (the count) and the heights of bars display the number of variants in that bin. Only TriNEx and -3 bp variant distributions are shown here, but -6 bp and -9 bp libraries show a very similar distribution to the -3 bp libraries.

2.2.5 Characterization of individual variants

In addition to testing the achieved resolution in FACS, as shown by the post-sorting distribution of fluorescence in each bin (Figure 2.9), I also individually measured the fluorescence of a set of randomly chosen variants across eGFP libraries. The known fitness (= cell lysate fluorescence intensity compared to eGFP) of these variants could then be used for calibration of statistical models that infer fitness from NGS counts.

The individually measured variant data originated from two sets of measurements:

- Randomly chosen variants from eGFP libraries, testing 48 variants each from four libraries (-3, -6, -9 and TriNEx). These were analysed by Dr. Stéphane Emond.
- Additionally, 46 variants each were selected from two high activity fractions (TriNEx and -6 bp) and two medium activity fractions each (TriNEx and -3 bp).

The fluorescence of these variants was measured in cell lysate in triplicate and Sanger sequenced by LGC Genomics. The resulting dataset was curated to eliminate repeated variants, variants with poor sequencing data, frameshifts and variants there were not observed in the sequencing dataset, giving a total of 218 sequences that were useful for validation of the NGS dataset.

Figure 2.13A shows the fluorescence distribution of the variants in the four sorted fractions and the combined dataset of 218 variants. The distribution of fluorescence in variants from the high bins is an order of magnitude higher than fluorescence in the medium fraction, where variants on average display ~5% of eGFP fluorescence. The combined set of variants spans the entire range from no fluorescence to increased fluorescence levels, with the median fluorescence $> 1\%$ of eGFP level, so it forms a balanced training set for evaluating different statistical interpretative models.

In statistical terms, estimating the fluorescence of variants in the NGS dataset is a process of fitting a continuous dependent variable (relative fluorescence) to multiple numerical independent variables (sequencing counts). Alternatively, the NGS data can be converted into categorical variables, for example by considering whether the enrichment in the high gate is above a certain threshold. To model these data, I considered multiple linear regression (MLR), partial least squares (PLS) regression, a generalised linear model, principal component analysis (PCA) and a combination of PCA with multiple linear regression.

Some need to be met for valid MLR analysis, such that the model needs to have:

- few determining factors (independent variable); this condition is met as the dataset only has two or three contributing variables - the sequencing counts in each gate or the proportion of reads in each gate,

- the factors must not be significantly collinear; if this condition is not met, PLS regression is more appropriate,
- the relationship between the factors and the output should be well understood; here, the prevalence of each variant in the FACS/sequencing bins has a clear link with variant fluorescence.

The condition that is not met by raw sequencing counts is non-collinearity, since more abundant variants tend to have more reads across the board. However, a simple normalisation of sequencing counts into the proportion of total variant reads in each bin meets this condition. Thus, the normalisation takes PLS off the list of methods that need to be considered.

Figure 2.13B shows the fit of a MLR model to NGS scores of 218 individually measured variants. Despite the large dispersion both in fluorescence and the proportion of reads in the three bins, the model shows a good correlation (adjusted $R^2 = 0.80$). In fact, the more complex models (PCA+MLR, PLS) perform no better and so there is little justification for their use here.

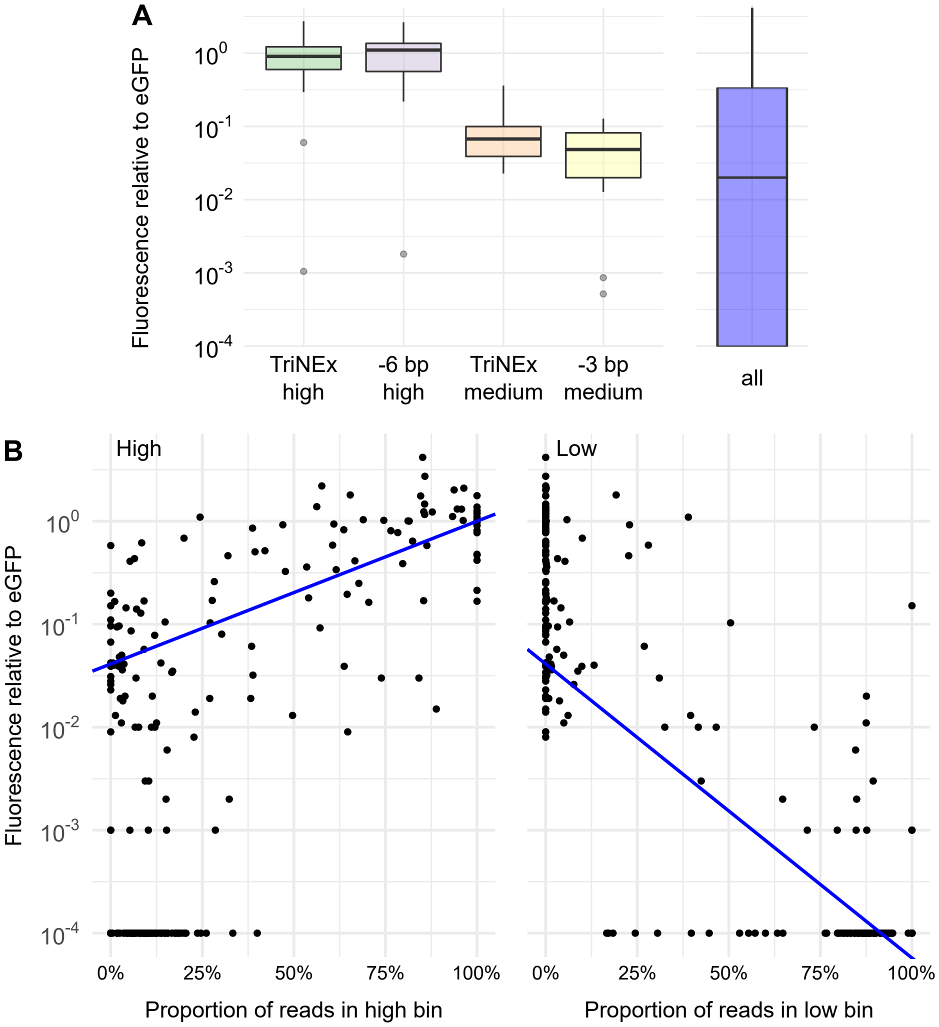


Figure 2.13 A) The distribution of fluorescence in randomly chosen individual variants. All fluorescence values are averages of triplicates and expressed relative to eGFP controls measured in the same experiments. Each box shows the median, the interquartile ranges and the outliers. B) A linear model fitting relative fluorescence (dependent variable) to the proportion of sequencing reads of each variant in high and low sequencing bins (independent variables). The proportion of reads in the medium bin is excluded because it is correlated with high and low bin proportions. Fluorescence of each variant was measured down to 0.1% of eGFP fluorescence and variants below that were indistinguishable from the background. To avoid issues on a logarithmic scale, such variants were assigned the relative fluorescence value 0.01%. The straight line shows the prediction of the linear model: as expected, predicted fluorescence increases with proportion of reads in the high bin (left) and decreases with the proportion of reads in the low bin (right).

2.2.6 The stability effect

Despite the high spread of sequencing counts for many variants, which makes it difficult to quantitatively predict their relative fluorescence, it is possible to categorically describe many variants as low, medium or highly fluorescent. Using that scheme, a variant is categorised into the gate which shows the highest enrichment compared to the baseline sequencing count, and only applied if the variant has at least 20 reads across the sorted bins. Figure 2.14 shows the placement of a selection of such mutations in GFP8 compared to eGFP. Only pure deletions without adjacent substitutions are displayed here.

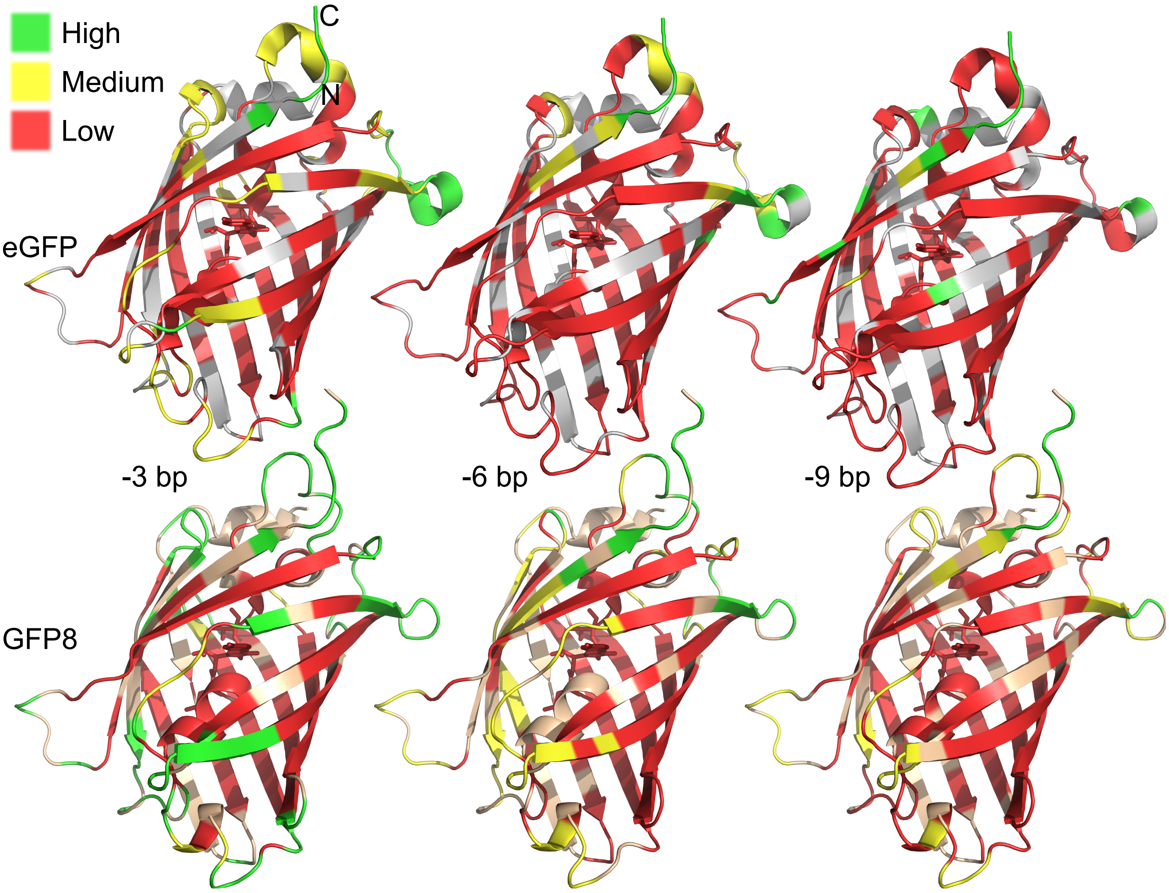


Figure 2.14 Tolerance of pure deletions in eGFP and GFP8 across the crystal structure. Top row: eGFP. Bottom row: GFP8. Lightly shaded positions do not have enough sequencing reads or show too much overlap between bins to assign fitness.

This figure essentially depicts in which sorting gate the variant is most strongly enriched, showing some suggestive trends:

- the improved stability of GFP8 substantially contributes to tolerance of deletions compared to eGFP - deletions that are enriched in the high gate (green) in GFP8 mostly exhibit reduced (yellow) or no (red depiction) fluorescence in eGFP. This is particularly true for deletions within β strands,

- as expected, deletions are more likely to be accepted in loops than in β-strands, and not tolerated at all inside the central helix,
- C-terminal deletions are well tolerated, while the low sequencing coverage prohibits interpretation of tolerance at the N-terminus,
- Some positions in β-strands accept the longer -6 bp deletions better than -3 bp deletions, likely because deletions removing two amino acids preserve the registry of β-strands.

Thanks to only a crude classification method employed here, this result is only preliminary. In the short term, the analysis of deletion effect can be improved by a formal comparison of tolerated position, through construction of a sequence similarity network comparing mutations of different length at different position, and through the use of a better scoring system (see previous section). However, the main limitation in this dataset is limited experimental resolution in FACS, so best improvement in the distinguishable detail can only come through improved sorting (Chapter 3).

2.3 Outlook

This chapter describes my early work on deep mutational scanning with TRIAD libraries, and as such it played a central role in methodological development of the project. In particular, setting up DMS in GFP libraries already shows the three strands of my projects:

1. the importance of a good experimental assay,
2. the impact of the chosen protein starting point on the outcome of the activity selection, and
3. the need for a customised data analysis approach that integrate the analysis of insertions, deletions and substitutions without preconceptions about library design.

In the work with GFP, I hoped to avoid the difficulties with point (1) by using a simple FACS assay, which sorted the bacteria according to their green fluorescence. This approach was experimentally relatively straightforward, such that I achieved moderate sorting resolution without extensive optimisation of expression or sorting conditions. Furthermore, I could have easily selected only eGFP-like or dark variants, by collecting only two gates and in this way deriving a very simple dataset. Instead, the simple selection was more successful in obtaining the third population (intermediate brightness), by simple means of repeating the cell sorting a second time.

While the sorting resolution was far from perfect and was a target for improvement in further work (see Chapter 3), it was good enough to provide some initial insights. First, the choice of the starting point clearly affected its tolerance for mutations: GFP8 was more robust towards mutagenesis both in substitutions and across all three lengths of deletions. Second, the GFP deletions NGS dataset was my test system for testing different software suites for DMS data analysis and for eventual development of InDelScanner scripts (Chapter 4). Third, despite the current popularity of complex modelling strategies and machine learning in sequencing data analysis, my pilot analyses showed that at least in this case, a simple model performs well (Figure 2.13).

A comparison of eGFP and GFP8

During this project, beyond performing a DMS experiment with TRIAD libraries that probes the effect of insertions or deletions, I wished to obtain multiple related datasets; in this way, a comparative analysis becomes possible. As outlined above, the choice of the protein starting point can make a major difference to the outcome of a screening campaign.

Unlike sfGFP, which has been extensively characterised in the initial publication and is frequently used in molecular biology work (Crivat and Taraska, 2012; Day and Davidson, 2009; Pedelacq and Cabantous, 2019; Pédelacq et al., 2006), GFP8 is a less well-described newer variant. To reduce this gap of information, I obtained a crystal structure of GFP8. The protein expressed well, and was purified with immobilised metal-affinity chromatography and gel filtration in good yield. The crystallisation experiments started with an overview of reported crystallisation conditions in literature, which did not yield useful hints for quick crystallisation. Therefore, I tested conditions with three commercial sparse-matrix crystallisation plates and further refined hits with manual screening of condition - though coincidentally, the final best-diffracting crystal grew in the buffer identical to the initial hit. The highest-resolution diffraction dataset was successfully solved to 1.29 Å resolution and deposited with the Protein Data Bank under the accession code 6HUT. Overall, the process of obtaining the crystal structure was straightforward, if somewhat time consuming.

Trying to understand the structural origin of differences in stability and mutational tolerance between GFP8 and sister constructs, I compared this crystal structure with the published crystal structures of eGFP, frGFP and sfGFP (Figure 2.3). The GFP8 backbone aligns very closely to eGFP and frGFP, in line with the notion that all GFP structures are essentially the same. Closer examination shows that the main difference amongst the altered side chains appears in position 30, occupied by Ser in eGFP and Arg in GFP8/sfGFP. Curiously, coordination of a polar electrostatic network by Arg30 in sfGFP has been credited with increased solubility and stability of sfGFP, yet Arg30 occupies an entirely different conformation in GFP8. This difference suggests that the stabilising effect of Ser30Arg may instead function through a general solubilising effect, increasing the polarity of the protein surface, and the exact surface hydrogen bond networks play a lesser role.

The temperature factors in the four structures (Figure 2.5) largely show similar trends in atomic motion across the structures: motion is highly constrained in the β-strands and the central helix, whilst loops show substantial flexibility. There may be some difference in the relative degree of motion between the loops and the protein core, but interpretation is complicated by varying crystallographic resolution in the four structures.

Overall, the new GFP8 crystal structure is very similar to crystal structures of related GFP constructs. On the other hand, GFP8 does show noticeably better mutational robustness and accepts a larger proportion of both InDels and random trinucleotide substitutions. These observations can be reconciled in two possible ones. One, the similarity in crystal structures suggests that the stability effect in GFP8 is an example of 'global' stabilisation, which improves the stability and solubility of the protein in general and thus causes overall higher tolerance of disruptive mutations, not constrained to a specific local region. Alternatively, the differences may originate from a change in protein dynamics in solution (difficult to interpret from the fixed 'snapshots' in crystal structure) or from changes in the properties of the protein unfolded state.

Chapter 3

Screening the effect of insertions in GFP in high throughput

In this chapter I build on the pilot DMS experiment presented in Chapter 2 by improving the process to achieve higher resolution in sorting and consequently in sequencing. Two substantial modifications are introduced: an expression control and UMIs placed on the library plasmids.

3.1 Introduction

Chapter 2 described the set of pilot experiments that were used to set up the deep mutational scanning work-flow on GFP. The deletions dataset was my test system on which I developed InDelScanner scripts, built my intuition on the effect of InDels and explored different strategies for statistical modelling. While the deletions dataset served its purpose in project development well, its interpretability was effectively bounded by the limited resolution in sorting. Here, I show my progress in establishing an improved protocol, which is designed to surpass those limitations.

First, I chose to move the libraries into a new fusion expression construct, which links expression of GFP variants to the far-red fluorescent protein mKate2 (Shcherbo et al., 2009). The design of the fusion construct was reproduced from Sarkisyan et al. (2016), who selected the fusion protein on multiple criteria: that it has a fast-maturing chromophore, exists in a monomeric state, lacks a green fluorescent state during chromophore synthesis and exhibits a large spectral distance from green fluorescence of GFP. They tested co-expression of GFP and mKate2 in a bicistronic construct, a fusion construct with a flexible linker and a fusion construct with a rigid α-helical linker, and evaluated the correlation in protein levels in individual cells. Neither linked construct generated measurable FRET transfer between mKate2 and GFP. Both linked constructs had reduced scatter of GFP expression relative to mKate2 expression, compared to the bicistronic system.

For my improved expression system, I adopted mKate2, the rigid α-helical linker, and took inspiration from their design of technical sequence surrounding the C-terminally placed unique molecular identifier (UMI). The mKate2 fluorescent spectrum (excitation maximum 588 nm, emission maximum 633 nm) does not overlap with the spectrum of eGFP (excitation maximum 488 nm, emission maximum 507 nm). Additionally, the mKate2 spectrum is well-suited to the yellow-green laser excitation (561 nm) available in FACS instruments, so it can be detected with filters designed for Texas Red or similar.

The protein expression construct used in the experiment described here has the following features designed into and around the open reading frame (Figure 3.3):

- an N-terminal 6×His-tag, which is in place to facilitate purification of individual variants in downstream work, if required,
- the mKate2 gene without interfering restriction sites,
- a rigid linker which has been shown to prevent interactions and FRET transfer between fluorescent proteins (Arai et al., 2001), flanked by EcoRI and NdeI restriction sites,
- a space for GFP: this is the eGFP sequence in the control construct or the TRIAD libraries, which can be directly cloned in with NdeI and HindIII restriction sites,
- a TAA stop codon followed by a T nucleotide, which enhances the strength of the stop signal in translation (Poole et al., 1995),
- a high quality UMI consisting of 21 random N nucleotides, and
- sequence matching the Illumina sequencing primer immediately following the UMI sequence, enabling easy amplification of the UMI after functional selection.

This construct is designed to improve the resolution of FACS fitness separation through measurement of protein expression in a separate fluorescent channel, and improve the utilisation of Illumina sequencing with the use of UMIs. Because each UMI encodes a distinct variant, only the UMIs need to be sequenced after sorting; thus saving on the amount of sequencing required. Furthermore, simpler post-sorting sequencing enables the comparison of sorting replicates, or the enrichment of variants under different selection conditions, without incurring the high cost and difficulty of full long-read library sequencing.

3.2 Results

3.2.1 Assembly of the mKate2::GFP::UMI construct

The initial mKate2::GFP construct

I first assembled the control construct, in which I combined the three protein elements (mKate2, the helical linker and eGFP) to assess the properties of this fusion protein. Since the individual components originated from different sources, I used Gibson assembly to link three fragments:

- mKate2: I obtained the mKate2 gene from a plasmid available on Addgene (pVHed02). The N-terminal His-tag was added in a primer-overhang during PCR amplification, creating a long linear DNA fragment. The amplification also introduced a substitution in the 5' region of the amplified gene, which removed an unwanted NdeI restriction site,
- the rigid linker sequence: this small fragment was added in a second small insert, created from annealing of two long oligonucleotides that overlap in the middle of the linker, each extending beyond the boundary to create the extra sequence required for assembly,
- GFP and the vector backbone: the vector fragment was amplified from the pID eGFP::N18 plasmid, that is the plasmid used in testing the deletion libraries. I used this construct because it does not contain a C-terminal His-tag, which I wished to avoid, and the 18-nucleotide random sequence originating from the UMI can be removed in a later step.

The two longer DNA fragments (eGFP with the plasmid backbone, mKate2) with appropriate overhangs were prepared with PCR reactions, the reactions checked with agarose gel electrophoresis and the fragments column purified (see Methods for details). The third, shorter fragment with the linker sequence was by denaturing and annealing two long oligonucleotides (Linker F and Linker R). The three DNA fragments were combined into a single construct with Gibson Assembly and a correct clone identified with Sanger sequencing (Figure 3.1).

Expression of the mKate2::eGFP construct in pID plasmid showed the desired red and green fluorescence in liquid culture. However, first testing of fluorescence with flow cytometry showed that the expression levels of mKate2 were variable between different days and dependent on the vector backbone. In order to address the question of optimal expression conditions, I expressed the fusion construct in both pID and pET plasmids, at three

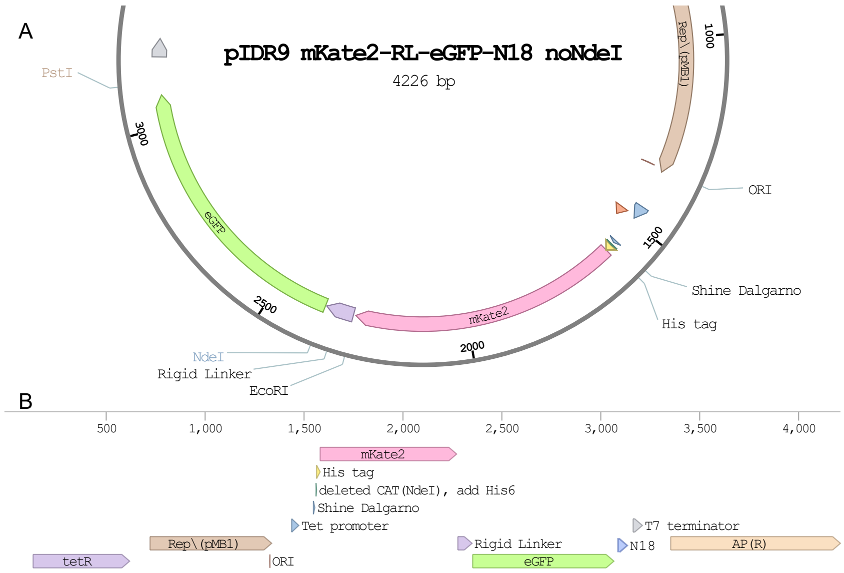


Figure 3.1 The plasmid map of the pID plasmid containing the mKate2::eGFP fusion construct. A) The relevant part of the plasmid map in the pID backbone, showing the His::mKate2::linker::GFP expression casette. B) The plasmid sequence in linear view, illustrating the small changes in this vector compared to pID::GFP. In this construct, the NdeI restriction site has been removed from mKate2 and the randomly chosen N18 sequence from previous designs remains.

different temperatures. The results (Figure 3.2A, B) indicated that expression levels of both proteins were optimal when expressed at 30°C, and that the pID anhydrotetracyclin-inducible expression of mKate2 was more robust than expression in the pET vector. The suitability of these expression conditions was confirmed with flow cytometry measurement: while the expression of mKate2 on its own remains variable between cells, the fusion construct is well behaved (Figure 3.2).

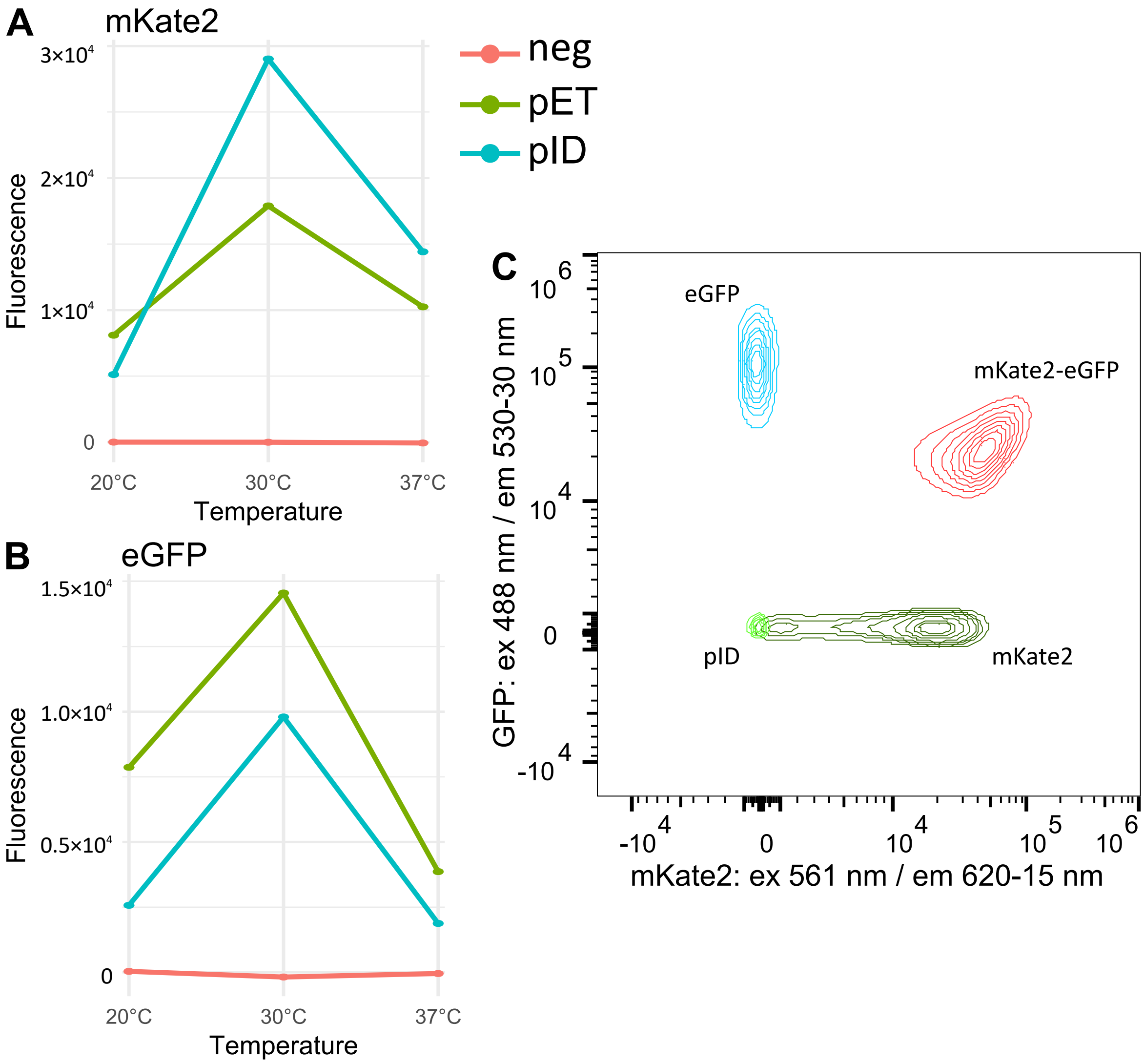


Figure 3.2 The levels of mKate2-eGFP fusion construct expression in bulk culture and at single cell level. A) The mKate2 fluorescence signal in bulk culture, excited at 580 nm and emission recorded at 630 nm. B) GFP bulk culture fluorescence signal, excited at 488 nm and recorded at 520 nm. Both bulk signals are normalised to the culture cell density, which accounts for the decrease in signal at 37°C. C) Contour plot of single cell expression of four controls samples. GFP is excited with the 488 nm laser and the emitted green light is recorded with the filter centred at 530 nm with bandpass width 30 nm, which captures light between 515 and 545 nm. Similarly, mKate2 is excited with the yellow-green 561 nm laser and emission recorded in the 620/15 nm bandpass channel.

Cloning TRIAD libraries into an UMI-tagged pasmid

Having established suitability of the mKate2::linker::GFP fusion construct, I started tagging the fusion construct plasmid with unique molecular identifiers and cloning TRIAD libraries into the new construct. Throughout the process of transferring the existing libraries I was careful to avoid PCR reactions *on the libraries*, because PCR on a library template creates a risk of unintentional shuffling between individual template molecules. The shuffling is especially problematic for insertion libraries, which contain 30% or more frameshifted variants - even a low proportion of shuffled sequences per cycle could further increase this proportions.

Instead, I introduced 21-nucleotide-long random unique molecular identifiers and the Illumina primer annealing sequence with PCR reactions into the plasmid backbone sequence. Additionally, I wished to keep the new UMI sequence as close to the end of the GFP sequence, to maximise the length of variants that can get sequenced in an NGS run. Therefore, my desired sequence required careful engineering of that part of the pID plasmid sequence.

Because of the restrictions of the existing sequence, the PCR reactions were rather challenging. The issues originated from the many constraints on primer design at the GFP C-terminal end, where the final GFP sequence is unfortunately repetitive and not ideal for specific primer annealing. Consequently, the PCR reactions on the full pID mKate2::GFP template (Figure 3.1) were very dirty.

Eventually, I solved the problem with a two step protocol:

1. Amplification of the plasmid backbone and most of the expression cassette in the pID mKate2::GFP template, except for most of GFP sequence and the unnecessary remnant N18 sequence (which is no longer random because of working from a single clone). This PCR reaction still gave multiple bands, but I could excise the correct band from an agarose gel and confirm its identity with Sanger sequencing. This step removed the difficult GFP C-terminal sequence, which enabled good primer design for introduction of UMIs.

2. Introduction of UMIs and the required cloning sites, so that the TRIAD libraries could be cloned directly into this new plasmid (Figure 3.3, where GFP stands for a library of variants).

In contrast, amplification of UMIs from the final construct was straight-forward, giving a clean amplicon bands using the Illumina sequence primer on one side and 30-nt primer matching the C-terminal sequence of GFP.

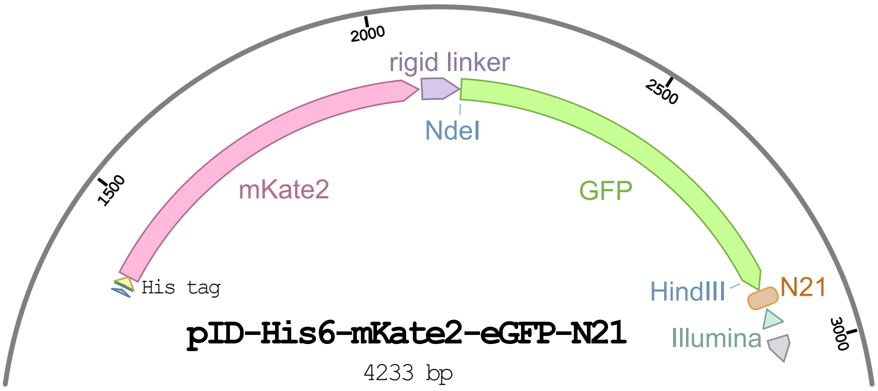


Figure 3.3 A plasmid map of the mKate2::eGFP::UMI fusion construct, which incorporates all desired features. The GFP TRIAD libraries were cloned into this construct using the NdeI and HindII restriction sites indicated. The stop codon after the HindIII site (not shown) is immediately followed by a 21 nucleotide long UMI, which contains hand-balanced random nucleotides. Finally, the Illumina technical sequence for annealing of read primer 1 directly follows the UMI.

After finding a successful set of PCR primers and conditions to create UMI-tagged vectors, I repeated the second step of construct taggind and prepared the master UMI-tagged plasmid with high quality, hand-mixed random N nucleotides. At that point, the existing GFP TRIAD and TriNEx libraries were cloned into the UMI-tagged construct, thereby avoiding PCR bias in the libraries.

Protein expression considerations

Repeated testing of expression conditions defined the following considerations:

- The fusion construct mKate2::GFP does not express properly in Lucigen 10G cells, although GFP on its own is not sensitive to the expression strain used. BL21(DE3) is preferred.

  The reason that this issue appeared was that the pilot sorts (Chapter 2) were performed in the 10G strain, which is optimised for general cloning work and therefore may or may not be suitable for protein expression. Using a single strain for both DNA manipulation and protein expression can be problematic, since DNA fidelity may be degraded after long propagation in a strain such as BL21(DE3). After I understood the issue with cell strains when working with mKate2, I used BL21(DE3) cells for fitness sorting and all further steps; in fact, using this strain it is possible to simplify the post-sorting flow cytometry measurements by avoiding a second transformation step.

- The exact level of mKate2 tends to show some day-to-day variation even when expressed in BL21(DE3) cells, but this does not negatively impact sorting resolution. The key element is ensuring that the expression time is long enough (3-5 hours after induction).
- Surprisingly, insertions are very poorly tolerated in both GFP backgrounds, to the extent that the histogram of GFP fluorescence closely resembles the negative control. This can create the misleading appearance that the sample is not expressing well or the protocol is otherwise faulty - but the positive variants are in fact present in the low frequency, high green fluorescence tail.

Reduction of UMI diversity

As shown in Figure 2.6, UMIs can be used for error correction, increasing the available sequencing length through subassembly, and to reduce the amount of sequencing needed by creating a variant-UMI dictionary. For all of these applications it is essential that within one library, a single UMI is only associated with a single variant, or the meaning of the UMI becomes scrambled.

However, the unique linkage between UMIs and variants is not bi-directional: a single library variant will typically be linked with multiple UMIs. The number of UMIs per variant is variable, so that more abundant variants are tagged with a greater number of UMIs. In this manner, UMIs have the added benefit that the fitness of each tagged variant is independently measured multiple times - the fitness of each variant-UMI pair is assessed independently. This independence of measurement also shows why the number of UMIs should not be too high: each UMI must still be sequenced enough times to assign it a meaningful fitness score. Therefore, the experimental plan must consider that the addition of UMIs to a library will raise the diversity by approximately an order of magnitude. In a DMS experiment, a good aim is between 5 and 20 UMIs per variant on average.

The theoretical diversity of 21 nt random N UMIs is 4^{21} or approximately 4×10^{12}, which exceeds the theoretical diversity of the largest insertion library (+9 bp, $> 10^8$ variants) by more than three orders of magnitude. These UMIs are therefore diverse enough that all variants in the TRIAD libraries can be uniquely tagged with these UMIs.

Deriving the target UMI diversity. An Illumina MiSeq sequencing run generates approximately 10 to 20 million pairs of reads recording the target sequence with a 20% PhiX spike-in. Since the post-sorting DNA is more efficiently sequenced with a short-read NextSeq run, the more expensive long-read MiSeq only needs to cover the input library sequence.

The six deletion libraries are small, containing about 500 variants per library. Allocating approximately 15 UMIs per variant and 10 sequencing reads per UMI, all deletions libraries can be sequenced with 0.5 million sequencing reads. Therefore, the total diversity of deletion libraries must be reduced to approximately $500 \times 15 = 7500$ uniquely tagged variants per deletion library, or around 2×10^4 variants in each pooled deletions library (one for eGFP and GFP8). I reduced the diversity of the deletions library to this level with a restrictive transformation into 10G *E. coli*, where I plated the post-recovery cell culture at increasing dilution and scraped the appropriate number of colonies from a large agar plate.

After allocating the reads to deletion libraries, that leaves approximately 19 million reads to insertion and substitution library sequencing. A similar calculation shows that ~200,000 UMIs per library can be efficiently sequenced and still fit into a single Illumina MiSeq run. While this value is too low to fully capture the starting diversity of insertion libraries, it is more than enough to sequence the insertion variants that do exhibit some GFP fluorescence. Because the proportion of positive variants in GFP insertion libraries is low ($< 5\%$), I decided to not reduce UMI diversity through restrictive transformation and instead performed an enrichment sort for non-negative variants. The diversity of the starting libraries is thus restricted during pre-sorting of the insertion libraries - this sort is very simple, selecting for all variants with GFP fluorescence above the negative control, but it does take several hours to collect enough sorted variants.

3.2.2 FACS with utilisation of the mKate2 expression control

After repeated testing of *E. coli* strains and expression conditions, I tested the FACS set-up and gating strategies before starting the intended four-way sorting (Figure 3.4). I observed that the variants could be roughly separated according to GFP fluorescence by first setting a gate on mKate, that is only sorting cells that express a certain level of mKate2. This gating strategy reduced the noise from stochastic variation in protein expression levels. However, there is some negative correlation between the mKate2 and GFP signal intensity: the variants with highest GFP fluorescence levels show mildly reduced mKate2 brightness. This effect is likely due to additional metabolic stress placed on the host bacteria when both fluorescent proteins undergo oxidative maturation of the two fluorophores. At the same time, mKate2 plays its intended role and shows correlation between expression and brightness at individual cell level.

In this first four-way activity sort, I inoculated expression cultures with cell suspension from a bacterial lawn grown on an agar plate, then expressed the fluorescent proteins at 30°C

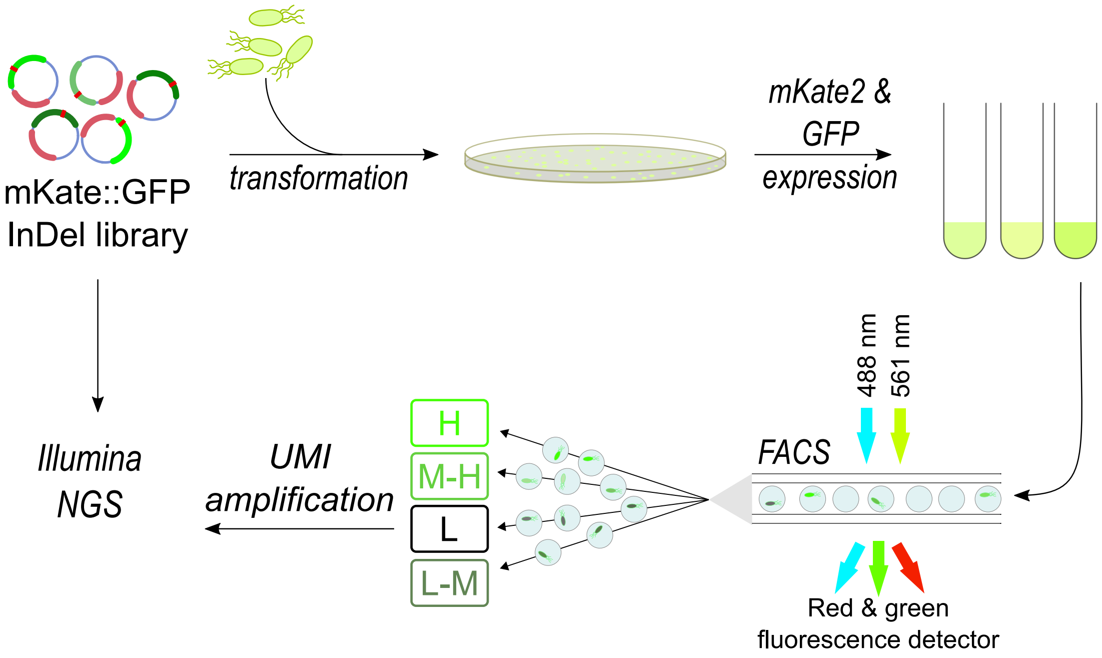


Figure 3.4 An outline of the improved DMS experiment with mKate2::GFP libraries. The starting InDel libraries, located in the mKate2::GFP::UMI construct, are separately transformed into the appropriate highly competent *E. coli* strain and grown overnight on large agar plates. The plates are scraped before mKate2 and GFP expression in liquid culture, then the individual cells are sorted with based fluorescence excited by the 488 nm (GFP) and 561 nm (mKate2) lasers. The plasmid DNA in each fraction is recovered, the UMIs amplified from the sorted fractions and sequenced.

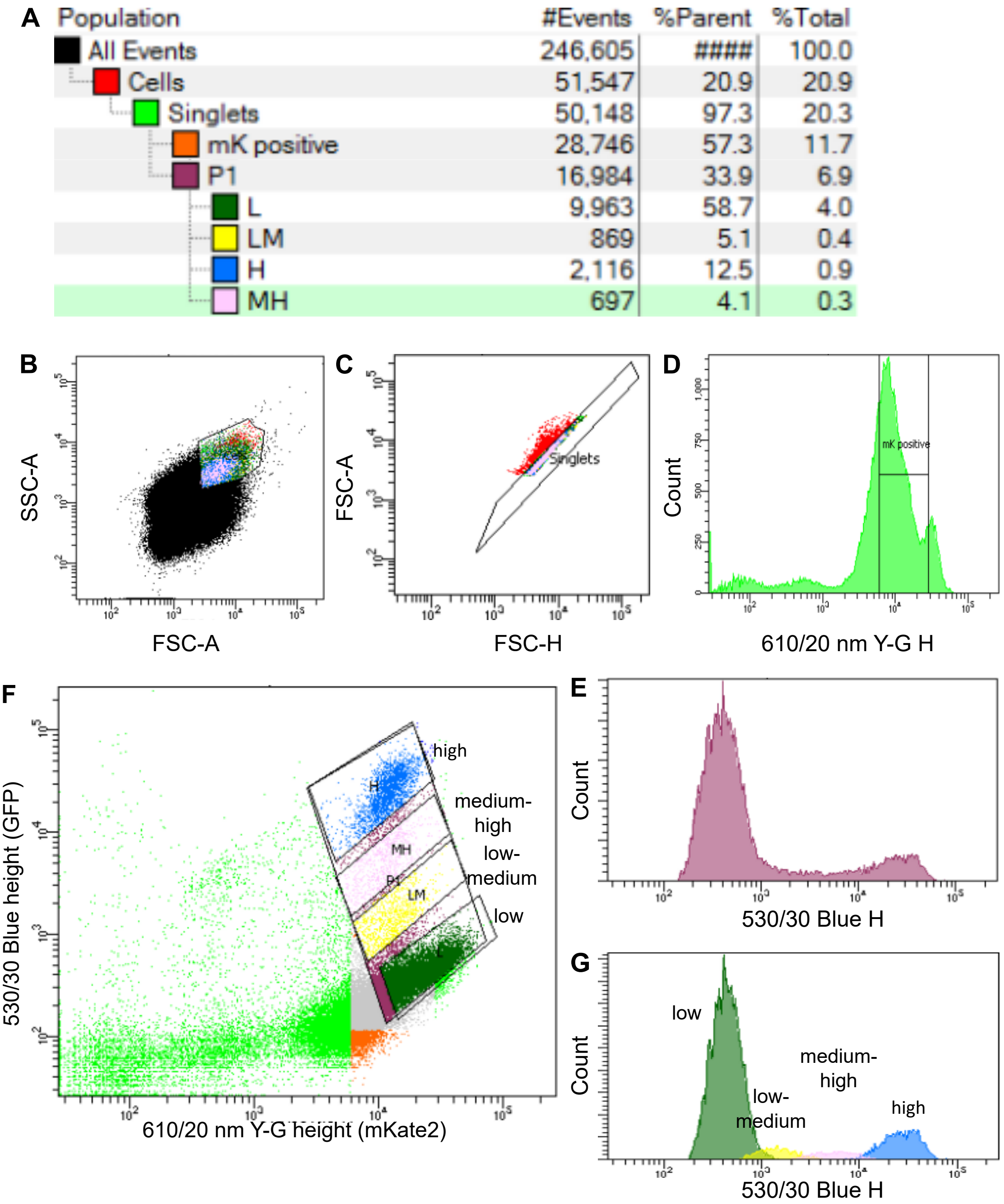
for three hours. After preparing the cell cultures for sorting, the cells were first isolated in the forward/side scatter plot and further gated to isolate singlet cells from doublets (Figure 3.5B, C). The singlet cell population exhibited somewhat variable expression of mKate2 (Figure 3.5D) and the expected two-peak distribution in GFP fluorescence (Figure 3.5E).

There is a distinct correlation in expression levels of both proteins, so I used diagonally slanted gates for the four-way sort (Figure 3.5F, see panel A for the gate hiearchy). The slant is clearest in the large population in the low gate, but it remains in the higher activity gates as well. Finally, just the GFP fluorescence of cells in the four sorting gates is shown for comparison (Figure 3.5G). The green fluorescence signal in the four sorting gates illustrates the value of slanted gates - the four populations have a similar width of distribution of fluorescence (in log scale), which suggests the gates are well defined. For both mKate2 and GFP signal, the cells were gated on signal height rather than area: while the two measures are similar, gating on signal height can give higher resolution when sorting small particles such as bacteria.

After sorting, I collected the cells in LB medium and re-grew them on large agar plates, which recovered between 10% and 60% of sorted events as bacterial colonies. I chose re-growth on agar plates rather than in liquid culture, because the agar plate method allows quantification of cell survival after sorting. Additionally, liquid culture can bias the proportion of different variants towards fast-growing cells, which should be avoided.

Finally, I scraped the bacterial lawns and used the majority of the suspension to isolate plasmid DNA, and a small aliquot of bacterial suspension to inoculate fresh expression cultures. Fluorescent proteins were expressed in the same manner (30°C, three hours) and analysed with flow cytometry (Figure 3.6). The post-sort fluorescence distribution shows that the four cell populations remain separated, although all non-negative fractions have a proportion of negative variants. Therefore, the position of medium fluorescence gates may need to be adjusted, or a two-step sorting strategy applied to narrow the distribution of low-medium and medium-high fractions.

This experiment demonstrates that the mKate2-GFP fusion construct is well behaved and can be efficiently sorted into four separate activity fractions. The experimental protocol for the sorting of both deletion, insertion and substitution libraries in GFP has been optimised and is effective for sorting the different TRIAD libraries. Although beyond the scope of this thesis and not all sorting outcomes are shown, the first sorting of all TRIAD libraries described in this chapter has been completed and is ready for NGS library preparation.



Population	#Events	%Parent	%Total
⬛ All Events	246,605	####	100.0
🟥 Cells	51,547	20.9	20.9
🟩 Singlets	50,148	97.3	20.3
🟧 mK positive	28,746	57.3	11.7
🟪 P1	16,984	33.9	6.9
🟩 L	9,963	58.7	4.0
🟨 LM	869	5.1	0.4
🟦 H	2,116	12.5	0.9
⬜ MH	697	4.1	0.3

Figure 3.5 Four-way sorting for GFP activity in the mKate2::GFP::UMI construct. A) The hierarchy of gates and the proportion of total cell population present in each gate. B) Gating for bacterial cells with forward and side scatter area, based on control samples. C) The bacterial cells were further separated into singlet and doublet cells, based on forward scatter area vs height. D) The singlet cells show a distribution of mKate2 expression that has a major peak with a shoulder of very high expression. The 'mK positive' gate shows the position of mKate2 expression in the positive control, but was not part of the four-way sort. E) The distribution of green fluorescence in sorted cells. F) The four sorting gates, showing how the slanted gate strategy incorporated relative normalisation of GFP signal according to the amount of mKate2 present. G) The distribution of green fluorescence in sorted gates during sorting. In all charts, 'H' refers to signal height (maximum intensity) and 'A' is the signal area (integrated signal).

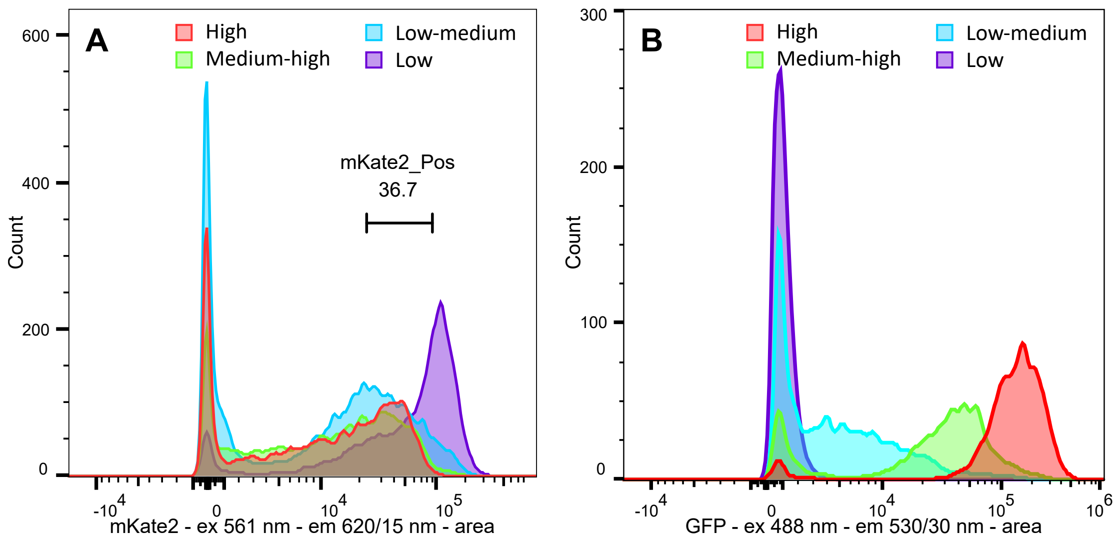


Figure 3.6 The fluorescence distribution of GFP8 pooled deletion libraries after four-way sorting, as analysed with flow cytometry. The cells were gated on forward and side scatter (not shown), then on expression of mKate2 (gate in panel A). All four post-sorting samples express mKate2 reasonably, more so if they express less GFP. Panel B shows the distribution in GFP fluorescence, which very well follows the expected order in the four sorted fractions. All libraries show a peak in the negative fractions, which is very minor for the high and medium-high libraries as expected.

3.3 Outlook

In this chapter I build on the pilot experiment presented in Chapter 2, improving the process to achieve higher resolution in variant sorting and consequently in deep sequencing.

First, I used the experience I gained during the pilot experiment with GFP deletion libraries to improve the experimental design, this time including the UMIs within carefully designed surrounding sequence. The new UMIs are more balanced, are placed immediately after the stop codon and are followed by the sequence of the Illumina sequencing primer. This technical sequence following the UMIs serves as an attachment point, first for isolation of the UMIs and then for addition of the full Illumina adapter sequence (comprising sequencing primer annealing sites, multiplexing indices and the flow-cell annealing sequence). Thus, the UMIs can be easily amplified from the sorted libraries and directly incorporated into Illumina amplicon sequencing libraries.

Second, I transferred the TRIAD libraries into a new expression construct that fuses GFP variants with the far-red fluorescent protein mKate2, adapted from Sarkisyan et al. (2016). After careful optimisation of expression conditions, I was able to perform four-way sorting of combined GFP deletion libraries, which resulted in better separated populations after a single sort.

While testing the expression and sorting conditions with TRIAD libraries in the mKate2::GFP construct, I observed that the insertion libraries display a much lower proportion of positive variants than deletion and substitution libraries. Because of the low tolerance of both eGFP and GFP8 for insertions, I adapted the sorting strategy for insertion libraries. Since the insertion libraries cannot be fully sequenced at the start (at least not at a reasonable cost), the library diversity needs to be reduced before the activity sorting. Instead of performing a restrictive transformation, I adapted the insertion library protocol to starts with an enrichment sort for non-negative variant, followed by a more stringent four-way sort and deep sequencing. In this way, the deep sequencing dataset will be enriched for tolerated mutations, rather than describing many ways for disrupting a protein fold.

The low tolerance of insertions in GFP was surprising, since insertions were better tolerated and more adaptive when tested in phosphotriesterase (Emond et al., 2020). Furthermore, it is conceivable that β-strands may be able to accommodate amino acids insertions by bulging out the extra sequence, but this is evidently not the case.

This project is still ongoing and will be completed in the near future. After extensive optimisation of the expression and sorting conditions, the main selection for maintenance of GFP green fluorescence is now complete for all TRIAD and TriNEx libraries in both eGFP and GFP8. After completion of this project with NGS data acquisition, the dataset will hopefully show which InDels are accepted across the entire protein, and therefore illuminate the specific requirements for insertion tolerance in the β-sheet rich GFP.

Chapter 4

InDelScanner: scripts for amplicon data containing InDels

This chapter describes the scripts and tools used to analyse the InDel libraries, which I sequenced with Illumina next-generation sequencing (NGS). The analysis required the use of custom scripts for a comprehensive analysis. In this chapter, I present the work leading up to development of InDelScanner scripts and their application in characterising the composition of PTE TRIAD libraries.

4.1 Introduction

The deep sequencing of both starting and sorted libraries is an essential step in the deep mutational scanning (DMS) work flow. Practically, deep sequencing analysis comprises three steps:

1. Pre-processing of sequencing reads: includes quality control, trimming of adaptor sequences, read merging (if using paired-end sequencing), and the removal of PCR duplicates,
2. Identification of mutations present in each read: can be done directly with amplicon sequencing of substitution libraries, but requires prior alignment of each read to reference sequence for InDels,
3. Statistical analysis and model building: the final step pulls the whole NGS dataset together, such that it is possible to estimate the fitness of individual variants from aggregated counts and create predictions about protein features that govern the properties of interest.

After the first publications in 2011, deep mutational scanning is now an established approach. Several variations of this workflow have been demonstrated on protein substitution libraries (Doud and Bloom, 2016; Romero et al., 2015; Starita et al., 2013) and some guidelines on the design of the statistical analysis have been proposed (Matuszewski et al., 2016; Starita and Fields, 2015b). Some groups have made their analysis scripts available for general use (Bloom, 2015; Fowler et al., 2011). However, including InDels in the published analysis procedures is not straightforward.

Primarily, the bioinformatic toolkit needs to (1) accurately identify mutations and (2) enable further statistical analysis. I knew from the start that especially point (1) - library variant identification - could be difficult, because existing software tools were developed for different applications. Therefore, this chapter first presents my attempts to utilise existing tools. Since those turned out to be unsuitable, I then describe the development of custom scripts that were targeted for working with InDel libraries. These scripts were primarily developed on the short read sequencing dataset obtained from GFP deletion libraries.

This chapter focuses on the computational methodology development, while Chapter 2 describes in detail how the GFP libraries were expressed, sorted and the results of the data analysis. Therefore, I only briefly describe the experimental protocol here. Four variant libraries (substitutions, -3 bp, -6 bp and -9 bp deletions) per construct (eGFP and GFP8) were sorted into three activity fractions (high, medium and low fluorescence) with FACS. The high activity fraction was gated on the positive control (eGFP) and the low activity fraction was gated on the negative control (empty plasmid).

The libraries were sorted separately, but for sequencing the DNA was extracted from each background-library-activity fraction and the fractions pooled across InDel libraries to give eight sequencing libraries: four for eGFP and four for GFP8, one each for the starting library, high, medium and low variants. For example, the GFP8-medium sequencing library contained the medium activity variants from all four GFP8 libraries: substitutions, -3, -6 and -9 bp deletions. I chose to pool the variant libraries in this pilot first sequencing run to substantially reduce the cost of sequencing library preparation, reasoning that the different types and lengths of mutations can later be separated computationally. Consequently, the data files that described this DMS experiment contained a mix of different InDel variants in the same file, at random positions.

The sequencing libraries

These sequencing files, a pair of FASTQ files for each sequencing library from a 2×75 bp MiSeq sequencing run, have the following key features:

1. Around 10% of reads contain a mutation, the rest are wild-type sequence only (each DNA fragment from a ~750 bp amplicon was only sequenced with two 75 bp long reads). Therefore, the proportion of bases that encode a mutation is very small.
2. The reads are short and do not share a common start/end point, that is they start and end in random positions in the amplicon. As a consequence, the reading frames differ and are unknown.
3. Since the sequencing libraries were prepared with the Nextera system, the position of the mutations (if any) in a read is likewise random and may be in the centre of the read or at the ends. The analysis scripts therefore must not make any strict assumptions about the location of reads and/or mutations, such as assuming that the first nucleotide in each read is also the first nucleotide in a codon (thus implying a reading frame).

4.2 Results

4.2.1 Finding InDels in DMS data with existing software suites

This section describes the attempts to analyse the initial GFP deletions Illumina-sequenced DMS dataset with existing software tools.

Bowtie2 and Samtools

In the first attempt, I used existing tools for working with FASTQ files that are typically used for analysis of genomes, exomes and transcriptomes. First, because genes containing InDels have a different length than the parental sequence, the reads had to be aligned to reference to find the variant position. At the time of the initial work (2017), the recommended state-of-the-art sequence alignment tool was Bowtie2 (Langmead and Salzberg, 2012). Bowtie2 uses an indexed reference sequence and supports alignment of paired-end sequencing reads directly, which was used to generate a SAM file (standing for Sequence Alignment/Map file) cf. BAM (binary SAM) file.

The SAM files describe the mapping location of each read and includes three key pieces of information: the position of the first aligned nucleotide; a CIGAR field, which uses a shorthand format that describes whether the read is a complete match, contains substitutions, insertions or deletions; and the alignment quality score.

After aligning, I attempted to examine the quality of the alignment and identify variants using the Samtools `mpileup` tool (Li et al., 2009). However, here it becomes apparent that this workflow is intended for a different use case, where the reference sequence is many orders of magnitude longer than reads; consequently the read depth is in the range 1-100× coverage. In amplicon sequencing, the reference is only one or two orders of magnitude larger than the sequence, so the read depth is very high ($> 50,000\times$). Inevitably, the `mpileup` program crashes because of excessive memory use.

Genome Analysis Toolkit (GATK)

The GATK suite has been developed and is continuously maintained by the Broad Institute and represents a leading software suite for NGS read alignment and analysis (der Auwera et al., 2013). In the recommended best practice protocol, the reads are aligned with BWA (Burrow-Wheeler Aligner), the SAM files are cleaned with the Picard software suite, followed by InDel realignment and base quality score recalibration. The files produced in this process are then used for 'calling' variants, that is using sophisticated statistical models to distinguish real variants from artefacts generated by sequencing errors.

While this protocol is very powerful when working with -omic datasets, in the case of the amplicon dataset the strength of GATK turned out to be a weakness. From the point of view of genomic analysis, an amplicon is a short sequence sequenced at extremely high coverage. In genomic data, a mammalian tissues sample will encode one (if homozygous), two (if heterozygous at the locus in question) or a handful (tumour or other abnormal sample) of genetic variants. Furthermore, each variant is present at a substantial level, typically >30%, and a rare variant would be detected when it is present at level above 1 in 100, or perhaps 1 in 1,000. In contrast, the amplicon dataset contains tens of thousands of variants, and the sequence alteration encoded by each variant is small (<10 nucleotides). Therefore, the overwhelming majority of bases match the reference sequence and each variant is very rare; 1 in 10,000 bases or rarer.

When I tried to use the GATK workflow with the amplicon dataset, the data processing required some additional set-up because of the different application, and was not sensitive enough to find any variants in the sequences. Changing the sensitivity setting in the variant caller (UnifiedGenotyper) did not improve the outcome. While the sensitivity could perhaps be improved by systematic optimisation of the complex settings, I abandoned this approach because it would inherently be arbitrary and opaque (it remains unclear which variants are actually called and why).

DMS Tools

The software `dms_tools` was first written and is maintained by the Bloom Lab for the purpose of analysing the results of DMS work with substitution libraries (Bloom, 2015). The `dms_tools` scripts are available at https://jbloomlab.github.io/dms_tools/, which also sets out common use cases.

There are two sections of `dms_tools` that were of particular interest:

- `dms_barcodedsubamplicons`. These scripts implement barcode-guided read sub-assembly to improve sequencing quality - a barcode (UMI) specifies individual DNA molecules from the library of interest, and each molecule is sequenced multiple times and the sequences 'collapsed' into a single consensus read. While I introduced barcodes in the GFP deletions libraries, the tools implemented in `dms barcodedsubamplicons` were incompatible.

  The first issue was that the sub-amplicon package makes the fundamental assumptions that all reads are the same length, and span the same region of the gene sequence. This assumption is obviously violated by InDel mutations. Second, the barcodes for sub-amplicon sequencing are added in a series of PCR steps. While the use of PCR makes the procedure relatively general, in the case of InDels it was undesirable because of the potential for shuffling different variants in the library together and creating 'chimera' variants.

- `dms_interprefs` scripts, which implement the calculation of the preference of every protein position for different amino acid residues. While the description of the statistics for DMS analysis was useful, the code itself was not compatible because of inherent assumptions about the lengths of sequence; any InDels are explicitly filtered out.

Since the attempt at use described above, an updated version `dms_tools2` has been published. It builds on the original programs with additional statistical methods and adaptations for different sequencing strategies - but still, the incompatibility of `dms_tools` with InDels remains.

4.2.2 Development of InDelScanner scripts

This section described the work flow I used for analysis of Illumina amplicon sequencing of InDel libraries, which were dubbed "InDelScanner". The scripts were initially developed on the GFP deletions DMS dataset (same as used for testing of other software solutions above) and the complete analysis of that dataset is presented in Chapter 2. Here, I instead demonstrate the use of InDelScanner scripts on the dataset from deep sequencing of phosphotriesterase (PTE) InDel libraries. The scripts are available at https://github.com/fhlab/TRIAD.

The PTE sequencing libraries were prepared for six libraries (-3, -6, -9 bp deletions; +3, +6, +9 bp insertions; no substitution library in this project). The plasmid minipreps containing each library were digested at endonuclease restriction sites located approximately 150 bp outside the ends of the gene, then processed using the Nextera library preparation protocol according to manufacturer's instructions (outlined in Section 2.2.3). Each of the six libraries was indexed separately, so each detected variant could be traced back to the source library. While this does not make a substantial difference for the analysis of desired variants, multiplexing each library separately enables better analysis of undesired variants for library quality control. The more diverse insertion libraries were sequenced with 3× higher loading on the flow cell compared to deletion libraries. This loading guaranteed full coverage for deletion libraries while generating more data for the more diverse insertion libraries.

The scripts for InDelScanner are primarily located in the files `indels/composition.py` and `indels/ind.py`, and accessed through shell scripts that automate the processing of multiple libraries. The three stages in the analysis of InDel libraries in a deep mutational scanning experiment are outlines in Figure 4.1 and described in detail below.

Read pre-processing and alignment

Input FASTQ files. Across all six libraries combined, the starting dataset contained 13 million pairs of forward and reverse reads, which corresponds to 26 million individual paired-end reads. According to Illumina specifications, a MiSeq flow cell sequenced with a v3 reagent kit (to which 2 × 75 bp kit belongs) should produce 44-50 million paired-end reads. Thus, the sequencing run produced approximately 55% of the total capacity of the flow cell. Part of the reduced output is caused by the use of 20% spike-in of PhiX genomic DNA, which increases the diversity of DNA loaded onto the flow cell at the cost of capacity. The Illumina detection of optical clusters on the flow cell is greatly hindered if all DNA clusters have the same sequence, so PhiX spike-in is recommended for low-diversity sequencing library including all amplicon sequencing. The remaining 25% must represent missing capacity of reads that did not pass quality filtering or a flow cell that wasn't loaded to full capacity - a reduced loading can increase sequencing quality.

Read assembly and filtering. In the InDelScanner workflow, the raw FASTQ reads are first assembled and prepared for further processing. The paired-end reads were assembled into a single longer contiguous read using the program PEAR (Zhang et al., 2014). Having

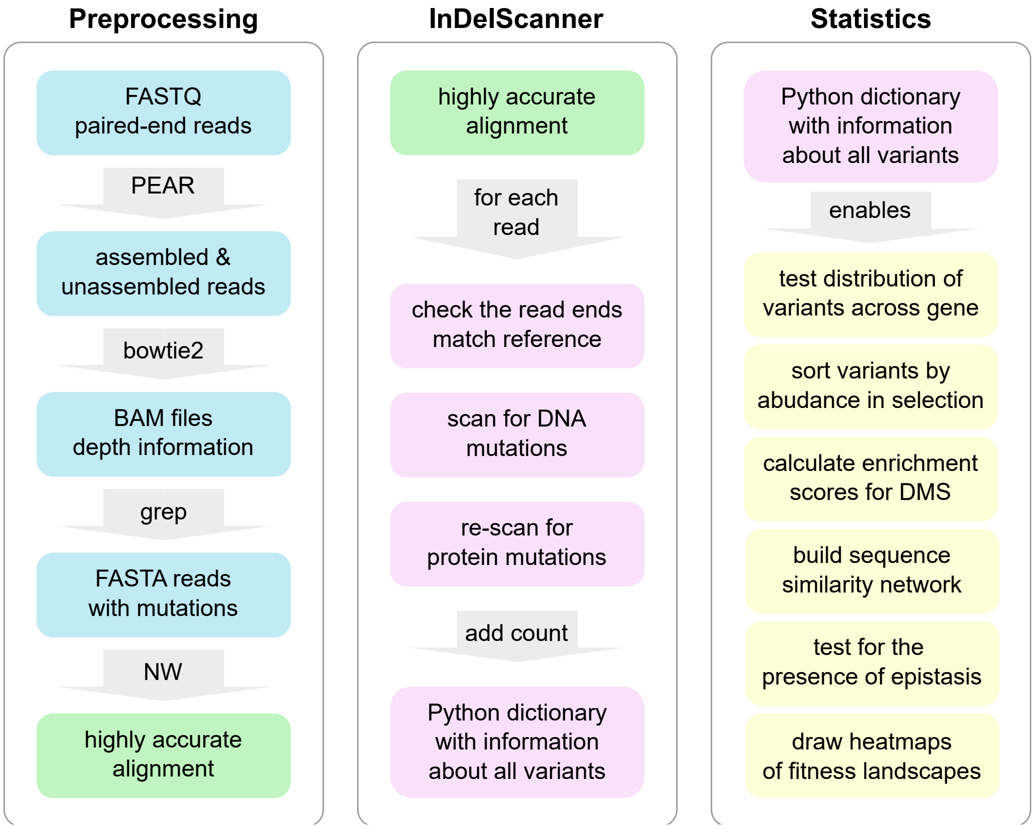

Preprocessing
FASTQ paired-end reads
PEAR
assembled & unassembled reads
bowtie2
BAM files depth information
grep
FASTA reads with mutations
NW
highly accurate alignment
InDelScanner
highly accurate alignment
for each read
check the read ends match reference
scan for DNA mutations
re-scan for protein mutations
add count
Python dictionary with information about all variants
Statistics
Python dictionary with information about all variants
enables
test distribution of variants across gene
sort variants by abudance in selection
calculate enrichment scores for DMS
build sequence similarity network
test for the presence of epistasis
draw heatmaps of fitness landscapes

Figure 4.1 A flow diagram of the InDelScanner scripts. The computational analysis starts with raw Illumina paired-end reads, which are trimmed, filtered, assembled and aligned to the reference sequence. This step generates a high-quality alignment file containing all reads of interest (left box). The second stage, variant calling, is done with the core IDS scripts and results in a dictionary of variants and their counts for all sequencing libraries (middle box). The pre-processing and variant calling are moderately computationally intensive and can be done on a personal machine or on a computational clusters. Finally, the dictionary of annotated variants is the source dataset for a variety of modular statistical analyses (right box).

inspected the sequencing quality and monitoring the assembly statistics with different options, I decided that the parameters for successful read assembly would be:

- an overlap of 5 or more bases between forward and reverse reads; this is decreased from the default value (10) because the 75-base pair reads are relatively short.
- no minimum length of assembled sequences; a permissive parameter that will assemble reads even if they are heavily trimmed. Changing this parameter did not have a noticeable impact on the percentage of assembled reads.
- quality threshold 15; if the quality scores of two consecutive bases are strictly less than the specified threshold, the rest of the read will be trimmed (not enabled by default). This option trims the reads before assembly if the quality is degraded and is part of good practice when handling NGS data.

Some reads do not assemble, either because they are too short, contain poor quality sequence or because they are not long enough to overlap each other. The latter reason is the most common and has a biochemical basis: the Nextera libraries are prepared with tagmentation, which requires the Nextera transposase to insert twice into the DNA. Due to crowding, there will be a minimal distance for two transposase insertions and therefore a minimum length of the resulting DNA fragment. Reads that sequence a DNA fragment longer than 150 bp cannot generate overlapping sequence, and therefore remain unassembled.

Alignment. Both assembled and unassembled reads were aligned to the gene reference, which was the entire length of the DNA fragment used for sequencing library preparation (gene + surrounding sequence). Alignment was done with Bowtie2 and the resulting SAM files used to obtain sequencing depth at each position. The assembly and alignment statistics for the PTE unsorted InDels library dataset are shown in Table 4.1. Of the total reads, 72%-80% reads were assembled, and both assembled and unassembled reads aligned to the reference sequence in a high proportion of cases (typically around 95%).

To calculate the sequencing coverage per base (i.e. sequencing depth), the SAM files generated by Bowtie2 were sorted with Samtools (Li et al., 2009) and the coverage was generated separately for assembled and unassembled reads with `samtools depth`. The total coverage per base in the sequenced fragment is shown in Figure 4.2A, and the aggregated coverage in the *wt*PTE gene is shown in Figure 4.2B.

The sequencing coverage is fairly even, averaging around 3×10^5 reads per base in the insertion libraries and around 1×10^5 reads per base in the three deletions libraries. While no region has zero or very low coverage, there is some systemic variation in depth and

Library	**-3 bp**	**-6 bp**	**-9 bp**
Total reads	1.04×10^6	1.09×10^6	8.99×10^5
Assembled reads	7.5×10^5	8.22×10^5	6.48×10^5
Assembled alignment rate	96.4%	95.1%	90.3%
Unassembled reads	2.86×10^5	2.67×10^5	2.51×10^5
Unassembled alignment rate	93.5%	95.9%	88.6%
% Assembled reads	72.4%	75.5%	72.1%
Total aligned reads	9.9×10^5	1.04×10^6	8.07×10^5
Mean coverage per base	8.59×10^4	8.84×10^4	6.99×10^4
Standard deviation	1.73×10^4	1.72×10^4	1.59×10^4
Library	**+3 bp**	**+6 bp**	**+9 bp**
Total reads	3.36×10^6	3.38×10^6	3.09×10^6
Assembled reads	2.56×10^6	$\times 10^6$	2.47×10^6
Assembled alignment rate	97.1%	95.0%	95.8%
Unassembled reads	8.08×10^5	8.67×10^5	6.28×10^5
Unassembled alignment rate	96.6%	94.0%	95.6%
% Assembled reads	76.0%	74.3%	79.7%
Total aligned reads	3.26×10^6	3.2×10^6	2.96×10^6
Mean coverage per base	2.76×10^5	2.73×10^5	2.38×10^5
Standard deviation	5.37×10^4	5.76×10^4	4.48×10^4

Table 4.1 The Illumina sequencing statistics for *wt*PTE InDel libraries. The insertion libraries were loaded onto the flow cell at 3× the amount of deletion libraries, which is reflected in the sequencing depth. The mean coverage per base was calculated across the PTE gene, discarding the flanking gene sequence, using the output of depth calculated by Samtools.

decreased coverage around position 250 in the *wt*PTE gene. This is probably due to sequence preference of the Nextera transposase, since the same pattern is observed in all six libraries.

Read filtering and re-alignment. Next, the SAM file is used to extract reads that a) are mapped well and b) contain a mutation, defined as any difference from wild-type sequence. These 'interesting' reads were then re-aligned to the reference sequencing with the Needleman-Wunsch algorithm in the EMBOSS implementation (Rice et al., 2000). Although using the alignment in the SAM file directly is faster, I tool a longer processing time at this stage as an acceptable trade-off so that I could be sure that the InDels are placed and identified correctly, which then enables the creation of accurate statistics of the library composition.

The first reason for this step is the desire to very accurately align the reads and correctly place the InDels. The Bowtie2 algorithms already align the gaps well, but the program does not allow exact control of alignment scoring, and there was uncertainty about the quality of the alignment given that genomic pipelines recommend additional fine-tuning of InDel alignment with more accurate tools in a separate step.

The Needleman-Wunsch is a widely used global, deterministic algorithm for very accurate sequence alignment (Needleman and Wunsch, 1970; Polyanovsky et al., 2011). It uses a scoring matrix to assess possible alignments between two sequences, such that a more positive score is better. In the case of DNA alignment, the scoring matrix 'DNAfull' is symmetric and equally penalizes any nucleotide substitution: each match is scored positively (+5) and every mismatch is penalized (-4). The scoring of gaps (i.e. InDels) is controlled by two parameters, the opening gap penalty (a fixed score that penalizes the creation of any InDel) and the extension gap penalty, which is proportional to InDel length. For a gap of length n, the total score is $gap\ opening\ penalty + (n-1) \times gap\ extension\ penalty$.

Depending on the balance between these parameters, the algorithm will choose between creating a long InDel with a mismatch at the end, or multiple short InDels with a small number of matched bases in-between. Typically, the gap extension penalty is set 5-10 times lower than the opening penalty. The default values for EMBOSS `needleall` program are gap opening penalty 10 and extension penalty 0.5. I observed that this was not appropriate for placing longer deletions (-9 and to an extent, -6 bp) in some sequence contexts, which were instead broken into multiple mutations. Therefore, I changes the scoring to a higher gap opening penalty (15) and manually verified the alignments .

Second, I decided to extract only non-wild-type reads from the SAM files and realign them with `needle-all`. Since this step accepts any difference from reference, it contains all reads with InDels, true substitutions or other incidental mutations in the libraries, as well

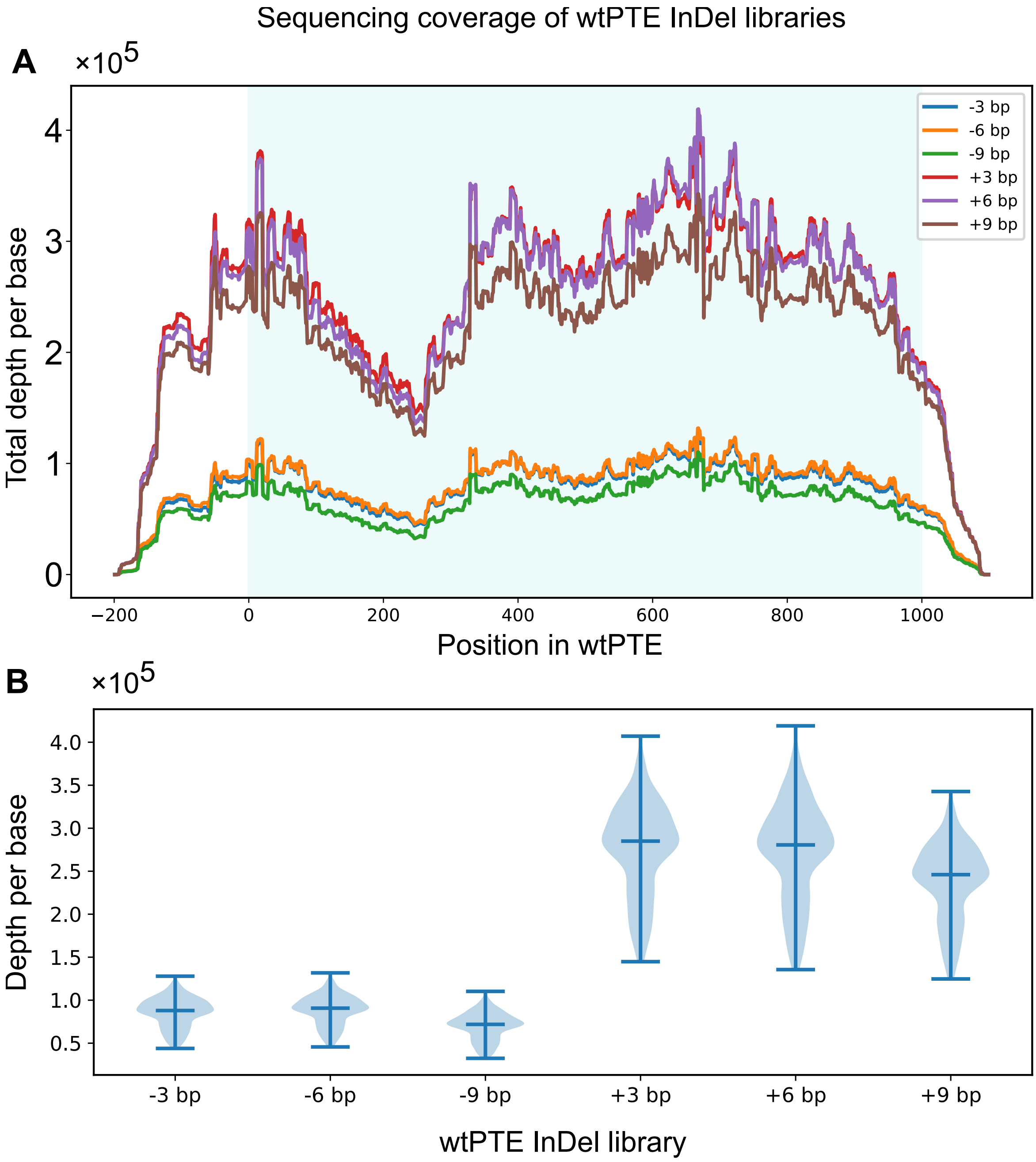


Figure 4.2 The combined sequencing coverage of *wt*PTE InDel libraries, calculated on the basis of all aligned reads (see Table 4.1). A) The coverage across the full DNA fragment submitted for Illumina sequencing, with the *wt*PTE gene highlighted in cyan. There is systematic variation in coverage across the gene, likely originating from Nextera transposase sequence preference. B) Violin plots of sequencing depth per base for the *wt*PTE gene region. The median depth and interquartile range are depicted with horizontal lines sequence depth distribution shown with the shaded area.

as reads containing single point substitutions from sequencing errors. Therefore, it is clear that the number of substitutions in the final statistics will be over-represented, but the rate of substitutions can later be adjusted by taking into account the (known) sequencing depth.

The choice to discard wild-type reads from the analysis was made for performance reasons. Although the read extraction is an additional step in the pipeline, the combination of specific read extraction and re-alignment it is still an order of magnitude faster than re-aligning all reads. However, the starting Bowtie2 alignment cannot be skipped, since it provides the depth information and quality filtering.

Pre-processing summary. This part of read processing encompasses all aspects of quality control (filtering, read adapter trimming, read assembly and initial mapping), followed by accurate alignment of reads to reference (left box in Figure 4.1). Special care is taken to correctly place InDels, which is made easier to verify by *a priori* knowledge of InDel lengths present. While I chose to use the Needleman-Wunsch (NW) algorithm in the EMBOSS implementation here, other options would be the Smith-Watermann algorithm and/or other implementations, for example the `scikit-bio` package . The EMBOSS implementation has the advantage of reliability, but it does not allow multi-threading. While this is not ideal, the time needed to process a dataset is relatively short (hours to days, depending on platform) compared by the time needed to collect the data and later data analysis (weeks to months).

Calling mutations

After the first stage prepared the data, the second stage parses the information contained in the alignments and aggregates it into a collection of annotated counts. The results is an organised data structure that links the positions of mutations, the number of occurrences, identifies the reading frame of each variant, links synonymous mutations and annotates them on the type of mutation (deletion, insertion, etc.). The two stages are done by separate scripts: pre-processing is organised with shell scripts that call external tools, while the mutation calling is done by the scripts `composition.py` or its variants. The analysis is done by scanning each sequence pair (aligned read and reference) with the BioPython tools.

End matching. First, the read is checked to ensure that both ends (last 3 nucleotides) match the reference sequence, so any mutations are located *within* the read. Reads with mismatches at the ends are discarded, because reads with mismatched ends cannot be accurately interpreted.

Consider an example: a sequencing read that perfectly matches the reference, then contains a -3 bp deletion, followed by a single matching nucleotide. The NW algorithm cannot possibly place this mutation, because that one matching nucleotide could come after a 3 nucleotide gap, or perhaps after a longer gap, or it could be the start of a substitution mutation. By default, the algorithm places it as a substitution, but that call is unreliable. A calculation of possible scoring outcomes determined that even -9 bp deletions and most insertions will be placed correctly as long as they are followed by 3 or more nucleotides. Therefore the alignment quality checkpoint requires that the first and last 3 nt in the read and reference match perfectly. Note that this length of matched and the scoring parameters were optimised for InDels up to 9 nucleotide long in TRIAD libraries and should be re-assessed for other library designs.

If the reads extend beyond the reference sequence, they are trimmed to the end of the reference sequence. This is non-essential for this analysis, but is useful for calling InDel mutations in results of Sanger sequencing.

Recording DNA mutations. First, the read pair is scanned and all mutations (regardless how complex) are recorded, nucleotide by nucleotide. The result is a list of mutations in the (approximate) standard format recommended by the Human Genome Variation Society (HGVS; den Dunnen et al. (2016)). The sequences are not altered at this point. The mutation list deviates from the standard in one noticeable way: the HGVS standard recommends that substitution mutations longer than one nucleotide are reported as ’deletion-insertion’ variants, while InDelScanner records them as a series of single nucleotide substitutions. Since the focus of these scripts is the analysis of InDel variants in amplicons, and the content of substitutions is still accurately reported, I decided that the effort to correctly comply with the standard was not worth the minor improvement in output.

Next, the pair of reads is scanned again and the DNA mutations are labelled in a custom format that is more useful for aggregating information across the dataset. Here, the InDels are described in detail if they preserve the reading frame (length a multiple of 3); if they are frame-shifting, that is noted and no further classification is performed.

Here, there is some duplication of processing time, which could be improved in the future for better performance. The processing is duplicated because the custom description code was implemented first, then used as a comparison when the HGVS description was added later. I implemented the two classifications separately to prevent the code from breaking or introducing errors into the existing code.

InDel redundancy and issues with placing InDels. In contrast with substitutions, which can be uniquely placed in most sequence contexts, InDels present some additional complications. Depending on the surrounding sequence, it is possible that distinct transposition events (i.e. the location of the InDel) and random sequence insertion can result in identical final sequence. Examples of two -3 bp deletions and four +3 bp insertions that create identical new sequence are illustrated in Figure 4.3.

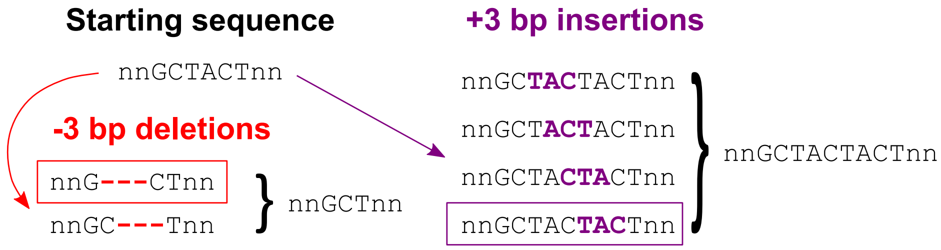


Figure 4.3 An illustration of two examples of InDel redundancy, which depend on the sequence context and in the case of insertions, the particular inserted random nucleotides. The sequence that influences redundancy is shown in capital letters and surrounding sequence is abbreviated with 'n'.

In the case of deletions, the NW algorithm arbitrarily and consistently assigns the deletion to the first position in the sequence where it can occur. In the example shown in Figure 4.3, two different deletions are shown, but both sequences will be assigned to the one highlighted in the red box. For insertions of NNN nucleotides, the range of types and locations that coalesce into one detected InDel is more varied.

Such InDel redundancy in the analysed sequences has some consequences:

- The diversity of mutations that can be observed is reduced compared to the maximal theoretical library diversity. In the example shown in Figure 4.3, the TRIAD libraries contain two deletions and four insertions, but only one of each is directly observed. In repetitive sequence, the number of 'hidden' InDels can be higher.
- The distribution of observed mutations across the gene cannot be directly used to infer the Mu transposon sequence preference, and the distribution must be adjusted first.

Recording protein mutations. In order to identify protein mutations, the read and reference sequence pair is parsed codon by codon. For this to work, the reading frame is first identified from the indexing of the reference read, which is controlled by the script's command line options (the first nucleotides in the reference need not be ATG, instead the reference reading framecan be padded with extra sequence).

When a gap is detected in the read, the codon is analysed as a deletion. If the codon is a complete deletion (i.e. '---' in the alignment), it is recorded as an amino acid deletion and the logic moves on to the next codon. If the length of the deletion is not a multiple of 3, the deletion is recorded as a frame-shift and the parsing ends for this sequence pair. Else, if the codon contains both nucleotides and a gap (i.e. for cross-codon deletions which align as 'N–' or 'NN-'), the nucleotides at the end of the gap in the alignment are shifted to the front until the first codon is complete, and that new codon translated to record a substitution. Then, the next codons will be pure deletions and the analysis of the read continues. Consequently, all cross-codon deletions are reported with the adjacent substitution (if there is one) reported *before* the deletion.

Insertions are treated in a similar way; the gap in the reference sequence is moved to align with the start of the next codon, so the substitutions are reported before the insertion, and if the insertion length is not a multiple of 3, a frame-shift is recorded. The inserted codons are translated into protein sequence and recorded with the index of the preceding residue and alphabetical labels (e.g. 23aV, 23bH, etc.). If the inserted sequence contains a stop codon, the variant will be a frame-shift and the analysis is terminated at that point.

All codons that contain substitutions are translated into protein sequence and recorded, either in the internal format for statistics or in the approximate HGVS format described earlier.

Summary. At the end of this process, all aligned pairs of sequences (read and reference) have been scanned and parsed into a dictionary, which aggregates information about the DNA changes, the protein mutations and counts (middle box in Figure 4.1). The information is structured with DNA variants nested under relevant protein variants, since multiple DNA variants can result in the same protein mutation. The variant information is stored both in internal format with a functional description (substitution/insertion/deletion/frame-shift, used to generate statistics) and approximately in line with the Human Genome Variation Standard.

4.2.3 The composition of TRIAD libraries in PTE

Here I describe the work on characterising the TRIAD libraries constructed in the phosphotriesterase (*wt*PTE) gene, which was included in the evaluation of the TRIAD method (Emond et al., 2020). The libraries had already been analysed with Sanger sequencing of randomly chosen variants, which showed that the libraries contained single mutations only (no incidental mutations) and the proportion of non-targeted variants (primarily frameshifting InDels) was higher in the insertion than in the deletion libraries. I decided to obtain this deep sequencing dataset to gain detailed insight into all variants, rare and common, and to better quantify library diversity and quality.

A theoretical 'ideal' dataset. Thanks to sequence dependence of InDel redundancy, the theoretical number of directly detectable variants is lower than the maximal theoretical diversity of libraries; the latter would be 1 deletion per bp gene length and 64^n (where n = 1, 2 or 3) variants per bp gene length for insertion libraries with one, two or three randomized NNN triplets. Because of sequence dependence of redundancy, calculating the detectable diversity of libraries analytically is not reasonably possible. Instead, I generated an 'ideal' dataset that simulates one read for every possible variant that would occur in a perfect library. These reads were then aligned with NW algorithm and the mutations called in the exact same way as the experimental dataset. Because of the immense maximal theoretical diversity of the +9 bp library ($> 10^8$ DNA variants), the accessible theoretical diversity of this library was estimated from the calculated diversity in the +6 bp library.

This comparison shows the theoretical diversity in *wt*PTE libraries as observed in DNA diversity, is ~0.75 deletion per bp gene length, 46 +3 bp insertions per bp (compared to 64 in a NNN codon), and 2.8×10^3 +6 bp insertions per bp gene length. Overall, the accessible diversity is 70-75% of maximal theoretical diversity across all TRIAD libraries (Table 4.2).

Diversity of detected variants

Here, the counts in the *wt*PTE libraries were aggregated and classified according to the type of mutation and compared with the theoretically accessible diversity calculated above (Table 4.2). In all libraries, the targeted in-frame InDels were found in high abundance, reaching more than 10^5 variants detected by deep sequencing in the most diverse +6 bp and +9 bp libraries ($> 10^3$ unique deletions total and $> 10^5$ unique insertions per library). In the following discussion, the numbers of variants are discussed without correcting for InDel redundancy unless otherwise specified.

Diversity of observed mutations: deletions. Thanks to the relatively small diversity of the deletion libraries, they were sequenced with high coverage, and so ~90% of possible (=accessible) DNA variants were observed (Figure 4.4A). The distribution of variant frequency is also relatively even: most deletions show similar frequencies, with 52% of all detected deletions having between 10 and 99 reads per variant, and only 11% of all deletions occurring more frequently (defined as 200 supporting reads or more per variant) across all

	Deletions		
TRIAD library	**-3 bp**	**-6 bp**	**-9 bp**
Observed unique in-frame DNA InDels	639	690	613
Accessible DNA diversity	748	747	730
Observed DNA diversity (% accessible)	85%	92%	84%
Observed unique in-frame protein InDels	530	562	492
Accessible protein diversity	590	589	551
Observed protein diversity (% accessible)	90%	95%	89%
InDels with no adjacent aa substitution	302 (57%)	320 (58%)	307 (63%)
InDels with adjacent aa substitution	223 (42%)	234 (42%)	180 (37%)
InDels resulting in truncated variants	5	8	5
Frameshifts	4.1%	20%	14%
	Insertions		
TRIAD library	**+3 bp**	**+6 bp**	**+9 bp**
Observed unique in-frame DNA InDels	20872	107165	103720
Accessible DNA diversity	4.6×10^4	2.8×10^6	1.8×10^8
Observed DNA diversity (% accessible)	45%	4%	$< 0.1\%$
Observed unique in-frame protein InDels	8400	58559	94303
Accessible protein diversity	13004	268159	5.63E+06
Observed protein diversity (% accessible)	65%	24%	2% [a]
InDels with no adjacent aa substitution	4671 (58%)	34008 (58%)	56086 (59%)
InDels with adjacent aa substitution	3359 (42%)	19561 (37%)	26691 (28%)
InDels resulting in truncated variants	370	4990	11526
Frameshifts	37%	29%	26%

Table 4.2 The observed diversity of detected variants in PTE InDel libraries, compared to the theoretical diversity of complete TRIAD InDel libraries. [a] The theoretical protein diversity of +9 bp library is estimated as 21× larger (20 amino acids and a stop codon) than the calculated diversity of +6 bp library. This estimate is in line with the ~21× difference in diversity between the +6 and +3 bp libraries.

three libraries combined. The distribution of the variants as a function of the number of supporting reads is shown in Figure 4.6.

It should be noted that the theoretical diversity per position depends on the sequence context: in PTE, the theoretical DNA diversity per position in the +3 bp library is between 42 and 48 variants and the mean number of observed single triplet insertions is 20.8 per position. The mean number of variants per position in the +6 and +9 bp insertion libraries is larger (107 and 103, respectively, Figure 4.4C,D), but the observed fraction of theoretical diversity is lower because of much higher theoretical diversity when more NNN triplets are inserted.

Mu transposon preferred sequence. I also used the distribution of -3 bp variants across the gene to re-visit previous estimates of Mu transposon insertion site preference. Previous analysis of the mini-Mu transposon target site preference was based on 806 observed transpositions and suggests a strong preference for pyrimidines in position 2 and purines in position 4 of the 5 bp transposition site (Haapa-Paananen et al., 2002). The -3 bp library is particularly informative, because the -3 bp deletion is located in the middle of the 5 bp Mu-transposon recognition sequence. Analysis of the -3 bp library shows that 85% of possible transposition positions are accessed by TransDel (Table 4.2).

To identify the transposon sequence preference, the distribution of -3 bp mutations across the gene was corrected for InDel redundancy by evenly distributing the counts of redundant InDels to all possible originating position (Figure 4.5A). Weighing the sequence context of each -3 bp deletion by the number of times each deletion was observed gave the weakly preferred transposition sequence to be 5'N-Py-G/C-Pu-N (see WebLogo insert in 4.5A, (Crooks, 2004)). This sequence bias of Mu transposons is less pronounced than previously thought (Haapa-Paananen et al., 2002) and does not clearly correlate with GC content (Figure 4.5B).

Diversity of observed mutations: insertions. The TransIns transposon accessed 95% of all possible insertion position (calculated for the +3 bp library), and this high coverage of gene positions translates into high library diversity at most positions in *wt*PTE (Figure 4.4B-D). The aggregated data shows that $\geq$10 distinct DNA insertions were observed at between 66% (+3 bp) and 80% (+6 and +9 bp libraries) of positions; furthermore, $\geq$100 insertions were detected in 34% (+6 bp) and 31% (+9 bp) of positions.

Precise quantification of diversity in +6 and +9 bp libraries is hindered by limited sequencing coverage, which was insufficient despite higher loading of the insertion libraries onto the flow cell. The constraints of the low coverage are evident in the observation that in

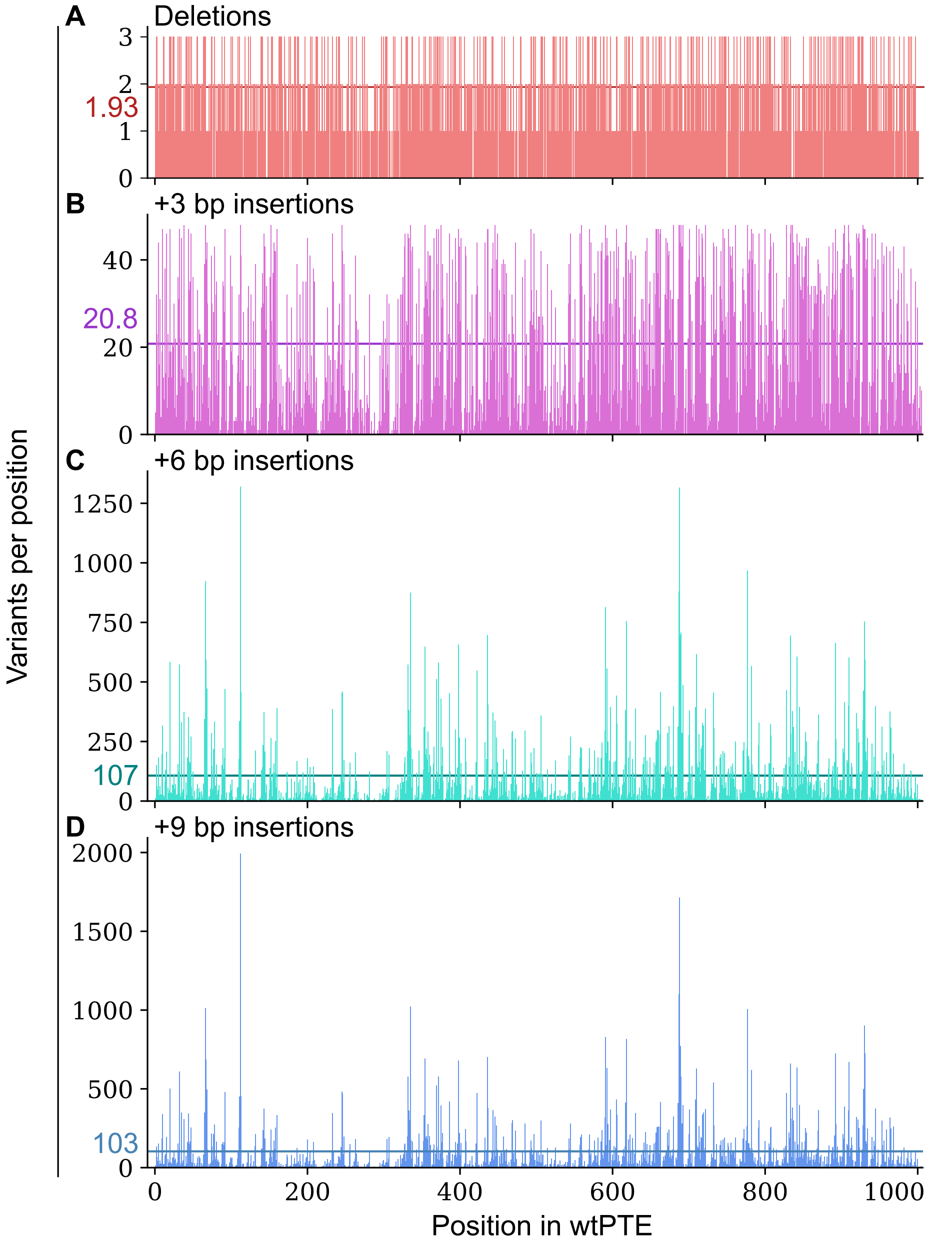


Figure 4.4 Diversity of observed InDels across the *wt*PTE gene. The horizontal line in each chart shows the mean number of detected variants per position in that library. A) The distribution and number of detected distinct DNA deletions in -3, -6 and -9 bp libraries combined per *wt*PTE position. Since each position either contains a deletion of a given length or not, the different lengths of deletions are combined in one panel. B-D) The number of observed +3 / +6 /+9 bp variants per DNA position in the corresponding library. The possible diversity of +6 bp and +9 bp insertions is much higher than for +3 bp library, which results in more pronounced high diversity “spikes” at positions where transposon insertion is favoured. These results are not corrected for codon ambiguity, which increases the unevenness of the distribution.

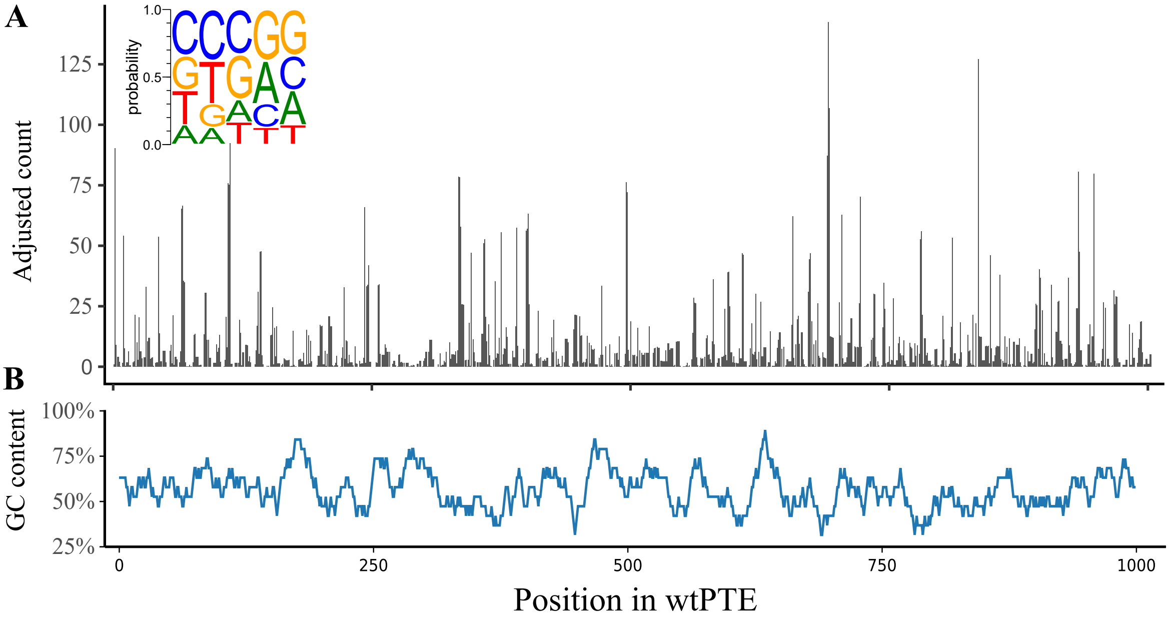


Figure 4.5 The sequence preference of the mini-Mu transposon. The composition of InDel libraries in the wtPTE gene was determined by deep sequencing and validated using Sanger sequences from randomly chosen variants. A) Relative frequency of TransDel transposon insertion across wtPTE, derived from -3 bp deletions observed in deep sequencing and normalized for sequencing depth and InDel redundancy in DNA sequence. WebLogo insert: The relative transposon insertion site preference was determined by extracting the five-nucleotide target sequence around each detected -3 bp deletion (in forward and reverse complement direction, since the direction of transposon insertion is unknown). The frequency of insertion at each position was used to weigh the contribution to consensus sequence, then normalized to give the proportion of each nucleotide per position in the Mu transposon consensus sequence. B) The GC-content in the *wt*PTE gene, calculated as the moving average in a 19 bp window.

these libraries, each variant was observed only once or twice. In a fully sequenced library, the distribution should average at higher number of observations, as was observed for deletion libraries (Figure 4.6).

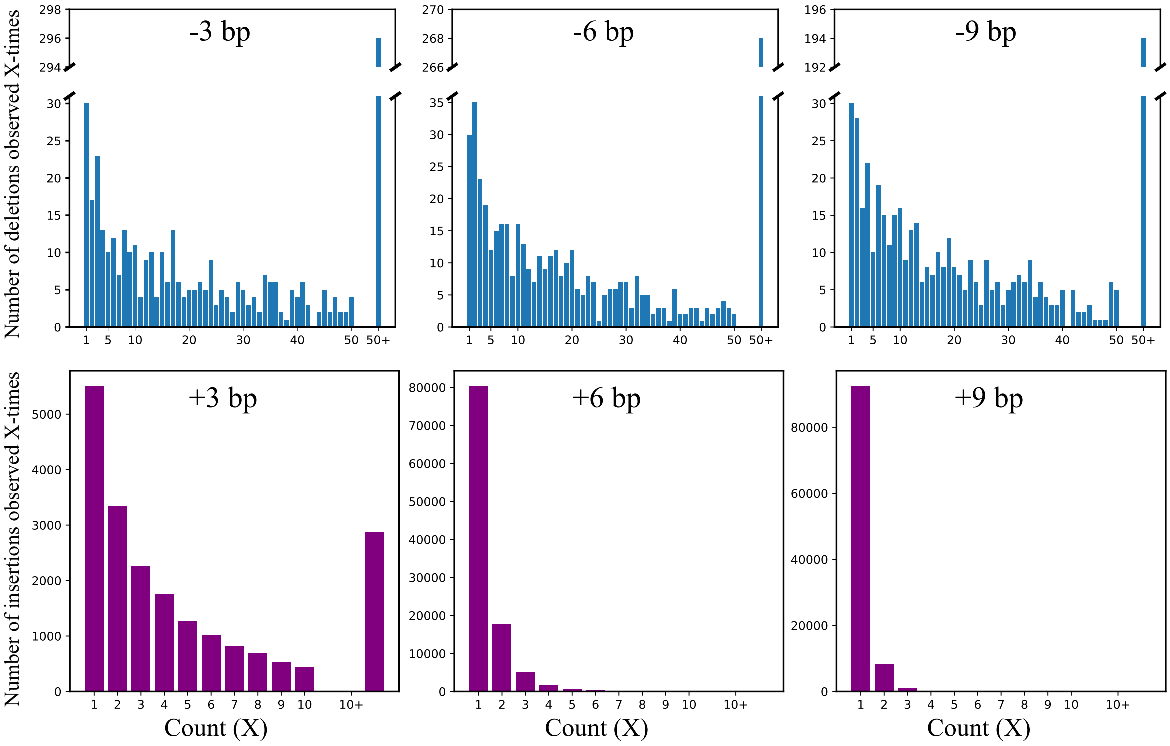


Figure 4.6 A histogram of the number of sequencing reads supporting each observed InDel. The histograms show how many mutations are observed once, twice, thrice, ten times or more. In deletion libraries most detected mutations are robustly supported by 10-40 observations (each observation is a single read in raw sequencing data). The bias of transposon site preference results in InDels being observed more often at some positions than other. In the deletion libraries, most mutations are supported by < 50 reads, but there is a long tail generated by positions that are close to the Mu transposon consensus – this is aggregated into one bin in the histograms in this Figure for clarity. Because of the large diversity of insertion libraries, variants are generally observed fewer times (x-axis) compared to the deletion libraries, which indicates under-sequencing of insertion libraries.

Substitutions and the insertion randomisation. When the transposition event occurs across two codons, the resulting InDels may exhibit an adjacent amino acid substitution: on protein level, an average of 39% of the InDels observed in the deep sequencing dataset of wtPTE variants exhibited such substitutions (Table 4.2). No significant bias was observed in the nucleotide composition of the in-frame insertions (Figure 4.7), indicating that TRIAD generates diverse insertion variants.

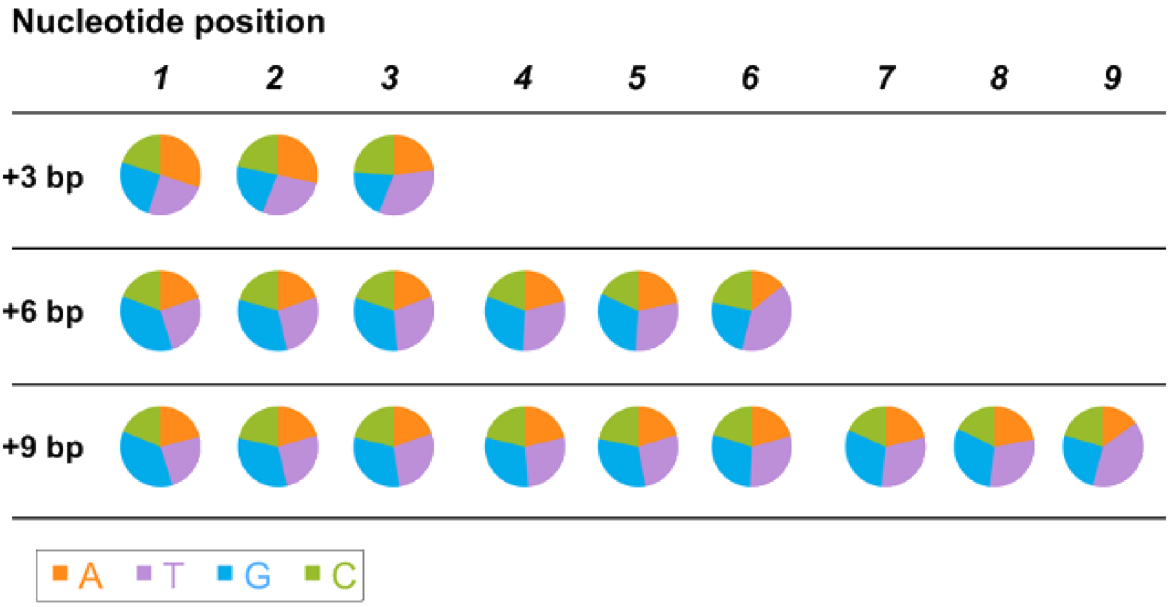


Figure 4.7 The distribution of random nucleotides in insertions by position in the inserted nucleotides, separated by length of the insertions. Each detected insertion contributes to the distribution equally, regardless of the frequency of the variant in the library.

The proportion of incidental substitutions and frame-shifts

As described above, the counts of different mutations that are generated during read pre-processing and parsing count *all* reads that are not wild-type, and this difference may be a genuine mutation, a PCR error or a sequencing error. The resulting counts are therefore artificially enriched for many variants with a single nucleotide substitution, which each appears only once or perhaps twice. This raises the question of whether these variants are genuine and infrequent - perhaps arising as polymerase errors during cloning steps in TRIAD library preparation - or if they are artefacts.

The true proportion of incidental mutations away from the InDels site can only be determined through long-read sequencing. The Sanger sequencing of >450 variants in total across six libraries did not detect such incidental mutations, which suggests that incidental mutations are very rare (Emond et al., 2020). To compare the results in the NGS dataset with this result, I estimated the true number of point substitutions in the library by calculating the background substitution frequency from reads that align outside wtPTE, in the plasmid backbone, where no mutations were deliberately introduced during library construction. These substitutions must therefore be sequencing artefacts. Such artefacts (with a single nucleotide substitution) occur in 3-4% of all reads, which corresponds to the error rate in the Illumina MiSeq NGS technology (0.25 to 0.40% error per base, Schirmer et al. (2016)).

An exact estimate is difficult due to a relatively low number of reads that align outside the fragment, as well as sequence dependence of polymerase errors – such that the error rate may be different inside and outside of the gene. Therefore, in the calculation of the proportion of frameshifting mutations in the libraries, such rare point mutations were simply counted as wild type (thus removing this 'noise').

Accounting for these considerations, the data shows that frameshifts were rare in the -3 bp deletion library (4%) and more frequent (>20%) in the others (Table 4.2) The -3 bp library requires fewer cloning steps, while the higher frameshift rate is likely because of accidental over-digestion by the Klenow fragment in the later stages.

4.3 Conclusions

Aims for the scripts. I started developing the InDelScanner scripts out of necessity, to answer a straightforward question: what variants are present in the TRIAD libraries? As is typical with bioinformatic work, the process of working with these scripts balanced accuracy (i.e. signal vs. noise) with computational performance (i.e. the time and processing power required to process a given NGS dataset).

While I developed these scripts specifically for analysis of TRIAD libraries, I chose a structure that is general enough to use for other amplicon sequencing projects. Early in tackling this project, I closely tailored the script structure to the expected TRIAD library design. While the TRIAD-specific design would make the data interpretation easier in some ways, the specific design quickly became too cumbersome for a clean implementation. Furthermore, the specific scripts design would require re-running the entire pipeline whenever a new question about the data appeared, which was unacceptable.

The InDelScanner scripts presented here are modular and can be used to work with Illumina DNA data in FASTQ format, Sanger sequences or NGS data converted into protein sequences. The data analysis structure does not make assumptions about the number or location of mutations in each variant, so it can be used for any length of read and reference sequence typical for amplicon sequence. The scripts record *all* variants observed in the dataset, and then the aggregated data is filtered in a non-destructive way to test hypotheses about the results.

As shown on the multiple datasets presented in this thesis, the scripts can be adapted to different experimental designs by altering individual modules, such as the alignment parameters.

Read pre-processing pipeline. The first step in the process is a straightforward sequence of event in a shell scripts, and versions are available both for processing on a PC and on a high-performance computational cluster. The pre-processing with the first hinge point of the work flow; the multiple-to-one sequence alignment. Each variant sequence is aligned to the reference sequence using the Needleman-Wunsch algorithm, chosen to provide the highest accuracy of the alignment.

There are multiple possible avenues for improving the performance of this pipeline. First, the Smith-Watermann (SW) algorithm for local alignment, using the same scoring matrix, may finish the alignment process faster, since it is optimised for local and not global alignment. The SW algorithm is available within the EMBOSS collection as the program `water`. However, before switching the output of the two algorithms should be compared on a collection of test sequences, to compare the output and computational time. Furthermore, alternative implementations of the alignment algorithms could be faster (e.g. SW implementation in `scikit-bio` package for Python).

Variant detection and composition analysis. The second hinge point of the process is the detection of mutations present in each aligned reference-read pair in the sequence alignment, which was generated in the first step. Each pair is scanned multiple times, first non-destructively to directly record DNA mutations and then on a copy that is modified in-place to describe mutations translated to protein sequence. The detection process tags each variant with useful information on the type of mutation (substitution, insertion, deletion) and links synonymous mutations under the same protein variants. The variants are aggregated into a dictionary of variant counts, which is then the basis for all further analysis and statistics.

The variant detection scripts were first set-up to work with DNA sequence, and report both DNA and protein variation, but can also work with protein sequence input (see Chapter 7. The scripts were primarily optimised for accuracy, so there is scope for improvement through grouping some repetitive functions and parallelisation of the file analysis. However, the sub-optimal speed of the scripts does not hinder the use of the scripts in practice, since the analysis of the results takes much longer than raw processing - the cause for this is that the data analysis is typically customised to the research questions in each project.

Outlook for data analysis. This chapter focuses on the use of the InDelScanner scripts to define the composition of single variant amplicon libraries, specifically with TRIAD libraries. Beyond the initial analysis of composition, a variety of statistical methods can be applied to further query the system. Some of those methods are presented in Chapter 2, while an example analysis of a multidimensional protein fitness landscape is shown in Chapter 7.

At one point, I attempted to use the advanced package for DMS statistical analysis Enrich2 (Rubin et al., 2017), since the package documentation describes a format which should allow the import of InDelScanner-generated variant descriptions into Enrich2. Unfortunately, it turned out that the import capability was limited, and the package author did not respond to requests for clarification - even though he originally recommended the use of his software to me.

In conclusion, the InDelScanner scripts have opened the path to exploration of deep sequencing experiments on libraries containing InDels. I have successfully used them to analyse multiple protein libraries, and thanks to the wealth of information in these datasets new questions are still appearing.

Chapter 5

The effect of small InDels on PTE stability

The question of the effect of InDels on PTE stability was raised during submission of our manuscripts on TRIAD libraries. This short chapter describes some experiments that probe the effect of InDels on thermostability of PTE and on soluble protein expression levels.

5.1 Introduction

During previous work in the research group on screening the effect of InDels on PTE (Emond et al., 2020), the TRIAD libraries were assayed in cell lysate in 96-well plate screening format. It was observed that InDels are on average more detrimental to the fitness of wtPTE by one order of magnitude in comparison to point substitutions. However, the host *E. coli* cells over-expressed both PTE variants of interest and GroEL/ES chaperones, which could mask additional disruptive effects of InDels on PTE structure. Consequently, I wished to at least partially disentangle the causes that cause the InDel disruptiveness to PTE cell lysate catalytic activity.

Generally, enzyme fitness is reflective of both enzyme catalytic activity and the concentration of soluble and functional enzyme, which itself relate to protein stability (DePristo et al., 2005). Before describing the experimental design, thermal and kinetic stability need to be defined, which are unfortunately often used interchangeably in cursory discussion of protein stability. To complicate matters further, protein stability can also refer to resistance to proteolytic degradation, to robustness of catalytic activity over a range of pH values and salt concentrations, or to tolerance of co-solvents in the reaction solution.

Thermal stability. The thermal stability of proteins is conveyed through two measures, the thermodynamic stability (ΔG) and thermal resistance (T_m). The first, ΔG, is defined as the difference in free energy between the native and unfolded state *in vitro*, by definition measured under equilibrium conditions - that is, reversible folding and unfolding. The favourable ΔG is the thermodynamic driving force for proteins folding, whether in the cell or in *in vitro*. However, since thermodynamic stability is a *difference*, when considering the effect of mutations, a change in ΔG may be brought about by changes to the folded or to the unfolded state.

PTE is a thermostable enzyme with a high T_m (78°C), somewhat surprising since it was originally isolated from the freshwater microorganism *Brevidomonas diminuta* that shows optimal growth at 35°C (Leifson and Hugh, 1954). PTE unfolds through a three-state mechanism ($N_2 \rightleftharpoons I_2 \rightleftharpoons 2U$), with $\Delta G = 4.3$ kcal/mol for the equilibrium between the active native state and the inactive dimeric intermediate (Grimsley et al., 1997). Futhermore, purified PTE shows high catalytic activity from 25°C up to 60°C (Armstrong, 2007). Between the high thermal resistance of the enzyme and its cold-environment origins, it is unsurprising that changes in PTE thermostability are a poor reporter of enzyme availability in the cellular environment; a mutation would need to be highly destabilizing to destabilise the enzyme to the point of inactivity at experimental conditions (specifically, protein expression at 30°C and functional screening at room temperature).

Kinetic stability. Separate from thermostability (i.e. the equilibrium measure), the amount of soluble enzyme in solution also depends on the ability of the protein to fold correctly, while avoiding misfolding or aggregation even at high concentration. Since protein folding in a cells does not occur under equilibrium conditions, kinetic stability covers common issues that prevent high soluble protein expression. Working with substitutions in PTE, the enzyme solubility correlates well with chaperone dependency (defined as the ratio of enzymatic activity in crude lysate with to without GroEL/ES), such that variants with low chaperone dependency show better expression levels. On the other hand, it has been shown that there is no correlation between the T_m and soluble expression levels, so T_m is not particularly informative for assessing the stability of PTE variants (Wyganowski et al., 2013). Since the substitution variants of PTE were typically folding-impaired, if impaired at all, buffering the kinetic stability issues was the reason that main screening campaign for the TRIAD libraries was done in the presence of GroEL/ES chaperones, co-expressed from a second plasmid present in the cell.

The results of both native and promiscuous activity screening indicated that while InDels are on average more disruptive, the TRIAD libraries simultaneously contain an equal proportion or more adaptive variants and are well worth screening, as long as sufficient variants are examined. While this positive result was encouraging, it is possible that the favourable impression of InDels was biased by the kinetic stability boost provided by presence of GroEL/ES chaperones. Thus, InDels may unfavourably affect the structure and the identified improved variants are now strongly handicapped, such that they can only function in the presence of chaperones, and are effectively evolutionary dead ends. If that were the case, we hypothesised that the best hits would likely be most strongly affected and make InDels much less attractive as a mutagenesis method for directed evolution.

I therefore set out to further characterise the interaction between the effects of small InDels on kinetic stability and catalytic activity of PTE. The investigation comprised two parts. First, I performed an additional investigation of stability in the four top hits in the phosphotriesterase screening campaign, in order to determine their chaperone dependence. Second, I set out to perform a small exploratory screen of paraoxonase activity of randomly chosen PTE variants, without pre-screening and in the absence of chaperones.

5.2 Results

5.2.1 Characterization of stability in improved arylesterase hits

First, I examined the hypothesis that InDels in TRIAD libraries confer a strong stability handicap. I therefore examined both the kinetic and thermal stability of the four variants (Table 5.1) and observed that the InDel hits do not significantly deviate in either direction; in fact two hits have a higher and two hits a mildly lower T_m, and all T_m's of the InDel variants are within ±7°C.

Considering the kinetic stability of PTE variants, the four hits show a similar soluble expression profile as *wt*PTE. This is true both for the proportion of solubly expressed protein, as well as for the ratio in soluble expression with and without the presence of overexpressed GroEL/ES chaperones. Three of the four hits show moderate solubility when overexpressed (much like *wt*PTE), which improves by approximately 20% when chaperones are present. The variant ΔA270L271L272G273 shows improved solubility, which is also less dependent on chaperone presence.

Finally, since the expression levels of the PTE variants is affected by GroEL/ES presence, the measured catalytic activity in cell lysate may also be affected. Following the definitions in Wyganowski et al. (2013) an enzyme's chaperone dependence is defined as the cell lysate catalytic activity in the presence of chaperones, divided by activity in the absence of chaperones, with both values appropriately normalised to cell density in the expression culture. This parameter is considered because while the total amount of protein might not be strongly affected by chaperone presence, it is conceivable that the change in folding pathway accessibility could affect the amount of active, fully folded enzyme. As long as the inactive intermediate form I_2 is soluble, it remains indistinguishable from native dimeric PTE on a denaturing SDS-PAGE gel. Three of the four variants exhibit a low chaperone dependency (<2-fold change in rates) while ΔA270L271L272G273 does show an increase in chaperone dependence. Thus, the properties of these hit variants are in direct contradiction with the expectation that the most active hits will also be the most compromised.

Taken together, the variants with the best catalytic activity identified in the arylesterase screen do not appear to have any stability or solubility handicap.

5.2.2 Small systematic screen of stability effects on randomly chosen variants

As a preliminary investigation of the stability effect of small InDels compared to point substitutions, I examined a group of randomly chosen variants from three libraries with mutations of equal length: -3 bp deletions, +3 bp insertions (TRIAD libraries) and the ±3 bp substitutions (TriNEx library) in wt•PTE. All three libraries were introduced into *E. coli* cells that did not contain GroEL/ES chaperones, which removes their effect as a potential confounder. The variants were randomly chosen and sequenced to verify they were in-frame and without incidental mutations, which could confound the effect of the InDels we were interested in. Approximately 30 in-frame variants per library were used for functional assays. For each variant, I measured two properties: the cell lysate initial reaction rate against paraoxon, which was the measure of variant fitness also used in the manuscript, and protein soluble expression level relative to *wt*PTE.

In the activity screen for this experiment, the kinetic assay was conducted at multiple PTE lysate concentrations, so that strongly deleterious variants that still maintain some catalytic activity can be distinguished from inactivating variants (defined as >2000-fold decreased activity). The results are shown in Figure 5.1B. For context, the distribution of activity in the screen without chaperones is compared to the activity distribution in the main screening campaign with chaperones by Stephane Emond (Figure 5.1A). The distribution of relative paraoxonase activity across variants in the chaperone-free screen largely follows

Table 5.1 The kinetic, solubility and thermostability properties of *wt*PTE, the four best InDels hits arising from the arylesterase screening, and the arylesterase substitution hit H254R for comparison. b) The catalytic activity data is reproduced from Emond et al. (2020). [a] The properties of H254R and *wt*PTE chaperone dependency were determined by Wyganowski et al. (2013)

PTE variant	$T_m(°C)$	- GroEL/ES	+ GroEL/ES	Ratio
wtPTE	78.1±0.2	63%	77%	1.22
H254R[a]	88.0±0.1	47%	82%	1.74
ΔA270L271L272G273	82±1	90%	81%	0.90
P256R/G256aA256b	84.3±0.3	56%	59%	1.05
S256aG256b	77.5±0.4	59%	70%	1.19
G311a	75.2±0.3	64%	83%	1.30

(a) Solubility and thermostability

	Paraoxon	**2-NH**	
PTE variant	k_{cat}/K_M $M^{-1}s^{-1}$	k_{cat}/K_M $M^{-1}s^{-1}$	Chaperone dependency
wtPTE	$2.2x10^7$	$4.2x10^2$	2.2[a]
H254R[a]	$8.9x10^6$	$1.1x10^3$	1.9
ΔA270L271L272G273	$2.70x10^5$	$4.03x10^3$	4.1
P256R/G256aA256b	$1.45x10^5$	$6.04x10^3$	1.3
S256aG256b	$2.70x10^5$	$6.83x10^3$	0.9
G311a	$1.35x10^5$	$7.36x10^3$	1.1

(b) Catalytic activity

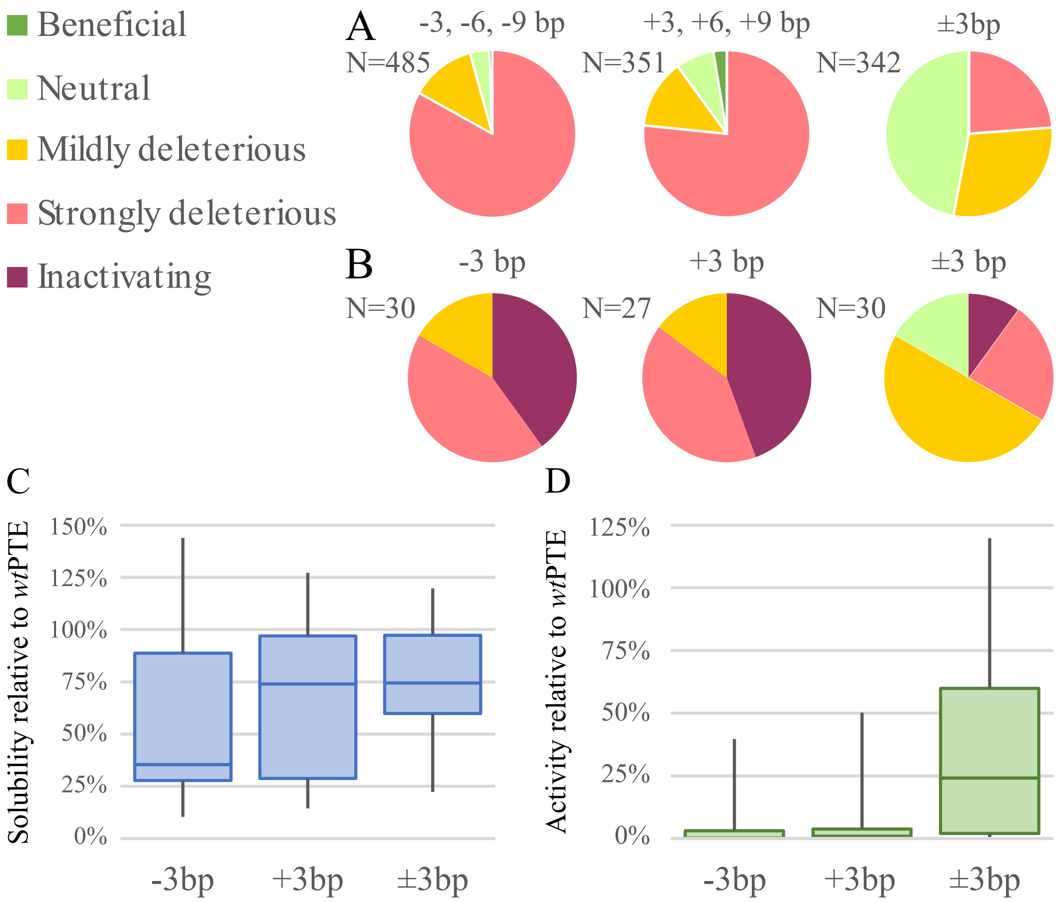


Figure 5.1 Distribution of relative paraoxonase activity in wtPTE variants. Beneficial variants show >1.5-fold improved activity compared to *wt*PTE, neutral variants show <1.5-fold change in either direction, mildly deleterious show between 1.5-fold and 10-fold decreased activity, strongly deleterious variants have between 10-fold and 2000-fold decreased activity and inactivating variants show no detectable paraoxonase activity. A) Activity distribution in the presence of GroEL/ES (reproduced from Emond et al. (2020), Figure 6). Inactivating variants are included with strongly deleterious variants. B) The effect of -3 bp deletions and +3 bp insertions compared to ±3 bp substitutions in the absence of chaperones. C) Change in solubility between randomly chosen variants in the absence of GroEL/ES. D) Aggregated change in activity, same data as panel B.

the distribution that was observed in the main screening campaign. For both deletions and insertions, the combined proportion of strongly deleterious and inactivating variants (which are grouped under 'strongly deleterious' in Figure 5.1A) are barely increased compared to when the variants were screened in the presence of chaperones. In the case of substitutions, the variants appear to be more noticeably impaired in catalytic activity by the absence of chaperones, as indicated by the predominance of mildly deleterious variants in this experiment compared to the screen with chaperones, where neutral variants predominated. Overall, the activity of InDels is an order of magnitude lower regardless of the presence of chaperones (Figure 5.1D).

I further investigated the kinetic stability of the same -3 bp, + 3 bp and ±3 bp variants by measuring their soluble expression levels compared to *wt*PTE (Figure 5.1C, see raw data in Figure 5.3). Here, the InDels are more deleterious to soluble expression and protein stability than substitutions: 17 out 30 deletions and 11 out of 27 insertions were found to be strongly destabilizing (<50% of soluble expression relative to *wt*PTE) while this was the case for only 6 out of 30 substitution variants. I also compared the average impact on soluble expression, measured as the geometric mean of solubility change, and observed that the mean solubility change in InDels was up to 1.5-fold lower than for substitutions. However, the median expression level of +3 bp insertion variants was similar to that of ±3 bp substitutions (5.1).

In the case of substitutions, it has been argued that the kinetic stability and catalytic activity are positively correlated (Wyganowski et al., 2013). While I have reproduced that trend in the TriNEx library, the link between soluble expression and catalytic activity appears less pronounced in the InDel libraries (Figure 5.2). In both libraries I found multiple variants that have >75% of *wt*PTE expression, yet very low or non-detectable catalytic activity; here, the impact of the InDels is most likely on active site organization.

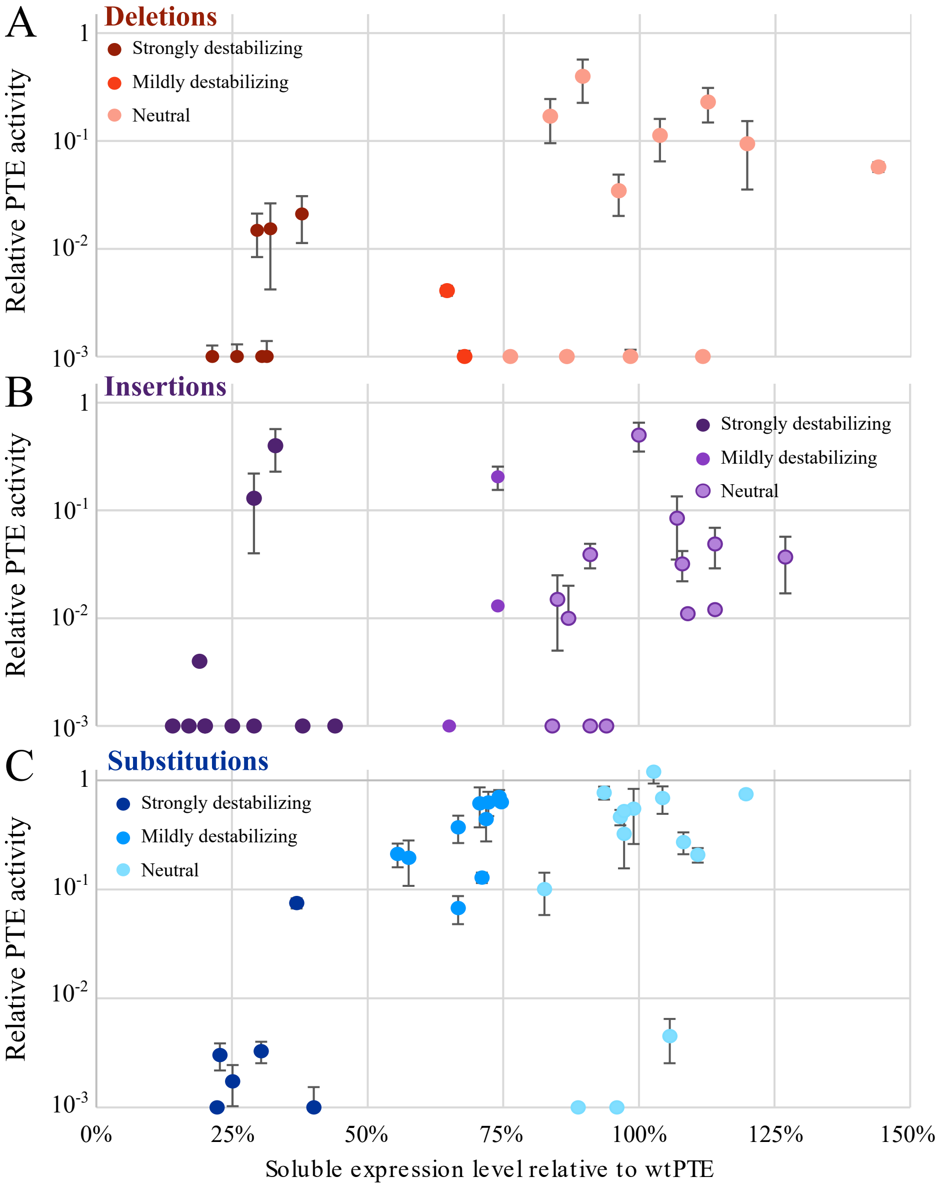


Figure 5.2 The relative paraoxonase activity of randomly chosen variants, as a function of soluble expression. Both parameters are normalised to *wt*PTE. A) -3 bp deletions. B) +3 bp insertion. C) ±3 bp substitutions.

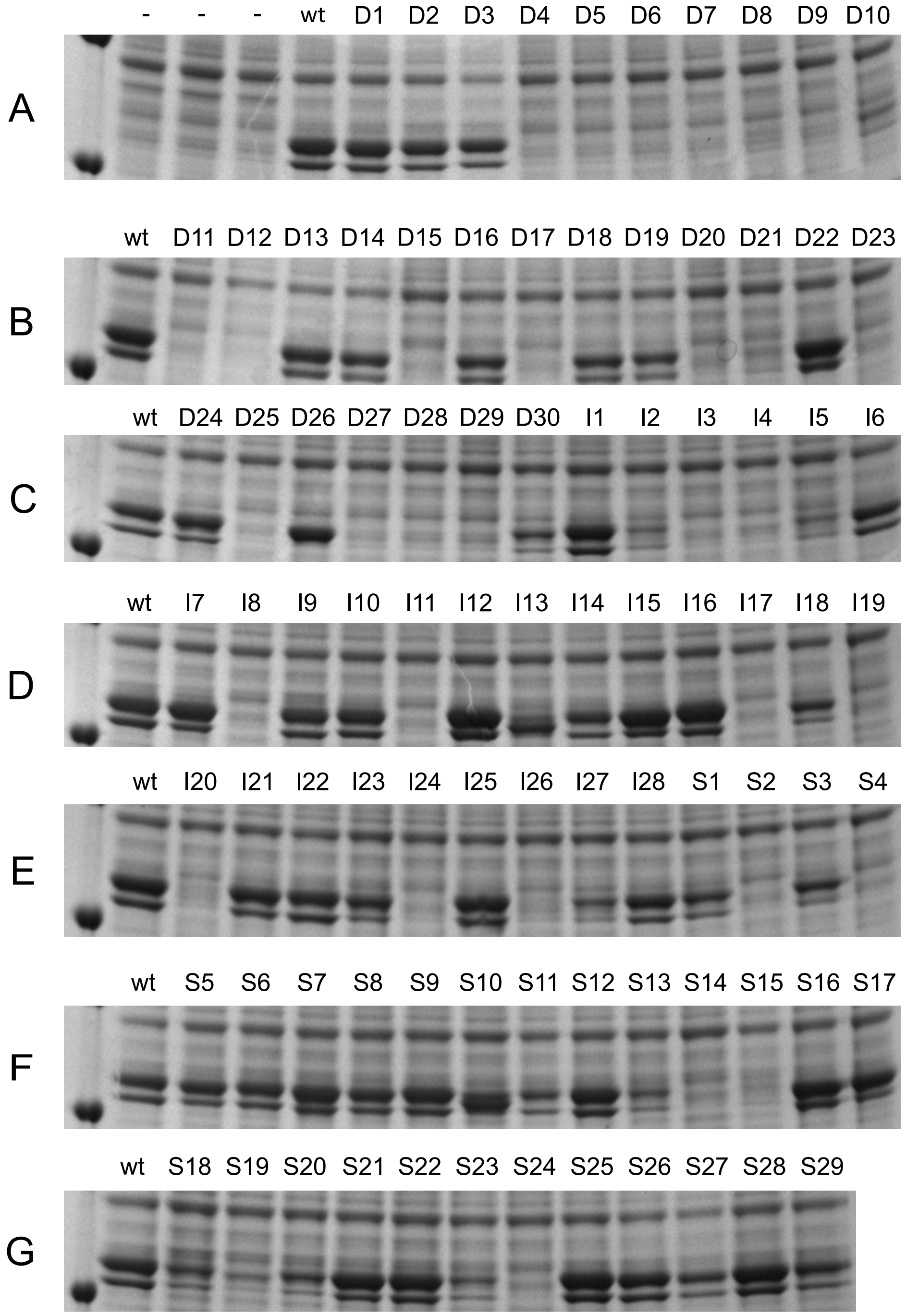

- - - wt D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
A
wt D11 D12 D13 D14 D15 D16 D17 D18 D19 D20 D21 D22 D23
B
wt D24 D25 D26 D27 D28 D29 D30 I1 I2 I3 I4 I5 I6
C
wt I7 I8 I9 I10 I11 I12 I13 I14 I15 I16 I17 I18 I19
D
wt I20 I21 I22 I23 I24 I25 I26 I27 I28 S1 S2 S3 S4
E
wt S5 S6 S7 S8 S9 S10 S11 S12 S13 S14 S15 S16 S17
F
wt S18 S19 S20 S21 S22 S23 S24 S25 S26 S27 S28 S29
G

Figure 5.3 The soluble expression levels of randomly chosen PTE variants, compared to *wt*PTE. D are deletion variants, I are insertion variants and S are TriNEx substitution variants. Each variant is numbered sequentially in arbitrary order. The leftmost lane in each gel shows the 35 kDa molecular wight marker (bottom) and in some lanes, the 40 kDa marker (top). The PTE protein appears as two bands around 37 kDa, and the intensity was quantified by normalization against the stronger band at 39 kDa.

5.3 Outlook

The postulate that protein stability promotes evolvability was first proposed in light of many proteins being only marginally stable, which limits their tolerance of folding-disrupting mutations (Bloom et al., 2006) before they cross over the stability cut-off. A common objection that plays a part in the rarity of studies using InDels for directed evolution is the perception that they are highly deleterious to both protein stability and activity. Here, I provide additional data that the reality is more nuanced.

Examining a set of randomly selected variants that carried a single 3 bp deletion, insertion or NNN triplet substitution, I again found that InDels are on average an order of magnitude more deleterious on enzymatic activity. However, the correlation with soluble expression levels was weaker and the InDel variants showed a greater distribution in observed solubility level. Surprisingly, the median solubility level in the +3 bp insertion library was similar to that of substitutions (Figure 5.1C), although the average solubility was lower. It appears that InDel libraries exhibit an all-or-nothing effect when it comes to solubility, while more substitutions show a medium level of soluble expression (Figure 5.2).

Considering only InDel variants with high soluble expression levels (>75% *wt*PTE), the data show a wide range of catalytic activity; these range from below detection level to close to *wt*PTE. This result suggests that the strongly deleterious average impact of InDel is only partially due to their reduced stability, but is instead a composite effect of i) disrupting folding and ii) sometimes disrupting the organization of the active site. This duality also indicates that it should be possible to find InDel catalysts that exhibit native or improved catalytic activity without compromising protein stability.

Indeed, I found this combination of properties when examining the stability of four best arylesterase hits. They exhibit similar levels of thermostability (T_m between 3°C lower and 6°C higher) compared to *wt*PTE (78°C), suggesting that thermostability of these variants is neither compromised nor necessarily a good proxy for mutational robustness, much as is observed with subsitutions. While the InDel variants were screened and selected as hits in the presence of chaperones, they for the most part did not acquire a high level of chaperone dependence: only one variant shows increased chaperone dependency as observed through catalytic activity, while all four show <30% increase in solubility when co-expressed with chaperones. Taken together, these results support our confidence that InDel libraries contain variants that are both highly active and not compromised in stability, which makes them good candidates for further rounds of directed evolution.

Chapter 6

Towards deep mutational scanning of InDels in PTE

This chapter describes the experimental assay development to monitor the phosphotriesterase reaction in microfluidic droplet format.

6.1 Introduction

Multiple lines of evidence show that small InDels have the potential to introduce valuable new or improved functionality in enzyme directed evolution campaigns. While so far, substitution mutagenesis combined with DNA shuffling has been the work horse of directed evolution, the potential of InDels remains insufficiently explored.

The colony and plate-based screen of PTE InDel variants performed by Dr. Stéphane Emond (Emond et al., 2020) showed that, although only a small proportion of the total diversity of insertion libraries could be screened, multiple adaptive InDels were still discovered. While it was encouraging to see that medium screening throughput is sufficient for finding some adaptive InDels (as would happen in the context of a directed evolution campaign), I also wishes to better understand the fitness landscape of InDels; for that, a high-throughput dataset is essential. Specifically, the three PTE insertion libraries contain between $8{\times}10^3$ variants for the +3 bp library, which is on the upper edge of screening capacity for a colony and plate-based screen, and $>10^5$ unique protein variants in the +6 and +9 bp libraries; this diversity can only be comprehensively screened with ultrahigh-throughput screening methods.

The previous results of GFP sorting (Chapter 2) show that some InDels can be tolerated in the compactly folded GFP variants, where the protein fold primarily comprises β-sheets.

Already having some data that PTE can tolerate InDels, I wished to extend the exploration of InDel tolerance in this model enzyme and generate a full high-throughput DMS dataset. Therefore, the first goal of this work was to develop an assay that can couple genotype (the PTE variant gene) and phenotype (an experimental measurement of enzyme catalytic activity) in a throughput above 10^6 enzyme variants per experiment.

Microfluidic assays have been shown to achieve that capacity, screening up to 10^8 droplets per day. However, the adaptation of fluorescence-activated droplet sorting (FADS) to PTE, needed in order to create a fitness landscape of the native paraoxonase activity, presented some unique challenges. An assay for PTE native activity, which would reflect on the inherent robustness of the enzyme, must cope with the high speed at which *wt*PTE operates. The standard microfluidic FADS workflow operates with droplet incubation times between one hour at minimum and multiple days (or however long the emulsion is stable), since it takes time to create and sort million of droplets - but this timeline is too slow for very active PTE variants. Consequently, the first consideration of the project was the creating and testing of new in-line FADS microfluidic chip designs which can cope with very short incubation times. In-line FADS designs were just shown to be possible at the time of this project and promised to allow monitoring of a reaction that is over in as little as five minutes.

6.2 Results

6.2.1 Feasibility of a microfluidic assay

In a microfluidic chip device, a suspension of cells is introduced onto the chip, where it is mixed with an aqueous solution containing the enzyme substrate and the lysis reagent. The now-mixed solution is immediately encapsulated and split into individual droplets. The cells are encapsulated at an appropriate concentration, such that most droplets contain either no cell or only a single cell. After mixing, the cell bursts open and releases the total amount of enzyme, and the enzyme concentration remains constant thereafter. The enzyme concentration (and through it the rate of reaction, the time for which the reaction remains in the linear range, and assay sensitivity for differences in activity) is thereby determined by the amount of enzyme produced per bacterial cell and the volume of droplets.

In a cell lysate activity assay in the 96-well format, the lysate must be diluted >1:1000-fold to be able to record a linear initial reaction velocity over ~15 minutes; at a lower dilution, the substrate is used up within seconds. Modelling an *E. coli* cell as a sphere with 1 μm radius, the cells has an approximate volume of 4×10^{-15} litres or 4×10^{-3} picolitres. Since microfluidic droplets typically have a volume betwee 1 and 10 pL, the lysis of one cell per droplet should achieve dilution to the correct order of magnitude. Testing of reaction rates in appropriate dilutions in cell lysate indicated that the lysate should be dilute enough in a droplet to be observed over 5-30 minutes. Comparing the two plasmids in which the libraries were available, a pET plasmid under the control of a T7 promoter and a pASK-derived pID plasmid under the control of a Tet repression system, I found that the lower background expression in the pID plasmid helped to improve the signal-to-noise ratio in the assay. In contrast, the T7 promoter was quite leaky and showed high enzymatic activity even when PTE expression was not induced. Since the challenge in creating a microfluidic assay for PTE was in having too much enzyme rather than too little, I chose to perform further work with enzymes in the pID-Tet plasmid.

Higher dilution is limited by the maximum droplet size that is compatible with droplet sorting, which prevented further exploration in that direction (too large droplets tend to break apart during sorting, which makes the outcome useless). Detailed exploration of different droplet sizes is also complicated by the requirement for many custom chips, since the droplet size is primarily determined by channel width and depth - manipulation of flow rates during an experiment can only modulate droplet size to a limited extent. Cooling the reaction in 96-well format also did not produce a substantial decrease in rate, while it would introduce experimental complexity in the microfluidic set-up, since it would require cooling the syringes, the tubing and the microfluidic chip device.

A standard modular approach to microfluidic droplet sorting consists of three steps: bacterial encapsulation in droplets (lasting 1-2 hours per sample), reaction incubation (from minutes to weeks, typically 4-48 hours) and droplet sorting (3-12 hours, depending on the number of droplets to be sorted). It should be noted that the order in which droplets are generated is lost during incubation, so the incubation time needs to be long compared to time needed for encapsulation/sorting or the variable incubation time between droplets can greatly increase the inherent noise in the system.

Since I established that a PTE native activity assay could be compatible with the microfluidic format as long as the incubation time were short enough, I started to explore the use of in-line microfluidics devices (Figure 6.1). These integrate encapsulation, incubation and sorting on a single chip. Their advantage is that a shorter incubation times are possible (1-30 minutes), depending on the chip design, and that the droplets are sorted in approximately the same order as they are generated. This maintains a mostly uniform incubation time across the sample. However, the in-line devices are a recent technical advance in microfluidic technology and are much more challenging to manufacture and operate.

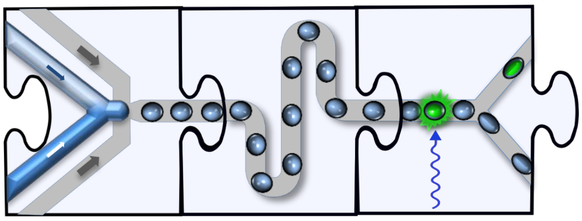


Figure 6.1 The steps in fluorescence activated droplet sorting. From left to right: bacterial encapsulation through flow focusing, incubation in a delay line, and droplet sorting.

Microfluidic device masters are most commonly manufactured with soft lithography techniques. During master fabrication, a thin polished silicon wafer is coated with a photoresistor material (e.g. SU-8) at a certain depth and exposed to UV-light through a patterned photomask printed on a transparency. Where the mask is clear, the photoresist material is exposed to light and it hardens during baking, while the rest remains soft and is washed off in the final steps. The hardened photoresist therefore forms a sharply defined pattern on the silicon support, and the pattern is then copied onto the PDMS elastomer. To create individual chips, the PDMS elastormer is mixed with a curing agent, poured over the patterned silicon master to solidify, then peeled off and bonded to a glass slide. Finally, the newly bonded chip is treated with silane to make the channel walls hydrophobic, which prevents aqueous droplets sticking to and breaking on channel walls (McDonald et al., 2000).

Fabrication of multiple-depth device masters require two or more photomasks, one for each depth. The first layer is manufactured in the standard way. Then, the second layer photomask must be very precisely aligned to the hard photoresist pattern created in the first step, and the process is repeated. The alignment requires additional features in the photomask (easy to achieve) and specialised skill and equipment (harder).

6.2.2 Evaluation of incubation line designs

At the time of this project, microfluidics chips that integrate droplet generation with a delay line have been used for measuring enzyme kinetics, but integration with a in-line sorting device had not yet been achieved. Before working to integrate all three chip components in a single design, I set out to replicate existing results with delay lines to validate the utility of this approach for measuring PTE activity. Thus, the first chip designs combined the flow focusing for droplet generation and different delay line styles. Two general approaches were available:

- Long, shallow delay lines: these were previously used in the research group to follow the reaction progress with arylsulfatase variants on the time-scale between one and

15 minutes (Kintses et al., 2012). The chip design was available in the laboratory for immediate fabrication.

- Deep delay lines: these are a more advanced design (Frenz et al., 2009) that integrate a shallow layer design (essential for flow focusing, or the droplets are too large) with a deeper delay line, which reduces the pressure in the delay line and slows down droplet travel speed. Consequently, the flow through the chip is more stable, the pressure in the delay line is reduced and longer incubation times can be achieved. However, the chip manufacturing process requires a specialised step where the two designs need to be precisely aligned. At the time of this work, this had not yet been achieved in the lab. The publications describing the use did not share the design files or describe the design in sufficient detail to reproduce the design, so they instead served as a starting point for re-development of these devices.

When I first started this project, the process for manufacture of two-depth devices was not yet in the technical scope of the group, so I started with exploration of different single-depth designs.

Shallow, narrow long delay line, uniform depth. The design by Kintses et al. (2012) is manufactured with uniform 25 μm channel depth, with a 20 μm flow focusing nozzle and a 300 μm wide delay line. One version of the device also incorporates evenly spaced 30 μm wide constrictions, which promotes mixing of the droplets. Droplet generation with these designs was possible, but the chip suffered from high back-pressure from the long delay line; this resulted in unstable droplet generation and a short lifetime of each experiment, before the bonding of the chip to the glass slide was disrupted and the chip de-laminated. I also observed that the droplets in the centre of the delay line travelled at a higher speed than those by the channel edges. This variability is concerning because a droplet with a longer incubation time will appear to carry a more active enzyme, even if two variant have identical true activity.

Wide delay line, uniform depth. This design was modelled on the wider, deeper chambers that were described as being resistant to issues with high pressure (Frenz et al., 2009) (Figure 6.2A). While we were not yet able to manufacture a device master wafer with two channel depths (this is determined by equipment available in the clean room facility and also requires specialised training), I tried a design with 20 μm flow focusing width and wider incubation chambers with a restriction point every 3 centimetres. The chips were manufactured at 25 μm and 50 μm channel depth and both depths evaluated. I also attempted to seal the two chips with different channel depths together with methanol to get deeper channels (instead of bonding to a glass slide), but this bonding was too weak to withstand normal operating pressure.

Compared to the Kintses narrow line design, the T-shaped flow focusing proved to be more reliable than the V-shaped design. The oil extraction outlet was incorporated at the beginning of the line and it did remove oil from the line, which packed the droplets into the delay line. However, while this feature was promising, the oil extractor pillar were too small to also block droplets from escaping from the chip (since pressure in the early oil collection tube is lower than in the delay line). A further issue with this design was that the channels in the delay line were positioned too close, so there was not enough PDMS between two channels to form a strong seal between the chip material and the supporting glass slide. Therefore, the chips easily de-laminated, either during silane treatment or during droplet generation.

I found that modifying the design of the oil extraction element using either thicker individual pillars or a T-shaped obstruction element. These designs improved the packing in the delay line, but were still too small to fully obstruct the flow of droplets.

Wide nozzle, wide delay line. This iteration used a single depth design that was set up to generate larger droplets (50 μm channel depth and nozzle width) and spaced out the delay line channels further apart to improve the strength of bonding (Figure 6.3A). Due to the larger volume of the channels, this chip was more reliable to operate and did not suffer from back pressure issues. Using the oil extraction feature, the incubation time was 90 seconds. Over this time, the droplets reach the outlet as a defined front without dispersing, which indicates that the constriction mixing is efficient (see Figure 6.3C and 6.3D for depiction of droplet movement through constrictions).

With this first usable in-line incubation device in hand, I assessed if the very short incubation time is long enough to reliably detect PTE activity on the fluorescein phosphotriester substrate (Figure 1.7 - **2**). Single cells expressing *wt*PTE from a pET plasmid were encapsulated in large droplets (~270 pL volume). I observed a strong signal and clear separation between empty and positive droplets (Figure 6.3E), even at high dilution inherent to large droplets.

Throughout chip design development, I was working towards approximating the conditions that should be available in a two-depth in-line incubation device: long, deep, wide delay lines with oil extraction which pack the droplets, so that a longer incubation time is achievable. I observed that the deep, wide delay design in Figure 6.3 generated stable droplets

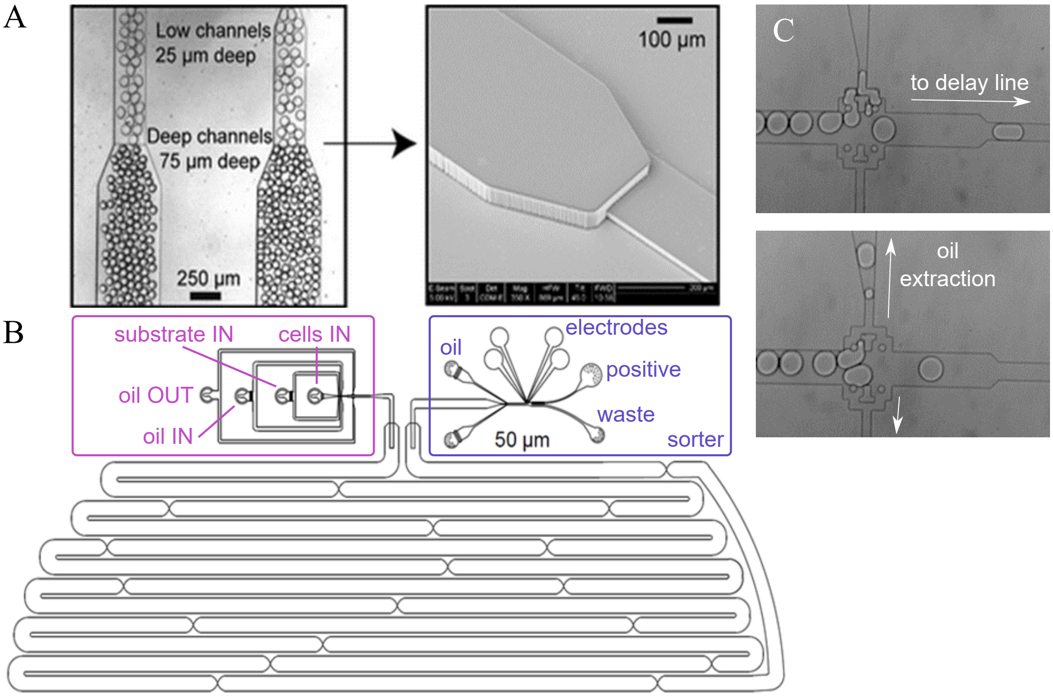


Figure 6.2 A) The transition between shallow and deep channels (reproduced from Frenz et al. (2009)) shows that droplets are squeezed in the shallow low channels and relax into spherical shapes when they reach the deep channels. B) The schematic outline of the wide delay line design, which can be manufactured in one or two layer depths. There is one outlet and three spherical inlets in the flow focusing module (magenta box): oil extraction outlet, oil inlet, substrate solution inlet and bacterial inlet. The latter three meet at a T-shaped flow-focusing junction, where droplets are generated. Droplets then flow past the oil extraction feature, where most of the oil is drawn out towards the oil extraction outlet, and the droplets continue into the wide delay line channels. Droplets mix every time they pass through a narrow constriction, facilitating an even travel rate. Finally, droplets reach the sorter on the top right. Channels are intended to be 25μm deep in the flow focusing and sorter module, then 50 μm deep in the delay line (1000 μm wide, 60 μm constrictions). C) Oil extraction after droplet generation. The pillars are too small to be reliably fabricated and do not prevent droplets from escaping.

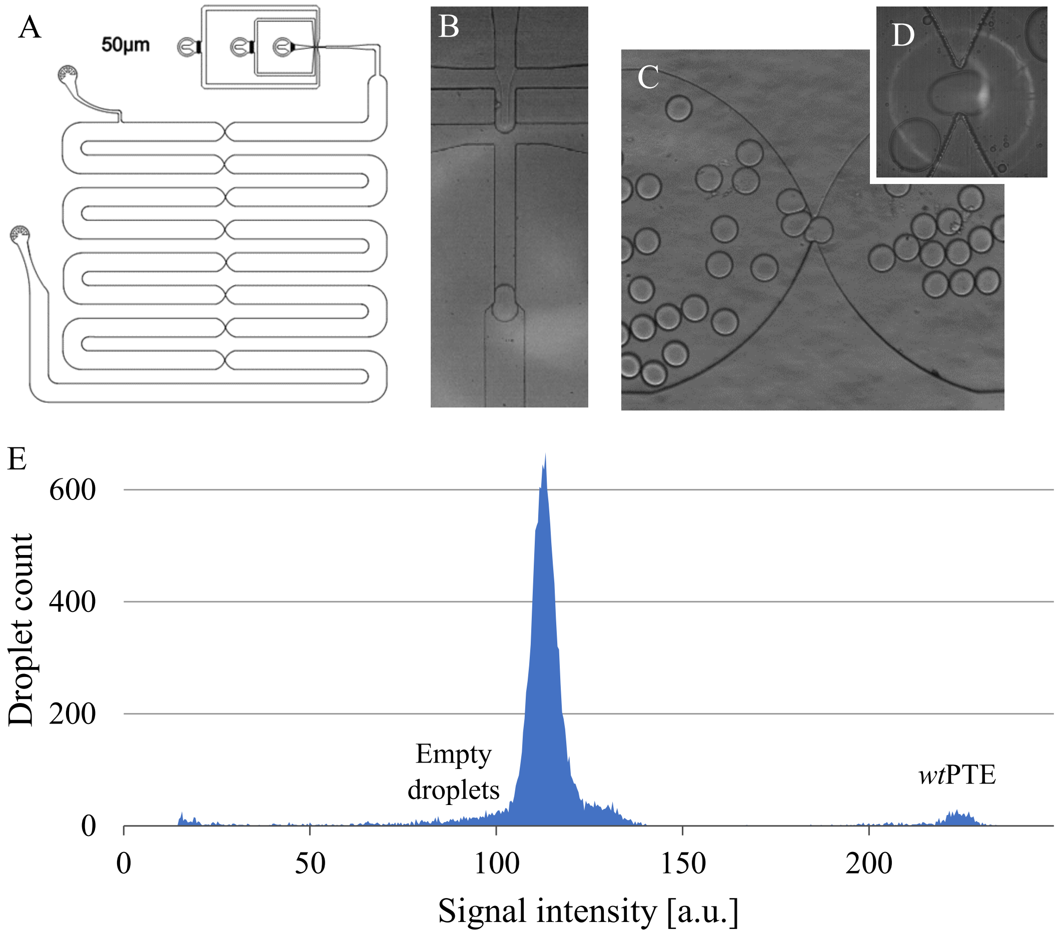


Figure 6.3 A) Single depth short delay line design with 50 μm nozzle size. B) Flow focusing. C) Droplets moving through a constriction in the chip and mixing. D) Close up of a droplet passing through. E) Histogram of fluorescent signal intensity with bacteria expressing *wt*PTE encapsulate in 270 pL droplets. Droplet occupancy was such that 95% droplets were empty and 5% contained a single cell per droplet.

and showed a strong fluorescent signal even at short incubations and high enzyme dilution. This result was encouraging for further development of a microfluidic device that could be used for PTE in-line incubation and droplet sorting. However, the challenge remained for achieving longer incubation times, smaller droplet sizes (which is a necessity for droplet sorting) and greater robustness of microfluidic device operation.The deeper delay lines are also more robust to operate, have much reduced back pressure (which makes chip operation unstable) and are less prone to de-lamination.

However, the large droplet size in this device is incompatible with sorting (270 pL in device cf. 2 pL in sorting chips), and smaller droplets must be generated in chips with smaller nozzle size and shallower channel depth. The droplet sorters also require shallow channel depth. Thus, while the chip development was progressing in the right direction, I concluded that these conflicting depth requirements (deep for the incubation line, shallow for flow focusing and sorting) could only be reconciled in a two-depth device. At the time the manufacture of two-depth devices was beyond the technical capability of the research group and the lithography facility in the Cambridge Nanoscience Centre, so I could not pursue the assay development further at that point and suspended the work.

Later, the process of creating microfluidic chips with two channel depths was within the technical capability of the group. This was thanks to changes in the technology available at the clean room facility, and thanks to technical expertise brought into the group by the arrival of Dr. Tomasz Kaminski. By then, David Schnettler's work on evolving a faster non-metal-dependent phosphotriesterase enzyme had progressed to a catalytic rate that also required a microfluidic device suitable for short incubation times. Together, they re-started the work on a microfluidic device suitable for in-line droplet generation, incubation and sorting and succeeded in creating a functional two-depth device.

Two-depth in-line flow focusing, delay line and sorter. This new design combines a $16 \times 25\mu$m flow focusing junction with shallow channel depth (15 μm), a deeper (25-32 μm) incubation line with either 5 or 20 wiggles, and finally a sorter device without bias oil at the same designed channel depth as the incubation line (measured 28 μm). The devices also performed better when operated with glass syringes, which give more stable flow rates, and the RAN surfactant (2% w/w). The devices generates droplets with approximately 11 pL volume and, depending on the length of the delay line, allows droplet incubation time between 5 and 30 minutes.

Using this device, I recorded the phosphotriesterase activity at the end of the short, 5-wiggle incubation line (7 min incubation time) and sorted the droplets with positive activity

(Figure 6.4B, signal strength >4V). While the droplet occupancy was unexpectedly low, there was a clearly detectable positive signal.

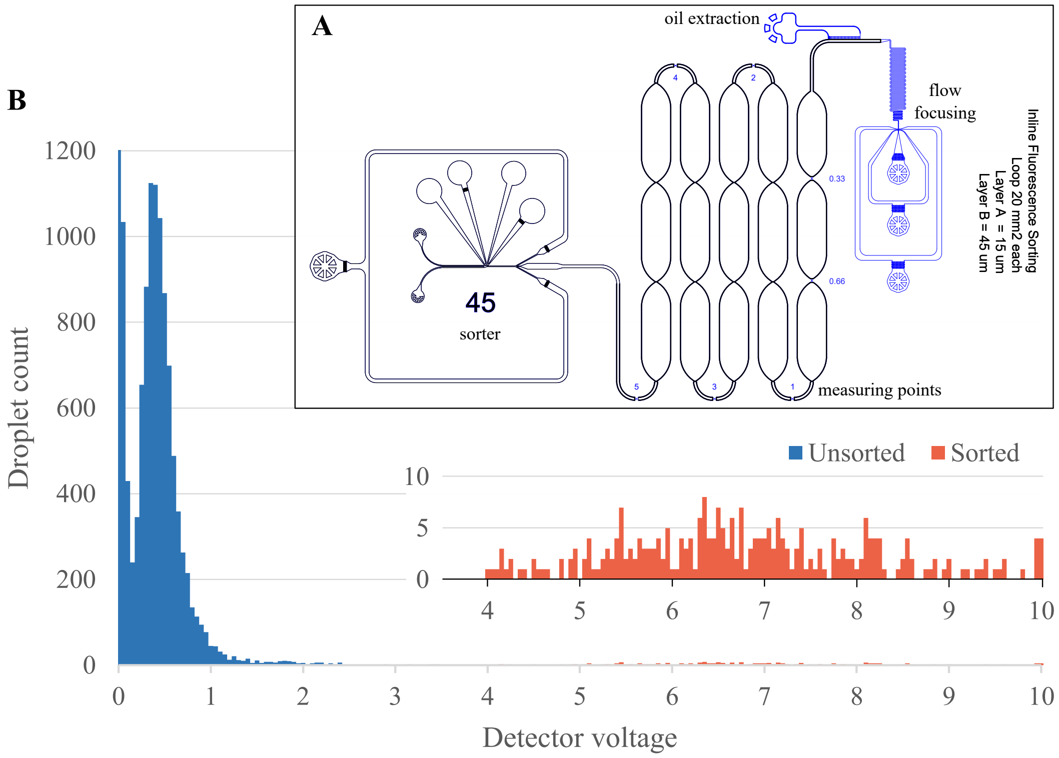


Figure 6.4 A) The two-depth integrated microfluidic chip design, integrating droplet generation with flow focusing, 5 lines of incubation line and a fluorescence activated droplet sorter. The blue line show the shallow mask design (aiming for 15 μm) and the black lines the deeper, additional layer on top (aiming for 45 μm total, achieved 30 μm). Droplets are generated in a three-inlet flow focusing device, then pass past an improved oil extraction feature (right) and entering the wide delay line. The delay line has narrow mixing points as before, as well as shallow measuring points that can be used to measure the fluorescence signal at different incubation times. Last, droplets are spaced out within the sorter feature by injection of additional oil. After spacing, individual droplets pass the final measuring point and are separated according to defined voltage gates. B) A histogram of *wt*PTE activity achieved with this device. The majority of droplets are empty (blue) and a small proportion of droplets show a positive signal (inset, orange).

6.3 Outlook for future work

This section briefly describes the steps required to bring this project to completion, as was originally proposed in the first year report.

Determining the optimal incubation line length. While I showed that there is a detectable signal when using the short two-depth integrated device, the rate of positive droplets was low and thus the experiment should be repeated. Additionally, the length of the incubation line in the current designs may not be optimal, so project revival should start with determination of optimal incubation line length. Using the two-depth integrated device design with a long incubation line, a population of droplets is produced (aiming for approximately 10% occupancy) and the fluorescence monitored at the shallow, narrow constriction point included in every turn in the chip design. The droplets are excited with the 488 nm laser and the fluorescence recorded in the same manner as during droplet sorting, on the microfluidic rig with the laser and fluorescent microscope. This experiment should produce a histogram showing the distribution of droplet fluorescence at each time point, with the positive population appearing at increasing voltages with longer incubation times. The flow rates and the incubation line volume can be used to calculate the incubation time. The optimal incubation line length for droplet sorting is one where the signal is strong, but not yet saturated (where all droplets with the enzyme have converted all of the substrate, so the signal is no longer changing).

Timeline: two weeks

Microfluidic chip re-design. If the reaction time-course shows that a different incubation line is needed, the chips can be re-designed in two ways. One, the photomasks used for creating of the existing two-depth device can be re-used to generate a chip master with a deeper incubation line: the design was intended to have a 45 μm deep incubation line, but only achieved approximately 30 μm channel depth in practice. The channel depth is an imprecise science and is controlled by the amount of SU-8 placed on the silicon wafer and the spin rate during coating, so it could be increased by reducing the spin rate. Two, the device design can be altered to increase or decrease the length of the delay-line to the optimum length. If the reaction is still too fast, then a new integrated design can be assembled with the 50×50 μm sorter and with the corresponding flow focusing design - then the incubation line depth can be increased to 50-60 μm.

Timeline: 0-3 months, depending on complexity and the amount of device designs required. The device design process is low-intensity and is ideally done in parallel with other projects.

Enrichment sorting. Once the device design is adequate, the protocol for activity sorting is fine-tuned with an enrichment sort. Here, a population of negative cells (expressing ACP protein) and a population of positive cells (*wt*PTE) are prepared separately, then mixed in a

99:1 negative:positive ratio by cell density. The mixed population of cells is encapsulated in droplets in the integrated in-line device, then sorted into a positive and negative channels. The DNA from the positive channel is recovered and transformed into electrocompetent cells, then the number of colonies on an agar plate is counted and compared with the number of sorted events to indicate % of event recovery. Given the variability in DNA recovery observed between protocols and operators, a large number of droplets should be sorted in this step ($> 10^4$ positive droplets). Randomly selected colonies are tested in cell lysate for enzymatic activity to measure the efficiency of the sorting.

Timeline: three to four weeks, depending on the efficiency of DNA recovery.

Library sorting. At this point, the libraries can be sorted and should be screened at the experimentally feasible coverage. Assuming 10% droplet occupancy, a +3 bp insertion library in PTE containing 50,000 protein variants would require screening 3 million droplets for 3-fold coverage, which is well within experimental feasibility. The +6 and +9 bp libraries are larger (>100,000 variants), so the screening may need to be done over two or three days. After screening, the DNA is recovered and again a random set of variants from each sort chosen for a cell lysate activity screen to measure the sorting efficiency and sensitivity (the activity level of the positive variants). A subset of these variants can also be Sanger sequenced to provide a dataset for validation of the NGS analysis.

Timeline: one month

DNA preparation and Illumina sequencing. Since the input libraries have already been sequenced and the composition analysed, only the enriched positive fraction needs to be sequenced at this point (sequencing of the negative gate is possible, but likely not cost-efficient). The appropriate DNA band is excised with restriction endonucleases, purified by gel electrophoresis and submitted to the Sequencing Facility.

Timeline: two weeks

Data analysis and validation. The NGS data can be analysed with existing scripts that have already been used for analysis of PTE InDel library composition, and for exploring enrichments in the GFP sorting experiment. After that, the project becomes more open ended and can explore the location of beneficial mutations, tolerance of insertions depending on the amino acid inserted, correlations between protein surface charge / hydrophobicity / flexibility and the effect of InDels, and more.

Timeline: one month to infinity...

Chapter 7

Sequence-function mapping the MKK1-ERK2 interaction

This project was done in collaboration with Remkes Scheele, who designed the MKK1 library and performed all experimental work. My contribution to the project was input to the sequencing design and all NGS data interpretation. The figures and methods described in this section were originally prepared for a manuscript and are reproduced or adapted for this chapter.

7.1 Introduction

The InDelScanner scripts were first developed out of necessity, to enable an analysis of the InDel libraries that were generated via the TRIAD mutagenesis method (Chapter 4). TRIAD libraries almost exclusively contain a single insertion or deletion per variant, at random positions, so the focus in the analysis of TRIAD libraries was in placing the InDels, analysing their composition and quantifying the biases in the finished libraries. I used the same approach with additional calculations to generate enrichment statistics on the GFP datasets (see Chapter 2). Here, I show the application of InDelScanner scripts in a deep mutational scanning experiment on a SpliMLiB library, which has a different, more complex design (Lindenburg et al., 2020).

7.1.1 The interaction between human kinases

In this project, we probed the functional interaction between two human kinases in the mitogen-activated protein kinase pathway, namely the kinase MKK1 and its downstream partner ERK2. A key element of the MKK1-ERK2 interaction is mediated by the docking domain (D-domain), located on the N terminus of MKK1 (Figure 7.1). The D-domain is a short peptide sequence which binds its cognate pocket in ERK2, and a productive interaction enables MKK1 to phosphorylate and thus activate ERK2 (Jacobs et al., 1999).

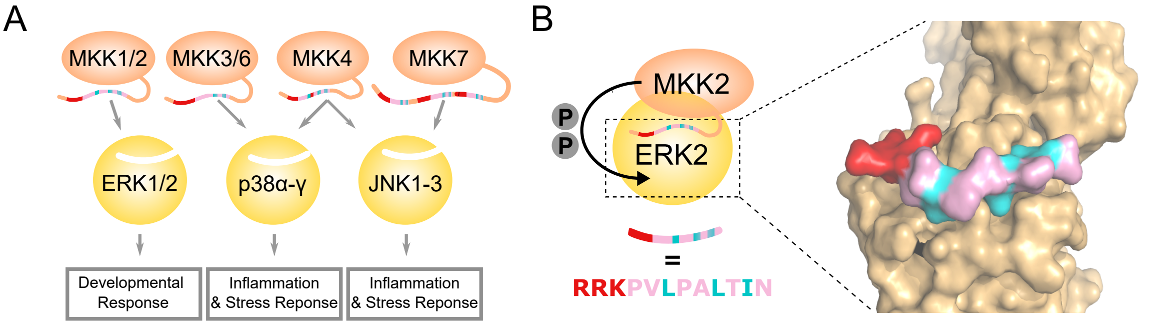


Figure 7.1 An overview of human mitogen activated protein kinases (MAPKs). A) A schematic overview of MKK cognate interactions with MAPKs and their downstream effectors. The D-domain is visualised as the multi-colour section of the unstructured N-terminal region of different MKKs. B) Phosphorylation of ERK by MKK is facilitated by MKK's D-domain binding the D-domain recruitment site (DRS) of ERK2. Shown here is the D-domain sequence of MKK2: basic residues (red) bind an acidic patch of residues in the ERK2 DRS, while the large hydrophobic residues (cyan) bind hydrophobic grooves within the ERK2 DRS. Spacer residues are shown in pink (PDB: 4H3Q).

7.1.2 The *in vitro* experimental assay

The core experiment of this project was the deep mutational scanning of the D-domain for ERK2 activation, which was inspired by the protocol by Podgornaia and Laub (2015). Within the D-domain, the key functional elements are a basic patch of residues at the start and some hydrophobic residues that follow, especially Leu or Ile at positions 9 and 11 (residue numbering follows MKK1 sequence). In the library design for this DMS experiment, the basic and spacer residues were left intact, while six residues in the hydrophobic region were randomised with the Split-and-Mix Library on Beads (SplitMLiB) protocol (Lindenburg et al., 2020). The gene library is assembled onto beads with sequential ligation of randomised DNA fragments. The randomised residues alternate in the sequence, such that they face the same side of the β-strand when the MKK1-ERK2 interaction occurs (in itself, the D-domain is an unstructured loop).

The *in vitro* assay. The assay for D-domain mediated activation of ERK2 (Figure 7.2) hinges on separation of individual MKK1 variants on monoclonal beads, which are isolated from each other in a droplet emulsion (Diamante et al., 2013). The phosphorylation reactions occur within the emulsion, such that the faster MKK1 variants phosphorylate more ERK2 (which in turn phosphorylates the substrate peptide subGFP on the bead), while the slower, less active MKK1 variants trigger a weaker cascade response (Rodems et al., 2004).

The starting point for the assay are paramagnetic beads, which each carry many copies of the same ^{ca}MKK1 gene (the genotype) and the subGFP target peptide (which will encode the phenotype at the end of the cascade). The subGFP target peptide is an N-terminally immobilised proline-directed substrate peptide of ERK2 (VA**PFSP**GGRAK), labelled with a C-terminal GFP. The beads are emulsified with oil, surfactant, purified ERK2 and an in-vitro transcription-translation (IVTT) kit in a polydisperse droplet format. While in droplets, MKK1 variants are expressed within individual droplets, which triggers the cascade reaction. The reaction is stopped after a fixed incubation time (3 hours), which separates faster and slower ^{ca}MKK1 variants. A chymotrypsin digest removes GFP from beads that carry less active variants, thus encoding the phenotype. Now, the ^{ca}MKK1 library was sorted with flow cytometry with sorting gates set against same-day positive (high ^{ca}MKK1 activity, high GFP retention) and negative (inactive ^{ca}MKK1 construct, low GFP) controls.

Randomly selected variants from all three activity gates were recovered, re-expressed and the activity tested in a secondary assay against a FRET peptide (Figure 7.3B). In the secondary assay the efficiency of the MKK1-ERK2 phosphorylation cascade is measured through the intensity of the FRET signal in the sensor: active ERK2 phosphorylates the linker peptide, which is thus resistant to chymotrypsin digestion and retains a high FRET signal.

The secondary screen validated the efficiency of the fluorescence-activated bead sorting and showed that there is a genuine difference in MKK1 activity levels between caMKK1 variants sorted into different fractions. Therefore, the D-domain sequence from the sorted libraries was recovered with PCR, where primers contained overhangs with adapter sequences for Illumina sequencing, and the libraries were submitted for Illumina Miseq targeted amplicon sequencing. Because of different diversity of the sorted fractions, the low fraction was sequenced at higher loading than medium and high activity fractions: the low fraction contained 85% of sorted beads, while the high activity fraction captured 3.2% of the sorted beads.

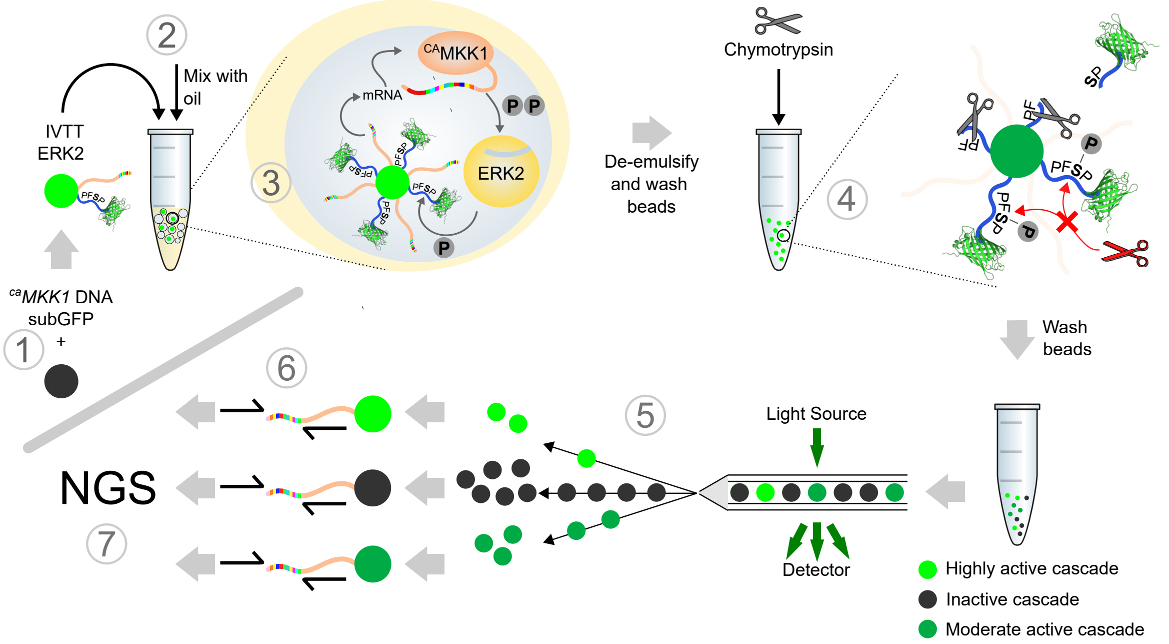


Figure 7.2 The *in vitro* screening assay used for probing the MKK1-ERK2 interaction. (1) Paramagnetic beads (black, then green circle) are functionalised with the substrate peptide subGFP and randomised ^{ca}MKK1 genes (orange wavy lines). (2) Beads are emulsified by mixing the oil-surfactant solution with an aqueous solution containing *in vitro* expression components (IVTT) and purified ERK2 protein. (3) Expression of ^{ca}MKK1 in the emulsion droplets starts the cascade by phosphorylation of ERK2. Active, phosphorylated ERK2 next phosphorylates serine within the substrate sequence of the subGFP construct. (4) After de-emulsification, beads are exposed to chymotrypsin protease, which cleaves off GFP from unphosphorylated subGFP. Beads encoding an active ^{ca}MKK1 mostly carry phosphorylated subGFP, which is resistant to chymotrypsin digestion, leaving subGFP attached to the bead. Thus, the bead links genotype and phenotype information. (5) Subsequent flow cytometric sorting of the beads based on subGFP fluorescence into activity-set gates followed by (6) PCR amplification of the D-domain and (7) deep sequencing allows for coupling of the cascades activity to the encoded MKK1 gene.

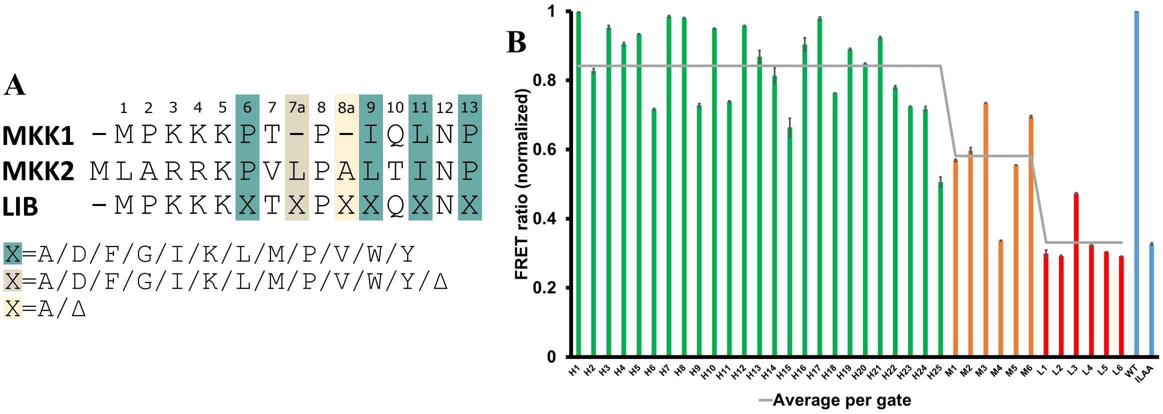


Figure 7.3 The library design and quality of activity sorting of MKK1 libraries. (A) Design of the randomised D-domain library of caMKK1, shaded residues are randomised with the residues shown below. The residue numbering is based on MKK1's wild type sequence and 'a' indicates an inserted residue. Δ indicates no residue was introduced at that position. (B) Randomly recovered ^{ca}MKK1 sequences from the high activity sorting gate (green), medium activity gate (orange) and low activity gate (red) were re-expressed and incubated in 96-well plate format with ERK2 and a FRET sensor containing the substrate for ERK2 in-between the fluorescent proteins. All values are normalised to ^{ca}MKK1. The horizontal line shows the average of each gate. There is a good correlation between the activity in the FRET assay and the sorted activity bin, mixed with some expected biological variation.

7.2 Results

7.2.1 Recording library composition

Initially, I attempted to analyse the composition of sequencing libraries from the three sorting fractions (high, medium and low activity) with the same pipeline as I used for TRIAD libraries. However, these libraries have very high randomisation in a short sequence (four randomised codons, one optional single codon insertion and one optional randomised codon insertion, all within an eight codon stretch - see Figure 7.3A), which made it impossible to accurately place the variants with the NW algorithm in DNA sequence, even with multiple adjusted scoring values. The NW scoring system roughly models natural evolution, where it is more likely that a complex sequence difference occurs through one or two mutagenic events (i.e. one longer substitution and one insertion) that mark gene divergence, than through multiple independent short mutations. Because of the nature of the scoring system, it is impossible to give the NW or similar algorithms "fixed" residues - while the fixed residues are obvious during manual inspection of the sequence, the algorithms evaluate alignments blindly, without recourse to library design.

Instead, the D-domain and surrounding DNA sequence were extracted from assembled reads using a regular expression (see Methods) and translated into protein sequence, which was then aligned with NW algorithm using a custom scoring matrix and a reference sequence with X in place of randomised residues. A subset of aligned sequences was manually verified to confirm that it contained all types of variants (substitution and insertion combinations in all randomised positions), then the full alignment was parsed to extract the variants and counts.

Unlike TRIAD libraries, the SpliMLiB libraries contain substitutions, so here sequencing errors are indistinguishable from true sequences. In theory, if a variant codes for a codon that was not designed into the library, it should be a sequencing artefact, but it could also have been introduced as a polymerase error during library assembly. The analysis therefore largely relies on filtering the variants by the number of supporting sequencing reads, which was done in two stages. In the first filtering step, only variants occurring $\geq 10\times$ in the high gate are considered (and 99.2% of 3.67×10^4 these variants are on-target). However, the medium and low gates are more diverse, so a lower cut-off ($\geq 3\times$) was used to detect more variants. This cut-off detected 4.90×10^5 variants in the low gate and 1.63×10^5, 95.5% of which had the correct design (Figure 7.4).

The input library was designed with equal proportion of all randomized amino acids for every position, which corresponds to 1/12 proportion of each amino acid at positions 6, 9 and 12; 1/13 at position 7a because the insertion is optional; and 1/2 at position 8a which was randomized with equal probability between an Ala insertion or no insertion. Out of theoretical diversity of 539,136 variants in the SpliMLib library ($12^4 \times 13 \times 2$), the deep sequencing detected 492,244 target variants (91%). The observed diversity in the low gate, which contains the majority of detected sequences including inevitable false negative sequences, shows the library was well balanced at all randomised positions (Figure 7.4A).

Next I created a curated dataset of active variants by further filtering the variants abundant in the high gate ($\geq$10 NGS reads). The curated dataset was made from two sets of variants with different requirements: a) variants with $\geq$50 reads in the high gate, with were selected for low reads representing less than 1/3 of all reads for that variant, and b) variants with a high gate read count 10-49, medium gate count lower than high gate count, and low reads representing less than 1/6 of all reads for that variant. This curation removed 11% of the abundant high gate sequences, which likely comprised mostly the expected 8% of false-positive sorting events.

Considering each position independently, the enrichment for different amino acid residues (defined as the amino acid proportion in the curated dataset divided by the proportion in

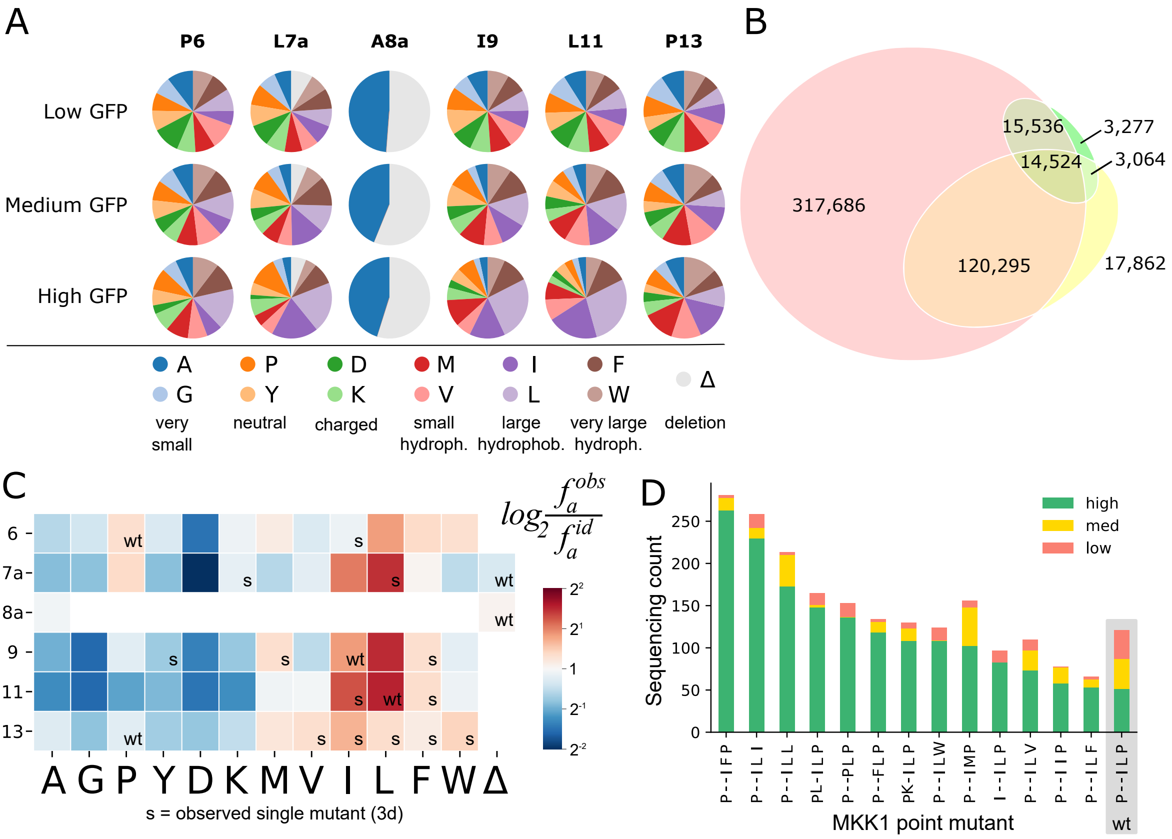


Figure 7.4 The amino acid distribution observed in deep sequencing of the activity-sorted MKK1 variants. A) A comparison of amino acid distribution of D-domain sequences recovered from the low activity gate (top, ≥3 sequencing reads), medium activity gate (≥3 sequencing reads) and the high activity gate (bottom, ≥10 sequencing reads). The low activity sequences show a balanced distribution of amino acid frequencies at each position, meeting expectations for a balanced starting library. Medium and high gate recovered sequences show a higher proportion of large hydrophobic amino acids, suggesting that they contribute to D-domain complementary to the ERK2 DRS. B) Venn diagram displaying the number of unique sequences found in each activity gate and their overlap. C) The enrichment of observed amino acids (f_a^{obs}) at each position in the curated set of 3.2×10^4 active variants found in the high gate, relative to expected frequencies (f_a^{id}) in a perfectly balanced library. The wild-type MKK1 residues (wt) and the active single site mutants (s) observed in the curated dataset are indicated. D) The observed distribution of sequencing read counts for caMKK1 and functional point mutants of caMKK1. Only the six randomised residues are listed in the sequences.

an ideally balanced library) is shown in Figure 7.4C. Overall, the D-domain sequences in the curated dataset show a strong enrichment for Leu and Ile at randomised positions 9 and 11. This trend follows the wild-type residues in MKK1/MKK2 and conforms with the highly conserved nature of these residues, and the complementarity of these residues to binding pockets on ERK2 (Figure 7.1B) (Bardwell and Bardwell, 2015). Similarly, there is a preference for the MKK2-like hydrophobic residue insertions at position 7a, favouring especially the insertion of Leu but also Ile and to some extent Pro. This third Leu in MKK2's D-domain (L7a) is complementary to an additional hydrophobic pocket in the ERK2 binding groove2 (the 'lower pocket). Thus, the large hydrophobic residues at MKK1 positions 7a, 9 and 11 form the core of the D-domain for targeting of the ERK2 DRS.

The randomised positions at the edge of the randomised region show mild overall sequence preferences: whereas Pro is present in positions 6 and 13 in both the wild type sequences of MKK1 and MKK2, the deep mutational scan reveals only a weak preference for Pro at position 6, and a general enrichment for any hydrophobic residue at position 13. As expected, the negatively charged residue Asp was depleted across the library, especially in positions 7a, 9 and 11, which form the core of the D-domain hydrophobic patch. Positively charged Lys was neutral in position 6, next to the already basic patch of amino acids in MKK1, and depleted otherwise. Similar to charged residues, the very small residues glycine and alanine were depleted in all positions, consistent with the known effect of alanine to disrupt MKK1 binding as used in the negative control construct ^{ca}MKK1-I9A/L11A (Figure 7.4C).

7.2.2 The MKK-ERK2 interaction exhibits widespread positive epistasis

General epistasis. After determining the single position preferences in the MKK1 D-domain, the next questions was whether the different positions showed epistatic interactions; that is, combinations of residues where the status at one position influences the preferences at another position. Particularly interesting was the presence of positive epistasis, which refers to a pair of mutations in different positions together exhibiting a more beneficial effect together than the sum their effects in isolation (that is, as point mutations) (Lehner, 2011).

I first investigated the presence of strong epistasis between positions in the D-domain through calculation of mutual information (see Methods), and found no overall covariation between the randomised positions in the D-domain. This is expected given the flexibility of the D-domain, and implies that positions can evolve independently. However, while the covariations are low when averaged across all genotypes, this does not preclude interactions between specific genotypes.

Specific positive epistasis widens D-domain active sequence space. The sequences in the curates set of active variants (3.2×10^4 variants, or 7.0% of all sorted sequences) contained 13 point mutants relative to wild-type MKK1 (Figure 7.4D), while 43 out of remaining 44 possible MKK1 point mutants were enriched in lower activity gates. Similarly to the overall preference for Leu/Ile residues in active D-domains, the point substitutions in these 13 variants tend to be hydrophobic. If only these substitutions were allowed in a combinatorial fashion, we would expect to observe at most $2 \times 3 \times 2 \times 4 \times 3 \times 6$ or 864 unique active variants in the curated deep sequencing dataset: [6: IP][7a: KLΔ][8a: AΔ][9: FIMY][11: FIL][13: FILPVW]. However, the observed number of active variants was an order of magnitude larger, which indicates the presence of many positive epistatic interactions. Thereby, viable sequence combinations of preferred hydrophobic and other non-preferred residues are only visible when embedded in a wider sequence context, i.e. dependent on non-additive effects beyond the hydrophobicity preference that dominates the single-residue fitness landscape (Figure 7.4C).

Having identified the presence of epistasis that underlies the widening of accessible sequence space, I further probed the curated dataset of active variants for the location of specific non-additive trends in amino acid preferences between different randomised positions. For some pairs of MKK1 positions, the preferred residues at one position are influenced by the amino acid present at the other (Figure 7.5). While most position pairs followed the trends typical for each randomised position (i.e. enrichment of hydrophobic residues), the two-dimensional fitness landscapes displays stripes of co-enrichment between a large hydrophobic residue (Ile/Leu) at one position and any other amino acid at the other. Further quantitative investigation of shifts in amino acid preference supports the interpretation that this system exhibits widespread epistasis and that the two-position enrichments are not just a reflection of individual position sequence preferences. This prevalence of positive epistasis indicates that the presence of an Ile/Leu residue allows anchoring to ERK2 in combination with otherwise deleterious residues, such as Ala or Asp. Interestingly, these co-enrichment patterns are also present for Leu in position 6 and all hydrophobic residues in position 13, which are positions with a weak individual sequence preference. This ability of the hydrophobic amino acid residues to expand the sequence space further supports the hypothesis that these residues are essential to D-domain binding and activation. Conversely, the preferred large hydrophobic residues show weak negative epistasis with each other, suggesting that a D-domain full of hydrophobic residues is not necessarily better.

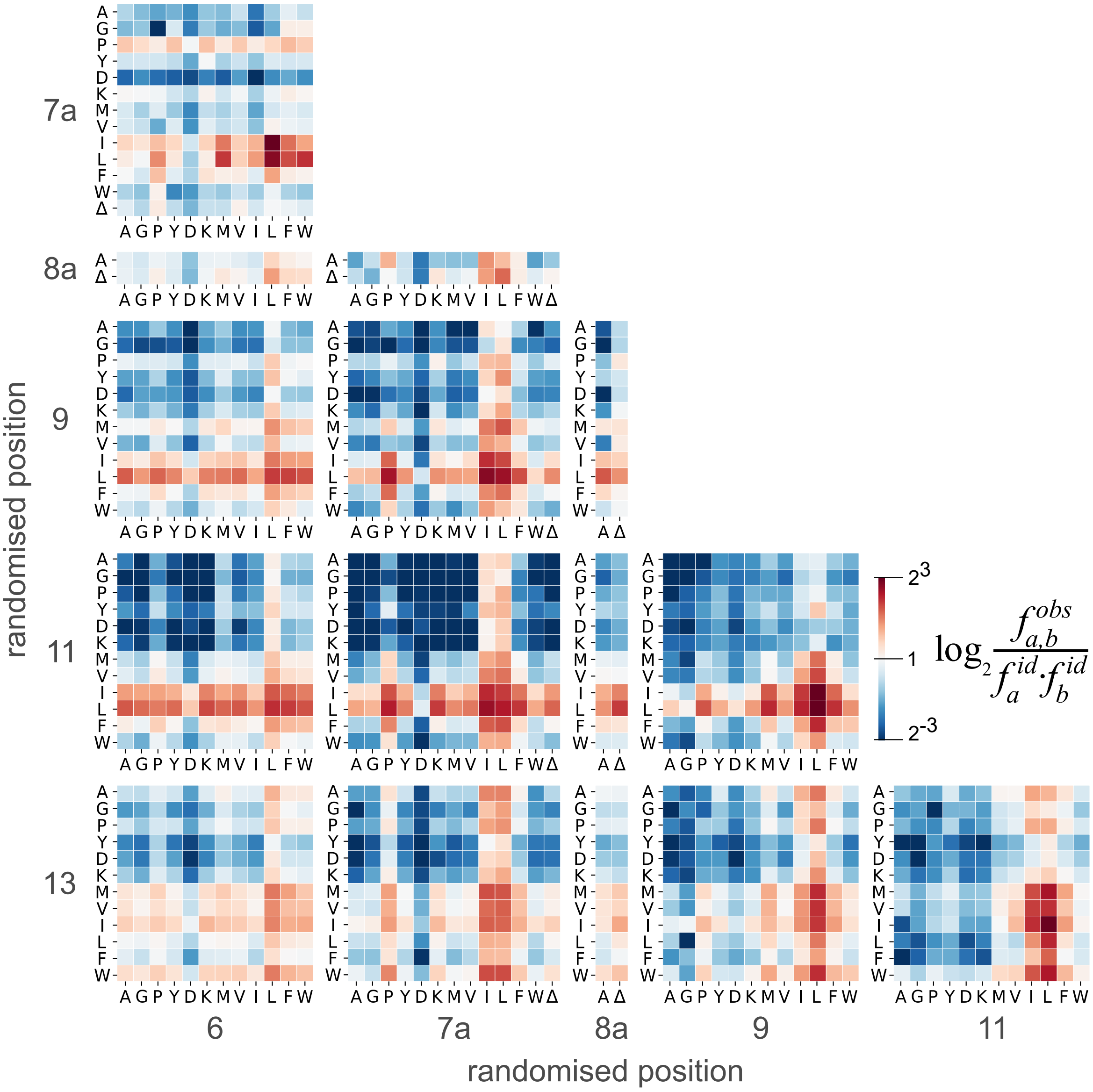


Figure 7.5 Hydrophobic residues in the centre of the D-domain epistatically expand the active fitness landscape. The heat maps show how the observed frequency of amino acid residues in two randomised positions ($f^{obs}_{a,b}$) compares to the expected frequency in an ideally balanced library ($f^{id}_a \cdot f^{id}_b$). A logarithmic scale is used to proportionately display both enrichment (red) and depletion (blue). Each panel represents one pair of randomised positions, such that the randomised position is labelled on the outside edge. The presence of large hydrophobic residues Leu/Ile at any one of positions 7a, 9 or 11 serves as an 'anchor', so that these residues appear in combination with non-preferred amino acids. The expansion of sequence space generates an appearance of red stripes and crosses in the two-dimensional fitness landscape.

7.2.3 Preferred motifs in the D-domain

Having shown that the amino acid preferences at the six randomised positions are interdependent and epistatically linked, the next step was a search for structural motifs within the fitness landscape that underpin this pattern. I first examined the change in preferences when one or more residues are restricted to preferred hydrophobic residues, comparing the change in amino acid distribution with the overall sequence preferences. If a Leu/Ile is fixed in any one position, there is a mild shift in the fitness landscape in the adjacent randomised position towards hydrophobic residues, creating an Φ-X-Φ motif (Φ denoting hydrophobic residues, especially Leu or Ile, and X the fixed spacer residue). Conversely, the sequence preferences at distant positions are relaxed, such that those positions become more tolerant of small or hydrophilic residues. This suggested that such an Φ-X-Φ motif could be the key functional element in the fitness landscape of the D-domain mediated MKK1-ERK2 interaction.

While a single large hydrophobic residue is not sufficient to anchor a D-domain, two adjacent such residues in an Φ-X-Φ motif greatly reduce constraints on the rest of the fitness landscape (Figure 7.6). Conversely, if two adjacent positions contain any residues but Leu/Ile, the remaining randomised positions are strongly enriched for these residues. The effect of the Φ-X-Φ motif is strongest in positions 7a/9, 9/11 and 11/13, where D-domains typically contain hydrophobic residues and the short Φ-X-Φ motif with two Leu/Ile residue appears sufficient. The activating effect of two hydrophobic residues is present regardless of the exact position of the motif, although the Φ-X-Φ motifs in positions 6/7a and 11/13 tend to be supplemented by a third hydrophobic residue nearby. On the other hand, the incorporation of two small or charged residues in positions 7a and 11 would prohibit the formation of any Φ-X-Φ motif, and can explain the strong selection pressure against these mutations co-occurring (Figure 7.6).

Thus, this DMS dataset shows that the D-domain fitness landscape is shaped by the requirement for a two-residue hydrophobic motif, although the location of this motif in the D-domain is flexible. Accommodating this shifting motif is thought to rely on the promiscuity of the ERK2 DRS, which naturally accommodates several unique D-domains with different orientations to maximise overlap with its hydrophobic pockets (Garai et al., 2012). The co-evolved amino acids which complete the motif are less constraint in chemical functionality, but are likely to optimise packing to the DRS – indicated by the strong negative epistasis of multiple small residues co-occurring.(Figure 7.5).

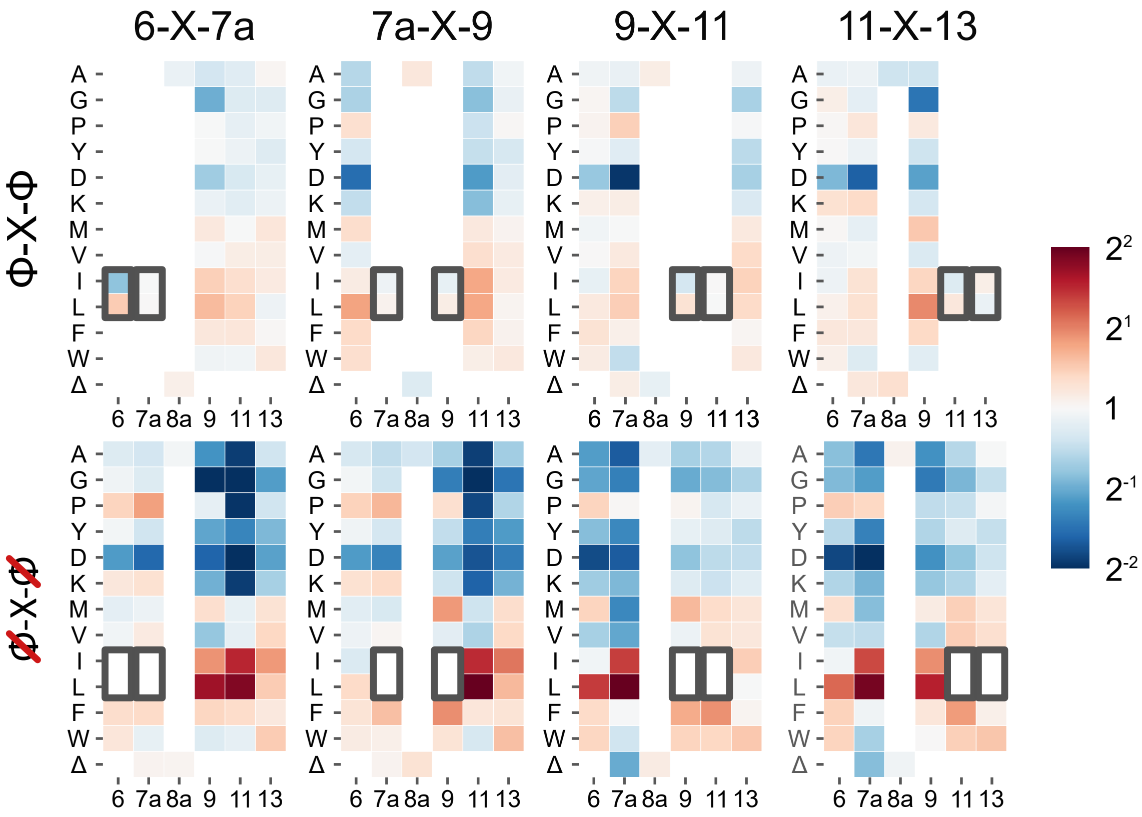


Figure 7.6 A Φ-X-Φ motif is sufficient to anchor the MKK1 D-domain to ERK2. The top row shows the enrichment fitness maps across D-domain, constructed from variants where two adjacent residues fit the Φ-X-Φ motif (where Φ=Leu/Ile). In those variants, the sequence preferences in other positions are strongly relaxed, indicated by pale colours across the heatmaps. Conversely, when Leu/Ile are excluded from a pair of positions (bottom row), the Φ-X-Φ motif re-appears at the other end of the D-domain (dark red squares).

7.2.4 A sequence similarity network identifies distinct communities in the fitness landscape

Having observed that active D-domain can accommodate the key hydrophobic Φ-X-Φ motif in more than one possible position, I was next interested in the clustering of active variants in the wider fitness landscape. I constructed a sequence similarity network (SSN) and visualised it with a force-directed layout (Figure 7.7). Each variant in the curated dataset is a node and edges connect all pairs of variants that differ by a single point mutation. The key question was whether the SSN contains any distinct regions or clusters within the network, which would indicate the possibility of distinct biological solutions to the problem of ERK2 activation. Therefore, the largest connected subgraph of the full SSN – spanning 97% of nodes in the full curated dataset – was partitioned using the Leiden algorithm to identify communities within the network (Traag et al., 2019). A community is a collection of nodes which are highly connected to each other but have relatively few links to nodes outside the community; in evolutionary terms, a community corresponds to a peak in the protein fitness landscape.

This SSN contains 11 distinct communities, which are further grouped into two similarly sized groups separated by the presence or absence of the 8aAla insertion (Figure 7.7). Thus, the 8aAla insertion appears to open a separate region of sequence space (Emond et al., 2020), likely through forcing a different conformation in the DRS binding pocket. Within each region of the fitness landscape, communities are distinguished by the nature and position of the defining single large hydrophobic residue. As expected from the anchoring effect of Leu/Ile residues described above, most communities are clustered around an anchor residue that is present in positions 7a, 9 or 11, yet the Φ-X-Φ motif is also observed in two communities at positions 6 and 7a (Figure 7.7, ochre and cyan sequence motifs). This ’walking’ of the anchor residue is consistent with our previous observation that the anchoring motif needs to be present, but is not confined to a single location.

In contrast with previously observed trends in protein-protein interaction fitness landscapes, this SSN is not scale-free; that is, there are no clear hub variants in this network. Instead, the D-domain SSN predominately features moderately well-connected variants, which indicate that the fitness landscape is fairly flat and smooth. Instead of a few hub variants, which would be expected in a highly rugged fitness landscape, this SSN suggests there are many similar solutions for binding and activation of human kinases. It amounts to a ‘rolling hills’ model with the possibility of multiple evolutionary pathways for interconversion of one specificity to another and no interference by canyons and mountains.

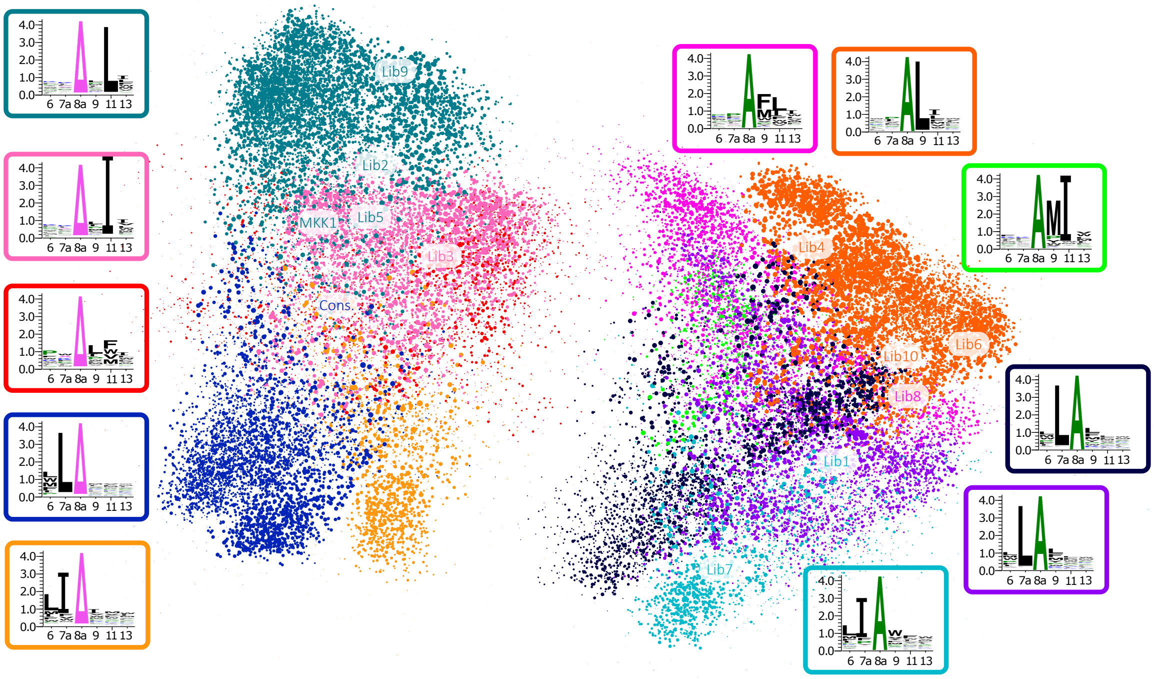


Figure 7.7 The key hydrophobic residues appears in different positions in the sequence similarity network (SSN). The SSN is constructed with each active variant as a node and an edge connecting every pair of variants that differ at exactly one position. Edges are hidden for clarity, because the large number of edges in the network overwhelms the visualisation of node grouping. Partitioning the SSN with the Leiden algorithm detected 11 distinct communities in this SSN, such that the variants are coloured according to community membership and node size is proportional to the number of edges around that node. The location of the library variants used in affinity screening is indicated. The 8aAla insertion causes a registry shift in which partitions the SSN in two equally-sized parts. Within each half, each community is distinguished by a sequence fingerprint (WebLogos in boxes) with different anchor hydrophobic residues. The largest communities (teal, pink, orange, green) show the Leu/Ile residues at positions 9 and 11, but communities with the key residues at the start of the D-domain are also present (yellow, cyan).

The most enriched D-domains are promiscuous binders. Finally, the NGS dataset was used to identify the most enriched D-domain for further characterisation of their ability to bind the ERK2 binding groove and activate the cascade. The 10 variants with the highest number of sequencing counts in the high gate were chosen for further analysis; they also stem from different regions of sequence space as shown in the SSN. Similar to the wild-type MKK1 D-domain, all 10 D-domains selected from the library out-competed a fluorescently labelled reporter peptide out from the binding groove. Interestingly, the affinity testing found a peptide sequence which bound ERK2 with an affinity higher than the wild-type MKK1 D-domain. In line with the low-level epistasis observed, this is not merely dependent on additivity of having multiple large hydrophobic residues: a consensus D-domain designed to contain additional leucines at positions 6 and 7a had a lower binding affinity for the ERK2 groove than the MKK1 D-domain or any other library selected D-domain.

While D-domains are believed to be crucial in regulating the specificity of interactions between kinases, synthetic D domains are known to be slightly promiscuous. They successfully bind grooves of non-cognate partners when isolated from the catalytic kinase domain; an effect most apparent between p38 and ERK2 targeting D-domains. Following this trend, we observed that the binding affinities of the ERK2 evolved D-domains are also promiscuous to bind p38, but rarely bind JNK. Interestingly, Lib1 which was the synthetic D-domain with the highest affinity for the ERK2 DRS, also has stronger affinity for the p38 DRS than cognate synthetic MKK4 and MKK6 D-domains and stronger affinity for the JNK DRS than the cognate MKK4 D-domain, making it the first known docking domain which binds all MAPK DRSs with affinities as good as their cognate synthetic D-domains.

7.3 Conclusions

In this project, I exploited the modularity of the InDelScanner computational approach in order to explore the sequence-function relationship in a new experimental system. The project was collaborative, split between Remkes Scheele performing all experimental work, and the data analysis on my part. The new *in vitro* assay for screening kinase cascade activity proved to enable ultrahigh throughput screening with high positive predictive value, enabling us to both identify interesting individual MKK1 variants and to build a fitness landscape of the docking domain mediated activation of ERK2.

A unique feature of this project was the use of the SpliMLiB library assembly, which independently randomises the MKK1 docking domain in six positions, two of which are optional insertions. The first goal for the NGS data analysis was to adapt the InDelScanner workflow for the complex SpliMLiB library design. Unlike the analysis for the TRIAD libraries (or any library design with random mutation placement), the MKK1 D-domain libraries concentrated a large number of mutations in a short stretch, which interfered with accurate identification of all possible mutation combinations. I solved this problem by first pulling the randomised DNA sequence out of the starting FASTA files and translating it into protein sequence before alignment. In combination with an alignment reference sequence with randomised (X) residues and a custom scoring matrix, this enabled accurate identification of any combination of mutations in a variant in this library.

The first goal for the data analysis was to identify the most active variants based on the NGS read distribution. There are some technical challenges in answering this question, since the activity assay primarily selected for wild-type-like activity and not for functional improvement: first, because most variants show an activity distribution in the FACS activity sorting, the variants' NGS reads are similarly spread across all three activity/sequencing bins. Second, quantitative interpretation of read counts at the level of individual variants is not straightforward, because the DNA for NGS was PCR-amplified without UMIs on the starting templates, creating PCR duplicates. Consequently, this experimental design was not ideal for finding the most active variant, and we simply chose the ten variants that were the most abundant in the high activity gate (judged by NGS read counts) for further characterisation. Interestingly, these ten variants spanned different regions of the fitness landscape and showed similar binding profiles to wild-type $^{\text{ca}}$MKK1, or in one case, a promiscuous generalist kinase profile.

Probing a multidimensional fitness landscape

While imperfect for identifying more active kinases, this experiment created a very rich dataset for understanding the fitness landscape of D-domain mediated ERK2 activation. It should be noted that technically, this experimental design is not deep mutational scanning, which compares the high activity variants to the input library, while we sequenced the three output libraries. In our sequence-function mapping experiment, sequencing of the low activity fraction rather than the input library was more informative on MKK1 activity at the same sequencing cost. Furthermore, the library design is known to generate very balanced libraries (Lindenburg et al., 2020) and the low activity gate contained the different residues in near-equal proportions, making the low gate sequences a good proxy for the starting library.

Since we sequenced all three activity gates, we could use not only the abundance of variants in the high activity gate but also the distribution across three bind to identify neutral MKK1 variants (i.e. variants with activity similar to $^{\text{ca}}$MKK1). This curated dataset of active variants comprised 32,375 variants, which is 20× more than used previously to map the fitness landscape of a bacterial kinase pair (Podgornaia and Laub, 2015). The single position fitness landscape of our system confirmed the strong effect of large hydrophobic residues in the D-domain, especially Leu or Ile at positions 9 and 11.

A strength of this dataset is that it probes the MKK1 fitness landscape in unprecedented six dimensions at high throughput. Therefore, we were able to interrogate this dataset to probe for both first order (interactions between two residues) and higher order epistatic effects that shape this fitness landscape. Consequently, the key features of the active D-domains are:

- a strong preference for Leu and Ile residues in positions 7a, 9 and 11; this was expected based on the prevalence of large hydrophobic residues in related kinases,
- first-order positive epistasis between Leu/Ile residues at one position and Leu/Ile in the next randomised position, which favours the formation of a Φ-X-Φ motif; this interaction would also have been discovered in a screen of all double mutants,
- second order epistasis, such that the presence of a Φ-X-Φ motif show positive epistasis towards amino acid residues in the rest of the D-domain which are otherwise not preferred, relaxing the fitness landscape constraints.

In this way, this multidimensional sequence-function mapping experiment showed that the Φ-X-Φ motif is not only found in the MKK1-like positions 9 and 11, but that it can ’walk’ across the docking domain while retaining function. Indeed, the sequence similarity network of the D-domain is highly connected without clear hubs, only divided in two regions by the presence or absence of the 8aAla insertion. Consequently, we can imagine that the docking domain sequence can wander across this fitness landscape in a number of directions, exploring new functionality while retaining the original MKK1-ERK2 interaction.

Chapter 8

Conclusions and Discussion

8.1 A deep mutational scanning workflow for InDel libraries

In this thesis, I present multiple projects around a central theme: exploring the sequence-function relationship in enzymes through deep mutational scanning experiments, focusing on the role of small InDels in protein function.

Since the first proof of concept for deep mutational scanning in 2010 (Fowler et al., 2010), this approach to sequence-function mapping became an established method for interrogating sequence determinants, which has been applied in a range of fields from antibody affinity maturation to structure determination and to understanding the evolution of clinically-relevant antibiotic resistance (Chen et al., 2020; Schmiedel and Lehner, 2019; Tabasinezhad et al., 2019). However, current applications of DMS are largely limited to investigation of all possible amino acid substitutions in a protein, typically through single-site saturation libraries or by analysing a subset of variants in an error-prone PCR library. As a result, the present body of knowledge is very limited with regards to high throughput studies on the effect of InDels, despite the fact that small InDel are common in natural evolution (Chothia et al., 2003).

In my PhD work I set out to adapt the standard DMS protocols for exploration of InDel libraries, specifically single position "InDel saturation" libraries. In the scope of mutagenesis, the TRIAD libraries are comparable to site saturation libraries in that the introduce every possible mutation of a given length (deletions and all NNN^x insertions of a given length, at all target gene positions). In parallel with the TRIAD libraries I worked with TriNEx substitution random site saturation libraries, which enable comparison of substitution with InDel tolerance in different parts of the protein.

The success or failure of a deep mutational scanning project hinges on two main elements:

- the quality of the fitness selection step: the functional selection must maintain the genotype-phenotype linkage, achieve ultrahigh throughput to cover the sequence space in a diverse starting library, be practically tractable and reproducible, with low experimental noise.
- the success of a deep sequencing strategy: while the cost of NGS is continuously falling, DMS experiments still need to be designed to achieve good statistical power while fitting into a reasonable number of sequencing runs. Therefore, strategies that reduce error rates and improve the amount of information gained per experiment, such as the use of UMIs, should be used.

8.2 Working with a fast enzyme requires a challenging microfludic assay

The most ambitious single project within my PhD plans, DMS on phosphotriesterase TRIAD libraries with a microfluidic droplet assay (Chapter 6), was challenging on both aspects and therefore a good long-term project to work on over multiple years. At the very beginning of the work, I investigated the feasibility of a microfluidic droplet assay for PTE - at the time, only colony screening or microtitre-plate assays were available for that very fast enzyme. More generally, as an enzyme PTE was an interesting target for testing the effect of InDels in high throughput than a model system that does not have a catalytic step.

I started with the assay development required to achieve sufficient throughput, using the monodisperse microfluidic emulsion format. I also briefly considered using gel-shell beads (Fischlechner et al., 2014), which can be sorted on a FACS instrument, and would not require new chip development. However, the multiple encapsulations used in that protocol are not compatible with fast enzymatic reactions, and so I did not pursue gel-shell beads further.

Having tested the cell lysate reaction rates in diluted bulk cultures, it was clear that PTE is too fast for the standard modular microfluidics workflow. Therefore the assay had to take the form of an integrated in-line droplet generation, incubation and sorting device. On the one hand, an integrated device is faster to operate because all steps are done during one experiment in a day; the modular workflow takes longer because droplet generation and sorting are generally done on different days. On the other hand, that advantage fades in comparison with the increased complexity of integrated chips, and their higher sensitivity to changes in experimental conditions. The reason for this is that a good delay line design, and good droplet generation and droplet sorting designs, require opposing properties:

- Flow focusing chips for droplet generation have narrow shallow channels (<25 μmu), which control the size of droplets - the droplets will generally be equal in size or larger to the cross-section of the flow focusing chip (larger because droplets can squeeze during flow focusing). The oil flow rate must be high (about 5-10× that of the aqueous phase) to fully separate the droplets from each other.
- Sorting devices similarly feature shallow, narrow channels with high flow rates. Here, this ensures proper droplet separation during signal detection, so droplets pass the laser focus one by one.
- In-line incubation lines, in contrast, feature wide and deep channels, in which the droplets are closely packed and flow through the line slowly. The large channel size reduces the pressure in the chip, which is proportional to $\frac{length}{width \times depth^3}$ and limited to ~3 bar before the chip de-laminates (Frenz et al., 2009).

An in-line integrated device unavoidably has relatively long channel length, because of the need to connect all three modules. There is some scope for varying the channel width, although the maximum width is limited by constrictions in the flow focusing and sorting elements, as well as the need for good droplet mixing, so they maintain an uniform incubation time. Since the pressure is exceptionally sensitive to channel depth, a two-depth device is the most effective way to reduce the pressure in the devices, and was the preferred way of solving this challenge from the start.

Fabrication of two-depth microfluidic chip device masters is technically more challenging, because it requires the use of two or more photomasks which must be aligned with each other with very high precision. While two-depth fabrication was not yet an option, I explored multiple single-depth designs with different incubation line length, width, depth and shape.

I observed that the designs with shallow and/or narrow, long incubation lines were very difficult to operate because of high back-pressure in the system. This caused difficulty with droplet generation, where the emulsion would typically 'escape' through one of the inlets rather than enter the high-pressure incubation line, or the chip would de-laminate after only minutes of operation. I soon switched to a deeper design (50 μm) featuring a wide bubble-shaped incubation line, which alleviated these problems, so I was able to record histograms of PTE activity after incubation time of 1.5 minutes or less, depending on the rate of oil extraction (Figure 6.3). The signal was surprisingly strong, given the very short incubation time and the high enzyme dilution in the larger 270 pL droplets. These results were promising indicators for the ability to detect PTE variants with medium and high activity during library sorting.

As the deep incubation line designs could not be combined with shallow flow focusing and sorting devices at the time, I suspended work on this project. However, the chip development was later re-visited by David Schnettler and Dr. Tomasz Kaminski. They developed a two-depth design that integrates all the features I hoped to combine at the start of this project: multiple depths, a bubble-shaped incubation line with constrictions used for fluorescence measuring, an efficient oil extraction system and a sorter. I later tested this device and observed stable droplet formation and a small fraction of positive droplets. Compared to my 50 μm design, this device creates much smaller droplets, which might account for the unexpectedly low proportion of positive droplets. Alternatively, not all cells might be lysed by the time they reach the detection point, again decreasing signal intensity.

Between the recent two-depth design and my previous work with deeper incubation lines, here I have shown that the detection of PTE activity in microdroplets is in principle feasible with chip designs and knowledge in the Hollfelder research group, and the project created the foundation for deep mutational scanning of fast enzymes.

Since the improved two-depth integrated microfluidic chip is now available, it would be relatively straightforward to revisit this project. Some interesting hits already emerged in low-throughput colony screening, yet that screening campaign only scratched the surface of the large diversity of the TRIAD insertion libraries. As the starting PTE TRIAD libraries have already been sequenced (Chapter 4, Emond et al. (2020)), completion of DMS on this enzyme would require only minor fine-tuning of experimental conditions and sequencing of the sorted libraries, which is outlined at the end of Chapter 6.

8.3 Setting up a high resolution DMS system with GFP

Once the development of a microfluidic device for assaying PTE proved intractable, I shifted the focus of my work to implementing DMS in an easier experimental system, specifically with GFP. I sorted the TriNEx and three deletions libraries with single-colour cell sorting on FACS and sequenced the resulting DNA fractions with a robust sequencing strategy. In this way, I obtained a pilot dataset that realistically represented the type of data I could expect out of a comprehensive deep mutational scan with InDel-containing libraries (Chapter 2). I primarily used this pilot dataset to develop the core InDelScanner analysis scripts (Chapter 4) and probe the general patters of InDel tolerance in GFP.

This dataset had both advantages and disadvantages:

- The experimental work was fast to perform and technically straightforward, at the cost of a limited resolution for intermediate fluorescence variants.

- The reads are short (2×75 bp paired end) and located in random locations in the GFP gene, which is a result of the Nextera sequencing library preparation. This library preparation technique was technically simple, since it avoided issues with sequencing a repetitive amplicon, but it makes placing mutations harder.
- The ends of the GFP gene, especially the N-terminal region, are poorly covered, so some known adaptive mutations (such as Gly4Δ) are missing from the dataset.
- Despite repeated sorting of the middle-fluorescence fractions in all libraries, some overlap is still present in the post-sorting fluorescence histograms. Consequently, many variants appear in more than one sequencing bin, some at substantial levels. This spillover obscures assignment of variant fitness.
- The sequencing strategy did not utilise UMIs, leading to a large number of sequencing substitution errors in the dataset. Removal of these requires additional arbitrary filtering steps and makes it difficult to assess TRIAD library quality through NGS alone. Furthermore, without UMIs the differences in read counts between variant bins can be masked by differences in (limited) PCR amplification in sequencing library preparation or by differences in bacterial growth rates after sorting.

Despite the limited FACS resolution in the pilot dataset, the results were sufficient to draw some preliminary conclusions regarding InDel tolerance in GFP. Some trends were as expected; short in-frame deletions are better tolerated at the protein C-terminus, in the flexible loops and of the subset in β-strands where deletions are accepted, the mutations were located at strand termini. A comparison of assigned fitness scores compared to those previously described confirms that the scoring of variants in my dataset agrees with those tested in a cell lysate assay by others (Arpino et al., 2014a).

In Chapter 2 I showed that a simple linear model is sufficient to assign the fitness of individual variants from NGS data. While that was shown with the subset of GFP deletion and substitution variants that were Sanger sequenced and present in high abundance in the pilot NGS dataset, the same approach will be implemented with the full NGS dataset with both deletions and insertion when available. Beyond simply listing fluorescent variants, some further analysis strategies will be possible on the complete DMS dataset:

- applying the linear model (Figure 2.13) to calculate the scores of a larger number of variants, which will generate a set of scores that describe the effect of pure deletions, pure insertions, pure substitutions and mixed variants,
- describe the single-residue fitness landscape for all these classes of variant in the form of heatmaps across the protein,

- compare the tolerance of this larger set of mutations between eGFP and GFP8, especially examining the positions where the two constructs differ from each other,
- create a sequence-similarity network to organise the numerous characterised variants.

Most of these would already be possible on the deletions pilot dataset, however the insertions are less investigated and present a more interesting problem that has not been previously described in GFP. Therefore, while additional valuable information is present in this deletion dataset, ultimately confident interpretation requires improvement of the experimental design so as to collect higher resolution data. This dataset was acquired with a shot-gun sequencing approach, which is reliable in generating *some* data but is too inefficient to scale well to larger insertion libraries ($> 100\times$ more diverse than deletions and substitutions). Instead of investing more time into the pilot dataset, I focused efforts on the improved experimental design that allows for screening insertions (Chapter 3).

8.3.1 GFP deep mutational scanning in an optimised experimental system

Having learned from the pilot experiment, I designed an improved experimental protocol for sequence-function mapping the effect of InDels in GFP (Figure 3.4). There were alterations both in the functional assay, through the use of two-colour activity sorting with mKate2 as an expression control, and in the approach to sequencing the library with the use of UMIs.

I adapted the protein fusion construct with mKate2 from Sarkisyan et al. (2016) and optimised the expression conditions for GFP in the new system. Because of the new expression construct, I was able to sort the deletions libraries into four activity fraction during a single sort. Once I started sorting the TRIAD libraries, I observed that the proportion of fluorescent insertion variants was orders of magnitude lower than for deletion libraries, such that the libraries were apparently non-fluorescent. Consequently, I adapted the screening protocol to use an enrichment sort for non-negative variants before proceeding to a four-way activity sort.

Once the expression and sorting conditions were set up, I was able to sort all TRIAD and TriNEx libraries in both eGFP and GFP8 within weeks. One significant modification from the pilot sorting experiment was that after sorting, here I re-grew the cells on antibiotic-supplemented LB agar plates rather than in liquid culture. This has two advantages, one practical and one technical. Plate re-growth avoids the need for laborious DNA isolation and re-transformation between sorting and analysis, or between two sorts, as long as subsequent steps are performed within about a week (while the colonies on a plate are still viable and

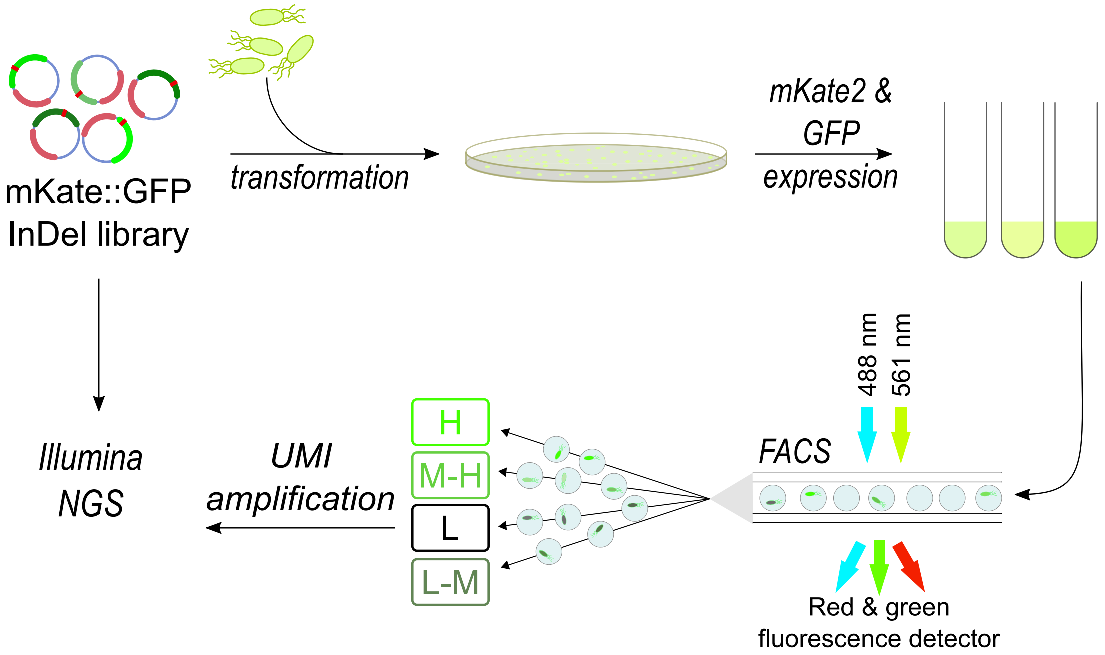


Figure 3.4 An outline of the improved DMS experiment with mKate2::GFP libraries, with mKate2 shown in red and GFP variants in green. The starting InDel libraries, located in the mKate2::GFP::UMI construct, are separately transformed into the appropriate highly competent *E. coli* strain and grown overnight on large agar plates. The plates are scraped before mKate2 and GFP expression in liquid culture, then the individual cells are sorted with based fluorescence excited by the 488 nm (GFP) and 561 nm (mKate2) lasers. The plasmid DNA in each fraction is recovered, the UMIs amplified from the sorted fractions and sequenced.

non-contaminated). Additionally, it is possible to quantify the approximate fraction diversity from re-growth and repeat the sort if only a small number of colonies grew - a liquid culture does not allow that check.

In addition to a substantially improved activity sorting process, this DMS workflow is designed around a UMI-guided sequencing approach (Figure 8.1). While even the relatively short GFP gene is too long to be captured within a single Illumina MiSeq paired-end read, by incorporating a diverse UMI in the library plasmid backbone, the UMIs can be used to computationally assemble sequencing reads from both the start and the end of the GFP gene. This data can then be used to set up a dictionary that links UMI sequences with variant information (i.e. which mutation is present and where) for all starting libraries, so that only UMIs need to be sequenced after the functional screening.

In addition to linking distant mutations in the same variant, this approach opens up the possibility of performing the functional screening in multiple conditions (e.g. protein expression temperatures, tolerance of a denaturing agent, buffer pH) and with technical replicates, with only a moderate increase in sequencing cost. Finally, the UMIs can facilitate recovery of specific gene sequences for further biophysical characterisation with dial-out PCR (Schwartz et al., 2012).

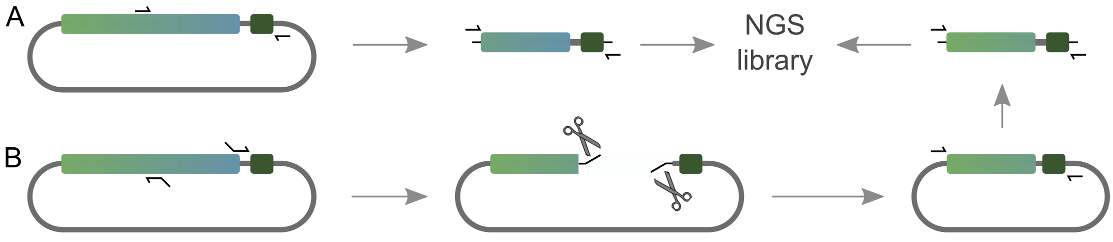


Figure 8.1 The UMI-guided sequencing strategy for GFP TRIAD starting libraries. The GFP gene is recovered from the starting plasmid in two halves, one directly and the other using a second PCR step. A) The part of the gene that is proximal with the UMI is amplified directly using one primer in the middle of the gene and the second primer on the Illumina technical sequence that follows the UMI. B) In order physically link the 5’ half of the gene with the 3’-located UMI, an inverse PCR step is used that amplifies the 5’ part of the gene, the plasmid backbone and the UMI. Primers have overhangs that enable a Type IIS restriction enzyme digestion and a seamless ligation of the PCR product. Finally, the now directly linked 5’ gene sequence and UMI are amplified as in part A). The resulting UMI-tagged gene fragments are subjected to further limited PCR reactions for indexing and NGS library preparation.

In conclusion, while the deep mutational scanning of TRIAD libraries on GFP is still a work in progress, the major technical milestones have been achieved. The experimental side of the project is nearly complete, the sequencing is imminent and the data analysis tools are ready for the NGS data. The substance of this project is the analysis of insertions variants, which are less well tolerated in GFP than substitutions or deletions, while the latter will form a good reference point for comparing different types of mutations.

8.4 Scripts for analysing InDels in amplicon sequencing libraries

The core InDelScanner scripts are concerned with detection and classification of DNA mutations in amplicon sequencing data, whether the variants are substitutions, insertions, deletions or a mix thereof. The core script design not make any assumptions about library composition - that is, it detects both desired mutations, sequencing errors, chimera variants, frameshifted variants and whatever else may be present in the dataset. In contrast, the data analysis scripts (typically in the form of Jupyter notebooks) are created separately for each project (GFP, PTE, MKK1) in a modular fashion, so that the analysis is customised to the library design and the research questions.

There is a trade-off in the script design between completeness of information collected in the core stage (during scanning of the alignment files) and the ease of later data interpretation. It would in principle be possible to already filter variants in the core scripts, discarding unwanted variants, and I attempted that approach early in script development. The advantage of that approach is that later data interpretation is easier, because the data is cleaner and variant counting is not complicated by very numerous but rare sequencing errors. However, there is a significant disadvantage: if the research question or library design changes, the scripts need to be substantially changed, which requires both coding time and processing time to re-run all analyses. Because of that issue, I soon switched to a modular design (Figure 4.1).

TRIAD library diversity in *wt*PTE. After developing the first iteration of InDelScanner scriptes, I used the scripts to analyse the diversity of TRIAD libraries prepared in *wt*PTE and contribute to the quality assessment of the TRIAD method. Previous to the NGS work, Sanger sequencing of randomly selected individual variants had been done, which suggested that the libraries were diverse, but could not access all gene positions with a limited number of sequences. The results provided by Sanger sequencing and high-throughput Illumina sequencing are complementary: while only long-read sequencing can confirm the co-occurrence (or absence thereof) of distant mutations in the same variant, the higher throughput of NGS sequencing provides a more complete view of the library composition. The NGS dataset detects both frequent and rare variants and quantifies their relative abundance.

The analysis of the *wt*PTE Illlumina dataset showed that the TRIAD protocol gives access to variants with mutations spread across the entire gene, accessing >85% positions in the deletions libraries and >95% position in the insertion libraries. Furthermore, including the rare variants in the analysis suggests that the sequence preference of the engineered Mu transposon is less biased than previously described (Haapa-Paananen et al., 2002).

8.5 Deep mutational scanning beyond single site saturation libraries

The modularity of InDelScanner scripts allowed a relatively straightforward adaptation of the same core computational process from the TRIAD libraries to sequencing a deep mutational scanning experiment on the docking domain of MKK1 kinase. In this project, the combination of a new powerful *in vitro* assay, the combinatorial SpliMLiB library design and adaptation of the InDelScanner scripts allowed us to push sequence-function mapping beyond the current standard.

Epistasis in deep mutational scanning. Intragenic epistasis, or non-additivity of mutation effects between mutations between different residues in a protein, has long been recognised as an important factor that shapes protein evolution and creates 'ruggedness' in fitness landscapes (Starr and Thornton, 2016; Weinreich et al., 2013).

Previous sequence-fitness mapping studies have largely scanned the effect of single site mutations across full genes (see Gray et al. (2017) and references therein, Chen et al. (2020); Wrenbeck et al. (2017)). The focus on mapping single site mutations either stems from the desire to map mutations across longer and longer DNA regions, up to full genes (~1,500 bp). For a typical gene length (330 amino acids), a library that randomises every single codon ranges in diversity from 7.2×10^3 to 2.1×10^4 variants depending on library design. This limited number of variants is within easy range for complete screening and sequencing. However, the diversity of combinatorial libraries increases exponentially with the number of residues that are diversified in any one variant: for a NNK design, a two-residue NNK combinatorial library contains 3.4×10^5 variant and a three- residue library 1.1×10^7 variants. In addition to prohibitive diversity in complete libraries, analysis of multiple mutations may be excluded even in datasets that contain multiple mutation variants, such as from error-prone PCR library preparation (Romero et al., 2015).

Datasets that do analyses epistastic effects in high-throughput have focused on pairwise interactions only, thereby probing first-order epistasis (Bank et al., 2015; Diss and Lehner, 2018; Olson et al., 2014). Such work is more tractable experimentally as it uses libraries that randomise only two residues, with moderate final diversity and easily created by shuffling a one residue saturation library. The datasets arising from these studies have also been used for determination of structural interactions (Rollins et al., 2019; Schmiedel and Lehner, 2019). Interestingly, they observed that epistatically-linked residues that do not form a structural link are often part of functional sites, highlighting the importance of understanding residue interactions in evolution of protein function.

More recently, advances in sequencing strategies have enabled identification of interacting mutations that are distant in DNA sequence (Yoo et al., 2020), which identified the best variants of a G-protein coupled receptor with multiple distant mutations. They showed that the associated sequence-similarity network for the best variants was poorly connected, such that the best variants were only accessible through a small number of paths due to widespread deleterious epistasis.

Higher order epistasis (i.e. interactions between more than two residues) has been recognised as important in adaptive evolution and key to determining evolutionary trajectories (Kaltenbach et al., 2015; Starr and Thornton, 2016). However, experimental investigations have focused on small recombination libraries (Weinreich et al., 2006, 2013; Yang et al., 2019), which reconstruct all possible intermediate variants that link two functional proteins, typically between the starting point and the end point in directed evolution. Each amino acid position is varied between the two residues that are present in the two endpoint proteins, so the total number of variants is small, typically <100. Since these studies closely examine epistatic interactions between a functional and an improved variants (where the average mutation effect is beneficial), it is not surprising that they tend to find numerous negative, restrictive interactions.

Far fewer groups have examined higher order epistasis exhaustively with high throughput experiments (Podgornaia and Laub, 2015; Wu et al., 2016) by simultaneously randomising four chosen residues. In both cases, analysis of the resulting 160,000 variants revealed abundant epistasis but also significant degeneracy in protein pathways, such that there is more than one option for a walk in fitness space between two functional variants. However, the widespread negative and sign epistasis also shape the fitness landscape into multiple fitness peaks, where transitions between them are restricted by epistatic ratchets.

Positive epistasis in the MKK1 docking domain. In contrast with widespread negative epistasis in adaptive evolution, our examination of the *neutral* fitness landscape of the MKK1 docking domain shows that this landscape is shaped by abundant positive epistasis. Consequently, the accessible sequence space that still maintains ERK2 activation is expanded, as long a the sequence contains a Φ-X-Φ motif with two Leu/Ile residues somewhere in the docking domain.

In the MKK1 docking domain, widespread positive epistasis generates evolutionary contingency, avoiding loss of function and enabling smooth transitions between residue combinations. Consequently, a type of fitness landscape emerges in which the functional variants are highly interconnected but do not contain any clear hubs. Such a 'rolling hills' fitness landscape is in stark contrast to the ruggedness observed elsewhere (Kondrashov and Kondrashov, 2015; Romero and Arnold, 2009) and is credited to positive epistasis reshaping the functionally accessible sequence space. Here, the Φ-X-Φ motif may promote kinase adaptability through positive epistasis. These features may help explain how diverse signalling cascades emerged in evolution. In addition to the versatility that the recognised modularity of kinases provides (Gordley et al., 2016), the results of this project suggest that epistatic effects contribute to robust intrinsic evolvability of protein function.

The experiments described in this work mimic a neutral drift scenario in evolution (a non-adaptive exploration of sequence space) that plays a prelude to future adaptive evolution by creating a multitude of starting points for functional innovation (Bloom et al., 2007; Wroe et al., 2007). When the trade-off between mutational burden and retention of function in this neutral phase is favourable, successful adaptive evolution becomes more likely (Aakre et al., 2015). The notion of smooth fitness landscapes arising from this project signals well for evolution of such networks and characterises important residues and their interdependencies.

8.6 Outlook

This thesis spans two experimental projects, deep mutational scanning on GFP and on PTE, and the computational project, that is the development and application of InDelScanner scripts to DMS datasets. Elements of these different projects overlapped: the InDelScanner scripts were originally developed on the pilot GFP dataset, applied to the TRIAD libraries in PTE and extended for the investigation of epistasis in the MKK1 kinase cascade experiment. Reflecting the limited timespan that was available for this work, the projects reach different stages of completion:

- The improved DMS experiment on the two GFP variants, eGFP and GFP8, using both insertion and deletion libraries, is a work in progress that is nearing completion. The experimental steps have been optimised, so the first iteration of sorted libraries will soon be sequenced and analysed. Beyond the single functional analysis, the experiments may be extended to functional screening at multiple conditions to obtain a more detailed snapshot of this fitness landscape.
- The InDelScanner scripts are mature enough to serve the original purpose, that is analysis of the composition and patterns in InDel-containing libraries, especially for deep mutational scanning experiments. The core scripts are functional, if a little slow, and may be improved in the future by switching to a different alignment algorithm during pre-processing and streamlining the core code for variant detection. On top of the core scripts, the InDelScanner design is modular and generally requires adaptation for any particular project, since the research questions are specific to each experiment. The next opportunity for improving my computational tool-set will open with analysis of the complete GFP TRIAD libraries, where I intend to integrate some approaches from the MKK1 D-domain and explore different correlations between related variants in the many libraries.
- Mapping the effect of InDels in high throughput on PTE was the more ambitious project in this PhD, due to the need for challenging microfluidic assay development. While I suspended work on this project myself, it could be restarted without much difficulty with the now-available two-depth integrated chip devices, and an outline of the possible project continuation is included in Chapter 6.

Finally, while the evidence on InDel tolerance suggests that short InDels provide at least different and perhaps advantageous starting points for directed evolution, this hypothesis has not yet been tested and investigating it would present a good extension to the work presented in this thesis. One possible experimental design would be through following the trajectories that directed evolution takes after first randomising the starting point in different directions: substitutions or InDels with comparable library diversity. After the initial diversification, the libraries could be screened for multiple round while introducing the same kind of mutations in every round, to maintain comparability, then selection a pool of improved variants in each variant. Such a limited pool of adapted variants (~1,000 to 10,000 variants per round) can be sequenced with third-generation sequencing technology such as Oxford Nanopore.

In this way, an approach that traces evolutionary trajectories in wider sequence space could complement well with insights from the immediate fitness landscape, generated by deep mutational scanning of the starting enzyme.

Chapter 9

Materials and Methods

Materials. The plasmids encoding eGFP, GFP8, PTE variants, ACP and empty plasmids (pID-Tet and pET backbones) were originally obtained from Stéphane Emond and further modified as described below. Restriction endonuclease enzymes were from ThermoFisher FastDigest line, or the regular ThermoFisher formulation when a FastDigest version is not available. The pGro7 plasmid for GroEL/ES over-expression was obtained from Takara Bio. Oligonucleotides were from Sigma Aldrich and Integrated DNA Technologies (IDT). Primer sequences are listed in Table 9.1.

9.1 Molecular biology

9.1.1 General cloning procedures

Competent cells. The competent cells were purchased from Lucigen and the strains were:

- E. cloni 10G ELITE Electrocompetent Cells ($2 \times 10^{10} cfu/\mu$g), referred to as 10G cells, used for all cloning and some protein expression. Genotype: F- mcrA Δ(mrr-hsdRMS-mcrBC) endA1 recA1 Φ80dlacZΔM15 ΔlacX74 araD139 Δ(ara,leu)7697galU galK rpsL nupG λ- tonA (StrR)
- E. cloni EXPRESS BL21(DE3) Electrocompetent Cells (($5 \times 10^{9} cfu/\mu$g), referred to as BL21(DE3) cells, used for protein expression. Genotype: F– ompT hsdSB (rB-mB-) gal dcm lon λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])

Both commercial strains were expanded to generate lab-made electrocompetent cells following standard protocols. These lower-efficiency cell stocks were used for routine cloning experiments and for smaller libraries, where the required transformation efficiency was <1,000 colonies (though the transformations usually yielded more).

Cloning. The restriction digest reactions were assembled with 1-2 μg plasmid DNA, 1.0 μL of each of two FastDigest restriction enzymes used in a double digest (commonly NotI, NcoI, MlyI, AcuI, HindII, PstI or AatII), 2 μL FastDigest Green Buffer (ThermoFisher) and milliQ water topped up to 20 μL. The reaction was incubated without shaking at 37°C for 1-2 hours, stopped by heat inactivation and separated by agarose gel electrophoresis, typically 0.8% agarose in TAE buffer, by running for 30-45 minutes at 100 V - 120 V. If needed, the desired bands were cut out from the gel with a razor blade, weighed (typically 80-150 μg) and the DNA extracted using the Gel DNA Recovery Kit (Zymo Research). If the band slice was large, the number of washes was increased to improve removal of the agarose and the agarose dissolving buffer. The concentration of DNA was measured using NanoDrop. The vector and insert fragments were mixed in a ligation reaction with 50-100 ng vector DNA (preferably 100 ng, but less if a small amount was available) in a 5:1 insert:vector molar ratio, in 20 μL reaction volume with 1 μL T4 DNA ligase and the appropriate buffer. The ligation reaction was incubated overnight at 16°C for high efficiency (library cloning) or for 1 hour at room temperature (routine cloning), then column purified using DNA Clean & Concentrator Kit by Zymo Research and eluted in 6 μL DNA elution buffer. The DNA concentration of the purified ligation reaction was generally below the useful detection level of the NanoDrop and was typically not measured, to avoid wasting an aliquot of DNA.

Transformation. The column-purified plasmid (~1 μL, 10-100 ng) or the purified ligation product (2 μL) were mixed with 25 μL of freshly thawed electrocompetent *E. coli* cells in a pre-chilled electroporation cuvette (1 mm gap). The mixture was incubated on ice for >15 minutes in the cuvette and then pulsed with 1800V, aiming for a time constant above 5.0 s. The cells were immediately mixed with 350 μL room-temperature or pre-warmed SOC buffer. If using commercial high-efficiency cells, the SOC medium was replaced with the included recovery buffer. After 30-60 minutes of recovery at 37°C with shaking, the cells were plated on LB-agar plates containing the appropriate antibiotic. For library transformations, 300 μL of the transformation culture was plated on a large LB-agar plate and a fraction of the transformation culture was plated on small plates at 1:10, 1:100 and/or 1:1000 volume dilution to estimate transformation efficiency and use it to estimate library diversity.

GroEL/ES chaperones. The cells co-expressing GroEL/ES chaperones from the pGro7 plasmid and the variant of interest were prepared in one of two ways. One, high efficiency electrocompetent BL21(DE3) cells were co-transformed with pGro7 and the variant plasmid by mixing equal molar amounts of each plasmid with the cells and plated on agar plates containing two antibiotics. Two, BL21(DE3) cells were first transformed with pGro7 alone and an overnight culture grown from a single colony. This culture was used to inoculate a 500 mL large growth culture, where cells were grown to OD_{600nm}, then pelleted and washed to give a stock of electrocompetent cells according to standard protocols. Then, the plasmid with the variant of interest was transformed into an aliquot of these electrocompetent BL21(DE3)-pGro7 cells following the above procedure.

DNA isolation. To isolate library DNA, the bacterial lawn on a large LB-agar plate was gently scraped with 5 mL LB until it formed a suspension with uniform consistency, the suspension was pelleted by centrifugation at 14,100 rcf for 5 minutes and the supernatant discarded. To isolate the DNA of a pure construct, a single colony was picked from an agar plate and grown in LB medium (5-10 mL) supplemented with the appropriate antibiotic overnight. If needed, an aliquot of the overnight culture was used to prepare a glycerol stock (12.5% final glycerol concentration), and the rest of the culture was pelleted at 4000 rcf for 15 minutes and the supernatant discarded. In both cases, the cell pellet was frozen at -80°C for >1 hour to promote cell lysis, then thawed and the DNA isolated with the GeneJET Plasmid Miniprep Kit (ThermoFisher), finally eluting in 50 μL elution buffer. The concentration of DNA was measured with NanoDrop and the DNA stored at -20°C.

Sanger sequencing. Individual constructs were Sanger sequenced by the Department of Biochemistry Sequencing Facility or by Source Bioscience, and the sequences analysed against plasmid maps with Benchling (https://www.benchling.com/) or in combination with InDelScanner scripts. When larger numbers of variants needed to be sequenced, they were grown in deep-well microtitre plates as overnight cultures in LB medium with ampicillin or carbenicillin, then stored as glycerol stocks in a second microtitre plate. The glycerol stocks were used to inoculate a plaque on LB-agar (200 μL / well, dispensed with an electronic multichannel pipette using 1 mL tips to prevent bubbles in the agar). The agar wells were incubated overnight at 37°C, then shipped to LGC Genomics for sequencing using the MTP Premium service, which includes isolation of DNA from cultures. GFP constructs were sequenced using the T7t primer. PTE variants in pET vector were sequenced by LGC Genomics with T7prom and T7term (= T7t) primers.

Preparation of DNA for Illumina sequencing. For deep sequencing, amplicon libraries were digested from pID-Tet with appropriate FastDigest restriction enzymes: Bpu1102I and

Van91I to give a pool of 1.3 kb linear fragments containing PTE, and the GFP libraries with PstI and NdeI, then purified with gel electrophoresis. The bands were cut out from the gels, purified and the amount of DNA quantified with NanoDrop and Qubit dsDNA HS Assay. The pure DNA was submitted to the Department of Biochemistry Sequencing Facility, which processed them using Nextera DNA Library Preparation Kit according to manufacturer's instructions and sequenced them on Illumina MiSeq using 2x75 bp paired-end sequencing.

9.1.2 Construction of pID mKate2::GFP::UMI and intermediate constructs

General PCR conditions. Unless stated otherwise, the PCR reactions were performed with Phusion polymerase in either HF or GC reaction buffer (ThermoFisher), in 20 μL or 50 μL reaction volume. A 20 μL reaction contained the plasmid template (5 ng), the forward and reverse primers (500 nM each final concentration), an equimolar mix of dNTPs (200 μM each), 1.5% DMSO and Phusion (0.2 U). The reactions were assembled on ice, typically in a master mix, and Phusion added immediately before starting thermocycling.

The typical cycling conditions were: initial denaturation at 98°C for 15 s, then 30 cycles of: 98°C denaturation 10 s, annealing 15 s, 72°C extension 120 s; final extension 72°C 5-10 min and hold at 10°C. The extension time and annealing temperature were adjusted for each reaction. For most PCRs, the annealing temperature was optimised with a set of reactions with a gradient annealing temperature.

pID plasmid. The pID plasmid (also named pID-Tet) is one of two workhorse plasmids in this thesis. It was originally constructed by Stéphane Emond from a pASK-derived backbone (Figure 9.1). It is the plasmid that is primarily used for TRIAD mutagenesis, where it is essential that the plasmid does not contain any MlyI or AcuI restriction sites.

pID GFP TRIAD libraries. I received the six TRIAD (-3, -6, -9, +3, +6, +9 bp) and the TriNEx library, each in the eGFP and GFP8 background, from Stéphane Emond at the beginning of this project. The GFP sequences were cloned into the pID plasmid using the 5' NdeI and the 3' HindIII restriction site, which are also present in the pID multiple cloning site, such that the GFP sequence was C-terminally tagged with the 6×His expression tag (Figure 9.2). Then, the transposon based mutagenesis protocols gave the aforementioned libraries.

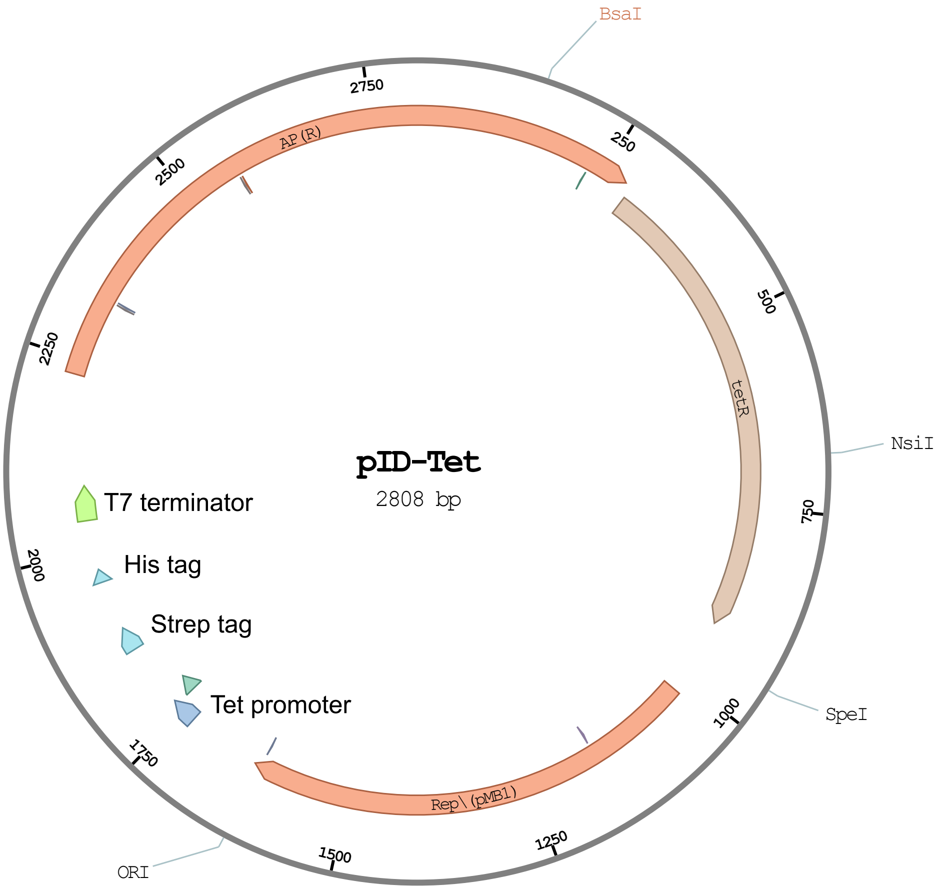


Figure 9.1 The plasmid map of the pID-Tet plasmid, abbreviated to pID. It features a pASK-derived backbone, ampicillin resistance, a multiple cloning site and the tet-repression system for controlling protein expression.

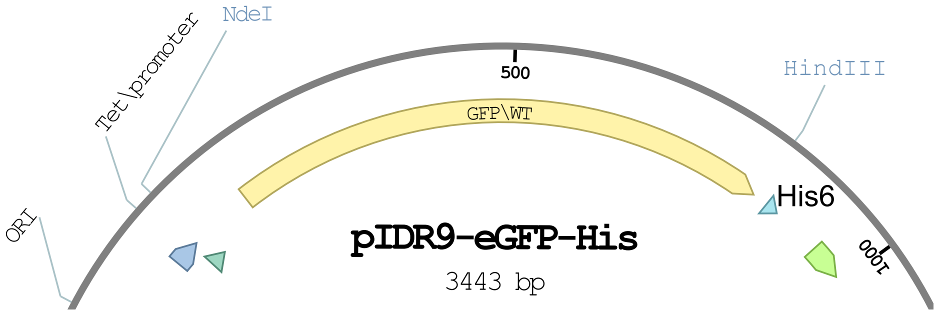


Figure 9.2 The eGFP sequence inserted into the pID cloning site. The ATG starting codon overlaps with the NdeI restriction site and the gene sequence ends with the HindIII restriction site. The Strep tag is no longer present in this plasmid, but the C-terminal His-tag remains.

pID eGFP::N18. The N18 UMI was introduced into the finished libraries, by replacing a part of the plasmid backbone with very similar sequence, which removed the His tag and placed an 18 nt UMI at the end of the GFP sequence. The new fragment, which included the pID ampicillin reistance gene, was amplified from the pID eGFP template with primers N18_F and NspI_R in a PCR reaction and purified by agarose gel electrophoresis. The new UMI-tagged fragment was cloned into the pID TRIAD libraries using NspI and HindIII restriction sites (replacing the part of the backbone, not the GFP library) following the standard cloning procedure.

This was my first construct that incorporated a UMI and presented several issues (Figure 9.3). First, the His tag was removed not because I did not wish to retain it, but because it presented issues for primer design (these returned later in the mKate2::GFP constructs, so the choice was justified). Second, having little experience at the time of designing these constructs, the GFP::N18 libraries now had a stop codon after the UMI, meaning the UMI would be translated. This was not ideal. However, since the C-terminal sequence has little effect on GFP folding or fluorescence, it was an acceptable stepping stone construct.

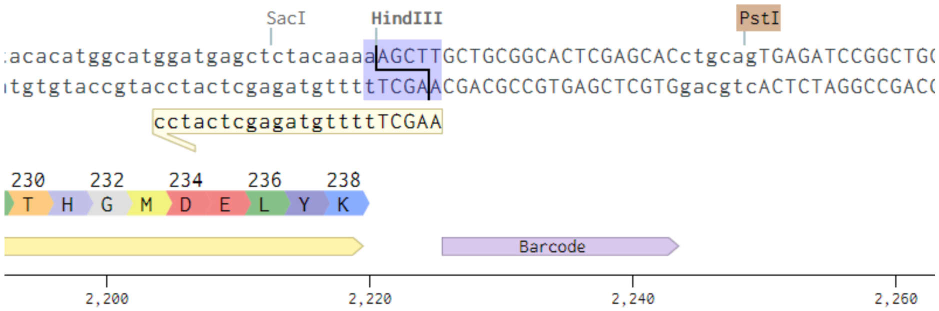

SacI
HindIII
PstI
:acacatggcatggatgagctctacaaaaAGCTTGCTGCGGCACTCGAGCACctgcagTGAGATCCGGCTG(
ıtgtgtaccgtacctactcgagatgttttTCGAACGACGCCGTGAGCTCGTGgacgtcACTCTAGGCCGAC(
cctactcgagatgttttTCGAA
230
232
234
236
238
T H G M D E L Y K
Barcode
2,200
2,220
2,240
2,260

Figure 9.3 C-terminal sequence in the GFP::N18 construct. The HindIII restriction site is followed by an 18 nt UMI (not the His tag), the PstI restriction site and a TGA stop codon.

pID mKate2::eGFP. The first construct that created the fusion protein mKate2::linker::GFP was prepared in the pID plasmid, still containing some sequence from the N18 UMI in the previous design in the pID eGFP::N18 construct (Figure 9.3). While the plasmid structure remains, this is no longer a functional UMI since I performed the PCR reactions from a single clone template.

The plasmid was constructed with Gibson assembly from three fragments. The vector fragment was amplified from a single clone pID eGFP::N18 plasmid with the primer pair GFP_F and Vector_R. The PCR conditions were optimised by screening a range of annealing temperatures, reaction buffers and the optimal DMSO concentration (0 or 3%). The optimal conditions were 3% DMSO in the GC buffer with annealing at 58°C, which gave a clean band at the expected 3.4 kb size. The PCR reaction was scaled up in these conditions and the product purified with Zymo Clean & Concentrate.

The mKate2 sequence was recovered from the plasmid pHed2 available from Addgene and amplified with primers mKate2_F and mKate2_R. The PCR optimisation was performed in the same manner. Since most reactions worked, the final fragment was purified from the pooled fractions in the HF buffer with 3% DMSO.

The linker sequence was adapted from Sarkisyan et al. (2016).Gibson assembly protocols recommend the use of multiple oligonucleotides for fragments that are longer than 60 bp, as is this linker. However, the linker sequence is repetitive and GC-rich, so I wished to avoid the potential for mis-priming of an oligonucleotide junction in the middle of the sequence. The fragment junctions in the Gibson assembly are designed with a T_m ~65°C.

The vector, mKate2 and oligonucleotides were mixed in 1:1:5 molar ratios and added to the Gibson Assembly master mix that was pre-warmed at 50°C. After a 60 minute incubation at 50°C, the mixture was column purified and transformed into 10G *E. coli* cells. Individual clones were selected and a correct variant identified with Sanger sequencing.

pID mKate2::trGFP . This plasmid was an intermediate construct, containing the correct start of the expression cassette (see Figure 3.3) but ending at residue G56 in eGFP (Figure 9.4). It was created from the pID mKate2::eGFP template, removing most of the GFP sequence and the unnecessary N18 sequence. The primer PNK_F introduced the Illumina sequencing primer sequence that forms the annealing sequence for the Read 1 primer in the Illumina system.

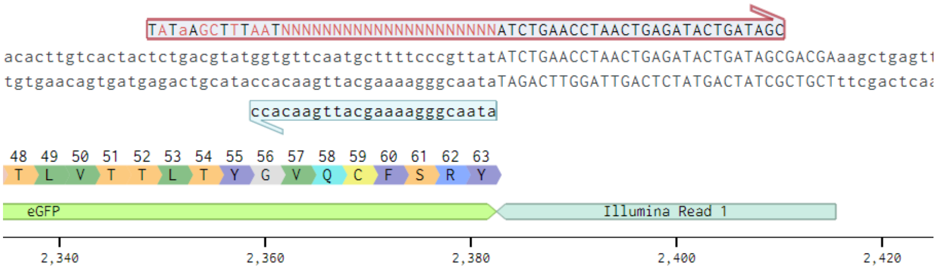

TATaAGCTTTAATNNNNNNNNNNNNNNNNNNNNNATCTGAACCTAACTGAGATACTGATAGC
acacttgtcactactctgacgtatggtgttcaatgcttttcccgttatATCTGAACCTAACTGAGATACTGATAGCGACGAaagctgagt
tgtgaacagtgatgagactgcataccacaagttacgaaaagggcaataTAGACTTGGATTGACTCTATGACTATCGCTGCTttcgactca
ccacaagttacgaaaagggcaata
48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
T L V T T L T Y G V Q C F S R Y
eGFP
Illumina Read 1
2,340
2,360
2,380
2,400
2,420

Figure 9.4 The technical sequence in pID mKate2::trGFP. This construct contains the His6::mKate2::linker::GFP casette, except the GFP sequence is truncated after residue G56. The Illumina primer annealing sequence is indicated. The annealing sites and overhangs for introduction of the 21 nt UMI are indicated. Red primer, UMI_N21_F. Blue primer, GFP_G56_R.

The pID mKate2::trGFP (truncated GFP) was created by a PCR reaction on the pID mKate2::eGFP template with primers PNK_F and PNK_R, which have designed annealing temperatures 63.0 and 63.6°C, respectively. The PCR reaction was optimised at annealing temperatures between 60°C and 70°C, with 3% DMSO and in both HF and GC buffers. At the end of the PCR reaction, the template was digested during a 2 h incubation with DpnI at 37°C. The correct bands were excised from an agarose gel and purified. The PCR product was circularised in a intramolecular blunt-end ligation with T4 ligase in the presence of phosphonucleotide kinase (PNK). After an overnight incubation at 18°C, the ligase mix was purified and transformed into 10G *E. coli* cells with electroporation. Colony PCR confirmed the presence of shorter mKate2::trGFP inserts in some of the colonies, which were Sanger sequenced to confirm successful creation of the desired variant.

pID mKate2::GFP::UMI. Next, the pID mKate2::trGFP plasmid was used as a template in a second PCR reaction, which added high quality 21 nt UMI and a HindIII restriction site into the plasmid backbone. The UMI-tagged linear product was prepared with pID mKate2::trGFP as a template, with primers UMI_N21_F and GFP_G56_R (Figure 9.4). After initial testing of the reaction conditions with a regular primer, a high quality version of the UMI_N21_F primer was ordered from Integrated DNA Technologies (IDT) - that primer had the same sequence, but the primer was synthesised with a better balanced hand-mix of random nucleotides at 4 nmol synthesis scale. The PCR reaction was performed in the HF buffer with 3% DMSO and 62°C annealing temperature, though this reaction gave a single band at a range of annealing temperatures. After a DpnI digest and checking the outcome of the reaction on an agarose gel, the PCR product was column purified using ZymoClean DNA Clean & Concentrate.

Finally, this UMI-tagged PCR product was used as the backbone for cloning in GFP TRIAD and TriNEx libraries (Figure 3.3), using the standard cloning procedure with HindIII and NdeI restriction endonucleases. Since high ligation efficiency is important when cloning large libraries, the purified fragments with sticky ends were ligated overnight at 16°C, column purified and transformed into commercial 10G E. cloni cells. Each library was plated on a large LB agar plate supplemented with carbenicillin and the transformation efficiency was estimated from counting colonies on dilution plates. If necessary, the transformation was repeated to give >10^6 colonies per insertion library.

UMI recovery. The recovery of UMIs was tested in a PCR reaction with the eGFP pooled deletion libray in the pID mKate2::GFP::UMI plasmid as a template, with primers Illumina_R1 and GFP_UMI (T_m 67°C for both). Without the need for a full optimisation screen, the reaction gave a single band at the expected size (89 bp), using the HF buffer, 3% DMSO and 65°C annealing temperature.

Table 9.1 Key primer sequences

Primer	**Sequence**
T7prom	TAATACGACTCACTATAGGG
T7t	GCTAGTTATTGCTCAGCGG
N18_F	CATAGTAAGCTTNNNNNNNNNNNNNNNNNNCTGCAGTGAGATCCGGCTGCTAACAAAGC
NspI_R	GCTAAGGCGTCGAGCAAAGC
mKate2_F	ATACTAATGCACCACCACCACCACCACTCAGAATTAATTAAAGAAAATATGCACATG
mKate2_R	CCGCCAGGCTGCCGAATTCACGATGTCCTAATTTCGACGG
Linker_F	GAATTCGGCAGCCTGGCGGAAGCGGCGGCGAAAGAAGCGGC
Linker_R	CTCATATGGCTCGCCGCCGCCGCTTTCGCCGCCGCTTCTTTCGTCGCCGCCGCTTCTTCCGCCGCTTCCGCCAGGCTGCCGAATTC
GFP_F	CGGCGGCGAGCCATATGAGTAAAGGAGAAGAACTTTTCAC
Vector_R	CTGAGTGGTGGTGGTGGTGGTGCATTAGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTC
GFP_G56_R	ATAACGGGAAAAGCATTGAACACC
PNK_F	TATCTGAACCTAACTGAGATACTGATAGCGACGAAAGCTGAGTGGCTGCTGC
PNK_R	ATAACGGGAAAAGCATTGAACACCATAC
UMI_N21_F	TATAAGCTTTAATGATC(N21)ATCTGAACCTAACTGAGATACTGATAGC
Illumina_R1	TCGTCGCTATCAGTATCTCAGTTAGGTTCAGAT
GFP_UMI	GGCATGGATGAGCTCTACAAAAAGCTTTAATGATC

9.2 Protein expression and crystallography

9.2.1 General expression conditions and activity assays

Small volume cultures. Cultures were started from either a single colony or a glycerol stock, and used to inoculate a 1 mL overnight culture in LB medium with 100 μg/mL ampicillin or carbenicillin. All media for growing PTE were supplemented with $ZnCl_2$ (200 μM final concentration). If the cells contained pGro7 for over-expression of GroEL/ES chaperones, the medium was supplemented with 0.02% w/v glucose (to repress GroEL/ES) or 0.02% w/v arabinose (to induce GroEL/ES expression) throughout. In the morning, 25 μL of the overnight culture were mixed with 975 μmL to start the expression culture and grown at 30°C or 37°C with agitation, and the cell growth was monitored with NanoDrop until OD_{600nm} reached 0.4-0.6. The culture was induced with addition of isopropyl β-D-1-thiogalactopyranoside (IPTG; 1 mM final concentration) or anhydrotetracycline (aTc; 200 μg/mL final concentration) and incubated at 30°C (unless specified otherwise) with continued agitation for 2-3 hours.

PTE initial rate assays. Following protein expression, the OD_{600nm} of the cultures was measured, the cultures were pelleted and the supernatant discarded. The pellet was resuspended in the lysis buffer (Bugbuster 1× or 0.1% (w/v) Triton-X100 detergent, 1μL Lysonase per 10 mL lysis buffer, 50 mM Tris·HCl pH 7.5, $ZnCl_2$ 200 μM) in 1:10 volume ratio, that is 100 μL lysis buffer for expression culture with 1 mL volume, and incubated at room temperature for 30-60 minutes with gentle agitation.

For initial rate assays used to estimate rate at a similar rate as occurring in microdroplets (section 6.2.1), the lysate was *not* clarified; when a cell lyses in a droplet, the droplet will contain both soluble and insoluble components. The lysate was pipetted into a black 96-well plate (FluoroNunc F96, Thermo Scientific) at 2× final concentration. Shortly before starting fluorescence measurements, an equal volume of substrate solution (20 μM fluorescein-triphosphate derivative shown in Figure 1.7, structure **2**) was quickly added with a multi-channel pipette. Fluorescence was measured on a microplate reader (Infinite 200 PRO, Tecan). Initially the plate was shaken for 5 seconds, followed by data collection. Samples were excited at 488 nm and emission recorded at 525 nm with 20 nm bandwidth. Each sample was measured every 30 seconds and data recorded with Magellan 7.2 software.

For initial rate assays assessing the relative activity of PTE variants, the variants were grown in a 96-well microtitre deep-well plate with 1 mL culture volume, as described above, in triplicate (three expression plates started from the same glycerol stock). The lysate was clarified by centrifugation and pre-diluted using serial dilutions to 1:10, 1:100 and 1:100 concentrations (20 μL lysate mixed with 180 μL 50 mM Tris·HCl pH 7.5 buffer containing 200 μM $ZnCl_2$). The pre-diluted lysates were mixed with the paraoxon substrate solution (180 μL of 200 μM paraoxon in 50 mM Tris·HClHCl, pH 7.5 supplemented with 0.02 %Triton-X100) in a clear-bottom microtitre plate. The reaction rate was monitored by recording absorbance at 405 nm, at the minimum possible time interval (16 s) in the SpectraMax 190 plate reader (Molecular Devices).

Initial rate data analysis. The data were analysed individually for each plate against controls recorded in three wells: no growth, negative activity control (cells containing a pET acylphosphatase (ACP) plasmid) and positive controls (pET *wt*PTE). The initial rate slopes were calculated for each well and each dilution using SoftMax Pro software, and the absorbance curves for each well were visually inspected to confirm linearity. For each well in each plate, the dilution was chosen that exhibited linear rate over the longest time interval (typically >600 s). Data from wells with obviously false signals were discarded. The initial rates were calculating by first correcting the average OD_{600nm} value for the no-growth control wells and subtracting it from all other wells, then the initial rate slopes were divided by this corrected OD_{600nm} for the corresponding well to correct for different bacteria growth rates between well. This gives the adjusted slope of each well. For each replicate plate separately, the activity of the negative and positive controls was calculated by averaging the adjusted slopes of triplicate controls in each plate at 1:1000 lysate dilution level. Then, the relative activity of individual variants was calculated as $\frac{Slope(variant)-Mean(ACP)}{Mean(wtPTE)}$. For variants measured at 1:100 and 1:10 dilution, the slope of the variant was divided by 10 or 100, respectively. Finally, the relative activities measured in the replicate plates were compared and averaged to give the activity of each variant relative to *wt*PTE.

Protein solubility measurement. Protein solubility was analysed by SDS-PAGE using NuPAGE Bis-Tris 4-12% pre-cast gels. The soluble fraction was assayed by analysing the clarified lysate, and to assay the insoluble fraction, the pellets obtained after lysis were re-suspended in lysis buffer with vigorous vortexing. The amount of sample (soluble and/or insoluble fractions of different variants) to be loaded on the gel was determined by normalization to the OD600. Each sample aliquot was mixed with NuPAGE LDS Sample Buffer and heated at 95°C for 10 minutes to denature the protein. The gels were run at 200 V for 60 min in MOPS buffer, then developed with InstantBlue Coomassie Protein Stain. The gels were photographed under equal aperture and exposure conditions and the intensity of the protein bands was measured using ImageJ.

Thermal denaturation assay. For this assay, the overnight cultures containing plasmids with PTE variants of interest in BL21(DE3) cells were used to inoculate a 200 mL expression culture in Overnight Express Instant TB Medium (Novagen) supplemented with carbenicillin and $ZnCl_2$. The expression culture was incubated at 30°C for 8 hours and then at 20°C overnight, both with shaking, then pelleted and lysed to give crude protein. The clarified lysates were passed through 20 μm filters (CellTrics) before loading onto a Strep-TActing

Superflow High Capacity column (1.5 mL resin). The columns were washed with 5 times with 1.5 mL wash buffer (100 mM Tris·HCl pH 8.0, 150 mM NaCl, 200 μM $ZnCl_2$) and eluted with 6 × 0.75 mL elution buffer (wash buffer with added 2.5 mM d-desthiobiotin). The purified proteins were exchanged into storage buffer (100 mM Tris·HCl pH 7.5, 200 μM $ZnCl_2$) with PD-10 desalting columns (GE Healthcare Life Sciences). Protein concentration was measured with absorbance at 280 nm.

Heat-induced unfolding of PTE variants was measured in triplicate over a range between 25 to 95 °C in a BioRad CFX Connect, using purified protein (5 and 10 μM final concentration in the storage buffer) and SYPRO™ Orange Protein Gel Stain (5X and 10X final concentrations). Protein unfolding was monitored by measuring the change in fluorescence caused by binding of the dye ($\lambda_{excitation}$= 488 nm; $\lambda_{emission}$= 500–750 nm) and the midpoint of denaturation (T_m) was determined as the maximum of the first derivative for each temperature–fluorescence curve, then averaged across all replicates that showed a defined melting curve (6 or more per variant).

9.2.2 GFP cell sorting

Single colour sorting. An aliquot of each deletion library in pID-Tet plasmid backbone was transformed into the laboratory stock of Lucigen 10G cells. The substitution libraries were similarly transformed into commercial 10G cells, to ensure sufficiently high transformation efficiency to maintain the diversity of the larger libraries. The transformed cells were allowed to recover at 37°C for one hour without shaking and then the transformation culture was diluted 1:20 into growth medium (LB-ampicillin) to grown overnight at 37°C.

The overnight culture was diluted 1:10 into fresh LB-ampicillin medium that was directly supplemented with anhydrotetracyclin and incubated at 30°C for 2 hours with shaking to express GFP. Cell density was calculated from $OD_{600\ nm}$ and diluted with PBS to 5×10^6 cell/mL ($OD_{600\ nm}$ = 1.0 corresponds to 10^8 cells/mL). Cultures containing the empty pID-Tet plasmid (negative) and pID-eGFP (positive) plasmids were grown from a single colony on agar plates and used as controls in setting sorting gates. All cultures were filtered through a 20 μm CellTrics filter prior to sorting.

Diluted cell solutions were sorted on Beckman Coulter Astrios EQ in the Flow Cytometry facility, Department of Pathology, with assistance by Nigel Miller. Cells were excited with 488 nm laser and fluorescence detected in the green channel (513/26 nm). Forward scatter and side scatter gating was used to separate single cells from clusters, debris and instrument noise. Cells were sorted at medium dilution (3000-5000 events/second) to avoid false positives from sorting two cells simultaneously. The fluorescent signal from positive and negative controls was used to set a sorting gate for high and low fluorescence, respectively. A medium gate was set as a narrow region in-between, leaving a range of fluorescence between gates that was not sorted to reduce mis-sorted events. Each library was sorted into LB-ampicillin solution until 10^5 events/gate were collected. Thus, each deletion library was over-screened at least 1000× and the substitution libraries at least 20×.

The cells were sorted into 1 mL of LB medium supplemented with ampicillin, to promote cell viability. The sorted cell solutions were diluted to 10 mL LB-ampicillin and grown overnight at 37°C with shaking. The plasmid DNA was isolated using GeneJET Plasmid Miniprep Kit and stored at -20°C for later use.

Two-colour sorting. The cell cultures for two-colour sorting were transformed into BL21 (DE3) cells - commercial or laboratory stock, depending on the library diversity - and after recovery at 37°C, the transformant cultures were plated on large LB-agar plates supplemented with carbenicillin (100 mg/mL) and incubated overnight to generate a bacterial lawn. The plates were gently scraped with 5 ml LB per plate to generate a homogenous cell suspension. The suspension (2μL) was used to inoculate 1 mL of expression culture per sample, which was grown at 30° or 37°C with shaking until $OD_{600\ nm}$ reached 0.4-0.6. The expression cultures were induced with anhydrotetracyclin and incubated at 30°C (unless otherwise specified) with shaking for 2-3 hours. The final cultures were diluted 1:50 v/v into PBS buffer (pH 7.5) supplemented with carbenicillin. The sorting samples were filtered through 20 μm CellTrics filters before sorting.

The diluted cell solutions expressing the GFP libraries in the pID mKate2::GFP::UMI fusion construct were sorted on the Becton Dickinson FACSAria III cell sorter in the Flow Cytometry facility in the Department of Pathology, with assistance by Joana Cerveira and Carl Bradford. A 70μm nozzle was used. Forward scatter area vs side scatter area gating was used to separate single cells from clusters, debris and instrument noise, and the events were additionally gated on forward scatter height vs area to separate singled cells from doublets.

The mKate2 was used as an expression level control in the fusion construct, with fluorescence excited with the 561 nm laser and emission recorded in the 610/20 nm height channel. GFP fluorescence was excited with the 488 nm laser and fluorescence detected in the 530/30 nm green channel. The cells were sorted with either 4-way purity mode (0/32/0) for separating into four activity gates, or with enrichment setting (16/16/0) for pre-sorting; all samples were were sorted at high dilution (1000-3000 events/second) to avoid false positives from sorting two cells simultaneously.

Analysis of sorting efficiency. The sorted plasmid DNA was re-transformed into lab stocks of electrocompetent cells (10G or BL21(DE3) as appropriate) and the overnight and expression cultures were prepared following the same procedure as for sorting. With cells expressing only GFP, the distribution of fluorescence was measured on the Guava easyCyte flow cytometer (excitation 488 nm blue laser, emission in the green channel).

The fluorescence of cells expressing the mKate2::GFP fusion constructs was measured on the Attune NxT flow cytometer in the flow cytometry facilities in the Department of Pathology or in the Cambridge BRC Cell Phenotyping Hub. The mKate2 fluorescence was excited with the 561 nm yellow-green laser and the emission recorded in the YL2 channel at 620/15 nm. All flow cytometry data was analysed with FlowJo software.

9.2.3 GFP8 crystallography

Large volume cultures. C-terminally 6-His tagged GFP8 in pID-Tet under control of a tetracyclin-repressor was transformed into E. coli 10G as described above. An overnight culture was grown from a single colony, the sequence verified by Sanger sequencing and the culture stored in glycerol stock at -80°C. The glycerol stock was used to start all further overnight cultures.

To prepare a batch of protein, an overnight culture was grown in 500 mL LB medium supplemented with ampicillin at 37°C, while shaking at 180 rpm. When $OD_{600\ nm}$ reached 0.4, the culture was cooled at 20°C for 30 minutes and anhydrotetracyline (aTc) was added to induce protein expression, followed by overnight incubation at 20°C with shaking at 180 rpm. In the morning the cells were harvested by centrifugation at 4000 rpm for 15 minutes at room temperature and the pellets lysed for 30 minutes in 50 mL lysis mix, which contained 10 uL Lysonase and 1× BugBuster in PBS buffer at pH 7.5.

Protein purification. The lysate was centrifuged at 7200 rpm for 30 minutes to remove cell debris, then loaded onto a pre-washed Ni-NTA affinity column in multiple portions, washed with 30 mM imidazole buffer and eluted in 5 mL elution buffer containing 500 mM imidazole. The eluted portion was concentrated, the buffer exchanged to storage/running buffer (50 mM Tris·HCl,pH 8.0, or 100 mM HEPES, pH 7.5) using PD-10 desalting columns (GE Healthcare Life Sciences) and the protein stored at -20°C or at 4°C overnight.

On the next day, the sample was thawed on ice, centrifuged at 14,100 rcf until clear and concentrated to 2 mL volume in preparation for gel filtration using the ÄKTApurifier fast protein liquid chromatography system. The sample was injected onto HiLoad 16/60 Superdex 75 prep grade size exclusion column (120 mL total volume) that had previously been equlibrated with water and then running buffer, kept at 8-10°C. The column was washed with one column volume of running buffer and fractions were collected in a 96-well plate, 1.25 mL each. Two peaks appeared in the UV-VIS chromatogram at 280 nm, and the second peak was green and showed the correct weight with SDS-PAGE. The appropriate fractions were collected, concentrated and stored at -20°C or at -4°C.

Crystallisation trials. Immediately before setting up crystallisation plates, the sample was thawed if necessary, centrifuged at 14,100 rcf for 15 minutes to remove any precipitate (if present) and the protein concentration adjusted to 20-50 mg/mL. The initial crystallisation screen was set up using three sparse matrix plates: Wizard I&2, Classics and JCSG+, screened against His6-tagged GFP8 purified to homogeneity as above, used at 15 mg/mL in 100 mM HEPES pH 7.5. The crystallisation trial was set up as sitting drop, two drops per well (1:1 and 2:1 protein:buffer mixture) prepared using Mosquito Crystal (TTPLabetch Ltd.) liquid dispensing system with disposable tips, then monitored within Rock Imager Hotel 1000 (Formulatrix, Inc.) at 19°C.

Three hits were identified and optimized in a 24-well plate in handing drop format. All wells were set up with two drops containing 1:1 and 2:1 protein : reservoir solution, where the reservoir contained 200 mM of the appropriate buffer, between 50 and 400 mM NaCl and the stated precipitant. Hit 1 (0.1 M HEPES pH 7.5, 200 mM NaCl, 10% v/v isopropanol) reservoir solution contained HEPES at pH 7.1, 7.5, 7.9 and 8.4, with 10% v/v isopropanol in reservoir. Hit 2 (100 mM CHES pH 9.5, 200 mM NaCl, 10% w/v PEG 8000) reservoirs contained 200 mM CHES at pH 8.9, 9.2, 9.6 and 9.85, and 10% m/v PEG 8000. Hit 3 (0.1M CAPS pH 10.5, 0.2M NaCl, 20% w/v PEG 8000) reservoirs contained CAPS at pH 9.1, 10.1, 10.5 and 10.9, and 10% m/v PEG 8000. Promising crystals were collected onto loops with 20% v/v glycerol as cryoprotectant, flash-frozen and stored in liquid nitrogen until data collection.

Diffraction data collection and analysis. Diffraction data was collected at Diamond Light Source, Oxford, UK, on the I04-1 beamline with wavelength 0.9282 Å. 7 out of 13 crystals diffracted. The best dataset (2:1 protein:reservoir, which contained 200 mM HEPES pH 7.5, 200 mM NaCl, 10% v/v isopropanol) was used to solve the crystal structure utilizing the CCP4 sofware suite. The Matthews coefficient indicated that the structure contained one monomer per unit cell and 39% volume was occupied by water. The initial model was derived with coordinates 4LQT using AMoRe (within CCP4) and the map density added by refining the structure with refmac5. The structure was manually edited in Coot to add missing residues, include the chromophore and water molecules. The full model was refined three times using Phenix until MolProbity validation reported no clashes and there was no improvement in the R scores.

9.3 Microfluidics

Microfluidic rig setup (2015-2016). Fluorescence measurements were carried out on a custom-built fluorescence microscopy rig. The light emitted from a 60W halogen light source was filtered (593 nm long-pass filter, BrightLine) and used to illuminate and image the chip using a high-speed camera (Miro eX4, Phantom). A diode-pumped solid-state laser (488 nm, 58-BCD-030-240, Melles Griot) was focused onto the microfluidic chip through an objective (Lplan Fluorphase x40 / 0.65 NA) which was incorporated in an inverted microscope (SP105F, Brunell Microscopes). The emitted light was collected by the same objective and was separated from the excitation light by a dichroic beam splitter (DI495LP). A second dichroic separated the red illumination light from the green fluorescence (555 nm single-edge, BrightLine). The emitted fluorescence was focused onto an avalanche photo diode after passing a bandpass filter (520/28 nm single-band, BrightLine). The detector signals were recorded using a data acquisition card (USB 6341, National Instruments) and analysed with a custom-made LabVIEW program (National Instruments).

Chip design and fabrication. Microfluidic devices were fabricated using standard soft lithography procedures (McDonald et al., 2000). New designs for single depth devices were prepared by myself with input from Phillip Mair and Dr. Fabrice Gielen using DraftSight CAD software. The master wafers for new designs were kindly fabricated by Dr Fabrice Gielen under conditions that give a 20 or 50 μm channel height.

To create individual devices, PDMS monomer and curing agent were mixed at a ratio 10:1, poured onto the lithographic plate and degassed by alternating between regular and reduced air pressure. The PDMS was solidified (65°C, min. 4 h) and individual chips cut out and peeled off the master. PDMS was activated by exposure to oxygen plasma and devices were sealed onto a microscope glass slide. Hydrophobic modification of the channels surface in the flow-focusing devices, which prevents the droplets breaking and the aqueous layer coating the channel walls, was achieved by injecting a solution of 1% (v/v) trichloro(1H,1H,2H,2H-perfluorooctyl)silane (Sigma) in HFE-7500 oil into the freshly bonded chip. The newly generated chips were heated to improve the strength of the PDMS-glass bonding.

Single depth chip operation. Droplets were generated using a flow-focusing device bearing three inlets. For the purposes of chip assessment, both aqueous inlets carried MilliQ water. When cells were encapsulated, these inlets carried aqueous solutions prepared in PTE storage buffer. The aqueous inlets supplied 1) a cell suspension with the required OD600nm cell density in Percoll (25%, v/v, Sigma) and 2) a solution of the fluorescein phosphotriester substrate (20 μM in PTE buffer). The third inlet contained a solution of block co-polymer surfactant AZ KryJeffa30 (2% w/w) in fluorinated oil HFE-7500 (Novec). All concentrations and flow rates are given for solutions before mixing. Aqueous and oil solutions were injected using plastic syringes (BD, 1 or 2.5 ml) on constant-flow pumps (neMESYS 14:1). The devices were imaged under a microscope (Brunell Microscopes) under a 10× or 40× bright-field objective.

PTE-expressing cells were encapsulated in droplets at flowrates 0.80 μl/min for each aqueous and 20 μl/min for the oil phase, to give large droplets (78 μm diameter, 250 pL) on the 50 μm deep channel device. Alternatively, smaller droplets (60 μm, 110 pL) were generated at flow rates 0.40 μl/min for each aqueous and 30 μl/min for the oil phase. Droplets passed through a short delay line and fluorescence was measured at the last restriction point.

Two-depth chip operation. The two-depth integrated device was operated in a similar manner, with two alterations. The surfactant used was 008-FluoroSurfactant (RAN Biotechnologies). Instead of plastic syringes, more reliable gas-tight SGE glass syringes were used (100 μL - 2.5 mL), preferably with a fixed needle. The syringes were dried at 65°C before use and washed with water and acetone after use.

9.4 Bioinformatic analysis

GitHub repository and code availability. All scripts are available in the GitHub repository at https://github.com/MayaPetek/InDelScanner, where they are organised in folders by dataset (PTE, GFP and MKK1). The base folder contains a conda environment file, which can be installed with 'conda env create -f InDelScanner.yml'. All other Python scripts assume this environment is active.

9.4.1 InDelScanner analysis of *wt*PTE TRIAD libraries

The PTE-specific scripts, in the exact form they were used, are available in the repository https://github.com/fhlab/TRIAD. While they are substantially the same as those in the development repository (InDelScanner), here they are presented exactly as they were used for data analysis in the Emond et al. (2020) publication. In this section, all scripts references are those in the fhlab/TRIAD repository.

Baseline dataset. The clean dataset for an ideal library of variants is generated with the script 'baseline.py', where each section of the code creates a separate file with a read for each possible variant. The ideal +6 bp library was generated in sections because of large file sizes and the counts later combined. The reads were then processed in the usual manner that is outlined below.

Library preparation. Libraries were digested from pID-Tet with FastDigest restriction enzymes Bpu1102I and Van91I and purified by gel electrophoresis to give a pool of 1.3 kb linear fragments. The gene fragments were submitted to Department of Biochemistry Sequencing Facility, where they were processed using Nextera DNA Library Preparation Kit according to manufacturer's instructions. The libraries were sequenced on Illumina MiSeq using 2x75 bp paired-end sequencing in a single sequencing run and the reads de-multiplexed and the adaptors trimmed to give the starting FASTQ files for each library.

Read processing. The raw read processing was pipelined with the script 'count.sh', which was run on the University of Cambridge computational cluster. Alternatively, the reads can also be processed with 'count_PC.sh', which is intended for processing on a PC machine. The raw forward and reverse reads were merged using PEAR version 0.9.10 (Zhang et al., 2014), with the options '- -keep-original - -min-overlap 5 - -min-assembly-length 0 - -quality-threshold 15 - -max-uncalled-base 0.01' and both assembled and unassembled reads were taken forward for the analysis.

An index of the reference sequence was created with Bowtie2 version 2.3.4 (Langmead and Salzberg, 2012), the reads were mapped to the reference and the resulting SAM files sorted with SAMtools version 1.9 (Li et al., 2009). At this point, 95% of the reads aligned to reference sequence. If the data is processed on a PC, at this point the non-wild-type reads are extracted from the SAM file based on the SAM flags, in order to reduce file sizes and RAM requirement. Working on the cluster, all reads were used.

The reads were re-aligned to reference using the Needleman-Wunsch algorithm with the scoring matrix 'EDNAFULL', gap open penalty 15 and gap extend penalty 0.5. The NW alignment was done with the *needleall* program implemented in EMBOSS version 6.6 (Rice et al., 2000).

Placing InDels in particular sequence contexts may be inherently ambiguous because of potential InDel redundancy: when two or more InDels inserted at different positions in the target gene result in identical final sequence, no algorithm will be able to distinguish between them and the resulting InDel is always assigned to a single arbitrarily chosen original insertion or deletion site. No attempt was made to correct for such ambiguity at this point.

Parsing variants. The resulting alignments were used to count the number of reads in which the mutations occur, their type and position using in-house developed Python scripts available in the GitHub repository. The main code is in the script 'PTE_composition.py' and the helper functions are drawn from the script 'ind.py'. The picked datasets are available in the folder 'manuscript_data'. Further statistics are presented in the IPython workbook 'TRIAD_composition_figure.ipynb', which reproduces the figures presented in this chapter. The workbook was composed with Jupyter Lab.

To analyse the sequence preference for TransDel transposition, the five-nucleotide target sequence around each detected -3 bp deletion was extracted in forward and reverse complement direction, since the direction of transposon insertion is unknown. The frequency of insertion at each position was used to weigh the contribution to consensus sequence, then normalized to give the proportion of each nucleotide per position in the Mu transposon consensus sequence.

9.4.2 SpliMLib peptide libraries

Read processing. Due to close spacing of substitutions and insertions in a short sequence, the reads could not be sensibly aligned to reference as DNA sequence. Therefore, all following steps were done on the protein sequences using a set of custom Python scripts, which are freely available at www.github.com/fhlab/Kinases. The initial pre-processing is similar as for TRIAD libraries, while the statistical analysis was optimised for the particular library design.

The raw forward and reverse reads were merged using PEAR and the unassembled reads were discarded. The assembled reads were filtered for reads containing the correct 5- or 6-base forward or reverse sequence, using the regular expression (line broken for clarity):

ATGCCCAAGAAGAAG[ACTG]3ACG([ACTG]3)?CCG(GCG)?
[ACTG]3CAG[ACTG]3AAC[ACTG]3GCCCCCGACGGC.

All matching sequences were translated into protein sequence and aligned to the reference sequence MPKKKXTPXQXNXAPDG using the Emboss implementation of Needleman-Wunsch algorithm (Rice et al., 2000) with a custom scoring matrix `PNULL` (located in the

’indels’ directory in the GitHub repository). In this matrix, each amino acid match is scored equally positive (5), all amino acid mismatches are penalized equally (-1) and a match to a randomized X position is scored as weakly positive (-1). Insertions and deletions are scored with gap open penalty 3 and gap extension penalty 5. This scoring scheme allows successful alignment of short, highly randomized sequences and correctly places the inserted residues (7a, 8a, if they are present) after the randomized substitution position. The aligned sequences were parsed to extract and count all present mutations and stored as a dictionary for later analysis.

Library composition. All three sequencing fractions were analysed to determine the number of observed variants, setting a cut-off of $\geq$10 sequencing reads for the high gate to reduce spurious observation of sequencing noise. However, because the medium and low gates are substantially more diverse, a lower cut-off of $\geq$3 supporting reads was used there. The amino acid diversity at each randomised position was calculated by extracting the amino acid position at each randomised position and weighing the composition by the number of times each variant was observed in the fraction of interest.

Mutual information. The analysis of covariation between randomized positions was adapted from Taylor and Sadowski (2011), using sequencing count in the filtered set of variants in place of amino acid frequency in an multi sequence alignment. The maximum alphabet size was for each position was the number of possible amino acids actually introduced (2, 12 or 13). This strategy was used both for single position Shannon entropy (S_i) and for the joint entropy of a pair of positions (S_{ij}), which then give the mutual information between two positions: $M_{ij} = S_i + S_j - 2S_{ij}$. Finally, mutual information between positions was normalized by dividing by joint entropy, such that $M^{'}_{ij} = \frac{M_{ij}}{M_{ij}}$.

Dataset curation. The variants observed in the high gate were filtered to generate a curated dataset of active variants, including two sets: a) all variants with $\geq$50 reads in the high gate, with reads in the low gate representing less than 1/3 of all reads for this variant, and b) variants with a high gate read count 10-49, medium gate count < high gate count, and low gate reads representing less than 1/6 of all reads for these variants.

The single position enrichment was derived by dividing the observed proportion (frequency) of each possible amino acid, f_a^{obs} (ranging between 0 and 1), by the ideal proportion f_a^{id} that is expected in balanced library (f_a^{id} = 1/12, 1/13 ,1/2 for positions 6, 9, 11, 13; 7a; and 8a, respectively). Similarly, the two-position enrichment was calculated by dividing the joint observed frequency of amino acids a and b, $f_{a,b}^{obs}$, by the ideal joint proportion$f_{a,b}^{id} = f_a^{id} f_b^{id}$ (since in the ideal case, the probabilities are independent). The presence of epistatic interactions between positions (i.e. deviation from independence of probabilities) is seen by dividing the $f_{a,b}^{obs}$ (= $f_{a|b}^{id} f_b^{id}$) by the joint proportions expected from the one-dimensional enrichment at each positions, given by $f_{a,b}^{obs} = f_a^{obs} f_b^{obs}$ without epistasis.

Sequence similarity network. A sequence similarity network was constructed with Python using the NetworkX package and the visualisation software Gephi. Each variant in the curated set of active variants was added as a node and the Hamming distance between all pairs of variants was calculated; an edge was added to the network if the Hamming distance was equal to 1. The degree, betweenness, closeness and eigenvector network centrality scores were calculated for all nodes, but they did not reveal any clear hub nodes.

I tested if the network can be partitioned in a meaningful way; this shows that a standard Louvain partitioning had a good modularity score of 0.7 (range -0.5 to +1.0, a positive value is better). However, the Louvain algorithm only guarantees that each node is well connected to its own community, yet some communities might be poorly connected to themselves (that is, missing a bridge within the community). This mis-assignment can happen if the bridge node within a community is also well connected to nodes outside of the community, so it gets reassigned to the outside community – leaving the original community divided. Therefore, I used the improved Leiden algorithm (Traag et al., 2019), which identified 11 large communities within the largest connected subgraph of the full network. The D-domain sequences in each community were extracted in FASTA format and used to create a WebLogo representation of sequence content (Crooks, 2004).

References

Aakre, C. D., Herrou, J., Phung, T. N., Perchuk, B. S., Crosson, S., and Laub, M. T. (2015). Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell*, 163(3):594–606.

Afriat, L., Roodveldt, C., Manco, G., and Tawfik, D. S. (2006). The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry*, 45(46):13677–13686.

Afriat-Jurnou, L., Jackson, C. J., and Tawfik, D. S. (2012). Reconstructing a Missing Link in the Evolution of a Recently Diverged Phosphotriesterase by Active-Site Loop Remodeling. *Biochemistry*, 51(31):6047–6055.

Ajawatanawong, P. and Baldauf, S. L. (2013). Evolution of protein indels in plants, animals and fungi. *BMC Evolutionary Biology*, 13(1):140.

Akiva, E., Copp, J. N., Tokuriki, N., and Babbitt, P. C. (2017). Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily. *Proceedings of the National Academy of Sciences*, 114(45):E9549–E9558.

Arai, R., Ueda, H., Kitayama, A., Kamiya, N., and Nagamune, T. (2001). Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Engineering, Design and Selection*, 14(8):529–532.

Armstrong, C. D. (2007). *Elucidating the Chemical and Thermal Unfolding Profiles of Organophosphotrus Hydrolase and Increasing Its Operational Stability*. PhD Thesis, Texas A&M University.

Arpino, J. A., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J., and Jones, D. D. (2014a). Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure. *Structure*, 22(6):889–898.

Arpino, J. A. J., Rizkallah, P. J., and Jones, D. D. (2014b). Structural and dynamic changes associated with beneficial engineered single-amino-acid deletion mutations in enhanced green fluorescent protein. *Acta Crystallographica Section D: Biological Crystallography*, 70(8):2152–2162.

Aubert, S. D., Li, Y., and Raushel, F. M. (2004). Mechanism for the Hydrolysis of Organophosphates by the Bacterial Phosphotriesterase. *Biochemistry*, 43(19):5707–5715.

Auerbach, D., Klein, M., Franz, S., Carius, Y., Lancaster, C. R. D., and Jung, G. (2014). Replacement of Highly Conserved E222 by the Photostable Non-photoconvertible Histidine in GFP. *ChemBioChem*, 15(10):1404–1408.

Baldwin, A. J., Busse, K., Simm, A. M., and Jones, D. D. (2008). Expanded molecular diversity generation during directed evolution by trinucleotide exchange (TriNEx). *Nucleic Acids Research*, 36(13).

Bank, C., Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. (2015). A Systematic Survey of an Intragenic Epistatic Landscape. *Molecular Biology and Evolution*, 32(1):229–238.

Bardwell, A. J. and Bardwell, L. (2015). Two Hydrophobic Residues Can Determine the Specificity of Mitogen-activated Protein Kinase Docking Interactions. *The Journal of Biological Chemistry*, 290(44):26661–26674.

Baret, J.-C., Miller, O. J., Taly, V., Ryckelynck, M., El-Harrak, A., Frenz, L., Rick, C., Samuels, M. L., Hutchison, J. B., Agresti, J. J., Link, D. R., Weitz, D. A., and Griffiths, A. D. (2009). Fluorescence-activated droplet sorting (FADS): Efficient microfluidic cell sorting based on enzymatic activity. *Lab on a Chip*, 9(13):1850–1858.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1993). Empirical and Structural Models for Insertions and Deletions in the Divergent Evolution of Proteins. *Journal of Molecular Biology*, 229(4):1065–1082.

Bloom, J. D. (2015). Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16(1):1–13.

Bloom, J. D., a Romero, P., Lu, Z., and Arnold, F. H. (2007). Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biology direct*, 2:17.

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874.

Boucher, J. I., Jacobowitz, J. R., Beckett, B. C., Classen, S., and Theobald, D. L. (2014). An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *eLife*, 3:e02304.

Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N. S., Zhu, C., Bagul, M., Goetz, E. M., Burgin, A. B., Garraway, L. A., Getz, G., Mikkelsen, T. S., Piccioni, F., Root, D. E., and Johannessen, C. M. (2016). Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Reports*, 17(4):1171–1183.

Bridgham, J. T. (2006). Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science*, 312(5770):97–101.

Briseño-Roa, L., Oliynyk, Z., Timperley, C. M., Griffiths, A. D., and Fersht, A. R. (2011). Highest paraoxonase turnover rate found in a bacterial phosphotriesterase variant. *Protein Engineering, Design and Selection*, 24(1-2):209–11.

Broom, A., Jacobi, Z., Trainor, K., and Meiering, E. M. (2017). Computational tools help improve protein stability but with a solubility tradeoff. *Journal of Biological Chemistry*, 292(35):14349–14361.

Caldwell, S. R., Newcomb, J. R., Schlecht, K. A., and Raushel, F. M. (1991). Limits of Diffusion in the Hydrolysis of Substrates by the Phosphotriesterase from. *Biochemistry*, 30(30):7438–7444.

Campbell, E., Kaltenbach, M., Correy, G. J., Carr, P. D., Porebski, B. T., Livingstone, E. K., Afriat-Jurnou, L., Buckle, A. M., Weik, M., Hollfelder, F., Tokuriki, N., and Jackson, C. J. (2016). The role of protein dynamics in the evolution of new enzyme function. *Nature Chemical Biology*, 12(September):1–13.

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, 26(7):1523–1531.

Chen, J. Z., Fowler, D. M., and Tokuriki, N. (2020). Comprehensive exploration of the translocation, stability and substrate recognition requirements in VIM-2 lactamase. *eLife*, 9:e56707.

Cherf, G. M. and Cochran, J. R. (2015). Applications of Yeast Surface Display for Protein Engineering. In Liu, B., editor, *Yeast Surface Display: Methods, Protocols, and Applications*, Methods in Molecular Biology, pages 155–175. Springer, New York, NY.

Choi, J. Y., Jang, T. H., and Park, H. H. (2017). The mechanism of folding robustness revealed by the crystal structure of extra-superfolder GFP. *FEBS Letters*, 591(2):442–447.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A. (2003). Evolution of the Protein Repertoire. *Science*, 300(5626):1701–1703.

Colin, P.-Y., Kintses, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., Hyvönen, M., Morgavi, D. P., Janssen, D. B., and Hollfelder, F. (2015). Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nature Communications*, 6:10008.

Cormack, B. P., Valdivia, R. H., and Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 173(1):33–38.

Costantini, L. M., Fossati, M., Francolini, M., and Snapp, E. L. (2012). Assessing the Tendency of Fluorescent Proteins to Oligomerize Under Physiologic Conditions. *Traffic*, 13(5):643–649.

Crameri, A., Whitehorn, E. A., Tate, E., and Stemmer, W. (1996). Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnology*, 14(3):315–319.

Crivat, G. and Taraska, J. W. (2012). Imaging proteins inside cells with fluorescent tags. *Trends in Biotechnology*, 30(1):8–16.

Crooks, G. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6):1188–1190.

Daggett, K. A., Layer, M., and Cropp, T. A. (2009). A General Method for Scanning Unnatural Amino Acid Mutagenesis. *ACS Chemical Biology*, 4(2):109–113.

Day, R. N. and Davidson, M. W. (2009). The fluorescent protein palette: Tools for cellular imaging. *Chemical Society Reviews*, 38(10):2887–2921.

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., Roux, A.-F., Smith, T., Antonarakis, S. E., and Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, 37(6):564–569.

Deng, Z., Huang, W., Bakkalbasi, E., Brown, N. G., Adamski, C. J., Rice, K., Muzny, D., Gibbs, R. A., and Palzkill, T. (2012). Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *Journal of Molecular Biology*, 424(3-4):150–167.

DePristo, M. A., Weinreich, D. M., and Hartl, D. L. (2005). Missense meanderings in sequence space: A biophysical view of protein evolution. *Nature Reviews Genetics*, 6(9):678–687.

der Auwera, G. A. V., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 43(1):11.10.1–11.10.33.

Diamante, L., Gatti-Lafranconi, P., Schaerli, Y., and Hollfelder, F. (2013). In vitro affinity screening of protein and peptide binders by megavalent bead surface display. *Protein Engineering, Design and Selection*, 26(10):713–724.

Diss, G. and Lehner, B. (2018). The genetic landscape of a physical interaction. *eLife*, 7:e32472.

Doud, M. B. and Bloom, J. D. (2016). Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*, 8(6):155.

Emond, S., Petek, M., Kay, E. J., Heames, B., Devenish, S. R. A., Tokuriki, N., and Hollfelder, F. (2020). Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nature Communications*, 11(1):3469.

Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31(6):1581–1592.

Fischlechner, M., Schaerli, Y., Mohamed, M. F., Patil, S., Abell, C., and Hollfelder, F. (2014). Evolution of enzyme catalysts caged in biomimetic gel-shell beads. *Nature Chemistry*, 6(9):791–6.

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. *Nature Methods*, 7(9):741–6.

Fowler, D. M., Araya, C. L., Gerard, W., and Fields, S. (2011). Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*, 27(24):3430–3431.

Frenz, L., Blank, K., Brouzes, E., and Griffiths, A. D. (2009). Reliable microfluidic on-chip incubation of droplets in delay-lines. *Lab on a Chip*, 9(10):1344–1348.

Fujii, R., Kitaoka, M., and Hayashi, K. (2006). RAISE: A simple and novel method of generating random insertion and deletion mutations. *Nucleic Acids Research*, 34(4):e30.

Ganini, D., Leinisch, F., Kumar, A., Jiang, J. J., Tokar, E. J., Malone, C. C., Petrovich, R. M., and Mason, R. P. (2017). Fluorescent proteins such as eGFP lead to catalytic oxidative stress in cells. *Redox Biology*, 12(March):462–468.

Garai, Á., Zeke, A., Gógl, G., Törő, I., Fördős, F., Blankenburg, H., Bárkai, T., Varga, J., Alexa, A., Emig, D., Albrecht, M., and Reményi, A. (2012). Specificity of Linear Motifs That Bind to a Common Mitogen-Activated Protein Kinase Docking Groove. *Science Signaling*, 5(245):ra74–ra74.

García-Fruitós, E., González-Montalbán, N., Morell, M., Vera, A., Ferraz, R. M., Arís, A., Ventura, S., and Villaverde, A. (2005). Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microbial Cell Factories*, 4:1–6.

Gerth, M. L., Patrick, W. M., and Lutz, S. (2004). A second-generation system for unbiased reading frame selection. *Protein Engineering, Design and Selection*, 17(7):595–602.

Gielen, F., Hours, R., Emond, S., Fischlechner, M., Schell, U., and Hollfelder, F. (2016). Ultrahigh-throughput–directed enzyme evolution by absorbance-activated droplet sorting (AADS). *Proceedings of the National Academy of Sciences*, 113(47):E7383–E7389.

Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R. L., Aharoni, A., Silman, I., Sussman, J. L., Tawfik, D. S., and Fleishman, S. J. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell*, 63(2):337–346.

Gonzalez, C. E., Roberts, P., and Ostermeier, M. (2019). Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β-Lactamase. *Journal of Molecular Biology*, 431(12):2320–2330.

Gordley, R. M., Bugaj, L. J., and Lim, W. A. (2016). Modular engineering of cellular signaling proteins and networks. *Current Opinion in Structural Biology*, 39:106–114.

Gray, V. E., Hause, R. J., and Fowler, D. M. (2017). Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics*, 207(1):53–61.

Griffiths, A. D. and Tawfik, D. S. (2003). Directed evolution of an extremely fast phosphotriesterase by in vitro compartmentalization. *The EMBO Journal*, 22(1):24–35.

Grimsley, J. K., Scholtz, J. M., Pace, C. N., and Wild, J. R. (1997). Organophosphorus Hydrolase Is a Remarkably Stable Enzyme That Unfolds through a Homodimeric Intermediate. *Biochemistry*, 36(47):14366–14374.

Grishin, N. V. (2001). Fold Change in Evolution of Protein Structures. *Journal of Structural Biology*, 2001(134):167–185.

Guo, M. T., Rotem, A., Heyman, J. A., and Weitz, D. A. (2012). Droplet microfluidics for high-throughput biological assays. *Lab on a Chip*, 12(12):2146–2155.

Gupta, R. D. and Tawfik, D. S. (2008). Directed enzyme evolution via small and effective neutral drift libraries. *Nature Methods*, 5(11):939–942.

Haapa-Paananen, S., Rita, H., and Savilahti, H. (2002). DNA Transposition of Bacteriophage Mu. *Journal of Biological Chemistry*, 277(4):2843–2851.

Hallet, B., Sherratt, D. J., and Hayes, F. (1997). Pentapeptide scanning mutagenesis: Random insertion of a variable five amino acid cassette in a target protein. *Nucleic Acids Research*, 25(9):1866–1867.

Harms, M. J. and Thornton, J. W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology*, 20(3):360–366.

Hashimoto, K. and Panchenko, A. R. (2010). Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proceedings of the National Academy of Sciences*, 107(47):20352–20357.

Hayes, F. and Hallet, B. (2000). Pentapeptide scanning mutagenesis: Encouraging old proteins to execute unusual tricks. *Trends in Microbiology*, 8(12):571–577.

Heim, R., Cubitt, A. B., and Tsien, R. Y. (1995). Improved green fluorescence. *Nature*, 373(6516):663–664.

Heinz, D. W., Baase, W. A., Dahlquist, F. W., and Matthews, B. W. (1993). How amino-acid insertions are allowed in an α-helix of T4 lysozyme | Nature. *Nature*, 361:561–564.

Hietpas, R. T., Jensen, J. D., and a Bolon, D. N. (2011). Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, 108(19):7896–7901.

Hoque, M. A., Zhang, Y., Chen, L., Yang, G., Khatun, M. A., Chen, H., Hao, L., and Feng, Y. (2017). Stepwise Loop Insertion Strategy for Active Site Remodeling to Generate Novel Enzyme Functions. *ACS Chemical Biology*, 12(5):1188–1193.

Jacobs, D., Glossip, D., Xing, H., Muslin, A. J., and Kornfeld, K. (1999). Multiple docking sites on substrate proteins form a modular system that mediates recognition by ERK MAP kinase. *Genes & Development*, 13(2):163–175.

Jones, D. D. (2005). Triplet nucleotide removal at random positions in a target gene: The tolerance of TEM-1 β-lactamase to an amino acid deletion. *Nucleic Acids Research*, 33(9):e80–e80.

Kaltenbach, M., Jackson, C. J., Campbell, E. C., Hollfelder, F., and Tokuriki, N. (2015). Reverse evolution leads to genotypic incompatibility despite functional and active site convergence. *eLife*, 4:e06492.

Kim, H. K. and Kaang, B. K. (1998). Truncated green fluorescent protein mutants and their expression in Aplysia neurons. *Brain Research Bulletin*, 47(1):35–41.

Kim, Y. C., Lee, H. S., Yoon, S., and Morrison, S. L. (2009). Transposon-directed base-exchange mutagenesis (TDEM): A novel method for multiple-nucleotide substitutions within a target gene. *BioTechniques*, 46(7):534–542.

Kintses, B., Hein, C., Mohamed, M. F., Fischlechner, M., Courtois, F., Lainé, C., and Hollfelder, F. (2012). Picoliter Cell Lysate Assays in Microfluidic Droplet Compartments for Directed Enzyme Evolution. *Chemistry & Biology*, 19(8):1001–1009.

Kondrashov, D. A. and Kondrashov, F. A. (2015). Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*, 31(1):24–33.

Kvikstad, E. M., Tyekucheva, S., Chiaromonte, F., and Makova, K. D. (2007). A Macaque's-Eye View of Human Insertions and Deletions: Differences in Mechanisms. *PLOS Computational Biology*, 3(9):e176.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8):323–331.

Leifson, E. and Hugh, R. (1954). A New Type of Polar Monotrichous Flagellation. *Microbiology,*, 10(1):68–70.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J., and Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., and Guo, J.-t. (2017). Effects of short indels on protein structure and function in human genomes. *Scientific Reports*, 7(1):1–9.

Lindenburg, L., Huovinen, T., van de Wiel, K., Herger, M., Snaith, M. R., and Hollfelder, F. (2020). Split & mix assembly of DNA libraries for ultrahigh throughput on-bead screening of functional proteins. *Nucleic Acids Research*, 48(11):e63–e63.

Liu, J. and Cropp, T. A. (2012). A method for multi-codon scanning mutagenesis of proteins based on asymmetric transposons. *Protein Engineering, Design & Selection*, 25(2):67–72.

Liu, S.-s., Wei, X., Dong, X., Xu, L., Liu, J., and Jiang, B. (2015). Structural plasticity of green fluorescent protein to amino acid deletions and fluorescence rescue by folding-enhancing mutations. *BMC Biochemistry*, 16(1):17.

Liu, S.-s., Wei, X., Ji, Q., Xin, X., Jiang, B., and Liu, J. (2016). A facile and efficient transposon mutagenesis method for generation of multi-codon deletions in protein sequences. *Journal of Biotechnology*, 227:27–34.

MacBeath, G., Kast, P., and Hilvert, D. (1998). Redesigning Enzyme Topology by Directed Evolution. *Science*, 279(5358):1958–1961.

Mankowska, S. A., Gatti-Lafranconi, P., Chodorge, M., Sridharan, S., Minter, R. R., and Hollfelder, F. (2016). A Shorter Route to Antibody Binders via Quantitative in vitro Bead-Display Screening and Consensus Analysis. *Scientific Reports*, 6(1):1–11.

Mathonet, P., Deherve, J., Soumillion, P., and Fastrez, J. (2006). Active TEM-1 β-lactamase mutants with random peptides inserted in three contiguous surface loops. *Protein Science*, 15(10):2323–2334.

Matuszewski, S., Hildebrandt, M. E., Ghenu, A.-H., Jensen, J. D., and Bank, C. (2016). A Statistical Guide to the Design of Deep Mutational Scanning Experiments. *Genetics*, 204(1):77–87.

Maynard Smith, J. (1970). Natural Selection and the Concept of a Protein Space. *Nature*, 225(5232):563–564.

McDonald, J. C., Duffy, D. C., Anderson, J. R., Chiu, D. T., Wu, H., Schueller, O. J., and Whitesides, G. M. (2000). Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis*, 21(1):27–40.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. (2014). Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):e112–e112.

Morelli, A., Cabezas, Y., Mills, L. J., and Seelig, B. (2017). Extensive libraries of gene truncation variants generated by in vitro transposition. *Nucleic Acids Research*, 45(10):gkx030.

Mullaney, J. M., Mills, R. E., Pittard, W. S., and Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, 19(R2):R131–R136.

Murakami, H., Hohsaka, T., and Sisido, M. (2002). Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nature Biotechnology*, 20(January):76–81.

Nagasundarapandian, S., Merkel, L., Budisa, N., Govindan, R., Ayyadurai, N., Sriram, S., Yun, H., and Lee, S. G. (2010). Engineering protein sequence composition for folding robustness renders efficient noncanonical amino acid incorporations. *ChemBioChem*, 11(18):2521–2524.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Neuenfeldt, A., Just, A., Betat, H., and Mörl, M. (2008). Evolution of tRNA nucleotidyltransferases: A small deletion generated CC-adding enzymes. *Proceedings of the National Academy of Sciences*, 105(23):7953–7958.

Odokonyero, D., Sakai, A., Patskovsky, Y., Malashkevich, V. N., Fedorov, A. A., Bonanno, J. B., Fedorov, E. V., Toro, R., Agarwal, R., Wang, C., Ozerova, N. D. S., Yew, W. S., Sauder, J. M., Swaminathan, S., Burley, S. K., Almo, S. C., and Glasner, M. E. (2014). Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. *Proceedings of the National Academy of Sciences*, 111(23):8535–8540.

Olson, C. A., Wu, N. C., and Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651.

O'Neil, K. T., Bach, A. C., and DeGrado, W. F. (2000). Structural consequences of an amino acid deletion in the B1 domain of protein G. *Proteins: Structure, Function, and Bioinformatics*, 41(3):323–333.

Osuna, J., Yáñez, J., Soberón, X., and Gaytán, P. (2004). Protein evolution by codon-based random deletions. *Nucleic Acids Research*, 32(17):e136–e136.

Palzkill, T. (2018). Structural and Mechanistic Basis for Extended-Spectrum Drug-Resistance Mutations in Altering the Specificity of TEM, CTX-M, and KPC β-lactamases. *Frontiers in Molecular Biosciences*, 5.

Pande, J., Szewczyk, M. M., and Grover, A. K. (2010). Phage display: Concept, innovations, applications and future. *Biotechnology Advances*, 28(6):849–858.

Park, H.-S., Nam, S.-H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J., and Kim, H.-S. (2006). Design and Evolution of New Catalytic Activity with an Existing Protein Scaffold. *Science*, 311(5760):535–538.

Pascarella, S. and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *Journal of Molecular Biology*, 224(2):461–471.

Pedelacq, J.-D. and Cabantous, S. (2019). Development and Applications of Superfolder and Split Fluorescent Protein Detection Systems in Biology. *International Journal of Molecular Sciences*, 20(14):3479.

Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006). Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology*, 24(1):79–88.

Pikkemaat, M. G. and Janssen, D. B. (2002). Generating segmental mutations in haloalkane dehalogenase: A novel part in the directed evolution toolbox. *Nucleic Acids Research*, 30(8):e35.

Podgornaia, A. I. and Laub, M. T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677.

Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M., and Tans, S. J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386.

Polyanovsky, V. O., Roytberg, M. A., and Tumanyan, V. G. (2011). Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms for Molecular Biology*, 6(1):25.

Poole, E. S., Brown, C. M., and Tate, W. P. (1995). The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *The EMBO Journal*, 14(1):151–8.

Raman, A. S., White, K. I., and Ranganathan, R. (2016). Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell*, 166(2):468–481.

Reetz, M. T., Höbenreich, H., Soni, P., and Fernández, L. (2008). A genetic selection system for evolving enantioselectivity of enzymes. *Chemical Communications*, pages 5502–5504.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277.

Rodems, S. M., Hamman, B. D., Lin, C., Zhao, J., Shah, S., Adams, D. J., Makings, L., Stack, J. H., and Pollok, B. A. (2004). A FRET-Based Assay Platform for Ultra-High Density Drug Screening of Protein Kinases and Phosphatases |. *ASSAY and Drug Development Technologies*, 1(1).

Rollins, N. J., Brock, K. P., Poelwijk, F. J., Stiffler, M. A., Gauthier, N. P., Sander, C., and Marks, D. S. (2019). Inferring protein 3D structure from deep mutation scans. *Nature Genetics*, 51(7):1170–1176.

Romero, P. A. and Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews. Molecular cell biology*, 10(12):866–76.

Romero, P. A., Tran, T. M., and Abate, A. R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164.

Roodveldt, C. and Tawfik, D. S. (2005). Shared Promiscuous Activities and Evolutionary Features in Various Members of the Amidohydrolase Superfamily. *Biochemistry*, 44(38):12728–12736.

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. A. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology*, 425(8):1363–1377.

Royant, A. and Noirclerc-Savoye, M. (2011). Stabilizing role of glutamic acid 222 in the structure of Enhanced Green Fluorescent Protein. *Journal of Structural Biology*, 174(2):385–390.

Rubin, A. F., Gelman, H., Lucas, N., Bajjalieh, S. M., Papenfuss, A. T., Speed, T. P., and Fowler, D. M. (2017). A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, 18(1):1–15.

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., Lukyanov, K. A., and Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):125.

Schmiedel, J. M. and Lehner, B. (2019). Determining protein structures using deep mutagenesis. *Nature Genetics*, 51(7):1177–1186.

Schwartz, J. J., Lee, C., and Shendure, J. (2012). Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nature Methods*, 9(9):913–915.

Sciambi, A. and Abate, A. R. (2014). Accurate microfluidic sorting of droplets at 30 kHz. *Lab Chip*, 15(1):47–51.

Shcherbo, D., Murphy, C. S., Ermakova, G. V., Solovieva, E. A., Chepurnykh, T. V., Shcheglov, A. S., Verkhusha, V. V., Pletnev, V. Z., Hazelwood, K. L., Roche, P. M., Lukyanov, S., Zaraisky, A. G., Davidson, M. W., and Chudakov, D. M. (2009). Far-red fluorescent tags for protein imaging in living tissues. *Biochemical Journal*, 418(3):567–574.

Shembekar, N., Chaipan, C., Utharala, R., and Merten, C. A. (2016). Droplet-based microfluidics in drug discovery, transcriptomics and high-throughput molecular genetics. *Lab on a Chip*, 16(8):1314–1331.

Shortle, D. and Sondek, J. (1995). The emerging role of insertions and deletions in protein engineering. *Current Opinion in Biotechnology*, 6(4):387–393.

Simm, A. M., Baldwin, A. J., Busse, K., and Jones, D. D. (2007). Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β-lactamase. *FEBS Letters*, 581(21):3904–3908.

Starita, L. M. and Fields, S. (2015a). Deep Mutational Scanning: A Highly Parallel Method to Measure the Effects of Mutation on Protein Function. *Cold Spring Harbor Protocols*, 2015(8):pdb.top077503.

Starita, L. M. and Fields, S. (2015b). Deep Mutational Scanning: Calculating Enrichment Scores for Protein Variants from DNA Sequencing Output Files. *Cold Spring Harbor Protocols*, 2015(8):781–783.

Starita, L. M. and Fields, S. (2015c). Deep Mutational Scanning: Library Construction, Functional Selection, and High-Throughput Sequencing. *Cold Spring Harbor Protocols*, 2015(8):pdb.prot085225.

Starita, L. M., Pruneda, J. N., Lo, R. S., Fowler, D. M., Kim, H. J., Hiatt, J. B., Shendure, J., Brzovic, P. S., Fields, S., and Klevit, R. E. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–72.

Starr, T. N. and Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, 25:1204–1218.

Stott, K. M., Yusof, A. M., Perham, R. N., and Jones, D. D. (2009). A Surface Loop Directs Conformational Switching of a Lipoyl Domain Between a Folded and a Novel Misfolded Structure. *Structure*, 17(8):1117–1127.

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: Detecting beneficial and pathogenic changes. *The Biochemical journal*, 449:581–94.

Tabasinezhad, M., Talebkhan, Y., Wenzel, W., Rahimi, H., Omidinia, E., and Mahboudi, F. (2019). Trends in therapeutic antibody affinity maturation: From in-vitro towards next-generation sequencing approaches. *Immunology Letters*, 212:106–113.

Taylor, M. S., Ponting, C. P., and Copley, R. R. (2004). Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. *Genome Research*, 14(4):555–566.

Taylor, W. R. and Sadowski, M. I. (2011). Structural constraints on the covariance matrix derived from multiple aligned protein sequences. *PLoS ONE*, 6(12):e28265.

Teşileanu, T., Colwell, L. J., and Leibler, S. (2015). Protein sectors: Statistical coupling analysis versus conservation. *PLOS Computational Biology*, 11(2):e1004091.

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J.-Q. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, 455(7209):105–8.

Tokuriki, N., Jackson, C. J., Afriat-Jurnou, L., Wyganowski, K. T., Tang, R., and Tawfik, D. S. (2012). Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nature Communications*, 3(1257):1–10.

Tóth-Petróczy, Á. and Tawfik, D. S. (2013). Protein insertions and deletions enabled by neutral roaming in sequence space. *Molecular Biology and Evolution*, 30(4):761–771.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233.

Tsien, R. Y. (1998). The Green Fluorescent Protein. *Annual Review of Biochemistry*, 67(1):509–544.

Weinreich, D. M., Delaney, N. F., DePristo, M. A., and Hartl, D. L. (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, 312(5770):111–114.

Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23(6):700–707.

Wijma, H. J., Fürst, M. J. L. J., and Janssen, D. B. (2018). A Computational Library Design Protocol for Rapid Improvement of Protein Stability: FRESCO. In Bornscheuer, U. T. and Höhne, M., editors, *Protein Engineering: Methods and Protocols*, Methods in Molecular Biology, pages 69–85. Springer, New York, NY.

Wrenbeck, E. E., Azouz, L. R., and Whitehead, T. A. (2017). Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature Communications*, 8:15695.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366.

Wroe, R., Chan, H. S., and Bornberg-Bauer, E. (2007). A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP Journal*, 1(1):79–87.

Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O., and Sun, R. (2016). Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965.

Wyganowski, K. T., Kaltenbach, M., and Tokuriki, N. (2013). GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates. *Journal of Molecular Biology*, 425(18):3403–3414.

Yanagida, H., Matsuura, T., and Yomo, T. (2010). Ribosome Display for Rapid Protein Evolution by Consecutive Rounds of Mutation and Selection. In Braman, J., editor, *In Vitro Mutagenesis Protocols: Third Edition*, pages 257–267. Humana Press, Totowa, NJ.

Yang, G., Anderson, D. W., Baier, F., Dohmen, E., Hong, N., Carr, P. D., Kamerlin, S. C. L., Jackson, C. J., Bornberg-Bauer, E., and Tokuriki, N. (2019). Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nature Chemical Biology*, 15(11):1120–1128.

Yang, G. and Withers, S. G. (2009). Ultrahigh-Throughput FACS-Based Screening for Directed Enzyme Evolution. *ChemBioChem*, 10(17):2704–2715.

Yang, T. T., Cheng, L., and Kain, S. R. (1996). Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein. *Nucleic Acids Research*, 24(22):4592–4593.

Yoo, J. I., Daugherty, P. S., and O'Malley, M. A. (2020). Bridging non-overlapping reads illuminates high-order epistasis between distal protein sites in a GPCR. *Nature Communications*, 11(1):690.

Zahnd, C., Amstutz, P., and Plückthun, A. (2007). Ribosome display: Selecting and evolving proteins in vitro that specifically bind to a target. *Nature Methods*, 4(3):269–279.

Zakas, P. M., Brown, H. C., Knight, K., Meeks, S. L., Trent Spencer, H., Gaucher, E. A., and Doering, C. B. (2016). Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nature Biotechnology*, 35(1).

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620.

Zinchenko, A., Devenish, S. R. A., Kintses, B., Colin, P. Y., Fischlechner, M., and Hollfelder, F. (2014). One in a million: Flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Analytical Chemistry*, 86(5):2526–2533.