Developing a nanopore sequencing workflow for protein engineering applications

Daniel Zhi Jun Tan  
National University of Singapore  https://orcid.org/0000-0003-1072-4304

Vincent Fung  
National University of Singapore  https://orcid.org/0000-0002-2124-6232

Tao Sun  
National University of Singapore

Kaiyuan Tian  
National University of Singapore

Zhi Li  
National University of Singapore

Kang Zhou (chezkg@nus.edu.sg)  
National University of Singapore  https://orcid.org/0000-0003-0606-7030

Research Article

Keywords: Nanopore sequencing, protein engineering, Escherichia coli, growth-based selection, secondary alcohol dehydrogenase, site-saturated mutagenesis

Posted Date: June 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3002995/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Sequencing plays a critical role in protein engineering, where the genetic information encoding for a desired mutation can be identified. We evaluated the performance of two commercially available NGS technologies (Illumina NGS and nanopore sequencing) on the available mutant libraries that were either previously constructed for other protein engineering projects or were constructed in-house for this study. The sequencing results from Illumina sequencing indicated that a substantial proportion of the reads exhibited strand exchange, which mixed information of different mutants. When nanopore sequencing was used, the occurrence of strand exchange was substantially reduced compared with Illumina sequencing. We then developed a new library preparation workflow for nanopore sequencing and was successful in further reducing the incidence of strand exchange. The optimized workflow was successfully used to aid selection of improved alcohol dehydrogenase mutants in cells where their activities were coupled with cell growth rate. The workflow quantified the enrichment fold change of most mutants in the library (size = 1,728) in the growth-based selection passaging. A mutant that was > 500% more active than its parent variant was identified based on the fold change data, but not with the absolute abundance data (random sampling of the passaged cells), highlighting the usefulness of this rapid and affordable sequencing workflow in protein engineering.

1 Introduction

Sequencing cost has been drastically reduced over the years by the new NGS technologies that have been gradually replacing Sanger sequencing, which is one of the most well-known sequencing methodologies and has been adopted by many researchers to study genetic information of DNA samples. However, it is the nature of Sanger sequencing that forbids multiplexing of different samples. This is because it would not be possible to differentiate which sample is associated with each nucleotide peak of the chromatogram. As a result, researchers who used Sanger sequencing in their protein engineering workflow can only sequence plasmids extracted from individual bacterial colonies. The data set size may not be large since it would be prohibitively expensive to pick many colonies for plasmid extraction and sequencing.

An alternative to Sanger sequencing would be Illumina NGS. At an affordable cost, millions of individual DNA reads can be sequenced. The immobilized DNA molecules were amplified to form clusters that are visible under fluorescence microscopes. The sequences of a large number of clusters could be determined simultaneously when they were used as template to synthesize complement strands. The technology could be used to identify mutations occurring in individual colonies after their materials are barcoded and pooled. Although Illumina NGS is about 38 times more expensive than Sanger sequencing per reaction at the time of writing, it is more cost-effective than Sanger sequencing, because Illumina NGS can allow multiplexing of many samples per reaction.

Several considerations must be made when planning to use Illumina NGS. Firstly, the sequencing read length is usually restricted to 250 bp while two reads may be paired, extending the coverage to 500 bp,
which is still much shorter than many coding sequences. Therefore, its general usability is reduced especially if the intended mutation locations are far apart and out of the restricted range. Within the past 10 years, a new generation of sequencing known as nanopore sequencing was commercialized. The method determines sequence of a DNA molecule when it passes through a porin in polarized membrane, based on the unique perturbation pattern to electric current across the membrane caused by the nucleobases in the porin. Although the nanopore individual sequencing accuracy per nucleotide (Qscore \( \sim 10 \) [accuracy: \( \sim 90\% \)]) is much lower than Illumina NGS (Qscore > 30 [accuracy: >99.9%]), nanopore sequencing should be able to provide reliable readout for protein engineering when enough coverage is achieved. One obvious advantage of nanopore sequencing is its long read length (easily > 5,000 bp), which can cover most coding sequences.\(^2,3\) Another critical advantage is that nanopore sequencing can be conducted in most standard research labs because of its affordable device and consumables costs, which enables very short turnover time (24 h) and its use in guiding protein engineering in a real-time fashion.\(^4–8\) As a comparison, it takes \( \sim 3 \) weeks to receive Illumina sequencing results in this study.

In this study, Illumina NGS sequencing service by commercial providers were evaluated in the beginning, but the extent of strand exchange between multiplexed samples were high (22%), which reduce the data reliability when results from paired reads are integrated. Standard nanopore sequencing workflow reduced the extent of strand exchange to 1–2%, and a new workflow based on chemical cleavage of phosphorothionate bond further reduced the occurrence probability of strand exchange extent to 0.3%. We subsequently used the new workflow and growth-based selection pressure to engineer a model protein (CpSADH) to be 520% more active than the wildtype when a non-preferred enantiomeric alcohol was used as the substrate.

2 Materials and methods

2.1 Chemicals and Materials

Complete Bacterial Protein Extraction Reagent (B-PER), the GeneJET Plasmid MiniPrep Kit for plasmid extraction and GeneJET gel extraction kit for gel extraction were purchased from Thermo Fisher Scientific. D-glucose and M9 medium broth powder were purchased from Bio Basic Asia Pacific Pte Ltd. 10X phosphate buffered saline solution (PBS) was purchased from Axil Scientific Pte Ltd. The Super Optimal broth with Catabolite repression (SOC) Outgrowth Medium was purchased from New England Biolabs. All other chemicals were purchased from Sigma Aldrich unless stated otherwise. Commercially available reagents were used as received without purification. All aqueous solutions were prepared using ultrapure water. All the oligonucleotides used in this study were synthesized by Integrated DNA Technologies.

2.2 Mutant library preparation

Mutant library \( Q \) (CpSADH L55NNK/W116NNK/F285NNK/W286NNK), library \( R \) (CpSADH L55NDT/F285NDT/W286NDT) and library \( K \) (CpSADH L55NNK/L262NNK/F285NNK/W286NNK) were
constructed in a previous study (unpublished). Library M (CpSADH L55WTC/C57KGC/T158ASC) and library ML03 (CpSADH L55NDT/C57NDT/T158NDT) were constructed for this study as described below.

### Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Plasmids</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV02</td>
<td>pDV02 (pMB1-Spect&lt;sup&gt;R&lt;/sup&gt;-P&lt;sub&gt;gyrA&lt;/sub&gt;-CpSADH)</td>
</tr>
<tr>
<td>Library Q</td>
<td>pQ, built from pDV02 (L55NNK/W116NNK/F285NNK/W286NNK)</td>
</tr>
<tr>
<td>Library R</td>
<td>pR, built from pDV02 (L55NDT/F285NDT/W286NDT)</td>
</tr>
<tr>
<td>Library K</td>
<td>pK, built from pDV02 (L55NNK/L262NNK/F285NNK/W286NNK)</td>
</tr>
<tr>
<td>Library ML03</td>
<td>pML03, built from pDV02 (L55NDT/C57NDT/T158NDT)</td>
</tr>
<tr>
<td>Library M</td>
<td>pM, built from pDV02 (L55WTC/C57KGC/T158ASC)</td>
</tr>
</tbody>
</table>

Spect<sup>R</sup>: spectinomycin resistance. The working concentration of spectinomycin in cell cultures was 50 µg/ml. gfp: green fluorescent protein, used as a control for CpSADH (secondary alcohol NAD-dependent dehydrogenase).

The gene encoding the *Candida parapsilosis* secondary alcohol dehydrogenase (CpSADH) was amplified using Q5 High-Fidelity DNA Polymerase (New England Biolabs). The cycling condition is as follows: 98°C for 30 s, 35 × (98°C for 30 s, 55°C for 30 s, 72°C for 30 s), 72°C for 120 s, held at 4°C. The PCR product was then assembled with a plasmid backbone containing PgyrA promoter, a terminator, pMB1 origin and spectinomycin resistance, based on the Guanine-Thymine standard. The resulting plasmid was named as pDV02 (Table 1). The mutant libraries were constructed via site saturation mutagenesis with primers encoding degenerate codons for the targeted mutation sites (refer to Table 1). For example, library Q was constructed with primers encoding for NNK targeting L55, W116, F285 and W286 (oligonucleotide information are listed in Table S1). *E. coli DH5α* (C2987H, New England Biolabs) was used as a cloning host. DH5α was transformed using the heat-shock method based on the manufacturer’s instruction. Transformants were plated on Luria Bertani (LB) agar with appropriate antibiotics (Table 1) after incubation at 37°C overnight. The resultant cells were washed off from the agar plate using 0.22 µm filtered 8.5 g/L NaCl solution and transferred into 2 mL Eppendorf tubes. The cells were pelleted by centrifugation at 14,000 rpm for 1 min. Their plasmids were extracted using the GeneJET Plasmid MiniPrep Kit (K0503, Thermo Fisher Scientific).

### 2.3 Illumina NGS and nanopore sequencing by commercial providers

Illumina NGS was performed for 6 samples from mutant library Q, while nanopore sequencing was performed for 4 samples from mutant libraries Q and R respectively. The library preparation workflow for Illumina and nanopore sequencing is as follows: The mutant library plasmids (~ 10 ng/ul) were used as
the DNA template for PCR. The PCR was performed in 0.2 mL tubes (AB1182: Thermo Fisher Scientific). Each 50 µL reaction was prepared by mixing 1 µL of DNA template, 0.3 µL of 100 µM forward primer solution, 0.3 µL of 100 µM reverse primer and 25 µL of 2X Q5 High-Fidelity DNA Polymerase Master Mix (New England Biolabs). PCR was conducted in individual PCR tubes for individual libraries using a unique pair of oligos that contain unique forward and reverse barcodes. The thermal cycler setting follows the protocol from New England Biolabs and is as follows: 98°C for 30 s, 35 × (98°C for 30 s, 55°C for 30 s, 72°C for 24 s), 72°C for 120 s, held at 4°C. The forward primers used in the PCR are barcoded oligos containing a 10 nucleotides long unique sequence attached to the 5’ end and a 17 nucleotides long binding region at the 3’ end. The reverse primers contain a 10 nucleotides long unique sequence at the 5’ end and a 21 nucleotides long binding region at the 3’ end. The binding region sequences were designed to ensure that the amplified region covers the mutation sites. Once the PCR has been completed, the amplicons were purified with gel electrophoresis using a 0.8% agarose gel with SYBR™ Safe DNA Gel Stain (S33102: Thermo Fisher Scientific). The PCR amplicons were then extracted from the gel and purified with the GeneJET Gel Extraction Kit. The purified PCR amplicons were normalised to the same concentration of 100 ng/µL and appropriate volumes of PCR amplicons were mixed to achieve the minimum DNA quantity required by Illumina NGS and nanopore sequencing. The samples were then sent to GENEWIZ Suzhou and Genomics Institute of Singapore for Illumina NGS (Novaseq PE250, 1 million reads/sample) and nanopore sequencing (GridION flow cell) respectively (Fig. 1). The sequencing results were analysed by our in-house algorithms.

2.4 In-house Nanopore sequencing

Mutant libraries K, M and ML03 were sequenced with the in-house nanopore device that was purchased from ONT. Mutant library K (three multiplexed samples) was sequenced using the original nanopore workflow, while mutant libraries M (four multiplexed samples) and ML03 (five multiplexed samples) were sequenced with the new nanopore workflow (Fig. 2). The general library preparation workflow for these mutant libraries were as follows: the cell cultures in either selection medium or non-selection medium (i.e., LB medium) were first harvested by centrifugation at 3,500 rpm for 10 mins. The supernatant was removed to obtain the cell pellets. The cell pellets were then resuspended in 1 mL of ultrapure H₂O and their optical density (OD₆₀₀) were measured and normalised to 0.1. The cell suspensions (OD₆₀₀ = 0.1) were used as the DNA template for PCR. The PCR was performed in 0.2 mL tubes (AB1182: Thermo Fisher Scientific). Each 50 µL reaction was prepared by mixing 1 µL of DNA template, 0.3 µL of 100 µM forward primer solution, 0.3 µL of 100 µM reverse primer and 1X Q5 High-Fidelity DNA Polymerase (New England Biolabs). The thermal cycler setting is as follows: 98°C for 8 min, 35 × (98°C for 30 s, 55°C for 30 s, 72°C for 24 s), 72°C for 120 s, held at 4°C.

The forward primers used for mutant libraries M and ML03 had the following structure: T*[Barcode] [Binding], where * indicates phosphorothioate (PS) linkage; [Barcode] indicates 10-nt barcode sequence; [Binding] indicates the 17-nt binding sequence. The forward primer for library K had the same structure except that it did not have T* at its 5’ end. Each reverse primer contained a 10-nt unique sequence at the 5’ end and a 21 nucleotides long binding region at the 3’ end.
The PCR amplicons from mutant library K was prepared using the ligation sequencing kit (SQK-LSK110) in accordance with ONT’s protocol (ACDE_9110_v110_revO_10Nov2020). The PCR amplicons from mutant libraries M and ML03 contained the PS bonded thymine attached to the 5’ end. This is to facilitate the generation of the 3’ adenine without the usage of T4 DNA polymerase. Since only the top DNA strand contains the 3’ adenine overhang, only the top strand will be sequenced. To cleave the PS bond, the iodine treatment was performed with the following steps: in a 1.5 mL Eppendorf microtube, 25 µL of purified amplicon with 5.5 µL of 1M Tris HCl buffer (pH 9.0) and 10 µL of 30 g/L iodine solution (solvent: ethanol) were mixed. The tube was incubated at 70°C in a thermomixer for 5 mins. Subsequently, 350 µL of the binding buffer (from GeneJET Gel Extraction Kit) and 250 µL of ultrapure H₂O were added. They were then purified with GeneJet spin columns. After the iodine treatment, the amplicons were ligated with 2.5 µL of the sequencing adaptors (AMX-F from ligation sequencing kit [SQK-LSK110: ONT]) by using the Blunt/TA ligase master mix (M0367S: NEB). The ligation reaction was incubated for one hour at 4°C. The amplicons were then purified using the AMPure XP beads as specified in the “Adapter ligation and clean-up” section from the protocol provided by ONT (ACDE_9110_v110_revO_10Nov2020). The prepared library sample was loaded onto a flongle flow cell (FLO-FLG001: ONT) as instructed by ONT for sequencing. The nanopore sequencing was performed using the MinION™ MK1B device connected to a Windows desktop with the MinKNOW™ software (version: 22.03.5) installed. The samples were sequenced until 1 Gb of raw data was obtained in the form of fast5 files or the sequencing activity stopped. The raw sequencing data containing read signals were converted to base sequences using a basecalling algorithm, Guppy (version: 6.2.1) with the super accuracy basecaller model (SUP). These base sequences were then written and compressed into FASTQ.gz files for further data analysis in MATLAB.

The MinION flow cell (FLO-MIN106D) was used to sequence mutant library K, while the flongle (FLO-FLG001) was used for libraries M and ML03. They contain the same type of nanopore (R9.4.1) and differ by the number of pores in the flow cell (FLO-MIN106D: at least 800 pores versus FLO-FLG001: at least 50 pores).

2.5 Post-sequencing analysis algorithm

The base sequences were analysed using a MATLAB App developed in house. The raw reads from the FASTQ.gz files contained the base sequences and their corresponding quality (Phred) score in ASCII characters (ASCII_BASE 33 format). Prior to de-multiplexing, the reads were first screened based on their sequence length. Any read longer than 500 bases was retained for further analysis. Demultiplexing was done aligning each read sequence with the library barcodes using the Smith-Waterman algorithm. This is done by first aligning the forward barcode oligo binding sequence with the first 100 bases of each read. A read is accepted for further sorting if the alignment score is greater than 0.6. Next, 15 bases before the oligo binding sequence were aligned with the forward barcodes. A read is successfully sorted into the corresponding library if both the barcode coverage (the number of perfectly aligned nucleotides per barcode length) and the identity percentage (the number of perfectly aligned nucleotides per alignment
length) are greater than 0.9. The sorted reads were then analysed to locate the mutation in the specified mutation sites. This was done by aligning the whole read sequence with the whole CpSADH gene sequence with the Smith-Waterman algorithm.

To determine the percentage occurrence of barcode mismatches, the following algorithm was implemented. The tail end of each sorted reads is first aligned with all the reverse barcodes that were used in the experiment. By using the same alignment scoring system as described above, the corresponding reverse barcode can be identified. Reads that have a pair of barcodes which are consistent with the assigned pair of barcodes are matched reads.

To determine the initial library composition, the index of each mutation site is first retrieved by using the nucleotide positions of the specified mutation site. The mutation at the site is accepted if all 3 nucleotides in the codon had error probability less than 0.05. The nucleotide quality score can be converted to error probability using the formula: error = 10^(-(ASC-33)/10), where ASC is the integer value of the ASCII character. A mutant is successfully identified after every mutation site in the read is accepted.

2.6 Minimal medium growth selection

The general workflow for the minimal medium growth selection in this study was as follows: prior to the growth selection experiment, the electrocompetent cells of *E. coli* MG1655 Δlpd strain was used for standard electroporation. The chromosomal integration of kanamycin resistant gene cassette into *lpd* locus of *E. coli* MG1655 was conducted in accordance with the lambda-red recombination-based method. After recovering in SOC for 1 h, the transformant was used to inoculate 10 mL of Medium 1 with appropriate antibiotics in a 50-mL falcon tube (Table 1). The initial optical density at 600 nm [OD$_{600}$] was 0.001. The tube was incubated at 30°C/250 rpm overnight. The overnight cell culture was then centrifuged at 3,500 rpm for 5 mins. The overnight grown seed culture was used to inoculate 10 mL of Medium 1 (initial OD$_{600}$: 0.001). In the 1st round of passaging, the *E. coli* culture was incubated in a 50-mL falcon tube for 72 h (30°C/250 rpm). During the 2nd round of passaging, the cell density from the 1st culture passage was normalised to OD$_{600}$: 0.1 with autoclaved ultrapure water. Subsequently, 100 µL of the normalised cell culture was used to inoculate 10 mL of Medium 1 (initial OD$_{600}$: 0.001) in a 50-mL falcon tube and cultured for 72 h (30°C/250 rpm).

To prepare 10 mL of Medium 1, the following solutions were mixed: 2 mL of the M9 basal medium, 2 mL of CSM stock solution (5 g/L, stock concentration, sterilized by using autoclave), 2 mL of sodium acetate stock solution (0.1 M, stock concentration, sterilized by using autoclave) and 17 µL of the K3 master mix. Antibiotic stock solutions were added if needed (refer to Table 1). Autoclaved ultrapure water was used to top up the volume to 10 mL. The M9 basal medium was prepared by dissolving 56.5 g of M9 medium broth powder in 1 L of ultrapure water, and the solution was autoclaved (more information included in Table S2). K3 master mix was prepared by mixing 2.5 mL of 0.1 M ferric citrate solution, 1 mL of 4.5 g/L thiamine solution, 3 mL of 4 mM Na$_2$MoO$_4$ solution, 1 mL of 1 M MgSO$_4$ solution, and 1 mL of 1000 X K3
trace elements stock. The 1000X K3 trace elements stock contained 5 g/L CaCl$_2$·2H$_2$O, 1.6 g/L MnCl$_2$·4H$_2$O, 0.38 g/L CuCl$_2$·2H$_2$O, 0.5 g/L CoCl$_2$·6H$_2$O, 0.94 g/L ZnCl$_2$, 0.03 g/L H$_3$BO$_3$, 0.4 g/L Na$_2$EDTA·2H$_2$O. All the stock solutions used to prepare the K3 master mix were autoclaved.

2.7 The fold change analysis based on nanopore sequencing

The library preparation workflow for nanopore sequencing is as follows: the cell cultures in the selection medium were first harvested by centrifugation at 3,500 rpm for 10 mins. The supernatant was removed to obtain the cell pellets. The cell pellets were then resuspended in 1 ml ultrapure H$_2$O and their optical density ($\text{OD}_{600}$) were measured and normalised to 0.1. The cell suspensions ($\text{OD}_{600} = 0.1$) were used as the DNA template for PCR. The PCR was performed as described in Section 2.4.

The mutant library composition was obtained following the procedure described in Section 2.4. The fold change analysis was performed by dividing the fraction of a mutant in a library by that of the mutant in the previous passage.

2.8 Diaphorase-coupled enzyme assay with secondary alcohol dehydrogenase

The diaphorase-coupled enzyme assay protocol was modified from a previous study.\textsuperscript{116} The in-vitro enzyme assay preparation was as follows: The cells from the growth-based experiment at 72 h was diluted by 10$^6$-fold with ultrapure water. Then, 200 µL of the diluent was plated on LB agar plate with 20 mM glucose and 20 mM sodium acetate and 50 µg/mL of spectinomycin. The plate was then incubated at 37°C overnight. Subsequently, isolated colonies were randomly picked from the agar plate and inoculated into LB liquid medium supplemented with 20 mM glucose, 20 mM sodium acetate, and 50 µg/mL spectinomycin. The culture was incubated at 30°C/250 rpm overnight. The grown cultures were washed with 1.1 mL of ultrapure water. The optical density at 600 nm [OD$_{600}$] of each culture was then measured and normalised to 10. Afterwards, each culture was harvested by aliquoting 1 mL of cell suspension into microtubes, centrifuged at 14,000 rpm for 1 min and discarding the supernatant. The harvested cells were lysed with 500 µL of B-PER in accordance to the manufacturer's protocol (Thermo Fisher Scientific). Subsequently, 10 µL of the cell lysate from each colony was pipetted into a 96-well black polystyrene microtiter plate (Costar 3603; Corning). For each reaction, the enzyme reaction reagent mixture contained 2 µL of 5 UN/ml diaphorase, 30 µl of 2 mM resazurin, 20 µl of 10 mM substrate ((R)-2-butanol or (S)-2-butanol), 10 µl of 1 mM NAD$^+$, 10 µl of 10X PBS and topped to 90 µl with ultrapure water. The assay reaction was initiated by adding the 90 µL of the enzyme reaction reagent into each sample well, and immediately followed by continuous measurement of the fluorescence of resorufin (Ex./Em. = 535/590 nm) by a plate reader (Infinite® 200 PRO; Tecan) for at least 20 mins. The final concentration of enzyme reaction contained 0.1 UN/mL of diaphorase, 0.6 mM resazurin, 2 mM substrate ((R)-2-butanol or (S)-2-butanol), 0.1 mM of NAD$^+$ and 1X PBS Buffer (pH 7.4).
3 Results

3.1 Paired-ends Illumina sequencing suffered from strand exchange

We first evaluated Illumina NGS on a library of CpSADH mutants (Library Q, Table 1). The region of interest was amplified using PCR and each amplicon was barcoded by unique sequences contained in forward and reverse primers. The amplicon fragments of each sample were pooled and sent to a service provider (GENEWIZ) for Illumina NGS. Paired-End 250 bp (PE250) PCR sample preparation and sequencing workflow was used. Four amino acids were mutated using NNK in the library and the mutation regions can be covered by the two 250-bp sequencing coverage ranges.

<table>
<thead>
<tr>
<th>Library name</th>
<th>Sequencing method</th>
<th>The percentage of mismatch reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library Q</td>
<td>Illumina NGS (Novaseq PE250)</td>
<td>21.8%</td>
</tr>
<tr>
<td>Library Q + R</td>
<td>Nanopore (GridION flow cell, standard workflow)</td>
<td>1.2%</td>
</tr>
<tr>
<td>Library K</td>
<td>Nanopore (MinION flow cell, standard workflow)</td>
<td>2.0%</td>
</tr>
<tr>
<td>Library M</td>
<td>Nanopore (Flongle flow cell, new workflow*)</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Various mutant libraries for CpSADH (secondary alcohol NAD-dependent dehydrogenase) were sequenced with either Illumina or nanopore sequencing. *Refer to Materials and methods (Section 2, Table 1) for the detailed protocol for the modified nanopore sequencing library preparation.

Through analysing the paired end reads, we discovered 21.8% (112,605 out of 516,249 reads) of the reads that contained both forward and reverse barcode had mismatched barcodes. This could be caused by the random strand exchanges during sample preparation since majority of the molecules share the same sequence except for the barcodes and the targeted mutations. This high error rate is problematic when paired reads are combined to investigate mutations spanning in a > 250 bp region.

3.2 Developing a Nanopore sequencing workflow to reduce strand exchange occurrence

As the extent of strand exchange was significant in Illumina NGS, we decided to explore nanopore sequencing, which could cover all the mutations in single read, thus avoiding any pairing operation during analysis. We analysed a sample that contained libraries Q and R (composition explained in Table 1). The sample was barcoded on both ends using PCR, mixed and processed using the standard workflow
recommended by the manufacturer. The sequencing results indicated that strand exchange occurrence was much lower than that of Illumina NGS (only 1.2%; 13,850 out of 1,181,177 reads).

In the standard nanopore sample prep workflow, T4 DNA polymerase is used to fill up gaps that may be present in the double stranded DNA amplicon fragments, as well as to add an additional adenine at the 3’ end. This is to facilitate the ligation of the sequencing adaptor which has a complementary 5’T end. We hypothesized that the strand exchange may occur during dA-tailing, where the polymerase may extend the 3’ end of the DNA strand which may include the wrong barcode information from another DNA strand that may be hybridized (Fig. 3). Since this enzymatic process is poorly characterized, we decided to generate the 3’A sticky end by using phosphorothioate (PS) modified oligonucleotide (oligo). Each forward oligo contained a T* at the 5’end before barcode. The * refers to the PS bond which can be cleaved using a simple chemical reaction involving iodine and ethanol at 70°C. As a result, a complementary 3’A end is produced without the use of any polymerase (Fig. 2b). Since only the forward oligo will generate the 3’A complementary sticky end, only the top strand is ligated to the adaptor for sequencing. The bottom strand would not be read. This new method allows for better control and understanding of the dA’ tailing process, which was not possible when the standard nanopore workflow involving T4 DNA polymerase was used.

When library M was sequenced this way, the strand exchange incidence was further reduced to 0.3% (21 out of 7,944 reads; we included unique barcodes on the reverse oligos for the purpose of studying strand exchange). Here, strand exchange was used as a feature to inspect if unexpected change is introduced to the samples during the workflow. A lower strand exchange rate should indicate that the obtained sequencing result is more reliable. The novel workflow we have developed here has substantially reduced the strand exchange rate and is easy to use. We have used the workflow in the rest of the study.

### 3.3 Performing quality control of mutant library generated for directed protein evolution with nanopore sequencing

To apply the modified nanopore sequencing workflow for actual protein engineering applications, we chose CpSADH, which is a useful secondary alcohol dehydrogenase that regenerates NADH when a secondary alcohol is reversibly oxidized to a ketone. The wildtype CpSADH predominately oxidizes aliphatic secondary S-alcohol into ketone.\(^{127}\) We sought to improve the activity of CpSADH towards (R)-2-butanol. Such mutants would be useful in converting racemic 2-butanol into 2-butanone for preparation of chiral 2-butanamine. Racemic 2-butanol could be derived from petroleum resources. In current practices of using them, half of such substrates were wasted because CpSADH could only oxidize (S)-2-butanol efficiently.

In 2018, Zhang et al. demonstrated the feasibility of using an NADH-auxotrophic system to assess NADH-producing formate and methanol dehydrogenases.\(^{138}\) By deleting the *lpd* gene in *E. coli*, cell growth was only possible when NADH was regenerated by a heterologous enzymatic reaction inside the cell. The authors have only used the wildtype dehydrogenases to demonstrate that the NADH regeneration from
these dehydrogenases could rescue cell growth from its NADH auxotrophy. They have not tested if this 
can be used to select more active enzyme mutants in protein engineering applications.

We then attempted to use the growth-based selection system for enriching CpSADH mutants with higher 
enzymatic activity towards \((R)\)-2-butanol but could not yield any positive hits. Hence, we hypothesized 
that the low success rate could be due to the biased library composition. Randomly picking colonies 
sample the population based on the composition, but the good mutants may not be the majority after 
selection even if they have been enriched, because their fraction in the initial library may be low. To test 
the hypothesis, we performed nanopore sequencing on the mutant library (ML03) before and after 
selection.

Site saturation mutagenesis via PCR is a common workflow to generate mutant libraries in protein 
engineering. However, it is unlikely all mutants can be delivered due to potential biasness during PCR, 
such as primer-template binding strength and preference. It is important to evaluate the quality of 
the mutant library, but this has not been thoroughly done in most protein engineering projects. A common 
practice is to use Sanger sequencing. A library would be deemed to be acceptable when mixed traces are 
observed in the desired sites. Such preliminary analysis cannot quantify the library composition.

Library ML03 was constructed by mutating three amino acid residues using NDT codon 
(L55NDT/C57NDT/T158NDT). Since three sites were chosen for mutation and each site can allow up to 
12 different amino acid substitutions due to degenerate codon (NDT), the theoretical library size was \(12^3\) 
\(= 1728\). Among them, \(~90\% (1560 unique mutants) were found in the nanopore sequencing analysis. 
The mutant library was discovered to be highly biased (Fig. 4c). For example, the most abundant mutant 
occupied 1.8% of the population despite that its expected abundance was only 0.06%. At all the three 
sites (L55, C57 and T158), cysteine (C) and serine (S) were underrepresented (Fig. 4b), which were 
attributed to the underrepresentation of G in the second nucleotide position of L55 and C in the first 
nucleotide position of T158 (Fig. 4a). This could be due to the different specific binding affinity of the 
NDT primers onto the template. Primers that have stronger binding tendencies would have higher 
sequence synthesis from PCR which may lead to an unequal distribution of mutants. This would not 
provide a fair starting point for all mutants during the selection process. Mutants that were low in % 
abundance in the beginning may not be able to outcompete mutants that were more abundant initially, 
even if they are enriched through the directed evolution. However, this issue could be mitigated by using 
the normalised fold change, which quantifies the enrichment of each mutant during the directed 
evolution.

### 3.4 Performing nanopore sequencing and fold change analysis on ML03

The cell library ML03 was cultured in a medium where they would only be able to grow when they could 
regenerate NADH using the CpSADH mutant and the provided substrate (R-2-butanol). The grown cells 
were passaged one more time in a modified chemically defined medium to strengthen the selection
pressure (the first growth medium contained low concentration of tryptone and yeast extract). The mutant population was then studied through nanopore sequencing (Section 2.7), and the number of unique mutants was found to decrease substantially between each passage (Table 3). This suggests that the mutants that are inactive or with low enzyme activities may have been out competed. For each after-passage sample, we ranked all the identified mutants based on their fold change with respect to the previous passage or original library. Figure 5 presents the fold change of top 10 mutants after each passage. The fold changes of top mutants after the second passage (up to 120 fold) were generally higher than those after the first passage (up to 60 fold). Many top mutants identified after the first passage were not substantially enriched during the second passage. Since the second passage did not include complex ingredients and had higher enrichment fold change, the top mutants identified after the second passage should be more reliable (having a higher success rate).

### Table 3
Nanopore sequencing results of ML03.

<table>
<thead>
<tr>
<th>Library</th>
<th>No. unique mutants</th>
<th>Total no. of qualified reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cells before selection</td>
<td>1560</td>
<td>12469</td>
</tr>
<tr>
<td>(R)-2-Butanol selected cells</td>
<td>1096</td>
<td>7080</td>
</tr>
<tr>
<td>(R)-2-Butanol selected 2nd passage cells</td>
<td>321</td>
<td>12459</td>
</tr>
</tbody>
</table>

After the second passage, the L55 position was dominated by either phenylalanine (F) or isoleucine (I) (Fig. 5b), suggesting that the residue preferred a hydrophobic amino acid residue to interact with the four-carbon chain in (R)-2-butanol; the C57 position was frequently mutated into glycine (G), which is substantially smaller than cysteine (C) and is considered less polar as compared to cysteine; the only residue found to replace T158 was serine (S). Serine and threonine (T) both carry hydroxyl group and only differ by one carbon, suggesting that the hydroxyl group is important in the active site.

### 3.5 Characterization of the mutants with high fold changes

To test the top ranked mutants after the second passage, we constructed four plasmids which would express the top four mutants (FGS, FGT, LFC, ICS) respectively and validated their enzymatic activity.

Mutant FGS exhibited the highest enzymatic activity towards (R)-2-butanol, followed by mutant FGT, while mutant LFC and ICS exhibited lower enzymatic activity than the wildtype (Fig. 6a). Further characterization of FGS using (R)-2-butanol showed that it has similar $K_M$ (42 mM) as the wildtype, while its $k_{cat}$ was 520% higher than the wildtype's $k_{cat}$. When (S)-2-butanol was used, all four mutants displayed similar activity as the wildtype (Fig. 6b). To further validate if the relationship between NADH-auxotroph growth and CpSADH's enzymatic activity exists, we grew *E. coli* strains expressing mutant FGS, the parent enzyme (LCT) or GFP. Mutant FGS indeed grew to a higher OD than the control strains (Fig. 6d), hence validating the combined workflows of nanopore sequencing and the growth-based selection system.
4 Discussion

Nanopore sequencing is more suitable for mutant library evaluation over Illumina NGS. There is almost no limit in read length when using nanopore sequencing, while Illumina NGS usually is restricted to 250–500 bp. This is an important consideration for protein engineering applications since mutant sites of interest may not be close together. Proteins fold onto themselves and the active pocket may comprise of amino acid residues that are far apart from each other when the protein is linearized but are close together in their native folding state.

At the time of writing this manuscript, many researchers had an impression that nanopore sequencing was not accurate enough for determining the mutant identity in protein engineering applications (personal communications). This study suggests that nanopore sequencing can be reliably used for quantifying library composition when high quality reads were selected. When nanopore sequencing was used to determine identity of homogenous plasmids derived from single colonies, ~90% of the reads report the same mutation combination (Table S3). When such consensus result is used, the conclusion would be even more convincing.

It was evident from nanopore sequencing that the initial mutant population was not distributed evenly. Many mutants were under-represented than others. When colonies were picked solely based on the library composition at the end point, many good mutants could be missed, because their fractions in the library could still be small at the end point even after they were substantially enriched during the passages. For example, the wildtype (LCT) was 28 times more abundant than the best performing mutant currently isolated (FGS) in the initial population before selection and was 7 times more abundant after selection. FGS only accounted for 5.7% after the second passage, which was quite low. This sheer demographic imbalance in the original mutant library made randomly picking highly active mutants at the end point challenging. In this study, we have managed to directly quantify the enrichment effectiveness using the nanopore sequencing workflow and successfully used it to isolate one highly active CpSADH mutant with respect to (R)-2-butanol.

5 Conclusion

In this study, we found that strand exchange occurred at high frequency (20%) when the Illumina technology was used to analyze multiplexed DNA samples that have high similarity. We developed a new Nanopore sequencing workflow that reduced the frequency to 0.3%. The workflow could be rapidly used at low cost to survey composition of protein mutant libraries, which were found to be biased when constructed using standard methods. Using the sequencing workflow enabled fairer evaluation of mutants of a Candida parapsilosis secondary alcohol dehydrogenase, leading to isolation of a mutant that was >500% more active than its parent enzyme. The workflow should be useful to the protein engineering community and inspire more Nanopore-sequencing applications in protein engineering.

Declarations
Author contributions

Daniel Tan: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft.
Vincent Fung: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft.
Tao Sun: Conceptualization, Methodology, Investigation, Formal analysis. Tian Kaiyuan: Investigation, Formal analysis. Zhi Li: Methodology, Supervision, Funding acquisition. Kang Zhou: Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition.

Conflict of interest statement

The authors declare no competing interest.

Acknowledgement

This work is financially supported by National Research Foundation (NRF) Singapore through a CRP project (identifier: A-0005297-01-00), and by Ministry of Education (MOE) Singapore through a Tier-2 research grant (identifier: R-279-000-594-112) and a Tier-1 research grant (identifier: A-8000193-01-00). V. F. was financially supported by PhD scholarship from National University of Singapore (NUS) and MOE Singapore.

Data availability

All data needed to evaluate the conclusions are present in the manuscript and its supplementary information.

References


**Illumina sequencing**

**Step 1: Amplification on individual samples**

![Diagram](image)

**Sample 1**
- A
- A
- a
- a

**Sample 2**
- B
- B
- b
- b

**Step 2: Samples were mixed together in equimolar**

![Diagram](image)

**Step 3: Paired-end sequencing**

![Diagram](image)

**Step 4: Demultiplexing**

![Diagram](image)

**Correctly demultiplexed samples**
- A
- a
- B
- b

**Incorrectly demultiplexed samples**
- A
- b
- B
- a

**Figure 1**

**Illumina NGS workflow for sequencing the libraries.** Letters and rectangles represent the inner adaptors. Combination of the same uppercase and lowercase letters represent correctly demultiplexed result, which any other combinations of letters represent incorrectly demultiplexed samples, which could be a result of strand exchange.
A new nanopore sequencing workflow for analyzing protein mutant library. In the original workflow provided by Oxford Nanopore Technologies, the end-prep and dA-tailing step adds 3'A overhangs to both strands of DNA. In this study, 3'A sticky end was generated using phosphorothionate oligos and a chemical cleavage method (for detailed protocol, refer to Section 2.4), and was added only to one strand.
of DNA. Subsequently, the sequencing adaptor with a 3’T sticky end could be ligated with the 3’A sticky end for nanopore sequencing.

**Figure 3**

**Strand exchange in nanopore sequencing.** Hybridization of single-barcoded PCR fragments during the multiplexing stage for dA-tailing in nanopore sequencing. T4 DNA polymerase is used to fill complementary sequence gaps and generate a 3’A sticky end for sequencing adaptor ligation.
Figure 4

Assessing library quality of ML03 in selected mutation sites. a) Nanopore readout percentages of all four nucleotides (A, T, C and G) in each mutation site before selection at the 55th, 57th and 158th positions are indicated by green, red, blue and yellow bars respectively. The expected percentages of nucleotides encoded by NDT codon during site saturated mutagenesis are indicated by corresponding coloured dotted lines. b) Nanopore readout percentage of all 12 amino acids encoded by NDT in each mutation sites before selection at the 55th, 57th and 158th positions. The average % readouts of 8.33% is shown using the red dotted line. c) Mutant library ML03 composition ranked based on % abundance. Every unique mutant before selection was counted and their % abundance ranked in ascending order.
Theoretical % of mutants is based on one in 1728 possible mutants in the library. The mutant population was discovered to be skewed and biased towards certain mutations.

**Figure 5**

Top 10 normalised fold change mutants of 1\textsuperscript{st} and 2\textsuperscript{nd} passaged ML03 cells from (R)-2-butanol selection. The normalised fold change of a mutant was calculated by dividing the percentage of each mutant in the selected library with the percentage of that mutant in the library before selection. The mutants were then named based on their amino acid residues found in L55, C57 and T158 positions respectively. For example, in **Figure 5a**, mutant IVS refers to L55I / C57V / T158S. The normalised logo plots were generated from MATLAB using the fold change values of the mutants in their respective libraries. 

- **a)** Normalised fold change of the top 10 ML03 mutants after 1\textsuperscript{st} passage of (R)-2-butanol selection with respect to the initial library before selection.
- **b)** Normalised fold change of the top 10 ML03 mutants after 2\textsuperscript{nd} passage of (R)-2-butanol selection with respect to the 1\textsuperscript{st} passage. Top 4 mutants, FGS, FGT, LFC and ICS were selected for enzymatic activity validation.
- **c)** Sequence motifs of 1\textsuperscript{st} passage ML03 mutants
from (R)-2-butanol selection. **d)** Sequence motifs of 2nd passage **ML03** mutants from (R)-2-butanol selection.

![Graphs and charts](image)

**Figure 6**

**Enzymatic activity validation of top 4 mutants with 2 mM (R)-2-butanol and 2 mM (S)-2-butanol, and growth curve of mutant FGS in 20 mM (R)-2-butanol selection medium.** Individual *E. coli* DH5α strains expressing top 4 fold change mutants from 2nd passaged ML03 was cultivated on LB agar plate and picked to validate their enzymatic activity. **DV02** (wildtype CPSADH) served as a performance reference, while **DV03** (eGFP) was the negative control. **a)** Red fluorescence signals (Ex./Em. = 535/590 nm) for the resazurin-diaphorase assay using 2 mM of (R)-2-butanol and **b)** 2 mM of (S)-2-butanol with the whole cell lysate. The *E. coli* MG1655 *Dlpd* strain containing the mutant **FGS** (L55F / C57G / T158S) was cultured in a chemically defined minimal media (composition specified in **Section 2.6**). **c)** Cell density of the fermentation cultures at every 24 h during the first passage and **d)** second passage was determined through measurement of optical density at 600 nm (OD$_{600}$). The experiment was conducted in 50-mL
falcon tubes with a final volume of 10 ml and were performed in duplicates. The size of the error bars (SE, n=2) may be smaller than the symbol sizes.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryinformation.docx](#)
- [TOC.png](#)