Methods for Multiplexed Biology

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Chemistry and Biochemistry

by

Nathan Barnard Lubock

2018
ABSTRACT OF THE DISSERTATION

Methods for Multiplexed Biology

by

Nathan Barnard Lubock

Doctor of Philosophy in Chemistry and Biochemistry

University of California, Los Angeles, 2018

Our ability to read and write DNA is fundamental for understanding Biology. While the past decade has brought about exponential improvements in our DNA sequencing and synthesis capabilities, major challenges remain. First, many DNA sequencers are hindered by their short read lengths, which has hindered genome assemblies, molecular haplotyping, and more recently, multiplexed functional assays. Synthetic Long Reads (SLRs) are a recently developed method that address this issue. SLRs leverage molecular barcodes to guide the computational reassembly of multiple short reads into a longer contiguous molecule. Here we present a novel SLR technology, BAGEL-seq, that can theoretically sequence molecules up to ∼40 kb, and achieve read lengths of ∼1 kb in a proof-of-principle experiment. Second, large-scale synthesis of gene-length, synthetic DNA is cost-prohibitive for many research applications. We present two complementary methods to address this limitation – one to quantify errors in synthetic gene constructs using next-generation sequencing (NGS), and another, DropSynth, to synthesize > 10,000 ∼1 kb genes using emulsions and DNA microarrays. Despite these limitations, researcher have recently leveraged DNA sequencing and synthesis to test the functional effects of thousands of variants in multiplex. Known as multiplexed functional assays (MFAs), these experiments have revolutionized the investigation of biological processes across the Central Dogma. In this dissertation we present three different MFAs. In the first, we used DropSynth to build homologogs of an essential E. coli protein, and tested their function in a complimentation assay. In the second, we measured the response of 39 murine olfactory receptors against hundreds of different odorants. Lastly, we assessed the effects of ∼7,800 single amino acid changes to the β2-adrenergic receptor in the presence of increasing agonist concentration. Taken together, this dissertation represents a fundamental improvement in our ability to read and write
DNA, and pushes the state of the art in combining these technologies for large-scale, multiplexed experiments.
The dissertation of Nathan Barnard Lubock is approved.

Todd O. Yeates
Leonid Kruglyak
Matteo Pellegrini
Jose A. Rodriguez
Sriram Kosuri, Committee Chair

University of California, Los Angeles
2018
To my family.
Contents

Acknowledgments xlii

Vita xlvi

1 Introduction 1

1.1 Background ................................................................. 1
1.2 Gene Synthesis ............................................................. 4
1.3 Synthetic Long Reads ..................................................... 6
1.4 Multiplexed Functional Assays ......................................... 8
1.5 This Work ................................................................ 9

References ................................................................. 10

2 BAGEL-seq – A Novel Method for Generating Synthetic Long Reads 15

2.1 Abstract ................................................................ 15
2.2 Introduction ................................................................. 16
2.3 Results ................................................................ 24
2.4 Future Directions ............................................................ 27
2.5 Materials and Methods .................................................. 29

References ................................................................. 33

3 A Systematic Comparison of Error Correction Enzymes by Next-generation Sequencing 38

3.1 Abstract ................................................................. 38
4  Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes 78
4.1  Abstract ................................................................. 78
4.2  Main Text ............................................................... 79
4.3  Materials and Methods ............................................... 87
4.4  Supplementary Information ....................................... 112
References ................................................................. 142

5  A Scalable, Multiplexed Assay for Decoding Receptor-Ligand Interactions 146
5.1  Abstract ................................................................. 146
5.2  Introduction ............................................................ 146
5.3  Results ................................................................. 148
5.4  Discussion .............................................................. 153
5.5  Materials and Methods ............................................. 153
5.6  Supplementary Information ....................................... 161
References ................................................................. 180

6  Deep Mutational Scan of the $\beta_2$-adrenergic Receptor 183
6.1  Abstract ................................................................. 183
6.2  Introduction ............................................................ 184
6.3  Results ................................................................. 186
6.4  Discussion .............................................................. 192
6.5  Materials and Methods ............................................. 194
6.6  Supplementary Information ....................................... 205
List of Figures

1.1 The cost of DNA Sequencing and Synthesis is decreasing exponentially. The advent of Next-generation Sequencing in 2005 resulted in \( \sim 10^6 \)-fold decrease in price in 10 years. Prices are reported per base sequenced or synthesized. At the time of writing (November, 2018) the price per base of microarray-base oligos is approximately \( 10^{-4} - 10^{-5} \) (data not shown). Figure ©Bioeconomy Capital [http://www.bioeconomycapital.com/bioeconomy-dashboard/](http://www.bioeconomycapital.com/bioeconomy-dashboard/) ................................................. 3

1.2 Typical gene synthesis workflow. An oligo library that tiles the gene of interest are synthesized and optionally purified. Using polymerase cycling assembly (PCA), overlapping oligos anneal together and the gaps are filled by a polymerase. After PCA, the full length product is amplified by PCR and optionally subjected to a round of error correction. Finally, the synthetic gene is cloned and verified by Sanger Sequencing. ©atdbio [https://www.atdbio.com/content/63/Gene-synthesis](https://www.atdbio.com/content/63/Gene-synthesis) .................................................. 4

1.3 Typical dilution-based SLR overview. In a dilution-based SLR, or haplotyping experiment, molecules of interest are diluted and amplified in 96- or 384-well plates. The dilution ensures sufficient diversity within each well and that haploids remain separated. The molecules are then fragmented and tagged (tagmented) with a unique barcode for each well. After sequencing, these barcodes are used to reassemble the molecules within each well. Due to the diversity of each well, it is unlikely, but not impossible that reads from one molecule could be incorrectly assigned to a different molecule. .................................................. 6
1.4 **Overview of a hypothetical multiplexed functional assay (MFA).** In this MFA, the activity of a given variant in the library is linked to a genetic reporter through a molecular barcode. Agonist stimulation of the expressed variant results in a signalling cascade that activates the transcription of its unique barcode. The level of activity can then be quantified by measuring the expression of the barcode with RNA-seq.

2.1 **Dilution-limited haplotyping overview.** Current methods rely on various techniques to separate long pieces of DNA into various compartments, where they are amplified, barcoded, and prepared for NGS. **A.** Historically, most haplotyping libraries were generated with fosmids. Briefly, genomic DNA is sheared and ligated with a fosmid backbone. Phages then package the $\sim40$ kb molecules and infect *E. coli*. The resulting library is diluted and grown in 96-well plates to amplify the selected DNA (purple and blue). **B.** Alternatively, the sheared genomic DNA can be diluted and amplified *in vitro* with multiple displacement amplification. The resulting product is then sheared and prepared for sequencing. **C.** Recently, a number of droplet-based technologies have been developed that obviate cumbersome plate-based protocols. These techniques separate the sheared genomic DNA into droplets containing Tn5 conjugated to beads, and perform an in-droplet tagmentation reaction.

2.2 **General overview of BAGEL-seq.** In the current iteration, we first clone a sequence of interest into a backbone containing a sequencing primer (purple), a unique barcode (blue), and a LoxP site. Next, we use dilute amounts of Tn5 transposase to randomly insert a single copy of a complimentary LoxP site on average throughout our sequence of interest. These LoxP sites will be inserted either in *cis* (left), *trans* (right), or not at all (middle). Next we dilute the sample and use Cre recombinase to perform intramolecular recombination between the two LoxP sites. This brings our sequencing primer and barcode adjacent to wherever Tn5 inserted the second LoxP site. Importantly, both excision (left) and inversion (right) result in readable sequence (orange). Finally, we tagment the products with Nextera and use PCR to amplify our sequences of interest.
2.3 **Illustration of BAGEL-seq primer walking.** Here, Tn5 has inserted LoxP sites at different locations in three copies of the same sequence. After Cre recombination, the sequencing primer (purple) is brought adjacent to wherever the second LoxP site was inserted. In addition, the sequencing primer reads through the barcode, maintaining the one-to-one linkage between barcode and molecule throughout the process. Lastly, the barcode enables reconstruction of the individual molecules by *in silico*.  

2.4 **In vitro validation of BAGEL-seq steps.**  
A. Serial dilution of Tn5 insertion with either full-length LoxP or Y-adapter LoxP. A supercoiled (−) and linearized (+) plasmid are provided for reference. At low concentration, both full-length LoxP and Y-adapter LoxP insertions linearize the plasmid. However, the full-length LoxP does not fragment the plasmid at higher concentrations and even has high-molecular weight species. This suggests that Tn5 insertion is not proceeding properly.  
B. Cre recombination control. The NEB control plasmid with two LoxP sites behaves as expected upon incubation with Cre (+). Similarly, a plasmid containing one Lox71 and two Lox67 sites recombines upon Cre addition regardless if it is supercoiled or linearized.  

2.5 **BAGEL-seq preliminary results.**  
A. Coverage of the top 5 most abundant barcodes. We see that reads are randomly distributed throughout the plasmid, albeit with some biases for specific positions.  
B. Overall coverage of the target plasmid combining all of the SLRs. Again, we find significant bias for a few positions on the plasmid. Due to the low diversity of the input DNA, this could be a result of Tn5’s insertion bias.  
C. Length distribution of SLRs. We find the majority of SLRs are shorter than the maximum 300 bp (red line) possible with paired 150 nt reads. However, this implies that we over-fragmented our input DNA before sequencing.
3.1 Schematic of Enzymatic Error Correction and Downstream Data Processing. We assembled our 142 bp product from two 113 nt oligos consisting of a 21 nt primer, a 64 nt payload, and a 28 nt overlap region. After annealing and overlap extension, we amplified our template via PCR, yielding 100 bp of template in-between the primer sites. We then denatured and re-annealed the PCR products to form heteroduplexes, thereby exposing any errors (shown in green). After, we subjected the pool of heteroduplexes to two successive rounds of ten different enzymatic error correction treatments. At each step, we took aliquots and sequenced the products on an Illumina MiSeq with fully overlapping forward and reverse reads. To mitigate sequencing errors, we used BBMerge to merge reads with a perfect agreement between the forward and reverse reads. We then aligned these sequences to the designed reference using an exhaustive Needleman-Wunsch aligner to minimize alignment artifacts. Finally, we further processed the alignments to quantitate the types and extent of different errors across all conditions.

3.2 Analysis of Model Gene Assembly Error Rates. A. The error rates per base are plotted across each position in our model separated by the four major classes of error types. We do not see strong positional effects for errors across the template. B. We find a majority of errors on the template are mismatches (MM), followed by single (Del.) and multiple base (M. Del.) deletions; Single (Ins.) and multiple base (M. Ins.) insertions occur at even lower frequencies. C. There are no significant differences between the median rate of mismatches at any base (Mann-Whitney U, NS). D. Similarly, there are no significant differences between transitions and transversions (Mann-Whitney U, NS), implying that the errors were doped uniformly into our oligos. Note: Blue line is a LOESS fit; box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.
3.3 **Effectiveness of Enzymatic Error Correction Methods.** Here we compare the error frequency (errors/kb) and number of perfect assemblies for ten different enzymatic error correction methods. We find that MutS is the most effective enzyme at increasing the percentage of perfect assemblies. However, ErrASE is the most effective at decreasing error frequency. Additionally, we see that the efficacy of T7 Endonuclease I is dependent on protocol, and that the addition of a ligase had detrimental effects on sequence quality. **Note:** the x-axis is ordered by decreasing number of perfect assemblies.

3.4 **Relative Decrease of Different Error Types.** **A.** All enzymes were able to correct both single- and multiple-base insertions and deletions. Additionally, we find that the best performing enzymes corrected the highest amount of mismatches. **Note:** the x-axis is ordered by increasing error frequency. **B.** We measure significant differences between the median decrease in C/G → G/C mismatches and the bulk median of all other mismatches after two treatments of ErrASE. Similarly, two treatments of T7 Endonuclease I results in a significant difference between the median decrease in A/T → T/A mismatches compared to the bulk median of all other mismatches (both Mann-Whitney U, $p << 0.001$).

3.5 **Effect of Polymerase on Assembly Quality.** We assembled two different 220 bp constructs (C1 and C2) from five 60 nt oligos with 20 bp overlaps with Q5 and Taq polymerase. **A.** We used our method to compare the error frequency (errors per kb) and percent perfect assemblies. We see that the average error frequency for both constructs is significantly higher for Taq than for Q5 (9.7 vs 2.5 errors/kb). We observe similar trends for the average percentage of perfect assemblies (60.5% for Q5 and 10.4% for Taq). **B.** Similar to the two-oligo assembly, we find that the Taq-based KAPA2G Robust polymerase also has a higher rate of transitions than transversions (mean of $5.32 \times 10^{-5}$ vs. $6.40 \times 10^{-6}$ over both constructs; Mann-Whitney U, $p << 0.001$). **C.** We find that the median rate of multiple base deletions per base in the overlap regions decreased ~2-fold relative to non-overlapping regions for both polymerases (Mann-Whitney U, $p << 0.001$). Similarly, the median rate of multiple base deletions per base also significantly decreases in the priming regions for both Taq (~6-fold) and Q5 (~13-fold) for both constructs (both Mann-Whitney U, $p << 0.001$). The difference in decrease between the polymerases was not significant.
3.6 **Effect of read aligner on error rates.** Here we mapped reads from the standard IDT oligo with **BBMap** (red), **Bowtie2** (green), and our Needleman-Wunsch aligner (blue), and quantified the error rates with our pipeline. We see that the choice of aligner affects the resulting error rates, especially for detecting multiple-base deletions. 63

3.7 **Distributions of error rates per position for the standard oligo assembly before and after ErrASE treatment.** We were unable to detect a significant change between the median error rate after two treatments for mismatches. **Note:** black bar is median value. 64

3.8 **In-depth analysis of standard assemblies.**

A) The error rates per base are plotted across each position in our model separated by the four major classes of error types. We do not see strong positional effects for errors across the template. B) We find a majority of errors on the template are mismatches (MM), followed by single (Del.) and multiple base (M. Del.) deletions; Single (Ins.) and multiple base (M. Ins.) insertions occur at even lower frequencies. C) We measure a significantly higher mismatch rate at A’s (4.33 × 10^{-3}) and T’s (4.25 × 10^{-3}) than at G’s (1.68 × 10^{-3}) and C’s (1.91 × 10^{-3}) (Mann-Whitney U, p << 0.001). D) We measure a significantly higher number of transitions (purple) than transversions (green) at each base (Mann-Whitney U, p << 0.001). The higher error rates at A’s and T’s is consistent with Taq polymerase errors. **Note:** Blue line is a LOESS fit; box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers. **Note:** here we performed the same analysis as Figure 2 in the main text with the error-doped assembly. 65

3.9 **Comparison of measured error rates from error-doped and standard oligos.** Here we plot the distribution of error rates per position and see that for every error sub-type the error rates are significantly higher for the error-doped oligos than those produced by the standard process (Mann-Whitney U Test, all p << 0.001). **Note:** Black bar is the median value. 66

3.10 **Mismatch correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments.** Error rates are plotted as the log_{2}-fold-change in error rate relative to the error-doped template. **Note:** box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers. 67
3.11 Single-base deletion correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments. Error rates are plotted as the $\log_2$-fold-change in error rate relative to the error-doped template. Note: box plots are first and third quartile for hinges, median for bar, and $1.5\times$ the inter-quartile range for whiskers.  

3.12 Single-base insertion correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments. Error rates are plotted as the $\log_2$-fold-change in error rate relative to the error-doped template. Note: box plots are first and third quartile for hinges, median for bar, and $1.5\times$ the inter-quartile range for whiskers.  

3.13 Correlations between error rates for five-oligo assembly technical replicates. We see that technical replicates are almost perfectly correlated (all $r > 0.995$), with the black line being $y = x$.  

3.14 Positional error rate distributions two assemblies using KAPA2G Robust and Q5 polymerase. We see that KAPA2G Robust, a Taq-based low-fidelity polymerase, incorporates Mismatches (MM) at nearly two-orders of magnitude higher than Q5, a high-fidelity polymerase. We find that both polymerases incorporate single base deletions (Del.), multiple base deletions (M. Del.), single base insertions (Ins.), and multiple base insertions (M. Ins.) at nearly identical rates. With the exception of multiple base insertions, these trends are robust to the different sequence contexts of the two constructs. We note that KAPA2G Robust incorporates a higher number of multiple base insertions around three tandem GGA repeats, likely due to polymerase slippage.
4.1 **DropSynth assembly and optimization.** A. We amplified array-derived oligos and exposed a single-stranded region that acts as a gene-specific microbead barcode. Barcoded beads display complementary single-stranded regions that selectively pull down the oligos necessary to assemble each gene. The beads are then emulsified, and the oligos are assembled by PCA. The emulsion is then broken, and the resultant assembled genes are barcoded and cloned. B. We used a model gene library that allowed us to monitor the level of specificity and coverage of the assembly process. We then optimized various aspects of the protocol including purification steps, DNA ligase, and bead couplings to improve the specificity of the assembly reaction. Enrichment is defined as the number of specific assemblies observed relative to what would be observed by random chance in a full combinatorial assembly. C. We attempted 96-plex gene assemblies with 3, 4, 5, or 6 oligonucleotides and the resultant libraries displayed the correct-sized band on an agarose gel. D. The distribution of read-counts for all 96 assemblies (4-oligo assembly) as determined by NGS.

4.2 **DropSynth assembly of 10,752 genes.** A. We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of library members with at least one perfect assembly and the median percent perfects determined using constructs with at least 100 barcodes is shown for each library. B. We observe that 872 PPAT homologs (75%) had at least one perfect assembly, and 1,002 homologs (87%) had at least one assembly within a distance of 5 a.a. from design. C. We assembled two codon variants for each designed DHFR homolog, allowing us to achieve higher coverage.
4.3 **PPAT complementation assay.** A. We used DropSynth to assemble a library of 1152 homologs of phosphopantetheine adenylyltransferase (PPAT), an essential enzyme catalyzing the second-to-last step in coenzyme A biosynthesis, and functionally characterized them using a pooled complementation assay. The barcoded library was transformed into *E. coli* Δ*coaD* cells containing a curable rescue plasmid expressing *E. coli* *coaD*. The rescue plasmid was removed allowing the homologs and their mutants to compete with each other in a batch culture. We tracked assembly barcode frequencies over four serial 1000-fold dilutions, and used the frequency changes to assign a fitness score. B. This phylogenetic tree shows 451 homologs each with at least 5 assembly barcodes, a subset of the full data set, where leaves are colored by fitness. Despite having a median 50% sequence identity, we find that the majority of PPAT homologs are able to complement the function of the native *E. coli* PPAT, with 70% having positive fitness values, while low-fitness homologs are dispersed throughout the tree without much clustering of clades.
4.4 Broad mutational scanning (BMS) analysis. A. The fitness landscape of 497 complementing PPAT homologs and their 71,061 mutants (within a distance of 5 a.a.) is projected onto the *E. coli* PPAT sequence, with each point in the heatmap showing the average fitness over all sequences containing that amino acid at each aligned position. Mutations are highly constrained at a core group of residues involved in catalytic function. Other positions show relatively little loss of function, when averaged over many homologs, despite known interactions with the substrates. The *E. coli* WT sequence is indicated by green squares, while the average position fitness, fitness of a residue deletion, mean EVmutation evolutionary statistical energy [20], site conservation, relative solvent accessibility, and secondary structure information is shown above. B. The average fitness at each position, with blue and red representing low and high fitness respectively, overlaid on the *E. coli* PPAT (PDB: 1QJC, 1GN8 [21]) structure complexed with 4’-phosphopantetheine and ATP. We observe loss-of-function for mutations occurring at the active site, while other residues involved with allosteric regulation by coenzyme A or dimer interfaces show large promiscuity, highlighting different strategies employed among homologs. C. In addition to complementing homologs, we can also analyze mutants of the 129 low-fitness (< -2.5) homologs, finding 385 gain-of-function (GoF) mutants across 55 homologs. We project this data onto the *E. coli* PPAT sequence and plot the number of GoF mutants at each position shaded by the number of different homologs represented. We find a total of 8 statistically significant positions (residues: 34, 35, 64, 68, 69, 103, 134, 135) corresponding to four regions in the PPAT structure.

4.5 The histogram of read distributions for six of the 96-plex 4-oligo assemblies shown in Fig 1B. A. T7 ligase and 20 ug beads. B. T4 and 20 ug beads. C. Taq ligase and 20 ug beads. D. T7 ligase and 100 ug beads. E. T4 ligase and 100 ug beads. F. Taq ligase and 100 ug beads.

4.6 A. A maximum likelihood phylogenetic tree for all 1,152 PPAT homologs as well as *E. coli* MG1655. Color scale represents percent amino acid sequence identity relative to E. coli PPAT (NP_418091.1). B. The gene length distribution for the 5,775 DHFR homologs assembled using either four or five 230-mer oligos with median gene lengths of 489 bp and 564 bp respectively.
4.7 **A.** Histogram of protein sequence lengths for all 1,152 PPAT library members. Lengths do not include start or stop codon. The longest, shortest, and median lengths are 516, 381, and 483 bp respectively. **B.** Although they share the same function, PPAT homologs have evolutionarily divergent sequences. The 662,976 pairwise percentage identities between the 1,152 members of the PPAT library at the amino acid level have a distribution with a median of 50% (σ = 5%). **C.** Without oligo isolation, amplification in bulk fails to produce the correct product [11]. A 4% agarose gel comparing the assembly products of a 24-member library of PPAT homologs (120 oligos) when the polymerase cycling assembly is done in bulk (BA) and in emulsion (EA). The expected product size upon correct assembly is between 520 bp to 550 bp. **D.** Each of the three 384-member PPAT libraries (1,920 oligos each) produced correct assembly products. A 4% agarose gel showing amplified assembly products, with the expected size for most amplicons around ~530 bp. Lane 1 and 2: High- and low-template PCR products for Lib 1. Lane 4 and 5: High- and low-template PCR products for Lib 2. Lane 7 and 8: High- and low-template PCR products for Lib 3. High- and low-template concentrations refer to either 2 uL or 0.2 uL of the purified assembly products from an emulsion used in a 50 uL PCR reaction.

4.8 **Agilent TapeStation gel image of DropSynth assembly of 28 384-member libraries of DHFR.** A total of 3 libraries of length 610bp (14, 15, 29) are assembled using 5 oligos while the remaining libraries of length 510bp are assembled using 4 oligos. Another 2 libraries (13, 30) are not shown with one having low yield on the oligo processing steps and another failing to amplify at the oligo stage.

4.9 **Agilent TapeStation gel image of 25 4-oligo DHFR libraries after assembly, digestion, ligation into barcoded plasmid and library preparation for sequencing.** 5-oligo libraries (14, 15, 29) were not prepared for sequencing due to limitations on Illumina read length capabilities.
4.10 **Sequencing statistics from sample S0.** These data are a set of paired end 600-cycle Miseq runs which read through the entire assembled gene and its assembly barcode for all three 384-member libraries. **A.** The number of reads per assembly barcode, with a median value of 2. S0 contains 7,038,274 unique assembly barcodes across 20,263,445 reads. Of these, 209,868 assembly barcodes 2.98% (739,771 reads 3.65%) mapped to the designed protein sequences without any amino acid mutations, of which 199,208 assembly barcodes contained at least one synonymous mutation. A total of 2,982,539 (42%) of the mapped assembly barcodes correspond to sequences containing a premature stop codon in the reading frame, of which the large majority (2,404,348) were due to indel mutations causing a frameshift while the rest were due to nonsense mutations. **B.** The long tail distribution of assembly barcodes per homolog, for assembly barcodes mapped to a perfect sequence. Median value is 56 and a total of 872 out of 1152 homologs are represented with at least one assembly barcode. **C.** The percentage of perfect protein sequences for constructs with at least 100 assembly barcodes. The solid line is the median value of 1.9%. **D.** Individually rank-ordered plots showing the number of barcodes with perfect assemblies, barcodes with assemblies within distance of 2 a.a., and all barcodes with an aligned homolog. **E.** The distribution of sequencing reads for the PPAT libraries. **F.** The coverage of the PPAT homologs as a function of the minimum percent identity. Most of the library members have assemblies with high identity to the respective designed homologs.

4.11 **A.** The library coverage shows strong correlation ($\rho=0.73$ (Pearson), p-value=3.4E-5) with the amount of DNA used to load the DropSynth beads prior to assembly. The coverage is defined as the number constructs with at least one perfect assembly. **B.** The number of constructs with the same barcode which dropout among different libraries. The red line is the level with an expectation value close to one for libraries of size 384 given a uniform dropout distributions. Values above this line are higher than would be expected by chance. About a dozen barcodes fall in this region.
4.12 **DropSynth assembly of 10,752 genes.** We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of barcodes per million representing assemblies within 5 a.a. of each gene is shown alongside the number of library members with at least one perfect assembly and the percent perfects determined using constructs with at least 100 barcodes.

4.13 **A.** The expected percentage of perfect assemblies for a given number of oligos and the amount of perfect oligos. **B.** The maximum gene assembly length possible for a given number of oligos and an oligo size ranging from (200 to 300bp).

4.14 **Error analysis of DropSynth Assemblies.** Using the error analysis pipeline developed by Lubock et. al [16], we randomly sampled one million reads from Miseq paired-end 600-cycle assembly barcode mapping data, performed an exhaustive alignment of each read against every perfect assembly and returned the best scoring alignment. **A.** Mismatches are the most common form of error, followed by multiple base deletions, single base deletions, and single base insertions. In particular, mismatches appear to be localized to the overlap regions. **B.** Raw counts of mismatches. A higher number of transitions than transversions were measured - in agreement with previous experiments where Taq-mediated amplification errors. This suggests that the majority of mismatches were likely introduced by KAPA2G Robust polymerase during assembly (evolved Taq variant).

4.15 **Phosphopantetheine adenyllytransferase (PPAT) metabolic pathway.** PPAT shown in red, catalyzes the second to last step in the five step biosynthesis of coenzyme A. It produces dephospho-coenzyme A from 4’-phosphopantetheine by transferring a adenylyl group from ATP [17], as shown. Either Mn$^{2+}$ or Mg$^{2+}$ acts as a cofactor. *E. coli* PPAT is hexameric and encoded by the 477 bp gene *coaD*. Several gene knockout [44, 45] and genetic footprinting [46] studies have confirmed *coaD* to be essential for growth on rich media in *E. coli* K-12 strains MC1061, MG1655, and DH10β. Both coenzyme A and dephospho-coenzyme A act as inhibitors of the forward reaction. PPAT’s low homology to its mammalian counterpart, which is encoded as one of the two domains on the bifunctional CoASy (CoA Synthase) enzyme, makes it a potential target for new antimicrobials [18]. At least a dozen different PPAT homologs have crystal structure data available.
4.16 A. Rescue plasmid pTKcoaD allows λ-red recombination of the essential coaD gene. Wild-type E. coli coaD is expressed constitutively along with GFP, which allows for confirmation of plasmid loss upon heat curing. B. High-copy expression plasmid pEVBC allows for IPTG-inducible expression of an homolog PPAT gene cloned in between the NdeI and KpnI sites. A 20-mer random assembly barcode is present downstream. C. Verification of the coaD gene knockout using colony PCR with two sets of internal primers. Four 42°C heat-cured colonies (c1-c4) are shown as well as four colonies (c5-c8) grown at 30°C which still contain the rescue plasmid. Red arrows indicate expected amplicon size when coaD gene sequence is present. D. Colony PCR verification of the coaD genomic knockout using external genomic primers for 9 knockout colonies and one wildtype control. Wildtype (no knockout) amplicon length is 590 bp while the knockout (KAN cassette knockin) amplicon length is 1150 bp, as marked by the red arrows. E. Comparison of E. coli DH10β Δ coaD pTKcoaD cells grown at 30°C (left) and 42°C (right). Cells were grown in LB+Kan for 15 hours at the corresponding temperature, to allow for sufficient outgrowth, before plating on LB+Kan and incubating at the corresponding temperature. By comparing the number of GFP-positive colonies seen in each case we estimated an escape frequency of 1 in 16,500 (σ = 1,600). We also tracked the escape frequency of cells after transformation with PPAT homologs and growth at 42°C, by determining the ratio of GFP negative to GFP positive cells, finding an escape frequency of 1 in 20,200 (σ = 9500) as determined by 8 independent transformations. These escape frequencies are similar to those previously reported for coaD (a.k.a. kdtB) upon heat curing of coaD expressing pMAK705 plasmid in a conditional knockout [44].
4.17 **PPAT complementation assay.** A. The fitness values for 651 homologs across two independent biological replicates shows strong correlation ($\rho=0.94$; Pearson). Six negative controls lacking the H/TxGH motif required for nucleophilic attack on the $\alpha$ phosphate of the ATP have very low fitness values (<3) in the assay. We colored each point based on the number of assembly barcodes that corresponded to errorless constructs, and find that reproducibility among replicates improves with increasing number of assembly barcodes (Fig. 4.18B). C. Despite having a median 50% sequence identity, distant homologs are typically still able to complement the function of the native *E. coli* PPAT (bottom row). This multiple sequence alignment table shows the fitness scores, percent sequence identity to *E. coli* PPAT, and source organism.

4.18 A. Fitness values of 329,897 individual assembly barcodes in each biological replicate, with a correlation of 0.948. A large number of low-fitness assembly barcodes correspond to assemblies with frameshifts due to indels. B. We see the reproducibility of the fitness values increase with the number of assembly barcodes. The absolute difference in homolog fitness values between the two biological replicates as a function of their number of assembly barcodes ($\rho=-0.34$; Spearman, p-value <2.2E-16). C. Fitness values are noisy with a median standard deviation of around 2.4. Box plots of individual assembly barcode fitness values for homologs in replicate A which have at least 50 assembly barcodes. Homologs are rank-ordered by their final fitness value.
4.19 **A.** Assembly barcode fitness for six of the homologs missing the H/TxGH motif required for catalytic activity. No simple mutation would be able to restore catalytic activity to these homologs, so they serve as a useful measure of the false positive rate for individual assembly barcodes. Of the 994 assembly barcodes only 9 assembly barcodes (0.9%) have a positive fitness value, indicating a low rate of false positives at the individual barcode level. **B.** Mean sequence fitness is reduced with increasing number of mutations ($\rho=-0.38$; Spearman, p-value $<2.2E-16$). Analysis of 144,573 sequences' fitness as a function of their a.a. distance from the designed homolog sequence. **C.** Very few sequences with less than $\sim$94% sequence identity show high fitness. For sequences represented by at least 2 assembly barcodes, we plot their fitness as a function of their sequence identity (relative to their corresponding designed sequences), within bins of 1%. ......... 125

4.20 The population of perfect and low mutational distance sequences expand as a function of time, while sequences with low sequence identity (primarily due to indels) are depleted. We see that non-functional assemblies are lost from the population primarily between the first two dilutions. Distribution of mapped assembly barcodes (top. and mapped reads (bottom., for each replicate (left & right., based on distance from the designed sequence. ....... 126
4.21 **Synthesis verification.** Sequence-verified clones were obtained for 37 of 49 homologs. A. The amount of colonies observed after transformation of amplified constructs into *E. coli* DH10β Δ coaD pTKcoaD cells grown at 30°C (positive control) and 42°C (complementation). Symbol indicates 42°C colony size relative to 30°C colonies. Dashed line shows slope of one and is not a fit. The presence of a cluster with low colony counts in both conditions made up primarily of low-fitness homologs suggests possible toxicity effects. Two false positives are observed which had positive fitness in the pooled assay but produced no colonies in this transformation. Both of these had a low number of assembly barcodes (1 and 25). The majority of high fitness homologs produced large numbers of colonies in both conditions with high correspondence between the two. B. Comparison of growth rate of individual homologs (log-scale) and gain-of-function mutants as determined on a plate reader with experimentally-determined fitness from pooled complementation assay, with a Spearman’s correlation of \( r_S = 0.86 \). Growth rate (hr\(^{-1}\)) is defined as the maximum slope of OD600 vs. time on a log/linear plot. Fit is carried out using log growth rate and does not include the eight homologs with a growth rate of zero. Wildtype PPAT *E. coli* had a growth rate of 0.132 indicative of gene dosage toxicity effects due to overexpression. C. Correlation between the residual error of the fit of growth rate to fitness and number of assembly barcodes in homologs (\( r_s = -0.50 \), Spearman, p-value 1.7E-3). Constructs with fewer assembly barcodes tend to have higher error between individual growth rate and fitness in the pooled assay, highlighting the need for many assembly barcodes to determine fitness.

4.22 **PPAT phylogenetic tree.** The majority of homologs listed complement wildtype *E. coli*, with low-fitness homologs randomly dispersed throughout the tree with minimal clustering. A phylogenetic tree of 451 homologs labeled, similar to Fig. 4.3D, with each leaf labeled with the organism name and shaded by fitness.
4.23 A. Phylogenetic tree of 411 homologs based on NCBI taxonomy rather than PPAT sequence, generated using phyloT (http://phylot.biobyte.de). The median fitness was used when multiple sequences were annotated with the same taxonomic ID. B. Fitness of PPAT homologs from organisms annotated as extremophiles. Of the different classes, alkaliphiles show a weak shift to lower fitness values (p=0.059 Wilcoxon rank sum test). Previous characterization of E. coli PPAT showed a maximum activity at pH 6.9 which was reduced to 68% of the maximum by pH 8 [47].

4.24 A. The average BMS position fitness compared to the conservation (Jensen-Shannon divergence). As expected mutations tend to be more constrained at highly conserved sites (ρ=-0.64; Pearson, p-value <2.2E-16). B. The average BMS position fitness compared to the relative solvent accessibility based on a DSSP analysis of the 1H1T crystal structure (dimer not hexamer). Buried residues tend to be more constrained (ρ=0.42; Pearson, p-value 3.9E-8). C. Mutational scanning coverage decreases at site of low fitness (ρ=0.76; Pearson, p-value <2.2E-16). This effect is due to assembly barcodes with low read numbers which, due to their low fitness, never pass the minimum 10 read threshold. D. Residues appearing in wildtype E. coli PPAT are associated with higher fitness values. The distribution of fitness values for residues present in the E. coli PPAT sequence (median = 2.16, σ = 0.24) compared to all others (median = 1.86, σ = 2.16).

4.25 Variant classifier. We implemented a classifier to predict how different BMS variants would perform in our assay. Each BMS variant was categorized into two bins based on whether or not their measured fitness score was greater than 0. We then performed a logistic regression using 6 features for our model - the amino acid mutation, secondary structure class as assigned by DSSP (loop, beta-sheet, or alpha-helix), relative solvent accessibility as assigned by DSSP, sequence conservation, evolutionary coupling as predicted by EVMutation, and the frequency of residue substitution from the sequence alignment used for EVMutation’s prediction. To assess the performance of our classifier, we performed 10 repeats of 5-fold cross-validation on our dataset and measured the precision and recall of each model on its respective hold-out set. We found that on average, our simple classifier has A. an average accuracy of 0.825 +/- 0.013, B. a precision of 0.853 +/- 0.009, and an average recall of 0.931 +/- 0.014.
4.26 **A.** The relative solvent accessibility and conservation of each of the eight gain of function positions. **B.** Weblogo showing the probability of each residue at the gain-of-function positions for low-fitness homologs. **C.** Weblogo of GoF residues for homologs which complemented. **D.** The mean fitness of each GoF mutation at the significant positions, with the number of mutants observed at each a.a. **E.** The same plot with the data derived from the broad mutational scan using complementing homologs and their mutants. **F.** *E. coli* PPAT structure with the eight GoF residues shaded in red. Glu-134 is involved in hydrophobic interactions with coenzyme A [42], suggesting a role for GoF mutations in modulating the inhibitory feedback, while Ala-103 participates in hydrophobic interactions between the PPAT dimers.

4.27 **A.** The oligo design process. Briefly, a.a sequences are assigned random weighted codons and appended with restriction and primer sites used in DropSynth assembly. Sequences are then split into five oligos with ~20-nt overlap regions. Individual oligo sequences are appended with restriction sites, padding sequences, gene-specific microbead barcodes flanked by nicking sites, and amplification primer sites leading to a library of 200-nt sequences. **B.** The DropSynth microbead barcoding process. Microbead barcode oligos are individually mixed with 3’ biotinylated ligation oligos and dual 5’ biotinylated anchor oligos, ligated using T4 ligase and phosphorylated with T4 PNK, exposing the microbead barcode sequence (NNNNNNNNNNNN). Biotinylated duplexes are then individually bound to M270 streptavidin Dynabeads and pooled together.
4.28 **Nick processing to generate single-stranded microbead barcode overhang.** A. A 10% TBE-Urea denaturing gel highlighting the steps in nick processing. Lanes 1, 5, 7: a 10 bp ladder. Lane 2: Before processing, all oligos should be 200 nt. Lane 3: After nick processing we expect fragments of 165 nt, 177 nt, 35 nt, and 23 nt. Lane 4: After streptavidin Dynabead cleanup of nick processed oligos we expect fragments of 165 nt and 177 nt. Lane 6: The captured Dynabead fraction after boiling at 90°C for 10 min in 10 mM EDTA pH 8.2. B. A non-denaturing 4% agarose gel showing the nick processing which takes a 200 bp duplex and leaves a 12-nt single-stranded microbead barcode overhang on a 165 bp dsDNA fragment. Lanes p1-p4 showing several samples after nick processing and also one before processing (NP). Lanes b1-b4 show the corresponding Dynabead fractions after denaturing at 80°C for 3 min. Full length oligos containing errors in the nt.BspQI sites will not have both strands nicked and are likely to be pulled down by the Dynabeads together with the short fragment.

4.29 **Characterization of the distribution of droplet sizes for the vortex emulsions.** Briefly, 100 uL of Kapa Robust buffer was added to an eppendorf tube with 600 uL of Bio-Rad Droplet Generation Oil and vortexed upright for 4 minutes on the highest setting of a Vortex-Genie 2. Samples were then taken from the bottom, middle, and top of the resulting emulsion and imaged under 40X magnification. The mode of the droplet diameter distribution peaks below 5 um. Scale bars are 100 um. Bottom right: Histogram of droplet diameters as determined by image analysis. Median droplet diameter is below 5 um.

5.1 **A Genomically Integrated Synthetic Circuit Allows Screening for Mammalian Olfactory Receptor Activation.** A. Schematic of the synthetic circuit for stable OR expression and function in an engineered HEK293T cell line (ScL21). Heterologous accessory factors expressed include (pink): RTP1S, RTP2, G_olf, and Ric8b. B. MOR42-3 reporter activation expressing the receptor transiently (left) or genomically integrated (right) at varying copy number, under constitutive or inducible expression in HEK293T cells. C. MOR258-5 reporter activation with/without accessory factors (A.F.s), RTP1S and RTP2, transiently coexpressed in HEK293T cells compared to stable receptor expression in ScL21. D. Reporter activation response curves for MOR258-5 and MOR41-1 genomically integrated in ScL21 with/without doxycycline induction of receptor expression.
5.2 Large-Scale, Multiplexed Screening of Olfactory Receptor-Odorant Interactions.

**A.** Experimental workflow for OR library generation and high-throughput screening. To perform assay, we cloned OR genes and barcodes into plasmids, engineered cell lines via individual transposition of plasmids, pooled cell lines and performed screen in 96 well plates. We assayed the equivalent of 81,012 wells of a screen where interactions are tested individually.

**B.** Heatmap of interactions from the screen clustered by odorant and receptor responses, and shaded by the minimum activating odorant concentration that triggered reporter activity. Only ORs and chemicals that registered at least one interaction are shown.

**C.** Chemical names and structures for odorants that activate MOR23-1 and MOR5-1.

**D.** Chemical names and structures for odorants that activate MOR258-5 and MOR13-1.

**E.** Chemical hits identified for MOR170-1 and MOR139-1 (black) mapped onto a PCA projection of the chemical space of our odorant panel (grey). Shaded areas highlight hits that cluster together in chemical space.

5.3 Schematic of the Synthetic Olfactory Activation Circuit in the Engineered Cell Line. Full graphical representation of the expressed components for expression/signaling of the ORs and the barcoded reporter system as shown in Fig. 5.1 of the main text. Receptor expression is controlled by the Tet-On system (Orange). After doxycycline induction, the OR is expressed on the cell surface with assistance from two exogenously expressed chaperones, RTP1S and RTP2 (pink). Upon odorant activation, G protein signaling triggers cAMP production. Signaling is augmented by transgenic expression of the native OR G alpha subunit, G_olf, and its corresponding GEF, Ric8b (pink). cAMP leads to activation of the kinase PKA that phosphorylates the transcription factor CREB leading to expression of the barcoded reporter.
5.4 **Engineering HEK293T Cells for Stable, Functional OR Expression.** A. Comparison of MOR42-3 activation under inducible receptor expression either transiently transfected (left) or integrated at single copy into the H11 genomic locus (right). B. Comparison of MOR42-3 reporter activation integrated at multiple copies in the genome with the PiggyBac Transposon System under constitutive or inducible receptor expression. C. Relative receptor/reporter copy number determined with qPCR for three transposed ORs relative to a single copy integrant. D. Comparison of MOR258-5 and MOR30-1 reporter activation (stimulated with 2-coumaranone and Decanoic Acid respectively) co-transfected with or without Accessory Factors (AF) Gαolf, Ric8b, RTP1S, and RTP2. E. Cell line generation for stable accessory factor expression. After transfection, clones were isolated and screened for activation of ORs, MOR258-5 and OR7D4, that require accessory factors for functional expression. The dark purple bar represents the clone (ScL21) selected for further experiments.  

5.5 **Design of a Multiplexed Genetic Reporter for OR Activation.** A. Annotated Vector map for the plasmid containing the OR expression cassette and genetic reporter for integration. B. MOR42-3 reporter activation in cells transiently co-expressing the receptor and genetic reporter on separate plasmids or together. C. Fold activation of MOR42-3 driven by an engineered CRE enhancer (7 CREB binding sites) compared to Promega’s pGL4.19 CRE enhancer. D. Genetic reporter basal activation upon inducible expression of MOR42-3 with or without a DNA insulator upstream of the CRE enhancer. 

5.6 **Evolutionary Tree of Mouse ORs.** Phylogenetic tree inferred from amino acid sequence of functional murine ORs. The length of lines indicate degree of divergence between ORs. Red dots indicate ORs that were selected for inclusion in this study. 

5.7 **Pilot-Scale Recapitulation of Odorant Response in Multiplex.** A. Heatmap displaying 39 pooled receptors’ activity against 9 odorants and 2 mixtures. Interactions are colored by the log₂-fold activation of the genetic reporter (see methods). Odorant interactions previously identified are boxed in yellow [12]. B. Dose-response curves for odorants or forskolin (adenylate cyclase stimulator) at 5 concentrations screened against the OR library. Curves for ORs known to interact with the odorant are colored. Stimulation with forskolin does not show substantial differential activity between ORs in our assay.
5.8 **Library Representation.** Representation of individual ORs in the library for the 39/42 ORs that had sufficient cellular coverage (see Methods). **A.** Frequency of each OR as a fraction of the library determined by the relative activation of each reporter stimulated with DMSO. **B.** The relationship between frequency of each OR in the library and the average coefficient of variation between biological replicate measurements of reporter activation for all conditions.

5.9 **Replicability of the Large-Scale Multiplexed Screen.** **A.** Histogram displaying the distribution of the coefficient of variation for the OR library when stimulated with DMSO. **B.** Histogram displaying the distribution of the coefficient of variation for the OR library against all conditions assayed. **C.** Dose-response curves for the control odorants included on each 96-well plate assayed. Each color represents a different plate.

5.10 **Significance and Fold Change of High-Throughput Assay Data.** **A.** The False Discovery Rate (FDR; Benjamini-Hochberg corrected, see Methods) plotted against the fold change for each OR-odorant interaction. The dashed line represents the 1% FDR, the cutoff used to identify positive interactions. **B.** The subset of interactions tested by a follow-up orthogonal luciferase assay (color indicates whether it was recapitulated in the orthogonal system). Of the interactions passing a 1% FDR, 20 of 27 also showed interaction in the orthogonal follow-up assay.

5.11 **Recapitulation of the Screen in a Transient, Orthogonal System.** Secondary screen of chemicals in a transient OR reporter activation system with a luciferase reporter gene readout [26]. Each plot shows the behavior of a control cell line expressing the reporter gene but no OR (black line), as well as a cell line expressing a specific OR and reporter gene. In addition, data from the high throughput screen (labeled as Seq) is plotted for reference.
5.12 Assay Correspondence with Previously Screened Odorant-Receptor Pairs. A. FDR plotted against fold induction for the 540 odorant-OR interactions that were previously tested by Saito et al.12. Points are colored by the EC$_{50}$ of the interaction in the previous work. Grey points represent interactions not identified in the previous screen. Comparing the results from transient versus integrated luciferase assays revealed that, in some cases, the integrated system required a higher concentration of odorant to achieve significant activation, likely because of the lower DNA copy number of the CRE-driven luciferase and receptor. Since the highest concentration of odorant assayed was 1 mM, low affinity interactions may not have been detectable in this screen. B. The FDR in the assay related to the EC$_{50}$ of the hit from the previous screen, colored by the fold activation from the multiplexed screen.

5.13 Location of Odors Tested with Respect to a Learned Chemical Space. Locations of the chemicals tested in this assay in chemical space. The molecular autoencoder was used to generate a 292-dimensional representations of 250,000 randomly sampled molecules from the ChEMBL 23 database (blue) as well as the chemicals tested in our assay (red) projected onto two dimensions with Principal Component Analysis (PCA)[26].

5.14 Clustering of Odorant Response for Receptors. The locations of any hits (black) with respect to other chemicals tested (grey) for each OR on the PCA on the 292-dimensional latent representation. PC1 explains 34.4% of variance and PC2 explains 14.0% of the variance.
6.1 Platform for Deep Mutational Scanning of GPCRs and Variant-Activity Landscape. A. Graphical Display of Multiplexed GPCR Activity Assay. ADRB2 variants with their barcoded genetic reporter are integrated into a defined genomic locus such that one variant is integrated per cell. Upon isoproterenol agonization, G protein signaling induces transcription of the cAMP-responsive genetic reporter and the barcode. The barcode sequence in the 3’ UTR of the reporter encodes the identity of the receptor within the same cell. B. Overview of workflow for Multiplexed GPCR Activity Assay. The variant library is generated, barcoded, and cloned into a vector with a genetic reporter. The library is then integrated into HEK293T cells and agonized with various concentrations of isoproterenol. After stimulation, mutant activity is determined by measuring the relative abundance of each variant’s barcoded cAMP-responsive genetic reporter transcripts with RNA-seq. C. Top: Secondary structure diagram represents the N and C termini in black, the transmembrane domains as blocks, and the intra- and extracellular domains in blue and green respectively. The EVmutation track displays average effect of every mutation as predicted by EVmutation. The Conservation track displays the sequence conservation of each residue. The shaded guides represent positions of the protein in the transmembrane domain. Bottom: The heatmap representation of the activity of every missense mutation and frameshift at each agonist condition. Cells are colored by the relative activity to the mean frameshift mutation.

6.2 Unsupervised Learning Elucidates Broad Structural Features and Critical Residues of the β2AR. A. We averaged amino acid substitutions into classes based on their physico-chemical properties. We then used Uniform Manifold Approximation and Projection (UMAP) to learn a 2D representation of every residue’s response to these classes of substitutions across all agonist conditions. Each residue is assigned into one of six clusters using HDBSCAN (see Supplementary Fig. 6.10). B. The class averages of each of these cluster reveals their distinct responses to mutation. The upper dashed line represents the mean of the Cluster 6 and the lower dashed line represents the mean activity of frameshifted mutants. C. A 2D snake plot representation of β2AR secondary structure with each residue colored by cluster.
6.3 Positional Constraint Restates the Significance of Known Structural Motifs and Suggests Novel Residues for Investigation. A. Residues within the transmembrane domain colored by their tolerance to particular substitutions. Teal residues are intolerant to both hydrophobic and charged amino acids (globally intolerant), and brown residues are tolerant to hydrophobic amino acids but intolerant to charged amino acids. These charge sensitive positions tend to point into the membrane, while the globally intolerant positions face into the core of the protein. B. The crystal structure of the hydroxybenzyl isoproterenol-activated state of the β2AR (PDB: 4LDL) with residues from the mutationally intolerant Clusters 1 and 2 highlighted in magenta. C-F Selected vignettes of residues from the mutationally intolerant UMAP clusters. C. W286 of the CWxP motif and the neighboring G315 are positioned in close proximity. Substitutions at G315 are likely to cause a steric clash with W286 (PDB: 4LDL). D. An inactive state water-mediated hydrogen bond network (red) associates N51 and Y326 (PDB: 2RH1). Disruption of this network may destabilize the receptor. E. The ligand-bound orthosteric site surface colored by mutational tolerance displays the assay’s discriminatory power between agonists (PDB: 4LDL). F. Mutationally intolerant β2AR residues at the G protein interface from the β2AR-Gs complex crystal structure (PDB: 3SN6), V222, I135, and Q229.

6.4 Conserved Extracellular Tryptophan-Disulphide ‘Structural Latch’ in Class A GPCRs is Rigid and Conformation-Independent. A. W99 is mutationally intolerant and appears to be contacting the C106-C191 disulfide bond of the ECL1. A structural comparison of Class A GPCR structures reveals the Trp-disulfide bond contact is conserved in 22 of the 25 receptors. B. The Trp-disulfide bond contact is maintained in both the inactive and active state structures for the β2AR and M2 muscarinic receptor.
6.5 **Schematic of Generation, Functional Assessment, and Analysis of All 7,828 Missense Variants of the β²AR.** We synthesized missense variants on an oligonucleotide microarray, amplified the oligos and appended random DNA barcode sequences, and cloned the variants into WT background vectors. We then mapped barcode-variant pairs with next-generation sequencing and cloned the remainder of the WT receptor and genetic reporter into the construct. Next, we integrated the variant library en masse into a serine recombinase landing pad engineered at the H11 locus of ∆ADRB2 HEK293T cells. The recombination strategy ensures a single receptor variant/genetic reporter is integrated per cell to avoid crosstalk between genetic reporters. After selection, we stimulated the library with various concentrations of ADRB2 agonist, isoproterenol. Finally, we determined mutant activity by measuring the relative abundance of each variant’s barcoded cAMP-responsive genetic reporter transcripts with RNA-seq.
6.6 Engineering HEK293T Cells for Clonal and Functional Integration of an ADRB2 Genetic Reporter. A. Schematic of functional assay to ensure the landing pad is present at single copy in the genome and can recombine a single donor plasmid per cell. Single copy integration is essential to ensure receptor’s of variable functionality do not activate barcoded reporters mapped to other variants. Upon co-transfection of the promoterless GFP and mCherry plasmids with bxb1 recombinase sites, a cell line with a single landing pad will exclusively integrate one cassette. Therefore, cells will be either GFP$^+$ or mCherry$^+$ but never both. B. Flow Cytometry plots detailing the percentage of GFP$^+$ and mCherry$^+$ cells when transfected with an equimolar ratio of promoterless GFP and mCherry expression cassettes with or without Bxb1 recombinase expression. C. Activation of a cAMP-responsive genetic reporter via a luciferase assay integrated in the landing pad when stimulated with isoproterenol in a WT or $\Delta$ADRB2 background. Activation of the reporter in the WT background emphasizes the importance for generation of the $\Delta$ADRB2 for the purpose of multiplexed experiment. D. Activation of a genetic reporter with or without exogenous ADRB2 expression via a luciferase assay integrated in the landing pad when stimulated with isoproterenol in $\Delta$ADRB2 cells. E. Activation of an equivalent integrated genetic reporter/ADRB2 cassette via qRT-PCR of the reporter transcript in $\Delta$ADRB2 cells. F. Schematic detailing the recombination of the reporter/receptor expression plasmid into the landing pad locus.
6.7 Individual and Global Multiplexed Assay Validation. A. To validate our genetic reporter, we compared the measured mutant activity screened individually in the landing pad locus via a luciferase assay to via the multiplexed mutational scan. We recapitulated results individually observed in the DMS for both null and hypomorphic mutations. B. The distribution of activity for frameshifts are significantly different that the distribution of our designed missense mutations across increasing isoproterenol concentrations ($p << 0.001$). C. We also find the relative activity for frameshift mutations mapped to each codon in the ADRB2 sequence is markedly decreased in the C-terminus of the protein (dotted line), and is consistent across agonist concentration. Blue line represents the LOESS fit. D. The measurements between barcodes at the RNA-seq level are well correlated ($r = 0.867, r = 0.871, r = 0.864, r = 0.868$) at all agonist concentrations (0, 0.150, 0.625, and 5 µM Iso). Similarly, the mean forskolin-normalized values for each variant are correlated at every concentration ($r = 0.657, r = 0.686, r = 0.729, r = 0.750$). Bars represent log$_{10}$ counts per hex-bin.

6.8 Activity of Proline Mutations by Protein Domain. Proline substitutions in the transmembrane domain results in lower activity across all agonist conditions.

6.9 Correlation with Sequence Conservation and Covariation. A. Mutational tolerance is highly correlated with sequence conservation and is maximized at EC$_{100}$ ($\rho = -0.689, \rho = -0.719, \rho = -0.747, \rho = -0.634$ for -Iso, 0.150 µM Iso, 0.625 µM Iso, and 5 µM Iso, respectively). Here we calculated sequence conservation using the Jensen-Shannon divergence from a multiple alignment of 55 ADRB2 orthologs from the OMA database. The blue line is the least squares fit. B. Similarly, our measure of relative activity for individual substitutions is well correlated with the predictions from EVMutation, and is maximized at EC$_{100}$ ($\rho = 0.370, \rho = 0.460, \rho = 0.521, \rho = 0.504$). The blue line is the least squares fit.
6.10 **Cluster Assignment is Robust Across Different UMAP Embeddings.** Given the high dimensionality of our data, we used UMAP to learn lower-dimension representations of our data before clustering with HDBSCAN (minimum cluster size = 10). To ensure that the clustering results are not biased by a particular UMAP embedding, we ran a hyperparameter search over the dimension and nearest neighbor parameters of UMAP. We then plot the HDBSCAN cluster assignments on a 2D UMAP embedding to ease visualization. Points that HDBSCAN does not assign to a cluster are colored powder blue. We find that groups of residues reliably cluster together regardless of the UMAP embedding, and manually assign all residues to six distinct clusters following the robust HDBSCAN assignment.

6.11 **Mutational Profile Suggests Side Chain Orientation and Environment.** A. The crystal structure of the hydroxybenzyl isoproterenol-activated state of the $\beta_2$AR (PDB: 4LDL) with residues colored by UMAP cluster identity. B. Distributions of Solvent Accessible Surface Area (SASA) for each cluster at EC$_{100}$. C. Hydrophobic versus Charge Sensitivity across all drug conditions. Points are colored by cluster identity. We define residues to be globally intolerant to substitution if their Hydrophobic and Charge Sensitivity is greater than 0. Similarly, we define residues to be uniquely charge sensitive if their Hydrophobic Sensitivity is less than 1 and their Charge Sensitivity is greater than 1. D. Distributions of SASA for intolerant and charge sensitive clusters are significantly different across all drug concentrations (all $p < 0.0005$).
6.12 **Inspection of Mutationally Intolerant Residues.** A. Rank order plot of mutational tolerance at 0.625 µM isoproterenol for all 412 β_2_AR residues mutagenized. Residues in known structural motifs (colored points) are significantly more sensitive to mutation than other positions on the protein (p << 0.001). B. Residues that interact with the head (orange) of hydroxybenzyl isoproterenol have significantly lower mutational tolerance than those that interact with the hydroxybenzyl functional group on the tail (purple). These differences are significantly different at EC_{50} (p = 0.028), EC_{100} (p = 0.016), and saturating agonist concentration (p = 0.008). C. Box plot displaying the mutational tolerance of all predicted contacts of the cholesterol binding pocket determined in the timolol-bound structure of the β_2_AR inactive state (PDB: 3D4S). The highly conserved W158^{4.50} is the most constrained residue. D. ECL1, with residues belonging to cluster 1 and 2 colored magenta, contains a region of sensitivity where R63 and L64 are both intolerant to substitution. However, neighboring E62 displays greater than WT activity at multiple individual mutations (PDB: 3SN6).

6.13 **Evaluation of Individual Missense Variants.** A. The inactive state β_2_AR structure highlighted in regions where residues display greater than WT activity without agonist stimulation for at least one individual mutation (yellow). These mutations localize to the extracellular membrane interface of TM1, TM2, and ECL1. B. Other concentrations of these mutants are found in the lower half of TM1, helix 8, and the TM5-TM6 interface. The blue colored structure represents the shift in TM6 upon adoption of the active state. C. 2-D snake plot with residues colored by the number of individual mutations that lead to greater than WT activity in the no agonist condition. These residues are enriched in the loops and termini which are truncated in the crystal structures. D. Activity of all ADRB2 mutants present in the gnomAD database plotted against to their allele frequency. We classified variants into four categories as follows: null mutants (purple) are variants whose mean plus a standard deviation (SD) are less than 1 (the mean frameshift); activating mutants (orange) are variants whose mean minus a SD are greater than the mean synonymous mutant (dashed line); hypomorphic mutants (periwinkle) are variants whose mean plus a SD are less than the mean synonymous variant; the rest of the variants are considered WT-like (white).
## List of Tables

3.1 **Estimated error frequencies for five-oligo gene assemblies.** Here, we averaged the errors/kb for both five-oligo assemblies using Q5 and KAPA2G Robust polymerases and their technical replicates across each error type (errors are standard error of the mean). We see that all error subtypes are similar except for mismatches. .................................................. 50

3.2 **Examples of where various aligners fail.** Here _ are padding for visualization, * are soft-trimming, and lower-case bases are inserts. ................................. 72

3.3 **Median error rate per position for assemblies using the error-doped oligos or the standard oligos.** We measure significant (Mann-Whitney U, \( p \ll 0.001 \)) differences between the median error rates of the error-doped and standard oligos for all error sub-types. .......................................................... 73

4.1 **Assembly barcode statistics for each serial dilution in the two biological replicates.** Barcodes for each sample were clustered using Starcode [35] to collapse barcodes within a Levenshtein distance of 1. ................................. 136

4.2 **Homologs and GoF mutants retrieved from the assembled library and individually tested in knockout (KO) PPAT cells.** Growth rate (hr\(^{-1}\)) is defined as the maximum slope of OD600 vs. time on a log/linear plot. Wildtype \( E. \ coli \) PPAT and 3 catalytically inactive wildtype mutants were also prepared and tested.137

4.3 **Cost to create pool of 384 barcoded DropSynth microbeads.** Creating the pool of barcoded beads is a one time cost and produces enough beads to carry out at least 210 assemblies of 384 genes, or over 80,000 genes, using the current protocol. . . 139

4.4 **DropSynth assembly costs per 384 gene library.** ................................. 139
4.5 Nick processing efficiencies for various conditions. ................. 140
4.6 The oligos required for the bead barcoding process. All oligos were ordered from Integrated DNA Technologies. .............................. 140
4.7 Primer sequences used in this study. ..................................... 141
5.1 Olfactory receptors screened in this study. ............................. 174
5.2 Odorants screened in this study ............................................. 175
5.3 Odorant-receptor pairs called as hits. .................................. 177
5.4 Primers used in this study. .................................................. 180
Acknowledgments

I could not have completed my thesis without the considerable help of others. First and foremost, I would like to thank my parents for instilling the value of education, and the power of leading a balanced life. Mom, thanks for singing my praises to anyone that would listen, and for forcing me to come up with digestible explanations for my work. Dad, thanks for the advice, support, and for waiting 3-6 months for 1-2 years while I finished up. I’d also like to thank my brother Ryan for dealing with my mess, fronting my considerable tab, and inspiring me to work harder.

I’ve been fortunate to have several amazing mentors who have influenced me along the way. Dr. Paul Yoo, thanks for taking me under your wing almost a decade ago and inspiring me to pursue a PhD. Dr. Lorena Beese, thank you for granting me the independence to take on meaningful projects in your lab. I’m always amazed how MutS keeps appearing in my research and how my career has come full circle since my first docking experiments. I’m also thankful for having two stimulating and influential rotations when I started at UCLA in the Eisenberg and Yeates labs. In particular I want to thank Dr. Lin Jiang for his guidance, advice, and ROSETTA mastery. I also want to thank Dr. Todd Yeates for, inspiring me to polish up on Linear Algebra and for our lively discussions about Reverse Polish Notation.

It goes without saying that none of this would have been possible if without the support, guidance, and mentorship of my advisor Professor Sriram Kosuri. You have profoundly shaped the way I approach research, provided me with countless opportunities, and granted me the freedom to tackle a number of different projects. I hope to emulate your creativity, quickness, and ability to embolden your students do audacious science.

Special thanks are in order for the Founding Fathers of the Kosuri Lab – Rocky Cheung, Eric Jones, and Hwangbeom Kim. Rocky, you are possibly the most unique human I’ve ever met and
I cherish the spiritual guidance you’ve given me over the years. Congratulations on the job and I’m excited to see what’s in store for you and Oscar. Jones, it’s a special privilege to be able to work with someone who is such a loyal, caring, and close friend. I’m looking forward to our coming adventures in the Pinnacle of Innovation. Hwangbeom, it was an honor to work with you. Your work ethic, ingenuity, and post moves were a constant source of inspiration for me. I hope you’re ready for our visit!

I’m also grateful to have worked with so many amazing people in the Kosuri Lab. Angus, thanks for saving me from having to get back into the lab for the error correction reviews. I’m looking forward to more chorizo, saunas, fresh powder, oligo peddling, and some wholesome family time. Calin, thanks for letting me bounce a bunch of half-baked ideas off you. I’m constantly inspired by your creativity, passion, and productivity. I can’t wait to see where your faculty search takes you. Christina, thanks for letting me pick your brain for all things Python related. Cliff, thanks for always humoring my wild speculation for stock movements and for having the Tapatio on deck. I’m sorry those three arrows on my CMB proposal ended up causing you such a headache. Guillaume, it’s been amazing to watch transform into a full-stack biologist as your computational skills have grown over the years. I’m excited to see what you’ll tackle next. Jess, I’m constantly impressed by your thoughtfulness and attention to detail when it comes to your research. Also thanks for the Soylent. Kim, thanks for helping keep my bash-fu sharp and my data tidy. The corner office is finally yours. PS you’re sudo now. Rishi, thank you for all of the stats/analysis advice, and for inspiring me to approach problems from first principles. Best of luck in the faculty search.

Special thanks also are in order for our Lab Managers/Research Assistants David, Danny, Suraj, and Jeff. Without you the lab would fall apart. Thank you to the wonderful undergrads Megan, Marcia, Tripp, and Joyce for bringing energy and enthusiasm to balance out us surly grad students, and Johnny for the spicy memes. Thank you to the people I mentored over the years – Alberto, Fay, Chari, Maeve, and Zach for all of your help and for keeping the BAGEL-seq dream alive. I’d also like to thank Michael for his assistance in the early BAGEL-seq days.

Thank you to Professors Todd Yeates, Leonid Kruglyak, Matteo Pellegrini, Jose Rodriguez, and Sabeeha Merchant for serving as my committee members and for their guidance over the course of my studies. I would like to thank my collaborators Di Zhang for his instrumental experiments on
both the error correction and DropSynth projects, and the Dror Lab at Stanford, particularly AJ Venkatakrishnan, for their $\beta_2$-adrenergic Receptor expertise.

I am especially in debt to the computational wizards Duilio Cascio, Luki Goldschmidt, and Alex Lisker. Thank you for the countless hours of help keeping our servers running despite my best attempts to brick them. I’d also like to thank Mark and Sum from the Protein Expression Technology Center for their help purifying Tn5 during my foray into the wet lab. Thank you also to Joshua Bloom and Aaron Cooper for your help with all things quantitative.

I would also like to acknowledge Jonathan Lowenson and the Cellular and Molecular Biology Training Grant program at UCLA, which is funded by the Ruth L. Kirschstein National Research Service Award GM007185.

I’d like to thank the team at Octant for transforming our projects into something special, for the LaCroix, and for somehow agreeing to hire me... I’d also like to thank Jason Kakoyiannis for his mentorship, Pizzana habit, and his infectious enthusiasm for the potentials of Synthetic Biology.

I’d also like to thank the people who have been there for me outside of the lab. Thank you Jeff for being my domestic partner, lifting buddy, and occasional shooting coach. George, I’m not quite a starving grad student thanks to your stock picks. Thank you to Cameron and Chase, for allowing me to golf in your presence and indulging in my Mai Tai habit. Thanks Dave for the latest biotech happenings and for the acts that I should see. Thank you to the Dukies Rohan, Hersh, Dylan, Brandon, Derek, Ray, Alex, Nam, and Bri for fielding my calls during rush hour and for giving me an excuse to get out of LA. I also want to send a special thanks to Catherine, Chris, Grace, Wenting, Costa, Rohan, Adriana, Andrew, Pia, Danny the Dog, and the rest of the LA wrecking crew for the late nights, camping adventures, Wine Wednesdays, and Friendsgivings. You’ve made my time in LA truly special.

Keely, thanks for everything. You’ve been by my side from basically the beginning, and I’m certain I couldn’t have done it without you. Words cannot express the gratitude I feel for everything that you’ve done for me.


Vita

EDUCATION

2013 B.S. (Chemistry), Duke University, Durham, North Carolina
2015 M.S. (Biochemistry), UCLA, Los Angeles, California

RESEARCH AND WORK EXPERIENCE

2014-2018 Graduate Researcher, Kosuri Lab, UCLA
2017-present Consultant, Bioscentric, Los Angeles
2017-present Consultant, Octant, Los Angeles

AWARDS AND HONORS

2014-2017 Ruth L. Kirschtein Pre-Doctoral National Research Service Award

PATENTS


PUBLICATIONS


Chapter 1

Introduction

“It is the writer who might catch the imagination of young people, and plant a seed that will flower and come to fruition.” — Issac Asimov

1.1 Background

Literacy is the fundamental underpinning of all human progress. The invention of writing enabled humans to store information over generations, and our ability to read enabled us to learn from it. The concept of literacy is even found in Biology itself. Life evolved its own language as outlined by the Central Dogma – genes are written in DNA, transcribed to RNA, and then translated into proteins. While we have made considerable progress understanding and engineering Biology, our ability to read and write DNA is still in its infancy.

By the early 1960’s, researchers understood the basics of the Central Dogma, but did not know its actual molecular mechanism. In particular, the question of how cells translated RNA into amino acids remained. To solve this problem, Har Gobind Khorana and colleagues developed methods to synthesize defined DNA oligonucleotides (oligos) [1]. They systematically synthesized all possible codons and used RNA polymerase to transcribe them to RNA. Extending the experiments of Marshal Nirenberg, they combined radio-labeled amino acids, the transcribed mRNAs, and translation machinery extracted from E. coli together. By measuring the amount of radioactive amino acid incorporation, they were able to associate each codon to its amino acid, and cracked the genetic
code in 1965.

Reading DNA proved to be much harder than writing it. In the 1950’s Fredrick Sanger invented a method for sequencing proteins by degrading the peptide strand one amino acid at a time and characterizing the results [2]. Researchers extended this idea to RNA in the 1960’s through sequential exonuclease digestion [3]. However, DNA sequencing by degradation was feckless, as it was too slow to scale to molecules longer than ~50 nucleotides. In 1977, almost 20 years after receiving the Nobel Prize for protein sequencing, Fredrick Sanger published an eponymous DNA sequencing method that enabled researchers to effectively read DNA [4, 5, 6]. Briefly, Sanger Sequencing leverages trace amounts of chain-terminating, fluorescently-labeled nucleotides to halt a primer extension reaction at every position in the sequence. The terminated fragments are separated with electrophoresis and the identity of the terminal base is determined from its fluorescent signal. Finally, the original sequence is recovered by concatenating the terminal bases together.

Since then, progress in both DNA synthesis and sequencing has proceeded exponentially (Figure 1.1). The development of solid-phase phosphoramidite chemistry by Beaucage and Caruthers in the 1980’s enabled the robust, scalable, and automatable synthesis of oligos [7]. In the early 1990’s, researchers modified the phosphoramidite synthesis process to enable oligo synthesis on the surface of microchips, vastly increasing the number of oligos synthesized in a single run [8, 9]. Today, researchers can purchase oligo libraries of up to a million members for $<0.00001 – 0.0001 per nucleotide [10].

Progress in DNA sequencing has improved even more rapidly. Spurred by the Human Genome Project, a number of major breakthroughs in DNA sequencing led to the development of Next-generation Sequencing (NGS) in the mid-2000’s [11, 12, 13, 14]. These instruments primarily differ from Sanger Sequencers in that they track the incorporation of each nucleotide, rather than measuring fragment lengths. In addition, they are highly multiplexed, and are now capable of reading billions of sequences simultaneously. These advances lead to an astounding rate of progress, with the cost of sequencing falling 10x every year between 2005-2010.

While these advances are nothing short of a revolution, our reading and writing capabilities are still limited in many respects. For reading DNA, it is still challenging to sequence long, contiguous pieces of DNA with the current NGS platforms. This has limited our understanding of many regions
The cost of DNA Sequencing and Synthesis is decreasing exponentially. The advent of Next-generation Sequencing in 2005 resulted in \( \sim 10^6 \)-fold decrease in price in 10 years. Prices are reported per base sequenced or synthesized. At the time of writing (November, 2018) the price per base of microarray-base oligos is approximately \$10^{-4} - 10^{-5}\) (data not shown). Figure ©Bioeconomy Capital http://www.bioeconomycapital.com/bioeconomy-dashboard/

in the genome, and has hampered the sequencing of novel genomes [15, 16]. For writing DNA, this is most apparent in gene synthesis, or the construction of gene-length fragments from synthetic oligos [17]. Currently, synthesizing even multi-kilobase genes remains challenging. Large-scale applications such as engineering synthetic gene pathways or even entire organisms require would require similar million-fold improvements to those brought about by NGS [18]. Lastly, these limitations in reading and writing DNA have hampered progress in an emerging technique know as multiplexed functional assays [19].
Figure 1.2: **Typical gene synthesis workflow.** An oligo library that tiles the gene of interest are synthesized and optionally purified. Using polymerase cycling assembly (PCA), overlapping oligos anneal together and the gaps are filled by a polymerase. After PCA, the full length product is amplified by PCR and optionally subjected to a round of error correction. Finally, the synthetic gene is cloned and verified by Sanger Sequencing. ©atdbio https://www.atdbio.com/content/63/Gene-synthesis

1.2 Gene Synthesis

Current methods in gene synthesis use $\sim 40 – 150$ nucleotide (nt) oligos to assemble longer genes (Figure 1.2)[17, 20]. Most often, these oligos are designed to have complementary overlapping sequences. In a PCR-like reaction known as Polymerase Cycling Assembly (PCA), a polymerase fills in the gaps between oligos and generates the full-length gene of interest. The full-length product is then amplified out of the reaction in a separate PCR reaction.

However, the error rate of oligo synthesis makes generating error-free genes almost impossible. For example, imagine we assembled a gene from 10 oligos that are 99% pure. The percent perfect assemblies would be $90.44\% = 0.99^{10}$. In reality, most column oligos today are $\sim 50\%$ pure, meaning
we can expect 0.098% perfect assemblies \[^10\]. This problem is exacerbated with cheaper, but lower quality array-based oligos. Thus, researchers have developed a number of error correction schemes to enable practical gene synthesis \[^21\].

Broadly, error correction methods can occur either at the level of the oligos, or on the synthesized genes themselves. The most common oligo-based error correction methods use either HPLC or PAGE to remove truncated oligos from the input stock \[^22, 23\]. Other more exotic methods use hybridization \[^24, 25\] or even sequencing to select oligos with perfect sequences \[^26, 27, 28\]. However, the most cost-effective and widely used error correction method employs a variety of different enzymes at the gene-level.

Enzymatic error correction leverages classes of proteins that recognize distortions in the DNA helix brought about by mishybridized bases \[^29, 30, 31, 32\]. In a typical experiment, researchers will denature and re-anneal their gene synthesis product, a mixture of perfect and imperfect assemblies, to expose errors through mishybridization. Next, mismatch binding proteins such as MutS are used to filter out imperfect assemblies \[^33\]. Alternatively, mismatch cleaving proteins such as T7 Endonuclease I recognize and cut at any errors \[^34\]. Perfect sequences can be recovered either through size selection or by a second round of PCA.

Lastly, a number of researchers are attempting to move away from the phosphoramidite chemistry given its inherent limitations. One of the most promising of these techniques leverages the enzyme terminal deoxynucleotidyl transferase (TdT) to synthesize oligos \[^35\]. Discovered in 1959 by F. J. Bollum, TdT randomly incorporates nucleotides in a template-independent manner, and can even produce oligos up to 8000 nt long \[^36, 37\]. Given these properties, F. J. Bollum proposed that it could be used for the synthesis of oligos with defined sequences in 1962 \[^38\]. However, controlling the incorporation of defined nucleotides has proven to be challenging, and the first practical demonstration occurred late 2018 \[^39\]. In it, researchers synthesized a 10 nt oligo with step-wise incorporation efficiencies ranging from 93.5 to 99.5%. While this is still worse than traditional oligo synthesis, further optimizations could make enzymatically synthesized oligos a reality.
Figure 1.3: **Typical dilution-based SLR overview.** In a dilution-based SLR, or haplotyping experiment, molecules of interest are diluted and amplified in 96- or 384-well plates. The dilution ensures sufficient diversity within each well and that haploids remain separated. The molecules are then fragmented and tagged (tagmented) with a unique barcode for each well. After sequencing, these barcodes are used to reassemble the molecules within each well. Due to the diversity of each well, it is unlikely, but not impossible that reads from one molecule could be incorrectly assigned to a different molecule.

### 1.3 Synthetic Long Reads

While NGS has undoubtedly revolutionized Biology, today’s sequencers still have a number of drawbacks. For Illumina, the *de facto* NGS platform, this is read length. At the time of writing, their sequencers can only read up to 600 bp at a time, and are unlikely to be able to read much more than that with their current technology. Since Illumina has a near monopoly on NGS, progress in areas that require long reads, such as genome assembly or haplotyping, has arguably been limited.

Recently, real-time single-molecule sequencers from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have emerged as potential competitors to Illumina. As the name suggests, real-time single-molecule sequencers directly read single molecules of DNA in real time. PacBio sequencers directly record the color of fluorescent dyes cleaved from derivatized nucleotides as they are incorporated into a single molecule of DNA in real time [40]. Alternatively, ONT detects changes in current as DNA molecules are pulled through a biological nanopore [41, 42]. Both methods have demonstrated read lengths of up to 100,000 bp, but have significantly higher error rates, lower throughput, and are more expensive than Illumina sequencing [43, 44].

Due to these limitations, researchers have developed a number of methods known as synthetic long reads (SLRs) as a compromise between the accuracy of Illumina sequencing and the read length of single-molecule sequencers. In general, SLRs use molecular barcoding techniques to computationally stitch together short reads into long, contiguous molecules [45]. The most common SLRs extend
methods used for molecular haplotyping, where molecules are segregated into compartments, diluted, fragmented, and tagged with a unique barcode for every compartment (Figure 1.3) [46, 47]. The dilution ensures that haplotypes remain separated, and aids in the barcode-guided reassembly of the molecules in the compartment.

However, the physical dilution step is cumbersome, inefficient, and does not scale well. In addition, PCR amplification of highly dilute molecules is often biased. Alternative SLR technologies that do not require dilution address some of these issues [48, 49, 50]. Nonetheless, every SLR technology to date relies on PCR to amplify long molecules, capping the maximum SLR length to approximately 10 kb. They also require an extreme amount of read depth to ensure each molecule is completely covered by reads, adding to the cost. Finally, no method currently exists that maintains a one-to-one linkage between a unique barcode and a single molecule. Thus, applications that require long, accurate reads will be hampered until single-molecule sequencers improve or better SLR methods are developed.

![Diagram](image-url)

Figure 1.4: **Overview of a hypothetical multiplexed functional assay (MFA).** In this MFA, the activity of a given variant in the library is linked to a genetic reporter through a molecular barcode. Agonist stimulation of the expressed variant results in a signalling cascade that activates the transcription of its unique barcode. The level of activity can then be quantified by measuring the expression of the barcode with RNA-seq.
1.4 Multiplexed Functional Assays

Understanding sequence-function relationships is a central goal of Biology. Typically, the functional consequences sequence changes are tracked in a one-by-one manner. For example, large-scale mutagenesis experiments such as alanine-scans require the time consuming cloning and testing of each individual mutant. Even the massive drug screening efforts by large pharmaceutical corporations measure drug binding through individual luciferase assays. However, by leveraging NGS and DNA microarray synthesis researchers are now capable of probing sequence-function relationships on unprecedented scales [19].

These experiments are broadly termed multiplexed functional assays (MFAs; Figure 1.4). First, researchers design and build libraries of genetic variants using DNA synthesis. Each variant is associated with a unique barcode that encodes its identity. Next, researchers deliver these variants either episomally or stably. Importantly, it is now possible to assay these variants in a native genomic context using recently developed genome engineering tools [51]. Finally, the functional consequences of these variants is quantified by measuring the abundance of the variant barcodes. Using this framework, researchers have used MFAs to investigate sequence-function relationships across the Central Dogma, such as promoters, regulatory elements, untranslated regions, splicing, and protein function [19]. One particular area where MFAs could be beneficial but have yet to been applied is drug development.

Progress in drug discovery has halted. Adjusted for inflation, the number of new FDA approved drugs per billion dollars of R&D spending has halved every 9 years since 1950 [52]. One way to address this issue is to screen drugs against un- or under-tested targets, as the vast majority of the “drugable genome” has yet to be investigated [53, 54]. This problem is exemplified in G protein coupled receptors (GPCRs), which are the targets of ~34% of all FDA approved drugs [55]. Only ~100 of the ~400 non-olfactory GPCRs are currently drugged, with > 65% of the drugs targeting only 36 receptors [56]. Additionally, new insights into the role of ~400 olfactory GPCRs outside olfaction suggest that they could also be valuable drug targets [57].

However, screening GPCRs in a comprehensive manner remains technically challenging. Their diverse role in modulating cellular physiology requires they operate through a number of downstream effectors. Thus, bespoke reporters are required to measure the activity of different classes of GPCRs.
Alternatively, the vast majority of GPCRs recruit $\beta$-arrestin upon receptor activation [58]. Recent techniques have leveraged this property to enable high-throughput, genome-wide screening of GPCR activity [59, 60, 61]. Regardless, both $\beta$-arrestin recruitment and traditional GPCR reporter assays typically use fluorescent read-outs, thereby prohibiting multiplexing and hampering scale.

Multiplexing GPCR screening technologies could enable large-scale screening of all GPCRs. This could be achieved by adding molecular barcodes to existing genetic reporters of GPRC activity. The abundance of each of these barcodes would correlate with GPCR activity and could be quantified in multiplex by RNA-seq. However, a number of technical challenges must be addressed before multiplexed GPCR screening is realized. First, one would have to extend existing GPCR genetic reporters to have barcoded outputs. Second, individual reporters and their target GPCR would have to be integrated concurrently into individual cell lines to avoid cross-activation by other GPCRs in the library. Lastly, these cell lines will likely need to be further engineered to enable robust expression of diverse GPCRs.

1.5 This Work

Here we describe methods for improving out base capacity to read (Chapter 1) and write (Chapters 2 & 3) DNA. We then leverage these capabilities to investigate protein function over evolutionary space (Chapter 3) and to test the function of GPCRs in multiplex (Chapters 5 & 6).

Chapter 1 describes a method called BAGEL-seq that generates synthetic long reads. In a proof of principle experiment, we achieved read lengths that exceed the capability of any Illumina sequencer. We also provide commentary on further optimizations and experiments, as this method remains in development.

In Chapter 2, we develop novel methodology to accurately measure error rates in DNA sequences using NGS. We use this method to characterize the most commonly employed enzymatic error correction methods in gene synthesis, and estimate the error rates of different polymerases.

Chapter 3 details a multiplexed gene synthesis method called DropSynth and used it to synthesize $>10,000$ genes of up to 669 bp in length. We then tested these genes in a multiplexed functional assay and explored the evolutionary and functional landscape of an essential enzyme in $E$. 
In Chapter 4, we present a method that enables the stable expression and screening of olfactory receptors in multiplex. We test hundreds of chemicals against 39 murine olfactory receptors and use recently developed chemical learning tools to begin to understand the chemical features that these receptors are recognizing.

Finally, in Chapter 5 we test 99.6% of all possible missense variants of the $\beta_2$-adrenergic receptor across several drug conditions in multiplex. We use unsupervised learning methods to identify known and novel mutants, including a “structural latch” that is conserved across Class A GPCRs.

References


Chapter 2

BAGEL-seq – A Novel Method for Generating Synthetic Long Reads

2.1 Abstract

Next-generation sequencing (NGS) has revolutionized our understanding of biology. However, NGS is still a nascent technique and consequently, the base technologies are limiting for many applications. In particular, the read length of most NGS platforms (<500 nucleotides) is too short, and the error rate of long read NGS technologies is too high, making it difficult to assemble and haplotype genomes. Synthetic long reads (SLRs) are a recently developed method that represent a compromise between the accuracy and throughput of current short read technologies, and the read length of long read technologies. SLRs are generated by using molecular barcodes to computationally assemble short reads back to the individual longer fragments of DNA from which they came. However, all current SLRs are limited to ~10 kb by PCR. Here we bypass this limit with a novel library preparation that can theoretically produce SLRs that are ~40 kb or greater.
2.2 Introduction

Long Read Technologies

The impact of Next-generation Sequencing (NGS) technologies on our understanding of genetics and biology is staggering [1, 2, 3]. Consequently, NGS technologies are now a fundamental tool in the analysis of genetic information across a broad spectrum of fields [4, 5]. However, these technologies are still in their infancy and suffer from a number of drawbacks. Of particular concern is the inability to read long (>10,000 bp), contiguous pieces of DNA accurately. This is most apparent in de novo genome assembly and haplotyping experiments, where long reads are required to span highly repetitive or variable regions of the genome, or are needed to phase sparsely populated heterozygous variants [6, 7].

Traditionally, long reads have been limited to single molecule sequencing platforms such as those developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). As the name suggests, single molecule sequencers aim to directly read the sequence of long DNA molecules, in contrast to the shotgun style approach of Illumina. In particular, the PacBio sequencer directly images fluorescent dyes cleaved from labeled nucleotides as they are incorporated into a single molecule of DNA. This process is parallelized by separating the reactions in zeptoliter ($10^{-21}$L) reaction chambers [8]. In contrast, ONT uses an array of biological pores conjugated to an integrated circuit to detect current changes brought about by individual nucleotides in parallel [9]. The theoretical maximum read length of PacBio sequencers is limited by the polymerase, while the read length of ONT sequencers is limited only by the length input DNA [10]. Although PacBio and ONT have demonstrated read lengths longer than 10,000 bp, the error rates are typically too high for practical use [11, 12]. Consequently, researchers leveraged the accuracy of existing short-read sequencers (Illumina, Ion Torrent, etc.) with molecular barcoding techniques to computationally stitch together short reads into long, contiguous molecules known as synthetic long reads (SLRs).

The first SLR technology, termed subassembly, was developed by Hiatt et al. in 2010, and extended the concept of hierarchical shotgun genome assembly [13]. Subassembly works by first ligating barcoded adapters to gDNA fragments, which are diluted then amplified by PCR. This effectively creates sub-libraries of a limited number of fragments. Next, the amplicons are concatenated and
fragmented, producing random break-points. A second primer is ligated to these break-points, and a sequencing library is prepared by amplifying on the barcode and break-point adapters. One read from a paired-end read is a barcode that identifies groups of reads originating from the same fragment, while the other is a small portion of a random portion of this fragment. Reads with the same barcode are then collapsed together into a long read in silico. The authors were able to generate ~700 bp reads from 36 bp paired end reads, on par with the state of the art pyrosequencing platforms of the time. However, subassembly could only produce SLRs on the order of ~1 kb due to limitations of the length of sequence that can fit on an Illumina flow-cell [14].

Since then, two different library preparations for generating SLRs have emerged. The first, leverages intramolecular circularization to bring a random sequence next to a barcoded primer. Specific implementations vary slightly in the way they barcode fragments and generate random ligation points. For example, Tile-seq and BAsE-Seq (Barcode-directed Assembly for Extra-long Sequences) use PCR to barcode and amplify targeted regions of various genomes. They then use exonuclease digestion to generate random sequences to bring in conjunction with the barcoded sequencing primer [14, 15]. These methods are limited to known sequences, and the resulting coverage is lessened in the 3′-end due to ligation bias. A more recent implementation developed by Stapleton et al. addresses these two fundamental weaknesses. First, they ligate two independent barcoded primers to either end of randomly sheared fragments, eliminating the need to know the sequences a priori. After amplification, a single break (on average) is introduced enzymatically, separating the two barcodes. Next, two fragments undergo an intramolecular circularization separately, and are subsequently fragmented and sequenced. Finally, a secondary sequencing step maps the two barcodes together, allowing the researchers to combine data from each of the barcodes, eliminating much of the ligation bias [16]. This method is able to generate reads of up to ~10 kb, but is unlikely to generate much higher due to the length limits of PCR and the inefficiency of longer intramolecular ligations.

In 2013, Stephen Quake’s group developed an alternative library preparation for generating SLRs and quickly spun out a company named Moleculo. Moleculo’s technology bypassed the need for circularization or concatenation, but does not barcode every molecule uniquely [17, 18]. Instead, their method amplifies, barcodes, and converts dilute pools of hundreds to thousands of molecules
into NGS libraries in plate format. The libraries are then pooled, sequenced in bulk, and the content of each well is reconstructed \textit{in silico}. Unlike similar dilution-limited haplotyping methods (discussed below), this method effectively over-sequences each well to ensure that every molecule can be reconstructed [17]. This method has been the most widely adopted SLR technology to date, likely due to its ability to generate SLRs from arbitrary sequences. As with previous SLRs, the length Moleculo’s SLRs are still limited by PCR. Additionally, this method is particularly susceptible to PCR biases and errors imparted by the highly dilute nature of the amplification step. Finally, Moleculo’s 384-well protocol is much more difficult and cumbersome than microtube based methods.

**Direct Haplotyping with NGS**

Haplotypes are the unique collection of mutations that co-occur along single chromosomes, and are crucial for understanding both human disease and genetics. For example, haplotype information is critical for determining the phenotype of many diseases, is used to increase the power of genome-wide association studies (GWAS), and is important for studying human history and migration [19, 20]. However, it has been difficult for researchers to directly obtain accurate haplotype information with short read sequencers. Consequently, inferential methods that use populations or pedigree information to impute haplotypes have been utilized more extensively [21]. Nevertheless, inferential methods are often unable to detect low-frequency, \textit{de novo}, and private variants; and are limited by linkage disequilibrium [20]. As a result, there has been a resurgence in the development of direct haplotyping methods over the last few years that leverage increased sequencing power.

Many direct haplotyping methods rely on variations of the same technique (Figure 2.1), which I will refer to as dilution-limited haplotyping for convenience [22, 23]. In essence, high molecular weight (HMW) gDNA is diluted into compartmentalized pools such that each pool is sub-haploid in content, or genomic fragments are either represented once or not at all same pool. Although each pool is sub-haploid, there are enough pools to cover a given genome sufficiently. Next, the pools are barcoded, amplified, combined together, and sequenced in bulk. Reads corresponding to each barcode are then assembled back together into longer blocks of coverage corresponding to a single haplotype. Since haplotype variants are often spaced multiple kbs apart, it is crucial that input DNA is as long as possible so that the size of these reconstructed blocks spans as many variants as
A. **Fosmid Library**

B. **Dilution and Amplification**

C. **Virtual Compartments**

---

Figure 2.1: **Dilution-limited haplotyping overview.** Current methods rely on various techniques to separate long pieces of DNA into various compartments, where they are amplified, barcoded, and prepared for NGS. **A.** Historically, most haplotyping libraries were generated with fosmids. Briefly, genomic DNA is sheared and ligated with a fosmid backbone. Phages then package the ∼40 kb molecules and infect *E. coli*. The resulting library is diluted and grown in 96-well plates to amplify the selected DNA (purple and blue).

**B.** Alternatively, the sheared genomic DNA can be diluted and amplified *in vitro* with multiple displacement amplification. The resulting product is then sheared and prepared for sequencing. **C.** Recently, a number of droplet-based technologies have been developed that obviate cumbersome plate-based protocols. These techniques separate the sheared genomic DNA into droplets containing Tn5 conjugated to beads, and perform an in-droplet tagmentation reaction.

These blocks are then fed into phasing algorithms that attempt to order the blocks into two different haplotype assemblies. Finally, the completeness of the resulting assemblies are often assessed by their N50, or the (in the context of haplotyping) “the smallest haplotype block in which the sum of that block and all larger blocks total to 50% (by length) of the complete haplotype assembly” [20].

Dilution-limited haplotyping methods differ primarily in the manner by which they compartmentalize and amplify the sub-haploid gDNA. Clonal techniques, like those pioneered by Burgtorf and Kitzman encapsulate 40 kb fragments in fosmids or 50-250 kb fragments in bacterial artificial chromosomes (BACs) before outgrowth in *E. coli* and pooling in plates (Figure 2.1A) [23, 24]. Alternatively, *in vitro* techniques partition raw gDNA into plates before amplification by multiple displacement amplification (MDA) [25, 26] or long range PCR (Figure 2.1B) [18]. These methods have a number of trade-offs. While clonal methods require complex and time consuming library preparations that do not scale for multiple samples, they generate the highest N50’s [7]. In contrast, *in vitro* methods (Figure 2.1B) have a lower experimental burden and are less expensive to perform, but are limited by their amplification methods. In the case of MDA, very deep sequencing (200-500
Gb) is required to rectify biases inherent to the amplification process [25, 26]. Long range PCR is less susceptible to these biases, but limits the input fragment length to <10 kb, making longer haplotype assemblies more difficult [20].

Importantly, the amount of dilution required to keep maternal and paternal DNA separated is proportional to the number of total compartments. The physical compartments of a plate are thus a major limitation to scaling these methods. To address this, researchers have devised a number of alternative methods that avoid cumbersome plate-based protocols. The simplest of these uses Tn5 conjugated to beads to distribute copies of a barcode along a long DNA molecule [27, 28, 29, 30]. The reaction is carried out such that one or few molecules associate with a single bead, creating a “virtual” compartment. By using $10^5 - 10^7$ beads, these methods obviate the need for dilution and amplification. Alternatively, microfluidic methods physically separate DNA and enzymatic mixtures into droplets, but they require expensive and complicated machinery [31, 32].

Eventually, SLRs could be useful for haplotyping, but as of now, their read length is too short. For example, before supplementation with population based phasing, Moleculo achieved haplotype blocks of $\sim$60 kb [18]. In contrast, similar dilution based haplotying methods have achieved $N50$s of up to $\sim$700 kb [26]. In an analysis of the parameters that underlie a successful haplotyping experiment, Lo et al. found the fragment length of DNA to be the key limiting factor for haplotype contiguity [7]. This likely explains why SLR methods with fragment lengths of $\sim$10 kb are unable to compete with MDA or clonal based haplotype methods that have fragments of 50-100 kb. Regardless, SLR methods have the distinct advantage of being able to directly phase multiple variants on a single read without the need for cumbersome dilution or cloning steps. Thus, an SLR method that could increase its read length to be on the same order of the fragment length of existing haplotyping methods would likely be competitive in this space.

**De Novo Assembly**

*De novo* assembly is the process of assembling an organism’s genome from sequencing data without any *a priori* information about its content. Like haplotyping, *de novo* assembly involves the assembly of short reads into longer contiguous pieces (contigs). Using connectivity information from mate-pair libraries or other techniques, these contigs can then be ordered into a genome. Longer contigs ease
the assembly process by better spanning highly repetitive genomic regions that are hard to map, and by lessening the long-range information needed to accurately order the contigs [33]. De novo assembly is in some ways easier than haplotyping, as genomes can be assembled into a single contig when the read length exceeds the longest repeat or structural variant in the genome. In contrast, haplotyping requires long stretches of heterozygous variants that are typically spaced ∼1.5 kb apart in order to stitch together haplotype blocks [7]. Nevertheless, many genomes are still incomplete as mapping these repetitive regions remains a challenge. Consequently, a deep understanding of these regions is limited, and tremendous amounts of effort must be expended to rectify them [6, 34, 35].

Typically, genomes are assembled de novo by using a combination of NGS library preparation techniques – often traditional whole genome sequencing (100-600 bp fragments), mate-pair libraries (2-20 kb inserts), and fosmid libraries (∼40 kb fragments) [36]. Since standard whole genome sequencing reads are often shorter than the structural variants and repeats they are attempting to span, longer-range connectivity information is needed to properly map the shorter reads. Mate-pair libraries provide this information by linking together ends of fragmented gDNA to a specially designed tag, forming a circle [37]. This circularized construct can then be enriched and converted into a sequencing library, with each of the resulting paired-end reads representing the beginning or end of the original fragment. Since the distance between these two fragments is known to an approximate degree, assembly algorithms can constrain contigs to certain regions. Fosmid libraries can also provide mate-pair like linkage information [38], but are often reconstructed into longer reads to span gaps in the whole genome assembly instead, or act as scaffolds on which to map short reads [39].

As long-read sequencers have matured, researchers have begun using them in-place of fosmid and mate-pair libraries for many assembly applications [40, 41]. This circumvents the labor-intensive mate-pair and fosmid library preparation processes and is particularly useful in finishing difficult regions. The error rate of these technologies is still too high to be used for the de novo assembly of larger genomes, but they have demonstrated success with smaller bacterial genomes [42, 43]. Alternatively, researchers have employed previously mentioned SLR technologies such as Moleculo to assemble genomes de novo [17, 44]. However, in a direct comparison, PacBio was able to generate a hundreds-fold more continuous assembly of the Drosophila genome [45]. This was attributed in part by PacBio’s increased read depth (90x vs 34x), but was mainly due to their longer reads (mean, max:
\( \sim 9, \sim 45 \text{ vs } \sim 4.5, \sim 15 \) \[45\]. As PacBio is currently limited by throughput and error rate, newer SLR methods with longer read lengths would be immediately beneficial for the de novo assembly.

Figure 2.2: General overview of BAGEL-seq. In the current iteration, we first clone a sequence of interest into a backbone containing a sequencing primer (purple), a unique barcode (blue), and a LoxP site. Next, we use dilute amounts of Tn5 transposase to randomly insert a single copy of a complimentary LoxP site on average throughout our sequence of interest. These LoxP sites will be inserted either in cis (left), trans (right), or not at all (middle). Next we dilute the sample and use Cre recombinase to perform intramolecular recombination between the two LoxP sites. This brings our sequencing primer and barcode adjacent to wherever Tn5 inserted the second LoxP site. Importantly, both excision (left) and inversion (right) result in readable sequence (orange). Finally, we tagment the products with Nextera and use PCR to amplify our sequences of interest.
BAGEL-seq

With their low error rates, SLRs should perform better than existing technologies for applications requiring long reads. However, direct comparisons against current haplotyping methods [20] and PacBio reads [45], reveal that the read length of current SLRs is too short. Since every SLR method to date relies on PCR to amplify its initial library, their maximum read length is limited to <20 kb. Moleculo has the additional limitation of not barcoding each molecule uniquely, making read assembly more difficult. This results in a bimodal read length distribution with a peak at 8-10 kb and a mass of shorter sequences. Many of these come from the library preparation itself, but some come from repetitive regions of the genome that cannot be mapped with short reads [46]. To increase read length, any future SLR method will need to bypass the length restriction of PCR. Additionally, uniquely barcoding each molecule will ease downstream read assembly, and unlock a variety of applications. We will address these issues with a novel library preparation.

Termed BAGEL-seq, or Barcode-directed Assembly of Genomes Enabled by LoxP, our method generally works as follows (Figure 2.2). First we, clone random fragments into a backbone with a unique barcode, a sequencing primer, and a LoxP site. For large ∼40kb fragments, this could be accomplished with a fosmid. After clonal amplification, we then use Tn5 transposase incorporate a second LoxP site at random positions along each copy’s sequence. Subsequent incubation with Cre recombinase results in a recombination even between the two LoxP sites, juxtaposing our sequencing primer and barcode next to the random sequence adjacent to the second LoxP site. This is functionally equivalent to primer walking every original molecule in multiplex, given that Tn5 integrates the second LoxP sites uniformly throughout their sequences (Figure 2.3). We then fragment and add an additional sequencing primer with Nextera. We can select for our sequences of interest through PCR amplification with the existing primer and the Nextera adapter. Finally, we group reads together from each barcode and assemble the original molecules computationally.

We believe BAGEL-seq has a number of practical advantages over existing SLR technologies. First, BAGEL-seq’s read length is not limited by PCR. This theoretically allows researchers to create synthetic long reads that are only limited by Cre-recombinase’s ability to circularize DNA. In the current embodiment, this comes at the cost of having to clone the initial library or use fosmids for extremely long constructs. Second, BAGEL-seq is not as susceptible to circularization biases
as other library preparations. These methods use a ligase under dilute conditions to circularize constructs. However, it is much more efficient to circularize shorter constructs (to a point), resulting in a bias towards shorter SLRs [15]. In contrast, Cre exhibits minimal bias towards circularizing small fragments and has practically unlimited range; researchers have used Cre to circularize ~ 20kb constructs *in vitro* [47, 48, 49], and ~ 100kb bacterial artificial chromosomes *in vivo*. Lastly, BAGEL-seq will not require cumbersome plate-based protocols [18, 27] or precisely timed digestion time courses [14, 15, 16].

Figure 2.3: *Illustration of BAGEL-seq primer walking.* Here, Tn5 has inserted LoxP sites at different locations in three copies of the same sequence. After Cre recombination, the sequencing primer (purple) is brought adjacent to wherever the second LoxP site was inserted. In addition, the sequencing primer reads through the barcode, maintaining the one-to-one linkage between barcode and molecule throughout the process. Lastly, the barcode enables reconstruction of the individual molecules by *in silico*.

2.3 Results

Tn5 Insertion

Although the final realization of BAGEL-seq will generally consist of the steps above, we have yet to determine its optimal reaction conditions. With respect to Tn5, we must tune its concentration so that performs a single insertion on average per molecule. Multiple insertion events will fragment our input DNA *a la* Nextera. In our initial tests, we performed a serial dilution of Tn5 loaded with LoxP and found that the resulting DNA was not fragmented at high concentrations (Figure 2.4 A).
Tn5 is known to insert into any double stranded DNA longer than $\sim 34$ bp, as this is approximately the length of DNA that is protected by Tn5’s molecular footprint [50]. At 34 bp, the LoxP sites are right on the edge of this length restriction. However, Tn5 requires a 19 bp recognition sequence (the ME sequence), bringing the total length of the inserted construct to 53 bp. We hypothesized the $\sim 20$ bp of extra DNA could serve as a substrate for Tn5 insertion, interfering with insertion into our target DNA.

Researchers have addressed this issue by loading Tn5 with Y-adapters. As the name might suggest, Y-adapters are formed by annealing two unevenly sized oligos together. In this case of Nextera, the shorter oligo only contains the ME sequence, while the longer oligo contains the ME Sequence and a 15 nt priming sequence. Unfortunately for BAGEL-seq, the LoxP site is an inverted repeat and is likely to be double stranded. We addressed this by extending the shorter oligo through the complimentary region of the LoxP site. This brings the double stranded portion to 32 bp. A serial dilution of Tn5 loaded with the LoxP Y-adapters resulted in the expected fragmentation pattern (Figure 2.4 A). While 32 bp is less than the proposed cutoff, it is possible that at least some of the double-stranded portion is serving as a template for Tn5.

Figure 2.4: In vitro validation of BAGEL-seq steps. A. Serial dilution of Tn5 insertion with either full-length LoxP or Y-adapter LoxP. A supercoiled (−) and linearized (+) plasmid are provided for reference. At low concentration, both full-length LoxP and Y-adapter LoxP insertions linearize the plasmid. However, the full-length LoxP does not fragment the plasmid at higher concentrations and even has high-molecular weight species. This suggests that Tn5 insertion is not proceeding properly. B. Cre recombination control. The NEB control plasmid with two LoxP sites behaves as expected upon incubation with Cre (+). Similarly, a plasmid containing one Lox71 and two Lox67 sites recombines upon Cre addition regardless if it is supercoiled or linearized.
Cre Recombination

The Cre recombination step also requires detailed consideration. First, we must bias intra-molecular recombination over inter-molecular recombination. Any inter-molecular recombination will scramble the one-to-one linkage between a barcode and its molecule. This is compounded by the fact that as the distance between LoxP sites increases (i.e. longer molecules), inter-molecular recombination becomes more favorable [51]. Fortunately, researchers have developed a rigorous set of mathematical models that predict the amount of dilution required to favor intra-molecular Cre recombination for arbitrarily spaced LoxP sites [49, 52, 53].

Second, Cre recombination is an equilibrium reaction, with excision efficiencies reaching a maximum of 65-70% [54]. In addition, increasing the reaction time or Cre concentration often results in the formation of high molecular weight aggregates [55]. We hypothesize that asymmetric LoxP sites could bias the reaction in a specific direction [56, 57, 58]. Briefly, these asymmetric sites contain mutations in one half of the inverted repeat region that do not affect Cre recognition or recombination. However, recombination results in a double mutant site that Cre cannot recognize. For BAGEL-seq, we used the asymmetric Lox71 on our cloning plasmid and had Tn5 insert its complementary Lox67. A successful intra-molecular recombination between Lox67 and Lox71 results in the double mutant Lox72. This drives the reaction forward and also stops inter-molecular recombination between successfully recombined products.

Next, we had to address issues arising from the Tn5 insertion. For one, the Y-adapters are not valid substrates for Cre. We had to develop a protocol to fill in the single stranded portion of the Y-adapter and the 9-bp gaps left over from the Tn5 insertion. Additionally, we had to confirm that Cre could recombine the asymmetric LoxP sites in vitro. This is complicated by the fact that a single Tn5 insertion event adds two LoxP sites – one on the 5′ and 3′ end of the linearized sequence. To test this, we cloned two separate Lox67 and a single Lox71 site into a pUC19 backbone. Cre was able to recombine both the supercoiled and linearized version of this plasmid in vitro (Figure 2.4 B).

After Cre recombination, we fragmented the products with Nextera to prepare them for Illumina sequencing. This adds a second primer to the 5′ and 3′ ends of every species in the reaction mixture (Figure 2.2). Unfortunately, our sequencing primer will be downstream of the Nextera primer (Figure 2.2 box). To ensure our sequencing primer is at the 5′ end of the resulting PCR product, we
performed 10 rounds of a primer extension with biotinylated primers. We purified the product with streptavidin beads and used that as an input for a standard PCR with our sequencing primer and Nextera primer.

**Proof of Principle**

For a proof of principle, we cloned a random fragment from the *S. cerevisiae* genome into a pUC19 containing our BAGEL-seq cassette. We successfully generated SLRs that were distributed throughout the plasmid (Figure 2.5 A). Interestingly, there were distinct biases in the coverage for all of the SLRs (Figure 2.5 B). This could be due to a combination of the low diversity of our library and the known insertion bias of Tn5 [59, 60]. While we were able to generate SLRs that had read lengths longer than what was possible on this Miseq (Figure 2.5 C), the vast majority were < 300 bp long. This implies that we over-fragmented our library during the NGS preparation step since we sequenced with paired-end 150 reads. Nonetheless, it is promising that we were able to produce up to ∼ 1000 bp reads in this preliminary experiment.

### 2.4 Future Directions

At this point BAGEL-seq is very much a prototype that would require numerous optimizations before it can be applied robustly. First, we need to improve the fragmentation and size selection step to increase the size of the read length distribution. With longer input molecules at the sequencing step, we can more accurately measure any potential bias for shorter recombination products. Next, we need to improve the overall efficiency of the current protocol. The limiting step appears to be Cre recombination step as Tn5 inserts with almost perfect efficiency (Figure 2.4A). Currently our efficiency, as measured by % recombined Lox72 site in the sequencing data, is only 3 – 10%, which is far from the reported ∼ 70% maximal efficiency. The lack of efficiency could come from the mutant Lox66 and Lox71. Alternatively, we may need to include extra sequence context on the end of our Y-adapter for proper Cre recognition. Lastly, we could try to perform the recombination in an emulsion, thereby increasing the local concentration of Cre and isolating individual DNA molecules from each other.
Figure 2.5: BAGEL-seq preliminary results. A. Coverage of the top 5 most abundant barcodes. We see that reads are randomly distributed throughout the plasmid, albeit with some biases for specific positions. B. Overall coverage of the target plasmid combining all of the SLRs. Again, we find significant bias for a few positions on the plasmid. Due to the low diversity of the input DNA, this could be a result of Tn5’s insertion bias. C. Length distribution of SLRs. We find the majority of SLRs are shorter than the maximum 300 bp (red line) possible with paired 150 nt reads. However, this implies that we over-fragmented our input DNA before sequencing.

Another important metric to quantify is the amount of intra- vs inter-molecular recombination. We could quantify this by using BAGEL-seq to sequence a library of plasmids with known sequences. If there is no inter-molecular recombination, all barcodes will only have reads that map to their unique sequences. We can use this metric to quantify inter-molecular recombination rates as a function of input DNA length, library complexity, and concentration. If we cannot minimize the amount of inter-molecular recombination, we will lose the ability to uniquely sequence individual molecules, and our data will resemble those of existing haplotyping methods.

It is also important to note that BAGEL-seq does not require Tn5 transposase and Cre recombinase – any combination of recombinase and transposase will suffice. For example we could use Mu
transposase to insert FLP recombinase sites. Nonetheless, a robust BAGEL-seq protocol that can uniquely sequence individual molecules will have profound ramifications. This is especially poignant in multiplexed assays of long regulatory regions or proteins, which currently require convoluted and time consuming methods to map variants to barcodes.

2.5 Materials and Methods

Tn5 Purification and Loading

All following methods are adapted from Picelli et al. [61]. After purification, we stored unassembled Tn5 at -20°C in a 55% glycerol stock by adding 1.1 volume of 100% glycerol, 0.33 volume of 2x Tn5 dialysis buffer (100 HEPES-KOH at pH 7.2, 0.2 M NaCl, 0.2 mM EDTA, 2 mM DTT, 0.2% Triton X-100, 20% glycerol). To load Tn5, we incubated this glycerol stock at 1:7 ratio with 100μM annealed oligos at room temperature for an hour. A drop-in Nextera replacement, these oligos are: Tn5MERev 5′-[phos]CTGTCTCTTTATACATCT-3′; Tn5ME-A (Illumina FC-121-1030) 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′; and Tn5ME-B (Illumina FC-121-1031) 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3′. We carried out all annealing reactions by adding equal volumes of 200μM single stranded oligos (e.g. Tn5ME-A and Tn5MERev), heating to 98°C for two minutes, then decreasing at 1°C a minute until 25°C.

Tn5 Insertion

We carried out Tn5 insertion reactions as in Picelli et al. [61]. Specifically the reaction conditions were:

Buffer (TAPS/DMF/MgCl2) - 4 uL
Tn5 (pre-loaded) - 1 uL
DNA (<= 50 ng) - x uL
H2O (to 20 uL) - 15 - x

We incubated this mixture at 55°C for 7 minutes, and halted the reaction by adding 5μL of 0.2% SDS. The reaction can be purified by either Zymo CC5 or Agencourt beads. The 5x TAPS-DMF
buffer consists of 50 mM TAPS-NaOH at pH 8.5 [RT], 25 mM MgCl2, and 50% DMF. A serial
dilution of the pre-loaded Tn5 should be performed with a test plasmid of similar size to determine
the optimal concentration to introduce a single insertion. This will result in a linearized species that
can be determined on a gel.

**Gap Fill**

After Tn5 insertion, it is important to fill the gap’s left behind before recombination. We attempted
a number of different protocols with varying effect. Since a more thorough characterization would
be necessary to determine the optimal one, I’ve included them all. All enzymes, unless noted were
purchased from NEB.

**Wang et al. Gap Fill**

This protocol is taken directly from Wang *et al.* [62].

Ampligase Buffer - 2 uL
dNTP (2.5 mM each) - 2 uL
T4 Pol (3 U/uL) - 1 uL
Ampligase (5 U/uL) - 2.5 uL
DNA (<50 ng) - x
H2O (to 20 uL) - 12.5 - x

Incubate at 37°C for 30 mins
Heat inactivate at 75°C for 20 mins

**T4 Polymerase and Taq Ligase**

This protocol is modified from Wang *et al.* by switching out Ampligase for Taq Ligase. Based on the
EpiCentre literature, the amount of Ampligase used in the protocol is equivalent to 187.5 cohesive
end units.

Taq Ligase Buffer - 2 uL
dNTP (2.5 mM each) - 2 uL
T4 Pol (3 U/μL) - 1 μL
Taq Ligase (40 U/μL) - 2.5 μL
DNA (<50 ng) - x
H2O (to 20 μL) - 12.5 - x

Incubate at 37°C for 30 mins
Heat inactivate at 75°C for 20 mins

It should be noted that we could substitute Cut Smart for NEB Buffer 2.1. It could also be useful to add NAD to the reaction. It should be noted that T4 Polymerase has strong 3′ → 5′ exonuclease activity at elevated temperatures. To reduce this, NEB recommends 15 mins at 12°C. Taq ligase will still be active at this temperature, but to a lesser extent.

**Klenow Fragment and Taq Ligase**

To eliminate any 3′ → 5′ exonuclease activity we can use Klenow fragment. Klenow shows optimal activity in Buffer 2, but can be used in all NEB buffers.

Taq Ligase Buffer - 5 μL
dNTPs (2.5 mM each) - 5 μL
Klenow (50 U/μL) - 1 μL
Taq Ligase (40 U/μL) - 2.5 μL
DNA (<50 ng) - x
H2O (to 50 μL) - 35.5 - x

Incubate at 37°C for 30 mins
Heat inactivate at 75°C for 20 mins

**Cre Recombination**

We purchased purified Cre recombinase from NEB. In order to avoid inter-molecular recombination, we must diluted the DNA in our sample. For most sequences < 10kb, nanomolar concentrations of DNA will typically suffice.

Buffer (10x) - 8.6 μL
Cre (NEB 15U/uL) - 1 uL
DNA (< 1 nM) - x
H2O (to 86 uL) - 76.4 - x

Incubate at 37°C for 30 mins, followed by 70°C for 10 mins.

It's important to note that Cre recombination is an equilibrium reaction that proceeds with ~ 20–30% efficiency. Additionally, adding more Cre or incubating for longer periods of time tends to bias the reaction to high molecular weight aggregates [54, 55, 63].

Next-generation Sequencing

After recombination, we use a Nextera reaction to fragment and add sequencing adapters to our library. We then perform a primer extension with a biotinylated primer that anneals to our sequencing primer to enrich for our barcoded species. As the standard Nextera amplification protocol suggests, we perform a 3 minute gap fill at 72°C. After a 30 second denature at 98°C, we then carry out the linear amplification with 10 cycles of 98°C for 10 seconds, 55°C for 30 seconds, 72°C for 60 seconds, and a 3 minute final extension at 72°C. We then used the M-270 Streptavidin beads and protocol from ThermoFisher to purify the biotinylated products. Next, we performed a 10 rounds of PCR on the purified biotinylated using our sequencing primer and one of the Nextera Primers. We quantified the concentration using an Agilent TapeStation and sequenced the resulting product on an Illumina Miseq.
References


Chapter 3

A Systematic Comparison of Error Correction Enzymes by Next-generation Sequencing

3.1 Abstract

Gene synthesis, the process of assembling gene-length fragments from shorter groups of oligonucleotides (oligos), is becoming an increasingly important tool in molecular and synthetic biology. The length, quality, and cost of gene synthesis are limited by errors produced during oligo synthesis and subsequent assembly. Enzymatic error correction methods are cost-effective means to ameliorate errors in gene synthesis. Previous analyses of these methods relied on cloning and Sanger sequencing to evaluate their efficiencies, limiting quantitative assessment and throughput. Here we develop a method to quantify errors in synthetic DNA by next-generation sequencing. We analyzed errors in a model gene assembly and systematically compared six different error correction enzymes across 11 conditions. We find that ErrASE and T7 Endonuclease I are the most effective at decreasing average error rates (up to 5.8-fold relative to the input), whereas MutS is the best for increasing the number of perfect assemblies (up to 25.2-fold). We are able to quantify differential specificities

---

†This chapter has been published as: N. B. Lubock, D. Zhang, G. M. Church, and S. Kosuri. “A systematic comparison of enzymatic error-correction methods using deep sequencing,” *Nucleic Acids Research*, vol. 45, no. 15, pp. 9206-9217, 2017
such as ErrASE preferentially corrects C/G transversions whereas T7 Endonuclease I preferentially corrects A/T transversions. More generally, this experimental and computational pipeline is a fast, scalable, and extensible way to analyze errors in gene assemblies, to profile error correction methods, and to benchmark DNA synthesis methods.

### 3.2 Introduction

Synthetic DNA is a central tool for biological research [1]. Notably, the initial development of nucleic acid synthesis led directly to the cracking of the genetic code [2]. Today, progress in biology is often limited by the difficulty in producing long, high-quality synthetic DNA [3, 4]. This bottleneck is particularly apparent in the assembly of gene-sized fragments of DNA known as gene synthesis [5].

Currently, gene synthesis relies on the assembly of many oligonucleotides (oligos) of ∼40-150 nucleotide (nt) into a single larger piece of DNA of >1,000 base-pairs (bp) [5]. A variety of methods to assemble oligos into gene-sized fragments exist, but ligation- and polymerase-based assembly methods are the most common [6, 7, 8, 9]. Regardless of the method, the quality of the final product is largely dependent on the quality of the oligos used in the assembly.

Oligos are primarily synthesized using phosphoramidite chemistry first developed by Beaucage and Caruthers in the 1980s [10]. Although these oligos are of high enough quality for common applications such as PCR, their error rates make practical gene synthesis challenging. Several groups have managed to synthesize genes from such oligos, but only find about 5-60% perfect products depending on the size and complexity of the template [11, 12, 13, 14]. This problem is further exacerbated when using lower-cost, but often lower quality oligos from array-based synthesis approaches [15, 16, 17, 18, 19, 20].

Consequently, researchers have developed a number of methods to ameliorate oligo error rate post-synthesis. Size selection methods such as HPLC or PAGE can filter truncated sequences, but are labor-intensive and ineffective against small errors such as single-base deletions, insertions, or substitutions [21, 22]. Hybridization-selection techniques can filter large pools of oligos, but are cost-prohibitive as the number of oligos needed effectively doubles [16, 23]. Sequencing-based retrieval methods can physically pick perfect sequences or separate them by barcoded PCR, but are
time-intensive and can require specialized equipment [24, 25, 26]. Enzymatic error correction is a more commonly-used technique that is relatively inexpensive and effective against most errors. This method employs a variety of different enzymes traditionally used for mutation detection to filter out by binding to or cutting at errors [27, 28, 29, 30].

Two particular classes of proteins are most prevalent in error correction: mismatch binding proteins and mismatch cleaving proteins. Generally, these enzymes recognize distortions in the DNA helix that are caused by mishybridized bases on either strand. In gene synthesis, a pool of perfect and imperfect sequences will be melted and re-annealed pairing perfect and imperfect strands to one another. This produces mishybridized bases that can be recognized by these enzymes. Mismatch binding proteins are used to enrich perfect sequences, while mismatch cleaving proteins are used (often in conjunction with exonuclease trimming) to remove imperfect sequences. The most commonly used mismatch binding protein, MutS, recognizes and binds to all single-base mismatches and a variety of small single stranded loops caused by insertions or deletions (indels) with varying affinity [31, 32, 33, 34, 35]. There are a number of different ways to bind and separate error-containing DNA with MutS including: gel-shift assays, MutS-functionalized columns, and MutS-functionalized magnetic beads [11, 20, 36]. Mismatch cleaving enzymes operate by cutting at or near an error and a variety of different mismatch cleaving enzymes are in use [37]. Broadly, these enzymes can correct errors in two different ways. Similar to mismatch binding methods, perfect sequences can be recovered by filtering them from those cut by mismatch cleaving enzymes. Alternatively, the exonuclease activity is used to trim the error-containing region left over by the mismatch cleaving enzymes. The full length sequences are then recovered by performing a PCR assembly with the trimmed sequences.

Previous assessments of different enzymatic error correction methods have relied on Sanger sequencing of finished gene synthesis products to determine their efficiencies [11, 12, 14, 19, 20]. These studies find that, broadly, the dominant mode of errors in gene synthesis products are single-base deletions and mismatches. However, the prohibitive cost of Sanger sequencing hundreds of thousands of bases has limited the effective characterization and comparison of existing methods. Alternatively, one can turn to the mutation detection literature to find biochemical characterizations of enzymes commonly used in error correction [30, 34, 38, 39, 40]. Although these reports provide
more detailed affinity data, they typically rely on electrophoretic methods and are thus similarly limited in sample size.

In order to overcome these limitations, we developed a custom experimental and computational pipeline that leverages Next-generation Sequencing (NGS) to characterize error rates. Here we report the first in-depth characterization via NGS of both the errors arising from the assembly process, as well as the ability of six of the most commonly used error correction enzymes to eliminate these errors across 11 total conditions. With sample sizes three to four orders of magnitude larger than previous reports, we are able to gain detailed insights into the modality of errors as well as each enzyme’s relative ability to correct them. We also used our method to assess the effect of polymerase on assembly quality by comparing a high-fidelity polymerase (Q5) to a low-fidelity one (KAPA2G Robust). We believe that our method can act as a generalizable platform to rapidly and cost-effectively test, characterize, and optimize oligo synthesis parameters or new enzymatic error correction methods.

### 3.3 Results

**Next-generation Sequencing Based Analysis of a Model Gene Assembly**

To assess different enzymatic error correction methods, we first constructed a constant reference sequence that served as the base for downstream analyses. We designed this sequence to have a length of 100 bp (not including two 21 bp priming regions for amplification and sequencing), a balanced nucleotide content (26:23:23:28 A:C:G:T content), good coverage of all nucleotide pairs and most triplets (80%) while limiting homo-polymer repeats greater than two, and a 28 bp region in the center that has good melting temperature and low secondary structure to facilitate overlap-extension assembly of the two primers. We assembled this sequence from two 85 nt oligos by a preliminary round of polymerase chain assembly (PCA). We then diluted the products of that reaction and used PCR to amplify the full-length 142 bp construct (Figure 3.1). We then subject the resulting assembly to multiple rounds of enzymatic error correction and sequence the products at each step.

We expect that errors arising during sequencing will convolute our true signal. In order to limit these errors as much as possible, we developed a stringent data processing pipeline briefly outlined
as follows: First, we cleaned our raw sequencing reads (509,717 per sample on average) by trimming sequencing adapters, removing any reads containing “N” base calls (212 reads on average), and filtering out any reads that aligned to either the PhiX or E. coli genomes with BBDuk (822 reads on average). This ensures that any spurious reads will not contaminate our alignments and lead to false-positive error calls. Next, we merged our paired end reads together with BBMerge, only keeping alignments with perfect correspondence between the forward and reverse reads. Since we sequenced our assembly with fully overlapping reads, each base is effectively sequenced twice. We found that an average of
95.2% of all bases in the merged reads had a Phred33 score (Q) of 41 (~1/12,600 chance of being miscalled), and 99.8% of all bases on average were above Q30 (1/1000 chance of being miscalled). It should also be noted that most bases were probably above Q41 as this is the default maximum Phred score for most read mergers to maintain backwards compatibility with legacy software. The merging step removed an average of 15.8% of input reads, resulting in an average of 426,514 reads per sample at the end of processing.

After pre-processing the reads, we used a Python implementation of the Needleman-Wunsch aligner, uta-align, to align our reads to the perfect reference sequence. We elected to use a Needleman-Wunsch aligner as it is guaranteed to converge on the optimal alignment for a given scoring system [41, 42]. In contrast, typical short read aligners such as BWA and Bowtie2 do not offer such guarantees as they use heuristics to trade accuracy for speed [43, 44]. We find that these heuristics often result in sub-optimal alignments and miscategorization of error sub-types (Figure 3.6, Table 3.2).

**Error-doped Oligos Enable Comparisons**

In order to assess the sensitivity of our assay, we treated our two-oligo assembly with the error correction cocktail ErrASE and measured the resulting error rates (Figure 3.7). Although we were able to measure significant (Mann-Whitney U, $p << 0.001$, Holm-corrected) reductions in the rate of single-base deletions, multiple-base deletions, and single-base insertions, we were not able to find a significant (Mann-Whitney U, NS, Holm-corrected) reduction between the median rate of mismatches. To ensure that we had a measurable change in error rates for mismatches after enzymatic treatment, we assembled our template from oligos that had errors doped into the sequence. Specifically, we ordered each base with 97% of the intended base, and 1% of the other three nucleotides (not including the 21 bp priming region and the last base of the oligo).

We found that the errors were doped uniformly into our assembly (Figure 3.2A), with the majority of errors being mismatches (90.9%), followed by single base deletions (3.1%), multiple base deletions (2.7%), single base insertions (1.9%), and multiple base insertions (1.5%; Figure 3.2B). Unlike the standard oligo assembly (Figure 3.8), we found no significant difference between the median mismatch rate ($3.99 \times 10^{-2}$) at any of the four bases (Mann-Whitney U, NS; Figure 3.2C).
Figure 3.2: Analysis of Model Gene Assembly Error Rates. A. The error rates per base are plotted across each position in our model separated by the four major classes of error types. We do not see strong positional effects for errors across the template. B. We find a majority of errors on the template are mismatches (MM), followed by single (Del.) and multiple base (M. Del.) deletions; Single (Ins.) and multiple base (M. Ins.) insertions occur at even lower frequencies. C. There are no significant differences between the median rate of mismatches at any base (Mann-Whitney U, NS). D. Similarly, there are no significant differences between transitions and transversions (Mann-Whitney U, NS), implying that the errors were doped uniformly into our oligos. Note: Blue line is a LOESS fit; box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.

Similarly, the median rate of individual transitions and transversions were not significantly different from each other (Mann-Whitney U, NS; Figure 3.2D). These data suggest that incorrect bases were doped in to our oligos at an approximately equal rate that exceeded the baseline error rate of KAPA SYBR Fast – the other potential source of mismatches. We note that the median rates of all error
Figure 3.3: **Effectiveness of Enzymatic Error Correction Methods.** Here we compare the error frequency (errors/kb) and number of perfect assemblies for ten different enzymatic error correction methods. We find that MutS is the most effective enzyme at increasing the percentage of perfect assemblies. However, ErrASE is the most effective at decreasing error frequency. Additionally, we see that the efficacy of T7 Endonuclease I is dependent on protocol, and that the addition of a ligase had detrimental effects on sequence quality. **Note:** the x-axis is ordered by decreasing number of perfect assemblies.

types were significantly higher in the error-doped assembly (Table 3.3, Figure 3.9; Mann-Whitney U, \( p << 0.001 \)). Although this is expected for mismatches, we suspect that the higher median error rates for the other error sub-types are a result of the non-standard synthesis required to dope the errors into our oligos.

**Enzymatic Error Correction Improves Assembly Quality**

Having established the error profile of the error-doped assembly, we evaluated 10 different enzymatic error correction methods using six different enzymes on their ability improve the quality of this assembly (Figure 3.3). As expected, consecutive rounds of enzymatic error correction improved both the relative error frequencies and the number of perfect assemblies. ErrASE was the most effective at decreasing the error frequency, with two rounds of treatment dropping the error frequency
from the doped oligo rate of 45.1 to 7.9 errors/kb. The next most effective enzyme at decreasing error frequency was T7 Endonuclease I (9.1 errors/kb). Based on previous reports in the mutation detection literature, we hypothesized that the addition of a ligase with T7 Endonuclease I would improve correction [39]. We find that the addition of T7 ligase actually decreased assembly quality relative to the no ligase control. In agreement with previous studies, we also find that T7 Endonuclease I is highly sensitive to protocol and concentration as exhibited by the wide range of error frequencies [12, 14]. After T7 Endonuclease I, we found MutS to be the third most effective enzyme at 10.9 errors/kb, with T4 Endonuclease VII, Surveyor, and Endonuclease V following.

However, when looking at number of perfect assemblies sequences, MutS was the most effective enzyme treatment. MutS increased the percentage of perfect sequences in the doped oligo from 1.9% to 47.8% (47.6% for 950nM), while ErrASE increased it to 45.6%, and T7 Endonuclease I increased it to 41.7%. In other words, the oligos that are imperfect after the MutS treatment have more errors on average than those after the T7 Endonuclease I and ErrASE treatments.

**Differences in Enzymatic Error Correction**

With an average of 426,514 reads per round of error correction, our method provides sample sizes three to four orders of magnitude higher than any previous study. This enabled us to compare the effectiveness of these enzymes on rarer errors such as insertions that would be inadequately sampled with Sanger sequencing. Using the error-doped template as a reference, we measured the relative change in error rates for each position across all different enzymatic error correction methods (Figure 3.4A).

We see that in general, all enzymes tested were able to correct insertions and deletions. We find that enzyme performance (as measured by error frequency or number of perfect assemblies) is directly related to the ability to correct mismatches. For example the best performing enzymes, ErrASE, T7 Endonuclease I, and MutS, were able to decrease the median mismatch error rate relative to the error-doped input by 6.2-, 5.1-, and 4.2-fold, respectively. In contrast, the worst performing enzyme, Endonuclease V, was unable to decrease the median mismatch error rate relative to the error-doped input.

We next sought to measure differences in affinity for specific errors between enzymes (Figures
Figure 3.4: Relative Decrease of Different Error Types. A. All enzymes were able to correct both single- and multiple-base insertions and deletions. Additionally, we find that the best performing enzymes corrected the highest amount of mismatches. Note: the x-axis is ordered by increasing error frequency. B. We measure significant differences between the median decrease in C/G → G/C mismatches and the bulk median of all other mismatches after two treatments of ErrASE. Similarly, two treatments of T7 Endonuclease I results in a significant difference between the median decrease in A/T → T/A mismatches compared to the bulk median of all other mismatches (both Mann-Whitney U, p << 0.001).

We were unable to measure any significant differences between bases for the median fold reduction of insertions and deletions (Kruskal-Wallis, NS) across all enzymes after two treatments. However, we were able to detect significant differences between the median fold reduction of different mismatches (Kruskal-Wallis, p << 0.001) across all enzymes after two treatments. Based on these data, we searched for specific mismatch correction biases in our best performing enzymes. For example, we found that two rounds of ErrASE or MutS treatment resulted in a significantly different change in the median fold reduction of C/G → G/C mismatches as compared to the bulk median of all other mismatches (15.2- vs 5.4-fold for ErrASE; 5.1- vs 4.1-fold for MutS; Mann-Whitney U, p << 0.001). In contrast, two rounds T7 Endonuclease I did not result in significant changes in the median fold reduction of C/G → G/C mismatches (5.6- vs 5.1-fold; Mann-Whitney U, NS). They did however, significantly change the median fold reduction of A/T → T/A mismatches as compared
to the bulk median of all other mismatches (12.7- vs 4.2-fold; Mann-Whitney U, $p << 0.001$).

Taken together, these data suggest that different enzymatic error correction methods could be used for different applications. For example, GC- or AT-rich constructs would be best corrected by ErrASE and T7 Endonuclease I, respectively. Alternatively, MutS can be used for applications such as protein libraries, where the proportion of perfect sequences are paramount. We also note that the relative rate of correction for transitions and mismatches in general is likely lower than what is measured here due to errors incorporated by the Taq-based KAPA SYBR Fast polymerase during the NGS preparation [45, 46, 47, 48, 49, 50]. For example, the median fold correction of A/T $\rightarrow$ G/C transitions (the most common Taq-based error) was significantly different than that of the bulk median for all other mismatches for ErrASE, MutS, and T7 Endonuclease I (2.6- vs 7.1-fold for ErrASE; 2.8- vs 4.4-fold for MutS; 2.5- vs 6.8-fold for T7 Endonuclease I; Mann-Whitney U, $p << 0.001$).

**Analysis of Two Five-oligo Assemblies**

In order to investigate the effect of polymerase fidelity on assembly quality, as well as the performance of our method on longer constructs, we assembled two 220-bp constructs from five 60 nt oligos with 20 bp overlaps. To facilitate annealing, we designed the overlap regions to have approximately 50% GC content and minimal secondary structure. We used random nucleotide sequences between the overlap regions with the single restriction being no single nucleotide repeats longer than 4. The resulting nucleotide content of the two constructs are relatively balanced (47:50:62:61 – A:C:G:T for construct one, and 52:53:58:57 – A:C:G:T for construct two). We assembled both constructs with either Q5 or KAPA2G Robust polymerases, and sequenced the assemblies in duplicate with an Illumina MiSeq (~242,000 reads per sample on average after the pipeline filtering). Technical replicates show high correspondence (Figure 3.13) and the error profiles were consistent for each polymerase across the two constructs (Figure 3.14).

As expected, constructs assembled with Q5, a high-fidelity polymerase, had lower error frequencies (2.5 vs 9.7 errors/kb) and a larger percentage of perfect constructs (60.5 vs 10.4%) than KAPA2G Robust, a Taq-based polymerase (Figure 3.5A). The majority of this difference is caused by the higher mismatch frequency in the KAPA2G Robust samples (Table 3.1). The frequencies of errors
Figure 3.5: **Effect of Polymerase on Assembly Quality.** We assembled two different 220 bp constructs (C1 and C2) from five 60 nt oligos with 20 bp overlaps with Q5 and Taq polymerase. **A.** We used our method to compare the error frequency (errors per kb) and percent perfect assemblies. We see that the average error frequency for both constructs is significantly higher for Taq than for Q5 (9.7 vs 2.5 errors/kb). We observe similar trends for the average percentage of perfect assemblies (60.5% for Q5 and 10.4% for Taq). **B.** Similar to the two-oligo assembly, we find that the Taq-based KAPA2G Robust polymerase also has a higher rate of transitions than transversions (mean of $5.32 \times 10^{-5}$ vs. $6.40 \times 10^{-6}$ over both constructs; Mann-Whitney U, $p << 0.001$). **C.** We find that the median rate of multiple base deletions per base in the overlap regions decreased $\sim$2-fold relative to non-overlapping regions for both polymerases (Mann-Whitney U, $p << 0.001$). Similarly, the median rate of multiple base deletions per base also significantly decreases in the priming regions for both Taq ($\sim$6-fold) and Q5 ($\sim$13-fold) for both constructs (both Mann-Whitney U, $p << 0.001$). The difference in decrease between the polymerases was not significant.

Other than mismatches are very similar between the two polymerases (Table 3.1). These errors are likely due to oligonucleotide synthesis, as polymerase and sequencing errors are most often mismatches. Using the previously measured error rates of $\sim 2 \times 10^{-4}$ errors/kb/cycle for Q5, we estimate the expected error frequencies of our assemblies to be $\sim$0.01 error/kb after 50 rounds of amplification with Q5 polymerase [48]. Since this value is an order of magnitude lower than our measured mismatch rate (0.21 mismatch/kb), we estimate the upper bound of mismatches in oligonucleotide synthesis to be 0.2 mismatches/kb.

In agreement with our two-oligo assemblies (Figure 3.8), the KAPA2G Robust amplified assemblies
Table 3.1: **Estimated error frequencies for five-oligo gene assemblies.** Here, we averaged the errors/kb for both five-oligo assemblies using Q5 and KAPA2G Robust polymerases and their technical replicates across each error type (errors are standard error of the mean). We see that all error subtypes are similar except for mismatches.

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Q5</th>
<th>KAPA2G Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mismatches</td>
<td>0.2131 ± 0.0019</td>
<td>7.1388 ± 0.0121</td>
</tr>
<tr>
<td>Single Base Deletions</td>
<td>2.0121 ± 0.0062</td>
<td>2.1891 ± 0.008</td>
</tr>
<tr>
<td>Single Base Insertions</td>
<td>0.0747 ± 0.0011</td>
<td>0.0816 ± 0.0014</td>
</tr>
<tr>
<td>Multiple Base Deletions</td>
<td>0.2326 ± 0.002</td>
<td>0.2342 ± 0.0029</td>
</tr>
<tr>
<td>Multiple Base Insertions</td>
<td>0.0014 ± 2e-04</td>
<td>0.0083 ± 4e-04</td>
</tr>
</tbody>
</table>

also had a higher median error rate per base for transitions ($5.32 \times 10^{-5}$) than for transversions ($6.39 \times 10^{-6}$) across both constructs (Mann-Whitney U, $p << 0.001$; Figure 3.5B). These errors agree with previous single-molecule studies of this polymerase, and suggest that KAPA SYBR Fast was indeed incorporating mismatches during our NGS preparation for the two-oligo assembly [46, 48]. We note that the KAPA2G Robust assemblies had a very high mismatch rate at the bases immediately before and after the third and fifth overlaps. We did not observe this issue in assemblies of the same oligonucleotide mixtures assembled by Q5.

Next, we measured the effect of the overlapping regions on the number of multiple base deletions (Figure 3.5C). In congruence with our data from the two-oligo assembly, we found that the median rate of multiple base deletions (for a given position in the assembly) was significantly different in the overlap regions than in the rest of the assembly with an average reduction of $\sim$2-fold for both Q5 and KAPA2G Robust across the constructs (Mann-Whitney U, Holm corrected; $p << 0.001$). We found no significant decrease in the rates of single base deletions in the overlapping regions. Since we added our sequencing primers by annealing to the first and last 15 bp of the constructs, we could also measure the effect of multiple base deletions in the priming region. Again, we found that the rate of multiple base deletions in the priming region was significantly different than both the overlap region and the rest of the assembly, with an average reduction of $\sim$13-fold for Q5 and $\sim$6-fold for KAPA2G Robust (Mann-Whitney U, Holm corrected; $p << 0.001$). The differences in reduction between Q5 and KAPA2G Robust were not significant, likely due to a small sample size (n≈25).
3.4 Discussion

One of the most promising methods to improve the quality of gene synthesis products is enzymatic error correction. Previous characterizations of error correction enzymes were limited by Sanger sequencing, which prohibited deep enough sequencing to adequately sample rare variants. Here we surpass this bottleneck by leveraging next-generation sequencing (NGS) and a custom computational pipeline to analyze errors in a model gene assembly. With sample sizes of three to four orders of magnitude greater than any previous study, we were able to accurately quantify error frequencies and sample rare errors such as insertions. In addition, NGS precludes the need for time consuming cloning steps. This enabled us to rapidly compare six of the most commonly used error correction enzymes in a total of eleven different conditions in a single experiment, and marks the first comprehensive comparison of enzymatic error correction methods via NGS.

We took multiple steps to minimize the number of false error calls resulting from our method. First, we sequenced our assembly with fully overlapping paired-end reads. Since each base is called independently twice and we only merge reads with a perfect match between the forward and reverse reads, it is unlikely that many sequencing errors made it through this filter. We compared the error profile of the Needleman-Wunsch alignment to two commonly used short-read aligners, BBMap and Bowtie2. As BBMap and Bowtie2 use heuristics that trade accuracy for speed, we found that their resulting alignments were sub-optimal and led to higher false error calls relative to the Needleman-Wunsch alignment.

We assessed the sensitivity of our method by comparing the error rates of a two-oligo assembly before and after ErrASE treatment. We could measure significant changes in all errors except for mismatches. We hypothesized that our polymerase had re-incorporated mismatches during the NGS preparation. To ensure that we could measure changes in the amount of mismatches, we re-assembled our model sequence with oligos synthesized with 3% of the incorrect base at every position. We expected that the net change in mismatches in the error-doped template after error correction would be larger than the basal error rate of the polymerase, enabling quantification. Additionally, increasing the error rate gives a more realistic number of errors (3-4) per assembly that might occur in a longer gene synthesis.

We then used our method to test the ability of six of the most common error correction enzymes...
in eleven total conditions to improve the quality of the error-doped assembly. As expected, we found that all error correction enzymes were able to decrease the error frequency and increase the number of perfect assemblies. We also found that two consecutive treatments of error correction were more effective than one. We then leveraged the large sample sizes generated by NGS to probe specific differences between different enzyme treatments. These data suggest that ErrASE would be the most effective at correcting GC-rich templates, and T7 Endonuclease I is the most effective at correcting AT-rich templates. Alternatively, MutS would be appropriate for the most common applications requiring a single sequenced-verified perfect assembly. The discrepancy of average error frequency and percentage of perfect sequences highlights the importance of using the metrics that are most appropriate for downstream application. In addition, we find that performance of these enzymatic treatments is sensitive to the protocol used as shown in the MutS and T7 Endonuclease I assays.

To test the effect of the polymerase on assembly quality, we assembled two 220 bp constructs from five oligos with both KAPA2G Robust and Q5 polymerases, and compared their error profiles. As expected, we measured a significantly higher number of mismatches in the KAPA2G Robust assemblies than in the Q5. Since the expected mismatch rate of Q5 is lower than our measured value, we estimated an approximate upper bound on the underlying error frequencies of column-synthesized oligos. This is corroborated by the fact that the frequencies of all error types except for mismatches agreed between the two polymerases. Thus, the most common errors in our assemblies were single base deletions, when controlling for polymerase effects. This agrees with previous studies of enzymatic error correction [11, 14, 19]. Two other studies found mismatches to be the most common error. In the first study, this is likely explained by the fact that they amplified their constructs with Taq-polymerase [12]. The second study assembled their genes from chip-synthesized oligos, which might have different error profiles [20]. Lastly, we found that the overlapping regions of our assembly were effective at decreasing the rate of multiple base deletions, but were ineffective for single base deletions.

Our method in its current iteration has limitations. For one, any polymerase misincorporations will convolute the true mismatch correction rate of a given enzyme. While we show that using a high-fidelity polymerase throughout the assembly and NGS library preparation steps ameliorates this issue, we might still be observing library preparation artifacts. Alternatively, we can incorporate
random barcoding strategies or utilize single molecule sequencing to further eliminate polymerase errors [46, 48, 50]. Second, Illumina sequencing limits our assessments to assemblies < 600 bp. We could extend our methodology to long-read technologies such as PacBio or Oxford Nanopore to assess kilobase-scale gene synthesis products [51]. At these lengths, we would likely have to switch from a Needleman-Wunsch alignment to more optimized versions in order to avoid a significant time penalty [52]. Lastly, our model two-oligo assembly used to analyze enzymatic error correction is not indicative of a typical gene synthesis product as it does not code for a gene, is shorter than standard assemblies (142 bp), is assembled from only two oligos, and has a contrived mismatch error rate.

Overall, our method is a fast and accurate method for looking at errors in arbitrary sequences. We believe that this method will be useful for not only rapidly profiling new enzymatic error correction methods, but for other applications such as assessing the quality of chip-synthesized oligos or developing new gene synthesis methods.

3.5 Materials and Methods

Pre-processing

To ensure that we only analyzed high quality reads, we first ran our sequencing data through a pre-processing pipeline. First, we used BBduk (part of the BBMap suite; version 36.14) to trim any Illumina adapters from our reads [53]. Next, we used BBduk to remove any reads with at least 26 bases that match to the PhiX (NC_001422) or E. coli (U00096.3) genomes. We also removed any read pairs that had an “N” base call in either one of the reads during this step. We then took the filtered reads and merged read pairs with perfectly overlapping regions with BBMerge (also part of the BBMap suite; version 36.14) using the pfilter=1 option.

Alignment and Parsing

After read pre-processing and merging, we use a custom Python script to align our reads to the reference oligo sequence, and parse the resulting alignments to get the positions of all errors. Our Python script uses the uta-align (version 0.1.6) package from the Python Package Index (PyPI) to perform a Needleman-Wunsch exhaustive global alignment of the input reads to the reference
sequence [54]. Our script can also provide functionality for performing any alignment supported by the uta-align library (e.g., Smith-Waterman local alignments), and allows for tunable gap penalties or match scores.

Once the alignment and parsing is complete, our script will output the results in a tidy csv file with the name of the read, the position of the error, the type of error, and the actual error itself [55]. The types of errors are as follows: M - Mismatch, D - single-base Deletion, I - single-base Insertion, P - multiPle-base deletion, and S - multiple-base inSertion. The errors are classified as: (Original Base)(Mutated Base) for mismatches; the reference base(s) that were deleted for deletions; and the base(s) that were inserted for insertions. Both single and multiple-base insertions are mapped to the “right” of the base in the reference sequence. For example, if the reference sequence was “GATTACA” and we inserted a C at position 3, the resulting alignment can be visualized as:

Position: 123-4567
Reference: GAT-TACA
Read: GATCTACA
CSV: Read_1, 3, I, C

Lastly, if there is a single-base deletion or insertion in a region where there is an identical base adjacent to the mapped position of the error, we distribute the fractional count of the total number of identical bases over each position. For example, if our alignment produced a deletion of A at position 2 in the sequence “TAAAG,” our software will note this as a deletion of A at positions 2, 3, and 4, with fractional counts of 1/3 at each of those positions. This compensates for the fact that there are three equally valid alignments in that region.

Error Frequency Calculations and Definitions

To be consistent with previous studies, we calculated the relative error frequency per kb ($f$) as

$$f = \frac{\sum_{i=1}^{n} x_i \frac{1000}{l_i}}{n}$$

(3.1)

where $x_i$ is the number of errors in read $i$, $l_i$ is the length of that read, and $n$ is the total number of
reads [12]. This is distinct from error rates, which are defined as the number of errors detected at a given base, divided by the total number of sequencing reads in the sample. Error rates can be further separated by the specific error sub-type.

**Reagents**

All the oligos were synthesized by Integrated DNA Technologies (IDT). The ErrASE Error Correction Kit was purchased from Novici Biotech and is now available as CorrectASE from ThermoFisher. The Surveyor Mutation Detection Kit was from Transgenomic. T4 Endonuclease VII was from Affymetrix. *Thermus aquaticus* MutS DNA mismatch repair protein was from Excellgen. Endonuclease V, T7 Endonuclease I, and T7 DNA Ligase were all from New England Biolabs.

**Error-enriched oligonucleotide synthesis and template assembly**

The 85-nucleotide (nt) forward and reverse oligos contains 21nt primer sites and 64nt template regions, 63 of which, except for the last base, were doped with 3% errors at each position (Supplementary File 1). This doping is achieved by hand-mixing 1% of every other base into the 97% of the reference base. For example, according to the reference sequence, if a position is supposed to be an A, then 1% of C, T, and G was mixed into 97% A during the initial oligo synthesis by IDT. With 28nt complementary regions, the two oligos were able to anneal and then assembled into a 142-base pair (bp) doubled-stranded template. This template consists of two 21bp primer regions and a 100bp region for error correction and for subsequent next-generation sequencing.

Specifically, to pre-assemble the forward and reverse oligos, 10.4µL nuclease-free water (Ambion), 4µL 5X HF Buffer (New England Biolabs), 0.4µL 25mM dNTP (New England Biolabs), and 0.2µL Phusion High Fidelity Polymerase (New England Biolabs) were added into 5µL 1µM mixed aforementioned forward and reverse oligos. Initially heated at 98C for 30 seconds, the reaction was then cycled 15 times: at 98C for 5 seconds, at 70C for 1 second, ramping down with a speed of 0.5C/second to 50C, at 50C for 30 seconds, and at 72C for 20 seconds. The final extension step was at 72C for 5 minutes. The product after the pre-assembly step was diluted 1:10 in nuclease-free water, 2µL of which, served as template, was added into 35.25µL nuclease-free water, 10µL 5X HF Buffer, 1µL 25mM dNTP, 0.5µL Phusion High Fidelity Polymerase, 1.25µL 10mM mixture of forward (5'
TACACGACGCTCTCCGATCT 3′) and reverse (5′ AGACGTGTCCTCTCCGATCT 3′) PCR amplification primers to make the total volume of this PCR 50 µL (Supplementary File 1). Initially heated at 98°C for 30 seconds, the reaction was then cycled 25 times: at 98°C for 5 seconds, at 62°C for 10 seconds, at 72°C for 10 seconds. The final elongation step was at 72°C for 5 minutes. Pooled PCR products were then cleaned using QIAquick PCR Purification Kit (Qiagen), and the purified products served as the template for subsequent error correction treatments and sequencing.

**Error correction of the synthetic DNA template**

**ErrASE**

Per the manufacturer’s instructions, 60 µL of ∼50 ng/µL template in 1X HF Buffer was re-annealed to form heteroduplex by heating at 98°C for 1 minute, cooling at 0°C for 5 minutes, and incubating at 37°C for 5 minutes. Next, 10 µL of this re-annealed heteroduplex was added into each well of the 6-well ErrASE tube and was incubated at room temperature for 1 hour. We then combined 2 µL from each well as template into the recovery PCR, whose setup and thermocycling conditions were the same as the assembly PCR in the section above. The PCR product using the treated heteroduplex from the first well of the ErrASE tube (presumably has the highest concentration of ErrASE) presented a band, indicating successful recovery after error correction. This product was thus cleaned-up using QIAquick PCR Purification Kit and served as the template for the second iteration of ErrASE treatment.

**Surveyor**

Per the manufacturer’s instructions, ∼50 ng/µL template in 1X HF Buffer was re-annealed to form heteroduplex by the following thermocycling conditions. First, the sample was heated at 95°C for 10 minutes. Then, the temperature was ramped down at 2°C/second, and was held at 85°C for 1 minute. Finally, the temperature was further cooled down to 25°C at 0.3°C/second, and was held for 1 minute at every 10°C interval. Per Saaem et al., 2 µL Surveyor Nuclease S and 1 µL Enhancer S were added into 8 µL re-annealed heteroduplex [19]. The reaction mixture was then incubated at 42°C for 60 minutes. After the treatment was concluded, 2 µL of the mixture served as the template in the recovery PCR, whose setup and thermocycling conditions were the same as the assembly PCR.
The product of this recovery PCR, once cleaned-up, entered the next round of Surveyor Nuclease treatment.

**Endonuclease V**

Similar to Fuhrmann *et al.*, 10µL of ~50ng/µL template in 1X HF Buffer was re-annealed using the cycling condition described in the ErrASE section [12]. We then added 5U of Endonuclease V, 2µL of NEBuffer 4, and nuclease-free water to the re-annealed heteroduplex to make the total volume 20µL. The reaction was incubated at 37°C for 24h, and 2µL of this mixture served as the template for the recovery PCR. The cleaned-up product then entered the next iteration of Endonuclease V treatment.

**T7 Endonuclease I (Fuhrmann)**

As in Fuhrmann *et al.*, 10µL of ~50ng/µL template in 1X HF Buffer was re-annealed using the cycling condition described in the ErrASE section [12]. We combined 2µL of NEBuffer 2, 25U of T7 Endonuclease I, and nuclease-free water to make the final volume 20µL. The reaction was incubated at 37°C for 24 hours, and 2µL of the mixture served as the template for the recovery PCR. The cleaned-up product entered the next iteration of T7 Endonuclease I treatment.

**T7 Endonuclease I with T7 DNA Ligase**

We first re-annealed 100ng of template in 1X HF Buffer according to the ErrASE protocol. Then we combined 2.5µL of T4 DNA Ligase reaction buffer, 10U of T7 Endonuclease I, T7 DNA Ligase (at 0, 1000U, or 10000U), and the appropriate amount of nuclease-free water to make the final volume 25µL. The reaction was then incubated at 25°C for 4 hours, and 2µL of the treated sample served as the template for recovery PCR. We used 100ng of the cleaned-up product for the next iteration of T7 Endonuclease I/T7 DNA Ligase treatment.

**T4 Endonuclease VII**

First, 10µL of ~50ng/µL template in 1X HF Buffer was re-annealed using the cycling condition described in the ErrASE section. Then, 1µL 1M Tris-HCl (pH 8.0), 4µL 50mM MgCl₂, 2µL 100mM
β-mercaptoethanol, 1μL 10mg/ml BSA, and 2μL T4 Endo VII (1000U) was added to the 10μL heteroduplex. The reaction mixture was incubated at 37C for 24 hours, and 2μL of which served as the template for the recovery PCR. Then the cleaned-up PCR product entered the next cycle of T4 Endonuclease VII.

**MutS**

Per the manufacturer’s instructions, 250ng/μL in 10mM Tris-HCl (pH=7.8) and 50mM MgCl₂ was heated to 95C for 5 minutes followed by cooling at 0.1C/second to 25C. To the re-annealed template, 207.39μL 1X binding buffer (20mM Tris-HCl (pH=7.8), 10mM NaCl, 5mM MgCl₂, 1mM Dithiothreitol and 5% glycerol) was added, making the concentration of DNA template to ~11.5ng/μL. This mixture was then aliquoted into two tubes with 109μL in each. Appropriate amount of MutS was added into each of the tubes so that the final MutS concentration was 950nM and 1900nM, respectively. The mixtures were then incubated at room temperature for 20 minutes. Equal volumes of Amylose Resin (New England Biolabs), washed and pre-equilibrated with 1X binding buffer, were added into the tubes. The mixtures were incubated at room temperature for 30 minutes, before being spun down. We purified the supernatants with a Qiagen MinElute kit, and eluted the product in 10μL EB. We used 2μL of the 1:100 diluted elution as the templates for the recovery PCR. Lastly, we pooled the PCR products, cleaned them up, and used them for the next iteration of MutS treatments.

**Next-Generation Sequencing using Illumina MiSeq**

Each of the control and enzymatically treated samples was prepared as an individual sequencing library. In summary, the sequencing libraries were prepared using two rounds of qPCR, with the first round appending the Illumina P5 sequence and the second appending the P7 sequence as well as the indices. We also note that the KAPA SYBR FAST kit is a Taq-based polymerase. Specifically, the first round of PCR was set up by mixing 25μL KAPA SYBR FAST Universal 2X qPCR Master Mix (KAPA Biosystems), 1μL 10μM Multiplexing PCR Primer 1.0, 1μL 10μM Multiplexing PCR Primer 2.0, 1μL ~100pg/μL error correction DNA template, and 22μL nuclease-free water. Per the manufacturer’s instructions, the 2-step thermocycling protocol was used for the qPCR reactions.
Once the signals reached the plateaus, the reactions were stopped and cleaned-up using Agencourt AMPure beads, according to the manufacturer’s instructions. The final elution volume was 30µL. To set up the second round of PCR, 25µL KAPA SYBR FAST Universal 2X qPCR Master Mix, 1µL 10µM Multiplexing PCR Primer 1.0, 1µL 10µM PCR Primer each with a distinct index, 1µL ~100pg/µL template from the first round PCR, and 22µL nuclease-free water. The thermocycling and cleaned-up procedures remained the same as those in the first round of PCR. Then, the individually prepared sequencing libraries were quantified using the Library Quantification Kit-Illumina (KAPA Biosystems), according to the provided protocol. Barcoded libraries were subsequently mixed to ~10nM concentration, and the mixed libraries were quantified again before being loaded onto an Illumina MiSeq with a V2 300 cycle kit.

Five-oligo Assembly with High- and Low-fidelity Polymerases

We designed two 220-bp constructs that can be assembled from five 60-nucleotide (nt) oligos each (Supplementary File 1). Each overlap region between adjacent oligos is 20-bp in length, and the first and last oligo contain 15-bp forward and reverse priming regions used for assembly. All overlap and priming sequences were taken from the set designed in Eroshenko et. al to minimize cross-hybridization and maximize $T_m$ similarity [56]. Each set of five oligos was synthesized by Integrated DNA Technologies (IDT) with no modifications and pooled into two 1µM five-oligo mixes.

To pre-assemble the five-oligo construct, 5µL of each 1µM five-oligo mix was added to 10µL of NEBNext Q5 HotStart HiFi PCR Master Mix or KAPA2G Robust HotStart ReadyMix and 5µL nuclease-free water. Initially heated at 98C for 30 seconds, the reaction was then cycled 15 times: at 98C for 5 seconds, at 70C for 1 second, ramping down with a speed of 0.5C/second to 50C, at 50C for 30 seconds, and at 72C for 20 seconds. The final extension step was at 72C for 5 minutes. The product after the pre-assembly step was diluted 1:10 in nuclease-free water, 2µL of which, served as template, was added into 20.5µL nuclease-free water, 25µL of Q5 or KAPA2G Robust master mixes, and 1.25µL 10mM mixture of forward and reverse amplification primers flanking the outer oligos of each construct. Initially heated at 98C for 30 seconds, the reaction was then cycled 20 times: at 98C for 5 seconds, at 62C for 10 seconds, at 72C for 10 seconds. The final elongation step was at 72C for 5 minutes. Pooled PCR products were then purified using a DNA Clean and Concentrator-5 (Zymo).
We prepared each assembly as an individual sequencing library with two technical replicates. The sequencing libraries were prepared using a single round of PCR, which appended both the Illumina P5 and P7 sequences as well as the indices. Specifically, 0.01ng of template was added to 20.5µL nuclease-free water, 25µL Q5 or KAPA2G Robust (depending upon initial condition), and 1.25µL 1µM forward and reverse sequencing primer with corresponding distinct indices. Each library was amplified for a small number of cycles (~12-14) empirically determined using KAPA SYBR FAST Universal 2X qPCR Master Mix (KAPA Biosystems). We estimate the total number of amplification cycles to be < 50 (< 15 for pre-amplification, 20 for amplification, and 12-14 for NGS prep). Individually prepared sequencing libraries were quantified using an Agilent TapeStation 2200. Barcoded libraries were subsequently pooled and mixed to 20nM concentration, and prepared for sequencing on a 500-cycle V2 MiSeq (Illumina).

3.6 Supplementary Information

Analysis of a Two Oligo Assembly

We applied our pipeline to quantify the different types of errors found in our two-oligo assembly of standard (not error doped) oligos (Figure 3.8). We find that on average about one-third of assemblies contain errors, with an overall error frequency of approximately 4.3 errors per kb. We find that mismatches account for the majority of errors (~75%), followed by single (~14%) and multiple-base deletions (~8%) (Figure 3.8A). The mismatches segregate into two significantly different populations, with the median error rate per base being higher at A’s (4.33×10^{-3}) and T’s (4.25×10^{-3}) than at G’s (1.68×10^{-3}) and C’s (1.91×10^{-3}) (Figures SB, C; Mann-Whitney U, p << 0.001, Holm-corrected). Furthermore, we find that the median rate of transitions was significantly higher than that of transversions for each base (Figure 3.2C; Mann-Whitney U, p << 0.001, Holm-corrected). All of these observations indicate that much of the mismatch error rate is due to polymerase misincorporation during the amplification steps for assembly and sample-preparation for sequencing. Specifically, we used the Taq-based KAPA SYBR Fast polymerase during next-generation sequencing library preparation steps. Consistent with our observations, misincorporations caused by Taq occur most often at A’s and T’s, and are preferentially A/T → G/C transitions (49–52). However, we
cannot completely rule out the effect of errors in the oligo synthesis as our error frequency of 4.3 errors per kb is higher than the 3 error/kb expected from 50 rounds of amplification at previously reported Taq error rates (52–54).

Next, we quantified the rates of single- and multiple-base deletions. We find that the median single-base deletion rate per position (5.64×10^{-4}), and that this rate did not vary significantly over the positions (Mann-Whitney U, NS). We also find that multiple-base deletions occur at a similar rate as single-base deletions (3.35×10^{-4}), and measure positional effects for where they occur. Some of this dependence can be explained by the fact that the positions of multiple-base deletions are mapped to the left-most deleted base. Thus, we expect the total number of multiple-base deletions to be highest at position one and decreasing after, since there are the most possible combinations of multiple-base deletions at that position. In addition, we measure a significant decrease in the median multiple-base deletion rate in the annealing region (positions 36-64) of our assembly (Mann-Whitney U, p << 0.001). Large deletions in this region would disrupt the hybridization of the initial assembly, leading to sequence drop-outs and a decrease in the measured number of deletions. We also expect the multiple deletion rate to drop towards the end of the sequence due to a “TATATAT” motif at positions 92-98. Any “TA” deletion (or other substring contained multiple times in the motif) will map to the left-most position, 92.

Finally, we quantified single-base insertions. These errors occur at median rate per position of 9.65×10^{-5}) and exhibit no positional dependence besides an outlier at position 1. An incomplete primer trimming by BBDuk can explain this outlier. Here, 57 of the 152 single-base insertions are a “T,” corresponding to the last base of the primer sequence directly upstream of our first base. Without these 57 bases, the rate of single-base insertions falls closer to the expected median value. Our method is also able to detect multiple-base insertions, which occur at a median rate of 6.16×10^{-6}).

**Availability**

The computational pipeline described above is open source, free to use under the MIT license, and available at [https://github.com/kosurilab/errorCorrect](https://github.com/kosurilab/errorCorrect). For the final analysis and figure production, we used R (version 3.3.*) and [ggplot2](https://ggplot2.tidyverse.org) [57, 58].
Accession Numbers

Sequencing data are available from the sequencing read archive (SRA) with the accession number SRP110084.

Funding

This work was funded by the funds from the US Department of Energy [DE-FC02-02ER63421 to S.K.], National Institutes of Health New Innovator Award [DP2GM114829 to S.K.], Searle Scholars Program [to S.K.], Office of Naval Research [N000141010144 to S.K. and G.M.C.] and a Ruth L. Kirschstein National Research Service Award [GM007185 to N.L.].

Conflict of interest statement.

None declared.

Acknowledgments

The authors would like to acknowledge members of the Kosuri Lab for comments on the manuscript, especially Rocky Cheung and Calin Plesa.
Figure 3.6: Effect of read aligner on error rates. Here we mapped reads from the standard IDT oligo with BBMap (red), Bowtie2 (green), and our Needleman-Wunsch aligner (blue), and quantified the error rates with our pipeline. We see that the choice of aligner affects the resulting error rates, especially for detecting multiple-base deletions.
Figure 3.7: Distributions of error rates per position for the standard oligo assembly before and after ErrASE treatment. We were unable to detect a significant change between the median error rate after two treatments for mismatches. Note: black bar is median value.
Figure 3.8: **In-depth analysis of standard assemblies.** A) The error rates per base are plotted across each position in our model separated by the four major classes of error types. We do not see strong positional effects for errors across the template. B) We find a majority of errors on the template are mismatches (MM), followed by single (Del.) and multiple base (M. Del.) deletions; Single (Ins.) and multiple base (M. Ins.) insertions occur at even lower frequencies. (C) We measure a significantly higher mismatch rate at A’s ($4.33 \times 10^{-3}$) and T’s ($4.25 \times 10^{-3}$) than at G’s ($1.68 \times 10^{-3}$) and C’s ($1.91 \times 10^{-3}$) (Mann-Whitney U, $p << 0.001$). (D) We measure a significantly higher number of transitions (purple) than transversions (green) at each base (Mann-Whitney U, $p << 0.001$). The higher error rates at A’s and T’s is consistent with Taq polymerase errors. Note: Blue line is a LOESS fit; box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers. **Note:** here we performed the same analysis as Figure 2 in the main text with the error-doped assembly.
Figure 3.9: Comparison of measured error rates from error-doped and standard oligos. Here we plot the distribution of error rates per position and see that for every error sub-type the error rates are significantly higher for the error-doped oligos than those produced by the standard process (Mann-Whitney U Test, all $p << 0.001$). Note: Black bar is the median value.
Figure 3.10: Mismatch correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments. Error rates are plotted as the log₂-fold-change in error rate relative to the error-doped template. Note: box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.
Figure 3.11: Single-base deletion correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments. Error rates are plotted as the log<sub>2</sub>-fold-change in error rate relative to the error-doped template. Note: box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.
Figure 3.12: Single-base insertion correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments. Error rates are plotted as the log$_2$-fold-change in error rate relative to the error-doped template. Note: box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.
Figure 3.13: Correlations between error rates for five-oligo assembly technical replicates. We see that technical replicates are almost perfectly correlated (all $r > 0.995$), with the black line being $y = x$. 
Figure 3.14: Positional error rate distributions two assemblies using KAPA2G Robust and Q5 polymerase. We see that KAPA2G Robust, a Taq-based low-fidelity polymerase, incorporates Mismatches (MM) at nearly two-orders of magnitude higher than Q5, a high-fidelity polymerase. We find that both polymerases incorporate single base deletions (Del.), multiple base deletions (M. Del.), single base insertions (Ins.), and multiple base insertions (M. Ins.) at nearly identical rates. With the exception of multiple base insertions, these trends are robust to the different sequence contexts of the two constructs. We note that KAPA2G Robust incorporates a higher number of multiple base insertions around three tandem GGA repeats, likely due to polymerase slippage.
Table 3.2: **Examples of where various aligners fail.** Here _ are padding for visualization, * are soft-trimming, and lower-case bases are inserts.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Ideal</th>
<th>Needleman-Wunsch</th>
<th>Bowtie2</th>
<th>BBMap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read:</td>
<td>__GCTGCCGATTT...</td>
<td>__GCTGCCGATTT...</td>
<td>__GCTGCCGATTT...</td>
<td>**GCTGCCGATTT...</td>
</tr>
<tr>
<td>Reference:</td>
<td>aGCTGCCGATTT...</td>
<td>aGCTGCCGATTT...</td>
<td>aGCTGCCGATTT...</td>
<td>**GCTGCCGATTT...</td>
</tr>
<tr>
<td>Read:</td>
<td>aaGCTGCCGATTT...</td>
<td>aaGCTGCCGATTT...</td>
<td>aaGCTGCCGATTT...</td>
<td>**GCTGCCGATTT...</td>
</tr>
<tr>
<td>Reference:</td>
<td>GCTGCCGATTT...</td>
<td>GCTGCCGATTT...</td>
<td>GCTGCCGATTT...</td>
<td>GCTGCCGATTT...</td>
</tr>
<tr>
<td>Read:</td>
<td>GCT___GATTT...</td>
<td>GCT___GATTT...</td>
<td>GCT___GATTT...</td>
<td>GCT___GATTT...</td>
</tr>
<tr>
<td>Reference:</td>
<td>GCTGCCGATTT...</td>
<td>GCTGCCGATTT...</td>
<td>GCTGCCGATTT...</td>
<td>GCTGCCGATTT...</td>
</tr>
<tr>
<td>Read:</td>
<td>GCTG_____TT...</td>
<td>GCTG_____TT...</td>
<td>GCTG_____TT...</td>
<td>GCTG_____TT...</td>
</tr>
<tr>
<td>Reference:</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
</tr>
<tr>
<td>Read:</td>
<td>...TGTATATATCGa</td>
<td>...TGTATATATCGa</td>
<td>...TGTATATATCGa</td>
<td>...TGTATATATCGa</td>
</tr>
<tr>
<td>Reference:</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
</tr>
<tr>
<td>Read:</td>
<td>...TGTATATATCatG</td>
<td>...TGTATATATCatG</td>
<td>...TGTATATATCatG</td>
<td>...TGTATATATCatG</td>
</tr>
<tr>
<td>Reference:</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
<td>...TGTATATATCG_</td>
</tr>
<tr>
<td>Read:</td>
<td>...TGTATATATgtCG</td>
<td>...TGTATATATgtCG</td>
<td>...TGTATATATgtCG</td>
<td>...TGTATATATgtCG</td>
</tr>
<tr>
<td>Reference:</td>
<td>...TGTATATATCG</td>
<td>...TGTATATATCG</td>
<td>...TGTATATATCG</td>
<td>...TGTATATATCG</td>
</tr>
<tr>
<td>Read:</td>
<td>...TGTATATA__G</td>
<td>...TGTATATA__G</td>
<td>...TGTATATA__G</td>
<td>...TGTATATA__G</td>
</tr>
</tbody>
</table>
Table 3.3: **Median error rate per position for assemblies using the error-doped oligos or the standard oligos.** We measure significant (Mann-Whitney U, \( p << 0.001 \)) differences between the median error rates of the error-doped and standard oligos for all error sub-types.

<table>
<thead>
<tr>
<th>Type</th>
<th>Error-Doped Oligo</th>
<th>Standard Oligo</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Errors</td>
<td>(4.38 \times 10^{-2})</td>
<td>(4.18 \times 10^{-3})</td>
</tr>
<tr>
<td>Mismatches</td>
<td>(3.99 \times 10^{-2})</td>
<td>(3.08 \times 10^{-3})</td>
</tr>
<tr>
<td>Single Base Deletions</td>
<td>(1.28 \times 10^{-3})</td>
<td>(5.64 \times 10^{-4})</td>
</tr>
<tr>
<td>Multiple Base Deletions</td>
<td>(1.17 \times 10^{-3})</td>
<td>(3.35 \times 10^{-4})</td>
</tr>
<tr>
<td>Single Base Insertions</td>
<td>(7.64 \times 10^{-4})</td>
<td>(9.65 \times 10^{-5})</td>
</tr>
<tr>
<td>Multiple Base Insertions</td>
<td>(6.16 \times 10^{-4})</td>
<td>(6.16 \times 10^{-6})</td>
</tr>
</tbody>
</table>

References


Chapter 4

Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes

4.1 Abstract

Improving our ability to construct and functionally characterize DNA sequences would broadly accelerate progress in biology. Here, we introduce DropSynth, a scalable, low-cost method to build thousands of defined gene-length constructs in a pooled (multiplexed) manner. DropSynth uses a library of barcoded beads that pull down the oligonucleotides necessary for a gene’s assembly, which are then processed and assembled in water-in-oil emulsions. We use DropSynth to successfully build >7000 synthetic genes that encode phylogenetically-diverse homologs of two essential genes in *E. coli*. We tested the ability of phosphopantetheine adenyllyltransferase homologs to complement a knockout *E. coli* strain in multiplex, revealing core functional motifs and reasons underlying homolog incompatibility. DropSynth coupled with multiplexed functional assays allow us to rationally explore sequence-function relationships at unprecedented scale.

†This chapter has been published as: C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, and S. Kosuri “Multiplexed gene synthesis in emulsions for exploring protein functional landscapes,” *Science*, vol. 359, no. 6373, pp. 343-347, 2018
4.2 Main Text

The scale at which we can build and functionally characterize DNA sequences sets the pace at which we explore and engineer biology. The recent development of multiplexed functional assays allows for the facile testing of thousands to millions of sequences across a wide array of biological functions [1, 2]. Currently, such assays are limited by their ability to build or access DNA sequences to test. Natural or mutagenized DNA sequences [3, 4] allow for large libraries, but are not easily programmed and thus limit hypotheses, applications, and engineered designs. Alternatively, researchers can use low-cost microarray-based oligo pools that allow for large libraries of designed ∼200 nucleotide (nt) sequences [5], but their short lengths limit many other applications. Gene synthesis is capable of creating long-length sequences, but high costs currently prohibit building large libraries of designed sequences [6, 7, 8, 9].

Here we develop a gene synthesis method we term DropSynth, a multiplexed approach capable of building large pooled libraries of designed gene-length sequences. DropSynth uses microarray derived oligo libraries to assemble gene libraries at vastly reduced costs. We and others have developed robust parallel processes to build genes from oligo arrays, but because each gene must be assembled individually, costs are prohibitive for large gene libraries [6, 10]. In these efforts, the ability to isolate and concentrate DNA from the background pool complexity was paramount for robust assemblies [11]. Previous efforts to multiplex such assemblies have not isolated reactions from one another, and thus suffered from short assembly lengths, highly-biased libraries, the inability to scale, and constraints on sequence homology [12, 13, 14, 15].

DropSynth works by pulling down only those oligos required for a particular gene’s assembly onto barcoded microbeads from a complex oligo pool. By emulsifying this mixture into picoliter droplets, we isolate and concentrate the oligos prior to gene assembly, thereby overcoming the critical roadblocks for proper assembly and scalability (Fig. 4.1A, Supplemental Movie S1). The microbead barcodes are unique 12 nt sequences that all oligos for a particular assembly share, and pair with complementary strands displayed on the microbead. Within each droplet, sequences are released from the bead using Type II restriction enzyme sites and assembled through polymerase cycling assembly (PCA) into full length genes. Finally, the emulsion is broken and the gene library is recovered. To test and optimize the protocol, we built model assemblies that were unique, but
Figure 4.1: **DropSynth assembly and optimization.** A. We amplified array-derived oligos and exposed a single-stranded region that acts as a gene-specific microbead barcode. Barcoded beads display complementary single-stranded regions that selectively pull down the oligos necessary to assemble each gene. The beads are then emulsified, and the oligos are assembled by PCA. The emulsion is then broken, and the resultant assembled genes are barcoded and cloned. B. We used a model gene library that allowed us to monitor the level of specificity and coverage of the assembly process. We then optimized various aspects of the protocol including purification steps, DNA ligase, and bead couplings to improve the specificity of the assembly reaction. Enrichment is defined as the number of specific assemblies observed relative to what would be observed by random chance in a full combinatorial assembly. C. We attempted 96-plex gene assemblies with 3, 4, 5, or 6 oligonucleotides and the resultant libraries displayed the correct-sized band on an agarose gel. D. The distribution of read-counts for all 96 assemblies (4-oligo assembly) as determined by NGS.
shared common overlap sequences. As a result, any contaminating oligo would still participate in the assembly reaction, allowing us to monitor assembly specificity and library coverage. We optimized each aspect of the protocol by trying to assemble 24-, 96-, and 288-member libraries composed of 3, 4, 5, and 6 oligos at once, based on how often we saw intended targets versus their expected frequency given random (i.e. bulk) assembly (Fig. 4.1B). Over many iterations we achieved high enrichment rates ($\sim 10^8$) by modifying the amount of beads, presence of size selection after assembly, ligase used for capture, and type of bead chemistry, testing both EDC crosslinking of carboxyl beads and streptavidin-coupled beads. We ultimately found that using streptavidin bead chemistry, Taq ligase for bead capture, and size-selection after assembly yielded the highest enrichment rates. Using these protocols, we were able to build libraries of up to 6 oligos that produced correct sized bands (Fig. 4.1C), and the resulting assembly distributions were not overly skewed (Fig. 4.1D, Fig. 4.5).

To test the scalability of DropSynth, we attempted assembly of 12,672 genes ranging in size from 381 to 669 bp which encode homologs of two bacterial proteins from across the tree of life (Fig. 4.2A, Fig. 4.6). A total of 33 libraries of 384 genes each encoded 5,775 homologs of dihydrofolate reductase (DHFR) with two different codon usages (11,520 DHFR genes), as well as 1,152 homologs of the enzyme phosphopantetheine adenylyltransferase (PPAT) (Fig. 4.7, A and B). DHFR genes were assembled from either four or five 230-mer oligos while PPAT genes were assembled from five 200-mer oligos. We obtained correctly-sized bands for 31/33 assemblies, with one failing due to oligo amplification issues and the other due to low yield on the oligo processing steps, in contrast to attempts using bulk assembly which produced shorter failed by-products (Fig. 4.7C). Three of the libraries (5x 230-mers) were too long to verify using our barcoding approach, but the resulting synthesis showed correct band formation (Fig. 4.8).

We cloned the libraries into an expression plasmid containing a random 20 bp barcode (assembly barcode) and sequenced the remaining 28 libraries consisting of 10,752 designs (Fig. 4.7D and Fig. 4.8, Fig. 4.9). For the PPAT 5x 200-mer assemblies, sequencing revealed that a total of 872 genes (75%) had assemblies corresponding to a perfect amino acid sequence represented by at least one assembly barcode, with a median of 2 reads per assembly barcode and 56 assembly barcodes per homolog (Fig. 4.2B, Fig. 4.10, A and B). This coverage increased when including sequences with deviations from the designed sequences, with 1,002 genes (87%) represented within 5 aa from the
Figure 4.2: DropSynth assembly of 10,752 genes. A. We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of library members with at least one perfect assembly and the median percent perfects determined using constructs with at least 100 barcodes is shown for each library. B. We observe that 872 PPAT homologs (75%) had at least one perfect assembly, and 1,002 homologs (87%) had at least one assembly within a distance of 5 a.a. from design. C. We assembled two codon variants for each designed DHFR homolog, allowing us to achieve higher coverage.
designed sequences (all homologs have some alignments regardless of distance) (Fig. 4.10D). For the DHFR 4x 230-mer assemblies we observed perfect sequences for 65% (6,271) of the designed homologs, and 75% have at least one assembly within 2 aa difference from design. Since there are two codon usages per homolog, when combined over homologs we observe 3,950 (79%) have at least one perfect, and 88% have at least one assembly in distance 2 aa (Fig. 4.2C). We see a strong correlation ($\rho=0.73$ (Pearson), p-value=3.4E-5) between the amount of DNA used to load the DropSynth beads and the resulting library coverage (Fig. 4.11A). We also found 15 microbead barcodes that have more dropouts than would be expected by chance (Fig. 4.11B). For constructs with at least 100 assembly barcodes, we observed a median of 1.9% ($\sigma=2.9\%$) and 3.9% ($\sigma=3.8\%$) perfect protein assemblies (Fig. 4.2A, Fig. 4.10C, Fig. 4.12) for PPAT and DHFR libraries respectively. The nearly double the rate of perfects for DHFR libraries compared to PPAT can be attributed to using longer oligos (230 vs. 200 nt) that only require 4 oligos instead of 5 to assemble the gene (Fig. 4.13A). Increasing the oligo length provides a way to assemble longer genes without significant decreases in the resulting yields (Fig. 4.13B). Furthermore, the distribution of perfect assemblies in the PPAT libraries is not overly skewed (Fig. 4.10D) and most library members have assemblies with high identity to their respective designed homologs (Fig. 4.10F). The resultant error profiles were consistent with Taq-derived mismatch and assembly errors that we have observed previously [16] (Fig. 4.14).

We sought to show how DropSynth-assembled libraries could be easily coupled as inputs into multiplex functional assays by probing how well the PPAT homologs of various evolutionary distance to *E. coli* could rescue a knockout phenotype. PPAT is an essential enzyme, encoded by the gene *coaD*, which catalyzes the 2nd to last step in the biosynthesis of coenzyme A (CoA) [17] (Fig. 4.15) and is an attractive target for the development of novel antibiotics [18]. Assembled PPAT variants on the barcoded expression plasmid were transformed into *E. coli* Δ*coaD* cells and screened for complementation by growing the library in batch culture through three serial 1000-fold dilutions (Fig. 4.3A, Table 4.1), while a rescue plasmid was simultaneously heat cured (Fig. 4.16). Assembly barcode sequencing of the resulting populations provided a reproducible estimate for the fitness of all homologs successfully assembled without error (biological replicates $\rho=0.94$; Pearson, p-value $<2.2\text{E-16}$) (Fig. 4.17A, Fig. 4.18A). Individual barcodes can display considerable noise, so having many assembly barcodes per construct improved confidence (Fig. 4.18, B and C). Negative controls
Figure 4.3: **PPAT complementation assay.** 

**A.** We used DropSynth to assemble a library of 1152 homologs of phosphopantetheine adenylyltransferase (PPAT), an essential enzyme catalyzing the second-to-last step in coenzyme A biosynthesis, and functionally characterized them using a pooled complementation assay. The barcoded library was transformed into *E. coli* Δ*coaD* cells containing a curable rescue plasmid expressing *E. coli* *coaD*. The rescue plasmid was removed allowing the homologs and their mutants to compete with each other in a batch culture. We tracked assembly barcode frequencies over four serial 1000-fold dilutions, and used the frequency changes to assign a fitness score. 

**B.** This phylogenetic tree shows 451 homologs each with at least 5 assembly barcodes, a subset of the full data set, where leaves are colored by fitness. Despite having a median 50% sequence identity, we find that the majority of PPAT homologs are able to complement the function of the native *E. coli* PPAT, with 70% having positive fitness values, while low-fitness homologs are dispersed throughout the tree without much clustering of clades.
and sequences containing indels show strong depletion (Fig. 4.17A, Fig. 4.19A, 4.20), and fitness is reduced with increasing numbers of mutations ($\rho=-0.38$; Spearman, p-value $<2.2E-16$) (Fig. 4.19, B and C). Pooled fitness scores also correlated well with measured growth rates of individually tested controls ($r_s=0.86$, Spearman, p-value $5.9E-12$) (Fig. 4.21). Approximately 14% percent of the homologs show strong depletion (fitness below -2.5) while 70% have a positive fitness value in the pooled assay. Low-fitness homologs are evenly distributed throughout the phylogenetic tree with only minor clustering of clades (Fig. 4.3B, Fig. 4.17B, Fig. 4.22, 4.23A) showing the high modularity of PPAT. There are several reasons homologs could have low fitness including environmental mismatches, improper folding, mismatched metabolic flux, interactions with other cytosolic components, or gene dosage toxicity effects resulting from improperly high expression [19] (Supplementary Text).

Errors during the oligo synthesis or DropSynth assembly give us mutational data across all the homologs, which we can further analyze to better understand function. We selected all 497 homologs that showed some degree of complementation (fitness greater than -1) as well as their 71,061 mapped mutants within distance 5 a.a. and carried out a multiple sequence alignment to find equivalent residue positions. For each amino acid and position, we found the median fitness among all of these homologs and mutants. The resulting data was projected onto the E. coli PPAT sequence (Fig. 4.4A and B), providing data similar to deep mutational scanning approaches [22, 23]. We term this approach broad mutational scanning (BMS). The average BMS fitness for each position shows strong constraints in the catalytic site, at highly conserved sites ($\rho=-0.64$; Pearson, p-value $<2.2E-16$), and at buried residues compared to solvent-accessible ones ($\rho=0.42$; Pearson, p-value $3.9E-8$) (Fig. 4.24, A and B, Supplementary Text). Surprisingly, some residues that are known to interact with either ATP or 4’-phosphopantetheine turn out to be relatively promiscuous when averaged over a large number of homologs. Furthermore, when mapped onto the E. coli structure (Fig. 4.4B), positions known to be involved with allosteric regulation by coenzyme A or dimer formation, show relatively little constraint, highlighting the diversity of distinct approaches employed among different homologs, while maintaining the same core function. We implemented a simple binary classifier to predict the sign of the BMS fitness value based on a number of features, achieving an accuracy of 0.825 (Fig. 4.25).

Additionally, we can search for gain-of-function (GoF) mutations amongst those homologs that
**Figure 4.4: Broad mutational scanning (BMS) analysis.**

**A.** The fitness landscape of 497 complementing PPAT homologs and their 71,061 mutants (within a distance of 5 a.a.) is projected onto the *E. coli* PPAT sequence, with each point in the heatmap showing the average fitness over all sequences containing that amino acid at each aligned position. Mutations are highly constrained at a core group of residues involved in catalytic function. Other positions show relatively little loss of function, when averaged over many homologs, despite known interactions with the substrates. The *E. coli* WT sequence is indicated by green squares, while the average position fitness, fitness of a residue deletion, mean EVmutation evolutionary statistical energy [20], site conservation, relative solvent accessibility, and secondary structure information is shown above.

**B.** The average fitness at each position, with blue and red representing low and high fitness respectively, overlaid on the *E. coli* PPAT (PDB: 1QJC, 1GN8 [21]) structure complexed with 4'-phosphopantetheine and ATP. We observe loss-of-function for mutations occurring at the active site, while other residues involved with allosteric regulation by coenzyme A or dimer interfaces show large promiscuity, highlighting different strategies employed among homologs.

**C.** In addition to complementing homologs, we can also analyze mutants of the 129 low-fitness (< -2.5) homologs, finding 385 gain-of-function (GoF) mutants across 55 homologs. We project this data onto the *E. coli* PPAT sequence and plot the number of GoF mutants at each position shaded by the number of different homologs represented. We find a total of 8 statistically significant positions (residues: 34, 35, 64, 68, 69, 103, 134, 135) corresponding to four regions in the PPAT structure.
did not complement. A total of 385 gain-of-function (GoF) mutants out of 4,658 were found for 55 homologs out of 129 low-fitness homologs (fitness < -2.5). By aligning these mutations to the E. coli sequence, the eight statistically significant residues (34, 35, 64, 68, 69, 103, 134, 135) shown in Fig. 4.4C localize to four small regions in the protein structure (Fig. 4.26, Supplementary Text). We retrieved six GoF mutants of six different homologs from the library, each with fitness determined from only a single assembly barcode, and individually tested their growth rates. Five of the six mutants showed strong growth and one failed to complement (Fig. 4.21B). We also tested two of the corresponding low-fitness homologs, finding increases in the growth rate of 10% and 42% for their GoF mutants (Table 4.2).

Broad mutational scanning using DropSynth is a useful tool to explore protein functional landscapes. By analyzing many highly divergent homologs, individual steric clashes, which might be important to a particular sequence, become averaged across the homologs. More broadly, DropSynth allows for building large designed libraries of gene-length sequences, with no specialized equipment, and estimated total costs below $2 per gene (Table 4.3 & 4.4). We also show that DropSynth can be combined with dial-out PCR [15], which could be expanded for gene synthesis applications where perfect sequences are paramount. The scale, quality, and cost of DropSynth libraries can likely be improved further with investment in algorithm design, better polymerases, and larger barcoded bead libraries.

### 4.3 Materials and Methods

#### Design of PPAT library

PPAT homologs were found by running a PSI-BLAST search with 1 iteration querying the NCBI RefSeq non-redundant protein database using E. coli PPAT (NP_418091.1). The resulting set of 11,062 homologs was further pruned to 10,277 by keeping only those with lengths ranging from 100 to 200 amino acids. T-Coffee (v11) [24] was used to align the sequences and RAxML (v8.2.10) [25] to infer a maximum-likelihood phylogenetic tree (Fig. 4.6). A custom Python script trimmed the tree by determining the distance from the root at which the number of nodes equaled the desired amount of homologs, and subsequently choosing a random descendant leaf for each node at that
distance. This reduced the tree around 1,300 homolog proteins. Each leaf on the pruned tree was then compared to its nearest neighbours to ensure neighbouring sequences differed by at least five amino acids. We then added in homologs from several model organisms and 34 pathogenic organisms. The final library was dispersed among three libraries of 384 homologs, with every other leaf on the tree distributed into a different library, for a total of 1,152 homologs. The final library contained members sourced from 3 Archeal, 9 Eukaryotic, and 1140 Bacterial organisms (of which the top four most represented Bacterial phyla were 414 Firmicutes, 337 Proteobacteria, 64 Actinobacteria and 38 Spirochaetes).

Design of DHFR library

DHFR homologs were found using the DHFR family (IPR012259) in InterPro database [26]. A total of 5,760 homologs were selected, with 4,992 having lengths less than 530 bp (which can be assembled using 4X 230-mer oligos) and the rest greater (requiring 5X 230-mer oligos). Each homolog was encoded with two codon optimizations, creating a total of 30 libraries with 384 variants in each. We modified the oligo design scripts to forbid the presence of any homopolymer repeats greater than 8 nt. We also added random buffer sequence in between the KpnI restriction site and the reverse assembly primer to bring all of the assembled sequences to within 100 bp of each other to facilitate size selection. This extra buffer sequence is removed upon cloning of the assembled sequence into the barcoded vector. A single pool of 47,616 230-mer oligos was synthesized on a microarray by Agilent Technologies.

Microbead barcode design

We took 2,000 20-mer primers whose design was previously described [27] and removed those containing NdeI, XhoI, EcoRI, KpnI, NotI, SpeI, BtsI, or BspQI restriction sites. All possible 12-mer subset primers were generated and screened for self-dimers, GC content between 45% and 55%, and melting temperature between 40°C and 42°C. Barcodes were further filtered to ensure a minimum modified Levenshtein distance of 3 between any selected barcodes [28]. The first 384 12-mer barcodes, were used in subsequent oligo designs, with the complementary barcode sequences used to generate the beads.
Oligo design

Our oligo design protocol is adapted from Eroshenko et al [27] and summarized in Fig. 4.27A. Briefly, protein sequences were assigned a random codon weighted by the frequency in the E. coli genome, in order to generate a nucleotide sequence compatible with the restriction sites required (NdeI, KpnI, BtsI, or BspQI). A KpnI restriction site (GGTACC) was added on the C-terminal end of the coding sequence, which encodes for a glycine and threonine, before the stop codon. The NdeI restriction site (CATATG) on the N-terminal defined the start ATG codon of the ORF. Immediately flanking these restriction sites, 20-mer assembly primer sequences were added, which are used in the emulsion PCA. These sequences were then split into five shorter overlapping fragments [27], with overlaps optimized to be around 20 bp with a melting temperature between 58°C and 62°C. Sequences which failed to split with these parameters had a new weighted random codon assignment generated, until a codon sequence was found which could be split successfully. BtsI sites were subsequently added on either side of the split sequences, which would release the sequences required for assembly from the bead inside the emulsion droplets, allowing the PCA to proceed. A padding sequence consisting of ATGC repeats was added on the 5’ end ahead of the first BtsI site, with the repeat length such that the final sequence length was 142 nt. Subsequently, an 8-nt Nt.BspQI site, the corresponding 12-mer microbead barcode (described above), and another Nt.BspQI site was prepended to the 5’ end of the sequence, with the restriction sites oriented to nick the top strand on the 5’ side of the barcode and the bottom strand on the 3’ side of the barcode. These Nt.BspQI sites facilitate the processing of the barcode region into a single-stranded top-strand overhang. The barcode was common to all five fragments for each gene, such that all fragments required for each gene assembly would be pulled down and localized onto the same beads. Finally a pair of 15-mer amplification primer sequences were added, with each pool of 384 genes (1,920 oligos) having a unique primer pair orthogonal to the other pools. BLAT [29] was used to screen these primers against the oligo sequences, removing those with homologies over 10 bp. After each of these design steps, we screened for the addition of illegal restriction sites, and modified the sequence if any were found. For PPAT three libraries of 384 genes as well as a small test library of 24 genes were ordered as a single pool of 5,880 oligos (200-mer), while for DHFR thirty libraries of 384 genes were ordered as a single pool of 47,616 oligos (230-mer) and synthesized on a microarray by Agilent Technologies. For the PPAT libraries the final assembly
length was 52 bp longer than the gene length due to the addition of restriction and primer sites. The scripts required to generate DropSynth oligos are available at https://github.com/kosurilab.

**DropSynth barcoded beads protocol**

The general strategy for creating the DropSynth barcoded beads is shown in Fig. 4.27B. Three oligos are used for each DropSynth barcoded microbead, with two of the oligos common to all beads. The anchor oligo attaches to the streptavidin bead surface through a double biotin modification on the 5’ end and has sequences necessary to hybridize with the ligation oligo and part of the barcode oligo. The ligation oligo has a biotin modification on the 3’ end and phosphate group on the 5’ which allows it to ligate to the microbead barcode oligo (Table 4.6). A different microbead barcode oligo is synthesized for each barcode with a common sequence on the 3’ end which can hybridize to the anchor oligo and the reverse-complement of the microbead barcode on the 5’ end which can pull down the gene fragments. This approach means only two synthesized oligos (anchor and ligation oligos) contain expensive modifications. Briefly the anchor oligo, ligation oligo, and each barcoded oligo are hybridized, ligated, and phosphorylated with T4 PNK. These are bound to streptavidin coated M270 Dynabeads, washed, and pooled together to form a uniform mixture of all 384 barcoded beads. This protocol can be scaled as necessary given the amount of multiplexing required. The current assembly protocol utilizes 18 uL of the final pooled bead mixture (∼3.25E5 beads/uL) for the capture of processed oligos, with the bead barcoding protocol provided producing enough pooled beads to carry out around 210 assemblies in 384-plex.

**DropSynth protocol**

DropSynth assembles gene-length fragments through the hybridization of oligos to barcoded microbeads and their resulting amplification. Briefly, individual oligo libraries are PCR-amplified using KAPA HiFi and 15-mer amplification primers. Oligo subpools are then bulk-amplified using the reverse amplification primer and a biotinylated forward amplification primer. After amplification, oligos are nicked using the nicking endonuclease Nt.BspQI, exposing a 12-nt ssDNA “barcode” overhang (Fig. 4.28, Table 4.5). The short biotinylated fragment that is cleaved following nicking is then removed by binding it to streptavidin M270 Dynabeads in a hot water bath. After a column
cleanup, each oligo subpool is mixed with the designed DropSynth barcoded beads and Taq ligase, and annealed overnight from 50°C to 10°C. In this process, all oligos required for each gene assembly are captured when each microbead barcode overhang anneals to a corresponding complementary microbead barcode on the bead. Captured beads are then mixed with KAPA2G Robust Mastermix, 20-mer forward and reverse assembly primers, BSA, BtsI, and BioRad Droplet Generation Oil. The mixture is immediately vortexed for 3 minutes, allowing for compartmentalization of captured beads in <5 um droplets (Fig. 4.29), which are subsequently heated allowing temperature-sensitive BtsI to release the sequences required for assembly from the bead. Droplets from each subpool are then loaded into PCR tubes and thermocycled, allowing PCA to proceed. The PCA products are then recovered by breaking the emulsion with chloroform, purified and re-amplified, providing sufficient assembled DNA for downstream applications.

**Optimization of DropSynth**

Significant optimization of the oligo processing and bead capture was required to achieve sufficiently high specificity to allow large multiplexing. Initial attempts to capture fully single-stranded oligos, generated using USER / λ exo / DpnII treatment [10], followed by primer extension of the missing complementary strand, performed poorly for three-oligo assemblies and failed altogether with four-oligo assemblies for all four polymerases tested (Kapa Robust, Kapa HiFi, Pfu Turbo, and Phusion). As an alternative approach, we nicked opposite strands on either side of the BC region with type IIS enzymes, before melting the microbead barcode strands apart and removing the unwanted biotinylated strand, leaving a single-stranded overhang along with the rest of the oligo, as shown in Fig. 4.1A. This eliminated the need for primer extension, and resulted in a 10-fold specificity improvement in tests on 96-plex assemblies of three to six oligos.

We also optimized the type of bead chemistry, testing both covalent carboxyl coupling and streptavidin coupling. Briefly, anchor oligos were covalently attached as follows. 100ul Dynabeads M-270 Carboxylic Acid were washed twice with 25 mM MES (pH 5). Next, 60μg anchor oligo in 25 mM MES (pH 5) was added to the washed Dynabeads and incubated at room temperature for 30 minutes. EDC was dissolved in cold 100 mM MES (pH 5) to a concentration of 100 mg/ml, after which 30μl EDC solution (3 mg) was added to the Dynabead/anchor oligo suspension. Next, 10μl
of 25 mM MES (pH 5) was added and the solution was incubated overnight at 4°C with slow tilt rotation. Finally, the coated Dynabeads were washed 4 times using PBS (0.1% Tween-20). Despite successful assembly from carboxyl-coupled beads, we observed significantly higher enrichment factors in streptavidin-coupled beads. Thus we proceeded using streptavidin-coupled beads in all DropSynth experiments.

We further optimized the amount of beads, ligation reaction, the ligase used in the capture step, nicking reaction, presence/absence of size selection after assembly, and different techniques to purify the emulsion assembly products before re-amplification to achieve an assembly enrichment factor of $10^8$, relative to the probability of a correct assembly by random chance, for a 288-plex five-oligo assembly (Fig. 4.1B).

**PPAT rescue plasmid and coaD knockout**

As PPAT (coaD) is an essential gene, we re-engineered plasmid pTKRED [30] and to constitutively express bicistronic wild-type (WT) coaD gene followed by sfGFP (Fig. 4.16A). The WT coaD gene from *E. coli* MG1655 was amplified with a strong constitutive promoter (TGACGGCTAGCTCAGTGCTAGGTACAGTGCTAGC) and RBS (TACGAGTGAAAGAGGAGAAATACTAG) on the 5’ end, and BamHI site on the 3’ end. This was ligated to a fragment containing a 5’ BamHI site, RiboJ self-splicing element [31], sfGFP [32], and a transcriptional terminator to create coaD_sfGFP. pTKRED was digested with BsaI and the larger fragment (8,391 bp) containing the λ–red genes was gel extracted. The coaD_sfGFP DNA fragment was then ligated into the larger pTKRED BsaI fragment to create pTKcoaD. This ligation was transformed into NEB 5-alpha electrocompetent *E. coli* and colonies were sequence verified. The pTKcoaD plasmid expresses PPAT and GFP constitutively while the λ–red recombinase genes are under IPTG induction. The temperature sensitive origin of replication can be used to heat cure the plasmid at 42°C, which can be confirmed through the loss of GFP fluorescence (Fig. 4.16E).

Knockout of the coaD gene in *E. coli* was carried out using standard techniques [30, 33]. Briefly pTKcoaD was transformed into both *E. coli* DH10B electrocompetent cells (ThermoFisher Scientific). Individual colonies were chosen and made electrocompetent. These were transformed with a recombination template containing a Kanamycin cassette flanked by homology arms to
the regions immediately adjacent to the *coaD* gene. This template was made by first amplifying the Kanamycin cassette from pZS2-123 [34] using primers coaD\_KO\_KAN\_FWD\_1 and coaD\_KO\_KAN\_REV\_1 (Table 4.7). The resulting amplicon was purified and further amplified using the primers coaD\_KO\_KAN\_FWD\_2 and coaD\_KO\_KAN\_REV\_2 (Table 4.7). The knock-in targeted only the PPAT coding region so as to not interfere with the essential *waaA* gene immediately upstream of *coaD*. Knock out strains were verified by Sanger sequencing and colony PCR (Fig. 4.16, C and D). We further verified that heat curing of the rescue plasmid suppressed cell growth and characterized the escape frequency.

**pEVBC expression plasmid**

The barcoded plasmid used to express PPAT homologs is a derivative of high-copy pUC19 with a pLac-UV5 promoter, NdeI and KpnI restriction sites for cloning, an in-frame stop codon, and a 20-mer random assembly barcode. This was made by first double-digesting pUC19 with AatII + BspQI and gel extracting the larger fragment. A gBlock DNA fragment was synthesized containing the promoter, several restriction sites, and an in-frame chloramphenicol acetyltransferase before the stop codon. We initially tried using this in-frame chloramphenicol resistance as a way to screen the library against frame-shifted products, but we found this highly biased the resultant libraries (*data not shown*) and thus we did not use this in-frame selection for the results presented here. This was ligated into the pUC19 AatII-BspQI backbone fragment to create plasmid pEV\_CMR. The plasmid pEV\_CMR was double digested with NcoI + KpnI and the long 2,209 bp fragment was gel extracted. Round-the-horn PCR was carried out using 1 ng of the pEV\_CMR digest as template, a forward primer pEVBC\_FWD with a 5' biotin and a NdeI site, and a reverse primer pEVBC\_REV1 with a 5' biotin (Table 4.7), a 20 N-mer random assembly barcode, and a KpnI site, for 5 cycles. This PCR product was further amplified with outer primers pEVBC\_FWD and pEVBC\_amp\_FWD for 15 cycles (Table 4.7). This amplicon was column purified, digested with NdeI + KpnI, treated with rSAP, cleaned up with Streptavidin coated Dynabeads to remove the small fragments, and column purified again to create the vector pEVBC (Fig. 4.16B).
Barcoded PPAT library in pEVBC

Assembled PPAT homolog genes for each library were digested with NdeI + KpnI and column purified. A ligation was then carried out for each PPAT library using 150 ng of NdeI + KpnI digested pEVBC vector and 100 ng of digested PPAT homolog genes using 3,000U of T7 ligase in a total volume of 30 uL. This reaction was column purified and concentrated to a volume of 16 uL. NEB 5-alpha electrocompetent E. coli cells were then transformed using 3-4 uL of the purified ligation, resulting in over 10 million cfus per transformation. Overnight cultures grown in LB with Carbenicillin were miniprepped, quantified, and an equimolar pool from all three PPAT homolog libraries was created, henceforth referred to as sample S0.

Barcoded DHFR library in pEVBC

Analogous to PPAT, assembled DHFR homolog genes for each library were digested with NdeI + KpnI and column purified. A ligation was then carried out for each library using 150 ng of NdeI + KpnI digested pEVBC vector and 100 ng of digested DHFR homolog genes using 3,000U of T7 ligase in a total volume of 30 uL. This reaction was column purified and concentrated to a volume of 8 uL. In order to overcome known DHFR overexpression issues in E. coli, we directly PCR-amplified ligation products using primers mi3_FWD and mi3_REV_N7## (Table 4.7) to add p5 sequencing adapters and library indexes, rather than transforming and miniprepping.

Assembly barcode mapping

The assembly barcoded PPAT libraries were sequenced on two Illumina Miseq paired end 600-cycle runs, and DHFR libraries were sequenced on three Illumina Miseq paired end 600-cycle runs. Each library was PCR amplified using primers mi3_FWD and mi3_REV_N7## (Table 4.7) to add p5 sequencing adapters and library indexes. The resulting amplicons were size-selected using gel-extraction and quantified using an Agilent 2200 Tapestation. Samples were then pooled and sequenced on a Miseq using custom primers mi3_R1, mi3_R2, and mi3_index (Table 4.7). This resulted in 27,822,356 total reads (for PPAT) after merging the runs together. Barcode read counts for the S0 (unselected) library were generated by extracting the 20 bp sequence corresponding to the barcode region from the Read 2 sequences and using Starcode [35] to collapse barcodes within
a Levenshtein distance of 1 (Fig. 4.10). Sequencing data are available from the sequencing read archive (SRA) under BioProject PRJNA421181.

Briefly, the subsequent data processing was carried out as follows. All Fastq files had adapters trimmed in bbduk followed by paired-end read merging using bbmerge (from the BBTools package version 36.14). All reads were then concatenated and piped into a custom python script which generated a consensus nucleotide sequence for each barcode. The script works as follows. First, we split reads into the 20 nt assembly barcode and the corresponding variant, and generate dictionary that maps every assembly barcode to a list of variants associated to it. To eliminate assembly barcodes that are associated with two different variants, we calculate the pairwise Levenshtein distance of every variant associated with a given assembly barcode. If a certain percentage of these assembly barcodes (5%) are greater than a distance cutoff (10) then we consider the assembly barcode contaminated and drop it from further analysis. Finally, we generate a consensus sequence by taking the majority basecall at every position. Mapped consensus sequences were then translated until the first stop codon and sequences perfectly matched to any designed homologs were annotated.

Analysis of the number of reads per assembly barcode as a function of dilution revealed a small number of assembly barcodes with very high number of reads, as many as 300,000 by the fourth dilution, attributed to the emergence of adaptive mutations conferring a growth advantage at 42°C, which occur stochastically. We also deduce from the lack of GFP positive colonies in the plates at various steps in the dilution that these adaptive mutations did not occur in cells still harboring the rescue plasmid. A total of 18 barcodes from serial dilution replicate A and 16 barcodes from replicate B were removed from further analysis.

Mutant homolog sequences were annotated by first aligning the consensus nucleotide sequence for each barcode against the 1,152 designed PPAT homologs using bbmap. The resulting SAM file was parsed to extract the closest alignment match. A pairwise alignment of the amino acid sequence was carried out for each mapped barcode sequence (until the first stop codon) against its best PPAT homolog alignment match. Mutants within a distance of 5 amino acids from the designed sequence had their individual a.a. mutations annotated for further analysis downstream.

We estimated the number of chimeric assemblies computationally. First, we used a custom python script to divide our merged reads into 5 equally sized chunks. We then used BBMap (v
36.xx) to perform a pseudo-local alignment to a reference fasta containing all of our designed constructs. We refer to these alignments as pseudo-local as BBMap first searches for an optimal global alignment, and clips the reads if they return a higher score. We then tallied the number of chunks successfully aligned, as well as the number of different unique references each chunk aligned to. We then categorized each construct as follows:

- **Perfect** - all 5 chunks align to the same reference
- **Chimeric** - all 5 chunks align, but to more than one reference
- ** Possibly Chimeric** - any number of chunks (not necessarily 5) align to more than one reference
- **Junk** - less than 5 chunks successfully aligned

**BMS analysis**

Briefly, we aligned all complementing homologs using MAFFT and created a lookup table for each residue of each homolog. For perfect homolog sequences we scanned through all residues and placed the homologs fitness into a BMS data table with the corresponding residue and E. coli position based on the alignment. For the mutants up to distance 5 a.a. from the perfect, we took only the mutated residues and added the fitness of the mutant into the BMS data table with the mutated residue and the corresponding E. coli position based on the alignment. For each residue and position in the BMS data table, the BMS fitness was determined as the median value of all of the corresponding data point at that position.

**Classifier**

We implemented a simple classifier to predict how different variants would perform in our assay. First, we categorized each variant into two bins based on whether or not their measured fitness score was greater than 0. We then selected 6 features for our model - the amino acid mutation, secondary structure class as assigned by DSSP (loop, beta-sheet, or alpha-helix), relative solvent accessibility as assigned by DSSP, sequence conservation, evolutionary coupling as predicted by EVMutation, and the frequency of residue substitution from the sequence alignment used for EVMutation’s prediction. We used the R package Caret to perform a simple logistic regression using these features. To assess
the performance of our classifier, we performed 10 repeats of 5-fold cross-validation on our dataset and measured the precision and recall of each model on its respective hold-out set. We then used the R package `precrec` to plot both the receiver operating characteristic (ROC) and precision recall curves [36].

**Complementation assay**

The complementation of synthesized homologs was carried out using a serial batch culture. After ligation into pEVBC, homologs from all three libraries were pooled together to create sample S0. Supercoiled S0 plasmid was then electroporated into electrocompetent *E. coli* DH10β ∆ coaD pTKcoaD. The serial batch cultures, consisting of two biological replicates, were initiated by making 10 transformations using 1 ng of S0 plasmid into 40 uL of cells and recovered at 30°C in 1 mL SOC + 1 mM IPTG for 1 hour. For each replicate, 5 transformations were pooled together and used to seed a fresh culture with between 7 million and 17 million cfus. Cells were grown in 1 L LB media supplemented with Kanamycin + Carbenicillin + 0.05 mM IPTG and grown to saturation at 42°C (8-10 generations) between each bottleneck. Cells were propagated through 3 bottlenecks for a total of 4 samples for each replicate, with 1000x dilutions at each bottleneck. DNA was miniprepped from each sample and cells were plated to ensure proper curing of the rescue plasmid, by screening for GFP+ colonies.

The barcodes from each of the 8 complementation samples were amplified using primers mi4_FWD and mi4_REV_N7## (Table 4.7) to add sequencing adapters and library indexes. The resulting 294 bp amplicon was size-selected using gel-extraction, purified, pooled, and loaded onto a Hiseq 2000 single-end 50 cycle run using custom sequencing primers mi4_R1 and mi4_index (Table 4.7), resulting in 138 million total reads. The barcodes for each sample were clustered using Starcode [35] to collapse barcodes within a Levenshtein distance of 1 (Table 4.1).

**Complementation data analysis**

In order to reduce noise in calculating the fitness change we pruned the barcodes leaving only those with at least 10 reads in S0 or at some point in the serial dilution. This reduced the total number of unique barcodes from 7,038,274 to 627,302. We calculated fitness scores for each mapped sequence.
with at least one barcode. First, the read counts at each dilution were normalized based on the total sequencing depth of the sample relative to S0. The log2 fold change between each dilution and sample S0 was then determined for each barcode using

\[ f_{x0} = \log_2(r_x + 1) - \log_2(r_0 + 1) , \]

where \( r_x \) is the number of normalized reads in the corresponding dilution. We then took the median value (to minimise effects of outliers) of the log2 fold change over all of the dilutions to determine the fitness for that barcode

\[ f_{BC} = \text{median}(f_{10}, f_{20}, f_{30}, f_{40}). \]

The median fitness for each barcode representing a sequence was determined for each replicate (A and B) individually

\[ f_{seqA} = \text{median}(f_{BC1A}, f_{BC2A}, f_{BC3A}, f_{BC4A}, ...), \]
\[ f_{seqB} = \text{median}(f_{BC1B}, f_{BC2B}, f_{BC3B}, f_{BC4B}, ...). \]

We then selected only those sequences represented in both replicates and took the median replicate fitness as the final fitness value

\[ f_{seq} = \text{median}(f_{seqA}, f_{seqB}). \]

Data analysis was carried out in R, with visualisations using ggplot2, ggtree [37], and UCSF Chimera [38]. Residue conservation was determined using Jensen-Shannon divergence [39], secondary structure and relative solvent accessibilities sourced from DSSP analysis [40, 41] of 1H1T [42]. The analysis scripts are available at https://github.com/kosurilab.

**Assembly Retrieval by Dialout Amplification**

The presence of a unique barcode on each assembly allows us to retrieve them from the library using PCR amplification [13, 15]. We attempted to amplify 48 unique homologs and 12 gain-of function mutants. As a positive control we also amplified the wild-type *E. coli* coaD gene from the pTKcoaD rescue plasmid. The designed primers flanked each construct, with reverse primers annealing to each gene-specific barcode. We observed correct size amplification products for 59 of 60, with 18 of these using lower complexity post complementation selection libraries as template, while the rest used the high-complexity sample S0. Individual amplicons were then gel-extracted, restriction digested with KpnI-HF and NdeI, ligated into empty pEVBC backbone, and transformed...
into chemically competent NEB DH5alpha *E. coli* cells. Colonies were verified via colony PCR and Sanger sequencing, and validated colonies were re-inoculated overnight and miniprepped. We successfully sequence-verified 43 of the 59 constructs (37 homologs and 6 gain-of-function mutants), in addition to the WT coaD gene (Table 4.2).

**Growth Rate Analysis of Dialed-out homologs and Gain-of-function Mutants**

Following successful dialout PCR and re-cloning, we transformed 1 ng of each construct in pEVBC into 7 uL of electrocompetent *coaD* knockout cells. We analyzed the presence of growth by counting dilution Carb + Kan plates at both 30°C and 42°C (Table 4.2, Fig. 4.21A). Four constructs had no colonies on the 42°C plates, of which two were low-fitness homologs, one was a gain-of function mutant with only one barcode (false positive), and another construct KOS35328 had good fitness (1.88) in the pooled assay determined using 25 barcodes. The lack of colonies for KOS35328 requires further investigation, and may be a transformation error. Six constructs had low colony counts on both plates, of which five correspond to low-fitness homologs (Fig. 4.21A). We noticed a trend in which homologs with enhanced fitness in the pooled complementation assay gave rise to greater numbers of colonies on the 42°C dilution plates. Furthermore, we also noticed that homologs with enhanced fitness in the pooled complementation assay typically gave rise to 42°C colonies that appeared larger than their corresponding 30°C colonies (Fig. 4.21A). Of the constructs with at least 10 colonies on the 42°C plates, we picked 3 colonies per homolog and re-inoculated them in 1 mL LB + Carb + Kan and grew overnight at 42°C. 2 uL of saturated culture was then diluted in 98 uL of LB + Carb + Kan in wells of a 96-well plate and loaded into a Tecan M1000 Plate Reader for 12 hours at 42°C. OD600 values, taken at 30-minute intervals, were measured at 9 points within each well and averaged. Resultant growth curves were plotted for all colonies and averaged on the construct level. Maximum slopes of each growth curve were calculated and plotted against fitness scores determined from the complementation assay (Fig. 4.21B). A strong correlation (Spearman $r_s = 0.86$, p-value 5.9E-12) was observed comparing homolog growth rate to fitness, validating our assay and analysis pipeline. Examining the residual errors of the fit of growth rate to fitness we observe that constructs with fewer barcodes tend to have larger errors (Fig. 4.21C) which agrees with the reproducibility of the fitness value among replicates as a function of the number of barcodes.
DropSynth bead barcoding protocol

Prepare 2X Binding and Wash buffer (2M NaCl, 1mM EDTA, 10mM Tris)

2X B&W 40mL:

- 4.675g NaCl salt
- 400 uL UltraPure 1M Tris-HCl, pH 7.5 (Invitrogen)
- 80uL UltraPure 0.5M EDTA, pH 8.0 (Invitrogen)
- UltraPure Distilled Water (Invitrogen) to 40 mL

This protocol can be done on a single 384 well plate or 4x 96 well plates, the latter protocol is provided. Reagents required:

- 384 uL 100 uM anchor oligo (Integrated DNA Technologies)
- 384 uL 100 uM ligation oligo (Integrated DNA Technologies)
- 1 uL 100 uM of each barcode oligo (Integrated DNA Technologies)
- 1,576 uL 10X T4 ligase buffer (New England Biolabs)
- 384 uL T4 PNK (10,000 U/mL) (New England Biolabs)
- 40 uL T4 ligase (concentrated 2,000,000 U/mL) (New England Biolabs)
- 1,920 uL stock Dynabeads M270 Streptavidin (Invitrogen)

For each of the four 96-well plates:

1. Mix 96 uL 100 uM anchor oligo and 96 uL 100 uM ligation oligo.

2. Prepare the 96 well plate. In each well add:

   - 2 uL of mixed anchor and ligation oligo
• 1 uL 100 uM barcoded oligo
• 4 uL 10X T4 Ligase buffer
• 33 uL UltraPure Distilled Water
• TOTAL: 40 uL

3. Anneal the mixed oligos on each plate using using the following conditions (30 min total):
   • 3 min at 70°C
   • Ramp down to 60°C for 1 min, 0.1°C/sec
   • Ramp down to 50°C for 1 min, 0.1°C/sec
   • Ramp down to 40°C for 1 min, 0.1°C/sec
   • Ramp down to 30°C for 1 min, 0.1°C/sec
   • Put plate on ice

4. Ligate the barcoded oligo to the ligation oligo:
   • Make a 1:10 T4 Ligase dilution:
     10 uL T4 Ligase (concentrated 2,000,000 U/mL)
     10 uL 10X T4 ligase buffer
     80 uL H₂O
     TOTAL: 100 uL
   • Add 1 uL T4 Ligase (1:10 dilution) to each well
   • Incubate plate at 16°C for 1 hr or longer, followed by 65°C for 20 min to heat inactivate the ligase

5. Phosphorylate the barcoded oligo:
   • Add 1 uL T4 PNK into each well
• Incubate the plate at 37°C for 40 min (or longer), followed by 65°C for 20 min to heat inactivate the PNK

5. Bind to beads:

• Prepare 480 uL stock Dynabeads M270 Streptavidin, washed, and resuspended in 960 uL B&W buffer

• Add 10 uL resuspended beads to each well. (∼3.25E6 beads/well and ∼18.5E6 molecules/bead)

• Mix overnight with shaking (2000 RPM) at room temperature.

7. Pool beads:

• Wash each well with 150 uL B&W buffer 5 times.

• Resuspend in 10 uL B&W buffer

• 1 uL of each well is mixed together, making a 96 uL mixed barcoded bead pool for each plate.

• Mix 96 uL from each plate to make a full 384 uL mixed barcoded bead pool. Store these at 4°C when not in use.

**DropSynth emulsion synthesis protocol**

The following protocol was used to assemble the PPAT library. All PCR steps were performed on a Bio-Rad C1000 Touch Thermal Cycler (Bio-Rad Laboratories).

1. Prepare the OLS pool

• Make 1/5, 1/10, and 1/20 dilutions of the OLS chip pool.

• Prepare mixtures of forward and reverse subpool amplification primers for each subpool, with 10µM final concentration of each primer.

2. Amplify subpools.
• For each subpool run a qPCR to determine the number of cycles required for amplification. Amplifications are stopped several cycles before plateauing to prevent over-amplification of the libraries.

• Amplify each subpool.

  – 1 uL template (test 1/5, 1/10, 1/20 OLS pool dilutions)
  – 1.25 uL subpool specific primer mix (10 uM FWD + 10 uM REV)
  – 22.75 uL UltraPure Distilled Water (Invitrogen)
  – 25 uL Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems)
  – TOTAL: 50 uL
  – PCR protocol:
    1. 3 min 95°C initial denaturation
    2. 45 sec 98°C denaturation
    3. 15 sec 58°C annealing
    4. 15 sec 72°C extension
    5. Go to step 2, repeat based on the number of cycles determined by qPCR.
    6. 1 min 72°C final extension

• Column purify amplified oligos using a Zymo Clean & Concentrator -5 (Zymo Research).

• Run PCR products on gel. Look for higher MW products, indicative of overamplification. Excessive low MW products may indicate chip synthesis issues.

• Size select, using gel extraction, if necessary.

• Create 20 pg/uL dilutions of each amplified subpool. (~91 million/uL)


• Run a second PCR using a biotinylated FWD amplification primer, with sufficient tubes to make 4.5 ug to 9 ug of PCR product.
  – 1 uL of 20 pg/uL subpool dilution
  – 1.5 uL subpool specific primer mix (10 uM biotinylated FWD + 10 uM REV)
- 22.5 uL UltraPure Distilled Water (Invitrogen)
- 25 uL Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems)
- TOTAL: 50 uL

- PCR protocol:
  1. 3 min 95°C initial denaturation
  2. 20 sec 98°C denaturation
  3. 15 sec 58°C annealing
  4. 15 sec 72°C extension
  5. Go to step 2, 18X
  6. 1 min 72°C final extension

- Pool and column purify PCR reactions using a Zymo Clean & Concentrator -5 (Zymo Research).


- Nick the bulk amplified subpools. Split the following across multiple tubes depending on the amount of DNA to be processed. In each 1.5 mL tube add:
  - 4.5 uL Nt.BspQI (10U/uL) (New England Biolabs)
  - 2 to 2.5 ug of DNA (final concentration ~16ng/uL)
  - 15 uL NEBuffer 3 (New England Biolabs)
  - UltraPure Distilled water (Invitrogen) to 150 uL
- Leave at 50°C overnight with shaking >1500 RPM.

5. Capture and remove the short biotinylated fragment.

- Wash 50 uL Dynabeads M-270 Streptavidin (Invitrogen) for each 1.5 mL tube in the nicking reaction, as per manufacturer’s instructions and resuspend in 2X B&W buffer.
- Add 50 uL of washed beads to the 150 uL nicking reaction in each tube.
- Incubate at 55°C with 800 RPM shaking for at least 1 hour.
• Move all 1.5 mL tubes to a 55°C water bath.

• Place the tube so that solution is just below the surface of the water. Hold a strong magnet underwater against the side of the tube to magnetically separate Dynabeads. Pipette the supernatant, which contains the processed oligos and save them in a new container. Remove the tube with the Dynabeads from the magnet. Add 100 uL of UltraPure Distilled water (Invitrogen) to the tube and resuspend the beads. Incubate these at 55°C for another 30 min and then repeat the procedure to recover the supernatant again while leaving the Dynabeads behind.

• Repeat this procedure for all tubes as necessary.

• Pool processed oligos (supernatant) for each subpool and column cleanup using a Zymo Clean & Concentrator -5 (Zymo Research).

6. Capture processed oligos with barcoded beads.

• Take 18 uL of the pooled barcoded beads. These are in stored in B&W buffer (high ionic concentration) which may interfere with ligation reaction. Resuspend them in 18 uL 10 mM Tris-HCl buffered solution.

• Mix the processed DNA with the barcoded beads:
  - 40 uL processed DNA (~1.3 ug, ~12 pmol)
  - 18 uL pooled barcoded beads (~5 million beads, binding capacity 1.2 ug DNA)
  - 10 uL 10X Taq ligase buffer (New England Biolabs)
  - 4 uL Taq ligase (40 U/uL) (New England Biolabs)
  - 28 uL UltraPure Distilled water (Invitrogen)
  - TOTAL: 100uL

• Overnight cycling (>2 hr incubation at each of the following temperatures) (13 hr), while shaking using an Eppendorf ThermoMixer C (Eppendorf):
  - 3 hours @ 50°C
  - Ramp to 40°C for 3h, 0.1°C/sec
- Ramp to 30°C for 3h, 0.1°C/sec
- Ramp to 20°C for 2h, 0.1°C/sec
- Ramp to 10°C for 2h, 0.1°C/sec

- Wash 3 times at 4°C using B&W buffer. This is important for removing unbound oligos in order to increase specificity.
- Wash twice at RT using B&W buffer
- Re-suspend in 100 uL Elution Buffer (Qiagen) (~50k beads/uL)

7. Emulsion assembly (ePCA).

- Setup emulsion. All of this procedure should be done in cold room on ice. Add Bts α I only at very last step. Try to minimize the time between adding the Bts α I and vortexing the emulsion.
  - 10 uL of loaded beads (~130 ng DNA)
  - 0.5 uL 100 uM FWD assembly primer
  - 0.5 uL 100 uM REV assembly primer
  - 50 uL Kapa2G Robust HotStart ReadyMix (2X) (KAPA Biosystems)
  - 1 uL BSA (New England Biolabs)
  - 31 uL UltraPure Distilled water (Invitrogen)
  - 7 ul Bts α I (New England Biolabs) (add last)
  - TOTAL: 100 uL

- Mix at low speed in vortexer to resuspend beads.

- Add 600uL Droplet Generation Oil for EvaGreen (Bio-Rad Laboratories) to a 1.5mL non-stick tube.

- Add 100uL aqueous phase to the bottom of the oil phase.

- Vortex at Max Speed in foam holder taped down for 3-4 minutes. If doing multiple emulsions, do this one at a time. We use a Vortex Genie 2 (Scientific Industries) at max speed.
• After vortexing all emulsions, place each emulsion into PCR tubes with 100 uL in each tube. Use a P1000 tip to avoid disturbing the emulsion. Most of the droplets will float to the top of the tube, try to get as much of this as possible and distribute this over multiple PCR tubes.

• PCR Cycling
  – 55°C for 90 min (allow Bts α I to cleave DNA from the beads)
  – 94°C for 2 min (initial denaturing)
  – 94°C for 15 sec (denaturing)
  – 57°C for 20 sec (annealing)
  – 72°C for 45 sec (extension)
  – Go to step 3 for additional 60 cycles
  – 72°C for 5min (final extension)
  – 4°C forever

8. Break the emulsion. Adapted from pg 69 of the Bio-Rad Droplet Digital PCR Applications Guide:

• After ePCA, pipet out the entire volume of droplets from each PCR tube into a 1.5 mL tube. Combine up to 400 uL, in each tube. Note: phase-lock tubes can also be used here to improve recovery.

• Carefully pipet and discard bottom oil phase after droplets float to the top. Press a P1000 down to its first stop, push through the droplets to the bottom of the tube, press down to the second stop to expel any droplets, then wait several seconds for the droplets to float back up to the droplet layer, and finally aspirate out the oil. You do not need to remove every last bit of oil - just remove most of it.

• Add 50 uL of TE buffer for each 100 uL of PCR reaction combined in the 1.5mL tube.

• In a fume hood, add 175 uL of chloroform for each PCR reaction in the tube. (If there are 4 PCR reactions in a tube than contents will be: <400uL PCR reactions, 200uL TE, 700 uL chloroform).
• Vortex at maximum speed for 1 min.

• In a centrifuge, spin down at 15,500 x g for 10 min.

• Remove upper aqueous phase by pipetting, avoiding the chloroform phase.

• Transfer this to a clean 1.5mL tube (this is the DNA).

• Proceed to column or SPRI bead cleanup (Beckman) for the recovered DNA.


• The amplicons will often be mixed with undesired lower-molecular weight assemblies. Removing these using size selection will increase final yield. Choose one of the following three approaches, ordered from highest yield to lowest yield:
  – Pippin Prep (Sage Science).
    1. Follow manufacturer’s protocol (calibration, checking currents, loading, etc...)
    2. Make sure to allow for a range broad enough to include every member of the library, yet narrow enough to exclude some of the shorter non-specific products (+/- 100 bp is usually fine).
    3. Collect the eluted product and column cleanup using a Zymo Clean & Concentrator -5 (Zymo Research).
  – or Gel extraction.
    1. Run amplicons on a gel and extract the correct range and purify.
    2. Note: Typically there is not enough DNA after the ePCA to visualize on a gel, so this is often a blind extraction.
  – or No size selection.
    1. Make a dilution of ePCA and use this as template for the re-amplification.

10. Re-amplification.

• Amplify ePCA products using Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems).
  – 0.2 - 2 uL template
- 1 uL 10 uM FWD assembly primer
- 1 uL 10 uM REV assembly primer
- 25 uL Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems)
- UltraPure Distilled Water (Invitrogen) to 50 uL
- TOTAL: 50 uL

PCR protocol:
1. 3 min 95°C initial denaturation
2. 15 sec 98°C denaturation
3. 20 sec 58°C annealing
4. 45 sec 72°C extension
5. Go to step 2, determine cycles using qPCR.
6. 3 min 72°C final extension

- Column purify re-amplified products using a Zymo Clean & Concentrator -5 (Zymo Research).
- Check size distribution on gel or tapestation.
- Quantify DNA and proceed to downstream applications.

Supplementary Text

PPAT complementation assay

There are several reasons homologs could have low fitness including environmental mismatches, improper folding, mismatched metabolic flux, interactions with other cytosolic components, or gene dosage toxicity effects resulting from improperly high expression. Of the homologs from extremophilic bacteria, only alkaliphiles showed slightly reduced fitness values which is not significant (p-value = 0.059 Wilcoxon) (Fig. 4.23B). Metabolic mismatch is unlikely since so many homologs were able to complement well and both CoA and dephospho-CoA act as inhibitors implementing negative feedback loops to control the metabolic flux through the pathway [17]. Control experiments revealed that high expression levels of wild-type *E. coli* PPAT result in growth defects, while similar levels of
expression for many other homologs had no impact (Table 4.2). This observation parallels similar findings for *E. coli* DHFR where wild-type overexpression was toxic while overexpression of homologs had no detrimental effects [19], an effect linked to evolved protein-protein interactions that confer benefits at physiological concentrations. PPAT interaction partners include several enzymes encoded by essential genes such as *leuS*, *murE*, and *rplD* [42, 43].

**Broad Mutational Assay (BMS)**

Although 87% of the 3,180 possible mutations are covered, the coverage is strongly correlated with position fitness ($\rho$=0.76; Pearson, p-value <2.2E-16) (Fig. 4.24C), implying that many mutations that are depleted in the pooled assay (and typically represented by a only a few assembly barcodes), never pass the 10-read threshold used to filter assembly barcodes, an issue that can be resolved by sequencing the initial library to a greater depth. Unlike traditional mutagenesis approaches, the presence of multi-bp deletions from the oligo synthesis process also allows us to evaluate the effect of removal of entire residues from the sequence (Del. in Fig. 4.4A).

**Gain-of-Function Mutants**

In *E. coli*, residue Glu-134 and proximal Leu-102 have hydrophobic interactions with the cysteamine moiety of CoA [42], suggesting that some GoF mutations play roles in tuning CoA inhibition, while Ala-103 participates in hydrophobic interactions contributing to dimer formation [17]. Residues 64, 68, 69 are surface-exposed in the hexameric PPAT complex and are possible candidates for interactions with other proteins. As many of these mutations had only a single assembly barcode, we estimated a false positive rate of 0.9% derived from the number of positive fitness mutants for negative controls (Fig. 4.19A).

**Acknowledgements**

This work was supported by the funds from the Human Frontier Science Program [LT000068/2016 to C.P.], Netherlands Organisation for Scientific Research Rubicon fellowship [to C.P.], National Science Foundation Graduate Research Fellowship under Grant No. 2016211460 [to A.M.S], a Ruth L. Kirschstein National Research Service Award [GM007185 to N.L.], National Institutes of Health
New Innovator Award [DP2GM114829 to S.K.], Searle Scholars Program [to S.K.], Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank Jeff Sampson and Paige Anderson at Agilent Technologies for oligo pools and critical advice. We thank George Church and Richard Terry for guidance during the early developments and Suhua Feng, the UCLA BSCRC Sequencing Core, and the Technology Center for Genomics & Bioinformatics for providing NGS services. S.K. and D.Z. are named inventors on a patent application on the DropSynth method (US14460496). The scripts required to generate DropSynth oligos are available at https://github.com/kosurilab/DropSynth. Sequencing data are available from the sequencing read archive (SRA) with the accession number SRP126669.
Figure 4.5: The histogram of read distributions for six of the 96-plex 4-oligo assemblies shown in Fig 1B. **A.** T7 ligase and 20 ug beads. **B.** T4 and 20 ug beads. **C.** Taq ligase and 20 ug beads. **D.** T7 ligase and 100 ug beads. **E.** T4 ligase and 100 ug beads. **F.** Taq ligase and 100 ug beads.
Figure 4.6: **A.** A maximum likelihood phylogenetic tree for all 1,152 PPAT homologs as well as *E. coli* MG1655. Color scale represents percent amino acid sequence identity relative to *E. coli* PPAT (NP_418091.1). **B.** The gene length distribution for the 5,775 DHFR homologs assembled using either four or five 230-mer oligos with median gene lengths of 489 bp and 564 bp respectively.
Figure 4.7: A. Histogram of protein sequence lengths for all 1,152 PPAT library members. Lengths do not include start or stop codon. The longest, shortest, and median lengths are 516, 381, and 483 bp respectively. B. Although they share the same function, PPAT homologs have evolutionarily divergent sequences. The 662,976 pairwise percentage identities between the 1,152 members of the PPAT library at the amino acid level have a distribution with a median of 50% (σ = 5%). C. Without oligo isolation, amplification in bulk fails to produce the correct product [11]. A 4% agarose gel comparing the assembly products of a 24-member library of PPAT homologs (120 oligos) when the polymerase cycling assembly is done in bulk (BA) and in emulsion (EA). The expected product size upon correct assembly is between 520 bp to 550 bp. D. Each of the three 384-member PPAT libraries (1,920 oligos each) produced correct assembly products. A 4% agarose gel showing amplified assembly products, with the expected size for most amplicons around ~530 bp. Lane 1 and 2: High- and low-template PCR products for Lib 1. Lane 4 and 5: High- and low-template PCR products for Lib 2. Lane 7 and 8: High- and low-template PCR products for Lib 3. High- and low-template concentrations refer to either 2 uL or 0.2 uL of the purified assembly products from an emulsion used in a 50 uL PCR reaction.
Figure 4.8: Agilent TapeStation gel image of DropSynth assembly of 28 384-member libraries of DHFR. A total of 3 libraries of length 610bp (14, 15, 29) are assembled using 5 oligos while the remaining libraries of length 510bp are assembled using 4 oligos. Another 2 libraries (13, 30) are not shown with one having low yield on the oligo processing steps and another failing to amplify at the oligo stage.

Figure 4.9: Agilent TapeStation gel image of 25 4-oligo DHFR libraries after assembly, digestion, ligation into barcoded plasmid and library preparation for sequencing. 5-oligo libraries (14, 15, 29) were not prepared for sequencing due to limitations on Illumina read length capabilities.
Figure 4.10: **Sequencing statistics from sample S0.** These data are a set of paired end 600-cycle Miseq runs which read through the entire assembled gene and its assembly barcode for all three 384-member libraries. 

**A.** The number of reads per assembly barcode, with a median value of 2. S0 contains 7,038,274 unique assembly barcodes across 20,263,445 reads. Of these, 209,868 assembly barcodes (2.98%) (739,771 reads 3.65%) mapped to the designed protein sequences without any amino acid mutations, of which 199,208 assembly barcodes contained at least one synonymous mutation. A total of 2,982,539 (42%) of the mapped assembly barcodes correspond to sequences containing a premature stop codon in the reading frame, of which the large majority (2,404,348) were due to indel mutations causing a frameshift while the rest were due to nonsense mutations.

**B.** The long tail distribution of assembly barcodes per homolog, for assembly barcodes mapped to a perfect sequence. Median value is 56 and a total of 872 out of 1152 homologs are represented with at least one assembly barcode.

**C.** The percentage of perfect protein sequences for constructs with at least 100 assembly barcodes. The solid line is the median value of 1.9%.

**D.** Individually rank-ordered plots showing the number of barcodes with perfect assemblies, barcodes with assemblies within distance of 2 a.a., and all barcodes with an aligned homolog.

**E.** The distribution of sequencing reads for the PPAT libraries.

**F.** The coverage of the PPAT homologs as a function of the minimum percent identity. Most of the library members have assemblies with high identity to the respective designed homologs.
Figure 4.11: **A.** The library coverage shows strong correlation ($\rho=0.73$ (Pearson), p-value=3.4E-5) with the amount of DNA used to load the DropSynth beads prior to assembly. The coverage is defined as the number of constructs with at least one perfect assembly. **B.** The number of constructs with the same barcode which dropout among different libraries. The red line is the level with an expectation value close to one for libraries of size 384 given a uniform dropout distributions. Values above this line are higher than would be expected by chance. About a dozen barcodes fall in this region.
Figure 4.12: **DropSynth assembly of 10,752 genes.** We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of barcodes per million representing assemblies within 5 a.a. of each gene is shown alongside the number of library members with at least one perfect assembly and the percent perfects determined using constructs with at least 100 barcodes.
Figure 4.13: **A.** The expected percentage of perfect assemblies for a given number of oligos and the amount of perfect oligos. **B.** The maximum gene assembly length possible for a given number of oligos and an oligo size ranging from (200 to 300bp).
Figure 4.14: Error analysis of DropSynth Assemblies. Using the error analysis pipeline developed by Lubock et. al [16], we randomly sampled one million reads from Miseq paired-end 600-cycle assembly barcode mapping data, performed an exhaustive alignment of each read against every perfect assembly and returned the best scoring alignment. A. Mismatches are the most common form of error, followed by multiple base deletions, single base deletions, and single base insertions. In particular, mismatches appear to be localized to the overlap regions. B. Raw counts of mismatches. A higher number of transitions than transversions were measured - in agreement with previous experiments where Taq-mediated amplification errors. This suggests that the majority of mismatches were likely introduced by KAPA2G Robust polymerase during assembly (evolved Taq variant).
Figure 4.15: Phosphopantetheine adenyllytransferase (PPAT) metabolic pathway. PPAT shown in red, catalyzes the second to last step in the five step biosynthesis of coenzyme A. It produces dephospho-coenzyme A from 4'-phosphopantetheine by transferring an adenylyl group from ATP [17], as shown. Either Mn$^{2+}$ or Mg$^{2+}$ acts as a cofactor. *E. coli* PPAT is hexameric and encoded by the 477 bp gene *coaD*. Several gene knockout [44, 45] and genetic footprinting [46] studies have confirmed *coaD* to be essential for growth on rich media in *E. coli* K-12 strains MC1061, MG1655, and DH10β. Both coenzyme A and dephospho-coenzyme A act as inhibitors of the forward reaction. PPAT’s low homology to its mammalian counterpart, which is encoded as one of the two domains on the bifunctional CoASy (CoA Synthase) enzyme, makes it a potential target for new antimicrobials [18]. At least a dozen different PPAT homologs have crystal structure data available.
Figure 4.16: A. Rescue plasmid pTKcoaD allows λ−red recombination of the essential coaD gene. Wild-type E. coli coaD is expressed constitutively along with GFP, which allows for confirmation of plasmid loss upon heat curing. B. High-copy expression plasmid pEVBC allows for IPTG-inducible expression of an homolog PPAT gene cloned in between the Ndel and KpnI sites. A 20-mer random assembly barcode is present downstream. C. Verification of the coaD gene knockout using colony PCR with two sets of internal primers. Four 42°C heat-cured colonies (c1-c4) are shown as well as four colonies (c5-c8) grown at 30°C which still contain the rescue plasmid. Red arrows indicate expected amplicon size when coaD gene sequence is present. D. Colony PCR verification of the coaD genomic knockout using external genomic primers for 9 knockout colonies and one wildtype control. Wildtype (no knockout) amplicon length is 590 bp while the knockout (KAN cassette knockin) amplicon length is 1150 bp, as marked by the red arrows. E. Comparison of E. coli DH10β Δ coaD pTKcoaD cells grown at 30°C (left) and 42°C (right). Cells were grown in LB+Kan for 15 hours at the corresponding temperature, to allow for sufficient outgrowth, before plating on LB+Kan and incubating at the corresponding temperature. By comparing the number of GFP-positive colonies seen in each case we estimated an escape frequency of 1 in 16,500 (σ = 1,600). We also tracked the escape frequency of cells after transformation with PPAT homologs and growth at 42°C, by determining the ratio of GFP negative to GFP positive cells, finding an escape frequency of 1 in 20,200 (σ = 9500) as determined by 8 independent transformations. These escape frequencies are similar to those previously reported for coaD (a.k.a. kdtB) upon heat curing of coaD expressing pMAK705 plasmid in a conditional knockout [44].
Figure 4.17: PPAT complementation assay. A. The fitness values for 651 homologs across two independent biological replicates shows strong correlation ($\rho=0.94$; Pearson). Six negative controls lacking the H/TxGH motif required for nucleophilic attack on the $\alpha$ phosphate of the ATP have very low fitness values (<3) in the assay. We colored each point based on the number of assembly barcodes that corresponded to errorless constructs, and find that reproducibility among replicates improves with increasing number of assembly barcodes (Fig. 4.18B). C. Despite having a median 50% sequence identity, distant homologs are typically still able to complement the function of the native E. coli PPAT (bottom row). This multiple sequence alignment table shows the fitness scores, percent sequence identity to E. coli PPAT, and source organism.
Figure 4.18: **A.** Fitness values of 329,897 individual assembly barcodes in each biological replicate, with a correlation of 0.948. A large number of low-fitness assembly barcodes correspond to assemblies with frameshifts due to indels. **B.** We see the reproducibility of the fitness values increase with the number of assembly barcodes. The absolute difference in homolog fitness values between the two biological replicates as a function of their number of assembly barcodes ($\rho=-0.34;\text{Spearman, } p\text{-value }<2.2\times10^{-16}$). **C.** Fitness values are noisy with a median standard deviation of around 2.4. Box plots of individual assembly barcode fitness values for homologs in replicate A which have at least 50 assembly barcodes. Homologs are rank-ordered by their final fitness value.
Figure 4.19: **A.** Assembly barcode fitness for six of the homologs missing the H/TxGH motif required for catalytic activity. No simple mutation would be able to restore catalytic activity to these homologs, so they serve as a useful measure of the false positive rate for individual assembly barcodes. Of the 994 assembly barcodes only 9 assembly barcodes (0.9%) have a positive fitness value, indicating a low rate of false positives at the individual barcode level. **B.** Mean sequence fitness is reduced with increasing number of mutations ($\rho$=−0.38; Spearman, p-value <2.2E-16). Analysis of 144,573 sequences’ fitness as a function of their a.a. distance from the designed homolog sequence. **C.** Very few sequences with less than $\sim$94% sequence identity show high fitness. For sequences represented by at least 2 assembly barcodes, we plot their fitness as a function of their sequence identity (relative to their corresponding designed sequences), within bins of 1%.
Figure 4.20: The population of perfect and low mutational distance sequences expand as a function of time, while sequences with low sequence identity (primarily due to indels) are depleted. We see that non-functional assemblies are lost from the population primarily between the first two dilutions. Distribution of mapped assembly barcodes (top) and mapped reads (bottom), for each replicate (left & right), based on distance from the designed sequence.
Figure 4.21: **Synthesis verification.** Sequence-verified clones were obtained for 37 of 49 homologs. **A.** The amount of colonies observed after transformation of amplified constructs into *E. coli* DH10β Δ coaD pTKcoaD cells grown at 30°C (positive control) and 42°C (complementation). Symbol indicates 42°C colony size relative to 30°C colonies. Dashed line shows slope of one and is not a fit. The presence of a cluster with low colony counts in both conditions made up primarily of low-fitness homologs suggests possible toxicity effects. Two false positives are observed which had positive fitness in the pooled assay but produced no colonies in this transformation. Both of these had a low number of assembly barcodes (1 and 25). The majority of high fitness homologs produced large numbers of colonies in both conditions with high correspondence between the two. **B.** Comparison of growth rate of individual homologs (log-scale) and gain-of-function mutants as determined on a plate reader with experimentally-determined fitness from pooled complementation assay, with a Spearman’s correlation of \( r_S = 0.86 \). Growth rate (hr\(^{-1}\)) is defined as the maximum slope of OD600 vs. time on a log/linear plot. Fit is carried out using log growth rate and does not include the eight homologs with a growth rate of zero. Wildtype PPAT *E. coli* had a growth rate of 0.132 indicative of gene dosage toxicity effects due to overexpression. **C.** Correlation between the residual error of the fit of growth rate to fitness and number of assembly barcodes in homologs (\( r_S = -0.50 \), Spearman, p-value 1.7E-3). Constructs with fewer assembly barcodes tend to have higher error between individual growth rate and fitness in the pooled assay, highlighting the need for many assembly barcodes to determine fitness.
Figure 4.22: **PPAT phylogenetic tree.** The majority of homologs listed complement wildtype *E. coli*, with low-fitness homologs randomly dispersed throughout the tree with minimal clustering. A phylogenetic tree of 451 homologs labeled, similar to Fig. 4.3D, with each leaf labeled with the organism name and shaded by fitness.
Figure 4.23: **A.** Phylogenetic tree of 411 homologs based on NCBI taxonomy rather than PPAT sequence, generated using phyloT (http://phylot.biobyte.de). The median fitness was used when multiple sequences were annotated with the same taxonomic ID. **B.** Fitness of PPAT homologs from organisms annotated as extremophiles. Of the different classes, alkaliphiles show a weak shift to lower fitness values (p=0.059 Wilcoxon rank sum test). Previous characterization of *E. coli* PPAT showed a maximum activity at pH 6.9 which was reduced to 68% of the maximum by pH 8 [47].
Figure 4.24:  
A. The average BMS position fitness compared to the conservation (Jensen-Shannon divergence). As expected mutations tend to be more constrained at highly conserved sites ($\rho=-0.64$; Pearson, p-value <2.2E-16).  
B. The average BMS position fitness compared to the relative solvent accessibility based on a DSSP analysis of the 1H1T crystal structure (dimer not hexamer). Buried residues tend to be more constrained ($\rho=0.42$; Pearson, p-value 3.9E-8).  
C. Mutational scanning coverage decreases at site of low fitness ($\rho=0.76$; Pearson, p-value <2.2E-16). This effect is due to assembly barcodes with low read numbers which, due to their low fitness, never pass the minimum 10 read threshold.  
D. Residues appearing in wildtype E. coli PPAT are associated with higher fitness values. The distribution of fitness values for residues present in the E. coli PPAT sequence (median = 2.16, $\sigma = 0.24$) compared to all others (median = 1.86, $\sigma = 2.16$).
Figure 4.25: Variant classifier. We implemented a classifier to predict how different BMS variants would perform in our assay. Each BMS variant was categorized into two bins based on whether or not their measured fitness score was greater than 0. We then performed a logistic regression using 6 features for our model - the amino acid mutation, secondary structure class as assigned by DSSP (loop, beta-sheet, or alpha-helix), relative solvent accessibility as assigned by DSSP, sequence conservation, evolutionary coupling as predicted by EVMutation, and the frequency of residue substitution from the sequence alignment used for EVMutation’s prediction. To assess the performance of our classifier, we performed 10 repeats of 5-fold cross-validation on our dataset and measured the precision and recall of each model on its respective hold-out set. We found that on average, our simple classifier has A. an average accuracy of 0.825 +/- 0.013, B. a precision of 0.853 +/- 0.009, and an average recall of 0.931 +/- 0.014.
Figure 4.26: A. The relative solvent accessibility and conservation of each of the eight gain of function positions. B. Weblogo showing the probability of each residue at the gain-of-function positions for low-fitness homologs. C. Weblogo of GoF residues for homologs which complemented. D. The mean fitness of each GoF mutation at the significant positions, with the number of mutants observed at each a.a. E. The same plot with the data derived from the broad mutational scan using complementing homologs and their mutants. F. *E. coli* PPAT structure with the eight GoF residues shaded in red. Glu-134 is involved in hydrophobic interactions with coenzyme A [42], suggesting a role for GoF mutations in modulating the inhibitory feedback, while Ala-103 participates in hydrophobic interactions between the PPAT dimers.
Figure 4.27: A. The oligo design process. Briefly, a.a sequences are assigned random weighted codons and appended with restriction and primer sites used in DropSynth assembly. Sequences are then split into five oligos with ~20-nt overlap regions. Individual oligo sequences are appended with restriction sites, padding sequences, gene-specific microbead barcodes flanked by nicking sites, and amplification primer sites leading to a library of 200-nt sequences. B. The DropSynth microbead barcoding process. Microbead barcode oligos are individually mixed with 3’ biotinylated ligation oligos and dual 5’ biotinylated anchor oligos, ligated using T4 ligase and phosphorylated with T4 PNK, exposing the microbead barcode sequence (NNNNNNNNNNNN). Biotinylated duplexes are then individually bound to M270 streptavidin Dynabeads and pooled together.
Figure 4.28: Nick processing to generate single-stranded microbead barcode overhang. A. A 10% TBE-Urea denaturing gel highlighting the steps in nick processing. Lanes 1, 5, 7: a 10 bp ladder. Lane 2: Before processing, all oligos should be 200 nt. Lane 3: After nick processing we expect fragments of 165 nt, 177 nt, 35 nt, and 23 nt. Lane 4: After streptavidin Dynabead cleanup of nick processed oligos we expect fragments of 165 nt and 177 nt. Lane 6: The captured Dynabead fraction after boiling at 90°C for 10 min in 10 mM EDTA pH 8.2. B. A non-denaturing 4% agarose gel showing the nick processing which takes a 200 bp duplex and leaves a 12-nt single-stranded microbead barcode overhang on a 165 bp dsDNA fragment. Lanes p1-p4 showing several samples after nick processing and also one before processing (NP). Lanes b1-b4 show the corresponding Dynabead fractions after denaturing at 80°C for 3 min. Full length oligos containing errors in the nt.BspQI sites will not have both strands nicked and are likely to be pulled down by the Dynabeads together with the short fragment.
Figure 4.29: **Characterization of the distribution of droplet sizes for the vortex emulsions.**

Briefly, 100 µL of Kapa Robust buffer was added to an eppendorf tube with 600 µL of Bio-Rad Droplet Generation Oil and vortexed upright for 4 minutes on the highest setting of a Vortex-Genie 2. Samples were then taken from the bottom, middle, and top of the resulting emulsion and imaged under 40X magnification. The mode of the droplet diameter distribution peaks below 5 um. Scale bars are 100 um. Bottom right: Histogram of droplet diameters as determined by image analysis. Median droplet diameter is below 5 um.
Table 4.1: **Assembly barcode statistics for each serial dilution in the two biological replicates.** Barcodes for each sample were clustered using Starcode [35] to collapse barcodes within a Levenshtein distance of 1.

<table>
<thead>
<tr>
<th>Biological replicates</th>
<th>Serial dilution</th>
<th>Total reads</th>
<th>Total barcodes</th>
<th>Total clustered barcodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>9,051,752</td>
<td>4,317,940</td>
<td>4,289,165</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9,790,924</td>
<td>2,319,457</td>
<td>2,231,361</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8,222,783</td>
<td>1,346,284</td>
<td>1,263,430</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7,947,874</td>
<td>970,291</td>
<td>892,753</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>9,136,919</td>
<td>4,259,319</td>
<td>4,228,531</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8,319,364</td>
<td>1,919,591</td>
<td>1,843,449</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10,036,601</td>
<td>1,393,886</td>
<td>1,292,371</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9,437,037</td>
<td>993,877</td>
<td>907,884</td>
</tr>
</tbody>
</table>
Table 4.2: **Homologs and GoF mutants retrieved from the assembled library and individually tested in knockout (KO) PPAT cells.**

Growth rate (hr⁻¹) is defined as the maximum slope of OD600 vs. time on a log/linear plot. Wildtype *E. coli* PPAT and 3 catalytically inactive wildtype mutants were also prepared and tested.

<table>
<thead>
<tr>
<th>Type</th>
<th>Construct ID</th>
<th>Assembly barcodes</th>
<th>42°C CFU</th>
<th>30°C CFU</th>
<th>Growth rate (hr⁻¹)</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homolog</td>
<td>CDD12392</td>
<td>699</td>
<td>312</td>
<td>234</td>
<td>0.198</td>
<td>2.02</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_041531153</td>
<td>63</td>
<td>343</td>
<td>334</td>
<td>0.178</td>
<td>-0.94</td>
</tr>
<tr>
<td>Homolog</td>
<td>CDA36762</td>
<td>834</td>
<td>324</td>
<td>237</td>
<td>0.205</td>
<td>2.76</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_051012154</td>
<td>63</td>
<td>0</td>
<td>281</td>
<td>0</td>
<td>-3.46</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_012984121</td>
<td>1302</td>
<td>372</td>
<td>315</td>
<td>0.205</td>
<td>1.83</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_028874703</td>
<td>462</td>
<td>357</td>
<td>371</td>
<td>0.208</td>
<td>1.69</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_009532117</td>
<td>220</td>
<td>349</td>
<td>313</td>
<td>0.218</td>
<td>3.15</td>
</tr>
<tr>
<td>Homolog</td>
<td>CDC50010</td>
<td>429</td>
<td>361</td>
<td>406</td>
<td>0.216</td>
<td>3.39</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_025936372</td>
<td>1150</td>
<td>207</td>
<td>168</td>
<td>0.215</td>
<td>2.53</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_050330521</td>
<td>89</td>
<td>20</td>
<td>24</td>
<td>0</td>
<td>-3.58</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_028844278</td>
<td>38</td>
<td>174</td>
<td>166</td>
<td>0.146</td>
<td>-3.03</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_012096847</td>
<td>710</td>
<td>383</td>
<td>383</td>
<td>0.214</td>
<td>2.49</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_050708028</td>
<td>1154</td>
<td>362</td>
<td>300</td>
<td>0.212</td>
<td>4.05</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_027397238</td>
<td>40</td>
<td>20</td>
<td>44</td>
<td>0.131</td>
<td>-2.99</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_007413164</td>
<td>172</td>
<td>497</td>
<td>300</td>
<td>0.176</td>
<td>1.68</td>
</tr>
<tr>
<td>Homolog</td>
<td>KOS35328</td>
<td>25</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>1.88</td>
</tr>
<tr>
<td>Homolog</td>
<td>KHS64893</td>
<td>506</td>
<td>96</td>
<td>210</td>
<td>0.218</td>
<td>2.49</td>
</tr>
<tr>
<td>Homolog</td>
<td>KGB86419</td>
<td>185</td>
<td>585</td>
<td>491</td>
<td>0.233</td>
<td>2.94</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_025369197</td>
<td>131</td>
<td>0</td>
<td>305</td>
<td>0</td>
<td>-3.70</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_021271192</td>
<td>256</td>
<td>10</td>
<td>360</td>
<td>0.178</td>
<td>1.81</td>
</tr>
<tr>
<td>Homolog</td>
<td>KJF18279</td>
<td>242</td>
<td>10</td>
<td>36</td>
<td>0</td>
<td>-3.52</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_029522041</td>
<td>128</td>
<td>2</td>
<td>640</td>
<td>0.124</td>
<td>-1.77</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------</td>
<td>-----</td>
<td>---</td>
<td>-----</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_014806494</td>
<td>143</td>
<td>523</td>
<td>451</td>
<td>0.177</td>
<td>-1.90</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_038558636</td>
<td>938</td>
<td>18</td>
<td>24</td>
<td>0.193</td>
<td>0.58</td>
</tr>
<tr>
<td>Homolog</td>
<td>CDC19414</td>
<td>310</td>
<td>638</td>
<td>541</td>
<td>0.236</td>
<td>3.78</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_029455214</td>
<td>49</td>
<td>43</td>
<td>35</td>
<td>0.167</td>
<td>-3.49</td>
</tr>
<tr>
<td>Homolog</td>
<td>BAN02173</td>
<td>56</td>
<td>290</td>
<td>487</td>
<td>0</td>
<td>-3.46</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_008711239</td>
<td>693</td>
<td>787</td>
<td>642</td>
<td>0.229</td>
<td>3.28</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_039669974</td>
<td>51</td>
<td>5</td>
<td>41</td>
<td>0</td>
<td>-3.52</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_013656808</td>
<td>167</td>
<td>326</td>
<td>403</td>
<td>0.147</td>
<td>-3.46</td>
</tr>
<tr>
<td>Homolog</td>
<td>KJS87341</td>
<td>2691</td>
<td>220</td>
<td>172</td>
<td>0.215</td>
<td>-0.70</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_005674855</td>
<td>1059</td>
<td>305</td>
<td>290</td>
<td>0.256</td>
<td>3.84</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_011140849</td>
<td>4757</td>
<td>278</td>
<td>216</td>
<td>0.214</td>
<td>1.68</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_009360218</td>
<td>2986</td>
<td>281</td>
<td>77</td>
<td>0.222</td>
<td>3.25</td>
</tr>
<tr>
<td>Homolog</td>
<td>EUC78355</td>
<td>317</td>
<td>413</td>
<td>388</td>
<td>0.217</td>
<td>1.84</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_011433776</td>
<td>828</td>
<td>206</td>
<td>170</td>
<td>0.191</td>
<td>0.24</td>
</tr>
<tr>
<td>Homolog</td>
<td>WP_006440043</td>
<td>20</td>
<td>297</td>
<td>174</td>
<td>0</td>
<td>-3.55</td>
</tr>
<tr>
<td>GOF</td>
<td>WP_013656808</td>
<td>1</td>
<td>148</td>
<td>124</td>
<td>0.209</td>
<td>3.31</td>
</tr>
<tr>
<td>GOF</td>
<td>WP_029455214</td>
<td>1</td>
<td>361</td>
<td>394</td>
<td>0.184</td>
<td>3.34</td>
</tr>
<tr>
<td>GOF</td>
<td>WP_023508997</td>
<td>1</td>
<td>179</td>
<td>141</td>
<td>0.217</td>
<td>2.41</td>
</tr>
<tr>
<td>GOF</td>
<td>WP_049662705</td>
<td>1</td>
<td>415</td>
<td>404</td>
<td>0.204</td>
<td>3.82</td>
</tr>
<tr>
<td>GOF</td>
<td>WP_05425071</td>
<td>1</td>
<td>0</td>
<td>319</td>
<td>0</td>
<td>4.06</td>
</tr>
<tr>
<td>GOF</td>
<td>WP_044825986</td>
<td>1</td>
<td>211</td>
<td>264</td>
<td>0.211</td>
<td>4.40</td>
</tr>
<tr>
<td>Wildtype</td>
<td>NP_418091</td>
<td>196</td>
<td>207</td>
<td>0.132 ND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inactive</td>
<td>NP_418091_H18Y</td>
<td>95</td>
<td>142</td>
<td>0.105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inactive</td>
<td>NP_418091_H18D</td>
<td>0</td>
<td>130</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inactive</td>
<td>NP_418091_H18W</td>
<td>0</td>
<td>113</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.3: **Cost to create pool of 384 barcoded DropSynth microbeads.** Creating the pool of barcoded beads is a one time cost and produces enough beads to carry out at least 210 assemblies of 384 genes, or over 80,000 genes, using the current protocol.

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>38.4 nmol anchor oligo (5’ dual biotin modification)</td>
<td>$300</td>
</tr>
<tr>
<td>38.4 nmol ligation oligo (5’ phosphorylation and 3’ biotin modifications)</td>
<td>$540</td>
</tr>
<tr>
<td>0.1 nmol of each of the 384 barcoded oligos</td>
<td>$1656</td>
</tr>
<tr>
<td>1,575 µL 10X T4 ligase buffer</td>
<td>$5</td>
</tr>
<tr>
<td>80E9 Units of T4 ligase (concentrated)</td>
<td>$40</td>
</tr>
<tr>
<td>1,920 µL stock M270 streptavidin Dynabeads</td>
<td>$456</td>
</tr>
<tr>
<td>3.84E9 Units T4 PNK</td>
<td>$344</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>$3341</strong></td>
</tr>
<tr>
<td>Cost per assembly</td>
<td><strong>$15.69</strong></td>
</tr>
<tr>
<td>Cost per construct</td>
<td><strong>$0.04</strong></td>
</tr>
</tbody>
</table>

Table 4.4: **DropSynth assembly costs per 384 gene library.**

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarray derived OLS pool (Agilent Technologies; ∼$0.10/oligo)</td>
<td>$192</td>
</tr>
<tr>
<td>3x 50uL rxn KAPA Real-time Library Amplification Kit (KAPA Biosystems)</td>
<td>$8.4</td>
</tr>
<tr>
<td>8x 50uL rxn KAPA HiFi HotStart ReadyMix (KAPA Biosystems)</td>
<td>$33.6</td>
</tr>
<tr>
<td>2x columns Zymo Clean &amp; Concentrator -5 (Zymo Research)</td>
<td>$2.5</td>
</tr>
<tr>
<td>1 biotinylated primer (Integrated DNA Technologies)</td>
<td>$40</td>
</tr>
<tr>
<td>3 primers (Integrated DNA Technologies)</td>
<td>$10</td>
</tr>
<tr>
<td>2x SPRI cleanup (Beckman)</td>
<td>$1</td>
</tr>
<tr>
<td>1,200 uL Dynabeads M-270 Streptavidin (Invitrogen)</td>
<td>$285</td>
</tr>
<tr>
<td>50 uL Nicking enzyme (Nt.BspQI) (New England Biolabs)</td>
<td>$27</td>
</tr>
<tr>
<td>50 uL Kapa2G Robust HotStart ReadyMix (KAPA Biosystems)</td>
<td>$2.7</td>
</tr>
<tr>
<td>7 uL BtsI (New England Biolabs)</td>
<td>$7.3</td>
</tr>
<tr>
<td>600 uL BioRad Droplet Generation Oil for EvaGreen (Bio-Rad Laboratories)</td>
<td>$2.3</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>$612</strong></td>
</tr>
<tr>
<td>Cost per Construct</td>
<td><strong>$1.59</strong></td>
</tr>
<tr>
<td>Cost per Construct with Barcoded Beads</td>
<td><strong>$1.63</strong></td>
</tr>
</tbody>
</table>
Table 4.5: **Nick processing efficiencies for various conditions.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>nt.BspQI digest (min)</th>
<th>nt.BspQI (U/uL) in 150 uL digest</th>
<th>[DNA] (ng/uL) in digest</th>
<th>M270 Dynabead incubation (min)</th>
<th>[M270 Dynabead] (ug/uL)</th>
<th>Molar Yield (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>190</td>
<td>0.266</td>
<td>9.5</td>
<td>200</td>
<td>2.2</td>
<td>56%</td>
</tr>
<tr>
<td>ii</td>
<td>240</td>
<td>0.266</td>
<td>14.4</td>
<td>280</td>
<td>1.2</td>
<td>35%</td>
</tr>
<tr>
<td>iii</td>
<td>985</td>
<td>0.3</td>
<td>16</td>
<td>375</td>
<td>2.2</td>
<td>46%</td>
</tr>
<tr>
<td>iv</td>
<td>985</td>
<td>0.32</td>
<td>14.4</td>
<td>375</td>
<td>1.2</td>
<td>40%</td>
</tr>
<tr>
<td>v</td>
<td>390</td>
<td>0.3</td>
<td>15.7</td>
<td>35</td>
<td>2.2</td>
<td>47%</td>
</tr>
<tr>
<td>vi</td>
<td>390</td>
<td>0.34</td>
<td>18.7</td>
<td>50</td>
<td>2.2</td>
<td>44%</td>
</tr>
<tr>
<td>vii</td>
<td>390</td>
<td>0.32</td>
<td>17.6</td>
<td>70</td>
<td>2.2</td>
<td>36%</td>
</tr>
</tbody>
</table>

Table 4.6: **The oligos required for the bead barcoding process.** All oligos were ordered from Integrated DNA Technologies.

<table>
<thead>
<tr>
<th>Oligo Name</th>
<th>Sequence</th>
<th>Modifications</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligation oligo</td>
<td>TCCGCGAGTAAACCTAACAA</td>
<td>3’ biotin, 5’ phosphorylation</td>
<td>38.4 nmol</td>
</tr>
<tr>
<td>Anchor oligo</td>
<td>TTGTAGGTTTTACTCGCGGAA-CACGTGCTATTAGATGCCCT</td>
<td>5’ dual biotin</td>
<td>38.4 nmol</td>
</tr>
<tr>
<td>384X barcoded oligos</td>
<td>12-mer microbead barcode</td>
<td></td>
<td>0.1 nmol each</td>
</tr>
<tr>
<td></td>
<td>reverse complement + AGGCATCTAATAGCACGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Sequence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coaD_KO_KAN_FWD_1</td>
<td>AACGCATTGAGGTTGTTGAAGTFCCTATACTTTCTTAGAATAGGAACTTCGG-AATAGGAACTTCGG-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATACCATCCGCAAAACGAGTTTTCTAACCTCCAGATTTTCGAGAATAGGGAACTTCGGGATACCATCCGCAAAACGAGTTTTCTAACCTCCAGATTTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coaD_KO_KAN_FWD_2</td>
<td>GCTTCAACTGCTGGAACCTTACCTCCAGCAGAACGCTTACCTCCAGAAGTTGCAATACCATCCGCAAAACGAGTTTTCTAACCTCCAGATTTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coaD_KO_KAN_fwd_2</td>
<td>GCACCAGAAGTAAATCCCTAGGCGCAGCCGAGCATGCAGTACTAGTTTGAGGAGTCCAGTTGAGGGAGTACCCGAGAAGTTGCAATACCATCCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pEVBC_FWD</td>
<td>Biotin-GCCGTCATATGAGCTGTTTCCTGATTGAAATTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pEVBC_REV1</td>
<td>Biotin-GTGGGATCTACATAGTGTGCTGCGGAACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


Chapter 5

A Scalable, Multiplexed Assay for Decoding Receptor-Ligand Interactions

5.1 Abstract

G-protein coupled receptors (GPCRs) are the primary means by which mammalian cells sense and respond to chemicals. Using next-generation sequencing of barcoded genetic reporters in engineered human cell lines, we developed a platform to screen large chemical panels against multiplexed GPCR libraries. We used this platform to map 39 mammalian olfactory receptors against 181 odorants, revealing many new associations and deorphanizing 15 receptors.

5.2 Introduction

Interactions between small molecules and receptors underpin an organism’s ability to sense and respond to its internal state and the environment. For many drugs and natural products, the ability to modulate many biological targets at once is crucial for their efficacy [1, 2, 3]. Thus, to understand the effect of many small molecules, we need to comprehensively characterize their functional interactions with biological targets. This many-on-many problem is laborious to study one interaction at a time, and is especially salient in the mammalian sense of smell [4, 5].

† This work has been published as: E. M. Jones and R. Jajoo, D. Cancilla, N. B. Lubock, J. Wang, M. Satyadi, R. Cheung, C. de March, J. Bloom, H. Matsunami, S. Kosuri. “A scalable, multiplexed assay for decoding receptor-ligand interactions” bioRxiv pp. 348739, 2018
Olfaction is mediated by a class of G protein-coupled receptors (GPCRs) known as olfactory receptors (ORs)\[6\]. GPCRs are a central player in small molecule signaling and are currently targeted by 34% of US Food and Drug Administration (FDA) approved drugs \[7\]. ORs are a large family of class A GPCRs with approximately 396, 1130, and 1948 intact receptors in humans, mice, and elephants respectively \[8\]. Each OR can potentially interact with many odorants, and inversely, each odorant with many ORs. The majority of ORs remain orphan – i.e. have no known ligand – because of this vast combinatorial space, further compounded by the fact that recapitulating mammalian GPCR function in vitro is challenging \[9, 10\]. In addition, no experimentally determined structure for any mammalian OR is available, hindering computational efforts to predict which odorants can activate each OR \[11\].

Most GPCR and OR assays test chemicals against each receptor individually \[12, 13\]. Multiplexed assays, where the activities of multiple receptors – often referred to as a library of receptors – are measured in the same well, would increase the throughput but have remained technically challenging. In such an assay, each cell expresses a single type of receptor and, upon activation, transcribes a short barcode sequence that identifies the particular receptor expressed in that cell. The enrichment of barcoded transcripts corresponding to each receptor’s activation are then measured by microarrays or next-generation sequencing. Such multiplexed GPCR activity assays have previously been attempted by transient transfection of individual receptors and subsequent pooled screening \[14, 15\]. However, these assays are difficult to perform, especially in olfaction, for several reasons. First, ORs, like many GPCRs, are difficult to express in their non-native contexts and often require specialized accessory factors and signaling proteins to function heterologously \[16\]. Second, transient transfection must be performed for tens to hundreds of individual cell lines each time an assay is performed. Thus, experimental protocols for such multiplexed screens are expensive, labor intensive, and often carried out in a low-throughput manner. Using stable lines would alleviate these burdens, but building stable OR reporter lines is challenging and has only worked in two reported cases \[17, 18\].
Figure 5.1: **A Genomically Integrated Synthetic Circuit Allows Screening for Mammalian Olfactory Receptor Activation.** A. Schematic of the synthetic circuit for stable OR expression and function in an engineered HEK293T cell line (ScL21). Heterologous accessory factors expressed include (pink): RTP1S, RTP2, G\textsubscript{olf}, and Ric8b. B. MOR42-3 reporter activation expressing the receptor transiently (left) or genomically integrated (right) at varying copy number, under constitutive or inducible expression in HEK293T cells. C. MOR258-5 reporter activation with/without accessory factors (A.F.s), RTP1S and RTP2, transiently coexpressed in HEK293T cells compared to stable receptor expression in ScL21. D. Reporter activation response curves for MOR258-5 and MOR41-1 genomically integrated in ScL21 with/without doxycycline induction of receptor expression.

### 5.3 Results

Here we report a new high-throughput screen to characterize small molecule libraries against mammalian OR libraries in multiplex. To do this, we developed both a stable cell line capable of functional OR expression (ScL21) and a multiplexed reporter for OR activity (Fig. 5.1a, Supplementary Fig. 5.3). Activation of each OR leads to the expression of a reporter transcript with a unique 15-nucleotide barcode sequence. Each barcode identifies the OR expressed in that cell; this enables OR activation to be measured by quantifying differential barcode expression with RNA-seq.
This technology enables the simultaneous profiling of a single chemical’s activity against a library of receptors in a single well.

We first engineered a stable cell line, ScL21, capable of functionally expressing ORs and responding to odorant stimuli by transcribing an RNA barcode. First, we found that multi-copy integration and inducible receptor expression are both essential for reporter activation, but individually neither of these features is sufficient to generate a response (Fig. 5.1b, Supplementary Fig. 5.4). Then, to allow larger OR repertoires to be assayed, we added features known to improve OR function [16, 19, 20]. We stably integrated a pool of 4 accessory factors at multi-copy under inducible expression: G_olf and Ric8b for signal transduction, and RTP1S and RTP2 to promote surface expression (Fig. 5.1c, Supplementary Figs. 5.3 and 5.4). To select a single line for further use, we isolated clones and screened for robust activation of two ORs known to require accessory factors to function heterologously (Supplementary Fig. 5.4). In addition, we then incorporated protein trafficking tags previously shown to increase surface expression [21, 22], included DNA insulator sequences to reduce background reporter activation, optimized the cAMP response element (CRE) to improve reporter signal, and combined these improvements into a single transposable vector to speed cell line development (Supplementary Fig. 5.5). We validated our system on two murine ORs with known ligands, and observed induction- and dose-dependent activation (Fig. 5.1d).

To pilot the platform, we chose 42 phylogenetically divergent murine ORs with both known and unknown chemical specificities and created a library of OR-expressing cell lines (Supplementary Fig. 5.6). To engineer the individual cell lines, we first cloned and mapped the ORs to their corresponding barcodes and transposed the plasmids individually into the genomes of HEK293T cells [23]. After selection we pooled the cell lines together, generating assay-ready libraries for repeated testing (Fig. 5.2a). Unlike a luciferase reporter assay, each well contains the entire OR library and a single chemical’s activity is measured against the entire library of ORs in a single well. We seeded the cell library in 6-well culture dishes and screened odorants known to activate ORs in our library (Supplementary Fig. 5.7); all but 3 ORs passed quality filtering to obtain reliable estimates of activation (See Methods). Analysis of the sequencing readout recapitulated previously identified odorant-receptor pairs, and chemical mixtures appropriately activated multiple ORs [12] (Supplementary Fig. 5.7). We found the assay was robust to chemicals such as the adenylate cyclase
stimulator, forskolin, which non-specifically induces barcode transcription independent of the OR each cell expresses. This is likely because our library-based approach measures the relative activation of ORs to each other, normalizing any global effects due to off-target reporter activation.

Next, we adapted the platform for high-throughput screening in 96-well format. To decrease reagent cost and assay time, we optimized an in-lysate reverse transcription protocol and used dual indexing to uniquely link barcode reads to the correct well once samples were mixed for sequencing (see Methods). With these improvements, the assay is able to recapitulate dose-response curves for known odorant-receptor pairs (Supplementary Fig. 5.7). We observed reproducible results between identically treated but biologically independent wells (Supplementary Figs. 5.8 and 5.9).

We subsequently screened 181 odorants with both known and unknown receptor specificity at three concentrations in triplicate against the 39-member OR cell library, or 81,012 wells if each combination had been tested individually including controls (Fig. 5.2a and Supplementary Tables 5.1 and 5.2). Each 96-well plate in the assay contained independent positive control odorants and solvent (DMSO) for normalization (Supplementary Fig. 5.9). We used a generalized linear model to determine OR-odorant interactions (see Methods)[24]. We found 112 significant interactions (out of > 7,000 combinations), of which 79 are novel, and 24 that target 15 orphan receptors (Benjamini-Hochberg corrected FDR = 1%; Fig. 5.2b, Supplementary Fig. 5.10a, and Supplementary Table 5.3)[25]. Overall, 28 of 39 receptors were activated by at least one odorant, and 67 of 181 odorants activated at least one OR (Supplementary Table 5.3).

To validate our assay, we compared results to a previous study and also analyzed individual interactions in a different experimental context. First, we chose 36 interactions with at least 1.2-fold induction to retest individually in a previously developed transient OR activation system [26] (Supplementary Fig. 5.11). Of the 27 significant interactions at an FDR of 1%, 20 of them replicated in this orthogonal system (Supplementary Fig. 5.10). Notably, some of the seven interactions which did not replicate in this orthogonal system, appear to be true hits. For instance, our assay registered two hits for MOR19-1 with high chemical similarity (methyl salicylate and benzyl salicylate), suggesting they are likely not false positives (Supplementary Fig. 5.11). Additionally, three of nine interactions not passing the 1% FDR threshold showed activation in the orthogonal assay, indicating that a conservative FDR threshold likely generated some false negatives. A previous large-scale OR
deorphanization study screened some of the same receptors and chemicals, and we found that 9/12 of their reported interactions with EC$_{50}$ below 100µM were also detected in our platform, though we did not identify most of the previous low affinity interactions [12] (Supplementary Fig. 5.12). Conversely, we also detected 14 positive interactions absent from the previous study. Finally, our assay replicated the vast majority of non-interacting odorant-OR pairs (493/507).

Using the data generated by this high throughput assay, we found that chemicals with similar features activate the same ORs, including those receptors we deorphanize in this study (Fig. 5.2c). For example, the previously orphan MOR19-1 has clear affinity for the salicylate functional group, while MOR13-1 is activated by four chemicals with hydrogen bond accepting groups attached and—in three cases—to stiff non-rotatable scaffolds (Supplementary Data 1). We also detect ORs with partial overlap in chemical specificity; MOR13-1 detects compounds with terminal carbonyls while MOR258-5 detects cyclic conjugated molecules (Fig. 5.2d). Benzaldehyde, an intermediate size carbonyl, activates both ORs.

To more systematically understand how chemical similarity relates to receptor activation, we used a recently developed molecular autoencoder to computationally map each tested chemical onto a 292-dimensional continuous representation of chemical space and visualized the results with Principal Components Analysis [27] (Supplementary Fig. 5.13). Chemicals for 11/17 multi-hit receptors cluster together across the first two principal components (Supplementary Fig. 5.14). For instance, of 13 aliphatic aldehydes or carboxylic acids with >5 carbons in our chemical panel, 10 activate MOR5-1 (Fig. 5.2c, Supplementary Data 1). Interestingly, this analysis also highlights the instances where ORs are sensitive to several distinct sets of chemicals (Fig. 5.2e). For example, MOR139-1 is activated by compounds that belong to two distinct clusters: one with benzene rings and the other with cyclohexane rings, hinting at the selective features of these odorants. Similarly, MOR170-1 exhibits a broad activation pattern: this receptor responds to 50% of all odorants in our panel that contains both a benzene ring, and either a carbonyl or ether group. Most of these odorants form a single cluster with the exception of the acetate compounds that form a separate cluster. Understanding the global chemical space that activates each OR establishes the groundwork for the prediction of novel odorant-OR interactions.
Figure 5.2: Large-Scale, Multiplexed Screening of Olfactory Receptor-Odorant Interactions. A. Experimental workflow for OR library generation and high-throughput screening. To perform assay, we cloned OR genes and barcodes into plasmids, engineered cell lines via individual transposition of plasmids, pooled cell lines and performed screen in 96 well plates. We assayed the equivalent of 81,012 wells of a screen where interactions are tested individually. B. Heatmap of interactions from the screen clustered by odorant and receptor responses, and shaded by the minimum activating odorant concentration that triggered reporter activity. Only ORs and chemicals that registered at least one interaction are shown. C. Chemical names and structures for odorants that activate MOR23-1 and MOR5-1. D. Chemical names and structures for odorants that activate MOR258-5 and MOR13-1. E. Chemical hits identified for MOR170-1 and MOR139-1 (black) mapped onto a PCA projection of the chemical space of our odorant panel (grey). Shaded areas highlight hits that cluster together in chemical space.
5.4 Discussion

We anticipate that this platform can be scaled to test the 396-member human OR repertoire and comprehensively define OR response to any odorant of interest. The approximate cost per well is on par with existing assays, but per receptor-ligand interaction interrogated, multiplexing dramatically reduces cost and labor. Our incomplete understanding for how ligands [28], drugs [1], hormones, natural products [29] and odors [12] interact with potential cellular targets limits our ability to rationally develop new molecules to modulate receptor activity. Multiplex methods like this platform offer a scalable solution to generate large-scale datasets that will help guide both empirical and algorithmic efforts to better dissect the complex interactions between small molecules and biological targets [30].

5.5 Materials and Methods

Odorant-Receptor Activation Luciferase Assay (Transient)

The Dual-Glo Luciferase Assay System (Promega) was used to measure OR-odorant responses as previously described26. HEK293T cells (ATCC #11268) were plated in poly-D-lysine coated white 96-well plates (Corning) at a density of 7,333 cells per well in 100 ul DMEM+10% FBS (Thermo Fisher Scientific). 24 hours later, cells were transfected using lipofectamine 2000 (Thermo Fisher Scientific) with 5 ng/well of plasmids encoding ORs and 10 ng/well of luciferase driven by a cyclic AMP response element or 10 ng/well of a plasmid encoding both the OR and the luciferase gene, and in both cases 5 ng/well of a plasmid encoding Renilla luciferase. Experiments conducted with accessory factors included 5 ng/well of plasmids encoding RTP1S (Gene ID: 132112) and RTP2 (Gene ID: 344892). Inducibly expressed ORs were transfected with 1 ug/ml doxycycline (Sigma-Aldrich) added to the transfection media. 10-100 mM odorant stocks were established in DMSO or ethanol. 24 h after transfection, transfection medium was removed and replaced with 25 ul/well of the appropriate concentration of odorant diluted from the stocks into CD293 (Thermo Fisher Scientific). Four hours after odorant stimulation, the Dual-Glo Luciferase Assay kit was administered according to the manufacturer’s instructions. Luminescence was measured using the M1000 plate reader (Tecan). All luminescence values were normalized to Renilla luciferase activity.
to control for transfection efficiency in a given well. Data were analyzed with Microsoft Excel and R.

**Odorant-Receptor Activation Luciferase Assay (Integrated)**

HEK293T and HEK293T derived cells integrated with the combined receptor/reporter plasmids were plated at a density of 7333 cells/well in 100 µL DMEM+10% in poly-D-lysine coated 96-well plates. 24 hours later, 1 ug/ml doxycycline was added to the well medium. Odorant stimulation, luciferase reagent addition, and luminescence measurements were carried out in the same manner as the transient assays. Constitutively expressed ORs were assayed in the same manner without doxycycline addition. Data were analyzed with Microsoft Excel and R.

**Odor Stimulation and RNA Extraction for Pilot-Scale Multiplexed Odorant Screening**

HEK293T and HEK293T-derived cells transposed with the combined receptor/reporter plasmid were plated at a density of 200k cells/well in a 6 well plate in 2 mL DMEM+10%FBS. 24 hours later, 1 ug/ml doxycycline was added to the well medium. 10-100 mM odorant stocks were diluted in DMSO or ethanol. 24 hours after doxycycline addition, odorants were diluted in OptiMEM and media was aspirated and replaced with 1 mL of the odorant-OptiMEM solution. 3 hours after odor stimulation, odor media was aspirated and 600 µL of buffer RLT (Qiagen) was added to each well. Cells were lysed with the Qiashredder Tissue and Cell Homogenizer (Qiagen), and RNA was purified using the RNEasy MiniPrep Kit (Qiagen) with the optional on-column DNase step according to the manufacturer’s protocol.

**Pilot Scale Library Preparation and RNA-seq**

5 ug of total RNA per sample was reverse transcribed with Superscript IV (Thermo-Fisher) using a gene specific primer for the barcoded reporter gene (OL003). The reaction conditions are as follows: annealing: [65°C for 5 min, 0°C for 1 min] extension: [52°C for 60 min, 80°C for 10 min]. 10% of the cDNA library volumes were amplified for 5 cycles (OL004F and R) using HiFi Master Mix (Kapa Biosystems). The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 5 cycles of 98°C for 20 seconds, 59°C for 15 seconds, and 72°C for 10 seconds, followed by an extension
of 72°C for 1 minute. The PCR products were purified using the DNA Clean & Concentrator kit (Zymo Research) into 10 µL and 1 µL of each sample was amplified (OL005F and R) using the SYBR FAST qPCR Master mix (Kapa Biosystems) with a CFX Connect Thermocycler (Biorad) to determine the number of PCR cycles necessary for library amplification. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. After qPCR, 5 µL of the pre-amplified cDNA libraries were amplified a second time at the same cycling conditions as the first amplification with the same primers used for qPCR for 4 cycles greater than the previously determined Cq. The PCR products were then gel isolated from a 1% agarose gel with the Zymoclean Gel DNA Recovery Kit (Zymo Research). Library concentrations were quantified using a TapeStation 2200 (Agilent) and loaded at equimolar ratios onto a HiSeq 3000 with a 20% PhiX spike-in and sequenced with custom primers: Read 1 (OL003) and i7 Index (OL006).

**Pilot Scale Data Analysis**

To determine fold activation of each OR treated with each chemical, we first calculated the fraction of barcodes (composition) corresponding to each OR in the control treatment (DMSO). Then, we calculated the fold change in the composition of each OR in each a specific condition. As the barcode reads from activated ORs can dominate the composition of all reads and change the effective library size, we then normalized the activation of each OR by the median activation for each well. To be effective, this normalization assumes that fewer than half of the ORs are activated by an odorant.

**OR Library Cloning**

The backbone plasmid (all genetic elements except the OR and barcode) was created using isothermal assembly with the Gibson Assembly HiFi Mastermix (SGI-DNA). A short fragment was amplified with a primer containing 15 random nucleotides to create the barcode sequence (OL007F and R) using HiFi Master Mix. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 35 cycles of 98°C for 20 seconds, 60°C for 15 seconds, and 72°C for 20 seconds, followed by an extension of 72°C for 1 minute. The amplicon and the backbone plasmid were digested with restriction enzymes MluI and AgeI (New England Biolabs) and ligated together with T4 DNA ligase
DH5α E.coli competent cells (New England Biolabs) were transformed directly into liquid culture with antibiotic to maintain the diversity of the barcode library.

OR genes were received as a gift from Hiro Matsunami. OR genes were amplified individually with primers (OL008) adding homology to the barcoded backbone plasmid using HiFi Master Mix. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 35 cycles of 98°C for 20 seconds, 61°C for 15 seconds, and 72°C for 30 seconds, followed by an extension of 72°C for 1 minute. The amplified ORs were purified with DNA Clean and Concentrator Kit (Zymo Research) and pooled together. The barcoded backbone plasmid was digested with NdeI and SbfI and the OR amplicon pool was cloned into it using isothermal assembly with the Gibson Assembly Hifi Mastermix. DH5α E.coli competent cells were transformed with the assembly and antibiotic resistant clones were picked and grown up in 96-well plates overnight. The plasmid DNA was prepped with the Zyp py -96 Plasmid Miniprep Kit (Zymo Research). Plasmids were Sanger sequenced (OL109-111) both to associate the barcode with the reporter gene and identify error-free ORs.

**OR Library Genomic Integration**

HEK293T cells and HEK293T-derived cells were seeded at a density of 350k cells/well in a 6-well plate in 2 mL DMEM+10% FBS. 24 hours after seeding, cells were transfected with plasmids encoding receptor/reporter transposon and the Super PiggyBac Transposase (Systems Bioscience) according to the manufacturer’s instructions. 1 ug of transposon DNA and 200 ng of transposase DNA were transfected per well with Lipofectamine 3000 (Thermo Fisher Scientific). 3 days after transfection, cells were passaged 1:10 into a 6-well plate, and one day after passaging 8 ug/mL blasticidin were added to the cells. Cells were grown with selection for 7-10 days. The OR library was transposed individually and pooled together at equal cell numbers.

**Accessory Factor Cell Line Generation**

HEK293T derived cells were transposed with plasmids encoding the accessory factor genes RTP1S, RTP2, G_olf (Gene ID: 2774), and Ric8b (Gene ID: 237422) inducibly driven by the Tet-On promoter pooled equimolar according to the transposition protocol in the OR Library Integration section. Cells were selected with 2 ug/mL puromycin (Thermo Fisher). After selection, cells were seeded in a
96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 3 days and expanded to 24-well plates after 7 days. Clones were screened for accessory factor expression by screening them for robust activation of MOR258-5/Olf62 and OR7D4 with a transient luciferase assay (Supplementary Fig. 5.3). The clone with the highest fold activation for both receptors and no salient growth defects was established for the multiplexed screen.

**Transposon Copy Number Verification**

gDNA was purified from cells transposed with the OR reporter vector and from cells containing the single copy landing pad with the Quick-gDNA Miniprep kit. 50 ng of gDNA was amplified with primers annealing to the regions of the exogenous DNA from each sample using the SYBR FAST qPCR Master Mix (Kapa Biosystems) on a CFX Connect Thermocycler using the manufacturer’s protocol. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. Cq values for the transposed ORs were normalized to the single copy landing pad to determine copy number.

**Lentiviral Transduction**

Lentiviral vector was produced by transient transfection of 293T cells with lentiviral transfer plasmid, pCMVΔR8.91 and pCAGGS-VSV-G using Mirus TransIT-293. HEK293T cells were transduced to express the m2rtTA transcription factor (Tet-On) at 50% confluency and seeded one day prior to transduction. Clones were isolated by seeding cells in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 7 days and expanded to 24 well plates. Clones were assessed for m2rtTA expression by screening for robust activation of MOR42-3 (Gene ID: 257926) with a transient luciferase assay.

**High-throughput Odorant Screening**

The OR library cell line was thawed from a liquid nitrogen frozen stock into a T-225 flask (Corning) three days before seeding into a 96-well plate for screening. The library was seeded at 6,666 cells per well in 100 µL of DMEM+10% FBS. 24 hours later a working concentration of 1 µg/mL of doxycycline in DMEM+10% FBS was added to the wells. 24 hours after induction, the media was
removed from each plate and replaced with 25 ul of odorant diluted in OptiMEM. Each odor was added at three different concentrations (10 µM, 100 µM, 1 mM) in triplicate with the same amount of final DMSO (1%). Each plate contained two control odorants at a three concentration (10 µM, 100 µM, 1 mM) in triplicate and three wells containing 1% DMSO dissolved in media. The library was incubated with odorants for three hours in a cell culture incubator with the lids removed.

After odor incubation, media was pipetted out of the plates and cells were lysed by adding 25 µL of ice-cold Cells-to-cDNA II Lysis Buffer (Thermo Fisher) and pipetting up and down to homogenize and lyse cells. The lysate was then heated to 75°C for 15 minutes and flash frozen with liquid nitrogen and kept at -80°C until further processing. Then 0.5 µL DNase I (New England Biolabs) was added to lysate, and incubated at 37°C for 15 minutes. To anneal the RT primer, 5 ul of lysate from each well was combined with 2.5 µL of 10 mM dNTPs (New England Biosciences), 1 µL of 2 µM gene specific RT primer (OL003), and 1.5 µL of H2O. The reaction was heated to 65°C for 5 min and cooled back down to 0°C. After annealing, 1 µL of M-MuLV Reverse Transcriptase (Enzymatics), 1 µL of buffer, and 0.25 ul of RNase Inhibitor (Enzymatics) were added to each reaction. Reactions were incubated at 42°C for 60 min and the RT enzyme was heat inactivated at 85°C for 10 min.

For each batch, qPCR was performed on a few wells (OL005F and OL013) with SYBR FAST qPCR Mastermix to determine the number of cycles necessary for PCR based library preparation. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. After qPCR, 5 µL of each RT reaction was combined with 0.4 µL of 10 µM primers containing sequencing adaptors (OL005F and OL013), 10 µL of NEB-Next Q5 Mastermix (New England Biosciences) and 4.2 µL H2O, the PCR was carried out according to the manufacturer’s protocol. The forward primer contains the P7 adaptor sequence and an index identifying the well in the assay and the reverse primer contains the P5 adaptor sequence and an index identifying the plate in the assay. PCR products were pooled together by plate and purified with the DNA Clean and Concentrator Kit. Library concentrations were quantified using a TapeStation 2200 and a Qubit (Thermo Fisher). The libraries were sequenced with two index reads and a single end 75-bp read on a NextSeq 500 in high-output mode (Illumina).
Analysis of Next-Generation Sequencing Data

Samples were identified via indexing by their PCR index adapters unique for each well (5’ end) and unique for each plate (3’ end). The well barcodes followed the 7bp indexing scheme in Meyer et al., Cold Spring Harb Protoc, 2010. The plate indexing scheme followed the Illumina indexing scheme. Sequencing data was demultiplexed, and 15bp barcode sequences were counted with only exact matches by custom python and bash scripts.

Statistical Methods for Calling Hits

Count data was then analyzed using the differential expression package EdgeR [24]. To filter out ORs with low representation, we empirically set a cutoff that an OR had to contain at least 0.5% of the reads from more than 399 of the 1954 test samples. This filtered out 3 of 42 ORs which were underrepresented in the cell library (MOR172-1, MOR176-1 and MOR181-1).

Normalization factors were determined using the EdgeR package function calcNormFactors, and glmFit was used with the dispersion set to the tagwise dispersion, since only 39 ORs were present in the library and trended dispersion values did fit the data well. By fitting a generalized linear model to the count data to determine if odorants stimulated specific ORs, we were able to determine both the mean activation for each OR-odorant interaction and the p-value. We then corrected this p-value for multiple hypothesis testing using the built in p.adjust function with the Benjamini & Hochberg correction, yielding a False Discovery Rate (FDR)[25]. We set a cutoff of 1% to determine interacting odorant-OR pairs. For each interaction between an odorant and an OR, we further required that an OR-odorant interaction was above the cutoff in two different concentrations of odorant or in just the 1000 µM concentration.

Molecular Autoencoder

We used an autoencoder as described in Gómez-Bombarelli et al. to visualize OR-chemical interactions in the context of chemical space [27]. Following the authors’ advice, we used a reimplementation of autoencoder as the original implementation requires a defunct Python package (https://github.com/chembl/autoencoder_ipython). This model comes pre-trained to a validation accuracy of 0.99 on the entire ChEMBL 23 database with the exception of molecules whose SMILES are longer than
120 characters. We used this pretrained model to generate the latent representations of the 168 chemicals for which we could find SMILES representations and 250,000 randomly sampled chemicals from ChEMBL 23. We then used scikit-learn to perform principal component analysis to project the resulting matrix onto two dimensions [31].
5.6 Supplementary Information

Figure 5.3: Schematic of the Synthetic Olfactory Activation Circuit in the Engineered Cell Line. Full graphical representation of the expressed components for expression/signaling of the ORs and the barcoded reporter system as shown in Fig. 5.1 of the main text. Receptor expression is controlled by the Tet-On system (Orange). After doxycycline induction, the OR is expressed on the cell surface with assistance from two exogenously expressed chaperones, RTP1S and RTP2 (pink). Upon odorant activation, G protein signaling triggers cAMP production. Signaling is augmented by transgenic expression of the native OR G alpha subunit, $G_{\alpha_{olf}}$, and its corresponding GEF, Ric8b (pink). cAMP leads to activation of the kinase PKA that phosphorylates the transcription factor CREB leading to expression of the barcoded reporter.
Figure 5.4: Engineering HEK293T Cells for Stable, Functional OR Expression. **A.** Comparison of MOR42-3 activation under inducible receptor expression either transiently transfected (left) or integrated at single copy into the H11 genomic locus (right). **B.** Comparison of MOR42-3 reporter activation integrated at multiple copies in the genome with the PiggyBac Transposon System under constitutive or inducible receptor expression. **C.** Relative receptor/reporter copy number determined with qPCR for three transposed ORs relative to a single copy integrant. **D.** Comparison of MOR258-5 and MOR30-1 reporter activation (stimulated with 2-coumaranone and Decanoic Acid respectively) co-transfected with or without Accessory Factors (AF) G$_{olf}$, Ric8b, RTP1S, and RTP2. **E.** Cell line generation for stable accessory factor expression. After transfection, clones were isolated and screened for activation of ORs, MOR258-5 and OR7D4, that require accessory factors for functional expression. The dark purple bar represents the clone (ScL21) selected for further experiments.
Figure 5.5: **Design of a Multiplexed Genetic Reporter for OR Activation.**

**A.** Annotated Vector map for the plasmid containing the OR expression cassette and genetic reporter for integration.

**B.** MOR42-3 reporter activation in cells transiently co-expressing the receptor and genetic reporter on separate plasmids or together.

**C.** Fold activation of MOR42-3 driven by an engineered CRE enhancer (7 CREB binding sites) compared to Promega’s pGL4.19 CRE enhancer.

**D.** Genetic reporter basal activation upon inducible expression of MOR42-3 with or without a DNA insulator upstream of the CRE enhancer.
Figure 5.6: **Evolutionary Tree of Mouse ORs.** Phylogenetic tree inferred from amino acid sequence of functional murine ORs. The length of lines indicate degree of divergence between ORs. Red dots indicate ORs that were selected for inclusion in this study.
Figure 5.7: Pilot-Scale Recapitulation of Odorant Response in Multiplex. A. Heatmap displaying 39 pooled receptors’ activity against 9 odorants and 2 mixtures. Interactions are colored by the log2-fold activation of the genetic reporter (see methods). Odorant interactions previously identified are boxed in yellow [12]. B. Dose-response curves for odorants or forskolin (adenylate cyclase stimulator) at 5 concentrations screened against the OR library. Curves for ORs known to interact with the odorant are colored. Stimulation with forskolin does not show substantial differential activity between ORs in our assay.
Figure 5.8: **Library Representation.** Representation of individual ORs in the library for the 39/42 ORs that had sufficient cellular coverage (see Methods). **A.** Frequency of each OR as a fraction of the library determined by the relative activation of each reporter stimulated with DMSO. **B.** The relationship between frequency of each OR in the library and the average coefficient of variation between biological replicate measurements of reporter activation for all conditions.
Figure 5.9: **Replicability of the Large-Scale Multiplexed Screen.** A. Histogram displaying the distribution of the coefficient of variation for the OR library when stimulated with DMSO. B. Histogram displaying the distribution of the coefficient of variation for the OR library against all conditions assayed. C. Dose-response curves for the control odorants included on each 96-well plate assayed. Each color represents a different plate.
Figure 5.10: **Significance and Fold Change of High-Throughput Assay Data.** **A.** The False Discovery Rate (FDR; Benjamini-Hochberg corrected, see Methods) plotted against the fold change for each OR-odorant interaction. The dashed line represents the 1% FDR, the cutoff used to identify positive interactions. **B.** The subset of interactions tested by a follow-up orthogonal luciferase assay (color indicates whether it was recapitulated in the orthogonal system). Of the interactions passing a 1% FDR, 20 of 27 also showed interaction in the orthogonal follow-up assay.
Figure 5.11: Recapitulation of the Screen in a Transient, Orthogonal System. Secondary screen of chemicals in a transient OR reporter activation system with a luciferase reporter gene readout [26]. Each plot shows the behavior of a control cell line expressing the reporter gene but no OR (black line), as well as a cell line expressing a specific OR and reporter gene. In addition, data from the high throughput screen (labeled as Seq) is plotted for reference.
Figure 5.12: **Assay Correspondence with Previously Screened Odorant-Receptor Pairs.**

A. FDR plotted against fold induction for the 540 odorant-OR interactions that were previously tested by Saito et al. Points are colored by the EC$_{50}$ of the interaction in the previous work. Grey points represent interactions not identified in the previous screen. Comparing the results from transient versus integrated luciferase assays revealed that, in some cases, the integrated system required a higher concentration of odorant to achieve significant activation, likely because of the lower DNA copy number of the CRE-driven luciferase and receptor. Since the highest concentration of odorant assayed was 1 mM, low affinity interactions may not have been detectable in this screen.

B. The FDR in the assay related to the EC$_{50}$ of the hit from the previous screen, colored by the fold activation from the multiplexed screen.
Figure 5.13: Location of Odors Tested with Respect to a Learned Chemical Space. Locations of the chemicals tested in this assay in chemical space. The molecular autoencoder was used to generate a 292-dimensional representations of 250,000 randomly sampled molecules from the ChEMBL 23 database (blue) as well as the chemicals tested in our assay (red) projected onto two dimensions with Principal Component Analysis (PCA) [27].
Figure 5.14: Clustering of Odorant Response for Receptors. The locations of any hits (black) with respect to other chemicals tested (grey) for each OR on the PCA on the 292-dimensional latent representation. PC1 explains 34.4% of variance and PC2 explains 14.0% of the variance.
Table 5.1: **Olfactory receptors screened in this study.**

<table>
<thead>
<tr>
<th>Mouse OR Convention</th>
<th>OR Convention</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOR102-1</td>
<td>Olfr1325</td>
</tr>
<tr>
<td>MOR110-1</td>
<td>Olfr812</td>
</tr>
<tr>
<td>MOR112-1</td>
<td>Olfr790</td>
</tr>
<tr>
<td>MOR119-1</td>
<td>Olfr214</td>
</tr>
<tr>
<td>MOR120-1</td>
<td>Olfr459</td>
</tr>
<tr>
<td>MOR13-1</td>
<td>Olfr644</td>
</tr>
<tr>
<td>MOR131-1</td>
<td>Olfr161</td>
</tr>
<tr>
<td>MOR132-1</td>
<td>Olfr24</td>
</tr>
<tr>
<td>MOR133-1</td>
<td>Olfr406</td>
</tr>
<tr>
<td>MOR8-1</td>
<td>Olfr575</td>
</tr>
<tr>
<td>MOR20-1</td>
<td>Olfr613</td>
</tr>
<tr>
<td>MOR203-1</td>
<td>Olfr992</td>
</tr>
<tr>
<td>MOR206-1</td>
<td>Olfr1098</td>
</tr>
<tr>
<td>MOR208-1</td>
<td>Olfr1413</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>Olfr599</td>
</tr>
<tr>
<td>MOR25-1</td>
<td>Olfr554</td>
</tr>
<tr>
<td>MOR30-1</td>
<td>Olfr569</td>
</tr>
<tr>
<td>MOR35-1</td>
<td>Olfr686</td>
</tr>
<tr>
<td>MOR4-1</td>
<td>Olfr620</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Olfr638</td>
</tr>
<tr>
<td>MOR168-1</td>
<td>Olfr916</td>
</tr>
<tr>
<td>MOR169-1</td>
<td>Olfr902</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Olfr895</td>
</tr>
<tr>
<td>MOR18-1</td>
<td>Olfr558</td>
</tr>
<tr>
<td>MOR180-1</td>
<td>Olfr1019</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>Olfr1079</td>
</tr>
<tr>
<td>MOR19-1</td>
<td>Olfr616</td>
</tr>
<tr>
<td>MOR194-1</td>
<td>Olfr1046</td>
</tr>
<tr>
<td>MOR199-1</td>
<td>Olfr1032</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>Olfr62</td>
</tr>
<tr>
<td>MOR134-1</td>
<td>Olfr356</td>
</tr>
<tr>
<td>MOR136-1</td>
<td>Olfr340</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Olfr1352</td>
</tr>
<tr>
<td>MOR142-1</td>
<td>Olfr1356</td>
</tr>
<tr>
<td>MOR144-1</td>
<td>Olfr39</td>
</tr>
<tr>
<td>MOR149-1</td>
<td>Olfr828</td>
</tr>
<tr>
<td>MOR158-1</td>
<td>Olfr362</td>
</tr>
<tr>
<td>MOR165-1</td>
<td>Olfr909</td>
</tr>
<tr>
<td>MOR9-1</td>
<td>Olfr609</td>
</tr>
</tbody>
</table>
Table 5.2: **Odorants screened in this study**

<table>
<thead>
<tr>
<th>Odorant</th>
<th>Odorant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentanoic Acid</td>
<td>Hexanoic Acid</td>
</tr>
<tr>
<td>1-nonanol</td>
<td>Nonanal</td>
</tr>
<tr>
<td>4-hydroxycoumarin</td>
<td>Dimedone</td>
</tr>
<tr>
<td>1-decanol</td>
<td>Decanal</td>
</tr>
<tr>
<td>4-Chromanone</td>
<td>(-)-Menthone</td>
</tr>
<tr>
<td>(+)-2-Heptanol</td>
<td>Citral</td>
</tr>
<tr>
<td>2-Butanone</td>
<td>beta-ionone</td>
</tr>
<tr>
<td>(+)-2-Octanol</td>
<td>Hydroxycitronellal</td>
</tr>
<tr>
<td>2-Hexanone</td>
<td>Pentyl acetate</td>
</tr>
<tr>
<td>(-)-B-Citronellol</td>
<td>Lyral</td>
</tr>
<tr>
<td>2-Heptanone</td>
<td>Allyl heptanoate</td>
</tr>
<tr>
<td>Geraniol</td>
<td>Acetophenone</td>
</tr>
<tr>
<td>3-Heptanone</td>
<td>Amyl hexanoate</td>
</tr>
<tr>
<td>Linalool</td>
<td>Control_1</td>
</tr>
<tr>
<td>2-Octanone</td>
<td>Nonanoic Acid</td>
</tr>
<tr>
<td>1-Undecanol</td>
<td>Control_2</td>
</tr>
<tr>
<td>3-Octanone</td>
<td>Amyl butyrate</td>
</tr>
<tr>
<td>Allyl phenylacetate</td>
<td>Decanoic_Acid</td>
</tr>
<tr>
<td>Propionic Acid</td>
<td>Butyl heptanoate</td>
</tr>
<tr>
<td>Benzene</td>
<td>DMSO</td>
</tr>
<tr>
<td>2_coumaranone</td>
<td>Heptyl isobutyrante</td>
</tr>
<tr>
<td>Benzyl acetate</td>
<td>Prenyl_Acetate</td>
</tr>
<tr>
<td>2-Nonanone</td>
<td>Hexyl acetate</td>
</tr>
<tr>
<td>Phenyl acetate</td>
<td>Vanillic_Acid</td>
</tr>
<tr>
<td>2,3-Hexanedione</td>
<td>Butyl formate</td>
</tr>
<tr>
<td>Octanethiol</td>
<td>a-Amylcinnamaldehyde</td>
</tr>
<tr>
<td>3,4-Hexanedione</td>
<td>Ethyl isobutyrante</td>
</tr>
<tr>
<td>Nonanedioic Acid</td>
<td>Eucalyptol</td>
</tr>
<tr>
<td>(-)-Carvone</td>
<td>1-butanol</td>
</tr>
<tr>
<td>Nonanethiol</td>
<td>Pentyl propionate (Amyl propionate)</td>
</tr>
<tr>
<td>(+)-Dihydrocarvone</td>
<td>Isovaleric Acid</td>
</tr>
<tr>
<td>Butanal</td>
<td>Dihydro Myrcenol</td>
</tr>
<tr>
<td>(+)-Camphor</td>
<td>1-propanol</td>
</tr>
<tr>
<td>Pentanal</td>
<td>Muscenoane</td>
</tr>
<tr>
<td>Dihydrojasmine</td>
<td>1-hexanol</td>
</tr>
<tr>
<td>Hexanal</td>
<td>ethyl maltol</td>
</tr>
<tr>
<td>Benzophenone</td>
<td>1-heptanol</td>
</tr>
<tr>
<td>Heptanal</td>
<td>calone</td>
</tr>
<tr>
<td>(+)-Pulegone</td>
<td>1-octanol</td>
</tr>
<tr>
<td>Octanal</td>
<td>Sandalwood Mysone</td>
</tr>
<tr>
<td>Iso E Super</td>
<td>w-Pentadecalactone</td>
</tr>
<tr>
<td>benzyl benzoate (Pentamethylbenzaldehyde)</td>
<td>Ethyl 2-methylbutyrate</td>
</tr>
<tr>
<td>Olibanum Coeur MD</td>
<td>2-Phenylethanol</td>
</tr>
<tr>
<td>Piperonyl alcohol</td>
<td>trans-2-Dodecalen</td>
</tr>
</tbody>
</table>
Turkish Rose Oil
Piperonyl acetate
Angel Eau de parfum (10 uM)
Tetrahydrofuran
α-Hexylcinnamaldehyde
Tetrahydrofuran
Dior Jadone Eau de parfum
Benzaldehyde dimethyl acetal
Flowerbomb Viktor and Rolf
2-Methyl-1-propanethiol
Chanel No 5
(+)-Dihydrocarveol
Axe
(-)-Dihydrocarveol
Aedione
(+)-Perillaaldehyde
Isobornyl acetate
(-)-Perillaaldehyde
α-Amylcinnamaldehyde dimethyl acetal
Benzyl salicylate
p-Tolyl isobutyrate
(+)-Limonene oxide,mixture of cis and trans
α-Tolyl isobutyrate
(-)-Limonene oxide,mixture of cis and trans
p-Tolyl phenylacetate
(R)-(+-)-Limonene
2-Methoxy-3-Methyl-pyrazine
(-)-Camphene
2-Methoxypyrazine
(+)-Camphene
Methyl salicylate
2,3-Diethyl-5-methylpyrazine
Anethole
Ethyl disulfide
Myrcene
Methyl disulfide
(±)-2-Butanol
trans-2-Methyl-2-butenal (2MB)
2-Isopropyl-3-methoxypyrazine
diacetyl
2-sec-Butyl-3-methoxypyrazine
galaxolide
cis-6-Nonenal
isobutyraldehyde
Cinnamaldehyde
Ethyl 2-methylpentanoate
beta-Damascone
2-Phenethyl acetate
Cedryl acetate
Piperonal
1-Octen-3-one
Pyrazine
2-Bromohexanoic acid
Sassafras oil
6-Bromohexanoic acid
thymol
2-Bromoocanonic acid
Triethylamine
Furfuryl methyl disulfide
L-Turpentine
Ethyl isovalerate
Anisaldehyde
Bis(2-methyl-3-furlyl)disulphide)
|D|ethyl sulfide
Dimethyl trisulfide
Eugenol
trans-2,cis-6-Nonadienal
Eugenol methyl ether
trans-2-Nonenal
4-Ethylphenol
Cinnamyl alcohol
Ethy| vanillin
n-Decyl acetate
Vanillin
Dimethyl anthranilate
2-Ethylphenol
t|s-2-Undecenal
Guaiacol
Neryl isobutyrate
2-bromophenol
cis-4-Decen|l
Benzaldehyde
Octyl formate
2,3-Diethylpyrazine
p-cymene
2-Methylbutyric acid
helional
Cyclobutanecarboxylic acid
1,9-nonanediol
Isopentylamine (1-Amino-3-methylbutane, Isoamylamine)
octanedioic acid (suberic acid)
Quinoline (1-Benzazine; 2,3-Benzopyridine)
decanedioic acid (sebacic acid)
Farnesene
Table 5.3: Odorant-receptor pairs called as hits.

<table>
<thead>
<tr>
<th>OR</th>
<th>Odorant</th>
<th>Minimum Activating Concentration (µM)</th>
<th>Previously Orphan Receptor?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOR102-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR112-1</td>
<td>Benzaldehyde</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR112-1</td>
<td>galaxolide</td>
<td>100</td>
<td>YES</td>
</tr>
<tr>
<td>MOR119-1</td>
<td>Axe (10 uM)</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR119-1</td>
<td>Furfuryl methyl disulfide</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR119-1</td>
<td>n-Decyl acetate</td>
<td>100</td>
<td>YES</td>
</tr>
<tr>
<td>MOR120-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR120-1</td>
<td>Lyral</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR120-1</td>
<td>Nonanethiol</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR13-1</td>
<td>Benzaldehyde</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR13-1</td>
<td>Cyclobutanecarboxylic acid</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR13-1</td>
<td>Pentanoic Acid</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR13-1</td>
<td>trans-2-Methyl-2-butenal (2MB)</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR131-1</td>
<td>(-)-Perillaldehyde</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR131-1</td>
<td>1-hexanol</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR131-1</td>
<td>3,4-Hexanedione</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR131-1</td>
<td>galaxolide</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR132-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR133-1</td>
<td>3-Octanone</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR134-1</td>
<td>Chanel No 5 (10 uM)</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR136-1</td>
<td>(+)-Dihydrocarveol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR136-1</td>
<td>(+)-Camphor</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR136-1</td>
<td>(+)-Dihydrocarveol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR136-1</td>
<td>2-Ethylphenol</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR136-1</td>
<td>Olibanum Coeur MD</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>(-)-Dihydrocarveol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>(+)-Dihydrocarvone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>(+)-Pulegone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>2-sec-Butyl-3-methoxy pyrazine</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>4-Chromanone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>beta-ionone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Butanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Dihydrojasmine</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Dimethyl anthranilate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Eugenol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Eugenol methyl ether</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>Code</td>
<td>Name</td>
<td>Concentration</td>
<td>Result</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------</td>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>helional</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Neryl isobutyrate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR139-1</td>
<td>Quinoline (1-Benzazine; 2,3-Benzopyridine)</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR142-1</td>
<td>Bis(2-methyl-3-furyl)disulphide)</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR142-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR158-1</td>
<td>Iso E Super</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR165-1</td>
<td>decanedioic acid (sebacic acid)</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR165-1</td>
<td>Octyl formate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>2-Bromohexanoic acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>2-Phenethyl acetate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>4-Chromanone</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>4-Ethylphenol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Anisaldehyde</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Benzyl acetate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>benzyl benzoate (Pentamethylbenzaldehyde)</td>
<td>10</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Chanel No 5 (10 uM)</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Cinnamyl alcohol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Dimethyl anthranilate</td>
<td>10</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>ethyl maltol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Eugenol methyl ether</td>
<td>10</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>helional</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Piperonal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Piperonyl acetate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Quinoline (1-Benzazine; 2,3-Benzopyridine)</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR170-1</td>
<td>Vanillin</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR180-1</td>
<td>a-Amylicinnamaldehyde dimethyl acetal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR180-1</td>
<td>Axe (10 uM)</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>4-Chromanone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>benzyl benzoate (Pentamethylbenzaldehyde)</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>beta-Damascone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>beta-ionone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>Eugenol methyl ether</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR189-1</td>
<td>Quinoline (1-Benzazine; 2,3-Benzopyridine)</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR19-1</td>
<td>Benzyl salicylate</td>
<td>10</td>
<td>YES</td>
</tr>
<tr>
<td>MOR19-1</td>
<td>Methyl salicylate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR199-1</td>
<td>ethyl maltol</td>
<td>100</td>
<td>YES</td>
</tr>
<tr>
<td>MOR203-1</td>
<td>helional</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR203-1</td>
<td>Piperonyl acetate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR208-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>YES</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>2-Bromooctanoic acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>6-Bromohexanoic acid</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>Heptanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>Hexanoic Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>Nonanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>Nonanoic Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>Code</td>
<td>Substance</td>
<td>Concentration</td>
<td>No.</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------</td>
<td>---------------</td>
<td>-----</td>
</tr>
<tr>
<td>MOR23-1</td>
<td>Octanal</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR25-1</td>
<td>(-)-Carvone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR25-1</td>
<td>Decanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR25-1</td>
<td>Decanoic-Acid</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR25-1</td>
<td>Nonanoic Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR30-1</td>
<td>Cedryl acetate</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR30-1</td>
<td>Decanal</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR30-1</td>
<td>Decanoic-Acid</td>
<td>10</td>
<td>NO</td>
</tr>
<tr>
<td>MOR30-1</td>
<td>Nonanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR30-1</td>
<td>Nonanoic Acid</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR4-1</td>
<td>Hexanoic Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR4-1</td>
<td>Pentanoic Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>2-Bromohexanoic acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>2-Bromooctanoic acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>6-Bromohexanoic acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>cis-4-Decenal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>cis-6-Nonenal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Decanoic-Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Hexanoic Acid</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Nonanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Nonanoic Acid</td>
<td>100</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Octanal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR5-1</td>
<td>Olibanum Coeur MD</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>2-coumaranone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>Benzaldehyde</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>Benzophenone</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>ethyl maltol</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>Piperonal</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR258-5</td>
<td>Quinoline (1-Benzazine; 2,3-Benzopyridine)</td>
<td>1000</td>
<td>NO</td>
</tr>
<tr>
<td>MOR9-1</td>
<td>galaxolide</td>
<td>1000</td>
<td>NO</td>
</tr>
</tbody>
</table>
Table 5.4: **Primers used in this study.**

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OL001</td>
<td>CCCTTTAATCAGATCGTCG</td>
<td>Gene Specific RT, Reporter Gene, for Q-RTPCR</td>
</tr>
<tr>
<td>OL002</td>
<td>CTGCCTGCCTTGACCAACCTCC</td>
<td>Gene Specific RT, GAPDH</td>
</tr>
<tr>
<td>OL003</td>
<td>AAGTGCAATCCTGCGCCTTTAACGAGATCGTCG</td>
<td>Gene Specific RT, Reporter Gene, for RNA-seq, Also NGS Read1 Primer</td>
</tr>
<tr>
<td>OL004F</td>
<td>CGCCGAAATGAAACCACCTA</td>
<td>Pilot-Scale RNA-seq Round 1 Library Prep Amplification</td>
</tr>
<tr>
<td>OL004R</td>
<td>AAGTGCCCTCCTGCCCTTTA</td>
<td>Pilot-Scale RNA-seq Round 1 Library Prep Amplification</td>
</tr>
<tr>
<td>OL005F</td>
<td>CAACGAGAAGGGCAGACGATCAG-NNNNNNNNAAGTGAAACCACCTA</td>
<td>P7+i7index+primer for RNAseq library amplification</td>
</tr>
<tr>
<td>OL005R</td>
<td>AATGATACGGCGACCACGAGATCTAACAAGTCCCTTACCTGCG</td>
<td>P5+Read1+primer for pilot-scale RNAseq library amplification</td>
</tr>
<tr>
<td>OL006</td>
<td>CGGGTTTCTTCGCTTCTGAGTGTGCTGTCTCTCTCT</td>
<td>i7 index read primer, pilot-scale experiment</td>
</tr>
<tr>
<td>OL007F</td>
<td>gggaataACGCGTNNNNNNNNNNNNNNNN- CGACGCATCTCGATTAAGGG</td>
<td>Amplification of fragment containing barcode to be cloned into reporter plasmid</td>
</tr>
<tr>
<td>OL007R</td>
<td>ggaaagaACGGGTATctagctggactaactat</td>
<td>Amplification of fragment containing barcode to be cloned into reporter plasmid</td>
</tr>
<tr>
<td>OL008F</td>
<td>tgtctctggccctgtgcctagctggctgct- CATATGATGCAAGCAAGACC</td>
<td>Amplification of fragment containing the OR to be cloned into the reporter plasmid</td>
</tr>
<tr>
<td>OL008R</td>
<td>AGTCGCGCCCTGAGTGACTTTCCACCTCGACTTTTAC- CATGTCGCTGGAA</td>
<td>Amplification of fragment containing the OR to be cloned into the reporter plasmid</td>
</tr>
<tr>
<td>OL009</td>
<td>CTTCTCCTGATCCCTTCTC</td>
<td>Sequencing and linking barcodes/ORs in the reporter vector</td>
</tr>
<tr>
<td>OL010</td>
<td>CCTGCGGCTCTTATCATGTC</td>
<td>Sequencing and linking barcodes/ORs in the reporter vector</td>
</tr>
<tr>
<td>OL011</td>
<td>TACAGGCAGGATGGAGCAG</td>
<td>Sequencing and linking barcodes/ORs in the reporter vector</td>
</tr>
<tr>
<td>OL012F</td>
<td>AAGTGAAACCCACCTACAAGG</td>
<td>QPCR of the transposon for copy number analysis</td>
</tr>
<tr>
<td>OL012R</td>
<td>CCGCTTACGATGACGGTCG</td>
<td>QPCR of the transposon for copy number analysis</td>
</tr>
<tr>
<td>OL013</td>
<td>AATGATACGGGCAGAACGAGACGATCTACAACNNNNNNAATGCGCTTCCCTGCCCTTTAA</td>
<td>P5+i5+Read1+primer, for large-scale library amplification</td>
</tr>
<tr>
<td>LP001F</td>
<td>TGGGCAGTTCCAGCTTATAGTC</td>
<td>Genomic Amplification of the H11 locus with the landing pad</td>
</tr>
<tr>
<td>LP001R</td>
<td>GGGCGTACTTGGCATGACGACGAG</td>
<td>Genomic Amplification of the H11 locus with the landing pad</td>
</tr>
</tbody>
</table>

**References**


Chapter 6

Deep Mutational Scan of the β2-adrenergic Receptor

6.1 Abstract

The G protein-coupled receptors (GPCRs) are a uniquely important protein family in human physiology. Their dynamic structure is critical to their myriad functions, but makes biophysical characterization challenging. We developed a platform to characterize large libraries of GPCRs in human cell lines, and use it to functionally assess all possible single amino acid substitutions to the β2-adrenergic receptor across several agonist concentrations. Cumulatively, we find that residues with similar mutational profiles reflect their structural and functional organization, and we identify both known and novel residues critical for function. In addition, we describe a previously uncharacterized, conserved extracellular “structural latch” maintained in both the inactive and active state of the receptor. Our approach enables mutational scanning for most GPCRs and other human proteins where function can be linked to a genetic reporter.

6.2 Introduction

G protein-coupled receptors (GPCRs) are central mediators of mammalian cells’ ability to sense and respond to their environment. In humans, the ∼800 GPCRs respond to a wide range of chemical stimuli such as hormones, odors, natural products, and drugs by modulating a set of prototypical pathways that affect cell physiology. Their central role in altering relevant cell states makes them ideal targets for therapeutic intervention, with ∼34% of all U.S. Food and Drug Administration (FDA)-approved drugs targeting the GPCR superfamily [1].

Understanding GPCR signal transduction is difficult for several reasons. First, GPCRs exist in a large conformational landscape, making traditional biophysical characterizations difficult. Consequently, most GPCR structures are truncated, non-native, or artificially stabilized. Even when structures exist, most are of inactive states, and only ∼18 receptors have active state structures available [2]. Second, GPCR dynamics are critical for their function. Static structures from both X-ray crystallography and Cryo Electron Microscopy do not directly probe receptor dynamics [3]. Towards this end, tools such as spectroscopy and computational simulation have aided our interpretation [4].

Alternatively, mutagenesis has long been a foundation of protein biochemistry and, when coupled with a phenotypic screen, provides a robust approach to directly investigate GPCR signaling and function [5, 6, 7]. Historically, technical constraints restricted the number of mutations that could be generated and characterized. Recent advancements in DNA synthesis, genome editing, and next-generation sequencing (NGS) have enabled Deep Mutational Scanning (DMS), a method to functionally assay all possible missense variants of a given protein [8, 9].

However, deep mutational scans often suffer from not being generalizable across different protein targets or screen a phenotype that is not informative of protein function. GPCRs are particularly susceptible to this problem because they bind a variety of signaling effectors that activate distinct pathways [10, 11, 12]. Here we report a novel platform to globally dissect the functional consequences of missense variation in GPCRs expressed in human cell lines. By constructing a method to assay genetic reporters in multiplex, we are able to link GPCR activation to cAMP production, a direct output of G protein signaling. Furthermore, genetic reporters are modular and can be easily exchanged to suit DMS targets of disparate function [13].
Figure 6.1: Platform for Deep Mutational Scanning of GPCRs and Variant-Activity Landscape.

A. Graphical Display of Multiplexed GPCR Activity Assay. ADRB2 variants with their barcoded genetic reporter are integrated into a defined genomic locus such that one variant is integrated per cell. Upon isoproterenol agonization, G protein signaling induces transcription of the cAMP-responsive genetic reporter and the barcode. The barcode sequence in the 3’ UTR of the reporter encodes the identity of the receptor within the same cell.

B. Overview of workflow for Multiplexed GPCR Activity Assay. The variant library is generated, barcoded, and cloned into a vector with a genetic reporter. The library is then integrated into HEK293T cells and agonized with various concentrations of isoproterenol. After stimulation, mutant activity is determined by measuring the relative abundance of each variant’s barcoded cAMP-responsive genetic reporter transcripts with RNA-seq.

C. Top: Secondary structure diagram represents the N and C termini in black, the transmembrane domains as blocks, and the intra- and extracellular domains in blue and green respectively. The EVmutation track displays average effect of every mutation as predicted by EVmutation. The Conservation track displays the sequence conservation of each residue. The shaded guides represent positions of the protein in the transmembrane domain. Bottom: The heatmap representation of the activity of every missense mutation and frameshift at each agonist condition. Cells are colored by the relative activity to the mean frameshift mutation.
6.3 Results

To comprehensively probe the structure-function relationship of the β2-adrenergic receptor (β2AR), we developed technology to generate and simultaneously profile the receptor’s 7,828 possible missense variants for differences in functional activity in HEK293T cells (Fig. 6.1a, b, Supplementary Fig. 6.5, 6.6). The activity of each variant is linked to expression of a cAMP responsive genetic reporter enabling an accurate depiction of receptor signaling capability. Variant identity is encoded in a short barcode sequence appended to the 3’ UTR of the reporter gene. Using RNA-seq we read the activity of the entire variant library in multiplex. We screened the mutant library under four conditions of the β2AR agonist isoproterenol: vehicle control, an empirically determined EC$_{50}$, EC$_{100}$, and beyond saturation of the WT receptor, and report measurements for 99.6% of possible missense variants.

To validate our assay, we recorded the activity of 6 mutants that are stably and individually expressed at single copy, the same configuration as the multiplexed assay, with a luciferase reporter gene [14, 15, 16]. These measurements largely agree with the results of our multiplexed assay (Supplementary Fig. 6.7). Our variant generation approach, microarray-based oligo synthesis, often produces single base deletions that introduced a plethora of frameshift mutations into our library [17]. As expected, frameshift mutations have consistently lower activity than missense mutations. Furthermore, frameshifts occurring in the C-terminus have a significantly diminished reduction of activity (Supplementary Fig. 6.7).

The heatmap representation of mutant activity reveals the helices are more sensitive to substitution than the termini or loops, and this effect becomes more pronounced at higher agonist concentration (Fig. 6.1c). In general, the transmembrane domain is especially sensitive to proline substitution (Supplementary Fig. 6.7). In lieu of large-scale functional data, two indications of the effect a potential mutation will have on protein function are sequence conservation and co-variation. While conservation is highly correlated ($\rho = -0.747$) with mutational tolerance, the aggregate fitness for all substitutions at a given position, it does not apply to specific substitutions([18, 19]). EVmutation, a a predictor of mutational effects from sequence covariation, correlates well ($\rho = 0.521$) with our variant-level data (Fig. 6.1c) [20]. Of note, correlation between our data and both predictors increases with agonist concentration up to EC$_{100}$, suggesting our phenotypic screen is evolutionarily relevant (Supplementary Fig. 6.8).
Figure 6.2: Unsupervised Learning Elucidates Broad Structural Features and Critical Residues of the $\beta_2$AR. A. We averaged amino acid substitutions into classes based on their physicochemical properties. We then used Uniform Manifold Approximation and Projection (UMAP) to learn a 2D representation of every residue’s response to these classes of substitutions across all agonist conditions. Each residue is assigned into one of six clusters using HDBSCAN (see Supplementary Fig. 6.10). B. The class averages of each of these cluster reveals their distinct responses to mutation. The upper dashed line represents the mean of the Cluster 6 and the lower dashed line represents the mean activity of frameshifted mutants. C. A 2D snake plot representation of $\beta_2$AR secondary structure with each residue colored by cluster.
Our data spans thousands of mutations of varying severity across multiple agonist conditions. We hypothesized unsupervised learning methods could reveal hidden regularities within groups' of residues response to mutation. We applied Uniform Manifold Approximation and Projection (UMAP) to learn multiple different lower-dimensional representations of our data and clustered the output with HDBSCAN [21, 22] (Supplementary Fig. 6.9). We found residues consistently separated into 6 clusters that exhibit distinct responses to mutation (Fig. 6.2a, b). Clusters 1 and 2 are globally intolerant to all substitutions, whereas Cluster 3 is affected by proline and hydrophilic substitutions. Cluster 4 is particularly inhibited by negatively charged substitutions, while Cluster 5 is uniquely intolerant to proline, and Cluster 6 is unaffected by substitution.

Mapping these clusters onto a 2D snake plot representation shows Clusters 1-5 primarily comprise the transmembrane helices, while Cluster 6 mainly resides in ICL3 and the termini (Fig. 6.2c) [2]. These flexible regions are often truncated before crystal structure determination to minimize conformational variability [23]. Surprisingly, a number of residues from Cluster 5 also map there. Given that residues in Cluster 5 are uniquely intolerant to proline substitutions, we hypothesize these regions may become structured in one or more receptor conformations.

Next, we projected the clusters onto the hydroxybenzyl isoproterenol bound structure (Supplementary Fig. 6.10; PDB: 4LDL). The globally intolerant Clusters 1 and 2 segregate to core of the protein, while Cluster 3, intolerant to polar residue substitution, is enriched in the lipid-facing portion. This suggested that differential response to hydrophobic and charged substitutions could correlate with side chain orientation within the transmembrane domain. Indeed, residues that are uniquely charge sensitive are significantly more lipid-facing than those that are sensitive to both hydrophobic and charged mutations (Fig. 6.3a; Supplementary Fig. 6.10) [24]. Taken together, DMS and unsupervised learning methods provide a way to determine patterns of mutational constraint between cohorts of residues. From this, we can learn structural features, such as side chain orientation and secondary structure, even without a crystal structure.

Decades of research have revealed many GPCRs couple ligand binding to G protein activation through a series of conserved motifs [25]. The globally intolerant UMAP clusters (1 and 2) highlight many residues from these motifs and suggest novel residues for further investigation (Fig. 6.2c, Fig. 6.3b). We can further resolve the significance of individual residues within these motifs by ranking
Figure 6.3: Positional Constraint Restates the Significance of Known Structural Motifs and Suggests Novel Residues for Investigation. A. Residues within the transmembrane domain colored by their tolerance to particular substitutions. Teal residues are intolerant to both hydrophobic and charged amino acids (globally intolerant), and brown residues are tolerant to hydrophobic amino acids but intolerant to charged amino acids. These charge sensitive positions tend to point into the membrane, while the globally intolerant positions face into the core of the protein. B. The crystal structure of the hydroxybenzyl isoproterenol-activated state of the $\beta_2$AR (PDB: 4LDL) with residues from the mutationally intolerant Clusters 1 and 2 highlighted in magenta. C-F Selected vignettes of residues from the mutationally intolerant UMAP clusters. C. W286 of the CWxP motif and the neighboring G315 are positioned in close proximity. Substitutions at G315 are likely to cause a steric clash with W286 (PDB: 4LDL). D. An inactive state water-mediated hydrogen bond network (red) associates N51 and Y326 (PDB: 2RH1). Disruption of this network may destabilize the receptor. E. The ligand-bound orthosteric site surface colored by mutational tolerance the unique sensitivity of the receptor-ligand contacts and displays the assay’s discriminatory power between agonists (PDB: 4LDL). F. Mutationally intolerant $\beta_2$AR residues at the G protein interface from the $\beta_2$AR-Gs complex crystal structure (PDB: 3SN6), V222, I135, and Q229.
the mutational tolerance of every position at EC\textsubscript{100} (Supplementary Fig. 6.11). In fact, 8 of the 10 most intolerant positions belong to known structural motifs. However, the most mutationally intolerant residue is the uncharacterized G315. In the active state, G315’s alpha carbon points directly at W286 of the CWxP motif, and any substitution at G315 will likely clash with W286 (Fig. 6.3c). Additionally, W286 is the second most intolerant position, reinforcing its essentiality for receptor function. Recent simulations suggest networks of water-mediated hydrogen bonds play a critical role in GPCR function [26]. Y326 of the NPxxY motif, the 5th most intolerant position, switches between two such networks during the active state transition. In the inactive state, Y326 networks with N51 and D79, two of the top 20 most intolerant positions (Fig. 6.3d).

Next, we wondered if residues in the orthosteric site that directly contact isoproterenol would respond differently to mutation than residues that contact other agonists. Using the crystal structure of the β\textsubscript{2}AR bound to hydroxybenzyl isoproterenol, we find that positions responsible for binding the derivatized hydroxybenzyl tail are significantly less sensitive to mutation than residues that contact the catecholamine head common to both molecules at EC\textsubscript{100} (p = 0.0162; Fig. 6.3e, Supplementary Fig. 6.11). Given this discrimination, we believe DMS can be a powerful tool for mapping ligand-receptor contacts.

The numerous β\textsubscript{2}AR crystal structures in various complexes and conformational states enable us to evaluate the functional consequences of predicted intermolecular interactions. For example, cholesterol is an important modulator of β\textsubscript{2}AR and the timolol-bound inactive state structure elucidated the coarse location of a cholesterol binding site (PDB: 3D4S; [27]). As previously predicted, W158\textsubscript{4.50x50} is the most mutationally intolerant of the residues (Supplementary Fig. 6.11). Furthermore, the relative contribution of most individual residues for stabilizing the G\textsubscript{s}-β\textsubscript{2}AR interface is unknown [7, 28, 29, 30, 31, 32, 33, 34]. Interestingly, most residues are tolerant to substitution, but three of the most intolerant positions are I135, V222, and Q229 respectively (Fig. 6.3f). Q229 appears to coordinate polar interactions between D381 and R385 of the α5 helix of G\textsubscript{s}, whereas V222 and I135 form a hydrophobic pocket on the receptor surface.

The mutationally intolerant UMAP clusters also highlight residues from tolerant regions of the structure. For instance, the uncharacterized W99\textsubscript{23.50x50}, of ECL1, is proximal to the disulfide bond C106-C191, which is important for stabilization of the high-affinity receptor state (Fig. 6.4a)
Aromatic residues are known to facilitate disulfide bond formation, but our data suggest only tryptophan is tolerated [37]. We hypothesize W99's indole group hydrogen bonds with the backbone carbonyl of neighboring G102, positioning W99 towards the disulfide bond. Other aromatic residues are unable to hydrogen bond and are less likely to be positioned properly.

This observation lead us to inquire whether the structural latching between W99 and the disulfide bond is specific to human β2AR or generic to all class A GPCRs. Comparing over 25 high-resolution structures of class A GPCRs from five functionally different sub-families and six different species revealed that position the trp and disulfide bond consistently contact each other (Fig. 6.4a). Three exceptions to this trend are the human S1P1 sphingosine receptor, A2A adenosine receptor and bovine rhodopsin. Expectedly, the S1P1 sphingosine receptor lacks the conserved disulfide bond and has a relatively long ECL1, uncharacteristic properties of class A GPCRs. For both the human A2A adenosine receptor and the bovine rhodopsin, the trp is substituted by another aromatic residue, phenylalanine. In addition to the trp-disulfide bond non-covalent interaction, we also observe the backbone geometry of ECL1 is highly similar among the class A GPCRs (data not shown). Based on the evolutionary coupling analysis and structural comparison of class A GPCRs, we find the trp in ECL1 together with the disulfide bond connecting ECL2 and the extracellular end of TM3 form a conserved “extracellular structural latch” that is maintained consistently in different GPCRs spanning diverse molecular functions and phylogenetic origins.

Next, we were interested in understanding the dynamics of the structural latch. While the overall RMSD between the inactive and active states for human β2AR (PDB: 2RH1 vs. 3P0G) and M2 muscarinic receptor (PDB: 3UON vs. 4MQS) are 1.32 and 1.78 respectively, the RMSD of the latch is nearly identical in both receptors (Fig. 6.4b). Additionally, we examined 100 frames each sampled from deactivating simulations of the human β2AR to investigate whether the latch is maintained during the conformational transition between the two states. The residues forming the structural latch in human β2AR (W99, C106, and C191) are locked in their chi1 angles, as seen in the inactive (2RH1) or active (3P0G) state structures, with standard deviations around 7-8 degrees each (data not shown). The low variability of the side chain geometry of these residues during the conformational transition between the active and inactive states asserts that the extracellular structural latch is rigid and indeed conformation independent. Furthermore, the majority of 15 other
residues that undergo very low chi1 rotamer changes are in proximity to the extracellular structural latch (data not shown). This suggests that the extracellular structural latch is a part of a larger rigid plug present at the interface of the transmembrane and extracellular region, which could be important for the structural integrity and thereby function of the receptor.

Conformation dictates GPCR function, therefore identifying individual mutations that stabilize particular states provides insight into the biochemistry of receptor activation. We filtered for mutations that lead to greater than WT activity without agonist stimulation to search for variants with increased basal activation rates or expression levels. Mapping these mutations onto the 2D snake plot, reveals they are not uniformly distributed throughout the protein, rather they are enriched in the termini, TM1, TM5, ICL3, and Helix 8 (Supplementary Fig. 6.12). Concentration at the N- and C-termini is unsurprising, as these regions have known involvement in surface expression [38]. Similarly, the enrichment of mutants in ICL3 reiterates its role in G protein binding [39, 40, 41]. Of note, a group of mutations in TM5 face TM6, which undergoes a large conformational change during receptor activation (Supplementary Fig. 6.12). The activating mutant E62R of ICL1 is also salient as R63 and L64 are both highly intolerant, suggesting an underappreciated role of ICL1 in receptor activation (Supplementary Fig. 6.11). Lastly, understanding how human variation affects β2AR signaling and GPCRs in general is critical. We find approximately 60% of reported variants in the genome Aggregation Database (gnomAD) result in a loss of function of the β2AR at EC_{100} (Supplementary Fig. 6.12).

6.4 Discussion

Our findings showcase a new generalizable approach for deep mutational scanning of human protein targets with transcriptional reporters. Genetic reporters enable precise measurements of gene-specific phenotypes that can be widely applied across the proteome. We show comprehensive mutagenesis can allude to the structural organization of the protein and the local environment of individual residues. These results suggest deep mutational scanning can work in concert with other techniques (e.g. X-ray crystallography and Cryo-EM) to augment our understanding of GPCR structure. Moreover, we identify key residues for β2AR function including uncharacterized positions that inform about
Figure 6.4: Conserved Extracellular Tryptophan-Disulphide ‘Structural Latch’ in Class A GPCRs is Rigid and Conformation-Independent.  

**A.** W99 is mutationally intolerant and appears to be contacting the C106-C191 disulfide bond of the ECL1. A structural comparison of Class A GPCR structures reveals the Trp-disulfide bond contact is conserved in 22 of the 25 receptors. **B.** The Trp-disulfide bond contact is maintained in both the inactive and active state structures for the β2AR and M2 muscarinic receptor.
Looking forward, our method is well poised to investigate many outstanding questions in GPCR biology. First, individual GPCRs are known to signal through multiple pathways: both through interactions with multiple G protein alpha subunits as well as beta-arrestin signaling [42]. Through systematic mutational interrogation across these pathways’ genetic reporters, we can understand the different mechanisms that underpin their signal transduction and the molecular basis for biased signaling [43]. Second, GPCRs are often targeted by synthetic molecules with either unknown or predicted binding sites. We find ligands imprint a mutational signature on their receptor contacts and each ligand’s mutational profile can reveal their molecular contacts either in the case of orthosteric ligands or allosteric modulators. Lastly, the identification of mutations that can stabilize specific conformations or increase receptor expression can aid in GPCR structural determination [44, 45].

6.5 Materials and Methods

Experimental Methods

Endogenous ADRB2 Deletion using CRISPR/Cas9

Cas9 and sgRNAs targeting the sole exon of ADRB2 were cloned and transfected into HEK293T cells according to the protocol outlined in Ran et al. [46]. After transfection, cells were seeded in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 3 days and expanded to 24-well plates after 7 days. Clones were screened for ADRB2 deletion by screening them for the inability to endogenously activate a cAMP genetic reporter when stimulated with the ADRB2 ligand isoproterenol. Clones were seeded side by side wild type HEK293T cells at a density of 7,333 cells/well in a pol 96-well plate. 24 hours later, cells were transfected with 10 ng/well of a plasmid encoding luciferase driven by a cyclic AMP response element and 5 ng/well of a plasmid encoding Renilla luciferase with lipofectamine 2000. 24 hours later, media was removed and cells were stimulated with 25µL of a range of 0 to 10µM isoproterenol (Sigma-Aldrich) in CD293 (Thermo Fisher Scientific) for 4 hours. After agonist stimulation, the Dual-Glo Luciferase Assay kit was administered according to the manufacturer’s instructions. Luminescence was measured using the M1000 plate reader (Tecan). All luminescence values were normalized to Renilla luciferase activity.
to control for transfection efficiency in a given well. Data were analyzed with Microsoft Excel and R.

**Landing Pad Genome Editing**

The H11 locus was edited using TALEN plasmids received from Addgene (#51554, #51555). HEK293T cells were seeded at a density of 75k cells in a 24-well plate. 24 hours after seeding cells were transfected with 50 ng LT plasmid, 50 ng RT plasmid, and 400 ng of the Linearized Landing Pad using Lipofectamine 2000. 2 days after transfection, cells were expanded to a 6-well plate and one day after expansion 500 µg/ml hygromycin B (Thermo Fisher Scientific) was added to the media. Cells were grown under selection for 10 days. After selection, cells were seeded in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 3 days and expanded to 24-well plates after 7 days. gDNA was purified using the Quick-gDNA Miniprep kit (Zymo Research) from the colonies and PCR was performed with Hifi Master Mix to ensure the landing pad was present at the correct locus (LP001F and R). The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 35 cycles of 98°C for 20 seconds, 63°C for 15 seconds, and 72°C for 40 seconds, followed by an extension of 72°C for 2 minutes. To ensure a single landing pad was present per cell, HEK293T cell lines with both singly and doubly-integrated landing pads along with untransduced (WT) HEK293T cells were plated at 4 x 10^5 cells per 6-well. All landing pad cells were transfected the next day with 1.094µg of both an attB-containing eGFP and mCherry donor plasmid and 0.3125µg of the BxB1 expression vector or a pUC19 control. Two singly-integrated landing pad cell samples were also transfected with 2.1875µg of either an attB-containing eGFP and mCherry donor plasmid with 0.3125µg of the BxB1 expression vector. Cells were transfected at a 1:1.5 DNA:Lipofectamine ratio with Lipofectamine 3000. 2 days later cells were passaged at 1:10 and were analyzed using flow cytometry 10 days later after 4 total passages. Samples were flown using the LSRII at the UCLA Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research Flow Cytometry Core. Cytometer settings were adjusted to the settings: FSC – 183 V, SSC – 227 V, PE-Texas Red – 336 V, Alexa Fluor 488 – 275 V.
Individual Donor Bxb1 Recombinase Plasmid Integrations

HEK293T derived cells engineered to contain the Bxb1 Recombinase site at the H11 locus were seeded at a density of 350k cells in a 6-well plate (Corning). 24 hours after seeding cells were transfected with 2 µg Donor plasmid and 500 ng plasmid encoding the Bxb1 recombinase using Lipofectamine 3000 (Thermo Fisher Scientific). 3 days after transfection cells were expanded to a T-75 flask (Corning) and 8 µg/ml blasticidin (Thermo Fisher Scientific) was added one day after expansion. Cells were kept under selection 7-10 days and passaged twice 1:10 to ensure removal of transient plasmid DNA.

Ligand-Receptor Activation Luciferase Assay for Genomically Integrated Receptor-Reporter Constructs

HEK293T and HEK293T derived cells integrated with the combined receptor/reporter plasmids were plated at a density of 7333 cells/well in 100 µL DMEM in poly-D-lysine coated 96-well plates. 48 hours later, media was removed and cells were stimulated with 25µL of a range of isoproterenol concentrations in CD293 for 4 hours. After agonist stimulation, the Dual-Glo Luciferase Assay kit was administered according to the manufacturer’s instructions. Luminescence was measured using the M1000 plate reader. Data were analyzed with Microsoft Excel and R.

Ligand-Receptor Activation q-RT PCR Assay for Genomically Integrated Receptor-Reporter Constructs

HEK293T and HEK293T derived cells integrated with the combined receptor/reporter plasmids were plated at a density of 200k cells/well in 2 mL DMEM in 6-well plates. 48 hours after seeding, media was removed and cells were induced with various concentrations of either forskolin (Sigma-Aldrich) or isoproterenol diluted in 1 ml of OptiMEM (Thermo Fisher) per plate for 3 hours. After stimulation, media was removed and 600 µL of RLT buffer (Qiagen) was added to each well to lyse cells. Lysate from each sample were homogenized with the QIAshredder kit (Qiagen) and total RNA was prepared from each sample using the RNeasy Mini Kit with the optional on-column DNase step (Qiagen). 5 µg of total RNA per sample was reverse transcribed with Superscript III (Thermo-Fisher) using a gene specific primer for the reporter gene and GAPDH (Supplementary Table X) according the
manufacturer’s protocol. The reaction conditions are as follows: Annealing: [65°C for 5 min, 0°C for 1 min] Extension: [52°C for 60 min, 70°C for 15 min]. 10% of the RT reaction was amplified in triplicate for both genes, the reporter gene and GAPDH (Supplementary Table X), using the SYBR FAST qPCR Master mix (Kapa Biosystems) with a CFX Connect Thermocycler (Biorad). The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. Reporter gene expression was normalized to GAPDH expression for each sample. Data were analyzed with Microsoft Excel and R.

**Variant Library Generation and Cloning**

The ADRB2 missense variant library was created by splitting the protein coding sequence into 8 distinct segments (∼52 a.a. each) and synthesizing all single amino acid substitutions for each segment separately as an oligonucleotide library (Agilent). 500 pg of the oligonucleotide library was amplified with biotinylated primers unique for each segment (Supplementary Table X) with the Real-Time Library Amplification Kit (Kapa Biosystems) on a CFX Connect Thermocycler (Biorad). The reaction and cycling conditions are as follows: 98°C for 45 seconds, X cycles of 98°C for 15 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, followed by an extension of 72°C for 1 minute. The number of cycles for the amplification was determined to ensure the amplification was in the exponential phase at least two cycles before the amplification reached saturation. The PCR products were cleaned up with the DNA Clean and Concentrator Kit (Zymo Research) and digested with restriction enzymes BamHI and BspQI, BbsI and BspQI, or BbsI and NheI (New England Biolabs). Digestions were cleaned up with the DNA Clean and Concentrator Kit and digested ends of the amplified library were removed by performing a streptavidin bead cleanup with the Dynabeads M-280 and the DynaMag (Thermo Fisher). Each library segment was to be cloned into a different vector that includes components of the ADRB2 reporter and the wild type sequence portion of ADRB2 upstream of the segment being cloned. These eight different base vectors were digested with restriction enzymes BamHI and BspQI, BbsI and BspQI, or BbsI and NheI. The base vectors were cleaned up with the DNA Clean and Concentrator Kit and the library segments were ligated into the base vectors with T4 DNA ligase (New England Biolabs). The ligations were transformed into 5-alpha Electrocompetent cells (New England Biolabs) directly into liquid culture. Cultures were
grown at 30°C overnight to maintain library diversity and dilutions were plated on agarose plates to ensure transformation efficiency was high enough to cover the entire library (>100 transformants per library member). DNA was prepared 16 hours later with the DNA Miniprep Kit (Qiagen). The vectors were digested with BspQI and AgeI or NheI and AgeI (Qiagen). Vectors containing unique sequences corresponding to each library segment that complete the ADRB2 protein sequence and reporter were digested with the same restriction enzymes. These fragments were gel isolated from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit (Zymo Research). These secondary fragments were cloned into the library vectors with the same protocol as the previous cloning step. DNA was prepared 16 hours later with the Plasmid Plus DNA Maxiprep Kit (Qiagen).

**Variant-Barcode Mapping**

After the initial cloning of the variant fragments from the oligonucleotide library into each segment’s corresponding base vector, the random barcode attached to each variant was associated to its variant with paired-end sequencing. Each plasmid was amplified with 2 rounds of PCRs with distinct primer sets for each segment (Supplementary Table X) with HiFi DNA Master Mix (Kapa Biosystems). For the first round of amplification, the reaction and cycling conditions were optimized as follows: 98°C for 30 seconds, 10 cycles of 98°C for 8 seconds, 64°C for 15 seconds, and 72°C for 10 seconds, followed by an extension of 72°C for 2 minutes. These amplicons were gel isolated from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit. Prior to the second round of amplification, the number of cycles to amplify was determined by performing qPCR with the SYBR FAST QPCR Master Mix (Kapa) on the CFX Connect Thermocycler according to the manufacturer’s instructions. The Cq determined from the QPCR plus an addition two cycles was used to as the number of cycles to amplify the libraries for the second round of amplification. For the second round of amplification, the reaction and cycling conditions were optimized as follows: 98°C for 30 seconds, X cycles of 98°C for 8 seconds, 62°C for 15 seconds, and 72°C for 10 seconds, followed by an extension of 72°C for 2 minutes. These amplicons were gel isolated from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit. Library concentrations were quantified using a TapeStation 2200 (Agilent) and a Qubit (Thermo Fisher). The libraries were sequenced with paired end 150-bp reads on a NextSeq 500 in medium-output mode and paired end 250-bp reads on a MiSeq (Illumina).
Variant Library Bxb1 Recombinase Plasmid Integrations

HEK293T derived cells engineered to contain the Bxb1 Recombinase site at the H11 locus and deletion of endogenous ADRB2 were seeded at a density of 2.13 million cells per dish in 6 100 mm x 20 mm tissue-culture treated culture dishes (Corning). 24 hours after seeding cells were transfected with 11.5 µg Donor plasmid and 2.9 µg plasmid encoding the Bxb1 recombinase using Lipofectamine 3000. 3 days after transfection cells were expanded to T-225 flasks (Corning) and 8 µg/ml blasticidin was added one day after expansion. Cells were kept under selection 7-10 days and passaged 1:10 four times to ensure removal of transient plasmid DNA.

Multiplexed Variant Functional Assay Agonist Stimulation, RNA Preparation and Sequencing

HEK293T derived cells engineered to contain the Bxb1 Recombinase site at the H11 locus, deletion of endogenous ADRB2, and integration of the ADRB2 mutagenic library were seeded at a density of 3,237,868 cells per dish in 150 mm x 25 mm tissue-culture treated culture dishes. 10 dishes were seeded for each biological replicate of each drug condition. 48 hours after seeding, media was removed and cells were induced with various concentrations of either forskolin or isoproterenol diluted in 9 ml of OptiMEM per plate for 3 hours. After stimulation, media was removed and 3.24 ml of RLT buffer was added to each well to lyse cells. Lysate from dishes belonging to the same replicate were pooled and vortexed thoroughly. 5 ml of lysate from each sample were homogenized with the QIAshredder kit and total RNA was prepared from each sample using the RNeasy Midi Kit with the optional on-column DNase step (Qiagen) and eluted into 500 ul H2O. 40 reverse transcriptase reactions were carried out for each sample using the Superscript IV RT kit (Thermo Fisher). For each reaction 11 ul of total RNA were added to 1 ul dNTPs (Qiagen) and 1 ul 2 uM RT primer (Supplementary Table X). The primers were annealed to the template by heating to 65°C for 5 minutes and cooling down to 0°C for 1 minute. After annealing, 4 ul of RT buffer, 1 ul DTT, 1 ul of RNaseOUT, and 1 ul SSIV were added to the mixture and cDNA synthesis was performed. The reaction and cycling conditions are as follows: 52°C for 1 hour, 80°C for 10 minutes. cDNA from the same sample was pooled together and treated with 100 µg/ml RNase A (Thermo Fisher) and 200 U of RNase H (Enzymatics) at 37°C for 30 minutes. cDNA was concentrated using the Amicon Ultra
0.5 mL 30k Centrifugal Filter (Millipore) according to the manufacturer’s instructions with a final spin step time of 15 minutes. To determine the number of cycles necessary for library amplification in preparation for RNA-seq, 1 ul of cDNA from each sample was amplified with SYBR FAST QPCR Master Mix according to the manufacturer’s instructions using primers for library amplification and adaptor addition (Supplementary Table X). Each sample was subsequently amplified for 4 cycles more than the Cq calculated in the QPCR run adjusting for sample volume. The entire volume of concentrated cDNA for each sample was amplified with sequencing adaptors using NEB-Next High-Fidelity 2x PCR Master Mix (New England Biolabs): 25 ul Master Mix, 2.5 ul of both 10 uM forward and reverse primer (Supplementary Table X), 4 ul of cDNA, and 16 ul H2O. The reaction and cycling conditions are as follows: 98°C for 30 seconds, X cycles of 98°C for 8 seconds, 66°C for 20 seconds, and 72°C for 10 seconds, followed by an extension of 72°C for 2 minutes. Amplified DNA was purified with the DNA Clean and Concentrator kit and gel isolated from a 1% agarose gel with the Zymoclean Gel DNA Recovery Kit. Library concentrations were quantified using a TapeStation 2200 and a Qubit. The libraries were sequenced with an i7 index read and a single end 75-bp read on a NextSeq 500 in high-output mode.

Quantification and Statistical Analysis

Barcode Mapping

We used the BBTools suite(https://jgi.doe.gov/data-and-tools/bbtools/) of programs to process our sequencing data using the default settings unless otherwise noted. First, we used BBDuk2 to filter out any reads matching PhiX (k=23, mink=11, hdist=1) and to trim off any Illumina sequencing adapters. We then used BBMerge to merge our paired end reads. We performed another round of trimming with BBDuk2 to ensure no adapters were left over after merging and to remove any sequence with an N base call. After merging and trimming the reads, we used a custom Python script (bcmap.py) to generate a consensus nucleotide sequence for each barcode.

Briefly the script works as follows. First, we split each read into the 15 nt barcode and its corresponding variant. We then generate a dictionary that maps each barcode to its list of unique sequences and their counts. To enable majority basecall we drop any barcode that has less than 3 reads. We then pass the barcodes through a series of filters to eliminate potential errors introduced by
barcodes that are mapped to multiple variants. Since we barcoded and mutagenized the ADRB2 gene in separate pieces, barcodes can be associated to variants from different pieces. We address this case by using **BBMap** to align every barcode’s sequences to the ADRB2 reference and consider that barcode to be contaminated if any sequence aligns >5 nt away from the most common sequence. Another source of contamination comes from our chip-synthesized library itself, which contains a significant number of single base deletions. We consider a barcode contaminated if it has any sequences of different lengths as it is unlikely that a single base deletion will come from an Illumina sequencer by chance. However, these filters would not catch the case where a barcode is contaminated with variants from the same piece of ADRB2. As we only synthesized the missense variants, we expect variants within the same piece of ADRB2 to be a Levenshtein distance of 4 from each other on average (approximately two changes to WT and two changes to a new codon). Thus, we drop any barcode that has a sequence with >1 read at a Levenshtein distance of 4 away from that barcode’s most common sequence. Lastly, we generate a consensus sequence by taking the majority base call at each position and call an N at any ties.

After we associate each barcode with its consensus sequence, we use a series of different alignments to determine that sequence’s identity. To find the designed missense variants in our library, we use **BBMap** to search for barcodes that an exact alignment to them. To find frameshift mutations, we use **BBMap** to align the consensus sequences to the ADRB2 reference and parse the resulting CIGAR strings for indels with a simple python script (**classify-negs.py**). Finding synonymous mutants required more processing as each sub-library did not start at a complete codon. We first used the rough **BBMap** alignment to determine what ADRB2 chunk each sequence was associated with. We then used a custom python script (**synon-filter.py**) to trim up to the last whole clonal codon, as the first few codons of each sequence were part of the clonally backbone and are unlikely to have any errors. Finally, we translated the resulting sequences, aligned the protein sequence to the ADRB2 coding sequence with a Smith-Waterman aligner from the Parasail library (**https://github.com/jeffdaily/parasail**), and retained perfect translations with the correct length.
Data Normalization

We incubated our cellular library with forskolin to activate the cAMP reporter in each cell, providing an agonist-independent measurement of maximal reporter activity. This measurement can be used to approximate cellular copy number. To ensure that barcodes with low cellular representation are excluded from our analyses we require that all forskolin barcodes be present in both repeats, we normalize our read counts to sequencing depth, we average the two repeats together, and filter out any barcodes less than 0.2 RPM (~8-10 reads at our sequencing depth). Next, we use this list of barcodes to control for copy-number variation in our measurements. We first require that all of the barcodes in the forskolin condition are also present in our drug conditions, and add a pseudocount that is scaled relative to the condition with the fewest number of reads ($\frac{N}{\min N}$). This explicitly sets missing barcodes to the pseudocount. We then normalize each condition to its read depth (including added pseudocounts) and divide this value by its associated forskolin value. We also excluded barcodes with high forskolin counts ($\geq 10$ RPM) as they are systematically less induced in the drug conditions relative to other barcodes.

With a filtered set of barcodes in place, we averaged together all measurements for each variant (median 11 barcodes per variant), keeping the repeats separate. To make our values more interpretable, defined activity as the ratio of these values to the mean frameshift. We then averaged the relative activities of the two repeats together and used propagation of uncertainty to combine their standard deviations.

Conservation, EVMutation, and gnomAD

To calculate sequence conservation, we aligned 55 ADRB2 orthologs from the OMA database (entry: HUMAN24043) using MAFFT with the default settings (`mafft -reorder -auto`). We then used the Jensen-Shannon Divergence to score this alignment and majority basecall to generate a consensus sequence ignoring any gaps if they made up $< 35\%$ of the alignment at that position. Using EMBOSS Needle to align this consensus sequence back to the ADRB2 reference, we found the consensus sequence had a two nt insertion at positions 360 and 361. We excluded these positions for the purposes of our analyses. For both EVMutation and gnomAD, we simply downloaded the results for ADRB2.
Unsupervised Learning

We performed a number of preprocessing steps before running UMAP on our data. First, we grouped amino acids into 8 different classes based on their physiochemical properties ((+) - H, K; (-) - D, E; Aromatic - F, W, Y; Amide - N, Q; Nucleophilic - C, S, T; Hydrophobic - I, L, V, M; Small - G, A; Proline - P) and averaged their relative activities. Next, we standardized the log2 relative activity values of each group and used mean imputation to model missing data for any missing AA groups at a given position. Finally, we combined the data from every drug condition into a 412 x 32 design matrix in which the columns are an AA group at a specific condition and the rows are the positions in the protein.

With our data processed, we used the R implementation of UMAP to run hyperparameter search of all combinations of UMAP embeddings with the parameters $n_{neighbors} = (4, 8, 16, 32)$ and $n_{components} = (2, 3, 4, 5, 6, 7, 8, 9, 10)$, holding $min\_dist=0$ and $n\_epochs=2000$ constant. This provided a variety of different representations of our data that we used HDBSCAN to search for clusters in these embeddings (R package dbscan; $minPts = 10$). To ease interpretation of the clustering, we plotted the HDBSCAN results onto a 2D UMAP embedding with the following parameters: $n\_neighbors=4$, $min\_dist=0$, $n\_components=2$, $n\_epochs=2000$, and $random\_state=3308004$ using the Python implementation (https://github.com/lmcinnes/umap).

We found the cluster assignments to be largely robust across the different embeddings, and used them to guide our manual cluster assignment.

Identification of Activating and Hypomorphic Mutations

We defined a variant to be “activating” if its mean activity minus its standard deviation were greater than the mean synonymous variant. Similarly, we defined a variant to be hypomorphic if its mean activity plus its standard deviation were less than the mean synonymous variant.

Mutational Tolerance

We defined mutational tolerance as a given residue’s ability to accommodate all amino acid substitutions. To calculate this, we first capped the maximum our activity values at WT-like activity (the mean of the synonymous barcodes). Similarly, we capped the minimum activity at the mean of the
frameshifts (1 on our activity scale). By limiting our activity measurements to this range, we ensure that individual substitutions do not have an outsized effect on the mutational tolerance. Next, we averaged the activities of every amino acid substitution for each position in the \( \beta_2 \)AR. Finally, we scaled the mutational tolerance values to lie on a 0-1 scale.

**Statistical Tests**

All statistical tests unless otherwise noted are the two-sided Mann-Whitney U and were performed in R (version 3.5.x) using the `wilcox.test` function.

**Structural Modeling and Solvent Accessible Surface Area**

All molecular graphics and analyses were performed with the UCSF Chimera package. To determine if a given in the \( \beta_2 \)AR points into the core of the protein or into the lipid membrane, we used FreeSASA (version 2.0.3) to calculate the Solvent Accessible Surface Area (SASA) of the \( G_s \)-bound \( \beta_2 \)AR (PDB: 3SN6). The \( G_s \) occludes the intracellular surface of the \( \beta_2 \)AR thereby reducing the SASA of residues on the intracellular surface. Similarly, the extracellular surface is mostly blocked by the extracellular loops. Finally, we used the Orientations of Proteins in Membranes (OPM) database to filter out any residues outside of the lipid membrane from our analyses. To quantify charge sensitivity, we calculated the average activity for H, K, R, D, and E substitutions at each agonist concentration for residues in the lipid membrane. We then multiplied the values by -1 and standardized the results within each concentration such that the values were mean-centered and scaled by their standard deviation. We calculated hydrophobic sensitivity (I, L, V, M) in an analogous manner. Next, we classified residues that had above average charge sensitivity and below average hydrophobic sensitivity as being exclusively charge sensitive. Conversely, we classified residues that had above average charge sensitivity and above average hydrophobic sensitivity as being intolerant.

**Code Availability**

All code is available at [www.github.com/KosuriLab/ADRB2](http://www.github.com/KosuriLab/ADRB2). Sequencing data can be accessed from the sequencing read archive (SRA) with the accession number XXXXXXXXXX
6.6 Supplementary Information

Figure 6.5: Schematic of Generation, Functional Assessment, and Analysis of All 7,828 Missense Variants of the β2AR. We synthesized missense variants on an oligonucleotide microarray, amplified the oligos and appended random DNA barcode sequences, and cloned the variants into WT background vectors. We then mapped barcode-variant pairs with next-generation sequencing and cloned the remainder of the WT receptor and genetic reporter into the construct. Next, we integrated the variant library en masse into a serine recombinase landing pad engineered at the H11 locus of ∆ADRB2 HEK293T cells. The recombination strategy ensures a single receptor variant/genetic reporter is integrated per cell to avoid crosstalk between genetic reporters. After selection, we stimulated the library with various concentrations of ADRB2 agonist, isoproterenol. Finally, we determined mutant activity by measuring the relative abundance of each variant’s barcoded cAMP-responsive genetic reporter transcripts with RNA-seq.
Figure 6.6: Engineering HEK293T Cells for Clonal and Functional Integration of an ADRB2 Genetic Reporter. **A.** Schematic of functional assay to ensure the landing pad is present at single copy in the genome and can recombine a single donor plasmid per cell. Single copy integration is essential to ensure receptor's of variable functionality do not activate barcoded reporters mapped to other variants. Upon co-transfection of the promoterless GFP and mCherry plasmids with bxb1 recombinase sites, a cell line with a single landing pad will exclusively integrate one cassette. Therefore, cells will be either GFP$^+$ or mCherry$^+$ but never both. **B.** Flow Cytometry plots detailing the percentage of GFP$^+$ and mCherry$^+$ cells when transfected with an equimolar ratio of promoterless GFP and mCherry expression cassettes with or without Bxb1 recombinase expression. **C.** Activation of a cAMP-responsive genetic reporter via a luciferase assay integrated in the landing pad when stimulated with isoproterenol in a WT or ΔADRB2 background. Activation of the reporter in the WT background emphasizes the importance for generation of the ΔADRB2 for the purpose of multiplexed experiment. **D.** Activation of a genetic reporter with or without exogenous ADRB2 expression via a luciferase assay integrated in the landing pad when stimulated with isoproterenol in ΔADRB2 cells. **E.** Activation of an equivalent integrated genetic reporter/ADRB2 cassette via qRT-PCR of the reporter transcript in ΔADRB2 cells. **F.** Schematic detailing the recombination of the reporter/receptor expression plasmid into the landing pad locus.
Figure 6.7: **Individual and Global Multiplexed Assay Validation.**

A. To validate our genetic reporter, we compared the measured mutant activity screened individually in the landing pad locus via a luciferase assay to via the multiplexed mutational scan. We recapitulated results individually observed in the DMS for both null and hypomorphic mutations. 

B. The distribution of activity for frameshifts are significantly different that the distribution of our designed missense mutations across increasing isoproterenol concentrations ($p << 0.001$). 

C. We also find the relative activity for frameshift mutations mapped to each codon in the ADRB2 sequence is markedly decreased in the C-terminus of the protein (dotted line), and is consistent across agonist concentration. Blue line represents the LOESS fit. 

D. The measurements between barcodes at the RNA-seq level are well correlated ($r = 0.867, r = 0.871, r = 0.864, r = 0.868$) at all agonist concentrations (0, 0.150, 0.625, and 5 $\mu$M Iso). Similarly, the mean forskolin-normalized values for each variant are correlated at every concentration ($r = 0.657, r = 0.686, r = 0.729, r = 0.750$). Bars represent log$_{10}$ counts per hex-bin.
Figure 6.8: **Activity of Proline Mutations by Protein Domain.** Proline substitutions in the transmembrane domain result in lower activity across all agonist conditions.

Figure 6.9: **Correlation with Sequence Conservation and Covariation.**

A. Mutational tolerance is highly correlated with sequence conservation and is maximized at EC$_{100}$ ($\rho = -0.689, \rho = -0.719, \rho = -0.747, \rho = -0.634$ for -Iso, 0.150 µM Iso, 0.625 µM Iso, and 5 µM Iso, respectively). Here we calculated sequence conservation using the Jensen-Shannon divergence from a multiple alignment of 55 ADRB2 orthologs from the OMA database. The blue line is the least squares fit.

B. Similarly, our measure of relative activity for individual substitutions is well correlated with the predictions from EVMutation, and is maximized at EC$_{100}$ ($\rho = 0.370, \rho = 0.460, \rho = 0.521, \rho = 0.504$). The blue line is the least squares fit.
Figure 6.10: **Cluster Assignment is Robust Across Different UMAP Embeddings.** Given the high dimensionality of our data, we used UMAP to learn lower-dimension representations of our data before clustering with HDBSCAN (minimum cluster size = 10). To ensure that the clustering results are not biased by a particular UMAP embedding, we ran a hyperparameter search over the dimension and nearest neighbor parameters of UMAP. We then plot the HDBSCAN cluster assignments on a 2D UMAP embedding to ease visualization. Points that HDBSCAN does not assign to a cluster are colored powder blue. We find that groups of residues reliably cluster together regardless of the UMAP embedding, and manually assign all residues to six distinct clusters following the robust HDBSCAN assignment.
Figure 6.11: Mutational Profile Suggests Side Chain Orientation and Environment. A. The crystal structure of the hydroxybenzyl isoproterenol-activated state of the $\beta_2$AR (PDB: 4LDL) with residues colored by UMAP cluster identity. B. Distributions of Solvent Accessible Surface Area (SASA) for each cluster at EC$_{100}$. C. Hydrophobic versus Charge Sensitivity across all drug conditions. Points are colored by cluster identity. We define residues to be globally intolerant to substitution if their Hydrophobic and Charge Sensitivity is greater than 0. Similarly, we define residues to be uniquely charge sensitive if their Hydrophobic Sensitivity is less than 1 and their Charge Sensitivity is greater than 1. D. Distributions of SASA for intolerant and charge sensitive clusters are significantly different across all drug concentrations (all $p < 0.0005$).
Figure 6.12: Inspection of Mutationally Intolerant Residues. A. Rank order plot of mutational tolerance at 0.625 µM isoproterenol for all 412 β2AR residues mutagenized. Residues in known structural motifs (colored points) are significantly more sensitive to mutation than other positions on the protein (p << 0.001). B. Residues that interact with the head (orange) of hydroxybenzyl isoproterenol have significantly lower mutational tolerance than those that interact with the hydroxybenzyl functional group on the tail (purple). These differences are significantly different at EC$_{50}$ (p = 0.028), EC$_{100}$ (p = 0.016), and saturating agonist concentration (p = 0.008). C. Box plot displaying the mutational tolerance of all predicted contacts of the cholesterol binding pocket determined in the timolol-bound structure of the β2AR inactive state (PDB: 3D4S). The highly conserved W158 is the most constrained residue. D. ECL1, with residues belonging to cluster 1 and 2 colored magenta, contains a region of sensitivity where R63 and L64 are both intolerant to substitution. However, neighboring E62 displays greater than WT activity at multiple individual mutations (PDB: 3SN6).
Figure 6.13: **Evaluation of Individual Missense Variants.**

A. The inactive \( \beta_2 \)AR structure highlighted in regions where residues display greater than WT activity without agonist stimulation for at least one individual mutation (yellow). These mutations localize to the extracellular membrane interface of TM1, TM2, and ECL1.

B. Other concentrations of these mutants are found in the lower half of TM1, helix 8, and the TM5-TM6 interface. The blue colored structure represents the shift in TM6 upon adoption of the active state.

C. 2-D snake plot with residues colored by the number of individual mutations that lead to greater than WT activity in the no agonist condition. These residues are enriched in the loops and termini which are truncated in the crystal structures.

D. Activity of all ADRB2 mutants present in the gnomAD database plotted against their allele frequency. We classified variants into four categories as follows: null mutants (purple) are variants whose mean plus a standard deviation (SD) are less than 1 (the mean frameshift); activating mutants (orange) are variants whose mean minus a SD are greater than the mean synonymous mutant (dashed line); hypomorphic mutants (periwinkle) are variants whose mean plus a SD are less than the mean synonymous variant; the rest of the variants are considered WT-like (white).
References


