High-Throughput Image-Based Screening of Pooled Genetic Variant Libraries

High-throughput image-based screening of pooled genetic variant libraries

A dissertation presented

By

George Alexander Emanuel

To

The Committee on Higher Degrees in Biophysics

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biophysics

Harvard University

Cambridge, Massachusetts

April 2018

Dissertation Advisor: Professor Xiaowei Zhuang George Alexander Emanuel

High-throughput image-based screening of pooled genetic variant libraries

Abstract

Image-based, high-throughput screening of pooled libraries of genetic perturbations will greatly advance our understanding of biological systems and facilitate many biotechnology applications. Here we report a high-throughput screening method that allows diverse genotypes and the corresponding phenotypes to be imaged in numerous individual cells. To facilitate genotyping by imaging, we introduced barcoded genetic variants into cells, each cell carrying a single genetic variant connected to a nucleic-acid barcode. We then performed live-cell imaging to determine the phenotype of each cell and massively multiplexed FISH imaging to measure the barcode expressed in the same cell for genotype identification. We demonstrated the utility of this method by screening genetic variants of the fluorescent protein YFAST. We imaged 20 million cells expressing ~60,000 YFAST mutants and identified novel YFAST variants that are both brighter and more photostable than the original protein.

Table of contents

Citation to previously published work

This dissertation was adapted from the following publication:

Emanuel, G., Moffitt, J.R., Zhuang, X. (2017) High-throughput image-based screening of pooled genetic variant libraries. Nature Methods 14(12) 1159-62.

Acknowledgements

My academic and personal growth over the course of my PhD was only possible with the help of all the inspiring people around me.

First, I thank my advisor, Professor Xiaowei Zhuang, for her mentorship and for the opportunity to conduct research in her laboratory. Xiaowei's astute guidance for selecting worthwhile problems combined with the freedom she provided to direct my research largely at my own discretion inspired me to execute my ideas from simple thoughts to functional technologies while learning from the countless mistakes I made along the way. This process was enjoyable in the environment she has created, where my experiments were only limited by the quality of my own ideas and the number of hours in a day rather than any financial constraint. This freedom together with her continuous demand for scientific rigor challenged me to plan and execute my work as logically as possible and has fundamentally changed the way I approach problems.

The biophysics program, carefully fostered by Professor Jim Hogle and Michele Eva-Pfeffer, provided an ideal infrastructure for my graduate studies through its flexibility and the broad range of research interests represented in the program. I would like to thank Jim and Michele for their continued commitment to the success and wellbeing of the students in the program. I also thank the faculty who served on my various committees: Professors Adam

Cohen, Sunney Xie, Ethan Garner, David Liu, Wesley Wong, Gary Yellen, and Eugene Shakhnovich,

Furthermore, I am grateful for all the scientist's I worked alongside every day in the Zhuang laboratory. In particular, I thank Dr. Jeffrey R. Moffitt for many productive conversations, where he readily shared both his expertise with *E. coli* and his enthusiasm for multiplexed FISH. I also thank Alec Goodman and Julie Szabo for keeping the lab running efficiently.

Finally, I would like to thank my parents, Jeff Emanuel and Stephanie Lyon, my brothers, extended family, and friends, including the many from Harvard Cycling, for their continued support.

Chapter 1. Identifying genetic variants using multiplexed in situ hybridization

Introduction

High-throughput screening of genetic variants or perturbations is playing an increasingly important role in advancing biology and biotechnology. For example, by simultaneously probing the effects of a large number of amino acid (or nucleic acid) changes within a target molecule, large-scale screening can enable efficient searches for fluorescent proteins better adapted for bioimaging[1] or protein and nucleic acid drugs with desired therapeutic properties[2,3]. In another area of applications, large-scale screening can also be used to efficiently probe the cellular responses to the inhibition or activation of individual genes or combinations of genes at the genomic scale[4], which helps decipher the roles of genes and gene networks on cellular behaviors.

Large-scale screening efforts are greatly facilitated by pooled, high-diversity libraries of genetic variants or perturbations because of the ease and scalability associated with the construction of these pools. Methods such as error-prone PCR[5] or cloning with large pools of array-synthesized oligonucleotides[6] allow pooled libraries with a very large number of genetic variants or perturbations to be created, often with a similar degree of effort as required to make any individual library member. However, unlike screening of individually constructed variants where the genotype is known *a priori*, screening of pooled libraries

requires methods to measure the genotype that produced the desired phenotype, and this is typically done by selecting/enriching library members with desired phenotypes and then using sequencing to determine their corresponding genotypes. Such approaches have been used to identify protein variants with desired properties, such as fluorescent proteins with improved brightness[1], reversible photoswitching[7,8] and increased lifetime[9–12]. At the genome-wide scale, RNAi or CRISPR-based approaches have been used in pooled library screens to measure the role of numerous genes in cellular phenotypes such as viability[13–20] and, more recently, in the expression of large fractions of the transcriptome[21–23]. However, there are many important phenotypes that cannot be measured easily with existing high-throughput screening approaches. Phenotypes ranging from cellular morphology and dynamics to the intracellular organization of proteins or RNAs require high-resolution, time-lapse imaging to be measured. Moreover, time-lapsed imaging can also facilitate the screening of many important photo-physical properties of fluorescent proteins. Unfortunately, it is challenging to combine pooled library screening with high-resolution imaging because it is difficult to isolate individual library members based on their imaged phenotype and then determine their genotype. If, however, one could measure the genotype of individual library members in situ by imaging, then library member isolation would not be required, making it possible to combine the ease and scalability of pooled library screening with imaging-based phenotype measurements. Moreover, with the ability to determine the phenotypes generated by all genotypes in the library, not just for select members with the desired phenotype, such an approach could map the full genotype-phenotype landscape, which is a powerful tool for both genome-wide screens and protein engineering.

Here we report a novel high-throughput, imaging-based screening method that allows the characterization of both the phenotype and genotype of individual cells in pooled populations of genetically diverse cells. In this method, we associate each genetic variant with a unique nucleic acid barcode. We then use imaging to determine both the phenotype of each cell as well as the corresponding genotype of the same cell by measuring the barcode expressed in the cell. We demonstrate the power of this method by screening 20 million *E. coli* cells containing ~160,000 unique barcodes and ~60,000 variants of YFAST, a recently developed fluorescent protein that becomes fluorescent upon binding to an exogenous chromophore[24]. From this screen, we identified a series of YFAST mutants with both increased brightness and photostability.

Barcoding genetic variants

The first step in our approach is the design of the nucleic acid barcodes. These barcodes are comprised of a series of nucleic acid hybridization sites, each corresponding to one bit in a *N*-bit binary code. For each bit, we designed two different sequences (termed readout sequences), one representing the value of "1" and the other representing "0" (Fig. 1A). We note that different designs of barcodes are possible, for example, a simplified binary code with "1" or "0" represented by the presence or absence of a readout sequence or a ternary code with three different readout sequences assigned to each bit. Because the number of unique barcodes grows exponentially with the number of bits, our barcoding scheme potentially allows the screening of millions of genetic variants with barcodes that contain just tens of hybridization sites.

We then randomly assigned a barcode to each genetic variant by randomly incorporating these barcodes into plasmids containing the desired genetic variants (Fig. 1B). To minimize the probability that the same barcode is assigned to multiple genetic variants, we designed a bottlenecking strategy: specifically, after generating the plasmid library containing the barcoded genetic variants, we selected a small, random subset of library members, the number of which is much smaller than the total number of possible *N*-bit binary barcodes. With this strategy, only a small fraction of the selected barcodes would be associated with more than one genetic variant by chance. In addition, when it was necessary to ensure that all, or nearly all, of the genetic variants are present in the final library, we further chose the selected barcode number to be substantially bigger than the number of genetic variants. We then used next generation sequencing to determine which barcode was associated with each variant and removed any remaining ambiguous barcode (i.e. a barcode assigned to more than one genetic variant) from further analysis. An added benefit of this bottlenecking strategy is that it provides error robustness to the barcode detection, i.e. if any bit of a barcode is mis-read, it will most likely produce an invalid barcode that is not present in the library—an error that can be detected and removed, as elaborated later.

The barcoded genetic-variant library was then incorporated into cells, where the genetic variant was expressed and the barcode sequence was transcribed, producing an RNA barcode. The phenotype of each cell was then determined via imaging (Fig. 1C), and the sequence of the RNA barcode expressed within each cell was determined using a modified version of multiplexed error-robust fluorescence in situ hybridization (MERFISH)[25]. Specifically, we used multiple rounds of hybridization to detect the barcodes, each round probing either a single readout sequence using a complementary FISH probe (readout probe)

or multiple readout sequences simultaneously using multiple readout probes linked to spectrally distinct dyes (Fig. 1C). Unlike our previous MERFISH experiments[25–27], where we exploit single-molecule FISH[28,29] to measure the copy numbers and locations of numerous RNA species within single cells, here each individual cell expressed only one type of barcode RNA in high abundance, and we measured the total signal from all barcode RNA molecules within each cell for each round of hybridization to determine the corresponding bit value for that cell. Thus, the substantially brighter signals should lead to low error rate, and we exploited the error robustness provided by the aforementioned bottlenecking strategy to detect any remaining errors.

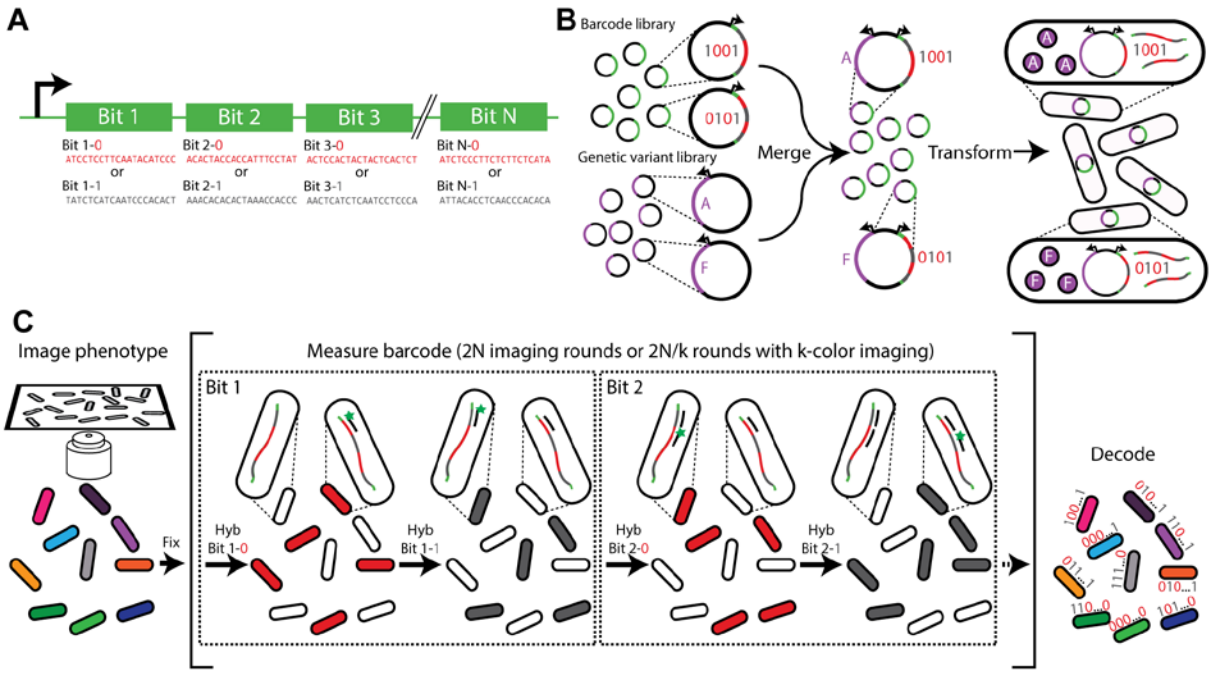

Figure 1. A high-throughput, image-based screening method using massively multiplexed fluorescence in situ hybridization. (A) Schematic depiction of a nucleic acid barcode. Each barcode consists of the concatenation of a series of hybridization sites, each of which is associated with a different bit in an *N*-bit binary barcode. Each hybridization site can utilize one of two readout sequences unique to that site, with one readout sequence representing the value of "1" at that bit and another representing the value of "0" at that bit. (B) Schematic depiction of barcoded genetic variant library construction. The library of barcodes is merged with a library of genetic variants and transformed into bacteria. The correspondence between the barcodes and genetic variants is determined by sequencing. (C) Schematic diagram of the image-based phenotype-genotype characterization. The phenotype is first characterized in surface-adhered cells. Then, the cells are fixed, and multiple rounds of hybridization are used to measure the RNA barcodes expressed in the cells. During the first round, readout probe 1-0 is added and cells with barcodes that read "0" in the first bit, i.e. which contain the readout sequence 1-0, should bind to the probe and become fluorescent, whereas cells with barcodes that read "1" in the first bit should remain dark. Once readout probe 1-0 is extinguished, readout probe 1-1 is added and the cells with barcodes that read "1" in the first bit, which contain the readout sequence 1-1, should become fluorescent. This process is repeated for the remaining bits. After measuring all bits, the barcode is determined for each cell, revealing the identity of the genetic variant contained in the cell and linking the measured phenotype to the genotype.

Measuring 2 phenotypes with 80,000 barcodes

To demonstrate the feasibility of our image-based screening method, we created a high-diversity, barcoded genetic variant library that contains ~80,000 distinct barcodes. To quantify the accuracy of our barcode measurements, we associated these barcodes with only two genotypes—the presence or absence of a fluorescent protein, which produced simple and clear phenotypes—the presence or absence of fluorescence in cells. To this end, we first created a library of all possible 21-bit binary barcodes as described above (Fig. 1A, B), and then cloned these into a high copy plasmid under the control of the high expression promoter, lpp. We inserted into this plasmid library a construct that expresses either the translational fusion of the blue fluorescent protein mTagBFP2[30] and the photo-switchable red fluorescent protein mMaple3[31] (mMaple3+) or mTagBFP2 alone (mMaple3-), and then cloned these plasmid libraries into *E. coli* (Fig. 2A). With 21 bits, there are over 2 million (2^{21}) possible barcodes. For error-robust barcode detection and unambiguous genotype identification, as described above, we bottlenecked each of these two libraries (mMaple3+ and mMaple3-) to roughly 40,000 members each and combined the two libraries to produce a final library that contained ~80,000 unique barcodes (only ~4% of the 2 million possible barcodes). We then used next generation sequencing to determine which barcodes were present in the final library and their corresponding (mMaple3+ or mMaple3-) genotype.

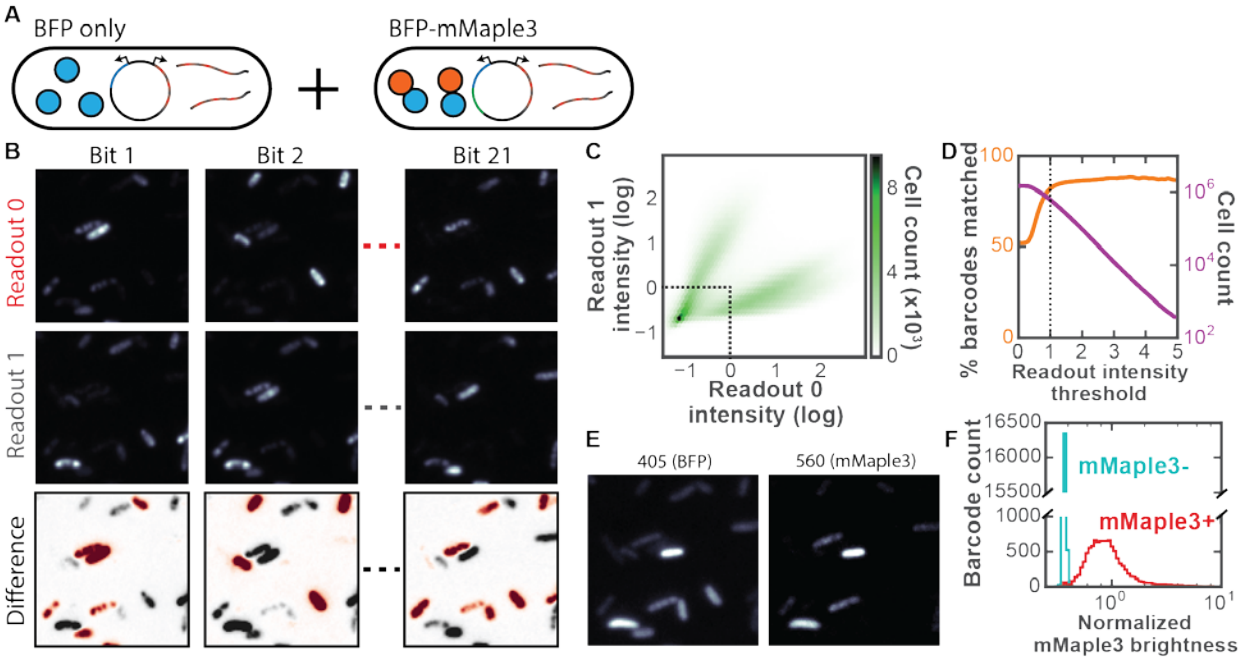

Figure 2. Performance characterization of the imaged-based screening method by measuring 1.5 million cells containing 80,000 unique barcodes associated with two known genotypes and phenotypes. (A) Schematic diagram of the library constituents. Among the 80,000 distinct 21-bit barcodes used here, approximately half are associated with the mTagBFP2 gene while the remaining are associated with the mTagBFP2-mMaple3 fusion gene. (B) Fluorescent images for each readout from a subset of the 21 bits. The top and middle panels for each bit show the images of cells after hybridization to the readout probes corresponding to "0" (top) or "1" (middle) at this bit. The difference image (bottom) indicates whether a "0" (gray) or "1" (red) value is assigned to the barcode within the cells for that bit. (C) Two-dimensional histogram of normalized fluorescence intensities for readout 0 and readout 1 of bit 1 for the measured cell. The fluorescence intensities are normalized to the median values of all cells. The dotted line depicts the threshold used for eliminating cells that appear dim in both readouts. The shade of green indicates the number of cells. (D) Orange: The percent of barcodes decoded in the imaging experiment that match valid barcodes present in the library as a function of the readout intensity threshold used to eliminate dim cells. Magenta: The number of cells above the readout intensity threshold. The dotted line corresponds with the threshold of 1 shown in (C). (E) Fluorescence image of mTagBFP2 and fluorescence image of post-activation mMaple3 of the same region as (B). (F) Histograms of median mMaple3 fluorescence intensity normalized to mTagBFP intensity for all barcodes associated with the mMaple3-mTagBFP2 fusion gene (mMaple3+, red) and for those associated with the mTagBFP2 gene (mMaple3-, cyan). Only those barcodes that were measured in at least five cells were analyzed, corresponding to in total 133,250 cells for mMaple+ and 194,329 cells for mMaple3-. These two numbers do not add up to the total number cells determined with valid barcodes because, for many barcodes, less than five cells were measured.

To determine the phenotypes of individual *E. coli* cells in the library, we adhered them to the surface of a coverslip and imaged the cells by illuminating with 405-nm light for 120 ms (one camera frame) to measure the mTagBFP2 fluorescence, illuminating with 405-nm light for an additional ~4 s in order to switch the mMaple3 protein to its red-fluorescent state, and illuminating with 560-nm light for 120 ms to determine the fluorescence intensity of mMaple3. Cells were then treated with a mixture of methanol and acetone to fix them and to permeabilize the membranes for fast hybridization to RNA[32]. To identify the RNA barcode expressed within each cell, we performed MERFISH imaging with three different readout probes in each round of hybridization, with each probe conjugated to a spectrally distinct fluorophore: ATTO565, Cy5, or Alexa750 (Table 1). In total, we screened 1.5 million *E. coli* cells in a measurement that in total lasted ~40 hours.

Table 1. Readout probes used for barcode readout. A list of the readout sequences used for the "0" and "1" values of each bit, the dye to which each probe was conjugated, and the specific hybridization round in which the probe was hybridized.

Bit number	Bit value	Readout sequence	Dye	Hybridization round
1	0	ATCCTCCTTCAATACATCCC	Cy5	1
1	1	TATCTCATCAATCCCACACT	Cy5	7
2	0	ACACTACCACCATTTCCTAT	ATTO565	1
2	1	AAACACACACTAAACCACCC	ATTO565	6
3	0	ACTCCACTACTACTCACTCT	Alexa750	1
3	1	AACTCATCTCAATCCTCCCA	Alexa750	6
4	0	ACCCTCTAACTTCCATCACA	Cy5	2
4	1	AATACTCTCCCACCTCAACT	Cy5	8
5	0	ACCACAACCCATTCCTTTCA	ATTO565	2
5	1	TCTATCATCTCCAAACCACA	ATTO565	7
6	0	TTTCTACCACTAATCAACCC	Alexa750	2
6	1	TCCAACTCATCTCTAATCTC	Alexa750	7
7	0	ACCCTTTACAAACACACCCT	Cy5	3
7	1	TTCCTAACAAATCACATCCC	Cy5	9
8	0	TCCTATTCTCAACCTAACCT	ATTO565	3
8	1	ATAAATCATTCCCACTACCC	ATTO565	8
9	0	TATCCTTCAATCCCTCCACA	Alexa750	3
9	1	ACCCAACACTCATAACATCC	Alexa750	8
10	0	ACATTACACCTCATTCTCCC	Cy5	4
10	1	TACTACAAACCCATAATCCC	Cy5	10
11	0	TTTACTCCCTACACCTCCAA	ATTO565	4
11	1	ACTTTCCACATACTATCCCA	ATTO565	9
12	0	TTCTCCCTCTATCAACTCTA	Alexa750	4
12	1	TTCTTCCCTCAATCTTCATC	Alexa750	9
13	0	ACCCTTACTACTACATCATC	Cy5	5
13	1	AATCTCACCTTCCACTTCAC	Cy5	11
14	0	TCCTAACAACCAACTACTCC	ATTO565	5
14	1	ACCTTTCTCCATACCCAACT	ATTO565	10
15	0	TCTATCATTACCCTCCTCCT	Alexa750	5
15	1	TCCTCATCTTACTCCCTCTA	Alexa750	10
16	0	TATTCACCTTACAAACCCTC	Cy5	6
16	1	TCAAACTTTCCAACCACCTC	Cy5	12
17	0	TTACCTCTAACCCTCCATTC	ATTO565	11
17	1	ACACCATTTATCCACTCCTC	ATTO565	12
18	0	TCCCAACTAACCTAACATTC	Alexa750	11
18	1	ACATCCTAACTACAACCTTC	Alexa750	12
19	0	ATCCTCACTACATCATCCAC	Cy5	13

(continued)

19	1	TCTCACACCACTTTCCTCAT	Cy5	14
20	0	TCCCTATCAATCTCCATAAC	ATTO565	13
20	1	TTATCCATCCCTCTTCCTAC	ATTO565	14
21	0	TCACCTCTAACTCATTACCT	Alexa750	13
21	1	TCCTACAACATCCTTCCTAA	Alexa750	14
22	0	ATCTCCCTTCTCTTCTCATA		Not measured
22	1	ATTACACCTCAACCCACACA		Not measured

As expected, for each bit, the majority of cells were either bright when stained with the readout probe representing the value of “1” and dim when stained with the readout probe representing the value of “0”, or vice versa (Fig. 2B, C). However, a fraction of cells appeared relatively dark in both the “1” and “0” channels, possibly because they were not expressing sufficient barcode RNA, or they were insufficiently permeabilized for readout probe hybridization. We therefore used a conservative thresholding strategy to remove these dim cells from further analysis. Specifically, we removed those cells whose “0” and “1” readout signals for any bit were both smaller than the respective median intensity observed for that readout signal across all cells (Fig. 2C). As expected, there was a strong correlation between cells that were below the threshold in different bits: if one bit was under threshold for a cell, the probability that the second bit was also under threshold was 80%. After applying this conservative thresholding, more than 600,000 of the 1.5 million measured cells remained (Fig. 2D). We then calculated the ratio of the detected “0” and “1” signals for each bit, and observed that cells were largely divided into two distinct populations based on this ratio with only a very small overlap between these two populations (Fig. 3A). We then used this “0”-to-“1” intensity ratio for each bit to determine the barcode sequence for each cell, calling the bit value “0” and “1” if the “0”-to-“1” intensity ratio was above or below a threshold (Online Methods), and found that 84% of the measured barcodes matched a valid barcode that was present in the library as determined by sequencing (Fig. 2D). The small fraction of cells in the overlapping region with nearly equal “0” and “1” intensity values would give relatively low-confidence bit-value assignment and could give rise to incorrect barcodes. Indeed, among the cells that have been assigned to valid barcodes, the distribution of the “0”-to-“1” intensity ratio showed two well separated populations of cells with

essentially zero overlap (Fig. 3B), consistent with the notion that incorrect calling of bits with ambiguous “0”-to-“1” ratios generated many of these incorrect barcodes. More stringent intensity thresholds for each bit discarded more cells without substantial improvement to the fraction of matching barcodes (Fig. 2D). In total, the bit intensity thresholding and rejection of invalid barcodes that did not match library barcodes led to the removal of ~66% of measured cells.

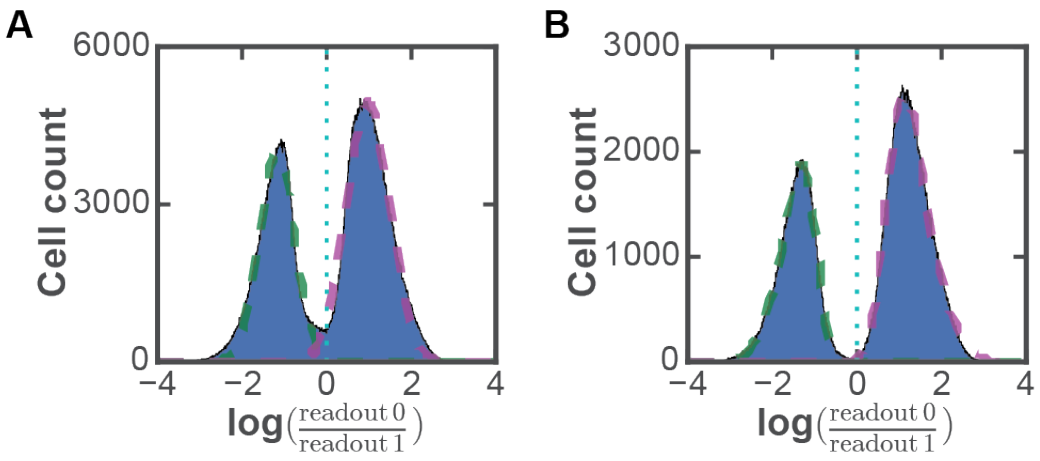

Figure 2. Distributions of the ratio of readout "0" intensity to the readout "1" intensity. (A) Histogram of the natural logarithm of the ratio of readout "0" intensity to readout "1" intensity for bit 1 for all cells that are above the intensity threshold for bit 1. The histogram was fit to a sum of two Gaussians and each fit Gaussian is depicted (green and magenta dashed lines). The overlap between the two Gaussian curves corresponds to 2% of the total number of cells presented here. (B) Histogram of the natural logarithm of the ratio of readout "0" intensity to readout "1" intensity for bit 1 for only those cells with assigned barcodes that match the valid barcodes that are determined to be in the library by sequencing. The histogram was fit to a sum of two skewed Gaussians and each fit Gaussian is depicted (green and magenta dashed lines). The overlap between the two Gaussian curves corresponds to 0.07% of the total number of cells presented here. The dotted cyan line depicts the bit-calling threshold. Cells above the bit-calling threshold were assigned a bit value of "0" for this bit while cells below the bit-calling threshold were assigned a bit value of "1". These distributions are representative of those observed for all other bits.

Error rate estimation

Next, we estimate our barcode misidentification rate using two different approaches. In the first approach, we estimate our error rate in barcode identification using the observation that 16% of the measured barcodes did not match the valid barcodes present in the library. We note that there are two types of errors. If the measurement error produces an invalid barcode that was not present in the library (type I error), this barcode read out error would be detected, and hence this type of error would not affect our accuracy in genotype identification. If, however, the measurement error produces a valid barcode that was present in the library (type II error), this error would not be detected and hence would cause a genotype misidentification. We recognize that the frequency of type I error occurrence is the product of the frequency of barcode error occurrence and the fraction of all possible 21-bit binary barcodes that are not present in the library. Since the frequency of type I error occurrence was measured to be 16% and 96% of all possible barcodes were not present in the library, the frequency of barcode error occurrence should be 16.7%. The frequency of type II error occurrence, which is the product of the frequency of barcode error occurrence (16.7%) and the fraction of all possible barcodes that are present in the library (4%), should then be only 0.67%. Hence our genotype misidentification rate was < 1%. This low error rate illustrates the benefit of our barcode bottlenecking strategy.

In the second approach, we verify our barcode measurement accuracy by taking the genotype determined from the phenotype measurement (presence and absence of mMaple3) as the ground truth and comparing this assignment to the genotype determined from the barcode measurement. To determine the phenotype of each cell, we normalized the mMaple3 fluorescence intensity of the cell measured with 560-nm illumination to the mTagBFP2

fluorescence intensity measured with 405-nm illumination (Fig. 2E) to remove differences in protein expression levels. We then calculated the median of the normalized brightness for all cells assigned to the same barcode and constructed a histogram of this normalized mMaple3 brightness for all barcodes associated with the mMaple3+ genotype as well as a histogram for all barcodes associated with the mMaple3- genotype. As expected, these two histograms are well separated with only a very small overlap (Fig. 2F). Next, we determined the fraction of barcodes that were misidentified by assuming that the overlap between the two histograms were solely due to barcode misidentification. To this end, we set a threshold based on the intersection point of the two histograms and assigned all cells with normalized mMaple3 brightness larger (or smaller) than this threshold as having a mMaple3+ (or mMaple3-) genotype. We then compared this genotype assignment to the genotype assignment based on the measured barcode and found that < 1% of genotype assignments disagree, which also suggests a <1% barcode misidentification rate. We note that this error rate is likely an overestimate since in addition to barcode misidentification, the natural spread in the intensity distribution of cells in each group should also contribute to the overlap of these distributions.

Given the excellent agreement between the barcode misidentification rate estimated by the two distinct methods described above, we conclude that our high-throughput imaging-based screening approach is capable of accurately determining the genotype in a large number of cells.

Chapter 2. Optimizing fluorescent protein photo-physics

YFAST introduction

To demonstrate the power of our approach for screening a large library of mutants to find proteins with desired properties, we screened for improved variants of a recently developed fluorescent protein, YFAST. YFAST is not itself fluorescent but becomes fluorescent upon binding to an exogenous, GFP-like chromophore, such as HMBR (Fig. 5A)[24]. By separating the chromophore from the protein tag, YFAST offers several potential advantages over traditional fluorescent proteins, such as the ability to make the sample fluorescent only when desired, the potential to sequentially image multiple proteins in the same color channel with different chromophores and different protein tags[24], as well as the potential to make an on-off switchable fluorescent protein through chromophore binding and unbinding. However, the relatively modest brightness and rapid photobleaching of YFAST limits the exploration of its full potential. For this reason, we sought to improve the YFAST brightness and its photobleaching kinetics. In particular, we observed that YFAST exhibited complex and reversible photobleaching behavior: (1) the photobleaching time course of YFAST shows a biphasic behavior with one decay component much faster than the other; and (2) after the illumination was stopped, the fluorescence of YFAST rapidly recovered to a substantial extent (Fig. 4).

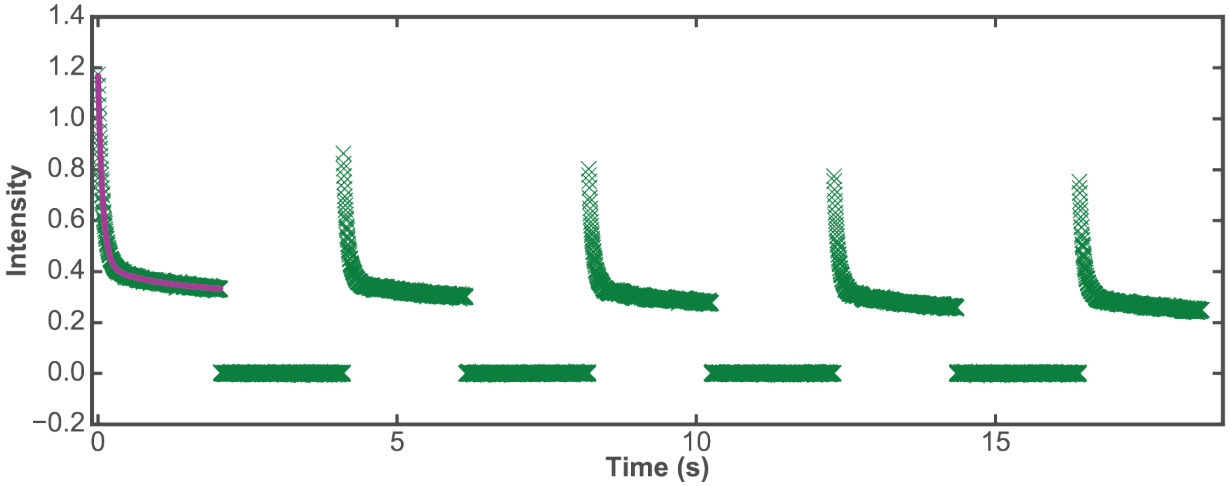

Figure 3. Reversible and biphasic photobleaching kinetics of YFAST. The normalized fluorescence intensity (green crosses) upon intermittent illumination with 488-nm light and a fit to a double exponential (purple line). E. coli expressing BFP-YFAST were adhered to a glass coverslip, immersed in 10 μM HMBR in PBS and imaged at 4-ms time resolution. The YFAST fluorescence intensity for each cell is normalized by the BFP fluorescence and averaged over multiple cells in the imaged area. The 488-nm illumination was switched on and off with a period of 2 seconds. Intensity values of zero represent the period of time when the illumination was off.

Screening strategy

We thus sought to identify YFAST mutants that are both brighter and more photostable. Specifically, because of the biphasic photobleaching behavior, we screened for mutants that exhibit a relatively large amplitude of the slow decay component as compared to the original YFAST, and preferably also with a slower decay rate of this component. We primarily focused on the amplitude and rate constant of the slow component as improving these properties will

benefit many imaging purposes. Additionally, we also reported the fast component amplitude in our library screens. We note that while brightness is a property that can be measured and selected via simple screening methods such as FACS, screening for photobleaching kinetics requires a time-resolved fluorescence measurement during screening, and, thus, would benefit from our image-based screening approach, especially given that YFAST exhibits a reversible photobleaching with complicated biphasic kinetics.

To screen for improved variants of YFAST, we adopted two basic strategies to design variants: a structurally naïve approach in which we systematically screened all single-point mutations and a structurally inspired approach in which we used the structure of a YFAST homology, Photoactive Yellow Protein[33], to focus multiple mutations near the chromophore region. These libraries were screened with our approach, as described below, and then we built additional libraries, iteratively, by combining beneficial mutations identified via these preliminary screens (See Online Methods for more detailed descriptions of the library designs). In total, we constructed and screened libraries containing ~60,000 unique YFAST variants associated with ~160,000 barcodes.

To measure the brightness of different YFAST variants while controlling for potential variations in the expression level, we fused YFAST variants to mTagBFP2 and imaged individual cells with both 405-nm and 488-nm illumination, respectively (Fig. 5A). The relative brightness of YFAST was calculated as the ratio of the YFAST fluorescence intensity measured with 488-nm illumination in the presence of the chromophore HMBR to the mTagBFP2 intensity measured with 405-nm illumination. To characterize the photobleaching kinetics of YFAST variants, we measured the intensity over time under constant 488-nm illumination alone with a 120-ms frame duration for at least 20 frames (Fig. 5B). For each mutant, we determined the two key parameters described earlier, the amplitude and rate constant of the slow bleaching component, for the purpose of screening for mutants that are brighter and more photostable than the original YFAST. In addition, we also determined the apparent amplitude of the fast bleaching component at our 120-ms time resolution. Because photobleaching did not go to completion during our measurements, we independently determined the background level (zero level) to better constrain the amplitude determination. The details of how these parameters were determined from the measured time courses are described in the following chapter. Under the illumination intensity used, faster (4-ms) time resolution measurements of the original YFAST showed that the decay rate constants for the fast and slow components were ~10 $s^{-1}$and ~0.1 s^{-1} respectively. Hence, we anticipate that the time-resolution, 120 ms, and duration, 2.5 s, of our library measurements, plus the independent determination of the background level, should allow us to determine the amplitude and rate constant of the slow bleaching component reliably. Though, we cannot rule out the possibility that beyond our measurement duration, YFAST displays a more complicated photodecay, in which case, our reported rate constant for the slow component

should be considered the initial decay rate of this component. On the other hand, given our limited time resolution, the apparent amplitude that we determined for the fast bleaching component is likely to systematically underestimate this amplitude. Nonetheless, it would still provide useful information for future imaging experiments using the YFAST variants at ~100 ms or slower time resolution. Because of the limited time resolution, we did not extract the rate constant of the fast bleaching component.

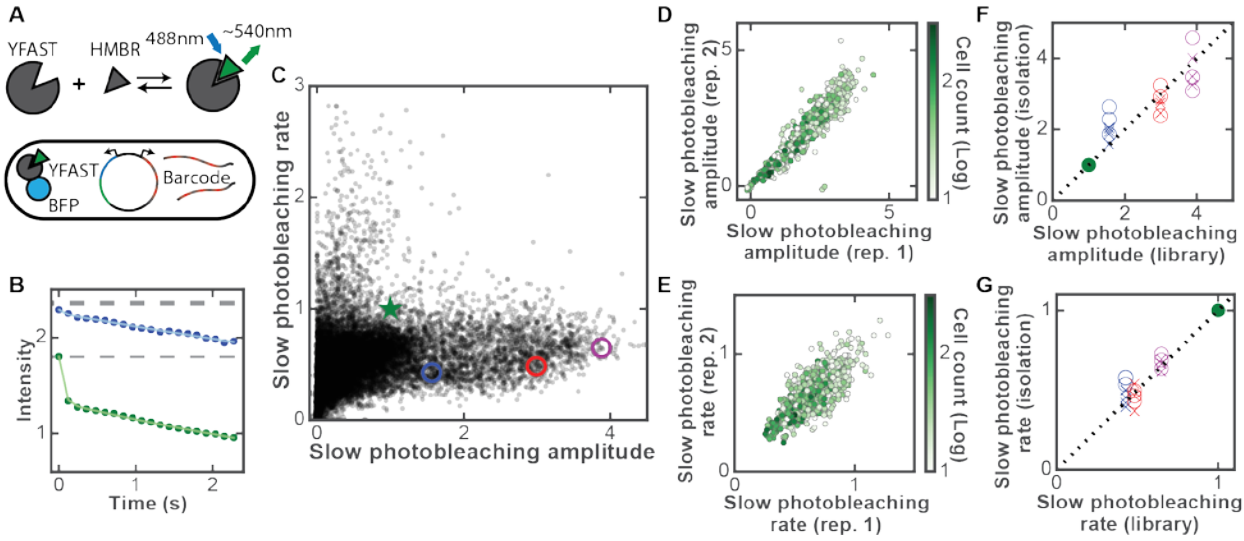

Figure 4. Screening YFAST mutant libraries for improved brightness and photobleaching kinetics by measuring 20 million cells containing 60,000 YFAST variants. (A) Schematic diagram of YFAST library design. YFAST is dark on its own, but it becomes fluorescent upon binding to a ligand such as HMBR24. A library of YFAST variants fused to mTagBFP2 for normalization is merged with a library of barcodes and transformed into E. coli cells. (B) The initial intensity (gray dashed line) and the photobleaching curve of the original YFAST (green circles) and a YFAST mutant (blue circles) measured from a single cell in the library screen measurement. The curve is fit to a double exponential decay (light green and blue line) with the background level independently determined by the intensity of cells that have dark YFAST variants (Online Methods). (C) Scatter plot of amplitudes and rate constants of the slow bleaching component for different YFAST mutants in all screened libraries. Each point depicts the median values of all cells associated with one mutant normalized to those of the original YFAST. Here only the mutants that contain at least 10 imaged cells are depicted. The library measurement results of the original YFAST and three selected mutants are indicated by the green star and colored circles, respectively. The blue, red and purple circles correspond to Mutant 1, Mutant 2, and Mutant 3 in Table 2, respectively. (D and E) Scatter plots of the amplitude (D) and rate constant (E) of the slow bleaching component for each mutant measured in two replicate library measurements containing a subset of the 60,000 mutants. Each point represents the median of all cells corresponding to each mutant and is colored by the minimum number of cells measured in either of the replicates. Only the mutants that contain at least 10 imaged cells in each replicate are depicted. (E) only contains mutants with a slow photobleaching amplitude that is at least half of that of the original YFAST. (F and G) The amplitudes (F) and rate constants (G) of the slow bleaching component for the original YFAST and the three selected mutants measured in isolation versus those measured in the library screen. Data from multiple replicates of isolation measurements conducted at the library-screen time resolution (120 ms; crosses) or at a higher time resolution (4 ms; circles) are shown. The values are normalized to those of the original YFAST. The mutants are depicted by the same color as those used in (C).

Screening results and validation

We screened a total of ~20 million cells containing the 60,000 YFAST mutants. After discarding cells by bit intensity thresholding and rejection of invalid barcodes, as described earlier, a total of ~6 million cells remained, which corresponds to ~100 cells per mutant on average. This oversampling was used to improve the phenotype measurement accuracy because we observed relatively large cell-to-cell variation of the quantities described above and the accuracy with these measurements increased with the number of cells measured per mutant. We then grouped cells based on the genotypes (YFAST mutants) measured and computed the median value of the three quantities mentioned above for each of these mutants (Fig. 5C and Fig. 6A). Across the ~60,000 variants of YFAST that were screened, we observed mutants that were substantially brighter and more photostable, with both larger amplitudes and slower decay rate constants of the slow bleaching component (Fig. 5C).

To test the accuracy with which we measured the phenotype for individual variants, we replicated the screen for a subset of our variants. We found that the measured amplitudes and rate constants are indeed reproducible between these replicate measurements (Fig. 5D, E and Fig. 6B).

Table 2. Sequences of the Isolated Mutants. A list of the amino acid sequences for the isolated YFAST mutants characterized in Figure 5 and Figure 5. The mutated residues are marked in bold.

Mutant name	Amino acid sequence
Mutant 1	MEHVAFGSEDIENTLAKMDDGQLDGLAFGAIQLDGDGNILQYNAAEGDI TGRDPKQVIGKN**L**FKDVA**C**GT**R**S**S**EFYGKFKEGVASGNLNTMFEWMIPT SRGPTKVKVHMKKALSGDSYWVFVKRV
Mutant 2	MEHVAFGSEDIENTLAKM**N**DGQLDGLAFGAIQLDGDGNILQYNAAEGDI TGRDPKQVIGKN**L**FKDVA**C**GT**R**S**S**EFYGKFKEGVASGNLNTMFEW**T**IPT **K**RGPTKVKVHMKKALSGDSYWVFVKRV
Mutant 3	MEHVAFGSEDIENTLAKMDDGQLDGLAFGAIQLDGDGNILQYNAAEGDI TGRDPKQVIGKN**L**FKDVA**E**GT**R**S**S**EFYGKFKEGVASGNLNTMFEW**T**IPT **K**RGPTKVKVHMKKALSGDSYWVFVKRV

To further confirm the accuracy of our library measurements, we selected a few improved variants that show both larger amplitudes and slower rate constants of the slow bleaching component (Table 2), and then cloned these YFAST variants and measured their properties in monoculture at the same time resolution as used in our library screen measurements. The results from these isolated mutant measurements are quantitatively comparable to the results that we obtained from the library screen for all three measured quantities, the amplitudes and rate constants of the slow bleaching component as well as the amplitude of the fast bleaching component (Fig. 5F, G and Fig. 6C). This agreement indicates that the massive scale of parallelization in our library measurement did not cause a substantial reduction in the measurement accuracy. Because the observed decay rates of the slow bleaching component of these YFAST variants are much slower than our measurement time resolution, the parameters that we determined for this component should reflect the true values for these variants. However, the near-zero values observed for the apparent amplitudes of the fast component for these mutants at our time resolution could be because the fast component was indeed diminished or because the decay rate of this component became much faster for these mutants, or both. To test these scenarios, we performed measurements of these mutants at a faster time resolution (4-ms) and found that the fast amplitudes of these mutants were indeed much smaller that of the original YFAST and, in addition, the decay rates of the fast component also became substantially faster (Fig. 6D, E). The combined effects of these changes produced the effective elimination of the fast component at 120-ms time resolution. As expected, for the two key parameters that we used to screened mutants, i.e. the amplitudes and rate constants of the slow component, the results obtained with the increased (4-ms) time resolution still agree with the results from our library screen measurements conducted at

lower time-resolution (Fig. 5F, G), supporting the appropriateness of the time resolution used in the screen for these parameters.

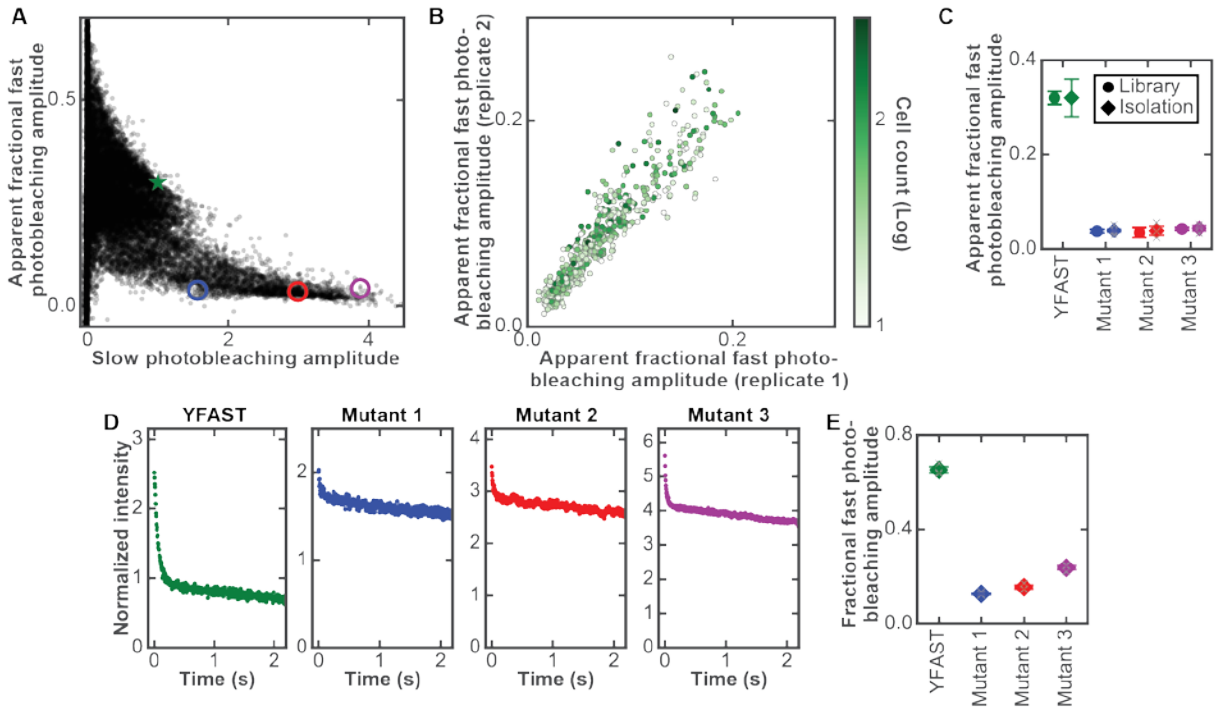

Figure 5. Additional quantifications of the YFAST library screen and comparisons between library measurements and isolated mutant measurements. (A) Scatter plot of the apparent fractional amplitude of the fast bleaching component and the amplitude of the slow bleaching component for different YFAST mutants in all of the libraries imaged. Each point depicts the median values of all cells associated with one mutant. The amplitudes of the slow bleaching component are normalized to that of the original YFAST. The apparent fractional amplitude of the fast bleaching component is defined as the amplitude of the fast bleaching component divided by the sum of the amplitudes of the fast and slow bleaching components. Here only the mutants that contain at least 10 imaged cells are depicted. The library measurements of the original YFAST and three selected mutants are indicated by the green star and colored circles, respectively. The color scheme is as in Fig. 3. (B) Scatter plot of the apparent fractional amplitudes of the fast bleaching component for each mutant measured in two replicate library measurements containing a subset of the 60,000 mutants. Each point represents the median of all cells corresponding to each mutant and is colored by the minimum number of cells measured in either replicate. Only the mutants with a slow component at least half as bright as the slow component for the original YFAST and that contain at least 10 imaged cells in each replicate are depicted. (C) The apparent fractional amplitude of the fast bleaching component for the original YFAST and the selected mutants (color coded as in (A)) determined from the isolated mutant measurements shown together with the results obtained from the library measurements. Solid circles represent the median from the library measurements (N=420, 136, 32, 220 cells for the original YFAST, mutant 1, mutant 2, and mutant 3, respectively). Diamonds represent the mean from the isolated mutant measurements (N = 12, 5, 3, and 3 replicates for original YFAST, mutants 1, mutant 2, and mutant 3, respectively).

(continued) Error bars represent SEM and individual measurements are depicted as gray crosses when N is less than 10. (D) The fluorescence intensity as a function of illumination time for the original YFAST and the three selected mutants (color-coded as in (A)) measured in isolation at the 4-ms time resolution. The fluorescence intensity is averaged over many cells. (E) The fractional amplitude of the fast bleaching component for the original YFAST and the selected mutants (color coded as in (A)) determined from the isolated mutant measurements shown in (D). Diamonds depict the mean and error bars represent SEM. Individual measurements are depicted as gray crosses (N = 3 replicates for the original YFAST and the three mutants). The amplitudes and rate constants of both fast and slow bleaching components are determined from these fast-time resolution measurements by a double exponential fit. The fractional amplitude of the fast bleaching component is shown in the plot here, the rate constants of the fast bleaching component are 13 ± 1 s, 11± 3 s, 34 ± 2 s, 48 ± 2 s for the original YFAST, mutant 1, mutant 2, and mutant 3 respectively (Mean ± SEM, N=3 replicate measurements), and the amplitudes and rate constants of the slow bleaching component are shown in Fig. 3F,G.

Chapter 3: Material and methods

Barcode library assembly

The barcode library consists of a set of plasmids, each containing a DNA barcode sequence that encodes a RNA designed to represent a single *N*-bit binary word. The expression of the RNAs is controlled by the lpp promoter. Every barcode in the library has *N* readout sequences, one corresponding to each bit, designed to be read out by hybridizing fluorescent probes with the complementary sequence. Although 22 bits are present in the barcode set that was constructed here, to reduce the number of hybridization rounds, experiments were conducted by reading out either 21 or 18 of the possible bits, depending on the library size. For each bit position, we assigned one 20-mer sequence to encode a value of "0" and another 20-mer sequence to encode a value of "1". To increase the rate of hybridization, these encoding sequences were constructed from a three-letter nucleotide alphabet, one with only A, T, and C, in order to destabilize potential secondary structures[34]. The utilized sequences were drawn from those previously used for MERFISH[26] with additional sequences designed using approaches described previously[26]. For each barcode, the bits are concatenated with a single G separating each.

We assembled this barcode library by ligating a mixture of short, overlapping oligonucleotides, each representing a pair of adjacent bits. For each pair of adjacent bits,

there are four unique combinations of bit values ("00", "01", "10", and "11"). Each corresponding sequence was synthesized as a single-stranded oligo. These oligos were then ligated to form complete, double-stranded barcodes that contain concatenated sequences of all bits with all possible bit values. For the ligation step, all oligos were mixed and diluted so that each oligo was present at a concentration of 100 nM. The mixture was phosphorylated by incubation with T4 polynucleotide kinase (16 µL oligo mixture, 2 µL T4 ligase buffer, 2 µL PNK [NEB, M0201S]) at 37 °C for 30 minutes and ligated by adding 1 µL T4 ligase (NEB, M0202S) and incubating for 1 hour at room temperature.

To prepare a plasmid library containing these barcode sequences along with the desired promoter, we diluted the ligation product 10-fold and amplified it by limited-cycle PCR on a Bio-Rad CFX96 using Phusion polymerase (NEB, M0531S0) and EvaGreen (Biotium, 3100). The PCR product was run in an agarose gel, and the band of the expected length was extracted and purified (Zymo Zymoclean Gel DNA Recovery Kit, D4002). The purified product was inserted by isothermal assembly[35] for 1 hour at 50 °C (NEB NEBuilder HiFi DNA Assembly Master Mix, E2621L) into a plasmid backbone fragment containing the colE1 origin, the ampicillin resistance gene, and other elements taken from the pZ series of plasmids[36]. The assembled plasmids were purified (Zymo DNA Clean and Concentration, D4003), eluted into 6 µL water, mixed with 10 µL of electro-competent *E. coli* on ice (NEB, C2986K), and electroporated using an Amaxa Nucleofector II. Immediately after electroporation, 1 mL SOC was added and the culture was incubated at 37 °C on a shaker for one hour. Subsequently, the SOC culture was diluted into 50 mL of LB (Teknova, L8000) supplemented with 0.1 mg/mL carbenicillin (ThermoFisher, 10177-012) and placed on the

shaker at 37 °C overnight. The following day, the culture was miniprepped (Zymo Zyppy Plasmid Miniprep Kit, D4019), yielding the complete barcode library.

Assembling protein mutant libraries

To create a library of mutant proteins, short nucleotide sequences containing regions of the protein with the desired mutations were synthesized as complex oligonucleotide pools. To then create the desired mutant genes from these pools, we amplified the pool and its corresponding expression plasmid via limited cycle PCR and assembled these fragments using isothermal assembly[35]. The expression backbone was derived from the colE1 origin and the chloramphenicol resistance gene from the pZ series of plasmids[36]. Oligo pool synthesis is prone to deletions, which could lead to frameshift mutations that produce non-viable proteins. To remove these variants prior to measurement, the protein variants were translationally fused upstream to the chloramphenicol resistance protein. These constructs were electroporated into *E. coli*, as described above, and these cultures grown in the presence of chloramphenicol to select only for protein variants that did not have frame-shift mutations and which could, thus, translate competent chloramphenicol resistance. These plasmids were re-isolated via plasmid miniprep and the genetic variants extracted via PCR prior to combination with the barcode library.

Merging mutation libraries with the barcode library

To merge a mutant library with the barcode library, the corresponding halves of each plasmid library were amplified by limited-cycle PCR. Of note, the forward primer for amplifying the barcode library contained 20 random nucleotides so that each assembled plasmid contained a 20-mer unique molecular identifier (UMI)[37,38]. Also, the protein mutant half contained the plasmid's replication origin (colE1) while the barcode half contained the ampicillin

resistance gene ensuring that only plasmids containing both halves were competent. The two halves were assembled by isothermal assembly and transfected into electrocompetent *E. coli* as described earlier. After incubating in SOC for 1 hour at 37 °C, the culture was again diluted into 50 mL LB and grown until it reached an optical density at 600 nm (OD600) of ~1. To limit the possibility that a single bacterium had taken up more than one plasmid, plasmids were extracted again from this culture and reinserted at a concentration where the number of *E. coli* cells significantly outnumbered the number of plasmids. Specifically, 2 µL of the plasmid library at 100pg/µL was re-electroporated into 10 µL of fresh electro-competent *E. coli*. This culture was then grown and diluted to a concentration of ~1000 cells/µL by using the OD600 to determine the number of cells in the culture and, thus, the appropriate dilution. From the diluted culture, a volume containing the desired number of cells, and hence the desired number of unique barcode-mutant pairs, was inoculated into a new culture. This culture was incubated at 37 °C overnight and the following day it was archived for future imaging experiments by diluting 1:1 in 50% glycerol (Teknova, G1796), separating into 100 µL aliquots, and storing at -80 °C. The remaining culture was mini-prepped to use as a PCR template for constructing the barcode to genotype lookup table.

Constructing the barcode-to-genotype lookup table

Since barcodes and gene variants were assembled randomly, next generation sequencing was used to construct a look-up table that links barcodes to their corresponding gene variant. The total length of the combined sequence of the gene variant and the barcode exceeded the read length of the sequencing platform used (Illumina MiSeq). To circumvent this challenge, multiple fragments were extracted from each library, sequenced independently and grouped computationally using the UMI.

The mini-prepped libraries were prepared for sequencing by two sequential limited-cycle PCRs. The first PCR extracted the desired region while adding the sequencing priming regions, and the second PCR added multiplexing indices and the Illumina adapter sequences. Between PCRs, the product was purified in an agarose gel and the final product was gel purified prior to sequencing.

For each sequencing read, the corresponding barcode or gene variant sequence was extracted. The reads were then grouped by common UMI, and the most frequently occurring barcode and gene variant seen for each UMI was assigned to that UMI, constructing the barcode-to-gene variant lookup table for every variant in the library. This analysis was conducted in custom software written in Matlab.

Library design of YFAST variants

Since YFAST is a recently developed fluorescent protein, the consequences of mutating different regions of the protein are not well characterized in the literature. Hence, we began our screen by concurrently designing libraries following two distinct strategies. In the first strategy, we took a structurally naïve view and constructed a library (library type 1, LT1) that consists of mutants corresponding to all possible single amino acid substitution, insertion, and deletion at each location within YFAST. The second strategy made use of structural information of the YFAST precursor, Photoactive Yellow Protein (PYP) (PDB: 1NWZ)[33] to target residues adjacent to the chromophore (library type 2, LT2-1), introducing up to 6 amino-acid substitutions per mutant. We screened these libraries using our screening method. Since many of the mutants in LT2-1 appeared dark, we refined the selection of mutations by redesigning the oligo pool to only include those amino-acid substitutions that appeared bright with relatively high frequencies in the LT2-1 library and created another library (LT2-2) that

combined these substitutions, containing up to 6 substitutions per mutant. We then screened this library with our method as well. We then created a library (library type 3, LT3) by combining mutations found to have favorable brightness and photostability (i.e. relatively large amplitude of the slow bleaching component) in LT1 with those mutations found to have favorable brightness and photostability in all LT2. Each variant in LT3 contains up to 10 mutations. We then screened LT3 and identified a mutant with 6 amino acid substitutions that is particularly photostable with a large amplitude of the slow bleaching component and a nearly eliminated the fast component at our measurement time resolution. Next, to further improve the fluorescent properties of this mutant, we created a new library (library type 4, LT4) that contained all possible single amino acid substitution, insertion, and deletion at every residue of this mutant. Finally, based on the screening results of LT4, we created the library type 5 (LT5) by splitting the entire protein sequence into 6 regions, selecting LT4 mutations with favorable brightness and photostability in each region, and creating all possible combinations of these mutations. LT5 contains 6-12 mutations per library member.

Some of the above libraries were constructed and measured concurrently while we were developing and optimizing our screening protocol. Therefore, we re-measured all of the libraries again, by mixing them into pools containing ~25,000 barcodes each. Instead of combining all libraries into a single pool and measuring a very large number of cells in a single screen over a long time, we opted to split the measurements into smaller pools and measured 1-2 million cells per experiment. Since the phenotype accuracy increases with the number of cells measured, we also included the results from the earlier measurements of individual libraries that were performed using the optimize protocol. Fig. 5 and Fig. 6 contain results from all library measurements performed with the optimized protocol.

Phenotype and barcode imaging

Each library was prepared for imaging by thawing the 100 µL aliquot from -80 °C to room temperature and diluting into 2 mL LB supplemented with 0.1 mg/mL carbenicillin. Imaging coverslips (Bioptechs, 0420-0323-2) in 60-mm-diameter cell culture dishes were prepared by covering them in 1% polyethylenimine (Sigma-Aldrich, P3143-500ML) in water for 30 minutes followed by a single wash with phosphate buffered saline (PBS). The *E. coli* culture was diluted 10-fold into PBS, poured into the culture dish, and spun at 100g for 5 minutes to adhere cells to the surface.

The sample coverslip was assembled into a Bioptech's FCS2 flow chamber. A peristaltic pump (Gilson, MINIPULS 3) pulled liquid through the chamber while three computer-controlled valves (Hamilton, MVP and HVXM 8-5) were used to select the input fluid. The sample was imaged on a custom microscope built around a Nikon Ti-U microscope body with a Nikon CFI Plan Apo Lambda 60x oil immersion objective with 1.4 NA. Illumination was provided at 405, 488, 560, 647, and 750 nm using solid-state single-mode lasers (Coherent, Obis 405nm LX 200mW; Coherent, Genesis MX488-1000; MPB Communications, 2RU-VFL-P-2000-560-B1R, MPB Communication, 2RU-VFL-P-1500-647-B1R; and MPB Communications, 2RU-VFL-P-500-750-B1R) in addition to the overhead halogen lamp for bright field illumination. The Gaussian profile from the lasers was transformed into a top-hat profile using a refractive beam shaper (Newport, GBS-AR14). The intensity of the 488-, 560-, and 647-nm lasers was controlled by an acousto-optic tunable-filter (AOTF), the 405-nm laser was modulated by a direct digital signal, and the 750-nm laser and overhead lamp were switched by mechanical shutters. The excitation illumination was separated from the emission using a custom dichroic (Chroma,

zy405/488/561/647/752RP-UF1) and emission filter (Chroma, ZET405/488/461/647-656/752m). The emission was imaged onto an Andor iXon+ 888 EMCCD camera. During acquisition, the sample was translated using a motorized XY stage (Ludl, BioPrecision2) and kept in focus using a home-built autofocus system.

Phenotype measurements were conducted immediately after cells were deposited onto the coverslip, inserted into the flow chamber, and immersed in PBS. For imaging *E. coli* cells expressing mMaple3-mTagBFP2 fusion or mTagBFP2 alone, an image was first acquired for 1 frame with 405-nm illumination to excite mTagBFP2 at a frame rate of 8.4 Hz (120 ms), followed by illumination with 405-nm light for 30 additional frames at 8.4 Hz to photoactivate mMaple3. Then an image was acquired with 560-nm illumination for 1 frame to detect mMaple3 fluorescence. For imaging *E. coli* cells expressing the YFAST mutants, images were first acquired in the absence of the chromophore with 405-nm illumination for 1 frame to measure the mTagBFP2 fluorescence to determine the position of each cell followed by an image with bright-field illumination for alignment between multiple imaging rounds. Then 10 μM of the chromophore HMBR (synthesized as described previously[24]) in PBS was flowed over the cells and a fluorescence image was acquired with 488-nm illumination for 1 frame to measure YFAST intensity, and a bright-field image was acquired again for alignment, followed by at least 20 frames at 8.4 Hz with constant 488-nm illumination to measure the decrease in intensity upon photobleaching. Since 8.4 Hz is the full field frame rate of the camera that we used, increasing the time resolution would require imaging a smaller field of view per frame and hence a reduction in the measurement throughput. Images were acquired at thousands of locations in the sample, each corresponding to a ~200x200 μm^2 field-of-view. All fields were imaged prior to the addition of the chromophore

to determine the position of each cell, and then after the chromophore was added, all of the subsequent exposure sequence described above was completed at each field prior to moving to the next. The illumination intensities at the back-focal plane used in these experiments were 1 W/cm^2, 3 W/cm^2, and 10 W/cm^2 for the 405-nm, 488-nm, and 561-nm lasers, respectively. Following the phenotype measurement, the cells were fixed by incubation in a mixture of methanol and acetone at a 4:1 ratio for 30 minutes. To prevent salts from precipitating and clogging the flow system, water was flowed before and after the fixation mixture. Once fixed, the cells were washed in 2x Saline Sodium Chloride (SSC) and hybridizations for MERFISH imaging began.

To determine the RNA barcode expressed within each cell, we performed multiple rounds of hybridizations. For each hybridization round, the sample was incubated for 30 minutes in hybridization buffer [2xSSC; 5% w/v dextran sulfate (EMD Millipore, 3730-100ML), 5% w/v ethylene carbonate (Sigma-Aldrich, E26258-500G), 0.05% w/v yeast tRNA, and 0.1% v/v Murine RNase inhibitor (NEB, M0314L)] with a mixture of readout probes labeled with either ATTO565, Cy5, or Alexa750 (Bio-Synthesis Inc) each at a concentration of 10 nM. In the readout probes, the dyes were linked to the oligonucleotides through a disulfide bond[26]. Then, the hybridization buffer was replaced by an oxygen-scavenging buffer for imaging[39] [2xSSC; 50 mM TrisHCl pH 8, 10% w/v glucose (Sigma-Aldrich, G8270), 2 mM Trolox (Sigma-Aldrich, 238813), 0.5 mg/mL glucose oxidase (Sigma-Aldrich, G2133), and 40 μg/mL catalase (Sigma-Aldrich, C100-500mg)]. Each position in the flow cell was imaged with 750-, 647-, and 560-nm illumination from longest to shortest wavelength followed by bright-field illumination for alignment before continuing to the next location. Following the imaging of all regions, the disulfide bonds linking the dyes to the oligonucleotides in the

readout probes were cleaved by incubating the sample in 50 mM tris (2-carboxyethyl)phosphine (TCEP; Sigma-Aldrich, 646547-10X1ML) in 2xSSC for 15 minutes. The sample was then rinsed in 2xSSC and the next hybridization round started. For each round of hybridization, three readout probes with spectrally discernable dyes (ATTO565, Cy5, and Alexa750) were hybridized simultaneously as described above (see Table 1). Altogether, with 14 hybridization rounds, all 42 readouts corresponding to 21 bits were measured in 40 hours. For smaller libraries, the imaging area was reduced, and the number of hybridization rounds was decreased to 12 (for 18-bit readout), reducing the measurement time to 22 hours.

Image analysis

To correct for residual illumination variations across the camera, a flat field correction was performed as follows. Every image was divided by the mean intensity image for all images with the given illumination color. Then, the images for different rounds corresponding to the same region were aligned using the image acquired under bright field illumination by up-sampled cross-correlation, creating a normalized image stack of all images at each position in the flow chamber. If the radial power spectral density of any given bright field image did not contain sufficient high frequency power, the image was designated as out-of-focus and all images for the corresponding region were excluded from further analysis.

To extract cell intensities, the edges of each cell were detected using the Canny edge detection algorithm on the image acquired with 405-nm illumination for mTagBFP2 imaging. The edges that formed closed boundaries were filled in and closed regions of pixels were extracted. If a given closed pixel region had a filled area of more than 20 pixels and the ratio of the filled area to the area of the convex hull was greater than 0.9, it was classified as a cell.

To increase the cell detection efficiency, the detected cells were then removed from the binary image, the image was dilated, filled, and eroded and cells were extracted again. This allowed cells where gaps exist in the detected edges to still be detected. For each cell, the mean intensity was extracted for the corresponding pixels in every image.

From the cell intensities, the phenotypes and barcodes were calculated. For each measured readout sequence, the measured intensity was normalized by subtracting the minimum and dividing by the median signal observed for that readout sequence across all cells. To determine whether a barcode contained a “1” or a “0” at each bit, the measured intensities of the “1” readout sequence and the “0” readout sequence for that bit were compared. Specifically, a threshold was selected on the ratio of these two values. If this ratio was above the threshold, the bit was called as a “1”. Otherwise, the bit was called as a “0”. Because the “1” and “0” readout sequences associated with each bit might be assigned to different fluorophores, it was necessary to optimize this threshold for each bit individually. This optimization was performed by randomly selecting 150 barcodes (a training set) from the set of known barcodes that were determined to be present in the library by sequencing. An initial set of thresholds was selected and the fraction of cells matching these barcodes was determined. The threshold for each bit was then varied independently to identify the threshold set that maximizes this fraction. This optimized threshold set was then used for determining the bit values for all cells.

Once the barcode was determined for each cell, cells were grouped by barcode and the median of the various phenotype values was computed to determine the measured phenotype for the genotype corresponding to that barcode. For the mMaple3 measurement, the normalized brightness was determined from the ratio of the mMaple3 intensity under 560-nm

illumination to the mTagBFP2 intensity under 405-nm illumination, as discussed above. For YFAST measurements, the normalized intensity was determined by the ratio of the YFAST fluorescence intensities under 488-nm illumination in the presence of the YFAST chromophore HMBR to the mTagBFP2 fluorescence intensities under 405-nm illumination. To account for the fluorescence background present in *E. coli* upon 488-nm illumination, the background was independently determined and subtracted before calculating the fluorescence ratio. The background was estimated by calculating the median intensity of all cells upon 488-nm illumination predicted to contain a non-fluorescent YFAST mutant. Specifically, cells, grouped by barcode, were assigned to the non-fluorescent population if the Pearson correlation coefficient between the fluorescence intensity measured under 488-nm illumination (YFAST channel) and those measured under 405-nm illumination (mTagBFP2 channel) for the grouped cells fell below a threshold of 0.2. Since the YFAST variant is translationally fused to mTagBFP2, when the two intensities are uncorrelated, it suggests that the number of YFAST proteins in the cells does not affect the brightness of the cell and hence the YFAST associated with that barcode should be dark.

Our initial high-time resolution (4-ms) measurements of the original YFAST variant revealed a biphasic decay of fluorescence with time (Fig. 4). To quantify this behavior, we fit the background-subtracted photobleaching curve, $b(t)$, to the sum of two exponentials:

$$b(t) = p_{\mathrm{fast}}\, \mathrm{e}^{-At} + p_{\mathrm{slow}}\, \mathrm{e}^{-Bt}$$

where p_{fast} and A represent the amplitude and decay rate constant for the fast photobleaching component and p_{slow} and B represent the corresponding values for the slow photobleaching component.

This double-exponential decay function was also used to characterize our library screen measurements. However, to increase the throughput of our screens, we utilized the full imaging frame of our camera, which required the use of a slower frame rate (8.4 Hz, ~120 ms). This frame rate was comparable to the decay rate observed for the fast component of the original YFAST variant; thus, we did not anticipate that the rate constant associated with the fast component would be well constrained by this double-exponential fit. To address this problem, we initially fixed the rate constant of the fast component to the value determined from the original YFAST and allowed the other three parameters to vary in the fit. The time resolution of our library measurements was much higher than the decay time constant of the slow component; thus, the parameters associated with the slow component, p_{slow} and B, were well constrained by this fit—a point confirmed by our observation that p_{slow} and B did not change appreciable (by <0.5%) when we varied the fixed value of A over a wide range or let A also being a fitting parameter. To estimate the fast component amplitude, p_{fast}, we utilized the well constrained value of the slow component amplitude, p_{slow}. Specifically, we calculated p_{fast} from the difference of the initial brightness of each variant ($p_{\mathrm{fast}} + p_{\mathrm{slow}}$) and the fit value for the slow component amplitude, p_{slow}.

The reported values for the slow component amplitude and decay rate are normalized to the corresponding values measured for the original YFAST, unless otherwise mentioned. The fast photobleaching component amplitude was not normalized in this fashion but rather was reported as the fraction of the total brightness, i.e. [$p_{\mathrm{fast}}/\ (p_{\mathrm{slow}} + p_{\mathrm{fast}})$], which we termed the fractional fast photobleaching amplitude.

This analysis was conducted in custom software written in Python.

Chapter 4. Concluding remarks

Considerations when designing a high-throughput screen

There are several issues that should be considered when designing a high-throughput screen, including, for example, the number of bits in the barcodes, the fraction of possible barcodes used, and the number of cells that should be measured per variant to allow phenotypes to be measured accurately. Here we summarize some important points that should be considered when designing a screen to measure the phenotypic variability within a given library.

Bottlenecking barcodes. We only include a small fraction of all possible *N*-bit binary barcodes in a library, and this bottlenecking strategy serves two purposes: (i) to limit the frequency with which the same barcode might be associated with two different genetic variants and (ii) to introduce an error robustness into our barcode-to-genotype identification process. In our construction of the barcoded genetic variants, barcodes are associated with individual genetic variants randomly, hence the probability that a given barcode could be assigned to multiple different genetic variant could be high. While this situation would be detected via next-generation sequencing when we build the barcode-to-genetic variant lookup table, these barcodes would have to be discarded from the library screen measurement since cells containing such barcodes could not be unambiguously assigned to a given genotype. If a large fraction of the used barcodes were associated with multiple genetic variants, the number of barcodes that would need to be discarded would be high. To overcome this problem, we restricted the number of barcodes used in the library to be less than 10% of the total number of possible *N*-bit binary barcodes. Specifically, after the barcoded genetic variants are assembled, we

bottleneck the size of the barcoded genetic variants library such that the number of genetic variant-barcode pairs in the library is <10% of the total number of possible *N*-bit binary barcodes. Because only such a small fraction of barcodes is included, most barcodes will be present only once in the library, and the chance that a barcode is present more than once (hence allowing the possibility of being paired with more than one genetic variant) will be very small (<10%). The remaining small fraction of barcodes that are paired with more than one variant can be detected by sequencing and discarded in further analysis.

The second reason why we bottleneck is to introduce error robustness into our genotype identification process. Specifically, if only a relatively small fraction of all possible barcodes are used, barcode measurement errors will more likely produce a barcode that is not present in the library, i.e. an invalid barcode. Because we know the exact barcodes that are present in the library via next-generation sequencing, it is possible to identify the invalid barcodes that resulted from errors during barcode imaging and discard them. This ability greatly reduces the rate at which we misidentify the genotype of a given cell. If we bottleneck the barcode number such that <10% of the total possible barcodes are present in the library, the chance that a barcode imaging error will lead to genotype misidentification will be reduced to <10%.

In our experiments, we chose a degree of bottlenecking such that only 1-10% of the possible 21-bit binary barcodes are present in our libraries. The bottlenecking was achieved experimentally by selecting a small, random subset of cells after transforming *E. coli* cells with the barcode-mutant plasmids under the condition that each cell contains a unique barcode-mutant pair (see Online Methods). For example, to achieve a bottlenecking degree of 4%, we select the number of cells that is 4% of the number of possible 21-bit binary barcodes.

Determining the number bits in the barcodes. The number of bits in the barcode is determined by the number of gene variants that we need to screen. While optimizing YFAST, we created mutant libraries in two ways: (1) The first type of libraries contained a defined, relatively small number of

mutants that we hope to screen exhaustively; (2) The second type of libraries contained a very large number of possible mutants where screening only a random subset of these mutants would already be very informative. When we created the first type of libraries, we chose a barcode diversity such that the number of barcodes in the library was 5 times more than the number of unique mutants to ensure that each mutant (or at least the vast majority of them) was present in the library at least once. Because of the bottlenecking strategy describe above, namely the number of barcodes in the library being < 10% of the total number of possible *N*-bit binary barcodes, we then needed the total number of possible barcodes to be 50 times more than the number of mutants to screen. Based on this number, we determined the desired number of bits. For example, if we plan to screen 20,000 specific mutants, we will need more than 1 million possible barcodes, and hence we would use a 21-bit barcoding scheme that can give ~2 million possible barcodes. When we created the second type of libraries in which only a subset of possible mutants will be screened, we selected a library size to be equal to the number of mutants that we intended to subsample from the larger library; in this case, each mutant in the library was only associated with a single barcode and the number of barcodes in the library was equal to the number of mutants to be screened. The number of possible *N*-bit binary barcodes and hence the number of bits required is then likewise determined based on the bottlenecking strategy.

Determining the desired number of measured cells per genetic variant. In the library screens, the number of cells that needs to be measured for each genetic variant is largely determined by the accuracy of the phenotype measurement. As the number of cells measured for each genotype increases, the accuracy with which that phenotype is measured improves. The desired cell number per genetic variant is set by the noise properties of the screened phenotype and the measurement accuracy that is needed to discriminate phenotype variations.

For our screen of YFAST variants, we observed a large cell-to-cell variance in the fluorescence intensity measurements between cells expressing the same genotype. This variance was observed even within a monoculture of the original YFAST. This observation indicated that the measurement

accuracy of this type of phenotype from a single cell was low and, thus, required us to screen many more cells than mutants to increase this accuracy. In addition, we found that different mutants appear in different abundance within our libraries, and this natural variation arose because of the random processes of constructing the plasmid-mutant libraries and transforming *E. coli*. To ensure that the majority of mutants are measured with a desired number of cells, this abundance variation further increases the oversampling requirement. For the YFAST measurements, we aimed to measure ~100 cells on average per mutant.

Finally, in our genotype (barcode) measurements, a substantial fraction of the cells are discarded by readout intensity thresholding and by the rejection of barcodes that do not match the valid barcodes present in the library, as described in the main text. In our measurements, ~66% of the measured cells were discarded because of this. As a result, we needed to measure, on average, 300 cells per YFAST variant to achieve of the goal of ~100 cells per mutant. We therefore measured 20 million cells to screen 60,000 YFAST variants.

Since the noise properties of the screened phenotype and the measurement accuracy that is needed to discriminate phenotype variations both depend on the phenotype to be screened, the number of cells that needs to be measured per genotype depends critically on the phenotype to be screened. Thus, we recommend that pilot measurements be conducted to determine the noise observed for the desired phenotype. With these measurements, it should then be possible to estimate the number of cells required to discriminate different phenotypes to a given accuracy. Also, it is worth noting that given the reproducibility between phenotypes measured for the same genotype in separate screens, it may also be possible to increase the number of cells measured on average for a given library by simply replicating the screen multiple times with the same library and pooling the results so as to improve the accuracy of phenotype variability if it is determined not to be sufficient from a single measurement.

Estimate of the maximum plausible library size of the genetic variants

There are multiple factors that determine the maximum library size of genetic variants that can be screened. The first potential limitation to the size of the library is the number of unique barcodes that can be measured. We have demonstrated the ability to image 21-bit barcodes, and we did not observe a degradation in the image quality between the last imaged bit and the first imaged bit. Thus, we envision that adding more bits to the barcode should be possible. For example, 25-bit barcodes are likely readily measurable. Moreover, given such a modest extension in the length of the barcode, we envision that there will be no challenges to constructing plasmids that contain 25-bit barcodes, or creating the barcode-mutant lookup table using existing next-generation sequencing approaches (Illumina HiSeq or NovaSeq). 25-bits would produce ~30 million possible barcodes. Based on our bottlenecking strategy, we typically select <10% of the possible barcodes to include in the library, which means <3 million barcodes to include in the library. By utilizing high-competency *E. coli* strains, as we have done here, it should be possible to create 10-fold more transformants than library members, a sufficient coverage level, by pooling a few transformation reactions. If we aim to see each mutant (or the vast majority of them) at least once, we would like to have 5 times more barcodes in the library than the number of genetic variants, which, in this case, means the library could contain up to ~600,000 genetic variants. Assuming that we would like to measure 10-100 cells per variant on average (depending on the phenotype measurement accuracy requirement), and based on our current settings of the barcode readout intensity threshold, in which 1/3 of cells pass the threshold and generate correct barcodes, we would need to measure ~18 - 180 million cells. In the measurements we demonstrated here, we characterized ~1-2 million *E. coli* cells in a 40-hour long screen. However, in these measurements we used a relatively low density of *E. coli* on our coverslips so as to minimize the chance of cells contacting each other. This density could be increased as high as 10-fold without producing substantial cell-cell contact, and improvement in cell-segmentation algorithms should also allow contacting cells to be properly segmented. Thus, we anticipate that it should be possible to measure ~18 – 180 million cells with a reasonable imaging time (2-18 days).

Moreover, there are multiple ways that our protocols could be modified so as to further increase throughput. For example, improved hybridization approaches might reduce the number of dim or dark cells, allowing more of the measured cells to be utilized in the screen. Alternatively, it may also be possible to use lower magnification objectives to measure much larger fields of view and hence allow substantial improvements in the measurement throughput. We can use the low magnification for genotype (barcode) imaging while keeping the use of high magnification for the high-resolution phenotype measurements because the phenotype measurements are typically fast and the total imaging time of the screen is dominated by barcode imaging which requires many rounds of hybridization. In parallel, we anticipate that it should also be possible to increase the number of barcodes by either increasing the number of bits in the binary barcode scheme or by using higher order barcoding schemes, such as ternary or quaternary schemes. Thus, we anticipate that further advances in methodology could extend the throughput of our image-based screening method substantially.

Conclusion

In summary, we developed a method for image-based screening of large, pooled genetic variant libraries by co-expressing the genetic variants and the barcodes that can identify these genetic variants in cells, and determining both the phenotypes of the genetic variants and the barcodes in the same cells using imaging. By reading out barcodes using massively multiplexed FISH, we demonstrated the ability to screen hundreds of thousands of barcodes that correspond to tens of thousands of unique genetic variations. Using this approach, we identified mutations in the YFAST protein, a recently discovered ligand-dependent fluorescent protein, with improved brightness and photostability. Compared to previous screening methods that have been developed to improve the time dependent properties of fluorescent proteins, which typically select a small number of winning mutants for

sequencing and genotype identification[9,12], our method based on in situ imaging for both phenotype measurements and genotype identification allows the genotype-phenotype correspondence to be determined for all examined mutants. Because a standard fluorescence microscope is used for both phenotype and genotype measurements, and because the barcoding scheme is independent of the specific genotypes and phenotypes to be probed, we envision that our method can be extended, with simple adaptations, to measure a broad range of cellular phenotypes in response to a wide variety of genetic variations, ranging from mutations of single proteins to inhibitions and activations of genes. We thus expect that this high-throughput, image-based screening method can be applied broadly to improve existing properties or identify new properties of proteins and nucleic acids, as well as to decipher the roles of genes and gene networks on cellular behaviors at the genomic scale.

References

1. Zhang J, Campbell RE, Ting AY, Tsien RY. Creating new fluorescent probes for cell biology. *Nat Rev Mol Cell Biol*. 2002;3(12):906-918. doi:10.1038/nrm976.

2. Bradbury ARM, Sidhu S, Dübel S, McCafferty J. Beyond natural antibodies: the power of in vitro display technologies. *Nat Biotechnol*. 2011;29(3):245-254. doi:10.1038/nbt.1791.

3. Keefe AD, Pai S, Ellington A. Aptamers as therapeutics. *Nat Rev Drug Discov*. 2010;9(7):537-550. doi:10.1038/nrd3141.

4. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet*. 2015;16(5):299-311. http://dx.doi.org/10.1038/nrg3899.

5. Cadwell RC, Joyce GF. Randomization of genes by PCR mutagenesis. *PCR Methods Appl*. 1992;2(1):28-33. doi:10.1101/GR.2.1.28.

6. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods*. 2014;11(5):499-507. doi:10.1038/nmeth.2918.

7. Grotjohann T, Testa I, Leutenegger M, et al. Diffraction-unlimited all-optical imaging and writing with a photochromic GFP. *Nature*. 2011;478(7368):204-208. doi:10.1038/nature10497.

8. Brakemann T, Stiel AC, Weber G, et al. A reversibly photoswitchable GFP-like protein with fluorescence excitation decoupled from switching. *Nat Biotechnol*. 2011;29(10):942-947. doi:10.1038/nbt.1952.

9. Shaner NC, Lin MZ, McKeown MR, et al. Improving the photostability of bright monomeric orange and red fluorescent proteins. *Nat Methods*. 2008;5(6):545-551. doi:10.1038/nmeth.1209.

10. Davis LM, Lubbeck JL, Dean KM, Palmer AE, Jimenez R. Microfluidic cell sorter for use in developing red fluorescent proteins with improved photostability. *Lab Chip*. 2013;13(12):2320. doi:10.1039/c3lc50191d.

11. Dean KM, Davis LM, Lubbeck JL, et al. High-speed multiparameter photophysical analyses of fluorophore libraries. *Anal Chem*. 2015;87(10):5026-5030. doi:10.1021/acs.analchem.5b00607.

12. Dean KM, Lubbeck JL, Davis LM, et al. Microfluidics-based selection of red-fluorescent proteins with decreased rates of photobleaching. *Integr Biol*. 2015;7(2):263-273. doi:10.1039/C4IB00251B.

13. Root DE, Hacohen N, Hahn WC, Lander ES, Sabatini DM. Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nat Methods*. 2006;3(9):715-719. doi:10.1038/nmeth924.

14. Chang K, Elledge SJ, Hannon GJ. Lessons from Nature: microRNA-based shRNA libraries. *Nat Methods*. 2006;3(9):707-714. doi:10.1038/nmeth923.

15. Bernards R, Brummelkamp TR, Beijersbergen RL. shRNA libraries and their use in cancer genetics. *Nat Methods*. 2006;3(9):701-706. doi:10.1038/nmeth921.

16. Shalem O, Sanjana NE, Hartenian E, et al. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science (80-)*. 2014;343(6166):84-87.

doi:10.1126/science.1247005.

17. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science (80-)*. 2014;343(6166):80-84. doi:10.1126/science.1246981.

18. Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol*. 2013;32(3):267-273. doi:10.1038/nbt.2800.

19. Zhou Y, Zhu S, Cai C, et al. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*. 2014;509(7501):487-491. doi:10.1038/nature13166.

20. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*. 2014;159(3):647-661. doi:10.1016/j.cell.2014.09.029.

21. Dixit A, Parnas O, Li B, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*. 2016;167(7):1853-1866.e17. doi:10.1016/j.cell.2016.11.038.

22. Adamson B, Norman TM, Jost M, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*. 2016;167(7):1867-1882.e21. doi:10.1016/j.cell.2016.11.048.

23. Jaitin DA, Weiner A, Yofe I, et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*. 2016;167(7):1883-1896.e15.

doi:10.1016/j.cell.2016.11.039.

24. Plamont M-A, Billon-Denis E, Maurin S, et al. Small fluorescence-activating and absorption-shifting tag for tunable protein imaging in vivo. *Proc Natl Acad Sci U S A*. 2016;113(3):497-502. doi:10.1073/pnas.1513094113.

25. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-)*. 2015;348(6233):aaa6090-aaa6090. doi:10.1126/science.aaa6090.

26. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc Natl Acad Sci*. 2016;113(39):11046-11051. doi:10.1073/pnas.1612826113.

27. Moffitt JR, Hao J, Bambah-Mukku D, Lu T, Dulac C, Zhuang X. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc Natl Acad Sci*. 2016;113(50):14456-14461. doi:10.1073/pnas.1617699113.

28. Femino AM. Visualization of single RNA transcripts in situ. *Science (80-)*. 1998;280(5363):585-590. doi:10.1126/science.280.5363.585.

29. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. 2008;5(10):877-879. doi:10.1038/nmeth.1253.

30. Subach OM, Cranfill PJ, Davidson MW, Verkhusha V V. An enhanced monomeric

blue fluorescent protein with the high chemical stability of the chromophore. Rao J, ed. *PLoS One*. 2011;6(12):e28674. doi:10.1371/journal.pone.0028674.

31. Wang S, Moffitt JR, Dempsey GT, Xie XS, Zhuang X. Characterization and development of photoactivatable fluorescent proteins for single-molecule-based superresolution imaging. *Proc Natl Acad Sci*. 2014;111(23):8452-8457. doi:10.1073/pnas.1406593111.

32. Shaffer SM, Wu M-T, Levesque MJ, Raj A. Turbo FISH: A method for rapid single molecule RNA FISH. Henrique D, ed. *PLoS One*. 2013;8(9):e75120. doi:10.1371/journal.pone.0075120.

33. Getzoff ED, Gutwin KN, Genick UK. Anticipatory active-site motions and chromophore distortion prime photoreceptor PYP for light activation. *Nat Struct Biol*. 2003;10(8):663-668. doi:10.1038/nsb958.

34. Zhang Z, Revyakin A, Grimm JB, Lavis LD, Tjian R. Single-molecule tracking of the transcription cycle by sub-second RNA detection. *Elife*. 2014;3:e01775. doi:10.7554/eLife.01775.

35. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345. doi:10.1038/nmeth.1318.

36. Lutz R, Bujard H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res*. 1997;25(6):1203-1210. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146584&tool=pmcentrez

&rendertype=abstract.

37. Kivioja T, Vähärautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*. 2011;9(1):72-74. doi:10.1038/nmeth.1778.

38. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci*. 2012;109(4):1347-1352. doi:10.1073/pnas.1118018109.

39. Rasnik I, McKinney SA, Ha T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Methods*. 2006;3(11):891-893. doi:10.1038/nmeth934.