Intensity Modulation of Trichromatic Split Fluorescent Proteins for Live Cell Mapping

Mamoru Ishii[1,2*], Tomoaki Kinjo[3], Yohei Kondo[4,5], Kenta Terai[6], Kazuhiro Aoki[1], Brian Kuhlman[3,7], Michiyuki Matsuda[1,8,9*]

[1]Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan.
[2]Institute of Industrial Science, The University of Tokyo, Tokyo, Japan
[3]Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Chapel Hill, NC, USA
[4]Center for One Medicine Innovative Translational Research (COMIT), Nagoya University, Nagoya, Aichi, Japan
[5]Graduate School of Medicine, Nagoya University, Tsurumai-cho, Nagoya, Aichi 466-8550, Japan
[6]Graduate School of Medicine, Tokushima University, Shinkura-cho, Tokushima 770-8501, Japan
[7]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[8]Affiliate Graduate School, Graduate School of Medicine, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan.
[9]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Chuo-shi, Yamanashi 409-3898 Japan.

*Corresponding authors. Center for Living Systems Information Science, Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
Lead correspondence: matsuda.michiyuki.2c@kyoto-u.ac.jp (M.M.), mishii@iis.u-tokyo.ac.jp (M.I.)

Running title
Multiplexed cell labeling with split FPs

Keywords
Split fluorescent proteins, Multiplexed spectral labeling, Cell labeling, Fluorescence microscopy, Protein tagging

Abstract

Multiplexed live cell-labeling technologies are essential tools for investigating complex biological systems, yet existing approaches face limitations in achieving reliable discrimination of multiple cell populations. Here, we present color and intensity modulation with split fluorescent proteins for cell labeling (Caterpie), a live cell labeling system that enables high-fidelity identification of multiple cell populations through rational design of split fluorescent proteins. Drawing inspiration from human trichromatic vision, we engineered optimized split fluorescent protein variants—split CFP2, split mNG3A, and split sfCherry3C—and developed a comprehensive library of the 11th β-strand (FP_{11}) tags comprising variable numbers of tandem repeats. Structure-guided protein engineering enhanced the performance of split mNG3A and split sfCherry3C through targeted modifications of their C-terminal sequences and binding interfaces, respectively. By systematically varying the composition and arrangement of FP_{11} repeats, we selected 20 distinct FP_{11} tags that enable robust cell population discrimination with 97% accuracy using conventional fluorescence microscopy. In these FP_{11} tag-expressing cells, we expressed multiple EGF ligands and their receptors and cultured them in bulk to examine the effects of EGF signaling on chemotaxis within the cell population. The results showed that chemotaxis changes according to the intensity of EGFR signaling. The Caterpie system holds significant potential for applications requiring precise identification and tracking of multiple cell populations in complex biological systems.

Introduction

Multiplexed spectral labeling techniques represent essential tools across diverse biological disciplines, including neuroscience, developmental biology, and immunology. These techniques enable critical applications such as discrimination of adjacent cells, specific cell populations labeling, and lineage tracing. The Brainbow system exemplifies this approach, utilizing multiple gene cassettes designed to stochastically express one of three fluorescent proteins (FPs), thereby generating distinct spectral signatures (*1*). While this system has proven particularly valuable for morphological analyses including connectomics, it presents limitations for specific cell labeling and lineage tracing due to potential variations in the relative FP expression levels across cellular generations. The development of Bitbow addressed certain limitations of earlier systems by enabling the generation of up to 31 (2^5-1) distinct color variations through random toggling of five FP genes within a single gene cassette (*2*). However, this approach faces challenges due to biased gene selection partners, resulting in fewer achievable color discrimination than theoretically predicted. These limitations underscore the advantages of rationally designed color palettes. However, conventional approaches to expanding color diversity through serial concatenation of multiple FPs face practical constraints due to increasing gene sizes. Split fluorescent proteins emerge as an elegant solution to these technical challenges.

Split fluorescent proteins (split FPs) are generated through strategic dissection of β-barrel fluorescent proteins (FPs) into two components: the 11th β-strand (FP_{11}) and a complementary segment comprising the first ten β-strands ($FP_{1\text{-}10}$) (*3*). While neither fragment exhibits fluorescence independently, their co-expression facilitates spontaneous reassembly, enabling chromophore maturation and subsequent fluorescence emission. This complementation system has been successfully employed for protein labeling applications, wherein target proteins are tagged with FP_{11} in cells expressing $FP_{1\text{-}10}$ (*4*, *5*). A particularly advantageous feature of this system emerges under conditions of abundant $FP_{1\text{-}10}$ expression: enhanced fluorescence intensity can be achieved by increasing the number of FP_{11} tags conjugated to the protein of interest, while maintaining relatively modest gene size requirements (*6*–*9*).

Drawing inspiration from human color perception, which discriminates diverse spectral signatures through the integration of three primary color intensities: blue, green, and red, we developed a novel multiplexed cellular labeling approach. This system exploits engineered arrays of FP_{11} fragments derived from cyan, green, and red split fluorescent proteins, each designed to complement with their cognate $FP_{1\text{-}10}$ fragments. This method, named color and intensity modulation with split fluorescent proteins for cell labeling (Caterpie), achieves robust discrimination of 20 distinct cell populations with 97% accuracy using conventional epi-fluorescence microscopy.

Results

Split fluorescent proteins for the trichromatic color palette

We aimed to establish a trichromatic color palette through the systematic engineering of tandem FP_{11} repeats derived from cyan, green, and red split fluorescent proteins, which undergo fluorescence complementation when co-expressed with their cognate FP_{1-10} fragments (Fig. 1A). For this, we sought to select split fluorescent proteins (split FPs) that achieve brightness comparable to their full-length FP ($FP_{Full-length}$) and exhibit fluorescence signal amplification through FP_{11} repeats. A critical consideration in designing this system is the requirement for sequence divergence among the trichromatic split FPs to prevent cross-complementation between fragments from different fluorescent protein pairs (*10*). To address this constraint, we evaluated split FPs derived from distinct evolutionary lineages: sfGFP, mNG2, sfCherry2, and mRuby4 (Fig. 1B). To quantitatively assess the relative brightness of $FP_{Full-length}$, FP_{11}(x1), and FP_{11}(x4) of reported split FPs, we developed a dual-reporter system. We engineered expression constructs encoding Histone H2B-tagged $FP_{Full-length}$, FP_{11}(x1), or FP_{11}(x4) fused to EBFP2-nls (nuclear localization signal) through a self-cleaving P2A site, enabling normalization of expression levels through nuclear EBFP2 fluorescence. These constructs were co-transfected into HeLa cells with a secondary plasmid encoding FP_{1-10} linked to iRFP670 via an internal ribosome entry site (IRES) at a DNA ratio of 1:4 (Fig. 1C).

After 48 hours, cells were observed under a confocal microscope, and split FP fluorescence intensities were quantified in iRFP670 positive cells (Fig. 1D). For each fluorescent protein variant, relative fluorescence intensity values were normalized to the full-length FP. In most cases, (x1) shows decreased fluorescence brightness compared to full-length, but sfGFP, mNG3A, and mNG3K were almost equivalent fluorescence brightness. Comparative analysis of (x1) versus (x4) constructs revealed that all sfGFP-derived split FPs exhibited more than two-fold enhancement in median brightness, with CFP2 demonstrating the most substantial increase of more than three-fold. Consistent with previous findings (*11*), split mNG3A demonstrated superior amplification from (x1) to (x4) compared to split mNG3K. Split sfCherry3C_{11}(x1) exhibited diminished fluorescence intensity relative to sfCherry3$C_{Full-length}$, attributable to reduced association efficiency, corroborating earlier observations (*12*). The truncated mRuby4$_{11}$ showed markedly decreased fluorescence intensity, highlighting the critical role of its C-terminal unstructured polypeptide chain (*10*). These preliminary findings indicated that while the combination of CFP2, mNG3A, and sfCherry3C provides a promising foundation for the trichromatic system, although further engineering optimization of mNG3A and sfCherry3C is necessary to achieve optimal performance across all spectral channels.

Fluorescence intensity-based cell classification using tandem CFP2 β-strand repeats

Before the engineering of mNG3A and sfCherry3C, we established a platform for cell

identification through fluorescence intensity modulation using $CFP2_{11}$ repeats. To this end, we constructed plasmids encoding repeated sequences of $CFP2_{11}$ (Fig. 2A). Initially, we engineered a plasmid encoding a Histone H2B-$CFP2_{11}$(x1) fusion construct, in which the $CFP2_{11}$(x1) sequence was flanked by a BglII restriction site at its 5' terminus and BamHI and NotI restriction sites at its 3' terminus. Through sequential ligation of BglII/NotI-digested fragments (insert) with BamHI/NotI-digested fragments (vector), we successfully amplified $CFP2_{11}$ to create (x2), (x4), and (x8) variants. Based on previous research (*6*), we incorporated GGSGG linker sequences between $CFP2_{11}$ fragments. At the BglII-BamHI junction, the six nucleotides encode glycine and serine residues, thereby serving as an integral part of the linker sequence. The $CFP2_{11}$ repeats [(x1), (x2), (x4), or (x8)] fused with Histone H2B were introduced into HeLa cells stably expressing $CFP2_{1\text{-}10}$ ($HeLa^{CFP2\ 1\text{-}10}$) (Fig. 2B). As a control, we expressed Histone H2B-tagged $CFP2_{\text{Full-length}}$ in the same $HeLa^{CFP2\ 1\text{-}10}$ cell line. When normalized to nuclear mCherry fluorescence, the CFP2 fluorescence intensity increased across consecutive constructs with a factor of approximately 1.7-fold, which was slightly lower than the theoretically expected 2-fold increment (Fig. 2C).

To identify cells expressing different numbers of $CFP2_{11}$ repeats, we implemented a Gaussian mixture model (GMM) for cell classification (Fig. 2D). The fluorescence intensities of mCherry and CFP2 were transformed into polar coordinates, where the $\text{Angle}_{\text{mCherry vs CFP2}}$ (ranging from 0° to 90°) represented the CFP2/mCherry intensity ratio. The angular data were divided into training and test datasets. We trained a GMM on the training dataset to establish classification parameters, determining the decision boundary of the $\text{Angle}_{\text{mCherry vs CFP2}}$. The model performance was validated using the test dataset, and the predicted repeat numbers were compared with the actual repeat numbers. The model achieved an overall accuracy of 96% (Fig. 2E). Among all constructs, $CFP2_{11}$(x8) exhibited the lowest accuracy (93.9%), with 6.1% of (x8)-expressing cells being misclassified as (x4).

To validate the system's discriminative capacity in a heterogeneous context, we analyzed mixed populations of cells expressing four different $CFP2_{11}$ repeat variants using fluorescence microscopy and generated CFP2/mCherry ratio images (Fig. 2F). The trained GMM demonstrated robust performance in classifying individual cells into (x1), (x2), (x4), or (x8) populations with 96% accuracy, which was validated through visual inspection (Fig. 2G).

Furthermore, the GMM provided membership probabilities for each cell, enabling quantitative assessment of classification reliability (Fig. 2H). These results established that four-level intensity modulation of a single fluorescent protein is sufficient for reliable population discrimination.

Computational design of split sfCherry3C for enhanced β-barrel stability and split-strand complementation

To optimize split sfCherry3C toward fluorescence intensities close to those of the full-length protein, we modelled the $\text{sfCherry3C}_{1\text{-}10}$/$\text{sfCherry3C}_{11}$ complex with ColabFold

(*13*) (based on AlphaFold-Multimer) (*14*), and used the structure model for subsequent design with Rosetta modeling suite (*15*). Because both global β-barrel stability and $\beta_{1\text{-}10}/\beta_{11}$ interface affinity are crucial for fluorescence of split FPs, we pursued two design strategies in parallel (*16*). First, to enhance global β-barrel stability, we employed a Rosetta script adapted from the previously described "Point-Mutation" (PM) workflow (*16*) and performed *in-silico* site-saturation mutagenesis (SSM) across all 224 residues, ranking each variant by the change in calculated energy (ΔE) calculated from the Rosetta *total_score* term (fig. S1). Next, to refine the $\beta_{1\text{-}10}/\beta_{11}$ interface, we repurposed the PM script for interface design; with the revised script, we performed SSM at every β_{11} position and calculated the binding energy of each single mutant from *dG_separated* score obtained with InterfaceAnalyzer protocol (Fig. 3A). Each single-mutant model was subsequently processed by the "Mutation Cluster (MC)" (*16*) workflow modified for interface design, in which FastDesign protocol with InterfaceDesign relax script stochastically samples sequence and rotamer space at residues within 5–7 Å of the "seed mutation" through Monte-Carlo design/minimization cycles, yielding multi-residue mutants that are re-evaluated with InterfaceAnalyzer (Fig. 3B). After filtering by the scores such as *total_score*, *dG_separated*, and mutation number, we conducted visual inspection to retain the variants that reinforce interface contacts while excluding ones that disrupted the chromophore-forming triad M67–Y68–G69 or introduced additional aromatic residues into β_{11} which could promote aggregation of multimeric β_{11} repeat. Following these criteria, we selected twenty-two candidate designs for subsequent in-cell validation.

Upon stable co-expression of $\beta_{1\text{-}10}$ and β_{11} in HeLa cells, three variants (M1, M2, M3) outperformed the parental split sfCherry3C in fluorescent intensity (Fig. 3C, D). We then combined the mutations from the three top performing variants to generate an optimized variant, named split sfCherry3Csp, which achieved a 2.5-fold enhancement in fluorescent intensity compared to the parental split sfCherry3C. The M1 and M3 mutations locate at the $\beta_{1\text{-}10}/\beta_{11}$ interface, where the models suggest the improved association of $\beta_{1\text{-}10}$ and β_{11} through the formation of reinforced side-chain contacts. The M2 mutation lies on the opposite face of the interface, implying an effect on β-barrel stability (Fig. 3E). The fluorescence spectra of split sfCherry3Csp were indistinguishable to those of the parental split sfCherry3C, indicating that the incorporated mutations enhanced the β-barrel stability and $\beta_{1\text{-}10}/\beta_{11}$ association without affecting the chromophore maturation (Fig. 3F).

Enhancement of complementation efficiency through structure-guided engineering of split mNG3A

As shown in Fig. 1D, split mNG3A demonstrated only modest signal amplification, with a 1.3-fold increase in fluorescence intensity from (x1) to (x4) variants. To gain insight into the structural factors underlying this limitation, we modelled the octameric $mNG3A_{11}$ peptide using ColabFold (*13*) (based on AlphaFold2) (*17*). The predicted structures suggested that

mNG3A$_{11}$ repeat adopts a loosely helical conformation, with hydrophobic residues (F5, W8, F12, M15, M16) converging to form a continuous hydrophobic interaction, despite only moderate model confidence. These features implied that multimeric mNG3A$_{11}$ may self-aggregate via its hydrophobic surfaces (Fig. 4A) (*17*). Additionally, inspection of the parental mNG crystal structure (PDB: 5LTP) and the structure model of split mNG3A suggested that the C-terminal residues D14–M15–M16 contribute minimally to the β-barrel fold (Fig. 4B). We therefore reasoned that removing these C-terminal residues could mitigate the aggregation propensity without perturbing β-barrel and chromophore maturation. Guided by these insights, we implemented three C-terminal modification approaches: (1) deletion of the C-terminal D14-M15-M16, (2) addition of one or two aspartic acid residue to preserve charge that could electrostatically repel neighboring fragments, and (3) substitution of F12 with tyrosine to introduce polarity (Fig. 4C). To evaluate these modifications, we expressed 8 repeats of the mNG3A$_{11}$ fragments fused to Histone H2B in HeLa cells expressing mNG3A$_{1\text{-}10}$ (Fig. 4D). Nuclear mNG3A fluorescence was quantified through microscopy, with expression levels normalized to nuclear mTagBFP2 fluorescence intensity (Fig. 4E). Among the modifications tested, we chose the mNG3Asp$_{11}$ variant for further analysis because of its highest fluorescence intensity, which was 9.1-fold brighter than the parental mNG3A$_{11}$. The fluorescence spectra obtained using the mNG3Asp$_{11}$ fragment were identical to those observed with the mNG3A$_{11}$ fragment (Fig. 4F).

Performance evaluation of split mNG3Asp and split sfCherry3Csp was conducted in parallel with split CFP2, following our previously established experimental framework (Fig. 4G). Quantitative analysis revealed that mNG3Asp$_{11}$(x4) achieved a 2.3-fold enhancement in fluorescence intensity compared to its monomeric counterpart. The engineered sfCherry3Csp demonstrated robust performance, with its monomeric variant exhibiting 80% of the fluorescence intensity of the full-length protein, while the tetrameric construct showed a 3-fold signal amplification relative to the monomer. These significant improvements in signal amplification efficiency indicate that both engineered variants, split mNG3Asp and split sfCherry3Csp, now achieve performance metrics comparable to split CFP2, providing the basis for trichromatic imaging.

Optimization of the fusion partner for the FP$_{11}$ tag

Prior to establishing the FP$_{11}$ tag library, we aimed to reduce construct size by identifying alternatives to the relatively large Histone H2B fusion tag. We evaluated multiple fusion protein configurations, comparing Histone H2B (126 amino acids) with two smaller alternatives: GB1 (56 amino acids) (*18*) and ΔSUMOstar (74 amino acids), a truncated version of SUMOstar (*19*), as well as constructs lacking fusion tags (fig. S2A). ΔSUMOstar was engineered by removing the unstructured polypeptide regions from both termini of SUMOstar; the resulting sequence is detailed in Figure S2B. In agreement with previous studies (*20*), constructs lacking fusion proteins demonstrated markedly reduced fluorescence

intensity, an effect that persisted even in the octameric (x8) variant (fig. S2C). Comparative analysis of fusion partners revealed that ΔSUMOstar-tagged constructs achieved substantially higher fluorescence intensity than GB1-tagged variants in octameric configurations, demonstrating the importance of fusion partner selection for optimal signal amplification. The ΔSUMOstar-tagged constructs exhibited consistent signal enhancement, with fluorescence intensity increasing 1.6-fold for each doubling of $CFP2_{11}$ repeats, ultimately enabling 93% discrimination accuracy between variants (fig. S2, D to F).

To optimize the modular arrangement of the three FP_{11} variants ($CFP2_{11}$, $mNG3Asp_{11}$, and $sfCherry3Csp_{11}$), we conducted a systematic analysis of possible configurations. In initial experiments, we generated ΔSUMOstar fusion constructs containing six total repeats, wherein each FP_{11} fragment was represented in duplicate (fig. S3A). The complete set of six possible permutations was generated and stably expressed in HeLa cells constitutively expressing all three $FP_{1\text{-}10}$ fragments. Quantitative analysis of relative fluorescence intensities was performed using EBFP2 fluorescence for normalization of expression levels (fig. S3B). Among all configurations evaluated, the $CFP2_{11}$(x2)-$sfCherry3Csp_{11}$(x2)-$mNG3Asp_{11}$(x2) arrangement demonstrated marginally superior fluorescence intensity across all spectral channels, leading to its selection for subsequent studies.

Selection of 20 optimal FP_{11} tags

Based on these preliminary findings, we established a comprehensive library of fusion constructs combining ΔSUMOstar with $CFP2_{11}$, $sfCherry3Csp_{11}$, and $mNG3Asp_{11}$ fragments in this order (Fig. 5A). Each fragment was represented in varying copy numbers [(x0), (x1), (x2), (x4), or (x8)] for each. In the first screening, we limited the FP_{11} tag combinations to 96 variants, with the total number of repeats not exceeding (x12). The integrity of all FP_{11} repeat sequences in the 96 plasmid constructs was verified by Sanger sequencing, thereby establishing a comprehensive FP_{11} tag library.

To generate a stable cellular platform for evaluating the FP_{11} tag library, we engineered an expression vector encoding all three $FP_{1\text{-}10}$ fragments ($CFP_{1\text{-}10}$, $sfCherry3Csp_{1\text{-}10}$, and $mNG3A_{1\text{-}10}$) separated by self-cleaving P2A peptides, followed by an IRES and iRFP670 (Fig. 5B). Through fluorescence-activated cell sorting (FACS) based on iRFP670 expression levels, we established a HeLa cell clone ($HeLa^{FP1\text{-}10x3_iRFP670}$) expressing high levels of three $FP_{1\text{-}10}$ fragments.

The complete library of 96 FP_{11} tags was stably expressed in $HeLa^{FP1\text{-}10x3_iRFP670}$ cells, with resultant populations analyzed by confocal microscopy. To quantitatively assess the fluorescence distribution patterns, we implemented spherical coordinate transformation to calculate two angular parameters: $Angle_{CvsG}$ and $Angle_{RvsCG}$, each ranging from 0° to 90° (Fig. 5C, left). To evaluate the discriminative power of these angular parameters, we performed pairwise population analysis using selected cell populations from the library. Cell classification was achieved using an unsupervised two-component Gaussian mixture model

without prior training data (Fig. 5C, right). Classification accuracy was determined by comparing model-predicted clusters against known population identities. Superior classification accuracy correlated with greater separation of populations in the $Angle_{CvsG}$ – $Angle_{RvsCG}$ parameter space. This analytical framework was systematically applied to all possible pairwise combinations within the 96 FP_{11}-tagged cell library (Fig. 5D).

To determine the number of distinguishable cell populations, we implemented an iterative sampling approach due to the vast number of possible patterns (approximately 2×10^{20}) when selecting 20 FP_{11}-tagged cells from a pool of 96 (Fig. 5E). Random subsets of predetermined sizes were drawn from the 96 FP_{11}-tagged cell populations, with pairwise classification accuracies extracted from the comprehensive analysis presented in Fig. 5D. For each FP_{11}-tagged subset, we identified the minimum classification accuracy, reflecting the subset's ability to distinguish the most similar cell population pairs. This sampling process was repeated 10^7 times, and the maximum achievable minimum accuracy was plotted as a function of subpopulation size (Fig. 5F). The analysis revealed that a subset of 20 cell populations maintained robust discrimination with a minimum pairwise classification accuracy of 90%. However, expanding to 30 populations resulted in a substantial decrease in minimum accuracy to 60%. Based on these quantitative insights, we selected an initial panel of 30 candidate FP_{11} tags (Fig. 5A, 1st) for subsequent refinement to establish the final optimized set of 20 tags (Fig. 5A 2nd) that would ensure maximal discriminative power.

We conducted comprehensive evaluation of the 30 candidate populations using both confocal microscopy and flow cytometry, with three independent experimental replicates (fig. S4). The relationship between population subset size and maximum achievable minimum pairwise classification accuracy was reassessed using these orthogonal detection methods (Fig. 5G). While fluorescence microscopy demonstrated marginally superior classification accuracy, flow cytometry maintained robust discrimination capabilities, achieving >85% accuracy across a set of top 20 FP_{11} tags. The final optimized set of 20 FP_{11} tags and their comprehensive pairwise discrimination analysis is presented in Figure S5. Here, to optimize the fluorescence signal intensity of single-color reference populations, we refined the copy number composition (Fig. 5A Final): $CFP2_{11}$(x8) and $mNG3Asp_{11}$(x8) were implemented at maximum repeat numbers, while $sfCherry3Csp_{11}$ was limited to (x4) to circumvent aggregation phenomena observed at higher copy numbers (fig. S6).

High-fidelity discrimination of twenty cell populations using optimized $FP_{1\text{-}10}$ expression platform

The $FP_{1\text{-}10}$ expression system was also refined by replacing the iRFP670 reporter with a puromycin resistance cassette. Through stringent puromycin selection, we established a HeLa clone cell ($HeLa^{FP1\text{-}10x3}$) expressing elevated levels of $CFP2_{1\text{-}10}$, $mNG3Asp_{1\text{-}10}$, and $sfCherry3Csp_{1\text{-}10}$. The optimized set of 20 FP_{11} tags was stably expressed in $HeLa^{FP1\text{-}10x3}$ cells, with individual cell populations analyzed by confocal microscopy. Fluorescence was

visualized with distinct spectral channels: CFP2 (blue), mNG3Asp (green), and sfCherry3Csp (red). Merged images revealed unique signatures based on both color composition and intensity distribution (Fig. 6A).

Cell segmentation was performed on maximum intensity projections of composite fluorescence signals using Cellpose (*21*). Mean fluorescence intensities were quantified across all three channels for each segmented cell. Following the methodology established in Figure 5C, we implemented spherical coordinate transformation to calculate $\text{Angle}_{\text{CvsG}}$ and $\text{Angle}_{\text{RvsCG}}$, generating a two-dimensional angular representation of relative fluorescence intensities (Fig. 6B, left). These parameters were used to train a Gaussian mixture model (GMM), which established both label assignment distributions and label membership probability distributions in $\text{Angle}_{\text{CvsG}}$ and $\text{Angle}_{\text{RvsCG}}$ (Fig. 6B, center and right). Validation of the trained GMM on independent test data demonstrated exceptional classification performance, achieving 97% accuracy in discriminating all 20 cell populations (Fig. 6C). Using unsupervised k-means clustering on the 20 components achieved 96% accuracy in distinguishing between the 20 cell populations, even without requiring pre-training data for individual fluorescence profiles (fig. S7).

To evaluate system performance under practical conditions, we used pooled samples containing all 20 cell populations and acquired tiled images (Fig. 6D). Application of the trained GMM enabled robust classification of individual cells, with 78% of cells being classified with more than 95% probability (Fig. 6E).

Interaction dynamics in multi-population cell cultures of MDCK cells expressing EGF-family ligands and their receptors

Finally, as an application of this system—named color and intensity modulation with split fluorescent proteins for cell labeling (Caterpie)—we studied the contribution of heterotypic cell interactions mediated by EGF-family ligands and their cognate receptors. For this, we used Madin-Darby Canine Kidney (MDCK) cells, which are widely used to study collective cell migration and cell competition regulated by the EGF signaling pathway. MDCK cells were labeled with the aforementioned $\text{FP}_{1\text{-}10}$ and 20 FP_{11} tags with 96% accuracy (fig. S8A, B). Mixed cells were classified by applying a trained Gaussian mixture model (fig. S8C), 75% of cells being classified with more than 95% probability. The cell densities of twenty different labeled cell types at 72 hours post-labeling ranged within a 1.4-fold difference between the minimum and maximum values (fig. S9).

Among the four EGF-family ligands expressed in MDCK cells, we focused on HBEGF and EREG. HBEGF, a high-affinity ligand, binds to heparan sulfate proteoglycans to provide strong signals restricted to short distances, promoting localized cell migration. In contrast, EREG, a low-affinity ligand, diffuses quickly and remotely, propagating signals approximately four times faster than high-affinity ligands and more efficiently to distant cells (*22*). Both HBEGF and EREG bind to EGFR (ErbB1) and ErbB4. Using lentiviral

transduction, we introduced EREG, HBEGF, EGFR (ErbB1), or ErbB4 in fluorescently labeled MDCK cells (Fig. 7A). To investigate how EGF ligands and receptors with different affinities interact between heterotypic cells and generate spontaneous patterns, these four cell types, along with wild-type cells, were co-cultured for 8 days under 3% low-serum conditions. The resulting cell population was analyzed by fluorescence microscopy to identify each cell type (Fig. 7B).

Before analyzing the positional relationships between heterotypic cells, we first examined the distribution patterns of homotypic cells, as cell types with limited motility and dispersal capability affect intercellular interactions with heterotypic cells. To examine homotypic cell clustering, each cell type is individually displayed (Fig. 7C). Dispersion of each cell type was analyzed by the Clark-Evans index; values less than 1 indicate cellular clustering. We did not find significant homotypic cell clustering in any cell types (Fig. 7D). Then, we investigated the frequency of adjacent heterotypic cells. When examining receptor-expressing cells adjacent to EREG ligand or HBEGF-expressing cells (Fig. 7E), we found EGFR cells displayed significantly higher adjacency to EREG cells than wild-type cells. Conversely, when examining ligand-expressing cells adjacent to EGFR receptor or ErbB4-expressing cells, HBEGF cells exhibited higher adjacency to EGFR cells and to a lesser extent to ErbB4 in comparison to wild-type cells. These results suggest that EREG and EGF function as chemoattractants, with EREG weakly attracting distant cells, while HBEGF strongly attracts nearby cells.

In conclusion, the Caterpie method provides a versatile platform for analyzing cell populations that may exhibit homotypic and/or heterotypic clustering.

Discussion

We have developed color and intensity modulation with split fluorescent proteins for cell labeling (Caterpie), a novel cell identification platform that achieves robust identification of 20 distinct cell populations with 96-97% accuracy. This system leverages fluorescence intensity modulation through engineered arrays of split fluorescent protein (split FP) fragments. The technological foundation of Caterpie comprises three complementary split FPs: split CFP2 and two newly engineered variants - split mNG3Asp and split sfCherry3Csp. To implement this system, we established a comprehensive toolkit consisting of 20 distinct FP_{11} tags, each constructed by fusing truncated SUMOstar protein to the 11th β-strand of the three split fluorescent proteins in varying tandem repeat configurations. These engineered FP_{11} tags demonstrate highly specific labeling of $FP_{1\text{-}10}$-expressing cells, enabling population discrimination with 96-97% accuracy through Gaussian mixture model (GMM) analysis of the resulting multidimensional fluorescence signatures.

Current multicolor fluorescent labeling approaches can be broadly categorized into two distinct methodological frameworks. The first relies on stochastic recombination events mediated by site-specific recombinase systems such as Cre/loxP and Flp/FRT (*1*, *2*, *23–27*). The second approach exploits the inherent randomness of transfection efficiencies or genomic integration sites (*28–32*). While both established approaches depend on stochastic processes to generate fluorescent protein expression patterns, Caterpie represents a paradigm shift through its implementation of rationally designed color palettes. This deterministic approach enables precise targeting of specific cell populations and lineages. The system holds particular promise for applications requiring high-fidelity identification and longitudinal tracking of distinct cell types within complex biological contexts. However, the full realization of Caterpie's potential necessitates further optimization of delivery methodologies. Specifically, development of refined strategies is required to achieve selective incorporation of distinct FP_{11} tag variants into target cell populations while maintaining variant segregation and preventing cross-contamination.

The theoretical prediction of an 8-fold signal enhancement for FP_{11}(x8) constructs relative to their monomeric counterparts was not fully realized in experimental measurements. Our studies with split CFP2 demonstrated approximately 5-fold enhancement in median fluorescence intensity (Fig. 2C). Further investigations revealed that the choice of fusion protein significantly impacts the signal amplification efficiency of octameric constructs (fig. S2C). In research using sfGFP, when x3, x4, or x7 $sfGFP_{11}$ was fused to β-tubulin or Lamin A/C, length and brightness correlated well, nearly achieving theoretical limits (*4*, *6*). However, when Teneurin-m was fused to x7 $sfGFP_{11}$, only a 4-fold increase in brightness was achieved (*9*). With the goal of reducing gene size, which is an advantage of our approach, we changed from the initially used Histone H2B to ΔSUMOstar and achieved almost the same amplification rate as Histone H2B, x1.6 by each repetition (fig. S2). This collective evidence

suggests that achieving linear brightness amplification through multimerization of the 11th β-strand requires careful optimization of the fusion partner.

The application of computational protein structure prediction proved instrumental in enhancing split fluorescent protein performance (Fig. 3). This approach represents a significant departure from conventional optimization strategies, which typically rely on random mutagenesis (*33*). Simultaneous introduction of mutations in both $\beta_{1\text{-}10}$ and β_{11} regions that enhance their mutual interactions proved challenging through random mutagenesis, because the statistical probability of obtaining cooperative mutations is exceedingly low. Our successful implementation of structure prediction-guided optimization suggests a broader applicability of this methodology across other split fluorescent protein systems.

The discriminative capacity of the Caterpie system holds significant potential for expansion through two primary avenues: the development of novel split fluorescent protein variants and the integration of subcellular targeting sequences. Furthermore, compatibility with the Landing Pad system (*34*), which enables precise single-copy genomic integration, presents opportunities for multiplexed cellular labeling. These technological developments collectively position Caterpie as a versatile platform for investigating complex biological processes requiring simultaneous tracking of multiple distinct cell populations.

Materials and Methods

Plasmids

The following cDNAs were synthesized with optimized codons by GeneArt (Thermo Fisher Scientific, Waltham, MA): sfGFP$_{1-10}$ (*3*), CFP2$_{1-10}$ (*35*), mNeonGreen3A$_{1-10}$ (*11*), sfCherry3Csp$_{1-10}$, mRuby4$_{1-10}$ (*10*), GB1 (*18*), ΔSUMOstar (*19*). Additional split fluorescent protein variants were generated through site-directed mutagenesis of existing templates: Cerulean$_{1-10}$ (*10*) and YFP1$_{1-10}$ (*35*) from sfGFP$_{1-10}$, mNeonGreen3K$_{1-10}$ (*11*) from mNeonGreen3A$_{1-10}$, and sfCherry3C$_{1-10}$ (*12*) from sfCherry2$_{1-10}$ (Addgene: #82603). The Histone H2B coding sequence was subcloned from Addgene (plasmid # 11680 for Histone H2B; Cambridge, MA).

Vector Construction for Stable Cell Line Generation

A base Tol2 transposon vector (pT2A-IRES-iRFP670) was constructed by subcloning iRFP670 cDNA (*36*) with an internal ribosome entry site (IRES) (*37*) into the pT2A vector (*38*). CFP2$_{1-10}$ cDNA was subsequently PCR-amplified and inserted into this base vector using In-Fusion assembly (Clontech, Mountain View, CA) to generate pT2A_CFP2$_{1-10}$-IRES-iRFP670. Two distinct vector backbones, pT2A_IRES-iRFP670 and pT2ADW (containing IRES-puro cassette) (*39*), were used for multicistronic construct assembly. The following elements were PCR-amplified and assembled into both vectors using In-Fusion: sfCherry3Csp$_{1-10}$, mNeonGreen3Asp$_{1-10}$ with self-cleaving P2A peptide (*40*), and CFP2$_{1-10}$ with P2A peptide. Stable cell lines were generated through co-transfection of the constructed pT2A vectors with pCS-TP (obtained from Kawakami et al., 2004 (*38*)).

To generate pPBpuro_EF1a constructs containing CFP2$_{11}$(x1), CFP2$_{11}$(x2), CFP2$_{11}$(x4), CFP2$_{11}$(x8) -mCherry-NLS, cDNAs encoding Histone H2B, self-cleaving peptide P2A sequences, mCherry, and the nuclear localization signal (NLS) of the SV40 large T antigen (PKKKRKV) (*41*) were PCR-amplified and assembled using In-Fusion into pPBpuro_EF1a vectors (a kind gift from K. Yusa), yielding pPBpuro_EF1a_Histone H2B-P2A-mCherry-NLS. Either synthesized CFP2$_{11}$ or CFP2$_{\text{Full-length}}$ cDNA (created by PCR amplification of CFP2$_{1-10}$ cDNA and CFP2$_{11}$) was then assembled into pPBpuro_EF1a_Histone H2B-P2A-mCherry-NLS, resulting in pPBpuro_EF1a_CFP2$_{\text{Full-length}}$/CFP2$_{11}$(x1)-mCherry-NLS. Restriction enzyme sites (BglII before FP$_{11}$, BamHI and NotI after FP$_{11}$) were introduced to facilitate the construction of tandem repeats. To generate CFP2$_{11}$(x2), the CFP2$_{11}$(x1) insert and vector were digested with BglII/NotI and BamHI/NotI, respectively, and then ligated. Subsequently, CFP2$_{11}$(x4) and CFP2$_{11}$(x8) were generated using the same approach. By substituting the CFP2$_{11}$ insert with sfCherry3Csp$_{11}$ or mNeonGreen3Asp$_{11}$ inserts, multiple tandem repeats of these fragments were constructed using the same strategy. Various modifications were made to customize the constructs for specific experimental requirements: the puromycin-resistance gene (puro) was replaced with the blasticidin S-resistance gene (bsr); Histone H2B was substituted with GB1,

ΔSUMOstar; mCherry was replaced with EBFP2 or mTagBFP2; and the NLS was substituted with the nuclear export signal (NES) of the HIV-1 rev protein (LPPLERLTLD) (*42*).

Mutant variants of sfCherry3C$_{1-10}$ were generated using overlap extension PCR and assembled into pPB-based vectors (*43*) containing IRES-bsr (blasticidin S-resistance gene) using In-Fusion. Additional mutant variants of mNeonGreen3A$_{11}$ and sfCherry3C$_{11}$ were synthesized and inserted into pPB-based vectors (*43*) containing IRES-puro (puromycin-resistance gene) using Ligation High Ver. 2 (TOYOBO). Stable cell lines were established through co-transfection of these constructs with pCMV-mPBase (obtained from the Wellcome Trust Sanger Institute).

Cell culture and establishment of cell lines

HeLa cells and Lenti-X 293T cells were obtained from the Human Science Research Resources Bank and Clontech, respectively. Both cell lines were maintained in DMEM (Wako Pure Chemical Industries) supplemented with 10% fetal bovine serum (Sigma-Aldrich) and 1% penicillin/streptomycin (Nacalai Tesque). MDCK (ECACC 84121903) cells were purchased from the European Collection of Authenticated Cell Cultures (ECACC) through the RIKEN BioResource Center (no. RCB0995) and maintained in DMEM (Wako Pure Chemical Industries) supplemented with 10% fetal bovine serum (Sigma-Aldrich), 1% penicillin/streptomycin (Nacalai Tesque).

For transposon-mediated gene transfer, pT2A_CFP2$_{1-10}$-IRES-iRFP670, or pT2A_FP$_{1-10}$x3-IRES-iRFP670 was cotransfected with pCS-TP into HeLa cells by using 293fectin (Thermo Fisher Scientific, Waltham, MA). These obtained HeLa cells were sorted using an FACS Aria IIu cell sorter (Becton Dickinson, Franklin Lakes, NJ) based on iRFP670 fluorescence to achieve a high expression level of the CFP2$_{1-10}$ or FP$_{1-10}$x3. Single-cell cloning of these sorted populations to yield HeLa$^{\text{CFP2 1-10_iRFP670}}$ cells and HeLa$^{\text{FP1-10x3_iRFP670}}$ cells.

pT2A_FP$_{1-10}$x3-IRES-iRFP670 was cotransfected with pCS-TP using either 293fectin (Thermo Fisher Scientific, Waltham, MA) for HeLa cells or electroporation with an Amaxa nucleofector (Lonza, Basel, Switzerland) for MDCK cells. The transfected cells were selected with 5 μg mL^{-1} puromycin (no. P-8833; Sigma-Aldrich), followed by single-cell cloning to yield HeLa$^{\text{FP1-10x3}}$ cells or MDCK$^{\text{FP1-10x3}}$ cell lines.

Fluorescence imaging

For evaluation of split fluorescent proteins, HeLa cells in 24-well plates were co-transfected with 100 ng of FP$_{11}$ plasmid [FP$_{\text{Full-length}}$, FP$_{11}$(x1), or FP$_{11}$(x4)] and 400 ng of FP$_{1-10}$ plasmid using 293fectin (Thermo Fisher Scientific, Waltham, MA). Cells were seeded onto collagen type I-coated (Nitta Gelatin, Osaka, Japan) glass-based 24-well plates (AGC Inc., Tokyo, Japan) and cultured for 48 hours. Prior to imaging, cells were equilibrated for at least 1 h in DMEM/F12 (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (Sigma-Aldrich), and penicillin/streptomycin (Nacalai Tesque).

For FP_{11} labeling, pPBbsr_EF1a-ΔSUMOstar-CFP2$_{11}$(xa)-mNeonGreen3Asp$_{11}$(xb)-sfCherry3Csp$_{11}$(xc) was cotransfected with pCMV-mPBase into HeLa$^{FP1-10x3}$ cells by using 293fectin or MDCK$^{FP1-10x3}$ cells using electroporation. Transfected cells were selected with 10 μg mL^{-1} blasticidin S (Calbiochem, San Diego, CA) and seeded onto collagen type I-coated glass-based 24-well plates.

Cells were observed with a Leica TCS SP8 FALCON confocal microscope (Leica-Microsystems, Wetzlar, Germany) equipped with an HC PL APO 20x/0.75 dry CS2 objective, an HC PL APO 40x/1.30 OIL CS2 objective, Leica HyD SMD detectors, a white light laser of 80 MHz pulse frequency, a Diode 405 (VLK 0550 T01; LASOS, Jena, Germany), a 440 nm diode laser (PDL 800-D; PicoQuant, Berlin, Germany), and a stage top incubator (Tokai Hit, Fujinomiya, Japan) to maintain 37 °C and 5% CO2. The following excitation wavelengths and emission band paths were used for the imaging: for EBFP2 imaging, 405 nm excitation, 410–450 nm emission; Cerulean or CFP2 imaging, 440 nm excitation, 460–490 nm emission; for sfGFP imaging, 488 nm excitation, 500–550 nm emission; for YFP1 imaging, 514 nm excitation, 520–550 nm emission; for mNG3A, mNG3K, or mNG3Asp imaging, 506 nm excitation, 510–550 nm emission; for sfCherry3C or sfCherry3Csp imaging, 594 nm excitation, 600–645 nm emission; for mRuby4 imaging, 561 nm excitation, 580–645 nm emission; for iRFP670 imaging, 650 nm excitation, 660–760 nm emission. To eliminate the background signal, the time gate for fluorescence detection was set from 1.0 ns to 6.0 ns.

In the multiplex imaging, the following excitation wavelengths and emission band paths were used: for CFP2 imaging, 440 nm excitation, 450–500 nm emission; for mNG3Asp imaging, 506 nm excitation, 515–580 nm emission; for sfCherry3Csp imaging, 594 nm excitation, 605–645 nm emission. To eliminate the background signal, the time gate for mNG3Asp and sfCherry3Csp fluorescence detection was set from 0.2 ns to 6.0 ns.

Images were processed and analyzed with ImageJ (*44*). For quantification of fluorescent intensity, images were segmented using Cellpose algorithm (*21*). Fluorescent intensity calculation was performed using custom MATLAB (MathWorks) and Python scripts.

Fluorescence spectral analysis

Fluorescence spectra were measured using a Leica TCS SP8 FALCON inverted microscope. Images were acquired with an HC PL APO 20x/0.75 dry CS2 objective using excitation at either 470 nm for split mNG3A or 546 nm for split sfCherry3C.

Flow cytometry analysis

Cells were suspended in PBS containing 3% FBS and analyzed or sorted with a FACS Aria IIu cell sorter (Becton Dickinson, Franklin Lakes, NJ). The following laser and emission filter combinations were used for fluorescence detection: CFP2, a 407 nm laser and an ET470/24m filter (Chroma Technology Corp., Bellows Falls, VT); mNeonGreen3Asp, a 488 nm laser, and

a DF530/30 filter; sfCherry3Csp, a 561 nm laser and a DF582/15 filter (Omega Optical); iRFP670, a 633 nm laser and a DF660/20 filter (Omega Optical). Cell debris and aggregates were excluded by gating for size and granularity. Data analysis was performed using FlowJo software (Tree Star, Ashland, OR).

Lentivirus infection

For lentivirus production, the EGF ligand or receptor expressing plasmid, psPAX2 (Addgene no. 12260), and pCMV-VSV-G-RSV-Rev (*45*) were co-transfected into Lenti-X 293T cells using polyethylenimine (no. 24765-1; Polyscience Inc.). The infected cells were selected with media containing the following antibiotics, depending on the drug resistance genes carried by the EGF ligand or receptor expressing plasmids: 200 μg ml^{-1} hygromycin (no. 31282-04-9; Wako).

Cell growth

For quantifying cell growth, MDCK cells labeled by Caterpie were seeded on collagen-coated 24-well glass-bottom plates (AGC Inc., Tokyo, Japan) at a density of 5×10^4 cells/ml. After 1 hour of incubation, the medium was replaced with DMEM/F12 supplemented with 100 units/ml penicillin, 100 μg/ml streptomycin, and 10% FBS. Following cell seeding, cells were observed using a confocal microscope (Leica SP8) every 24 hours for 3 days. For observation conditions, please refer to the "Fluorescent imaging" section. Cell numbers were counted using the Cellpose segmentation algorithm.

Computational design for enhanced β-barrel stability and split-strand complementation

The tertiary structures of split mNG3Asp and split sfCherry3Csp were modeled using ColabFold (*13*) implementation of AlphaFold Multimer (*14*) or AlphaFold2 (*17*) and visualized with PyMOL (http://www.pymol.org/). The highest-confidence model of the split sfCherry3Csp was then adopted as the starting coordinate for downstream Rosetta-based design.

To enhance global β-barrel stability of split sfCherry3Csp, *in silico* site-saturation mutagenesis (SSM) of all 224 positions in sfCherry3C□–□□ and sfCherry3C□□ was performed with a Rosetta-based workflow adapted from Thieker et al (*16*). For each variant, side-chain and backbone sampling were confined to a 10□Å sphere around the mutated residue, while the remainder of the protein was held fixed under coordinate constraints (sd = 1 outside the design sphere; sd = 2 within a surrounding “soft sphere”). Each mutant underwent three cycles of Cartesian FastRelax (MonomerRelax2019), and the change in Rosetta energy (ΔE) was calculated as the difference in total_score between each mutant and the wild-type model using the ref2015 score function.

To reinforce β□–□□/β□□ interface, two design workflows, Point-Mutation (PM) and Mutation Cluster (MC), were adapted from the workflow of Thieker et al. and modified for interface optimization. (1) In the PM workflow, SSM of the β□□ fragment (residues 1–16)

was carried out, with the addition of an InterfaceAnalyzer to compute binding energy (dG_separated) and packing statistics. Residues within 10 Å of each mutation site were sampled under the same constraint scheme used for stability design, while non-neighbor residues were held fixed; each point mutant was relaxed by Cartesian FastRelax prior to interface analysis, and final dG_separated values were calculated using the unconstrained ref2015 Rosetta score function. (2) In the MC workflow, the seed mutations from the PM step were proceeded into a FastDesign protocol using the InterfaceDesign2019 relax script to generate clustered mutations. The neighbor-selection radius was extended to 12 Å, and an inner design shell was defined by residues both in direct atomic contact (≤ 5–7 Å, via a CloseContact selector) and satisfying the InterfaceByVector geometric filter. A ResidueTypeConstraintGenerator was applied to inner-shell, biasing retention of wild-type identities. Critical chromophore-interacting residues were protected by marking them non-designable. Multi-residue combinations sampled via FastDesign, and each cluster variant was evaluated with InterfaceAnalyzer for dG_separated and packing quality. The scores were obtained under the unconstrained ref2015 function.

Calculation of angular data from individual cells

Three-dimensional fluorescence intensity data were acquired from individual cells by measuring CFP2, mNG3Asp, and sfCherry3Csp signals with background subtraction. These measurements were represented in a 3D coordinate system, with CFP2 intensities plotted on the x-axis, mNG3Asp on the y-axis, and sfCherry3Csp on the z-axis. For each cell, the Cartesian coordinates (x, y, z) were transformed into spherical coordinates (r, $Angle_{CvsG}$, $Angle_{RvsCG}$) using a custom-made Python script. The coordinate transformation was performed using standard spherical coordinate conversion equations.

$$r = (x^2 + y^2 + z^2)^{1/2}$$

$$Angle_{CvsG} = \tan^{-1}(y/x)$$

$$Angle_{RvsCG} = \cos^{-1}(z/r)$$

Two-component Gaussian mixture model (GMM) for FP_{11} tag screening

Fluorescence intensity measurements from individual cells across 96 distinct cell populations were transformed into angular coordinates ($Angle_{CvsG}$ and $Angle_{RvsCG}$). To enhance clustering robustness, outliers were removed by excluding the top and bottom 10% of data points that exhibited the greatest deviation from their respective cluster centers in both angular measurements. An unsupervised two-component GMM was implemented using scikit-learn (version 0.24.2; Pedregosa et al., 2011 (*46*)) to perform pairwise population comparisons. For random sampling, a predetermined number of cell populations were selected from a total of 96 populations using simple random sampling without replacement. All computational analyses were performed using custom Python scripts.

Cell classification using Gaussian mixture model (GMM)

Prior to angular calculations, we subtracted background and normalized the fluorescence intensities of CFP2, mNG3Asp, and sfCherry3Csp by dividing each value by the median intensity across all 20 cell types. Using these normalized fluorescence values, we then calculated $Angle_{CvsG}$ and $Angle_{RvsCG}$ measurements for subsequent analysis.

The classification model was trained using GMMs implemented in the scikit-learn library. we first fitted a single-Gaussian distribution to the training subset of each class individually; the resulting class-specific means and covariance matrices were then concatenated to obtain the GMM. The training dataset comprised 80% of the total data (~2.9x10^4 cells in total, n > 10^3 cells per cell population), consisting of two-dimensional features ($Angle_{CvsG}$ and $Angle_{RvsCG}$ measurements) from 20 distinct cell populations. The trained GMMs were then applied to angular data from pooled cells, and the predicted cell labels were visualized. For each class k, we calculated the probability of each input data point belonging to the corresponding GMM model, representing the likelihood of class membership:

$$P(x|\theta_k) = N(x|\mu_k,\Sigma_k)$$

where x is the input data point, θ_k represents the parameters of the k-th GMM model, and $N(x|\mu_k,\Sigma_k)$ denotes the normal distribution with mean μ_k and covariance Σ_k. Subsequently, the probabilities were normalized such that the sum of probabilities across all classes equals 1 for each data point:

$$P(k|x) = P(x|\theta_k) / \Sigma_k P(x|\theta_k)$$

We visualized the membership probabilities of the predicted classes. All computational analyses were performed using custom-made Python scripts.

Cell classification using K-means

We used $Angle_{CvsG}$ and $Angle_{RvsCG}$ as the feature space coordinates for clustering. K-means clustering was performed on the resulting dataset using scikit-learn's KMeans algorithm with 21 clusters (random_state = 0). As shown in Figure 6B, cells labeled with tag (BGR)=(0,0,4) corresponding to ID #4 exhibited a wide distribution of values in the $Angle_{CvsG}$-$Angle_{RvsCG}$ space, resulting in their identification as two distinct clusters that were subsequently merged. This accounts for the selection of 21 clusters in our analysis. The predicted cluster labels were then remapped to match the original cell population labels for validation. Classification accuracy was assessed by comparing the remapped cluster assignments with ground truth labels and visualized using a confusion matrix. The overall classification accuracy was calculated as the percentage of correctly classified cells. All calculations were performed using Python.

Clark-Evans Index

We analyzed the spatial distribution patterns of five cell types. Individual cells were first segmented using Cellpose, with centroids calculated and classified into five cell types via the Caterpie method. For spatial analysis, we extracted the coordinates of each cell type, calculated nearest neighbor distances using a k-dimensional tree algorithm, and computed the Clark-Evans index (R). This index compares observed mean distances to those expected under complete spatial randomness, where $R = 1$ indicates random distribution, $R < 1$ indicates clustering, and $R > 1$ indicates regularity. All calculations were performed using Python.

Cell count of the nearest neighbor

We analyzed the spatial relationships between different cell types using a nearest-neighbor approach. Cells were classified into five distinct populations based on fluorescent labels. For each cell of the target population (label 1), we calculated the distance to every cell of other populations (labels 2-5) and identified the closest neighboring cell using a k-dimensional tree algorithm. To assess the statistical significance of the observed patterns, we generated 100 randomized distributions for each dataset by shuffling cell positions while preserving the total number of each cell type. To determine whether the observed spatial associations between cell types differed significantly from random expectations, we calculated the ratio of the frequency with which each non-target cell type appeared as the nearest neighbor to target cells in the original distribution divided by the mean frequency observed across 100 randomized distributions. All calculations were performed using Python.

Statistics

All statistical analyses were carried out using Python. No statistical analysis was used to predetermine the sample size. An unpaired Welch's t-test was used to evaluate statistically significant differences. Data are expressed as the mean ± s.d. *P*-values of less than 0.05 were considered to be statistically significant in two-tailed tests and were classified into four categories: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and n.s. (not significant, i.e., $p \geq 0.05$).

Figure Legend

Figure 1. Caterpie: A cell identification system based on split fluorescent protein arrays
(A) Schematic overview of the Caterpie approach, illustrating cell identification through modular arrays of split fluorescent proteins. **(B)** Evolutionary relationships among split fluorescent proteins utilized in this study, presented as a phylogenetic tree. **(C)** Molecular architecture of expression constructs: Histone H2B fusions containing either full-length FP ($FP_{Full\text{-}length}$), single FP_{11} fragment, or tetrameric FP_{11} arrays, shown alongside the complementary $FP_{1\text{-}10}$ fragment. **(D)** Quantitative analysis of relative fluorescence intensity in HeLa cells expressing $FP_{Full\text{-}length}$, FP_{11}(x1) or FP_{11}(x4), with $FP_{1\text{-}10}$ fragments (1:4 transfection ratio). Data presented as bee swarm plots with median values indicated by black lines (n > 130 cells per condition). Measurements obtained through confocal microscopy.

Figure 2. Signal amplification and cell classification using split CFP2 arrays
(A) Molecular strategy for generating tandem $CFP2_{11}$ repeats, including detailed amino acid and nucleotide sequences at the $CFP2_{11}$(x1) insert-vector junction. **(B)** Construct architecture and expression analysis: Upper panel shows schematics $CFP2_{1\text{-}10}$ and Histone H2B fusions containing either full-length CFP2 ($CFP2_{Full\text{-}length}$) or varying copy numbers of $CFP2_{11}$ [$CFP2_{11}$(x1), (x2), (x4), and (x8)]. Lower panel presents representative confocal micrographs of HeLa cells stably co-expressing $CFP2_{1\text{-}10}$ with either $CFP2_{Full\text{-}length}$ or $CFP2_{11}$ variants. Scale bar, 20 µm. **(C)** Quantitative analysis of normalized CFP2 fluorescence in HeLa cells expressing Histone H2B-tagged $CFP2_{11}$ [(x1), (x2), (x4), or (x8)]. Data presented as bee swarm plots with median values (red lines); >1300 cells analyzed across three independent experiments. **(D)** Workflow schematic for Gaussian mixture model (GMM) and implementation. **(E)** Classification performance matrix showing prediction accuracy against true labels (rows). Data represented as mean ± SD of prediction accuracy from three independent experiments, with color intensity indicating classification accuracy. **(F)** Representative ratio image of pooled cells expressing different $CFP2_{11}$ copy number variants [(x1), (x2), (x4), or (x8)]; scale bar: 100 µm. **(G)** Cell classification map derived from panel (F), showing GMM-based assignment of individual cells to specific copy number variants [(x1), (x2), (x4), or (x8)]. **(H)** Visualization of GMM classification confidence through membership probability mapping of cells shown in panel (G).

Figure 3. Computational design of split sfCherry3C
(A) In-silico site-saturation mutagenesis of sfCherry3C . Left: schematic of the β residues targeted for saturation mutagenesis (black spheres). Right: heatmap of mean dG_separated values in Rosetta energy unit (REU) across three independent models for each amino acid substitution at each position. (B) Cluster-mutation design of sfCherry3C . Left: schematic showing a representative seed point mutation (yellow dot) and the shell of neighboring residues (yellow spheres) targeted for sequential cluster mutagenesis. Right:

heatmap of the mean dG_separated across ten cluster-mutated models derived from each seed mutation at every β□□ site; dot size indicates the median number of cluster-mutated residues across these ten models. (C) Schematics of mutant variants of sfCherry3C$_{1\text{-}10}$ (sfCherry3C$_{1\text{-}10M}$) and Histone H2B-tagged mutant variant of sfCherry3C$_{11}$ (sfCherry3C$_{11M}$). (D) List of mutant variants of sfCherry3C$_{1\text{-}10M}$ and sfCherry3C$_{11M}$. Bee swarm plot showing normalized sfCherry3C fluorescence intensity of HeLa cells stably co-expressing Histone H2B-tagged sfCherry3C$_{11M}$(x1) with sfCherry3C$_{1\text{-}10M}$. Red lines represent the median. (E) Rosetta models of split sfCherry3Csp. The overall β-barrel is shown as a cartoon colored by fragment (gray, β□–□□; magenta, β□□). Each inset zooms on an engineered site, showing the parental conformation on the left and the corresponding design on the right, with key residues rendered as sticks representation. All models were visualized in PyMOL. (F) Emission spectra of HeLa cells expressing split sfCherry3C or split sfCherry3Csp with 546 nm laser excitation.

Figure 4. Optimization of split mNG3A through structure-guided C-terminal engineering

(A) Predicted structural model of native mNG3A11(x8) generated using ColabFold (AlphaFold2) and visualized with PyMOL. Colors represent hydrophobicity according to the Eisenberg hydrophobicity scale. **(B)** Predicted structural model of parental split mNG3A generated using ColabFold (AlphaFold Multimer) and visualized with PyMOL. Color-coded domains: mNG3A$_{1\text{-}10}$ (green) and mNG3A$_{11}$ (light green). **(C)** Comparative sequence analysis of original mNG3A$_{11}$ and engineered mutant variants (mNG3A$_{11M}$). **(D)** Schematic representation of expression constructs: mNG3A$_{1\text{-}10}$ and Histone H2B fusion with the mNG3A$_{11M}$(x8) array. Quantitative analysis of normalized mNG3A fluorescence in HeLa cells expressing Histone H2B-tagged mNG3A$_{11}$ [(x1), (x2), (x4), or (x8)]. Data presented as bee swarm plots with median values (red lines). **(E)** Quantitative analysis of normalized mNG3A fluorescence intensity in HeLa cells stably co-expressing Histone H2B-tagged mNG3A$_{11M}$ (x8) with mNG3A$_{1\text{-}10}$. Median values indicated by red lines. Values normalized to set the median value of mNG3A$_{11M1}$(x8) to 1. **(F)** Emission spectra of HeLa cells expressing split mNG3A or split mNG3Asp with 470 nm laser excitation. **(G)** Comparative analysis of relative fluorescence intensity of HeLa cells expressing FP$_{\text{Full-length}}$, FP$_{11}$(x1) or FP$_{11}$(x4), with cognate FP$_{1\text{-}10}$ fragments (1:4 transfection ratio). Data presented as bee swarm plots with median values (red lines); n > 130 cells per condition. Measurements obtained through confocal microscopy.

Figure 5. Systematic selection and validation of FP$_{11}$ tags for Caterpie implementation

(A) Schematic representation of FP$_{11}$ tags: ΔSUMO-tagged FP$_{11}$ arrays comprising paired repeats of CFP2$_{11}$, mNG3Asp$_{11}$, and sfCherry3Csp$_{11}$ [(x0), (x1), (x2), (x4), (x8) each]. Comprehensive catalog of the FP$_{11}$ tag library comprising 96 distinct combinations, detailing copy numbers of CFP2$_{11}$ (labeled as "C"), mNG3Asp$_{11}$ (labeled as "G"), and sfCherry3Csp$_{11}$

(labeled as "R") fragments. The marks indicate the FP_{11} tags selected during primary and secondary screening processes, and those finally chosen. **(B)** Schematic representation of expression constructs: $FP_{1\text{-}10}$x3. **(C)** Analysis scheme for classification of cell populations expressing various FP_{11} tags. **(D)** Matrix analysis of pairwise classification accuracy across 96 cell populations expressing various FP_{11} tags, arranged by FP_{11} tag numbers shown in panel (A). (n > 230 cells per population). (**E)** Analysis scheme for determining the highest achievable minimum pairwise classification accuracy among randomly selected FP_{11}-tagged subpopulations. **(F)** Graph showing the highest achievable minimum pairwise classification accuracy among randomly selected FP_{11}-tagged subpopulations. For each subpopulation size, minimum accuracy between all possible pairs was determined from 10^7 random selections, with maximum values plotted. **(G)** Comparative performance analysis of 30 candidate populations (detailed in fig. S4) using both confocal microscopy (blue) and flow cytometry (orange). Analysis methodology follows that of panel (F). Data presented as mean ± SD from three independent experiments.

Figure 6. High-fidelity discrimination of twenty distinct cell populations using Caterpie
(A) Representative multicolor fluorescence micrographs of 20 distinct cell populations expressing unique FP_{11} tag combinations. Fluorescence channels: CFP2 (blue), mNG3Asp (green), and sfCherry3Csp (red). Copy numbers of each FP_{11} variant ($CFP2_{11}$, $mNG3Asp_{11}$, and $sfCherry3Csp_{11}$) are indicated in the upper left corner of each panel. Scale bars: 30 µm. **(B)** Left: Scatter plot showing $Angle_{CvsG}$ versus $Angle_{RvsCG}$ distributions for 20 distinct cell populations from panel (A). (n > 1000 cells per population) Center: Distribution of label assignments in $Angle_{CvsG}$ and $Angle_{RvsCG}$ based on Gaussian mixture model (GMM) classification. Right: Distribution of label membership probabilities in $Angle_{CvsG}$ and $Angle_{RvsCG}$. **(C)** Classification performance matrix showing prediction accuracy against true population identities (rows). Color intensity indicates classification accuracy. Overall average accuracy, 97 %. **(D)** Large-field composite image of pooled cell populations from panel (A), displaying CFP2 (blue), mNG3Asp (green), and sfCherry3Csp (red) fluorescence channels. Scale bar: 500 µm. **(E)** Detailed analysis of region indicated by dashed box in panel (D). Left: Higher magnification of selected region. Center: Population assignment map following GMM-based classification into 20 distinct categories. Right: Visualization of GMM classification reliability through membership probability mapping.

Figure 7. Interaction dynamics in multi-population cell cultures of MDCK cells expression EGF-family ligands and their receptors
(A) Schematic diagram showing the association pairs between EGF ligands (HBEGF, EREG) and EGF receptors (EGFR, ErbB2). **(B)** Representative pseudocolored image of five distinct cell types cultured together and identified using the Caterpie method. Yellow: WT, Pink: EREG, Cyan: HBEGF, Green: EGFR, Orange: ErbB4; Copy numbers of each FP11 variant (CFP211, mNG3Asp11, and sfCherry3Csp11) are (1,2,4), (2,0,4), (4,0,1), (2,4,0), and (8,2,1),

respectively. Scale bar: 1 mm. **(C)** Individual display of the five cell types shown in panel E. The upper left corner of each panel indicates the re-expressed EGF signaling molecules and the copy numbers of FP_{11} tags. **(D)** Graph showing the Clark-Evans Index calculated for each of the five cell types. Data from three independent experiments. Black lines represent the mean values across the three experiments. **(E)** Normalized nearest neighbor cell analysis for evaluating spatial relationships between cells. This figure shows the ratio of heterotypic nearest neighbor counts observed in the actual data to those counted after randomly reassigning the labels of five cell types in the same dataset. Data from three independent experiments. Black lines represent the mean values across the three experiments.

Supplementary Figure S1. Comprehensive mutational scanning of split sfCherry3C

Heatmap displaying the ΔEnergy values for saturating mutations in both $sfCherry_{1-10}$ and $sfCherry_{11}$.

Supplementary Figure S2. Optimization of signal amplification using truncated SUMOstar fusion strategy

(A) Schematic representation of expression constructs: $CFP2_{Full-length}$-NES and $CFP2_{11}$ variants [(x1) or (x8)]-NES with or without GB1 or ΔSUMOstar (truncated SUMOstar) tags. **(B)** Amino acid sequences of SUMOstar, with ΔSUMOstar region highlighted in bold. **(C)** Bee swarm plots showing quantitative analysis of normalized CFP2 fluorescence in HeLa cells co-expressing $CFP2_{1-10}$ with either $CFP2_{Full-length}$-NES or $CFP2_{11}$(x1)/$CFP2_{11}$(x8)-NES variants, with or without GB1 or ΔSUMOstar tags (constructs shown in panel A). Median values are indicated by red lines. **(D)** Construct architecture and expression analysis: Upper panel shows schematics of ΔSUMOstar fusions containing varying copy numbers of $CFP2_{11}$ [$CFP2_{11}$(x1), (x2), (x4), and (x8)]. Lower panel presents representative confocal micrographs of HeLa cells stably co-expressing $CFP2_{1-10}$ with either $CFP2_{Full-length}$-NES or ΔSUMOstar-tagged $CFP2_{11}$ variant. Scale bar: 20 μm. **(E)** Comparative analysis of normalized CFP2 fluorescence in HeLa cells expressing ΔSUMOstar-tagged $CFP2_{11}$ variants [(x1), (x2), (x4), or (x8)] shown in panel D. Data presented as bee swarm plots with median values indicated by red lines. **(F)** Classification performance matrix showing predictive accuracy against true population identities (rows), with color intensity indicating classification accuracy.

Supplementary Figure S3. Influence of FP_{11} fragment order on fluorescence signal distribution

(A) Schematic representation of expression constructs: ΔSUMO-tagged FP_{11} arrays comprising paired repeats of $CFP2_{11}$, $mNG3Asp_{11}$, and $sfCherry3Csp_{11}$ [(x2) each, total x6]. **(B)** Quantitative analysis presented as bee swarm plots showing normalized fluorescence intensities (CFP2, mNG3Asp, and sfCherry3Csp) in HeLa cells expressing six different permutations of FP_{11} repeat arrangements, with concurrent FP_{1-10}x3 expression. Data obtained

by confocal microscopy, with median values indicated by black lines.

Supplementary Figure S4. Comparative analysis of binary classification performance across 30 selected cell populations

(A) Catalog of 30 selected FP_{11} tag variants (detailed in Fig. 5A), detailing identification numbers and copy numbers of $CFP2_{11}$, $mNG3Asp_{11}$, and $sfCherry3Csp_{11}$ fragments. **(B)** Matrix analysis of pairwise classification accuracies for 30 cell populations expressing distinct FP_{11} tag combinations (listed in panel A), evaluated by both microscopy (upper panel) and flow cytometry (lower panel). Data represent results from three independent experiments.

Supplementary Figure S5. Binary classification performance analysis of final 20 cell population panel

(A) Catalog of optimized FP_{11} tag set comprising 20 selected variants (detailed in Fig. 5A), detailing identification numbers and copy numbers of $CFP2_{11}$, $mNG3Asp_{11}$, and $sfCherry3Csp_{11}$ fragments. **(B)** Matrix analysis of pairwise classification accuracy for 20 cell populations expressing distinct FP_{11} tag combinations (defined in panel A). Comparative evaluation performed using both microscopy (left panel) and flow cytometry (right panel). Data represent mean values from three independent experiments in Supplementary Figure S3.

Supplementary Figure S6. Aggregation character of $sfCherry3Csp_{11}$(x8)

Representative images of sfCherry3Csp fluorescence in HeLa cells expressing ΔSUMOstar-tagged $sfCherry3Csp_{11}$ variants [(x4), or (x8)], with concurrent $FP_{1\text{-}10}$x3 expression.

Supplementary Figure S7. Unsupervised k-means clustering of 20 cell populations

(A) Left: Scatter plot showing the distribution of $Angle_{CvsG}$ versus $Angle_{RvsCG}$ for 20 distinct cell populations (n > 1,000 cells per population). Right: Scatter plot showing the distribution of $Angle_{CvsG}$ versus $Angle_{RvsCG}$ for individual cells grouped into 20 distinct populations using unsupervised k-means clustering. **(B)** Classification performance matrix of unsupervised k-means classification in panel (A). Color intensity indicates classification accuracy. Overall average accuracy, 96 %.

Supplementary Figure S8. High-fidelity discrimination of twenty distinct cell populations in MDCK cells using Caterpie

(A) Representative multicolor fluorescence micrographs of 20 distinct cell populations expressing FP_{11} tags. Fluorescence channels: CFP2 (blue), mNG3Asp (green), and sfCherry3Csp (red). Copy numbers of each FP_{11} variant ($CFP2_{11}$, $mNG3Asp_{11}$, and $sfCherry3Csp_{11}$) are indicated in the top of each panel. Scale bars: 30 μm. **(B)** Classification performance matrix showing prediction accuracy against true population identities (rows). Color intensity indicates classification accuracy. Overall average accuracy, 96%. **(C)** Top: Large-field composite image of pooled cell populations from panel (A), displaying CFP2

(blue), mNG3Asp (green), and sfCherry3Csp (red) fluorescence channels. Scale bar: 500 µm. Middle: Population assignment map following GMM-based classification into 20 distinct categories. Bottom: Visualization of GMM classification confidence through membership probability mapping.

Supplementary Figure S9. Cell growth of Caterpie labeled MDCK cells

Growth curves showing cell density of 20 different MDCK cell types fluorescently labeled with the Caterpie method measured every 24 hours. The lower panels display individual growth curves for each cell type.

References

1. J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, J. W. Lichtman, Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature 450, 56-62 (2007).
2. M. W. Veling, Y. Li, M. T. Veling, C. Litts, N. Michki, H. Liu, B. Ye, D. Cai, Identification of neuronal lineages in the Drosophila peripheral nervous system with a "digital" multi-spectral lineage tracing system. Cell Rep. 29, 3303-3312.e3 (2019).
3. S. Cabantous, T. C. Terwilliger, G. S. Waldo, Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. Nat. Biotechnol. 23, 102-107 (2005).
4. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, A scalable strategy for high-throughput GFP tagging of endogenous human proteins. Proc. Natl. Acad. Sci. U.S.A. 113, E3501-E3508 (2016).
5. N. H. Cho, K. C. Cheveralls, A. D. Brunner, K. Kim, A. C. Michaelis, P. Raghavan, H. Kobayashi, L. Savy, J. Y. Li, H. Canaj, J. Y. S. Kim, E. M. Stewart, C. Gnann, F. McCarthy, J. P. Cabrera, R. M. Brunetti, B. B. Chhun, G. Dingle, M. Y. Hein, B. Huang, S. B. Mehta, J. S. Weissman, R. Gómez-Sjöberg, D. N. Itzhak, L. A. Royer, M. Mann, M. D. Leonetti, OpenCell: Endogenous tagging for the cartography of human cellular organization. Science 375, eabi6983 (2022).
6. D. Kamiyama, S. Sekine, B. Barsi-Rhyne, J. Hu, B. Chen, L. A. Gilbert, H. Ishikawa, M. D. Leonetti, W. F. Marshall, J. S. Weissman, B. Huang, Versatile protein tagging in cells with split fluorescent protein. Nat. Commun. 7, 11046 (2016).
7. S. He, A. Cuentas-Condori, D. M. Miller, Strategy for bright, tissue-specific GFP labeling of endogenous proteins in C. elegans. Genetics 212, 387-395 (2019).
8. J. Goudeau, C. S. Sharp, J. Paw, L. Savy, M. D. Leonetti, A. G. York, D. L. Updike, C. Kenyon, M. Ingaramo, Split-wrmScarlet and split-sfGFP: tools for faster, easier fluorescent labeling of endogenous proteins in Caenorhabditis elegans. Genetics 217, iyab014 (2021).
9. R. Kamiyama, K. Banzai, P. Liu, A. Marar, R. Tamura, F. Jiang, M. A. Fitch, J. Xie, D. Kamiyama, Cell-type-specific, multicolor labeling of endogenous proteins with split fluorescent protein tags in Drosophila. Proc. Natl. Acad. Sci. U.S.A. 118, e2024690118 (2021).
10. R. Tamura, F. Jiang, J. Xie, D. Kamiyama, Multiplexed labeling of cellular proteins with split fluorescent protein tags. Commun. Biol. 4, 257 (2021).
11. S. Zhou, S. Feng, D. Brown, B. Huang, Improved yellow-green split fluorescent proteins for protein labeling and signal amplification. PLoS One 15, e0242592 (2020).
12. S. Feng, A. Varshney, D. Coto Villa, C. Modavi, J. Kohler, F. Farah, S. Zhou, N. Ali, J. D. Müller, M. K. Van Hoven, B. Huang, Bright split red fluorescent proteins for the

visualization of endogenous proteins and synapses. Commun. Biol. 2, 344 (2019).

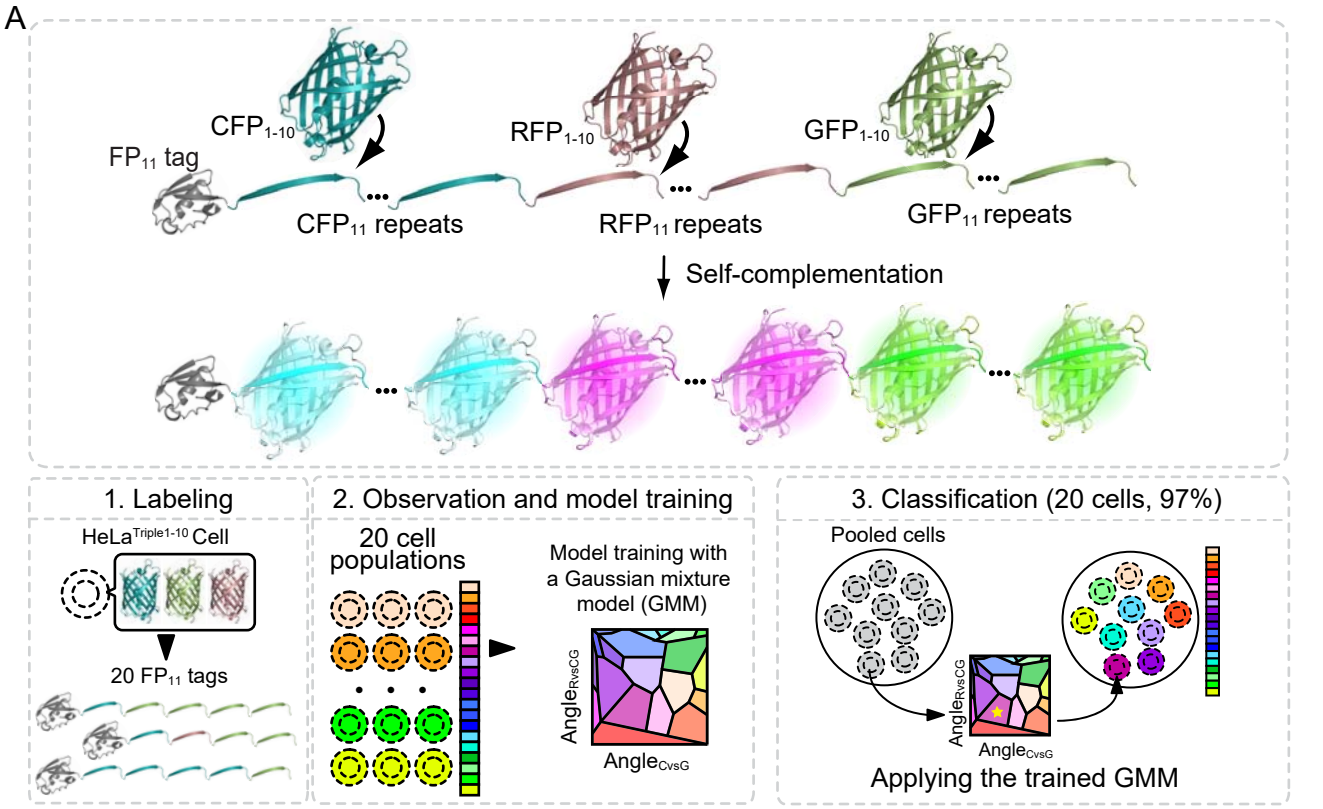
13. G. Kim, S. Lee, E. Levy Karin, H. Kim, Y. Moriwaki, S. Ovchinnikov, M. Steinegger, M. Mirdita, Easy and accurate protein structure prediction using ColabFold. Nat. Protoc. 20, 620–642 (2025).
14. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 [Preprint] (2022); https://doi.org/10.1101/2021.10.04.463034.
15. A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, P. Bradley, ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
16. D. F. Thieker, J. B. Maguire, S. T. Kudlacek, A. Leaver-Fay, S. Lyskov, B. Kuhlman, Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. Protein Sci. 31, e4428 (2022).
17. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589 (2021).
18. P. Zhou, A. A. Lugovskoy, G. Wagner, A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. J. Biomol. NMR 20, 11-14 (2001).
19. L. Liu, J. Spurrier, T. R. Butt, J. E. Strickler, Enhanced protein expression in the baculovirus/insect cell system using engineered SUMO fusions. Protein Expr. Purif. 62, 21-28 (2008).
20. N. Milech, B. A. Longville, P. T. Cunningham, M. N. Scobie, H. M. Bogdawa, S. Winslow, M. Anastasas, T. Connor, F. Ong, S. R. Stone, M. Kerfoot, T. Heinrich, K. M. Kroeger, Y. F. Tan, K. Hoffmann, W. R. Thomas, P. M. Watt, R. M. Hopkins, GFP-complementation assay to detect functional CPP and protein delivery into living cells. Sci. Rep. 5, 18329 (2015).
21. C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, Cellpose: a generalist algorithm for cellular segmentation. Nat. Methods 18, 100-106 (2021).
22. E. Deguchi, S. Lin, D. Hirayama, K. Matsuda, A. Tanave, K. Sumiyama, S. Tsukiji, T.

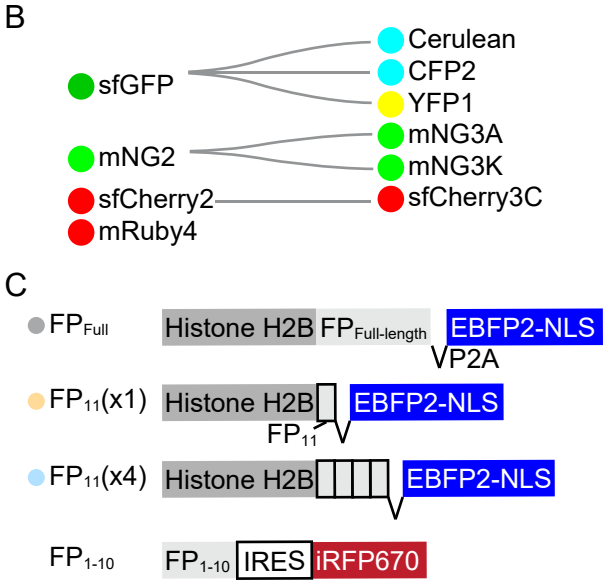
Otani, M. Furuse, A. Sorkin, M. Matsuda, K. Terai, Low-affinity ligands of the epidermal growth factor receptor are long-range signal transmitters in collective cell migration of epithelial cells. Cell Rep. 43, 114986 (2024).

23. H. J. Snippert, L. G. van der Flier, T. Sato, J. H. van Es, M. van den Born, C. Kroon-Veenboer, N. Barker, A. M. Klein, J. van Rheenen, B. D. Simons, H. Clevers, Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. Cell 143, 134-144 (2010).
24. I. Tabansky, A. Lenarcic, R. W. Draft, K. Loulier, D. B. Keskin, J. Rosains, J. Rivera-Feliciano, J. W. Lichtman, J. Livet, J. N. H. Stern, J. R. Sanes, K. Eggan, Developmental bias in cleavage-stage mouse blastomeres. Curr. Biol. 23, 21-31 (2013).
25. C. Ghigo, I. Mondor, A. Jorquera, N. J. Nowak, S. Wienert, S. P. Zahner, B. E. Clausen, H. Luche, B. Malissen, F. Klauschen, M. Bajénoff, Multicolor fate mapping of Langerhans cell homeostasis. J. Exp. Med. 210, 1657-1664 (2013).
26. D. Cai, K. B. Cohen, T. Luo, J. W. Lichtman, J. R. Sanes, Improved tools for the Brainbow toolbox. Nat. Methods 10, 540-547 (2013).
27. K. Loulier, R. Barry, P. Mahou, Y. Le Franc, W. Supatto, K. S. Matho, S. Ieng, S. Fouquet, E. Dupin, R. Benosman, A. Chédotal, E. Beaurepaire, X. Morin, J. Livet, Multiplex cell and lineage tracking with combinatorial labels. Neuron 81, 505-520 (2014).
28. K. Weber, U. Bartsch, C. Stocking, B. Fehse, A multicolor panel of novel lentiviral "gene ontology" (LeGO) vectors for functional gene analysis. Mol. Ther. 16, 698-706 (2008).
29. J. García-Marqués, L. López-Mascaraque, Clonal identity determines astrocyte cortical heterogeneity. Cereb. Cortex 23, 1463-1472 (2013).
30. F. García-Moreno, N. A. Vasistha, J. Begbie, Z. Molnár, CLoNe is a new method to target single progenitors and study their progeny in mouse and chick. Development 141, 1589-1598 (2014).
31. R. Sakaguchi, M. N. Leiwe, T. Imai, Bright multicolor labeling of neuronal circuits with fluorescent proteins and chemical tags. eLife 7, e40350 (2018).
32. T. Kumamoto, F. Maurinot, R. Barry-Martinet, C. Vaslin, S. Vandormael-Pournin, M. Le, M. Lerat, D. Niculescu, M. Cohen-Tannoudji, A. Rebsam, K. Loulier, S. Nedelec, S. Tozer, J. Livet, Direct readout of neural stem cell transgenesis with an integration-coupled gene expression switch. Neuron 107, 617-630.e6 (2020).
33. S. Feng, S. Sekine, V. Pessino, H. Li, M. D. Leonetti, B. Huang, Improved split fluorescent proteins for endogenous protein labeling. Nat. Commun. 8, 370 (2017).
34. K. A. Matreyek, J. J. Stephany, M. A. Chiasson, N. Hasle, D. M. Fowler, An improved platform for functional assessment of large protein libraries in mammalian cells. Nucleic Acids Res. 48, e1 (2020).
35. T. Köker, A. Fernandez, F. Pinaud, Characterization of split fluorescent protein variants and quantitative analyses of their self-assembly process. Sci. Rep. 8, 5344 (2018).

36. D. M. Shcherbakova, V. V. Verkhusha, Near-infrared fluorescent proteins for multicolor in vivo imaging. Nat. Methods 10, 751-754 (2013).
37. T. Akagi, K. Sasai, H. Hanafusa, Refractory nature of normal human diploid fibroblasts with respect to oncogene-mediated transformation. Proc. Natl. Acad. Sci. U.S.A. 100, 13567-13572 (2003).
38. K. Kawakami, H. Takeda, N. Kawakami, M. Kobayashi, N. Matsuda, M. Mishina, A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. Dev. Cell 7, 133-144 (2004).
39. N. Komatsu, K. Terai, A. Imanishi, Y. Kamioka, K. Sumiyama, T. Jin, Y. Okada, T. Nagai, M. Matsuda, A platform of BRET-FRET hybrid biosensors for optogenetics, chemical screening, and in vivo imaging. Sci. Rep. 8, 8984 (2018).
40. J. H. Kim, S. R. Lee, L. H. Li, H. J. Park, J. H. Park, K. Y. Lee, M. K. Kim, B. A. Shin, S. Y. Choi, High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. PLoS One 6, e18556 (2011).
41. D. Kalderon, B. L. Roberts, W. D. Richardson, A. E. Smith, A short amino acid sequence able to specify nuclear location. Cell 39, 499-509 (1984).
42. U. Fischer, J. Huber, W. C. Boelens, L. W. Mattaj, R. Lührmann, The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. Cell 82, 475-483 (1995).
43. K. Yusa, R. Rad, J. Takeda, A. Bradley, Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. Nat. Methods 6, 363-369 (2009).
44. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671-675 (2012).
45. Miyoshi H., Blomer U., Takahashi M., Gage F. H., Verma I. M., *J. Virol.* ***72***, 8150 (1998).
46. D. K. Barupal, O. Fiehn, Generating the blood exposome database using a comprehensive text mining and database fusion approach. Environ. Health Perspect. 127, 97008 (2019).

Fig.1


A
FP11 tag
CFP1-10
RFP1-10
GFP1-10
CFP11 repeats
RFP11 repeats
GFP11 repeats
Self-complementation
1. Labeling
HeLaTriple1-10 Cell
20 FP11 tags
2. Observation and model training
20 cell populations
Model training with a Gaussian mixture model (GMM)
AngleRvsCG
AngleCvsG
3. Classification (20 cells, 97%)
Pooled cells
Applying the trained GMM


B
sfGFP
mNG2
sfCherry2
mRuby4
Cerulean
CFP2
YFP1
mNG3A
mNG3K
sfCherry3C
C
FPFull
Histone H2B
FPFull-length
EBFP2-NLS
VP2A
FP11(x1)
FP11
FP11(x4)
FP1-10
IRES
iRFP670

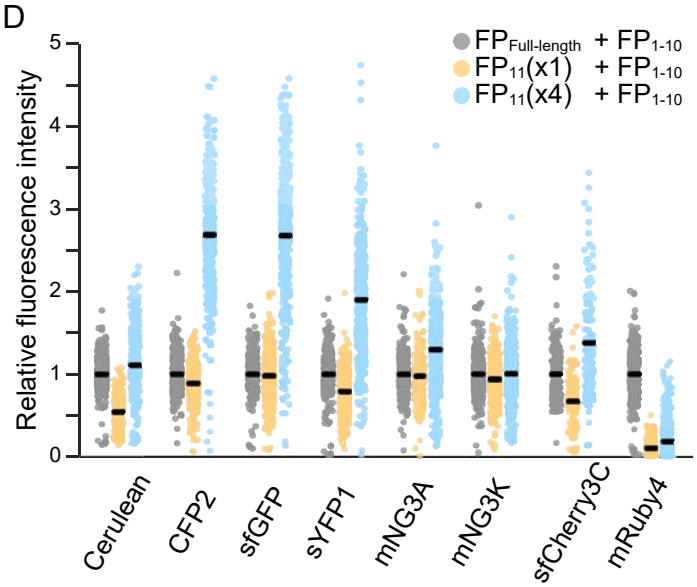

D
Relative fluorescence intensity
FPFull-length + FP1-10
FP11(x1) + FP1-10
FP11(x4) + FP1-10
Cerulean
CFP2
sfGFP
sYFP1
mNG3A
mNG3K
sfCherry3C
mRuby4

Fig.2

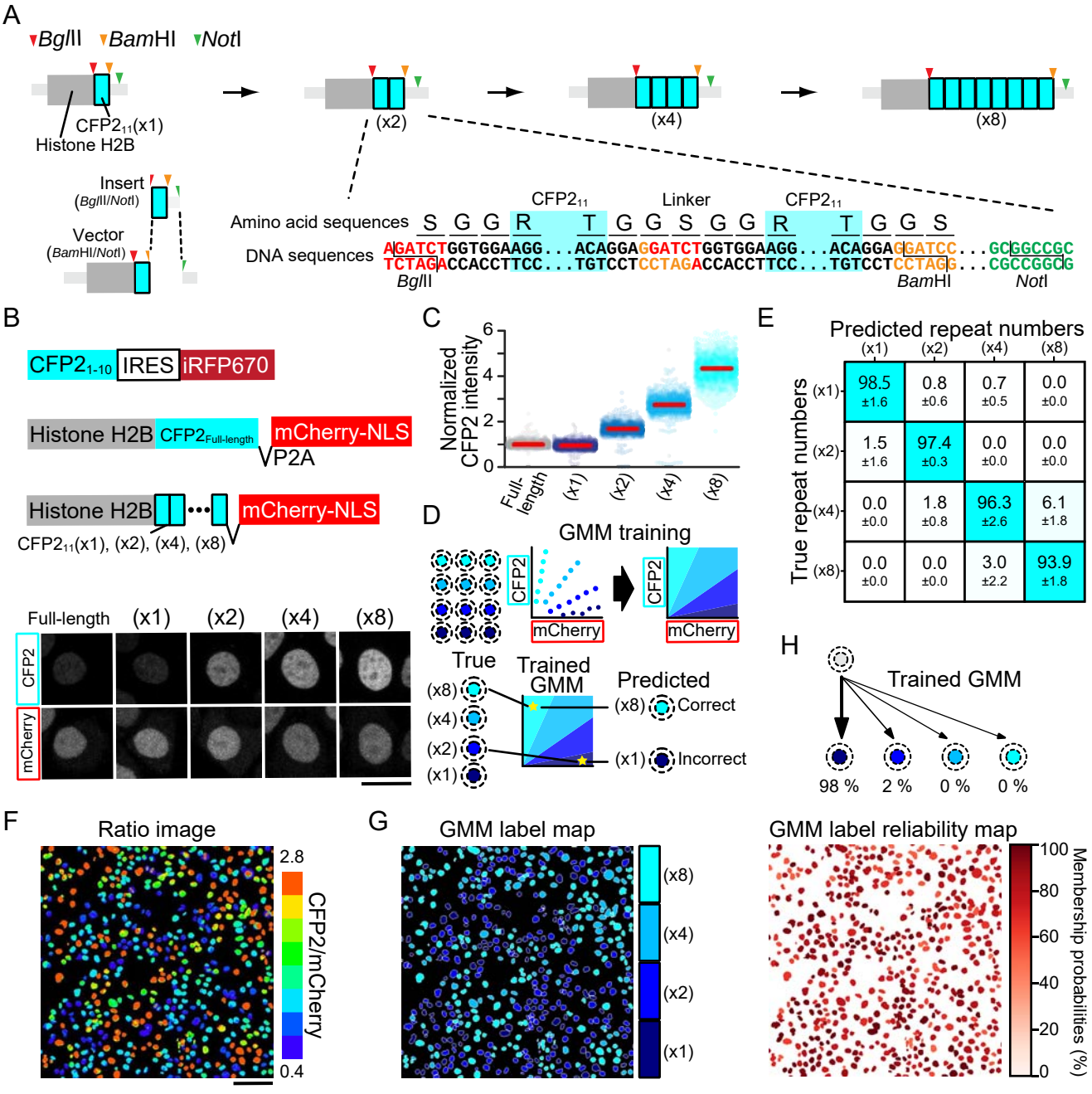
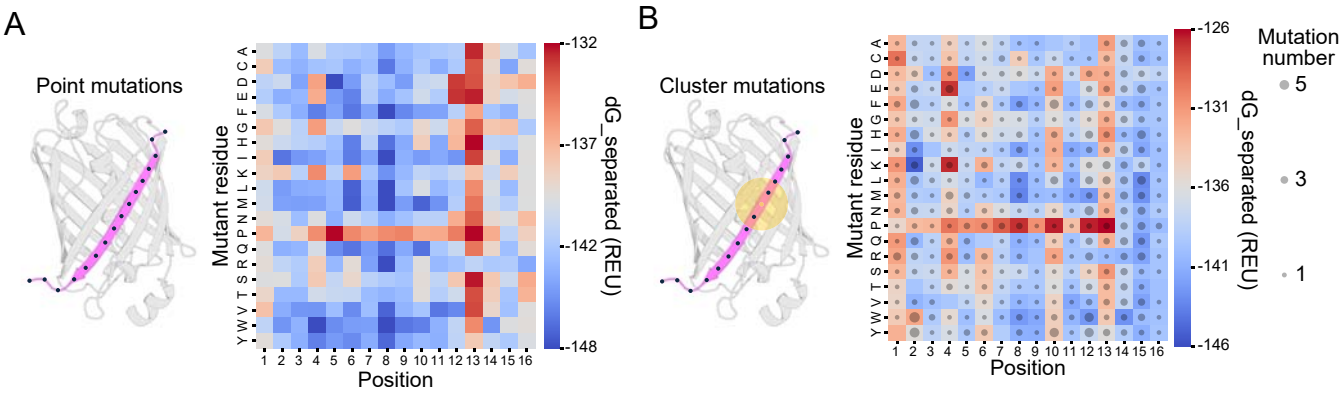
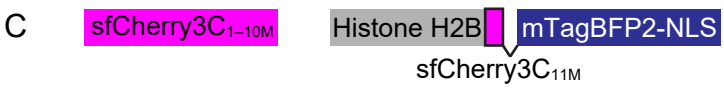
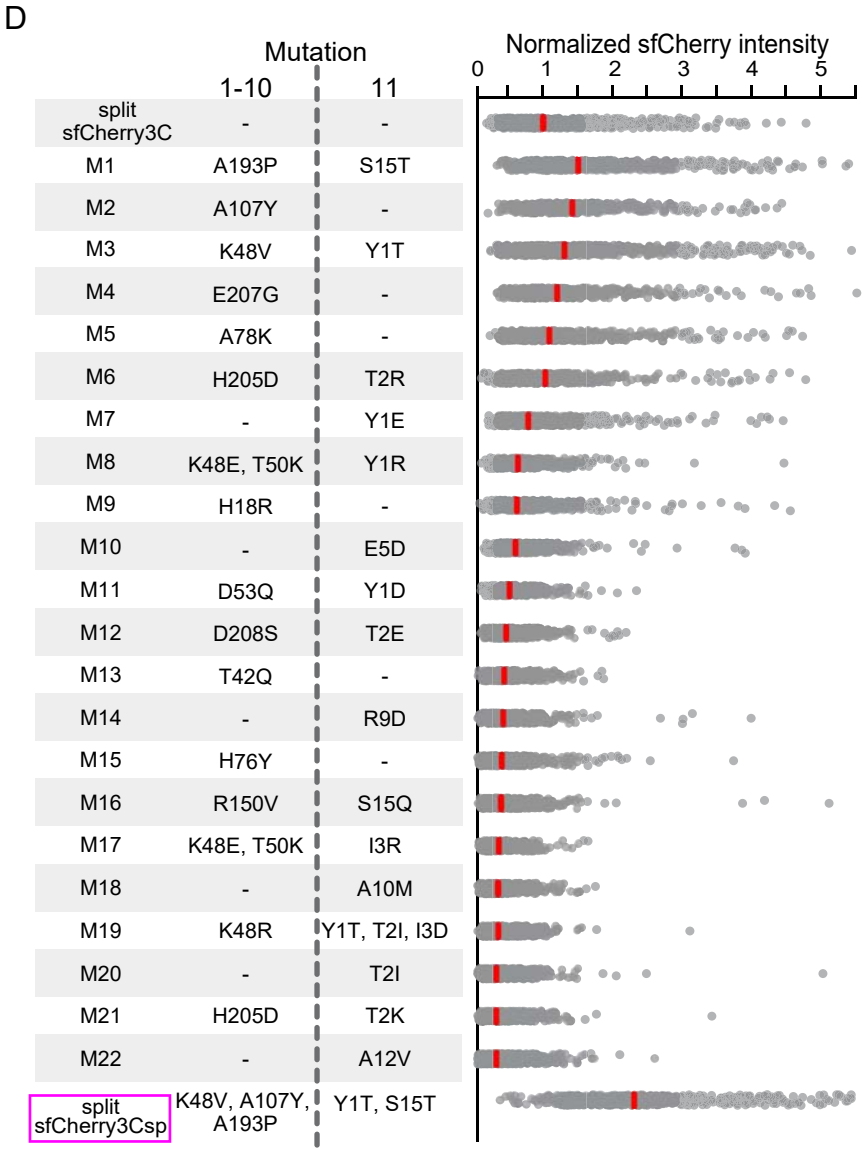
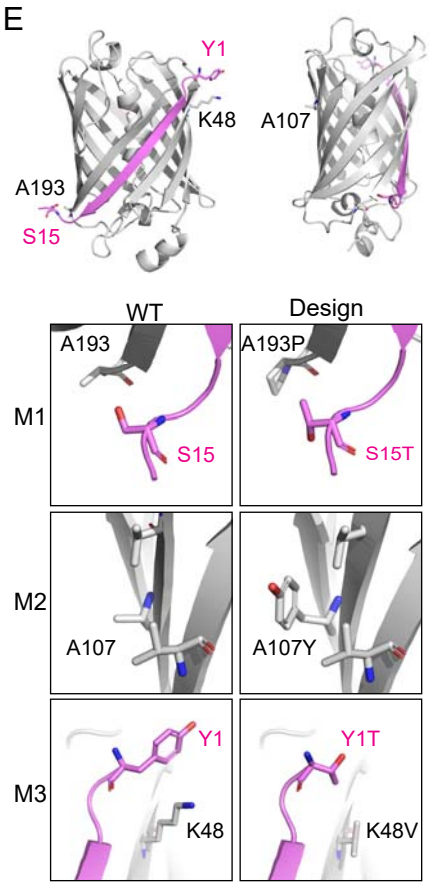

A
BglII
BamHI
NotI
CFP2₁₁(x1)
Histone H2B
(x2)
(x4)
(x8)
Insert
(BglII/NotI)
Vector
(BamHI/NotI)
Amino acid sequences
CFP2₁₁
Linker
CFP2₁₁
S G G R T G G S G G R T G G S
DNA sequences
AGATCTGGTGGAAGG...ACAGGAGGATCTGGTGGAAGG...ACAGGAGGATCC...GCGGCCGC
TCTAGACCACCTTCC...TGTCCTCCTAGACCACCTTCC...TGTCCTCCTAGG...CGCCGGCG
BglII
BamHI
NotI
B
CFP2₁₋₁₀
IRES
iRFP670
Histone H2B
CFP2Full-length
mCherry-NLS
P2A
Histone H2B
mCherry-NLS
CFP2₁₁(x1), (x2), (x4), (x8)
Full-length
(x1)
(x2)
(x4)
(x8)
CFP2
mCherry
C
Normalized CFP2 intensity
Full-length
(x1)
(x2)
(x4)
(x8)
D
GMM training
CFP2
mCherry
True
Trained GMM
Predicted
(x8)
(x4)
(x2)
(x1)
Correct
Incorrect
E
Predicted repeat numbers
True repeat numbers
(x1) (x2) (x4) (x8)
98.5 ±1.6 | 0.8 ±0.6 | 0.7 ±0.5 | 0.0 ±0.0
1.5 ±1.6 | 97.4 ±0.3 | 0.0 ±0.0 | 0.0 ±0.0
0.0 ±0.0 | 1.8 ±0.8 | 96.3 ±2.6 | 6.1 ±1.8
0.0 ±0.0 | 0.0 ±0.0 | 3.0 ±2.2 | 93.9 ±1.8
H
Trained GMM
98 %
2 %
0 %
0 %
F
Ratio image
2.8
0.4
CFP2/mCherry
G
GMM label map
(x8)
(x4)
(x2)
(x1)
GMM label reliability map
100
80
60
40
20
0
Membership probabilities (%)

Fig.3


A
Point mutations
Mutant residue
Position
dG_separated (REU)
-132
-137
-142
-148
B
Cluster mutations
Mutant residue
Position
dG_separated (REU)
-126
-131
-136
-141
-146
Mutation number
5
3
1


C
sfCherry3C1–10M
Histone H2B
mTagBFP2-NLS
sfCherry3C11M


D
Mutation
1-10
11
Normalized sfCherry intensity
0
1
2
3
4
5
split sfCherry3C - -
M1 A193P S15T
M2 A107Y -
M3 K48V Y1T
M4 E207G -
M5 A78K -
M6 H205D T2R
M7 - Y1E
M8 K48E, T50K Y1R
M9 H18R -
M10 - E5D
M11 D53Q Y1D
M12 D208S T2E
M13 T42Q -
M14 - R9D
M15 H76Y -
M16 R150V S15Q
M17 K48E, T50K I3R
M18 - A10M
M19 K48R Y1T, T2I, I3D
M20 - T2I
M21 H205D T2K
M22 - A12V
split sfCherry3Csp K48V, A107Y, A193P Y1T, S15T


E
Y1
K48
A107
A193
S15
WT
Design
M1
A193
S15
A193P
S15T
M2
A107
A107Y
M3
Y1
K48
Y1T
K48V

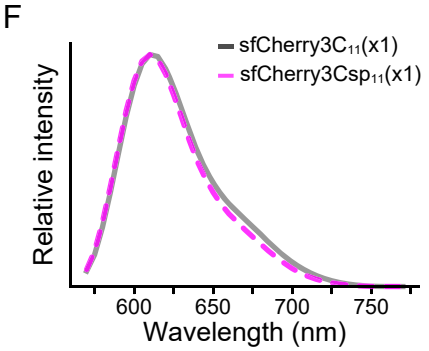

F
sfCherry3C11(x1)
sfCherry3Csp11(x1)
Relative intensity
600
650
700
750
Wavelength (nm)

Fig.4

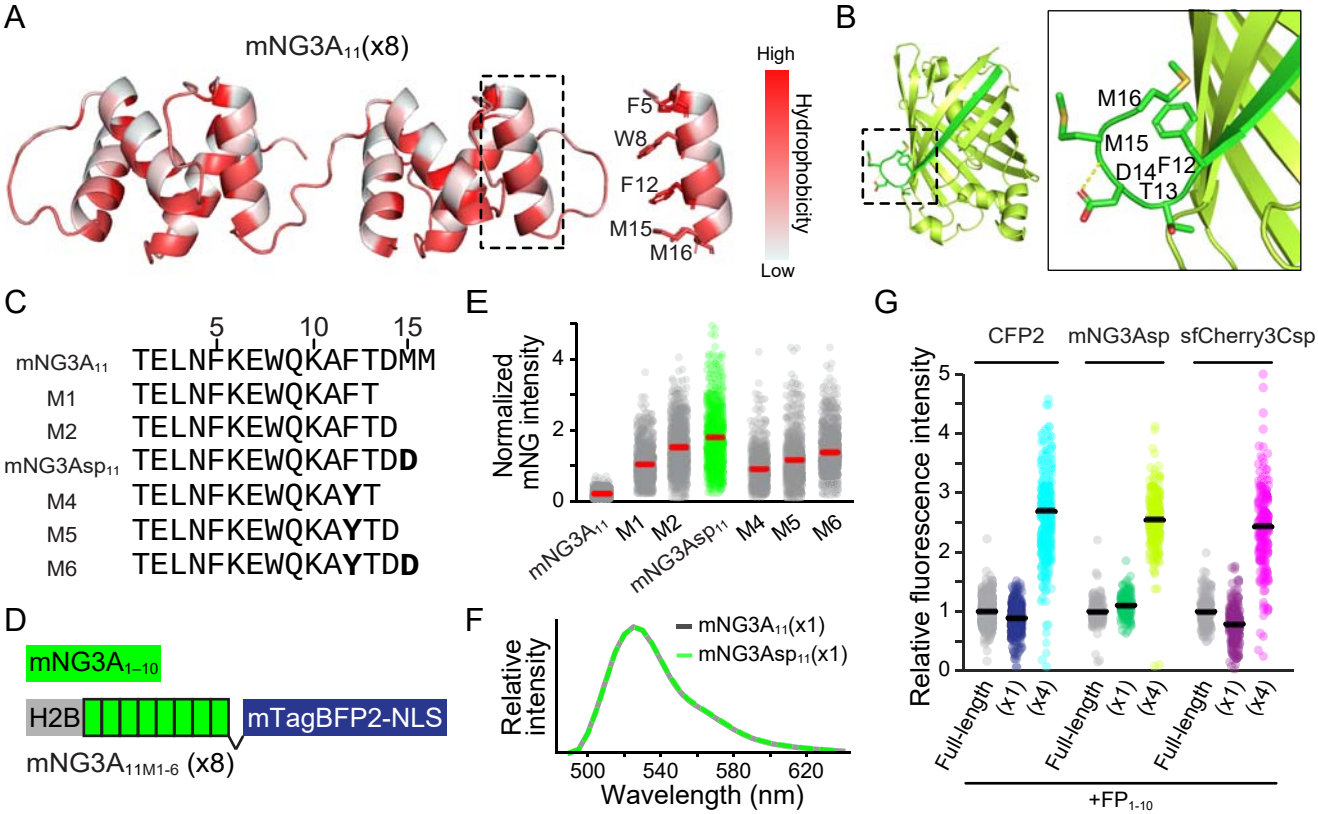
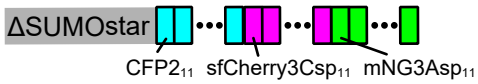

A
mNG3A11(x8)
F5
W8
F12
M15
M16
High
Hydrophobicity
Low
B
M16
M15
D14
F12
T13
C
5 10 15
mNG3A11 TELNFKEWQKAFTDMM
M1 TELNFKEWQKAFT
M2 TELNFKEWQKAFTD
mNG3Asp11 TELNFKEWQKAFTDD
M4 TELNFKEWQKAYT
M5 TELNFKEWQKAYTD
M6 TELNFKEWQKAYTDD
D
mNG3A1–10
H2B
mTagBFP2-NLS
mNG3A11M1-6 (x8)
E
Normalized mNG intensity
mNG3A11
M1
M2
mNG3Asp11
M4
M5
M6
F
Relative intensity
mNG3A11(x1)
mNG3Asp11(x1)
Wavelength (nm)
G
CFP2
mNG3Asp
sfCherry3Csp
Relative fluorescence intensity
Full-length
(x1)
(x4)
+FP1-10

Fig5

A



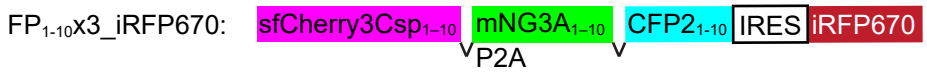
Each FP_{11} repeat number: (x0), (x1), (x2), (x4), (x8)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
C	1	2	4	8	0	0	0	0	0	0	0	0	1	1	1	1	2	2	2	2	4	4	4	4	8	8	8	1	1	1	1	2
G	0	0	0	0	1	2	4	8	0	0	0	0	1	2	4	8	1	2	4	8	1	2	4	8	1	2	4	0	0	0	0	0
R	0	0	0	0	0	0	0	0	1	2	4	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	4	8	1
1st			O			O		O		O		O				O		O	O						O						O	
2nd			O			O						O				O		O	O						O							
Final				O				O			O					O		O	O						O							

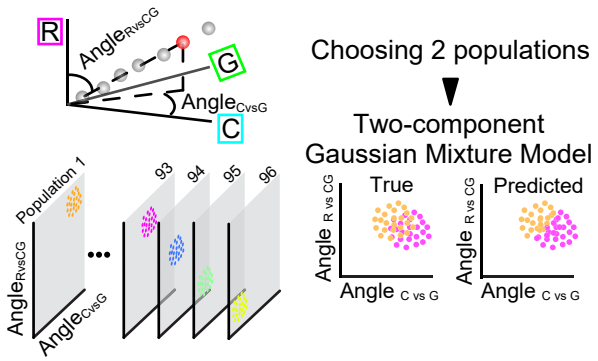
ID	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
C	2	2	2	4	4	4	4	8	8	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
G	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	2	2	2	4	4	4	4	8	8	8	1	1	1	1	2	2	2
R	2	4	8	1	2	4	8	1	2	4	1	2	4	8	1	2	4	8	1	2	4	8	1	2	4	1	2	4	8	1	2	4
1st		O		O														O				O	O	O	O		O					O
2nd		O		O														O				O		O								O
Final		O		O														O				O		O								O

ID	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
C	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	4	4	4	4	4	4	4	4	4	8	8	8	8
G	2	4	4	4	8	8	1	1	1	1	2	2	2	2	4	4	4	8	8	1	1	1	2	2	2	4	4	4	1	1	2	2
R	8	1	2	4	1	2	1	2	4	8	1	2	4	8	1	2	4	1	2	1	2	4	1	2	4	1	2	4	1	2	1	2
1st				O	O	O	O	O						O						O		O		O				O			O	
2nd				O		O		O						O						O								O			O	
Final				O		O		O						O						O								O			O	

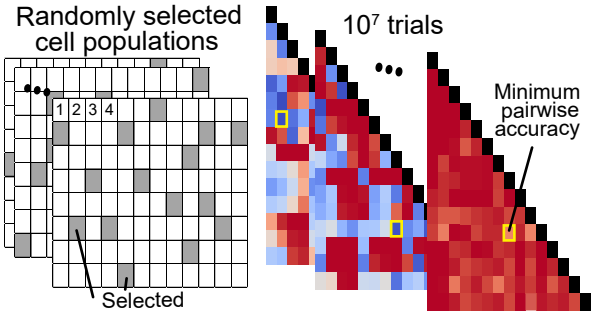
B



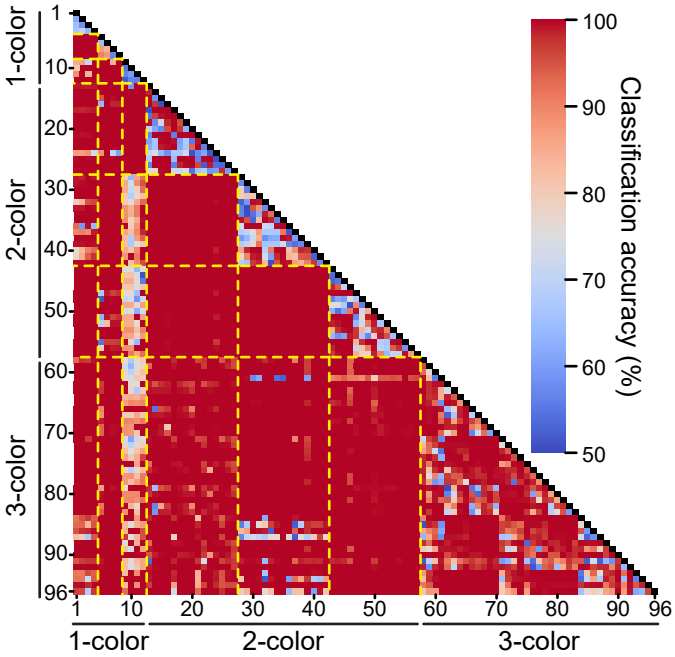
C



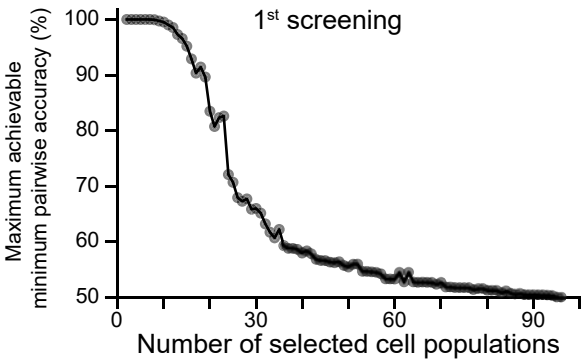
E



D



F



G

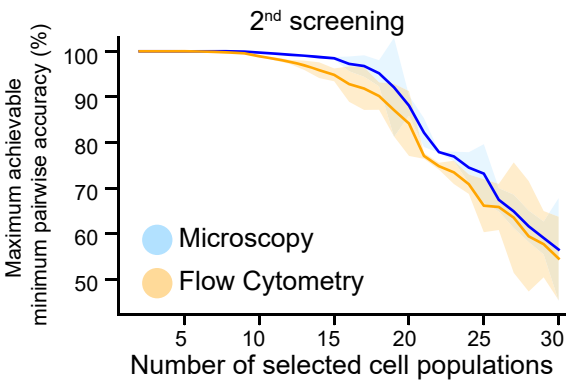


Fig.6

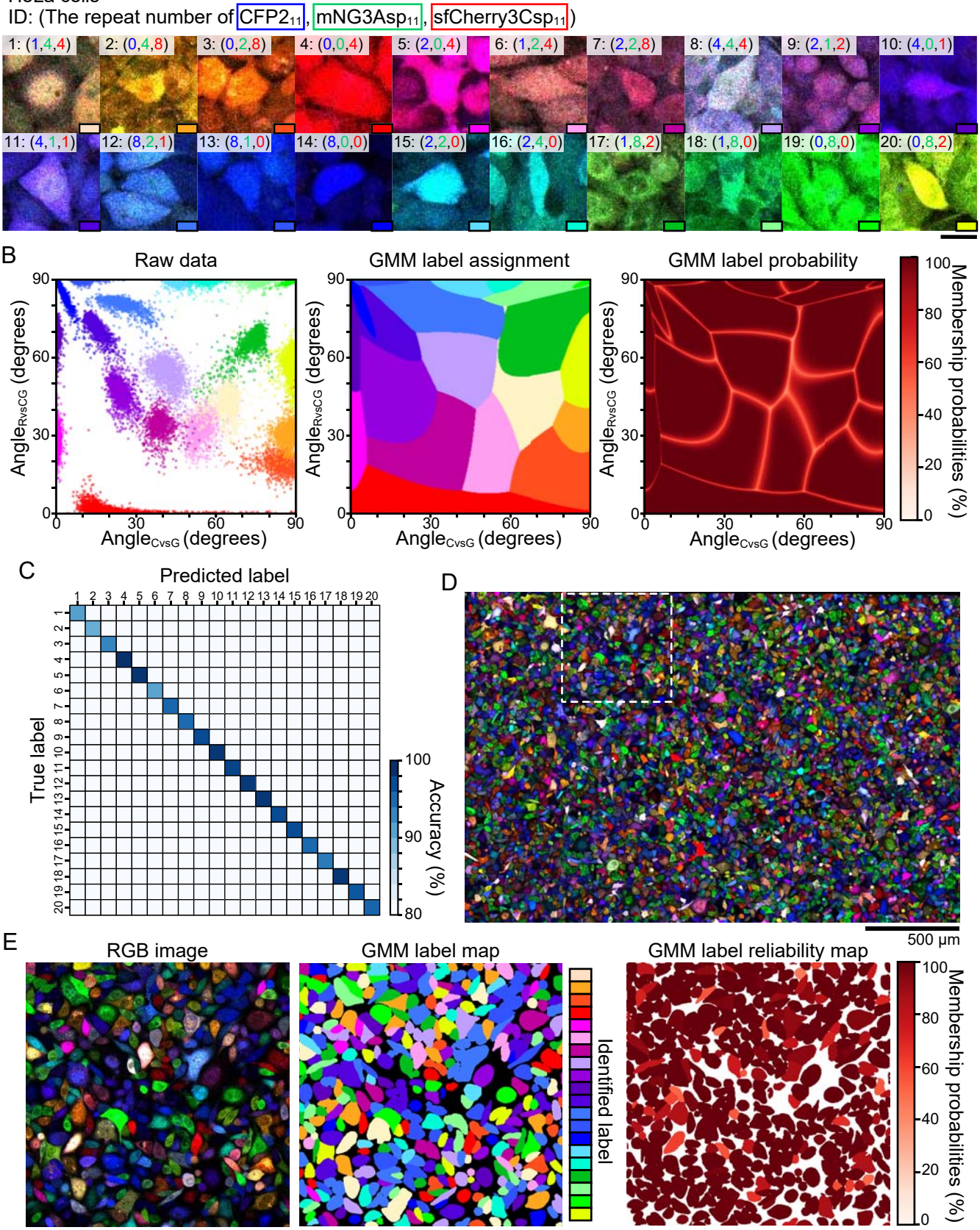
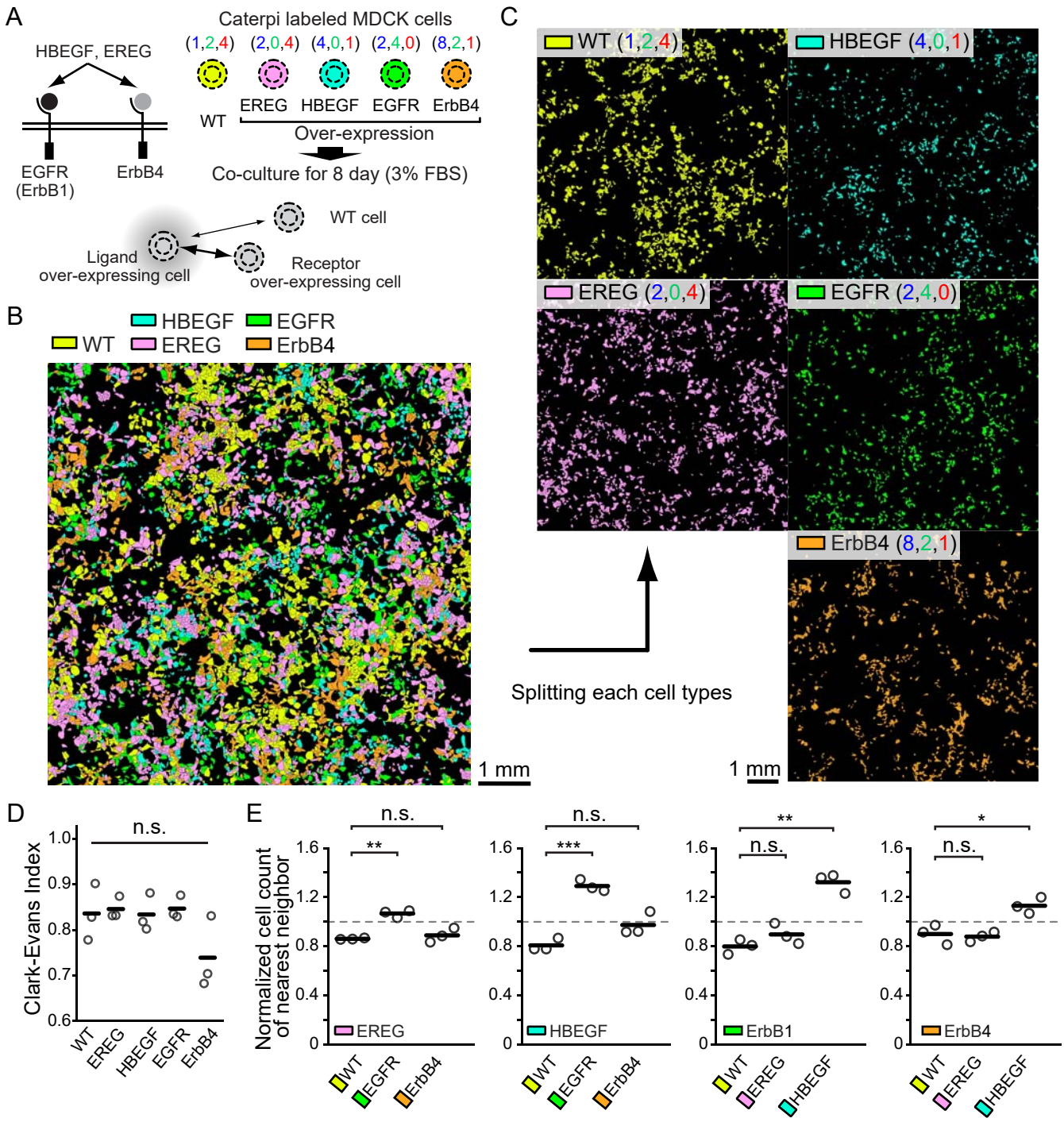
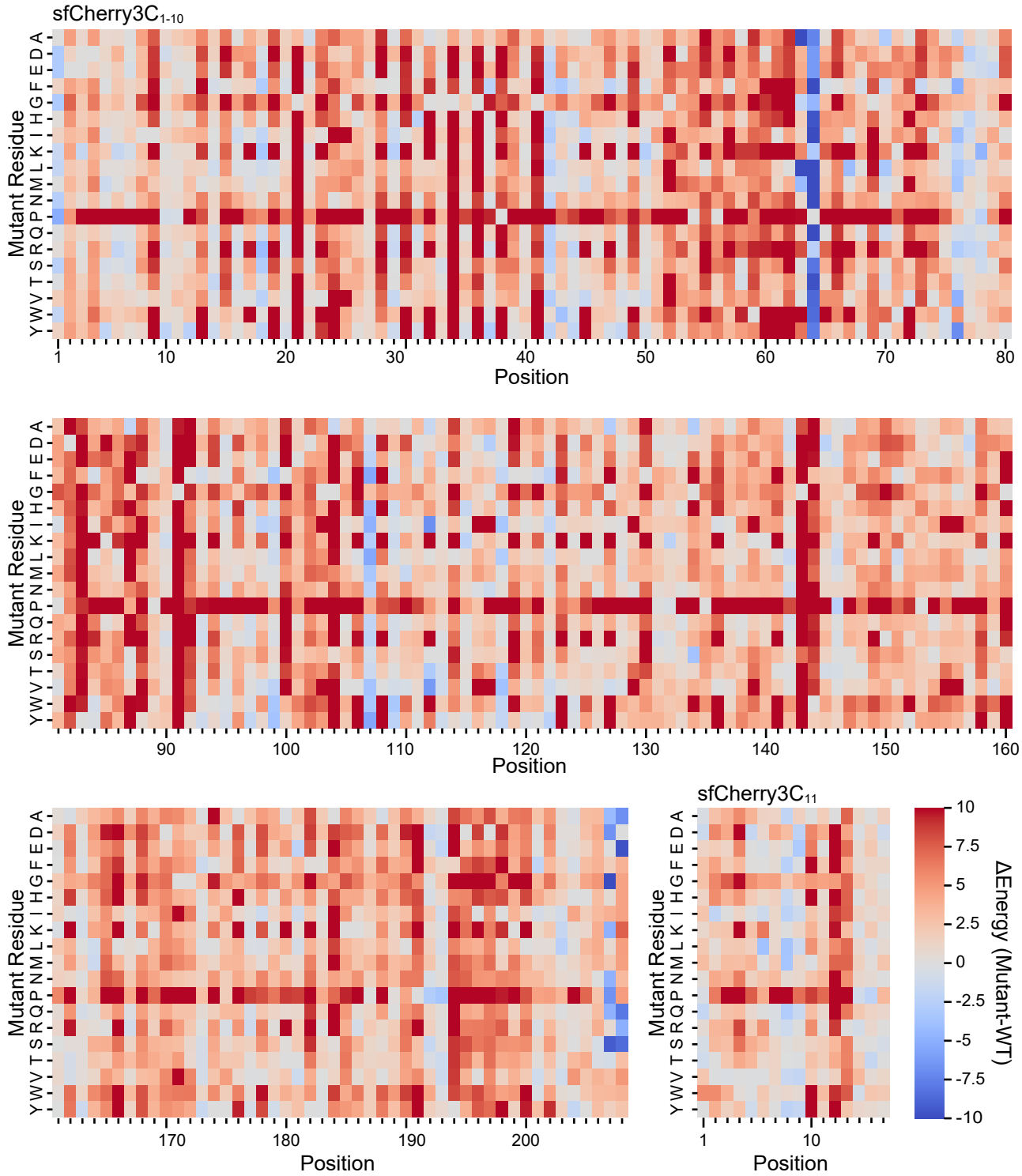

A
HeLa cells
ID: (The repeat number of CFP2$_{11}$, mNG3Asp$_{11}$, sfCherry3Csp$_{11}$)
1: (1,4,4)
2: (0,4,8)
3: (0,2,8)
4: (0,0,4)
5: (2,0,4)
6: (1,2,4)
7: (2,2,8)
8: (4,4,4)
9: (2,1,2)
10: (4,0,1)
11: (4,1,1)
12: (8,2,1)
13: (8,1,0)
14: (8,0,0)
15: (2,2,0)
16: (2,4,0)
17: (1,8,2)
18: (1,8,0)
19: (0,8,0)
20: (0,8,2)
B
Raw data
GMM label assignment
GMM label probability
Angle$_{RvsCG}$ (degrees)
Angle$_{CvsG}$ (degrees)
Membership probabilities (%)
C
Predicted label
True label
Accuracy (%)
D
500 μm
E
RGB image
GMM label map
GMM label reliability map
Identified label
Membership probabilities (%)

Fig.7

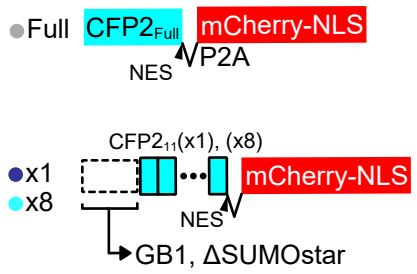

A
HBEGF, EREG
EGFR (ErbB1)
ErbB4
Caterpi labeled MDCK cells
(1,2,4) (2,0,4) (4,0,1) (2,4,0) (8,2,1)
WT
EREG HBEGF EGFR ErbB4
Over-expression
Co-culture for 8 day (3% FBS)
WT cell
Ligand over-expressing cell
Receptor over-expressing cell
B
WT
HBEGF
EREG
EGFR
ErbB4
1 mm
C
WT (1,2,4)
HBEGF (4,0,1)
EREG (2,0,4)
EGFR (2,4,0)
ErbB4 (8,2,1)
Splitting each cell types
1 mm
D
Clark-Evans Index
n.s.
WT EREG HBEGF EGFR ErbB4
E
Normalized cell count of nearest neighbor
n.s.
**
EREG
***
HBEGF
ErbB1
*
ErbB4
WT EGFR ErbB4
WT EREG HBEGF

Sup1


sfCherry3C1-10
Mutant Residue
Y W V T S R Q P N M L K I H G F E D A
1
10
20
30
40
50
60
70
80
Position
90
100
110
120
130
140
150
160
170
180
190
200
sfCherry3C11
ΔEnergy (Mutant-WT)
10
7.5
5
2.5
0
-2.5
-5
-7.5
-10

Sup2

A

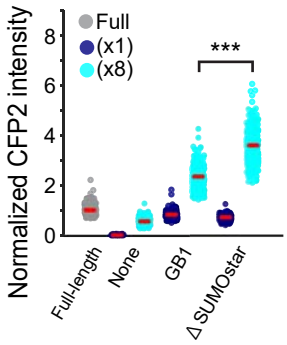


B

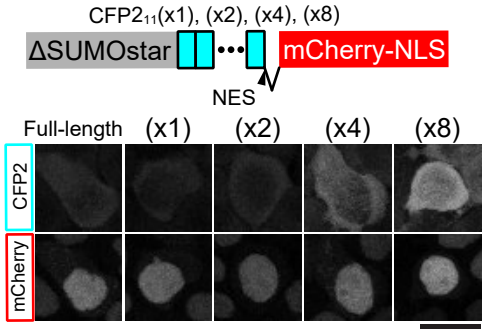
ΔSUMOstar

(MGHHHHHHGSLQDSEVNQEAKPEVK
PEVKPE)**THINLKVSDGSSEIFFKIK**
KTTPLRRLMEAFAKRQGKEMDSLTF
LYDGIEIQADQTPEDLDMEDNDIIE
AHREQ(IGGRDDLPSL)

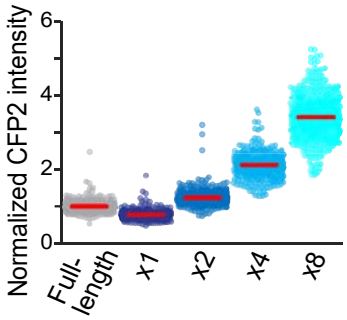
C



D



E

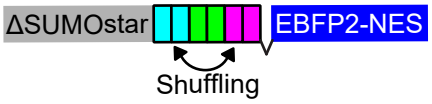


F

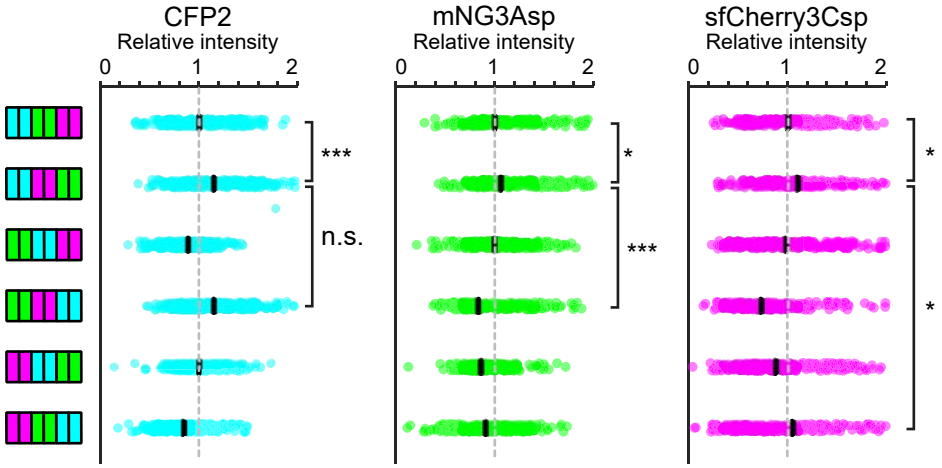
True labels \ Predicted labels	x1	x2	x4	x8
x1	87.8	2.3	0.0	0.0
x2	11.0	95.4	6.2	0.0
x4	1.2	1.1	91.2	2.5
x8	0.0	1.1	2.5	97.5

Sup. 3

A


ΔSUMOstar
EBFP2-NES
Shuffling

B


CFP2
Relative intensity
0
1
2
***
n.s.
mNG3Asp
Relative intensity
0
1
2
*
***
sfCherry3Csp
Relative intensity
0
1
2
*
*

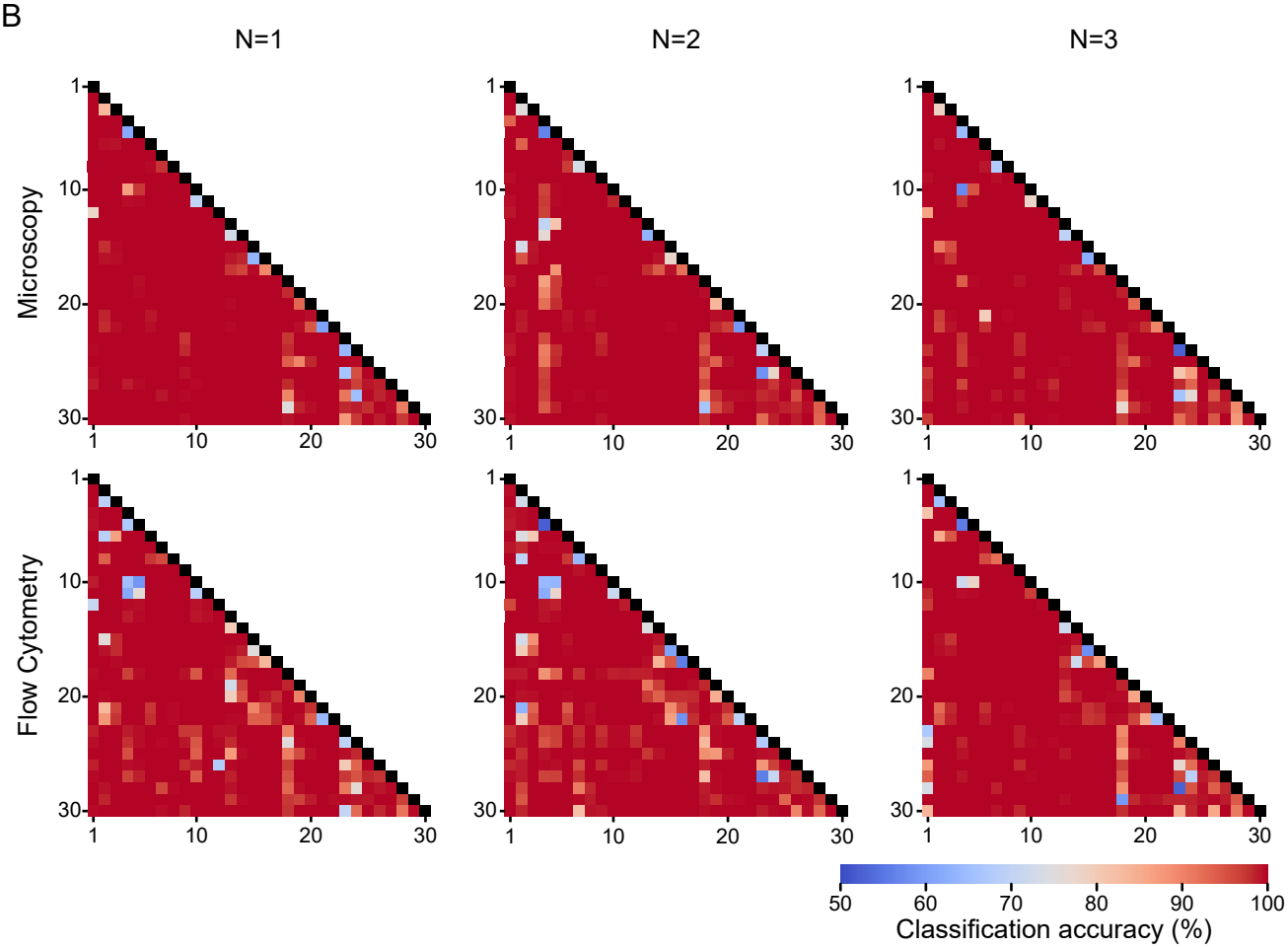
Sup. 4

A

ID: (The repeat number of CFP2$_{11}$, mNG3Asp$_{11}$, sfCherry3Csp$_{11}$)

1: (4,0,0)	6: (1,8,0)	11: (2,0,4)	16: (0,8,2)	21: (1,8,1)	26: (4,1,1)
2: (0,2,0)	7: (2,2,0)	12: (4,0,1)	17: (0,8,4)	22: (1,8,2)	27: (4,1,4)
3: (0,8,0)	8: (2,4,0)	13: (0,2,8)	18: (1,1,2)	23: (2,1,1)	28: (4,2,2)
4: (0,0,2)	9: (8,1,0)	14: (0,4,8)	19: (1,2,4)	24: (2,1,2)	29: (4,4,4)
5: (0,0,8)	10: (1,0,8)	15: (0,8,1)	20: (1,4,4)	25: (2,2,8)	30: (8,2,1)

B



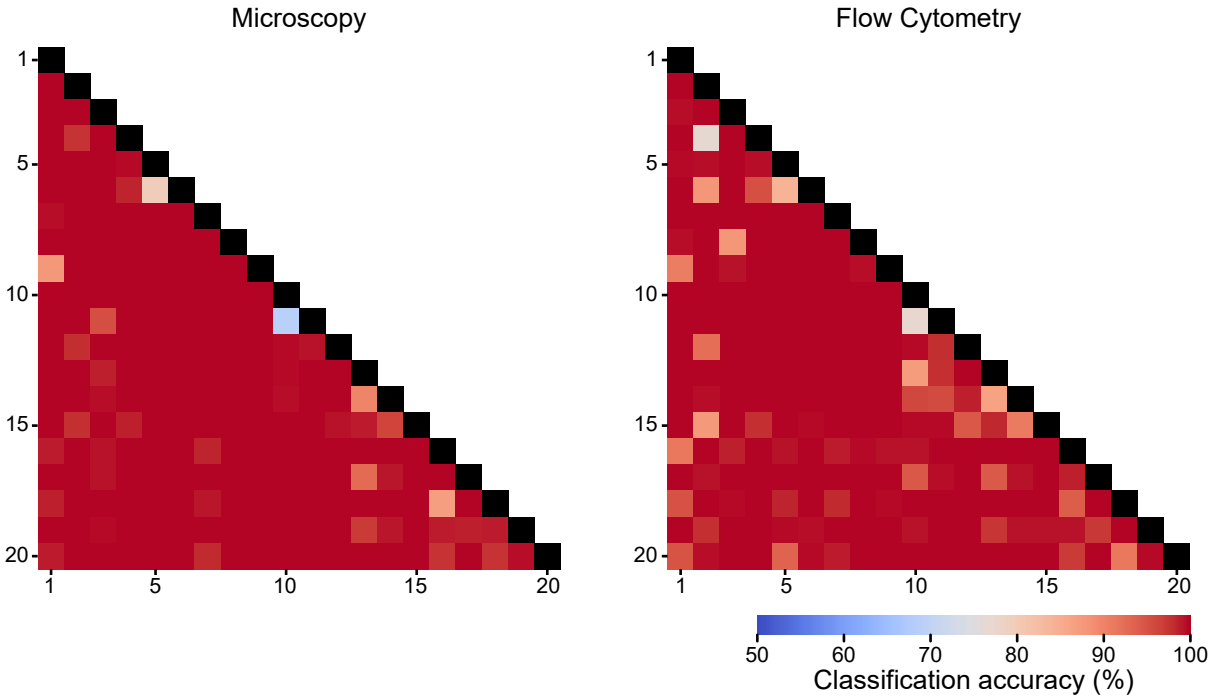
Sup. 5

A

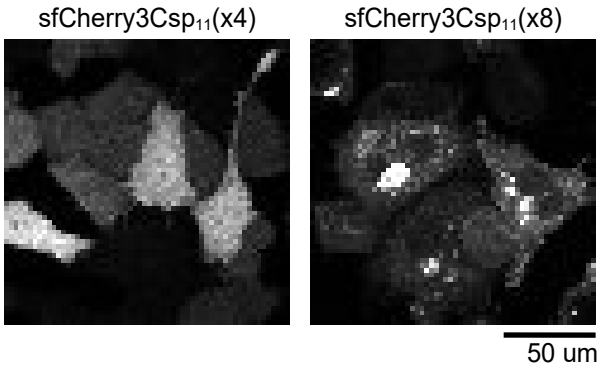
ID: (The repeat number of CFP2$_{11}$, mNG3Asp$_{11}$, sfCherry3Csp$_{11}$)

1: (4,0,0)	5: (2,2,0)	9: (4,0,1)	13: (1,2,4)	17: (2,2,8)
2: (0,2,0)	6: (2,4,0)	10: (0,2,8)	14: (1,4,4)	18: (4,1,1)
3: (0,0,8)	7: (8,1,0)	11: (0,4,8)	15: (1,8,2)	19: (4,4,4)
4: (1,8,0)	8: (2,0,4)	12: (0,8,2)	16: (2,1,2)	20: (8,2,1)

B

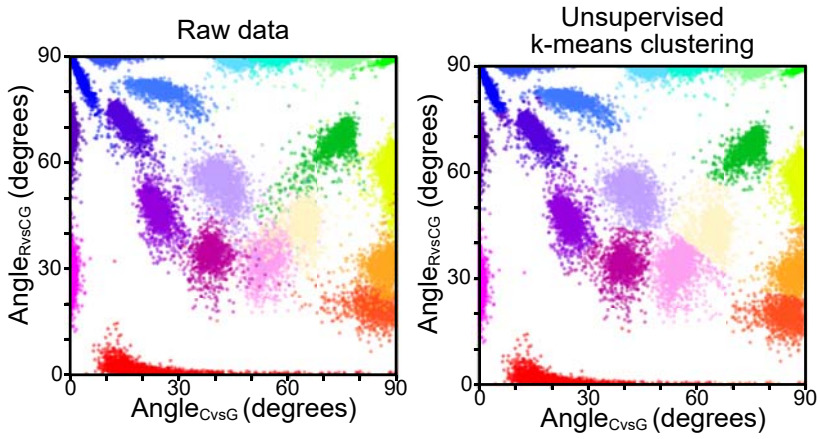


Sup.6

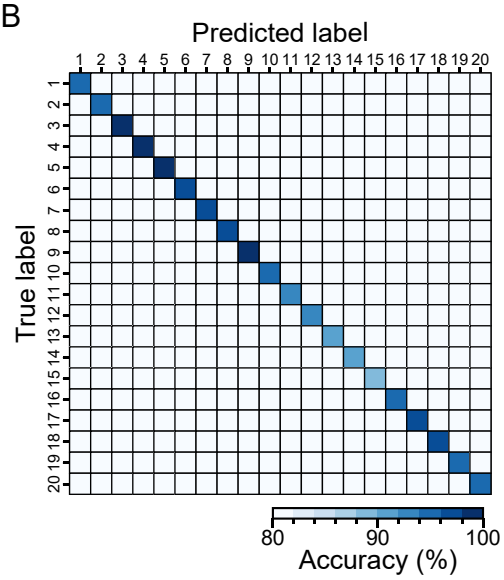


Sup.7

A

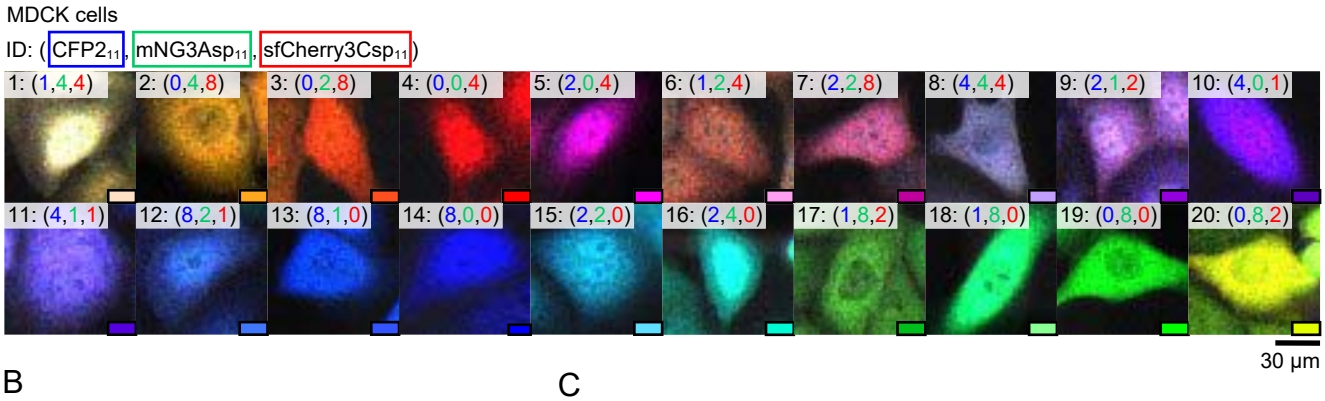

Raw data
Unsupervised
k-means clustering
Angle$_{RvsCG}$ (degrees)
Angle$_{CvsG}$ (degrees)
0
30
60
90

B

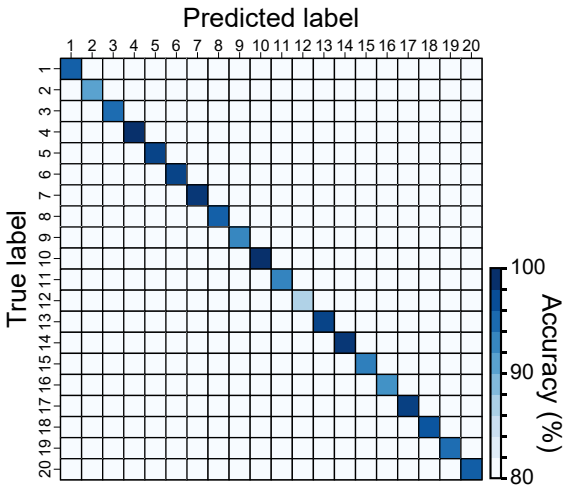

Predicted label
True label
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
80
90
100
Accuracy (%)

Sup.8

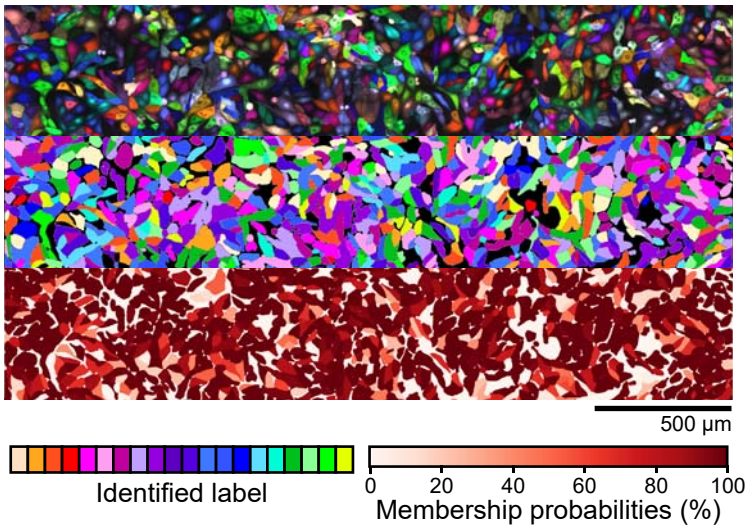
A


MDCK cells
ID: (CFP2$_{11}$, mNG3Asp$_{11}$, sfCherry3Csp$_{11}$)
1: (1,4,4)
2: (0,4,8)
3: (0,2,8)
4: (0,0,4)
5: (2,0,4)
6: (1,2,4)
7: (2,2,8)
8: (4,4,4)
9: (2,1,2)
10: (4,0,1)
11: (4,1,1)
12: (8,2,1)
13: (8,1,0)
14: (8,0,0)
15: (2,2,0)
16: (2,4,0)
17: (1,8,2)
18: (1,8,0)
19: (0,8,0)
20: (0,8,2)
30 μm

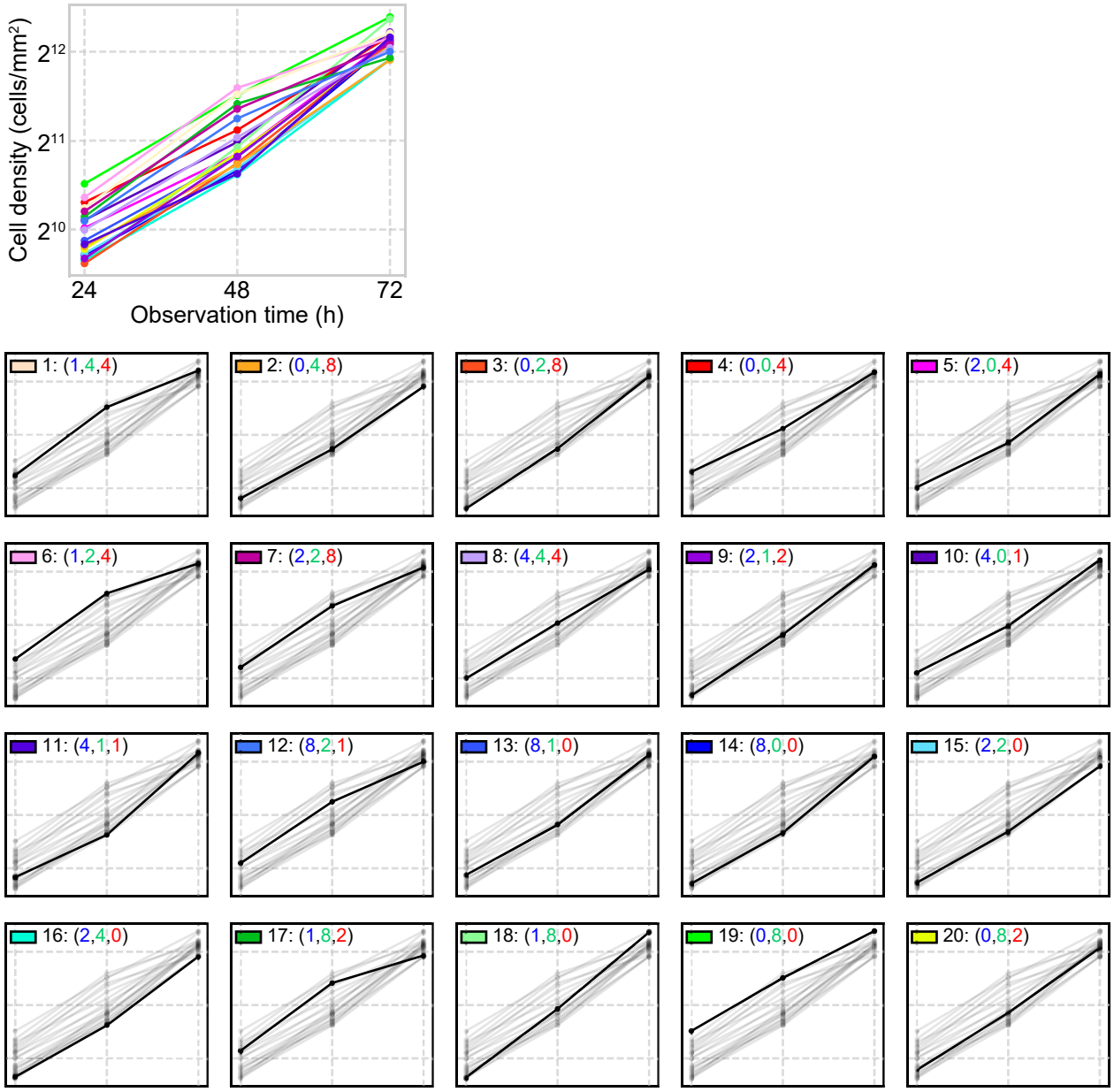
B


Predicted label
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
True label
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
100
90
80
Accuracy (%)

C


500 μm
Identified label
0 20 40 60 80 100
Membership probabilities (%)

Sup.9


Cell density (cells/mm²)
2¹²
2¹¹
2¹⁰
24
48
72
Observation time (h)
1: (1,4,4)
2: (0,4,8)
3: (0,2,8)
4: (0,0,4)
5: (2,0,4)
6: (1,2,4)
7: (2,2,8)
8: (4,4,4)
9: (2,1,2)
10: (4,0,1)
11: (4,1,1)
12: (8,2,1)
13: (8,1,0)
14: (8,0,0)
15: (2,2,0)
16: (2,4,0)
17: (1,8,2)
18: (1,8,0)
19: (0,8,0)
20: (0,8,2)