Supplementary Information

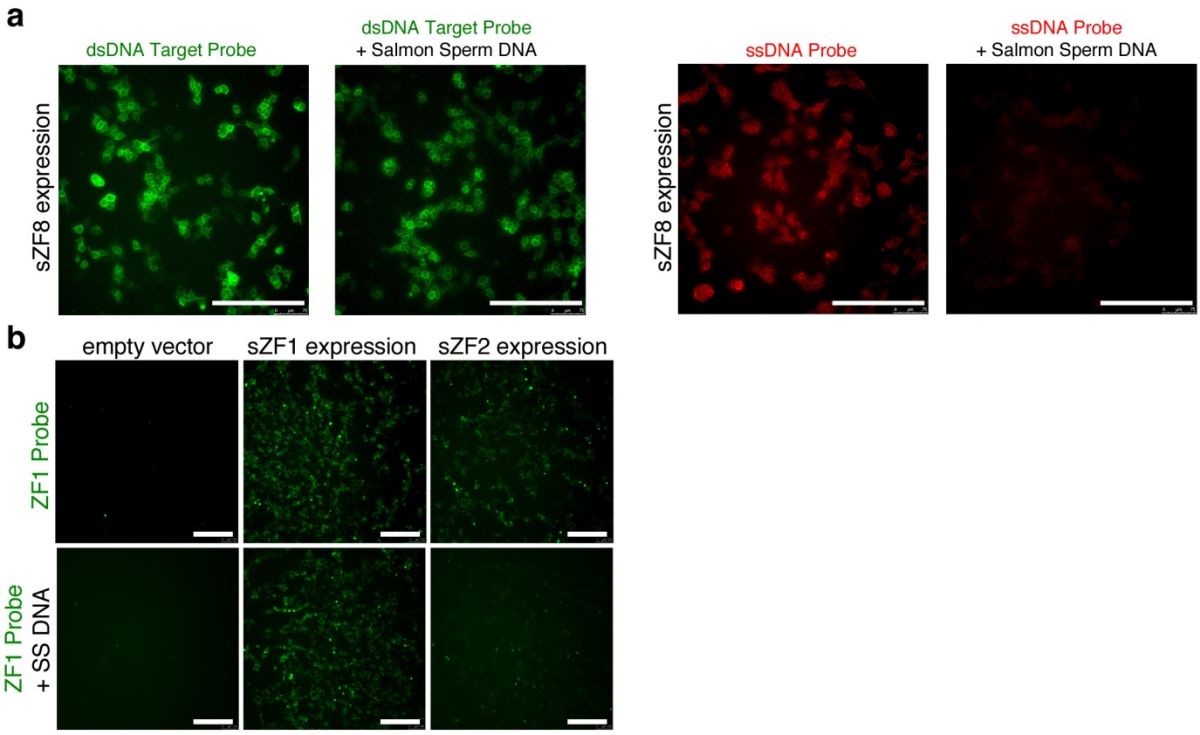
Barcoding cells using cell-surface programmable DNA-binding domains

Prashant Mali[1,3], John Aach[1,3], Jehyuk Lee[1,2], Daniel Levner[1,2], Lisa Nip[2], George M. Church[1,2,4]

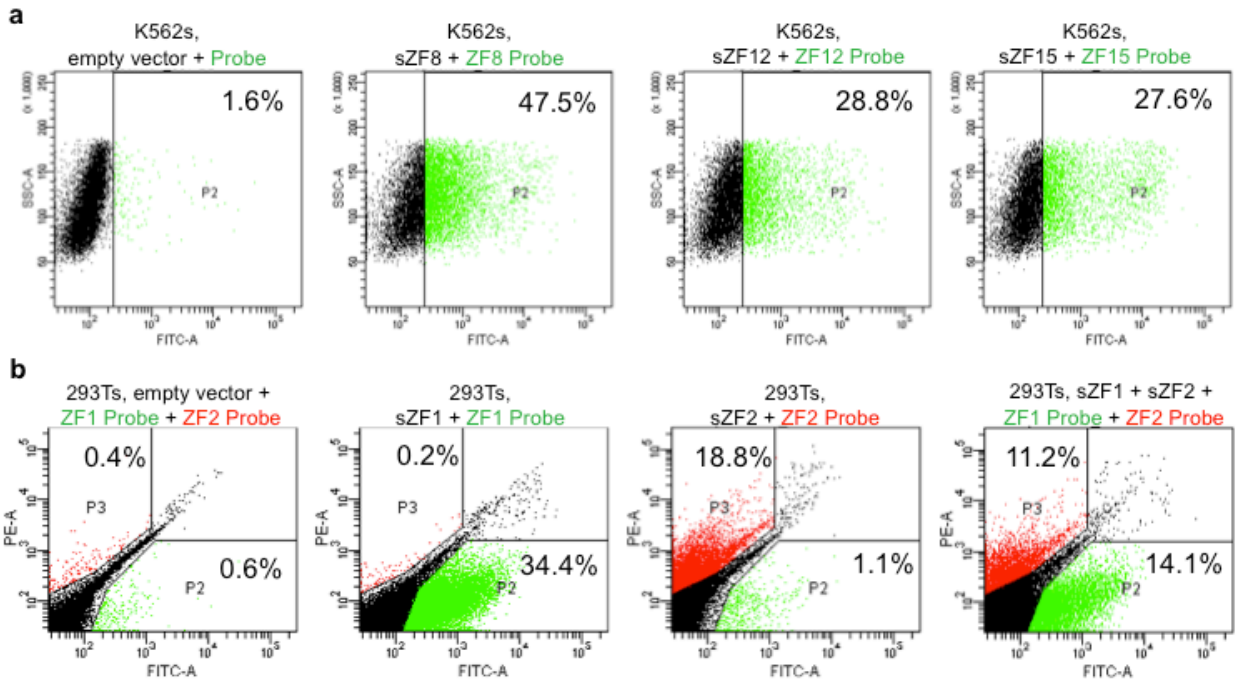
[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts, USA.

[3]These authors contributed equally to this work, [4]Correspondence should be addressed to gchurch@genetics.med.harvard.edu.

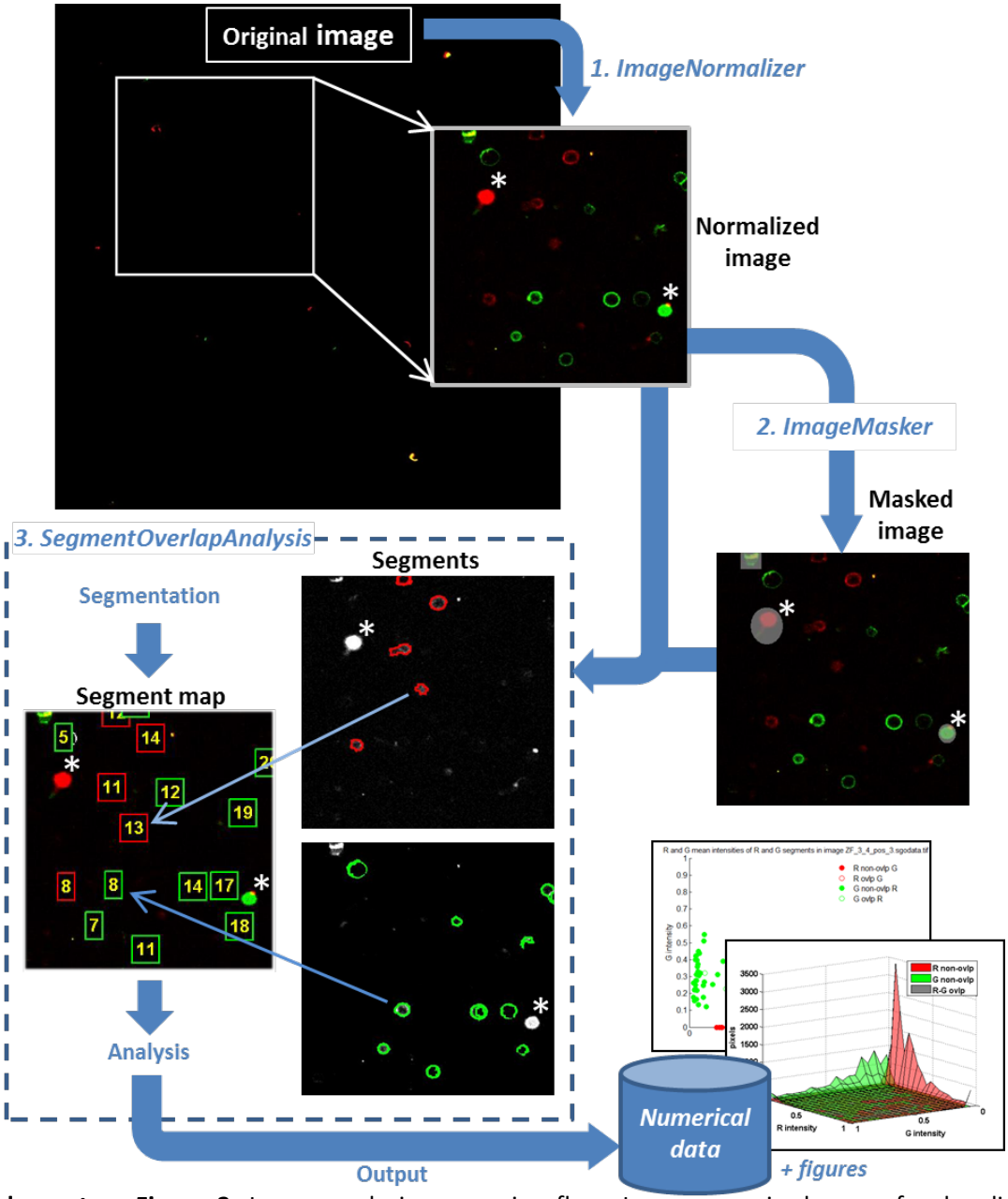
Table of Contents



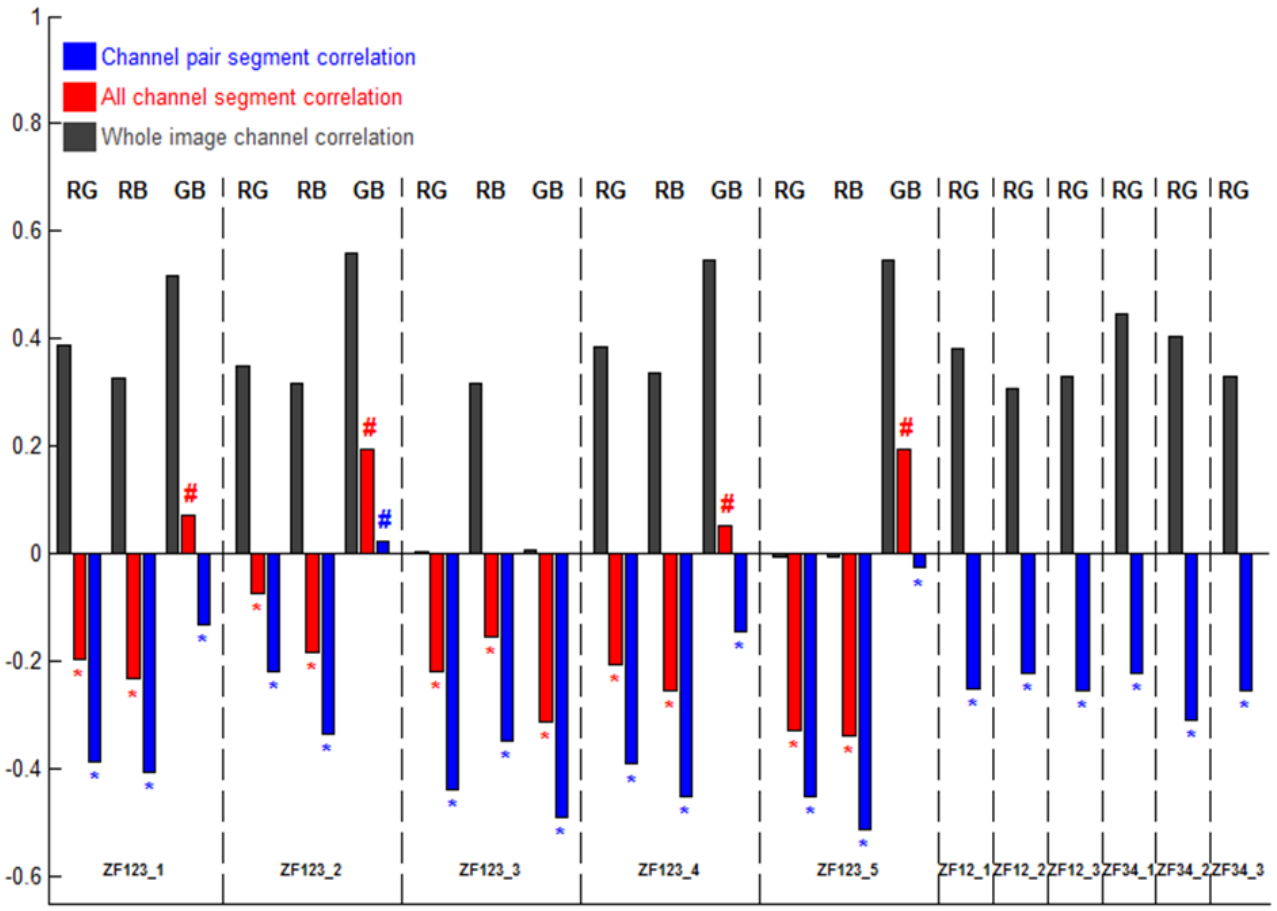
Supplementary Figure 1. Target dsDNA specific labeling of sZF expressing cells. **(a)** ssDNA vs. dsDNA: sZF expressing cells bind both single-stranded and double-stranded DNA, however in the presence of a dsDNA competitor (SS DNA, Salmon Sperm DNA), binding to only the dsDNA is retained. **(b)** Target vs. non-specific dsDNA: Cells expressing different sZFs (here sZF1 and sZF2) can bind to dsDNA molecules non-specifically (both bind the ZF1 probe), however in the presence of SS DNA competition only binding to the target dsDNA (ZF1 probe by sZF1 expressing cells) is retained, and non-specific interactions (ZF1 probe by sZF2 expressing cells) are competed out. Thus sZF expressing cells specifically bind their target dsDNA probes in the presence of appropriate competitor dsDNA molecules. The scale bar is 100microns.



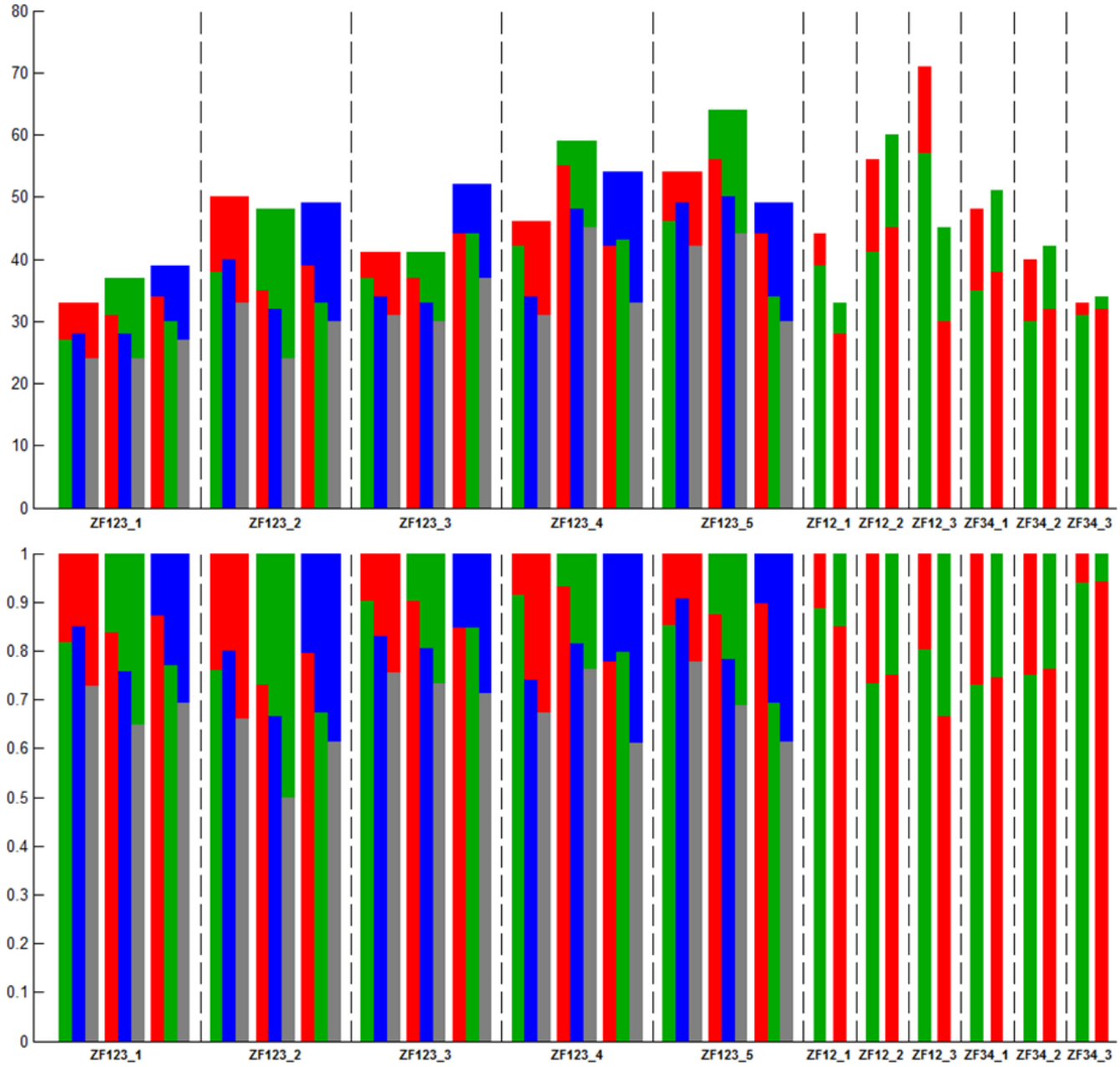
Supplementary Figure 2. FACS analysis of sZF expressing cells. **(a)** K562 cells nucleofected with either an empty vector or sZF8, sZF12 and sZF15 expressing vectors were confirmed via FACS analysis for their ability to bind their target dsDNA probes. **(b)** Similarly, 293T cells expressing either sZF1 or sZF2 were tested for their ability to bind their target dsDNA probes in both simplex and multiplex formats.



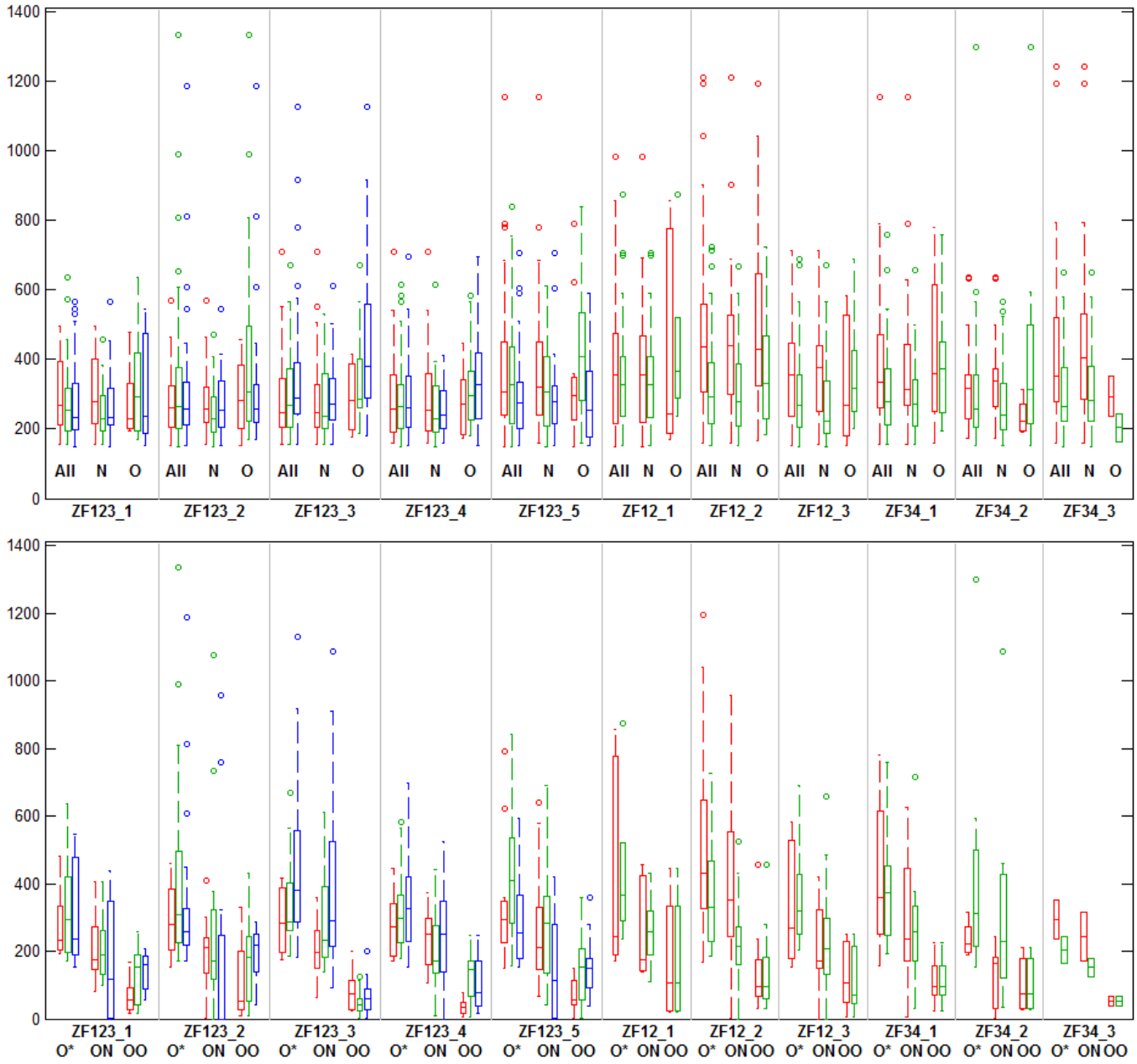
Supplementary Figure 3. Image analysis processing flow. Images acquired as confocal z-slices for each fluorescent label are consolidated into a single multi-channel image and each channel separately normalized using *ImageNormalizer*. A 1024x1024 unnormalized input image is shown (top) along with a 400x400 region of normalized output that is tracked through the rest of the processing flow. Next, dead cells and cell debris are masked out using *ImageMasker* to exclude them from subsequent analysis. Viable cells appear as rings in the confocal z-slices while dead cells and debris appear as dense spots or diffuse smears. Two apparently dead cells marked with white asterisks in the normalized image section (one red and one green) have been masked in the masked image. Finally, *SegmentOverlapAnalysis* is used to segment and analyze the image. Segmentation is performed separately for each channel. Each segment is assigned a channel and segment number (see Segment map) and is shown above by a colored segment boundary. Masked cells do not appear as segments. Segmented images are then analyzed (see text for details), and results are output as figures and in a computer-readable data file in which individual segments are listed by their channel and segment number.



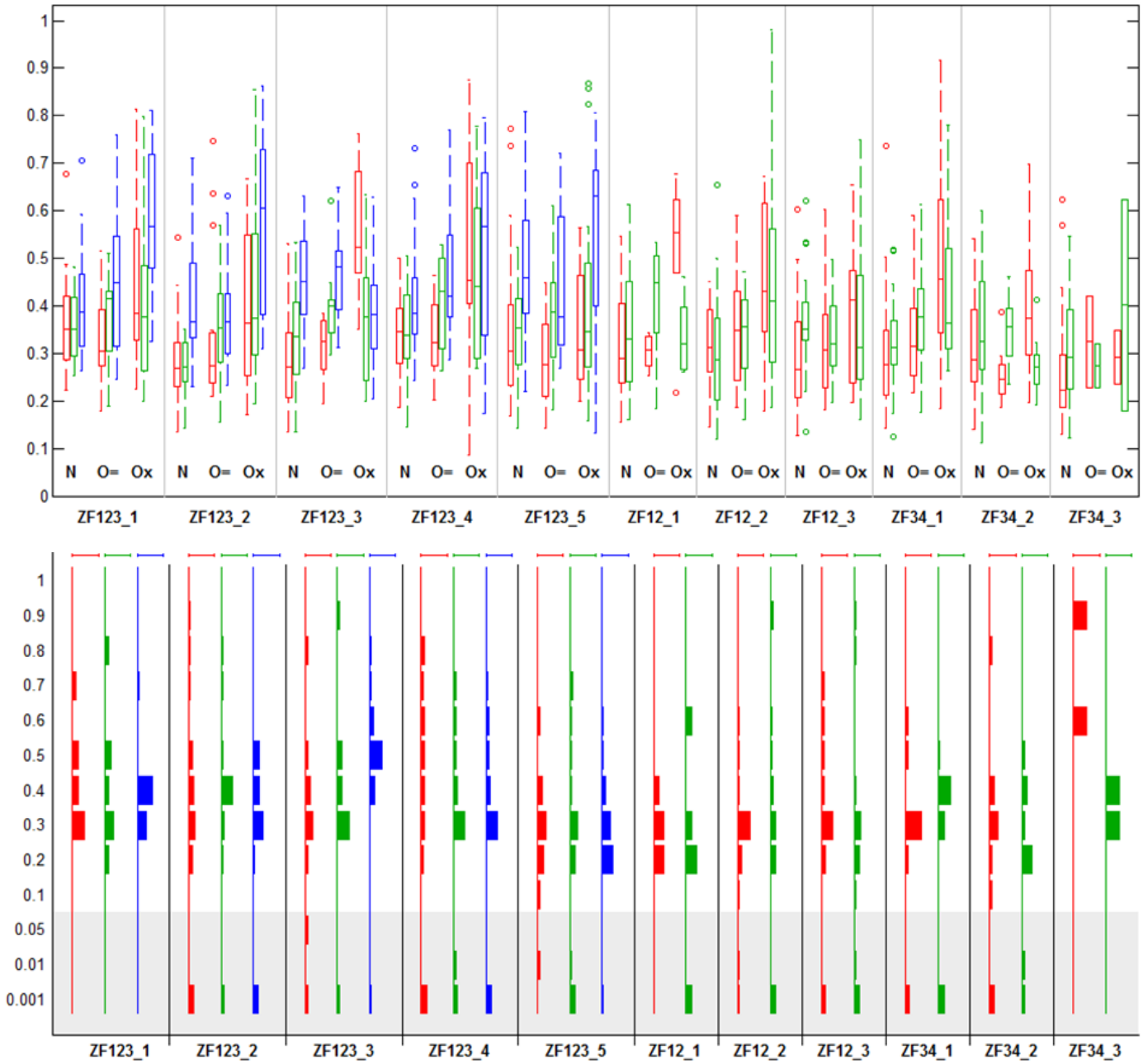
Supplementary Figure 4. Pearson correlation coefficients between channel pairs in images of sZF cells. For each pair of channels, *whole image correlations* (dark gray) were computed over all pixels in the image and c*hannel pair segment correlations* (blue) were computed within the union of segment regions of the channel pair. For three channel images, *all segment channel correlations* (red) were also computed within the union of segment regions of *all three* channels. Channel pairs are indicated as RG, RB, and GB. *Whole image correlations* are all large positive values, likely because they include a very high number of pixels of low intensity that are outside of any segment regions that may represent correlated background fluorescence. *Channel pair segment correlations* are all negative (excepting the small positive ZF123_2 GB value), indicating that within segments, background cross-talk between channels is overcome by stronger ZFP-related signals. The result is consistent with the hypothesis that the ZFPs corresponding to the channels bind their labeled oligos specifically. *All channel segment correlations* are intermediate between whole image and channel pair segment correlations, probably because they blend both specific labeling in the channel pair segments and correlated background in the third channel segment regions. Correlation P values were computed for all *channel pair* and *all channel* correlations from 1000 random shuffles of pixel intensities in their respective segment regions. All P values were < .001 (# = actual correlation > all random correlations; * = actual correlation < all random correlations).



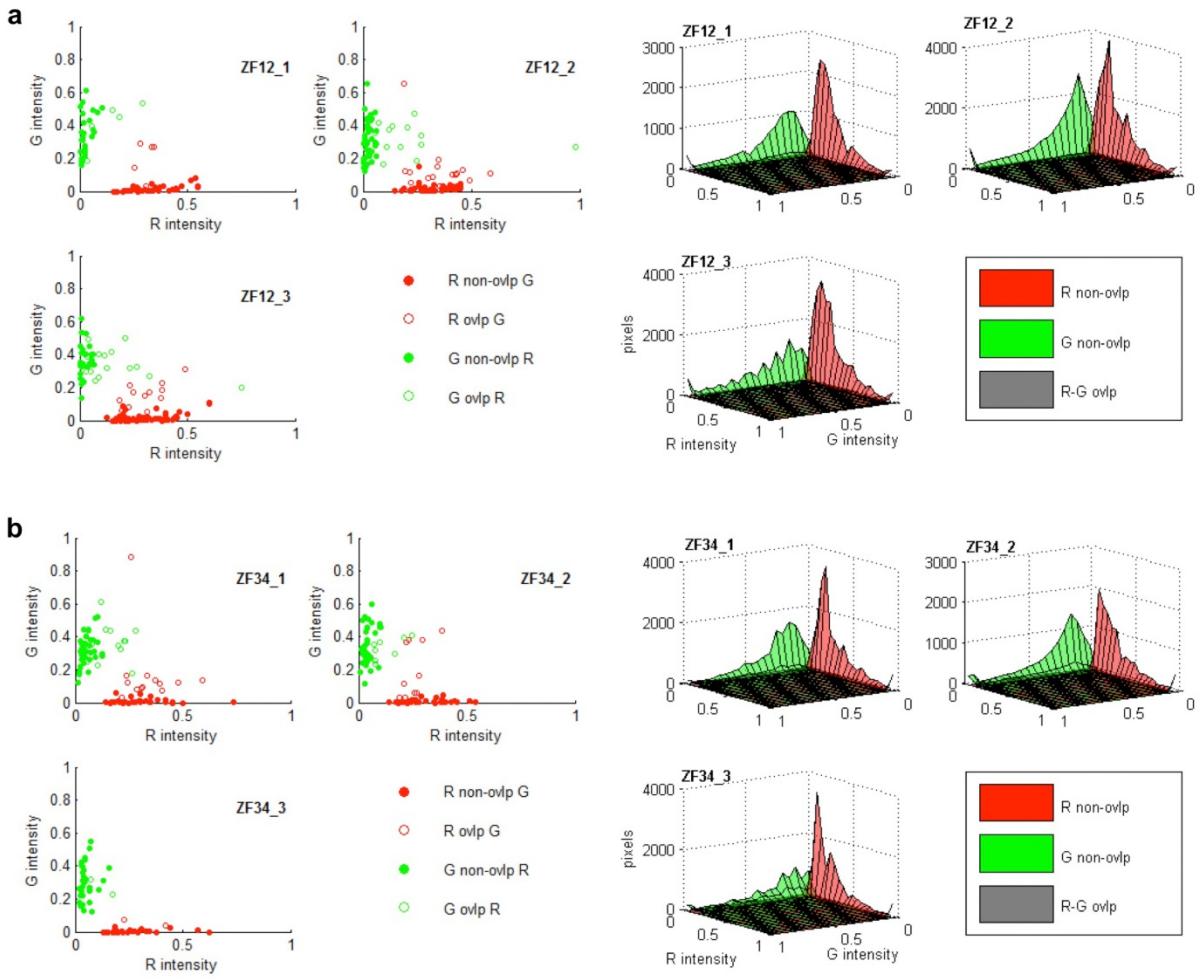
Supplementary Figure 5. Overall segment and NOVS counts and fractions from each channel in each image. Bars in this figure are either stacked on top of bars of other colors (for two-channel images, right), or are on top of multiple bars of other colors (three-channel images, left). By its color, a top bar in either case represents a channel in which segments have been generated. The bars under the top bar represent other channels which have also been segmented and some of whose segments overlap those corresponding to the top bar. The size of the lower bar indicates the count (*Top chart*) or fraction (*Bottom chart*) of segments in the top bar channel that are NOVS for the top bar with respect of lower bar segments. For instance, for ZF123_1 (leftmost image in each chart), the *top chart* indicates that there were 33 segments overall in the R channel (height of wide red bar). The height of the green bar within the red bar is 27, which indicates that 27 of those 33 were NOVS R segments that did not overlap any G segment; similarly, the height of the blue bar within the red bar is 28 and indicates that 28 of the 33 R segments were NOVS that did not overlap any B segment. The gray bars present for the three-channel images indicate the number of NOVS segments in the top channel that did not overlap segments of *either* of the lower bar channels; thus 24 (gray bar height) of the 33 R segments overlapped neither a B nor a G segment. The bottom chart gives corresponding fractions: ~82%, ~85% and ~73% of the R segments were NOVS with respect to these other channels, respectively.



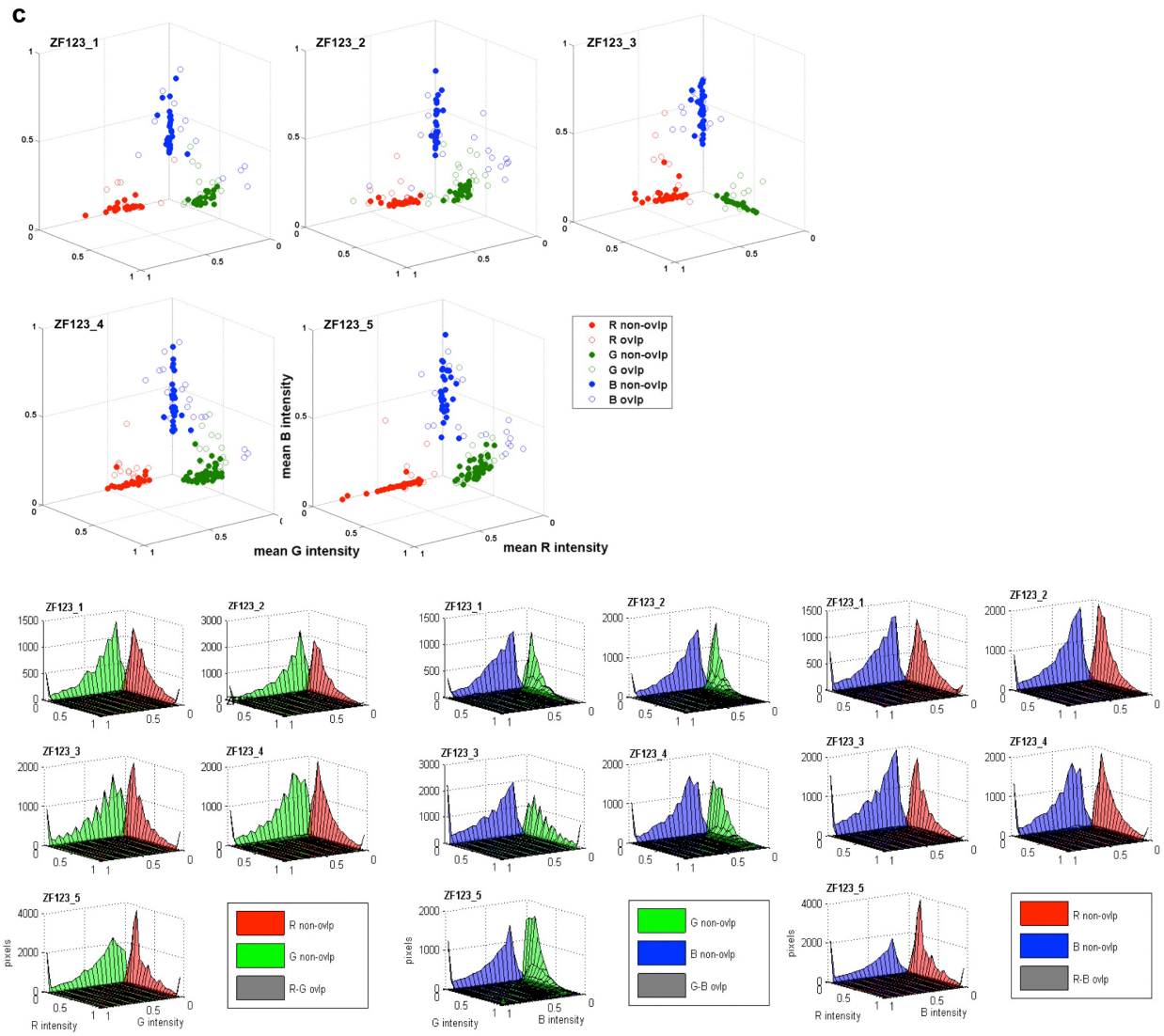
Supplementary Figure 6. Segment area size distributions for all images, presented as box plots. *Top:* Segment areas for all (All) NOVS (N), and OVS (O) segments in all channels of all images. In this figure, NOVS and OVS are relative to all other channels in the image, e.g., for ZF123_1, a three-channel image, R NOVS are segments that overlap neither a G nor a B segment, and R OVS are segments that overlap either a G or a B segment. *Bottom:* Box plots of complete OVS segment areas (O* = O segments from the top boxplot), and then of the OVS-nonovlp (ON) and OVS-ovlp (OO) portions of these OVS segments, in all channels in all images. The O*, ON, and OO distributions reflect the fact that the total area of any OVS segment is the sum of the areas of its OVS-nonovlp and OVS-ovlp parts. Here it can be seen that, in general, OVS-ovlp areas are usually small compared to their OVS-nonovlp counterparts.

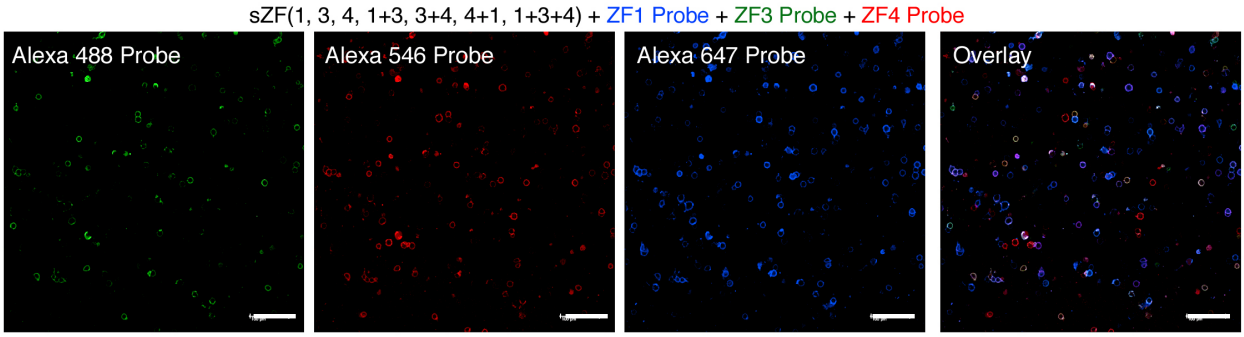


Supplementary Figure 7. Segment intensity distribution information for all images. In this figure, , where NOVS and OVS are relative to all other channels in the image—e.g., an NOVS R segment in a three-channel image is an R segment that overlaps neither a G nor a B segment. *Top:* Boxplots for distributions of mean intensities within segments of each channel: *N:* Mean intensities in segment channel for NOVS segments (e.g., mean R intensity for each NOVS R segment); *O=:* mean intensities for segment channel in all OVS segments (e.g., mean R intensity for each OVS R segment); *Ox:* mean maximum non-segment-channel intensity in all OVS segments (e.g., mean (max(B,G)) intensity for each OVS R segment). *Bottom:* Empirical probability distribution functions (EPDFs) of Wilcoxon rank sum p-values comparing distributions of pixel-level segment channel and maximum non-segment channel intensities for all OVS-ovlp segments. E.g., for each R OVS segment, Wilcoxon p-values were computed comparing the R intensities of all pixels in the segment and the max(B,G) intensities of those pixels, a histogram assembled of the Wilcoxon p-values of all the OVS R segments, and histogram counts were then sum-normalized to 1. Note that p-value bins of unequal size are used to highlight nominally significant p-values in the ranges $p \leq .001$, $.001 < p \leq .01$, and $.01 < p \leq .05$; these bins shown against a light gray background. At the top are scale bars showing P(EPDF p-value bin)=1.

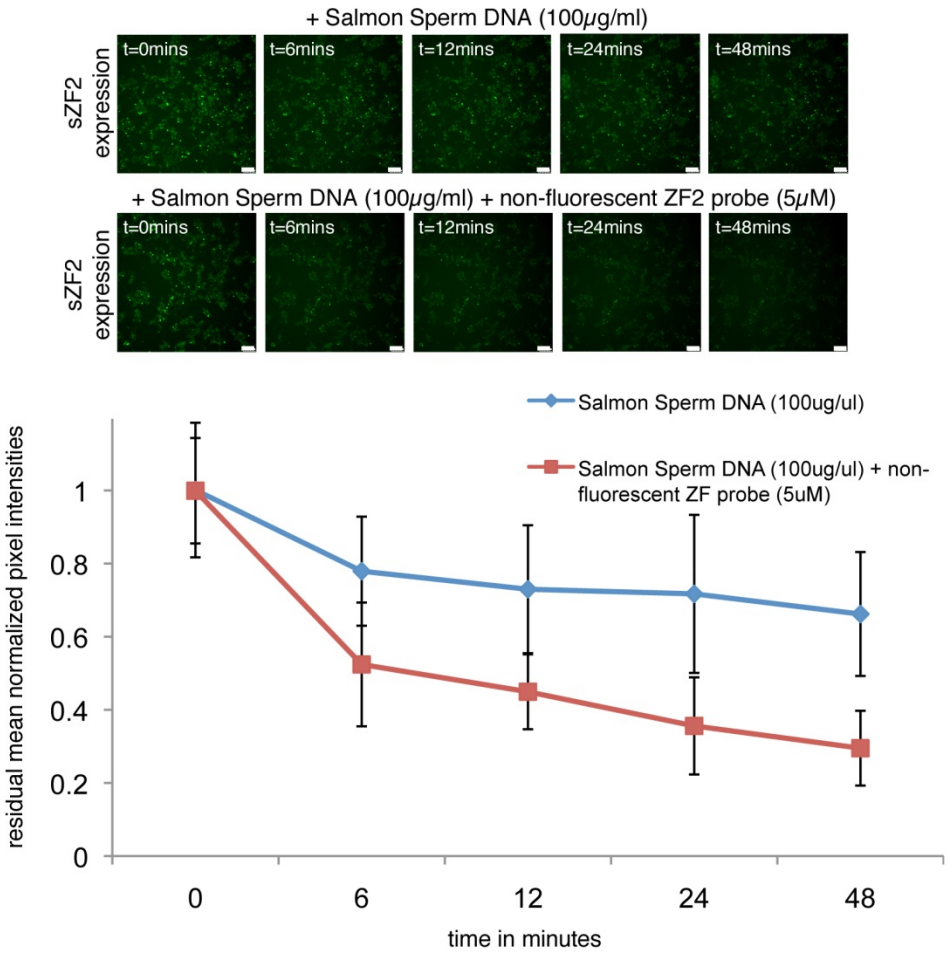


Supplementary Figure 8. Scatterplots and histograms of mean segment and pixel intensity. **(a)** Scatterplots of mean intensity of segments in the R and G channels for the images from the ZF12 set of cell samples, and 2D histograms of pixel intensities of the non-overlapping regions of the R and G segments derived from these images is shown. **(b)** Scatterplots of mean intensity of segments in the R and G channels for the images from the ZF34 set of cell samples, and 2D histograms of pixel intensities of the non-overlapping regions of the R and G segments derived from the images is shown. **(c)** 3D scatterplots of mean intensity of segments in R, G and B channels for the images from the ZF123 set of cell samples, and 2D histograms of pixel intensities of the non-overlapping regions of the R and G, R and B, and G and B segments derived from the images is shown. NOVS segments are shown with filled markers and OVS segments are shown with unfilled markers.

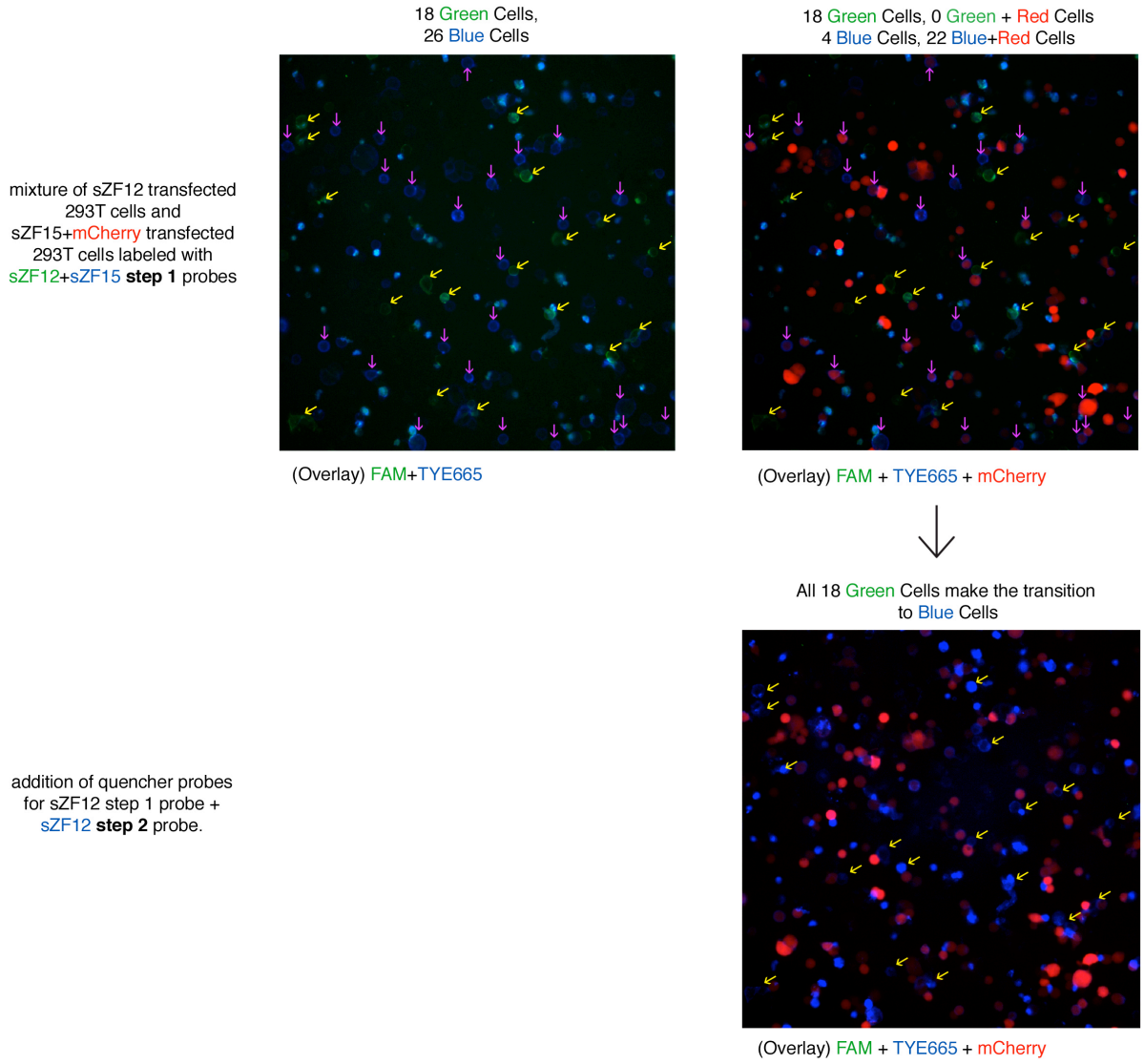

c
ZF123_1
ZF123_2
ZF123_3
ZF123_4
ZF123_5
R non-ovlp
R ovlp
G non-ovlp
G ovlp
B non-ovlp
B ovlp
mean B intensity
mean G intensity
mean R intensity
pixels
R intensity
G intensity
B intensity
R non-ovlp
G non-ovlp
R-G ovlp
G non-ovlp
B non-ovlp
G-B ovlp
R non-ovlp
B non-ovlp
R-B ovlp



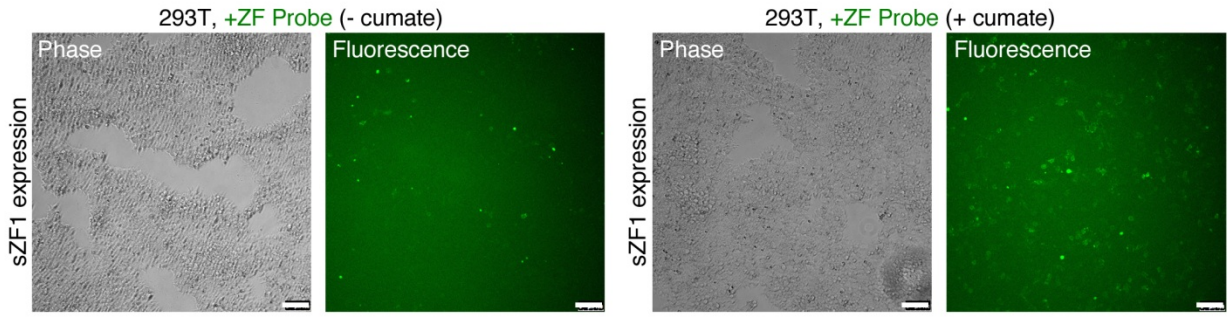
Supplementary Figure 9. Cell labeling through combinatorial expression of sZFs. Since 3 colors can be used to label up to 7 distinct cell types, cells bearing all possible combinations of sZF1, sZF3 and sZF4 were mixed and stained. Singly, doubly and triply stained cells can be visually distinguished easily in the resulting images showing that 7 distinct cell types can be successfully labeled in this manner. The scale bar is 100microns.



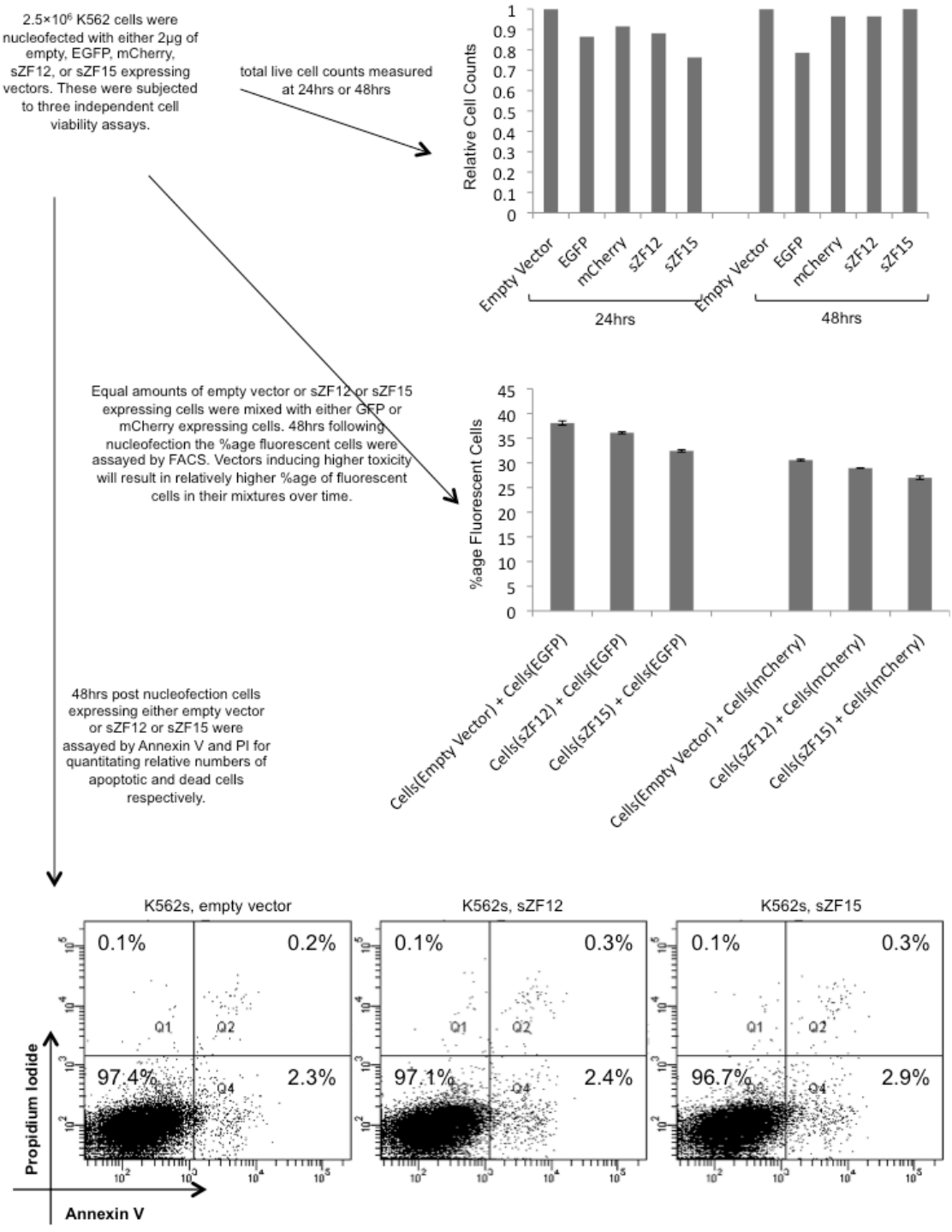
Supplementary Figure 10. Re-probing sZF expressing cells. For certain applications, the ability to re-probe a cell with different labels or functional tags in a sequential manner (which is not feasible using fluorescent proteins) is also desired. Since the zinc-finger dsDNA interaction is a non-covalent interaction it should be feasible to displace the latter using a competing dsDNA ligand. Towards this, we examined the dsDNA dissociation kinetics from sZFs. Specifically, the dsDNA dissociation kinetics from sZF2 was examined by assaying residual fluorescence intensity from a bound dsDNA probe at 6 minute intervals and for 48 minutes total. Due to the high affinity of sZF2 for its target dsDNA it demonstrated low rates of dissociation (top image series, and blue curve in plot). Thus to enable re-probing, dissociation of bound dsDNA was promoted using high concentrations of a non-fluorescent target dsDNA in solution. This indeed resulted in a rapidly diminishing fluorescence signal over time (bottom image series, and brown curve in plot). In conclusion, while the high affinity of sZFs to their target dsDNA greatly facilitates ease of imaging, bound probes can be actively displaced and hence sZFs enable dynamic re-probing of cells. The scale bar is 100microns.



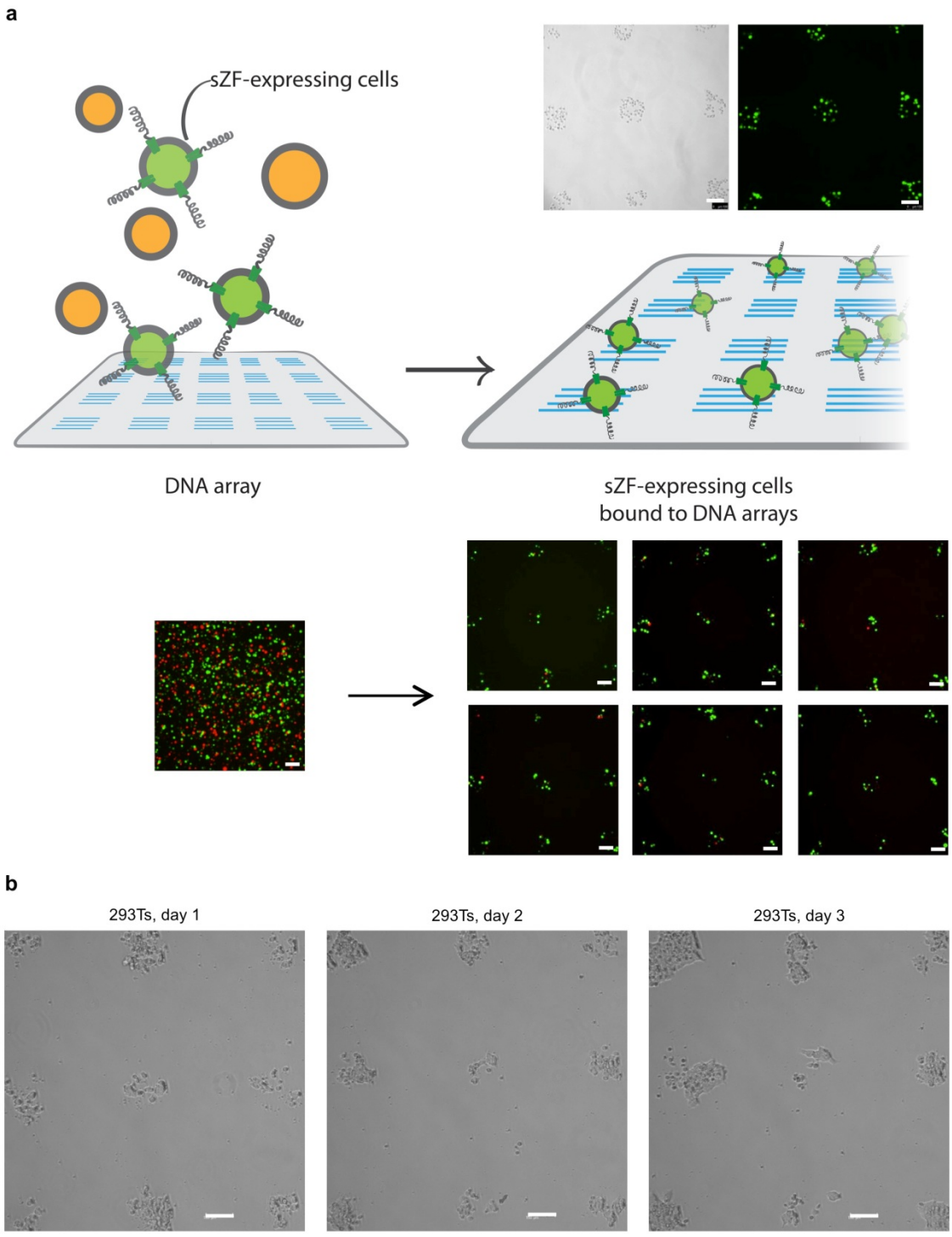
Supplementary Figure 11. Correlating genotype to labeling association in sZF expressing cells. Here cells expressing sZF12 or sZF15+mCherry were mixed and labeled using the scheme in **Fig. 2f**. Upon labeling with the step 1 probes it is evident that live cells are either labeled green or blue, and upon overlay of mCherry signal only the blue cells co-localize with it. Furthermore upon addition of step 2 probes, all green cells now change signal to blue. Together these observations confirm that cells expressing sZFs are correspondingly labeled by their cognate target DNA probes.



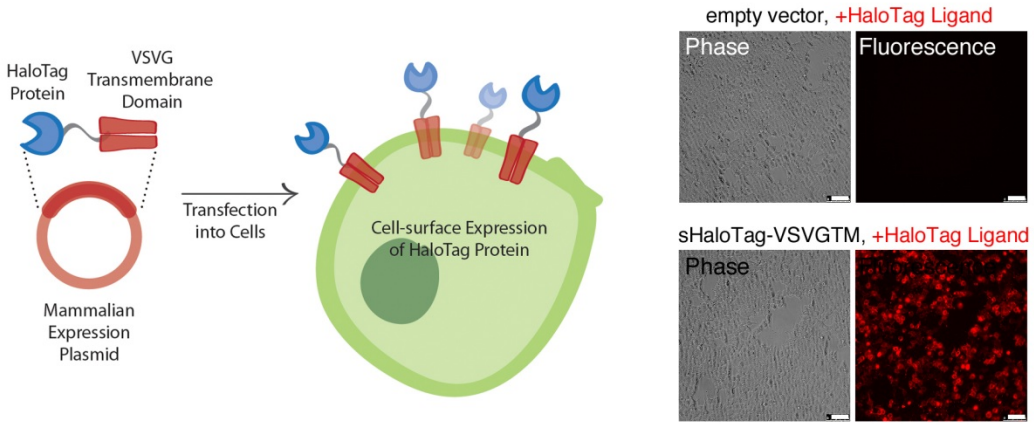
Supplementary Figure 12. Small molecule (cumate) inducible sZF expression. A lentiviral vector with a cumate inducible promoter to drive sZF expression was constructed and stably transduced in 293T cells. Upon small molecule induction sZF expression could be readily detected by the ability of the cells to bind dsDNA molecules. However, expression of sZFs from the tet responsive promoters (refer **Fig. 3a**) was observed to be significantly higher than from the cumate inducible promoters, but both inducible systems demonstrated small molecule responsive induction and can thus be used as versatile tools for barcoding cells. The scale bar is 100microns.



Supplementary Figure 13. Toxicity analysis of sZFs expressed in K562s. K562 cells nucleofected by different sZFs or control plasmids were analyzed for their relative viability using three different approaches as indicated. It is evident that in these cells sZF expression does not adversely affect their viability.



Supplementary Figure 14. Capture of sZF expressing cells on dsDNA arrays. (a) We show that from a mixture of K562 cells (expressing either mCherry or sZF8+GFP), the sZF expressing constituents can be selectively captured on a dsDNA array. (b) In experiments with 293Ts which are normally adherent cells, we also show their adherence to the dsDNA arrays does not adversely affect their viability over a period of 72hrs. The scale bar is 100microns.



Supplementary Figure 15. Cell-surface HaloTag expression using a VSVG transmembrane domain. The HaloTag protein was fused at its N-terminus to a Ig κ-chain leader sequence and at the C-terminus to the Vesicular Stomatitis Virus protein G (VSVG) transmembrane domain. This enabled cell surface expression of the HaloTag protein as evidenced by binding to its cognate fluorescent HaloTag ligand (right panel). This format of cell-surface expression also facilitated successful incorporation of the HaloTag protein on the membrane of lentiviruses during viral production in 293T cells. The scale bar is 100microns.

Name	Target Sequence	Protein Sequence	Vectors Constructed
sZF1	gGTCGGGGTAg	SRPGERPFQCRICMRNFSQDSSLRRHTRTHTGEKPFQCRICMRNFSRQEHLVRH LRTHTGEKPFQCRICMRNFSDPTSLNRHLKTHLRGS	transient, stable (Lentivirus: tet, cumate inducible)
sZF2	tGAAGCAGCAc	SRPGERPFQCRICMRNFSSQTQLVRHTRTHTGEKPFQCRICMRNFSQSTTLKRHL RTHTGEKPFQCRICMRNFSQRNNLGRHLKTHLRGS	transient, stable (Lentivirus: tet, cumate inducible)
sZF3	tGTGGCGGATa	SRPGERPFQCRICMRNFSTRQNLDTHTRTHTGEKPFQCRICMRNFSRRDTLERH LRTHTGEKPFQCRICMRNFSRPDALPRHLKTHLRGS	transient
sZF4	cGAGGACGGCa	SRPGERPFQCRICMRNFSAPSKLDRHTRTHTGEKPFQCRICMRNFSDESNLRRHL RTHTGEKPFQCRICMRNFSRVDNLPRHLKTHLRGS	transient
sZF5	aGAAGATGGTg	SRPGERPFQCRICMRNFSTNQKLEVHTRTHTGEKPFQCRICMRNFSVRHNLQRH LRTHTGEKPFQCRICMRNFSQHPNLTRHLKTHLRGS	transient
sZF6	gGACGACGGCa	SRPGERPFQCRICMRNFSAPSKLDRHTRTHTGEKPFQCRICMRNFSLGENLRRHL RTHTGEKPFQCRICMRNFSDGGNLGRHLKTHLRGS	transient
sZF7	aGTCGATGCCc	SRPGERPFQCRICMRNFSHQRDLMRHTRTHTGEKPFQCRICMRNFSVRHNLTR HLRTHTGEKPFQCRICMRNFSDPTSLNRHLKTHLRGS	transient
sZF8	gGAGGACGGCa	SRPGERPFQCRICMRNFSAPSKLDRHTRTHTGEKPFQCRICMRNFSLVENLRRHL RTHTGEKPFQCRICMRNFSRVENLHRHLKTHLRGS	transient
sZF9	aTTAGAAGTGa	SRPGERPFQCRICMRNFSRNFILQRHTRTHTGEKPFQCRICMRNFSQGGNLVRH LRTHTGEKPFQCRICMRNFSQQTGLNVHLKTHLRGS	transient
sZF10	aTTATGGGAGa	SRPGERPFQCRICMRNFSRQSNLSRHTRTHTGEKPFQCRICMRNFSRNEHLVLHL RTHTGEKPFQCRICMRNFSQKTGLRVHLKTHLRGS	transient
sZF11	cGAAGACGCTg	SRPGERPFQCRICMRNFSGRQALDRHTRTHTGEKPFQCRICMRNFSDKANLTRH LRTHTGEKPFQCRICMRNFSQRNNLGRHLKTHLRGS	transient
sZF12	tGAGGACGTGt	SRPGERPFQCRICMRNFSRNFILQRHTRTHTGEKPFQCRICMRNFSDRANLRRH LRTHTGEKPFQCRICMRNFSRHDQLTRHLKTHLRGS	transient
sZF13	aGACGCTGCTc	SRPGERPFQCRICMRNFSTGQILDRHTRTHTGEKPFQCRICMRNFSVAHSLKRHL RTHTGEKPFQCRICMRNFSDPSNLRRHLKTHLRGS	transient
sZF14	aGAGTGAGGAc	SRPGERPFQCRICMRNFSRQDRLDRHTRTHTGEKPFQCRICMRNFSQKEHLAG HLRTHTGEKPFQCRICMRNFSRRDNLNRHLKTHLRGS	transient
sZF15	aTGGGTGGCAt	SRPGERPFQCRICMRNFSNKTDLGRHTRTHTGEKPFQCRICMRNFSRMDVLTRH LRTHTGEKPFQCRICMRNFSRSDHLSLHLKTHLRGS	transient
sZF16	cTGGGGTGCCc	SRPGERPFQCRICMRNFSSKKSLTRHTRTHTGEKPFQCRICMRNFSEAHHLSRHL RTHTGEKPFQCRICMRNFSRSDHLSLHLKTHLRGS	transient

Supplementary Table 1. List of 16 zinc fingers used in this study, their protein & target DNA sequences, and vectors constructed for mammalian gene expression.

	mean	(std)	min	max
Number segments	47.1	(9.8)	33	71
Number NOVS (1 channel)	38.2	(7.9)	27	57
Fraction NOVS (1 channel)	0.81	(0.08)	0.67	0.94
Number NOVS (2 channel)	32.3	(6.9)	24	45
Fraction NOVS (2 channel)	0.65	(0.05)	0.61	0.71

Supplementary Table 2. Counts and fractions of segments and NOVS formed per channel aggregated over all images and channels. When NOVS are identified from R, G, or B segments, they can either be identified relative to another single channel (NOVS "1 channel" segments, e.g., R NOVS identified as R segments that do not overlap any G channel segment), or, for three-channel images, two other channels (NOVS "2 channel" segments, e.g., R NOVS identified as R segments that overlap neither a G nor a B channel segment). Separate statistics are given here for the former *vs.* the latter case because there are likely to be fewer NOVS relative to 2 channels *vs.* 1 channel.

Name	Sequences
5' Alexa488 probe	5'-TATGAGGACGAATCTCCCGCTTATA-3'
5' Alexa546 probe	5'-GTTTATCGGGCGTGGTGCTCGCATA-3'
5' Alexa547 probe	5'-TAGTAGTTCAGACGCCGTTAAGCGC-3'
sZF1_Alexa488 probe	5'-atgtgGTCGGGGTAgcggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACACCAGCCCCATCGCCG-5'
sZF1_Alexa546 probe	5'-atgtgGTCGGGGTAgcggc-3'
	3'-CAAATAGCCCGCACCACGAGCGTATTACACCAGCCCCATCGCCG-5'
sZF1_Alexa647 probe	5'-atgtgGTCGGGGTAgcggc-3'
	3'-ATCATCAAGTCTGCGGCAATTCGCGTACACCAGCCCCATCGCCG-5'
sZF2_Alexa488 probe	5'-cacatGAAGCAGCAcgact-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATGTGTACTTCGTCGTGCTGA-5'
sZF3_Alexa488 probe	5'-atgttGTGGCGGATattga-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACAACACCGCCTATAACT-3'
sZF4_Alexa488 probe	5'-acatcGAGGACGGCagcgt-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTGTAGCTCCTGCCGTCGCA-5'
sZF5_Alexa488 probe	5'-ttgaaGAAGATGGTgcgct-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATAACTTCTTCTACCACGCGA-5'
sZF6_Alexa488 probe	5'-tcaagGACGACGGCaacta-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATAGTTCCTGCTGCCGTTGAT-5'
sZF7_Alexa488 probe	5'-ttgaaGTCGATGCCcttca-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATAACTTCAGCTACGGGAAGT-5'
sZF8_Alexa488 probe	5'-tcaagGAGGACGGCaacat-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATAGTTCCTCCTGCCGTTGTA-5'
sZF9_Alexa488 probe	5'-atgtaTTAGAAGTGacggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACATAATCTTCACTGCCG-5'
sZF10_Alexa488 probe	5'-atgtaTTATGGGAGacggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACATAATACCCTCTGCCG-5'
sZF11_Alexa488 probe	5'-atgtcGAAGACGCTgcggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACAGCTTCTGCGACGCCG-5'
sZF12_Alexa488 probe	5'-atgttGAGGACGTGtcggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACAACTCCTGCACAGCCG-5'
sZF13_Alexa488 probe	5'-atgtaGACGCTGCTccggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACATCTGCGACGAGGCCG-5'
sZF14_Alexa488 probe	5'-atgtaGAGTGAGGAccggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACATCTCACTCCTGGCCG-5'
sZF15_Alexa488 probe	5'-atgtaTGGGTGGCAtcggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACATACCCACCGTAGCCG-5'
sZF16_Alexa488 probe	5'-atgtcTGGGGTGCCccggc-3'
	3'-ATACTCCTGCTTAGAGGGCGAATATTACAGACCCCACGGGGCCG-5'

Supplementary Table 3. Sequences for sZF DNA probes, I. List of sZF probes used in **Figs. 1,2(a,b,c),3** and **Supplementary Figs. 1, 9, 10, 12**. To avoid synthesis of new fluorescent probes for each tested sZF they all share a binding domain for a common fluorophore bearing oligonucleotide (shown here for a Alexa488 fluorophore bearing probe). sZF probes for binding Alexa546 and Alexa647 bearing fluorophores were similarly constructed.

Name	Sequences	Label
Step1Color1	5'-TATGAGGACGAATCTCCCGCTTATA-3'	5' 6-FAM
Step1Color2	5'-TAGTAGTTCAGACGCCGTTAAGCGC-3'	5' Hex
Step1Color3	5'-TATCCCGTGAAGCTTGAGTGGAATC-3'	5' TYE 665
Step2QuenchBHQ	5'-ATGCACTATTTTACGTATCCCGTGC-3'	3' Black Hole-1
Step2QuenchFQ	5'-TATGTTGTGCCTTACGCCTCGATTA-3'	3' IowaBlackFQ
Step2QuenchRQ	5'-TTAACCGAACTGACGGCCATCAAGG-3'	3' IowaBlackRQ
Step3Color1	5'-TTCTATTCTAAGCCGGCGGTCATAT-3'	5' 6-FAM
Step3Color2	5'-TCCAAGTTAGCTTACTCCATGCCCC-3'	5' Hex
Step3Color3	5'-TCCATAGATTTCTCCGTGAGTCTTT-3'	5' TYE 665
MC_sZF02	5'-cacatGAAGCAGCAcgact-3'	
	3'-AGGTTCAATCGAATGAGGTACGGGGTACGTGATAAAATGCATAGGGCACGATACTCCTGCTTAGAGGGCGAATATGTGTACTTCGTCGTGCTGA-5'	
MC_sZF03	5'-atgttGTGGCGGATattga-3'	
	3'-AGGTATCTAAAGAGGCACTCAGAAAATACAACACGGAATGCGGAGCTAATATCATCAAGTCTGCGGCAATTCGCGTACAACACCGCCTATAACT-5'	
MC_sZF06	5'-tcaagGACGACGGCaacta-3'	
	3'-AAGATAAGATTCGGCCGCCAGTATAAATTGGCTTGACTGCCGGTAGTTCCATAGGGCACTTCGAACTCACCTTAGAGTTCCTGCTGCCGTTGAT-5'	
MC_sZF12	5'-atgttGAGGACGTGtcggc-3'	
	3'-AGGTATCTAAAGAGGCACTCAGAAATACGTGATAAAATGCATAGGGCACGATACTCCTGCTTAGAGGGCGAATATTACAACTCCTGCACAGCCG-5'	
MC_sZF14	5'-atgtaGAGTGAGGAccggc-3'	
	3'-AAGATAAGATTCGGCCGCCAGTATAATACAACACGGAATGCGGAGCTAATATCATCAAGTCTGCGGCAATTCGCGTACATCTCACTCCTGGCCG-5'	
MC_sZF15	5'-atgtaTGGGTGGCAtcggc-3'	
	3'-AGGTTCAATCGAATGAGGTACGGGGAATTGGCTTGACTGCCGGTAGTTCCATAGGGCACTTCGAACTCACCTTAGTACATACCCACCGTAGCCG-5'	

Supplementary Table 4. Sequences for the sZF DNA probes (sequential labeling), II. List of sZF probes used for the sequential labeling experiments in **Fig. 2(d,e,f,g,h)** and **Supplementary Fig. 11**.

Supplementary Note 1. Image Analysis Methods

We used image analysis to obtain quantitative measures of the specificity of binding of cell surface-expressed ZFPs (sZFs) to their corresponding oligos. Whereas mixture experiments were conducted with mixtures of up to 9 samples of cells expressing distinct sZFs (here termed *sZF cell types*), only three fluorescent oligo labels could be distinguished during a single scan so that processing >3 oligos required multiple rounds of oligo treatments and scans. Because cell movements during multiple rounds complicated image analysis, we confined our quantitative analysis to experiments involving two or three sZF cell types.

Preliminary work with initial images established key functional requirements: (i) Images acquired for each of the (up to three) fluorophores required independent normalization due to different profiles of background noise and highly bright regions. The presence of small regions of very high intensity from cell debris or aggregated labeled oligos caused cells to appear very dim when images were normalized to their maximum intensities in each channel, so that intensity clipping was needed to establish appropriate dynamic range for the live oligo-labeled cells. (ii) Dead cells and some cell debris among the live cells needed to be excluded from analysis since they non-specifically adsorbed oligos of all species. (iii) With suitable corrections for (i) and (ii), live sZF-expressing cells could be identified as image segments with intensity above appropriate thresholds in at least one image channel. However, because this mode of segmentation automatically forces identified cells to have high intensity in one channel but not (at least, not necessarily) in other channels, it introduced a potential bias into measures of labeling specificity based on direct comparisons of within-cell intensities across channels. We therefore focused attention on developing measures of specificity that avoided direct comparisons within cells of intensity levels across channels. Each of these three requirements was taken into account in designing our image analysis.

We decided early on that the relatively small number of images combined with substantial variability of appearance of cells within and between images was more compatible with human interactive *vs.* fully automated image processing. However, we sought to impose discipline on our procedures by developing a suite of three MatLab (The Mathworks, Natick, MA) GUI applications using MatLab's GUIDE tools and Image Analysis and Statistics Toolboxes: *ImageNormalizer*, *ImageMasker*, and *SegmentOverlapAnalysis*. These applications provided a framework by which human actions were constrained to setting parameters within a structured set of processes. The applications were designed to be used in turn *via* the processing flow depicted in **Supplementary Fig. 3**. These applications have been made freely available for non-commercial research and can be downloaded from our supplemental web site http://arep.med.harvard.edu/sZF_cell_barcode/ along with documentation on their usage. Briefly:

ImageNormalizer is used normalize the individual channels of the images acquired from the fluorescence microscope, and produce a consolidated three channel TIF image that is used as input by the other two applications. For each sample, the images produced by the microscope comprise a 3-channel 1024x1024 pixel JPG per fluorophore, only one channel of which contains actual data. Normalization options include specification of an upper clip intensity (provided as a percentile intensity) to adjust for small bright regions (cf. requirement (i) above), and also a choice of one of three background subtraction algorithms (None, Linear, and Quadratic). To process Linear or Quadratic background subtraction, a user-specified background threshold is used to identify the image foreground and a morphological closure operation is applied to de-noise the edges of the foreground. The background is then recalculated as the complement of the morphologically closed

foreground, and a best-fit linear or quadratic of background intensity *vs.* 2D coordinates is subtracted from the clipped image. The result is then renormalized to maximum intensity 1 in each channel. A histogram of upper level intensities and several viewing options for the image (including composite, single channel, and false color) are available to the user to help guide selection of normalization options and settings of parameters. The user saves the normalized image when satisfied with the result to the final normalized TIF image along with a log that records the parameters and settings used.

ImageMasker is used to mask out dead cells, cell detritus, or other irregular patches of intensity from subsequent analysis. As cell images were acquired as single confocal z-slices, normal individual live cells could be distinguished by roughly circular or elliptical ring-like appearance while dead cells appeared as bright, completely filled masses. However, live cells were also often in small clumps and sometimes adjacent to dead cells, so that automated identification of live cells would be computationally challenging. *ImageMasker* was thus designed to make it easy for a human to scan an image and draw small mask regions over dead cells or other irregular features. By maintaining a file of mask coordinates for each image, *ImageMasker* enables users to iteratively review and reprocess masks. *ImageMasker* also allows users to magnify parts of the image so that smaller features can be accurately masked, and like *ImageNormalizer* also supports many modes of viewing the images. In addition to the coordinate file, a binary TIF image of masked out regions is written by *ImageMasker* when a user saves a mask. This binary TIF image is used as an input to the subsequent application.

SegmentOverlapAnalysis, the final application in the suite, allows the user to interactively control the segmentation of the image into cell regions, and, when the user is satisfied with the segmentation, to submit the segmented image to statistical analysis. Segmentation is performed separately for each channel in three steps: (i) A user-specified intensity threshold is first used to separate foreground segments from background pixels. (ii) A morphological closure operation of user-specified size consolidates segments separated by short distances. Finally, (iii) a user-specified area threshold is used to filter away small segments. To facilitate user evaluation of the segmentation, statistics on the numbers of segments and segment area means and standard deviations for each channel are displayed, and (as with the other applications) numerous options for viewing the image are provided. Regions masked *via ImageMasker* are prevented from being included within segments by setting their intensities to zero prior to applying the operations above. Because sZF expressing cells were sometimes in small clumps that emerged from the segmentation process as single large segments, the segment number and area statistics were not always useful in setting intensity thresholds. Instead we generally looked by eye for a compromise threshold that appeared to maximize production of ring-like shapes (including clumps of rings) without yielding many segments of aberrant size or other shape.

When the user is satisfied, clicking on a button causes the system to calculate the statistics. Statistical calculation is not interactive and cannot be manipulated directly by the user within the application. This was a design objective intended to increase the discipline of the analysis by making it hard for users to either directly or unconsciously adjust image processing to generate more favorable statistical results. For similar reasons, image normalization and masking were isolated from segmentation in separate applications.

Processing results: Eleven image sets were processed with this series of applications. Under *ImageNomalizer*, all channels were clipped for bright regions at either the 99.6^{th} or 99.8^{th} percentile of

channel intensity; we found background did not significantly impact analysis and so no background subtraction was used. Using *ImageMasker*, between 21 and 59 masks were drawn per image (mean=37.0, standard deviation=12.7). Total masked area ranged between 1.36% and 4.95% (mean = 2.94%, standard deviation=1.43%). Segment numbers and areas generated by use of *SegmentOverlapAnalysis* are given below in **Supplementary Table 2** and **Supplementary Figs. 5 and 6**. Output files generated by all applications for all images have been made available as Supplemental Data and can be obtained at http://arep.med.harvard.edu/sZF_cell_barcode/. The output files contain all application parameter values we set for all images in the course of the analysis.

Supplementary Note 2. Image Analysis Statistics

The statistics implemented in *SegmentOverlapAnalysis* were designed to provide quantitative measures of the ability of sZFs to label cells and the specificity of the sZFs. However, as noted above, the identification of cell regions by intensity thresholding potentially biases comparative numerical measures of channel intensities in cell regions. To mitigate this bias, we computed correlations between channel intensities in segment regions, and also devised measures that assessed the degree to which segment areas generated in different channels overlap with each other. Slight differences in analysis are required for two-channel *vs.* three channel images.

Two-channel images

Correlation measures included:

(i) *Whole image correlations*, i.e., Pearson correlation coefficients computed for the intensities in the two channels at every pixel in the entire 1024x1024 image.

(ii) *Channel pair segment correlations,* i.e., Pearson correlation coefficients for the intensities in the two channels at every pixel in the union of all segment areas derived for both of the two channels.

For channel pair segment correlations, the statistical significance of the correlation was estimated from the distribution of correlations computed from 1000 random shuffles of the intensities of pixels within the union of the segment areas of the two channels. If the sZFs are specific for their oligos, we would expect to see negative correlations between the intensities of the pixels.

Segment overlap measures attempt to quantify the degree to which segments formed in one channel are distinct from segments formed in the other. Given a two-channel RG image, a segment in the R channel might either be entirely disjoint from (i.e., contain no pixels also contained in) segments in the G channel: we termed such segments *non-overlap segments* (NOVS). The alternative is that a segment in the R channel might overlap one or more segments in the G channel (*overlap segments*, OVS). Within an R OVS, the area of overlap with G segments is called the *OVS-overlap region* (OVS-ovlp), while the area that does not overlap G segments is called the *OVS-non-overlap region* (OVS-nonovlp). OVS segments could be formed either (a) from sZF cell types whose sZFs are specific for their oligos that happen to at least partly overlap sZF cells of the other type, or (b) from sZF cell types whose sZFs are *not* specific to their oligos. However, if (b) obtains, we would expect to see OVS segments exclusively; therefore specificity is indicated by the presence of substantial numbers of NOVS in each channel. This gave rise to several measures, including:

(iii) *Counts of NOVS per channel*, and *fractional abundance of NOVS among all segments in a channel*.

In addition to counting NOVS according to the strict criterion above, in which to be an NOVS requires that *no* pixels be present in the segments of the other channel, we also considered counts and fractional abundances of segments for which a *below-threshold fraction* of area might be contained in segments of the other channel. The intent of this measure was to allow us to distinguish and count cells that might overlap a cell bearing the other ZFP by a small amount. In support of this measure, users are allowed to set an "overlap area threshold" parameter within the *SegmentOverlapAnalysis* application. However, we ultimately did not find this capability useful and computations generated by this feature were not further analyzed.

(iv) *Channel intensity measures in NOVS and OVS:* On the hypothesis that NOVS represent cells with specific sZFs and OVS represent random overlaps from adjacent cells of different sZF cell types, we would expect to see channel intensities in the segment channel in the NOVS that are similar to the channel intensities seen in the OVS-nonovlp regions of OVS segments. Due to the potential bias in comparing intensities in the channel in which a cell has been segmented with the other channels (described above), we generally avoided analyzing intensity differences across channels within these regions. However, within OVS-ovlp regions, channel intensities in both channels will be above their segmentation thresholds, so that we would expect relative parity of channel intensities in these regions.

To support evaluation of these hypotheses, we therefore computed for each segment: (iv.a) the *mean and standard deviation of pixel intensities* across entire NOVS and OVS segments for all channels (e.g., we computed for R segments the means and standard deviations of R and G intensities); (iv.b) the mean and standard deviation of the channel intensities within the OVS-ovlp regions (e.g., the mean R intensity in the OVS-ovlp regions of OVS R segments); and (iv.c) a Wilcoxon rank sum p-value comparing the two channels in OVS-ovlp regions (e.g., a Wilcoxon p-value of the R *vs.* G intensities of all pixels in the OVS-ovlp region). Wilcoxon rank sum p-values were computed using the MatLab ranksum function.

(v) *2D pixel intensity histograms across segment regions:* The intensity and ranksum measures in (iv) attempt to capture the relative similarity of channel intensities in OVS-ovlp regions *vs.* their distinctness in NOVS and OVS-nonovlp regions on a segment-by-segment basis. As a second way of evaluating these relationships, we constructed three 2D histograms of the pixel intensities in both channels: one each for the sum total of NOVS and OVS-nonovlp regions in each channel (in which pixels are entirely within segment areas of one channel exclusively), and one for the OVS-ovlp region (which is common between the channels). Thus, in an RG image, a 2D histogram of R and G intensities was constructed for R NOVS and OVS-nonovlp regions, for G NOVS and OVS-nonovlp regions, and for the common OVS-ovlp region. If sZFs are specific, we expect to see strong concentration of pixels with high R and low G in the R NOVS and OVS-nonovlp regions, a strong concentration of pixels with low R and high G in the G NOVS and OVS-nonovlp regions, and a more even distribution of high R and G intensities across the OVS-ovlp region.

Three-channel images

The measures described above all analyze intensities or areas obtained from a pair of channels. These same measures are computed for three-channel (RGB) images *for each of the three pairs of channels* (RG, RB, and GB), but are then also extended in a number of ways to include consideration of the third channel.

Correlation measures

(i) *Whole image correlations* are computed *for each pair* of channels.

(ii) *Channel pair segment correlations* are also computed *for each pair* of channels.

(ii-a) Additionally, correlation coefficients are also computed in three-channel images for each pair of channels across the union of all segment areas from all three channels (*all segment channel pair correlations*). Thus, while (ii) comprises correlations of R and G in the union of R and G segment areas, of R and B in the union of R and B segment areas, and of G and B in the union of G and B segment areas, the three *all segment channel pair correlations* are between

R and G in the union of R, G, *and B* segment areas, between R and B in the same union, and between G and B in that union.

Statistical significances are computed for both the channel pair segment correlations (ii), *and also for the all segment channel pair correlations (ii-a)*, by randomly shuffling intensities of the channels in the prescribed areas.

Segment Overlap measures

(iii) *Counts and fractional abundances of NOVS per channel*. These are computed as above *for each pair of channels*, but then also extended to consider segments in one channel that overlap *either* of the other channels. For instance, counts and fractional abundances are computed for R segments that do not overlap any G segment, and for R segments that do not overlap any B segment, but, in addition, counts and fractional abundances are computed for R segments that do not overlap *either* any B or G segments.

(iv) Channel intensity measures in NOVS and OVS: Again, in addition to computing the two-channel intensity measures for the three possible pairs of channels RG, RB, and GB, an additional intensity measurement is computed that compares a channel against an aggregate of the other two channels. Specifically, for a given channel, OVS segments are identified against the union of segments from the other two channels, in which each pixel of the OVS-ovlp region is assigned the maximum of the other two channel intensities, and a Wilcoxon rank sum p-value is computed in this region comparing the first channel against the maximum of the other two. Thus, for instance, for each R segment that overlaps either G or B segments, a Wilcoxon rank sum p-value is computed for the union of the G and B segment overlaps that compares R intensity against the maximum of the G and B intensities in this overlap region.

(v) *2D pixel intensity histograms across segment regions:* Three sets of pairwise 2D pixel intensity histograms across pairs of segment regions are produced for the RG, RB, and GB channel pairs.

Supplementary Note 3. Image Analysis Results Summary

Before summarizing results derived from the analysis of these measures, we make a few general observations:

- *SegmentOverlapAnalysis* reports the numerical values of all these measures in an output file, but, for a subset of these measures, also generates figures portraying them. The results presented below comprise a selection of figures generated from the numerical data to summarize results, and a compilation of selected *SegmentOverlapAnalysis*-generated summary figures. As noted above, the complete set of output files and figures is provided as Supplemental Data on http://arep.med.harvard.edu/sZF_cell_barcode/.
- Except for the correlation analyses above, few of the measures above can be formally tested for statistical significance for two main reasons: (i) In many cases no appropriate null hypothesis can be stated. For instance, although as noted above, the presence of NOVS segments in a channel is evidence that the sZFs are specific for their oligos *vs.* the hypothesis that the sZFs are non-specific, there is no clear choice of null hypothesis regarding the distribution of the number of NOVS segments that might arise for non-specific sZFs. Rather, it seems likely that given non-specific ZFPs, there should simply be no NOVS. (ii) Intensity data analyzed in the context of OVS and NOVS segments is subject not only to the bias noted above that relates to use of thresholds in the image segmentation, but also to additional complications, such as the fact that the segmentation of the different channels may employ different thresholds in each channel, and these thresholds may also differ on an image-by-image basis. Therefore, even in the OVS-ovlp analysis comparing intensities of two channels in regions where both should be above-threshold, there can be no consistent expectation regarding whether one channel should be higher than the other. Therefore, the Wilcoxon p-values computed for these channel intensities are intended to be used only to indicate the degree of difference between the channel intensities in these regions *vs.* as actual formal tests of statistically significant differences.

Correlation measures: **Supplementary Fig.4** summarizes all correlation coefficients and CDFs computed for the images within the study. Pixel intensity correlations between channels that reflect binding of labeled oligos corresponding to sZFs are all (with one exception) negatively correlated with very high statistical significance in regions of the image containing cells. This is evidence of the high specificity of the sZFs.

Counts and fractional abundance of NOVS: **Supplementary Fig. 5** summarizes the counts of segments of all channels in each image, and the counts and fractional abundance of NOVS in all channels. As noted above, the presence of substantial numbers and fractions of NOVS in each channel are evidence that the sZFs are highly specific to their oligos. Aggregated information on segment and NOVS counts and fractions is given in **Supplementary Table 2**. In general, ~38 NOVS were formed in any one channel.

Channel segment area data for all segments: **Supplementary Fig. 6** describes segment areas for all, all NOVS, and OVS segments in all images of this study, as well as the breakdown of OVS into OVS-nonovlp and OVS-ovlp parts. As described above, area filters were applied in the course of image segmentation, but these could be different across channels and images; therefore no overall relationship is expected to be observed for segment areas. However, in general OVS-ovlp areas tend to be small compared to OVS-nonovlp areas, possibly consistent with the hypothesis that cell overlaps tend to be small regions confined to the peripheries of cells.

Channel segment intensity data and OVS-ovlp rank sum intensity comparisons: **Supplementary Fig. 7** provides distribution-level information on segment intensities OVS-ovlp rank sum p-values (see above). Segment intensity distributions exhibit considerable variation across channels and images, possibly reflecting the differences in intensity thresholds that were used to segment channels across the image set as well as the sensitivity and saturation characteristics of the different fluors used to label the oligos. Nevertheless, it is of note that despite these factors, rank sum p-values in OVS-ovlp regions, in which the intensities of the channels compared are all above their segmentation thresholds, mostly indicate the absence of statistically large differences in central values. Although these p-values cannot be used to formally determine statistical significance for reasons noted above, it is also noteworthy that p-values as low as 0.001 are seen in a fraction of OVS-ovlp regions.

Scatterplots of mean intensities: For each pair of data-containing channels, scatterplots were generated showing the mean intensities of all segments within each of the channels, enabling comparison of the intensities in the two channels for each segment. **Supplementary Figs. 8a, 8b** show RG scatterplots for the three two-channel image sets acquired for cell samples ZF12 and ZF34. **Supplementary Fig. 8c** show the five three-way RGB channel plots for ZF123. In all these scatterplots, filled plot markers are used for NOVS segments and unfilled markers for OVS segments. In general, the filled markers for a channel tend to align closely to an axis that represents zero intensity for the other channel(s), indicating high specificity of sZFs to their oligos. Unfilled markers tend to align with the same axis except for being further away, which results from the intensity in the other channel conferred by the cell overlap. In general, all sZFs show high specificity except for some cross talk between the sZFs and oligos corresponding to the G and B channels in **Supplementary Fig. 8c**.

2D histograms of pixel intensities: **Supplementary Fig. 8a and 8b** show the RG 2D pixel intensity histograms for the three two channel image sets acquired for cell samples ZF12 and ZF34, respectively. **Supplementary Fig. 8c** shows the five RG, RB, and GB 2D pixel intensity histograms for the five 3 channel image sets acquired for cell sample ZF123. For each pair of channels with actual data, three 2D histograms were generated from pixel intensities corresponding to non-overlapping segment regions in each channel (all pixels from NOVS segments plus all pixels from OVS-nonovlp regions of OVS segments), and for the OVS-ovlp regions of the two channels. These are displayed as surface plots within the same figure, with colors corresponding to the channel non-overlapping regions and gray for the overlap regions. For instance, for the R and G channels of each image, one 2D histogram of R and G intensities is generated for all NOVS and OVS-nonovlp R segments and displayed as red, one 2D histogram of R and G intensities is generated for all NOVS and OVS-nonovlp G segments and displayed as green, and a 2D histogram of all R and G intensities in all R and G segment overlap regions is shown in gray. All 2D pixel histograms show high specificity for each single channel in the non-overlapping segment regions, and very dispersed intensities for the overlap regions.

Segment Maps of all images: For each image, each segment in each channel is assigned a numerical identifier which is reported in the numerical data on individual segments generated by *SegmentOverlapAnalysis*. Segment maps are copies of the original normalized input images in which the segment identifier numbers are printed as numbers within small boxes at their segment centroid locations. Segment maps are generated so that the cell regions corresponding to segments of interest can be visually located in the original images. Segment maps are generated both as grayscale images for each individual channel, and as a composite image showing all channels and all segments. Segments for a particular channel are shown as numbers within a box, where the color of the edge of the box indicates the channel in which the segment was derived. A portion of a segment map is shown in **Supplementary Fig. 3**. All segment maps generated in this study are available within the Supplemental Data provided at http://arep.med.harvard.edu/sZF_cell_barcode/.