Title

Image3C: a multimodal image-based and label independent integrative method for single-cell analysis

Authors

Alice Accorsi[1,2,†], Andrew C. Box[1,†], Robert Peuß[1,3,†], Christopher Wood[1], Alejandro Sánchez Alvarado[1,3,#,*], & Nicolas Rohner[1,4,#]

[†]These authors contributed equally to this study

[#]These authors share senior authorship

*Author for correspondence: Alejandro Sánchez Alvarado (asa@stowers.org)

Affiliation

[1]Stowers Institute for Medical Research, Kansas City, MO, United States

[2]Howard Hughes Medical Institute, Stowers Institute for Medical Research, Kansas City, MO, United States

[3]Current address: Institute for Evolution and Biodiversity, University of Münster, Münster, Germany

[4]Department of Molecular & Integrative Physiology, KU Medical Center, Kansas City, KS, United States

Running headline

Image3C, tool for *de novo* morphology-based cell clustering

Abstract

Image-based cell classification has become a common tool to identify phenotypic changes in cell populations. However, this methodology is limited to organisms possessing well characterized species-specific reagents (e.g., antibodies) that allow cell identification, clustering and convolutional neural network (CNN) training. In the absence of such reagents, the power of image-based classification has remained mostly off-limits to many research organisms. We have developed an image-based classification methodology we named Image3C (Image-Cytometry Cell Classification) that does not require species-specific reagents nor pre-existing knowledge about the sample. Image3C combines image-based flow cytometry with an unbiased, high-throughput cell cluster pipeline and CNN integration. Image3C exploits intrinsic cellular features and non-species-specific dyes to perform *de novo* cell composition analysis and to detect changes in cellular composition between different conditions. Therefore, Image3C expands the use of imaged-based analyses of cell population composition to research organisms in which detailed cellular phenotypes are unknown or for which species-specific reagents are not available.

Impact statement

Image3C analyzes cell populations through image-based clustering and neural network training, which allows single-cell analysis in research organisms devoid of species-specific reagents or pre-existing knowledge on cell phenotypes.

46 **Introduction**

47 Single-cell analysis have proven crucial to our understanding of fundamental biological
48 processes such as development, homeostasis, regeneration, aging and disease (Goolam et al., 2016;
49 Kimmel et al., 2019; Pepe-Mooney et al., 2019; Philippeos et al., 2018; Tirosh et al., 2016). High-
50 throughput analyses of these and other biological processes at single cell-resolution require
51 technologies capable of describing individual cells and subsequently clustering them based on
52 similarities of features like morphology, cell surface protein expression or transcriptome profile.
53 Recent advances in image-based cell profiling and single-cell RNA sequencing (scRNA-seq) allow
54 quantification of differences between cell populations and comparisons of cell type composition
55 between samples (Caicedo et al., 2017). Single-cell studies that use traditional research organisms
56 (*e.g.*, mouse, rat or fruit fly) benefit from the availability of genomic platforms and established
57 antibody libraries. However, the same cannot be said for a growing number of important, yet
58 understudied research organism lacking such reagents and whose biological interrogation would
59 benefit immensely from single-cell analyses. In these cases, classical histochemical methods are
60 often used to identify and characterize specific cells. Yet, the successful identification and
61 enumeration of biologically meaningful cell types in such studies can be harmed by both the
62 limited number and variety of cellular attributes (few features or low dynamic range) available for
63 determination of cell types, and by observer bias when using traditional, hand-counting approaches
64 (*e.g.*, hemocytometer and Giemsa stain) (van der Meer, Scott, & de Keijzer, 2004). These
65 shortcomings, together with the lack of extensive knowledge on cell-specific phenotypes available
66 for training or for *a priori* assumptions usually results in the underestimation of the complexity of
67 cellular composition or interactions among cell types within tissues.

68 Automated classification of cells using convolutional neural networks (CNN, machine learning
69 method specialized in image recognition and classification) has become a promising approach for
70 accurate high-throughput cell analysis that is free from observer bias (Blasi et al., 2016; Eulenberg
71 et al., 2017; Kobayashi et al., 2017; Lei et al., 2018; Nassar et al., 2019). To date, CNN-based
72 automated clustering and classification techniques require pre-existing knowledge about the
73 organism or cell type of interest (*e.g.*, cell specific morphological traits within an image set) or the
74 availability of cell-specific reagents (*e.g.*, antibodies), or genomic sequence (*e.g.*, single-cell
75 sequencing) (Baron et al., 2019; Blasi et al., 2016; Eulenberg et al., 2017; Kobayashi et al., 2017;
76 Lei et al., 2018; Nassar et al., 2019). This means that to make effective use of artificial intelligence
77 (AI) approaches for single-cell analysis, one must have information available to train the algorithm
78 or for machine learning (ML) models, which often arises in the form of information gleaned from
79 the use of reagents like antibodies. Research areas that rely on inter-species comparisons or studies
80 on emerging research organisms would benefit from single cell-based analyses that do not require
81 pre-existing knowledge of cell types (*i.e.*, which is required for training a CNN for example) and/or
82 availability of antibodies or molecular databases. For example, within the interdisciplinary field of
83 eco-immunology, a growing number of researchers is investigating immune system adaptation to
84 different environments by studying immune cell compositions in diverse animals (Maizels &
85 Nussey, 2013). Given the influences of immune cell composition on the immune system response
86 of an organism (Kaczorowski et al., 2017), applying modern single cell analysis in eco-
87 immunological research would substantially increase our knowledge about the plasticity and
88 conservation of immune responses in a variety of different animals and conditions(Peuß et al.,
89 2020).

90 To make sophisticated cellular composition analysis available to any research organism without
91 the need for either pre-existing knowledge about the cell populations or species-specific reagents,
92 we developed Image-Cytometry Cell Classification (Image3C). This method analyzes, visualizes,
93 and quantifies, in a high-throughput and unbiased way, the composition of cell populations by
94 using cell morphological traits and non-species-specific fluorescent probes (*e.g.*, nuclear staining
95 or dyes for metabolic states that function well in a variety of organisms). By taking advantage of
96 cell morphology and/or fluorescent dyes related to function or metabolic state, Image3C can
97 analyze single cell suspensions derived from any experimental design, identify different cell types
98 and cluster them. Once the cell types are identified by *de novo* clustering on cell intrinsic features,
99 Image3C employs a CNN that uses these clusters as training sets, avoiding in this way user bias
100 or manual classification. This produces a CNN-based cell classifier 'machine' used to quantify
101 subsequently acquired image-based flow cytometry data and to compare cellular composition of
102 samples across multiple experiments, in a high-throughput and unsupervised manner without the
103 need for repeating time-consuming steps for *de novo* clustering. In comparison to existing label-
104 free cell clustering methods, Image3C does not require initial antibody staining (Lippeveld et al.,
105 2020; Nassar et al., 2019), pre-existing knowledge of specific cell morphology (Yakimov et al.,
106 2019) and is not limited to a specific cellular phenotype (Blasi et al., 2016) for *a priori*
107 identification of certain cell types. This makes Image3C extremely versatile and applicable to
108 virtually any research organism and tissue from which dissociated single cells can be obtained. In
109 sum, Image3C combines modern high-throughput data acquisition by image-based flow
110 cytometry, advanced and unbiased clustering analysis, statistics to compare cellular compositions
111 across different samples and a CNN classifier component to easily determine changes in cell
112 composition across multiple experiments.

113 **Results and Discussion**

114 *Image-Cytometry Cell Classification (Image3C)*

115 Image3C is an imaging tool developed to study tissue composition at single-cell resolution in
116 research organisms for which antibodies and pre-existing knowledge about cell types are not
117 readily available. Image3C allows for high-throughput and unbiased analysis in scenarios where
118 manual counting and observer-based cell identification are currently the only options. Image3C
119 includes all the components required for compensating captured images, quantifying multiple
120 features for each event, clustering the events, visualizing and exploring the data, training and using
121 the CNN for analyzing subsequent samples and integrating multiple experiments (Figure 1 and
122 S1).

123 Once a single-cell suspension is prepared from the organism of interest, the cells are stained
124 with a combination of dyes that are expected to function independently irrespective of the species
125 used, and which have high affinity for specific cellular organelles such as nuclei, or molecules
126 associated with metabolic states such as reactive oxygen species. We validated reagents
127 experimentally by determining that nuclear dyes stain intracellular material matching expected
128 characteristics of nuclear DNA or by activation of cells with drugs to change their metabolic state.
129 The labelled samples are then run on the ImageStream®^{X} Mark II (Amnis Millipore Sigma) and
130 images of individual events are collected for each channel of interest (Figure 1A).

131 Feature values from both morphological and fluorescent data, such as cell size and nuclear size,
132 are extracted from the cell images using IDEAS software (Amnis Millipore, free for download
133 upon creation of Amnis user account) (Figure 1B, Table S1a and S1b for feature description).
134 Correlation between features is calculated and redundant features are trimmed as well as samples
135 that, among replicates, are outliers (Figure S2 and S3). This prevents clustering artifacts potentially

136 caused by having multiple features providing the same information or including samples that are
137 not representative (Figure 1B). Finally, fluorescence intensity features are transformed to improve
138 homogeneity of variance of distributions and, if used, DNA staining is normalized to remove
139 intensity drift between samples and thus align the 2N and 4N DNA content histogram peaks
140 (Figure 1B and S4).

141 Exported feature quantifications are used for clustering the events. Dimensionality reduction
142 and visualization of clusters is achieved by generating force directed layout (FDL) graphs into
143 VorteX clustering environment (Figure 1C) (free to install) (Samusik, Good, Spitzer, Davis, &
144 Nolan, 2016). Cell images for events within each cluster can be visualized using FCS Express
145 together with custom R scripts (Figure 1D). These visualization tools and the cluster feature
146 averages (*i.e.*, the mean value of each feature for each cluster) (Figure 1E) allow to explore the
147 images of selected groups of events and the features that differ between cells belonging to separate
148 clusters. If control and treatment samples are included, a statistical analysis using negative
149 binomial regression to compare cell counts per cluster between samples is also available in the
150 Image3C pipeline. This high-throughput and unbiased analysis provides a comprehensive
151 understanding of a cell population composition at higher resolution than what is possible with
152 traditional histological methods.

153 Once this pipeline is run on a first set of samples (*e.g.,* homeostatic state) and the cell clusters
154 are defined for the tissue of interest, the images and the relative clustering IDs can be used to train
155 a CNN classifier in an unsupervised way (Figure 1F), including the ability to score frequency of
156 "new" cell types that do not match any of the clusters identified at homeostasis. Therefore, future
157 experiments in the same tissue used for training the CNN classifier can be analyzed directly
158 through the CNN (Figure 1G). This significantly reduces the number of steps and time required to

159 process data collected from following experiments with treated conditions. An even greater
160 advantage is represented by the fact that, in the absence of CNN, every time new experimental sets
161 are run it would be necessary to go again through the *de novo* clustering part of the pipeline (Figure
162 1B-1E) and the new set of clusters would need to be re-annotated to be compared with cell
163 population composition observed in previous experiments. Manually matching clusters between
164 different experimental sets might be a source of mistakes, mainly if the user is not familiar with
165 the cell types present in the sample and if specific biomarkers or pre-existing knowledge about cell
166 types and morphology are not available. The CNN splits all the cell images in the classes defined
167 during the training step and allows to compare the abundance of cells with same morphology
168 among different samples without the need to cross-annotate clusters (Figure 1F and 1G). The CNN
169 inclusion in Image3C and the reproducibility of image acquisition through image-based flow
170 cytometry allows use of the clusters defined from one experiment (*e.g.,* homeostatic state) to setup
171 a classifier in an unsupervised and unbiased way for later use as a reference in analyzing
172 experimental manipulations of these cell populations. We conclude from these results that
173 Image3C can perform *de novo* high-throughput characterizations and define specific cell type
174 behavior or population composition in both homeostatic and experimentally perturbed cell
175 populations across multiple experiments.
176
177 *Image3C recapitulates cell composition of zebrafish whole kidney marrow (WKM) tissue*
178 To test whether Image3C could identify homogeneous and biologically meaningful cell
179 populations, we used the research organism *Danio rerio*. We obtained cells from adult zebrafish
180 WKM (location of the hematopoietic tissue) in homeostasis condition, stained them and run on the
181 ImageStream®^{X} Mark II. We analyzed intrinsic morphological and fluorescent features, such as

182 cell and nuclear size, shape and darkfield signal (side scatter, SSC). Feature values were extracted
183 from each cell image and processed through our pipeline (Table S1a for feature description).
184 Clustering by the final set of normalized and non-redundant morphological and fluorescent
185 features produced distinct cell populations (Figure 2A, 2B, 2C and S2).

186 Image3C can distinguish between the major classes of cells present in zebrafish WKM (Figure
187 2, Data File S1 and S2) that were described using standard sorting flow cytometry and
188 morphological staining approaches (D. Traver et al., 2003). It is noteworthy that Image3C can
189 clearly identify dead cells and debris (Figure 2A and 2B) allowing to optimize experimental
190 protocols in order to minimize cell death and to run the subsequent analysis only on the intact, live
191 cells. Image3C can identify cells with outstanding morphological features, such as neutrophils
192 from other myelomonocytes (Figure 2B and 2C). Based on zebrafish neutrophil characteristics
193 such as high granularity, high complexity and low circularity of the nuclei (Lugo-Villarino et al.,
194 2010), this type of granulocytes can be easily distinguished. Other types of myelomonocytes, such
195 as monocytes and eosinophils are here merged in the same cluster, since in zebrafish they share
196 many morphological characteristics (Lugo-Villarino et al., 2010). Similarly, using only intrinsic
197 morphological features for the clustering, different lymphocytes (B and T-cells) and hematopoietic
198 stem cells cannot be separated from each other, but they can be clearly distinguished from the
199 myelomonocytes (Figure 2A and 2B). Within the Lymphocytes/Progenitors fraction we find two
200 clusters (*Dr*1 and *Dr*7) that mainly differ in cell diameter (Figure 2C). Whether this morphological
201 difference has a biological implication needs to be studied in future experiments.

202 Image3C also enables the quantification of cell population (clusters or CNN classes) relative
203 abundance, an important tool for comparing population composition across different treatment
204 groups under different environmental conditions (Peuß et al., 2020). Here, we compared our results

205 with previously published data to validate our method. Although a direct comparison with results
206 from classical approaches (David Traver et al., 2003) is not possible since we gated out (removed
207 analytically) mature erythrocytes before clustering (see Material and Methods), the
208 myelomonocyte to lymphocyte ratio (M/L ratio = 1.59) is similar to the one obtained with classic
209 histological approaches (mean M/L ratio = 1.35) (Figure 2D) (David Traver et al., 2003).
210
211 *Image3C identifies professional phagocytes in zebrafish WKM tissue*
212 Next, we sought to determine whether Image3C could be used to characterize and quantify
213 biological processes by identifying a tissue of interest and then comparing cellular composition
214 dynamic, function, and physiological responses of specific cell types across a range of
215 experimental conditions. Our goal was to detect statistically significant changes in cluster relative
216 abundance between control and treated samples to gain a more detailed understanding of cell
217 population dynamics and individual cell function.
218 As proof-of-concept, we performed a standard phagocytosis assay using WKM tissue from
219 adult zebrafish. The single cell suspension was incubated with CellTrace Violet labeled
220 *Staphylococcus aureus* (CTV-*S. aureus*) and with dihydrorhodamine-123 (DHR), a reactive
221 oxygen species that becomes fluorescent if oxidized to report oxidative bursting following
222 phagocytosis. As controls, we inhibited phagocytosis through cytoskeletal impairment with CCB
223 incubation or through incubation with bacteria at lowered temperature by placing tubes on ice.
224 Events collected on the ImageStream®^{X} Mark II were analyzed with Image3C and clustered in 26
225 distinct clusters using quantifications of morphological and fluorescent features (Table S1b for
226 feature description), including nuclear staining, phagocytized *S. aureus* and DHR positivity
227 (Figure 3A). Professional phagocytes are defined by their ability to take up *S. aureus* (CTV

228 staining lies within the cell boundary) and induce a reactive oxygen species (ROS) response (bright
229 DHR signal) (Rabinovitch, 1995). In zebrafish, professional phagocytes are mainly granulocytes
230 and monocytic cells and can be discriminated from each other based on morphological differences,
231 such as cell size, granularity and nuclear shape (Wittamer, Bertrand, Gutschow, & Traver, 2011).
232 To compare samples incubated with CTV-*S. aureus* and the samples where phagocytosis is
233 inhibited (CTV-*S. aureus* + CCB and CTV-*S. aureus* + Ice), we used the statistical analysis
234 included in Image3C based on a negative binomial regression model (Figure 3B, 3C and S5 and
235 Table S2 and S3). Statistical analyses reported clusters with differences in relative abundance
236 between phagocytosis and phagocytosis-inhibited samples. Visualizing these clustered event
237 images (Data File S3) while considering the values and intensities of morphological and
238 fluorescent features for these clusters (Data File S1) allowed identification of 4 clusters of
239 professional phagocytes: granulocytes within clusters *Dr*4_P, *Dr*12_P and *Dr*13_P and monocytic
240 cells in cluster *Dr*21_P (Figure 3A, 3B). The morphology of cells in cluster *Dr*12_P is
241 characteristic of phagocytic neutrophils (Figure 2B, 3A) that become adhesive and produce
242 extracellular traps upon recognition of bacterial antigens (Palić, Andreasen, Ostojić, Tell, & Roth,
243 2007). Overall, the relative abundance of professional phagocytes is 5-10% (Figure 3C), which is
244 in line with previous studies that estimated the number of professional phagocytes in WKM tissue
245 of adult zebrafish using classical morphological approaches (Wittamer et al., 2011). It is also
246 noteworthy that in line with other studies (Page et al., 2013), we did not observe a cluster of
247 lymphocytes (*e.g.*, B-cells) that actively phagocytize CTV-*S. aureus* bacteria (Figure 2, Data File
248 S3). Compared to the classical morphological approaches, Image3C allows to analyze thousands
249 of events in a high-throughput and unbiased fashion, allowing the study of rare cell morphology
250 and increasing results confidence and reproducibility. These results show that Image3C can

251 successfully analyze biological processes since we were able to recapitulate the presence, cell type
252 and frequency of professional phagocytes in adult zebrafish WKM organ.

253 A new aspect that Image3C highlighted is that CCB selectively affects cell viability based on
254 cell identity, introducing artifacts and cell damage, actions not specific to inhibition of
255 phagocytosis (Figure 3B). All mature erythrocyte containing clusters had a significantly higher
256 cell count in the CTV-*S. aureus* samples compared to the CTV-*S. aureus* + CCB ones (Figure 3B,
257 Table S2, Data File S1). Cluster analysis revealed that erythrocytes were almost absent in samples
258 incubated with CCB (Data File S1), while there was a significant increase in the relative abundance
259 of clusters containing dead and apoptotic cells (Figure 3B, Table S2). Both outcomes are likely
260 due to reduced cell viability of erythrocytes upon CCB incubation. Moreover, we excluded the
261 possibility of higher cell death in the professional phagocytes upon CCB incubation, since pseudo-
262 phagocytes (phagocytes with DHR response but no internalized CTV-*S. aureus*) were significantly
263 more abundant in the CTV-*S. aureus* + CCB sample (Figure 3B, Table S2). These results are
264 remarkable since Image3C allowed us to observe an effect of CCB on erythrocytes viability that,
265 as far as we know, was not described before.

266 Image 3C analysis also uncover another important biological observation. When we inhibited
267 phagocytosis by incubating the single cell suspension on ice (CTV-*S. aureus* + Ice) and compared
268 the specificity of inhibition with the CTV-*S. aureus* + CCB sample (Figure 3C, Table S3), we
269 discovered that the inhibition of phagocytosis through low temperature only affects adhesive
270 neutrophils (cluster *Dr*12_P) (Figure 3C). This is suspected to occur as ice inhibits adhesion, while
271 CCB effectively blocks phagocytosis in all professional phagocytes in zebrafish WKM tissue by
272 acting on the cytoskeleton. The use of Image3C allowed us to specifically identify cell types that
273 are sensible to low temperature and those that are not, confirming the existence of different

274 phagocytosis mechanisms and providing additional knowledge about pro and cons of different
275 protocols that can be applied to inhibit phagocytosis based on specific goals and needs.
276
277 *Image3C recapitulates cell composition of a freshwater snail hemolymph*
278 Since we aim to provide a tool that is widely applicable, we tested Image3C versatility on the
279 apple snail *P. canaliculata* an emerging organism for which molecular and cell biological tools
280 have yet to be fully developed. As such, we repeated the same experiments done in zebrafish on
281 the hemolymph of *P. canaliculata.* For morphological examination of the cellular composition of
282 the hemolymph collected from adults in homeostasis conditions, we stained the single cell
283 suspensions with Draq5 (DNA dye) and ran on the ImageStream®^{X} Mark II. We used Image3C to
284 analyze the images of the events and we identified 9 cell clusters (Figure 4A). Two of these clusters
285 were comprised of cell doublets, debris and dead cells (clusters *Pc*5 and *Pc*8) and the other clusters,
286 based on inspection of cell images, were grouped into 2 main categories (Figure 4A and Data File
287 S4). The first category includes small blast-like cells (cluster *Pc*4) and intermediate cells (clusters
288 *Pc*2 and *Pc*3) with high nuclear-cytoplasmic (N/C) ratio. These cells morphologically resemble
289 the Group I hemocytes previously described using a classical morphological approach (Accorsi,
290 Bucci, de Eguileor, Ottaviani, & Malagoli, 2013). The second category was comprised of larger
291 cells with lower N/C ratio and abundant membrane protrusions (clusters *Pc*1, *Pc*6, *Pc*7 and *Pc*9).
292 Likely, these cells correspond to the previously described Group II hemocytes that include both
293 granular and agranular cells (Accorsi et al., 2013).
294 To identify which of these clusters were enriched for granular cells, we looked at the heatmap
295 with feature values for each individual cluster (Figure 4B, Table S1a for feature description).
296 Cluster *Pc*6 had the highest values for the features related to cytoplasm texture and granularity

297 (*i.e.,* area granularity, intensity granularity and signal granularity) amongst all clusters other than
298 cell doublets (Figure 4B, Data File S1 and S4). Based on these data, we identified cluster *Pc*6 as
299 the one containing the granular hemocytes. The clusters obtained by Image3C were not only
300 homogeneous and biologically meaningful, but were also consistent with published *P. canaliculata*
301 hemocyte classification obtained by classical morphological methods (Accorsi et al., 2013). Such
302 remarkable consistency was observed both in terms of identified cell morphologies and their
303 relative abundance in the population of circulating hemocytes (Figure 4C, Data File S4). For
304 example, the relative abundance of the previously reported small blast-like cells is 14.0%, a value
305 almost identical to the abundance of the corresponding cluster *Pc*4 (13.8%).

306 Similarly, the category of larger hemocytes, or Group II hemocytes represents 80.4% of the
307 circulating cells as measured by traditional morphological methods, while clusters *Pc*1, *Pc*6, *Pc*7
308 and *Pc*9 combined represent 72.4% of the events analyzed with Image3C (Figure 4C, Data File
309 S1). A sub-set of these cells are the granular cells (cluster *Pc*6), which correspond to 7.7% of all
310 hemocytes by classical histological methods and 8.9% by Image3C. The intermediate cells
311 (clusters *Pc*2 and *Pc*3) are less well represented in both approaches, with a relative difference in
312 abundance of 5.6% versus 10.6% of the manually and Image3C analyzed events, respectively
313 (Figure 4C, Data File S1). This difference is likely best explained by the remarkable difference in
314 both the number of cells and the number of features that can be considered for analysis by
315 Image3C. Only a few hundred hemocytes were visually analyzed using traditional histological
316 methods based only on cell diameter and N/C ratio (Accorsi et al., 2013). In contrast, the automated
317 pipeline used by Image3C facilitated the analysis of 10,000 nucleated events for each sample and
318 considered 25 morphological features for each cell. The significantly higher number of
319 morphological features simultaneously considered by Image3C also explains the higher number

320 of clusters and improved resolution to distinguish cell types compared to the traditional methods.
321 Hence, Image3C, not only can properly analyze cells obtained from an emerging research
322 organism generating biologically meaningful and informative clusters but also represents an
323 unprecedented increase in the accuracy of cell type identification over traditional histological
324 methods, while also allowing high-throughput capability.
325
326 *Image3C identifies phagocytosis competent cells in the hemolymph of a freshwater snail*
327 As with zebrafish, we also performed a phagocytosis experiment on hemocytes from *P.*
328 *canaliculata*. Our goal was to test if it is possible with an emerging research organism to
329 successfully discover cell phenotypes and functions and to obtain information about specific
330 biological processes of interest by using Image3C to compare cell populations among treated and
331 control samples.
332 Here, we setup the phagocytosis assay incubating the cells with CTV-*S. aureus* and DHR at
333 room temperature. The phagocytosis was inhibited, as control, either adding EDTA (CTV-*S.*
334 *aureus* + EDTA) or using low temperature by incubating samples on ice (CTV-*S. aureus* + Ice).
335 Events collected on the ImageStream®^{X} Mark II were analyzed with Image3C and clustered in 20
336 distinct clusters using quantifications of morphological and fluorescent features (Table S1b for
337 feature description), including nuclear staining, phagocytized *S. aureus* and DHR positivity
338 (Figure 5A). We compared the phagocytosis permissive samples (CTV-*S. aureus*) with samples
339 where phagocytosis was inhibited by EDTA incubation or low temperature using the statistical
340 analysis included in Image3C based on a negative binomial regression model (Figure 5B, 5C and
341 S6, Table S4 and S5). The clusters with relative abundance significantly higher in the phagocytosis
342 samples (Figure 5B, Data File S1 and S5) and with high intensities of both DHR and bacteria

343 signals (Figure S7), are the two clusters that we identify as enriched with professional phagocyte
344 (cluster *Pc5*_P and *Pc17*_P) (Figure 5B and S7, Data File S5). The two clusters show a different
345 DHR signal intensity (ROS response) from one another upon bacteria exposure (cluster *Pc5*_P
346 with high DHR signal, cluster *Pc17*_P with low DHR signal) (Figure S7, Data File S1 and S5).
347 Both *Pc5*_P and *Pc17*_P relative abundance is significantly higher in the phagocytosis samples
348 compared to the EDTA treated sample (Figure 5C, Table S4), showing that EDTA inhibits
349 phagocytosis for both types of professional phagocytes. In the sample where the phagocytosis was
350 inhibited by low temperature, however, only cluster *Pc17*_P had a significantly lower relative
351 abundance compared to the phagocytosis sample (Figure 5C and S6, Table S5). We can conclude
352 that like CCB inhibition in the zebrafish phagocytosis experiment, EDTA is a more effective and
353 generalized (not cell-type-specific) inhibitor of phagocytosis than low temperature. These results
354 show that also in an emerging research organism, Image3C allowed discovery of new aspects of
355 this biological process and highlighted differences among professional phagocytes that would have
356 been difficult to detect with other methods.
357 The data analysis with Image3C clearly highlighted that CCB and EDTA, two classical
358 phagocytic inhibitors commonly used in controls for phagocytosis experiments in vertebrates and
359 invertebrates, respectively, result in a drastic change of cell morphology and cause cell death and
360 necrosis. This consequence is not easily detectable by other methods and is therefore often
361 overlooked. In the present work, these changes significantly modified the overall cell cluster
362 number and distribution and indicates that the effects of CCB and EDTA on cell morphology
363 should be taken into consideration in any study of morphological features of cells with
364 phagocytosis properties because artifacts might be significant.
365

366 *Convolutional Neural Network (CNN) allows unbiased comparison between experiments*

367 When determining differences between control and experimental treatments, Image3C
368 necessarily combines images and data from all samples and then clusters the cells. This must be
369 taken into consideration for experimental planning. Experiments meant to analyze cell
370 composition and morphological diversity in one biological domain (*e.g.*, homeostasis condition as
371 presented in Figure 2 and 4) should be carried out separately from those in other domains that are
372 likely to introduce changes in the cell population composition or cell morphologies that would be
373 a confounding factor for the *de novo* clustering in homeostasis condition. Image3C clustering
374 works best when used, at the same time, only on samples belonging to a single experimental
375 domain, such as homeostasis or the phagocytosis assay. An issue that emerges when analyzing
376 different experimental sets independently is the increase of time requirement for analytical steps,
377 the likelihood of introducing errors, and the need to repeatedly annotate the clusters in the FDL
378 graph obtained from each experimental set. This last element is required for comparing cell type
379 behaviors among multiple experiments and have a global understanding of their functions and
380 response to treatments (*i.e.*, cluster #1 from one analytical run cannot be expected to match cell
381 morphologies with cluster #1 from another run, since there is a stochastic element to the process).

382 This last point is probably the most challenging since mistakes can easily be introduced based
383 on user-biases or lack of sufficient pre-existing knowledge about cell morphologies or of cell
384 biomarkers that would allow a confident cross-annotation between multiple FDL graphs. In
385 addition, we observed that the number of clusters drastically increases when including treatments
386 that influence cell morphological properties of the cell. As an example, while we detected 9 unique
387 clusters in naïve hemolymph samples, we detected 20 clusters in the phagocytosis experiment
388 (Figure 3A and 4A). This is in part due to the fact that professional phagocytes change their

389 morphology upon detection of pathogens (Palić et al., 2007), thus creating new clusters. Similarly,
390 the complexity of the clustering is also increased by treatments, such as CCB and EDTA
391 incubations, that are necessary to ensure identification of professional phagocytes, but have a
392 strong impact on the morphology of the cells making the clustering and annotation steps more
393 challenging and prone to mistakes since treated samples contain aberrant populations not found at
394 homeostasis (Figure 5A, Data File S5).

395 To provide an alternative for streamlining the analysis of multiple experimental sets upon initial
396 *de novo* clustering and cell type identification in homeostasis samples, we included in Image3C
397 the possibility to use these initial images and their cluster IDs to train a CNN in an unsupervised
398 way (Figure 1). This trained classifier can then be used to assign the cell images subsequently
399 collected from additional experimental sets to one of the clusters defined and annotated in the
400 homeostasis condition in a high-throughput way. In this way it will be possible to determine the
401 behavior of a specific cell type through multiple experimental sets without re-clustering whenever
402 new data is acquired. A crucial element that allows this approach is also represented by the
403 ImageStream®^{X} Mark II system that provides highly reproducible and comparable images of cells
404 coming from different experiments and acquired in different days, introducing much less
405 variability then standard light or electron microscopy.

406 For our pipeline, then, a CNN (LeCun et al., 1989) based on the architecture of DenseNet
407 (Huang, Liu, Maaten, & Weinberger, 2017) was deployed to: 1. use, as training set, images and
408 clusters obtained from a first group of samples (*e.g.*, homeostasis conditions, naïve cells or WT
409 samples) analyzed in an unbiased way by *de novo* clustering and 2. assign new cell images acquired
410 through ImageStream®^{X} Mark II system to their respective classes in an unsupervised way. As
411 proof-of-concept, we used the clusters identified for *P. canaliculata* hemocytes in homeostasis

412 condition with the first part of the pipeline (Figure 4A) for training and setting up the CNN
413 classifier. This approach would define the classes based on the unbiased *de novo* clustering of
414 thousands of cells with no need for formal annotation or previous knowledge about cell types and
415 tissue composition. To prepare a data set for training the classifier, we first combined clusters that
416 strongly overlapped with one another in terms of morphological characteristics (*e.g.*, doublets and
417 dead cells) to increase accuracy of the classifier (Figure 6A). We used 80% of the cells obtained
418 in the original *P. canaliculata* dataset together with their cluster IDs to train the classifier through
419 over 25,000 iterations. After each iteration, we tested the training with 10% of the original dataset
420 and we determined the relative accuracy by scoring numbers of cells whose cluster ID assigned by
421 the classifier matched the original cluster ID (Figure 6B and 6C). The remaining 10% of the
422 original dataset was used to calculate the precision of the trained classifier. Clusters with higher
423 support numbers obtained higher precision scores. The weighted average precision score (f1-score,
424 precision average score across clusters controlling for support numbers) of 0.75 is relatively high
425 considering the complexity of the phenotype (BF, darkfield and Draq5 images) (Figure 6D) and
426 comparable to other studies using machine learning for cell classification (Blasi et al., 2016). The
427 true probability match for each individual cell (probability for each cell that the class assigned by
428 the classifier would match the original cluster ID) demonstrated that lower true probability matches
429 occurred where clusters strongly overlapped or where cell phenotypes are intermediate between
430 clusters, providing an additional layer of information about our dataset (Figure 6D).

431 To test the efficiency of this pipeline, we extracted all the images belonging to the two clusters
432 identified in the phagocytosis assay as cluster-containing phagocytes and determined to which
433 naïve cell-type they correspond using the CNN classifier and only the BF, SSC and Draq5 channels
434 (*i.e.*, DHR and labeled bacteria signals where not used). We found that 59.4%, 6.2% and 9.2% of

435 the phagocytes belonged to cluster *Pc*1_CT, *Pc*6_CT and *Pc*7_CT, respectively (Figure 6E), where
436 CT stands for Classifier Training. These results confirmed a previously published result that used
437 classical morphological staining and manual annotation to conclude that the hemocytes able to
438 phagocytize were primarily Group II hemocytes (Accorsi et al., 2013). Only 8% of the phagocytes
439 were clustered in the Group I hemocytes, here represented by cluster *Pc*2_CT, *Pc*3_CT and
440 *Pc*4_CT, while the remaining 17.2% were assigned by the CNN to the cluster *Pc*5_CT, constituted
441 by doublets and dead cells (Figure 6E). This result can be explained by the fact that *in vitro*
442 phagocytosis triggers microaggregate formation (hemocyte-hemocyte adhesion) in invertebrate
443 hemocytes that resembles the nodule formation observed *in vivo* (Walters, 1970). It is important
444 to observe how this analysis allowed us to assign phagocytes to cell types using the annotation
445 already performed in Figure 4A (*de novo* clustering of hemocytes in homeostasis condition)
446 without the need to re-annotate the FDL obtained during the phagocytosis assay (Figure 5A).
447 To test the adaptability of the trained CNN to new datasets, we collected hemocytes from male
448 apple snail specimens, we stained the cells with Draq5 and recorded BF, SSC and nuclei images
449 from 10,000 cells on the ImageStream®^{X} Mark II as previously described. We extracted the
450 images of the cells and we used our CNN classifier to determine the relative abundance of
451 hemocytes collected from male snails in the 7 classes of the classifier (Figure 6F). First of all, we
452 visually compared the female hemocytes clustered by the *de novo* clustering with the male
453 hemocytes that were run on the ImageStream®^{X} Mark II at a different time and were assigned to
454 a class by the classifier (Figure 6F). This comparison shows that the female and male hemocytes
455 belonging to the same cluster are morphologically extremely similar and different from the
456 hemocytes assigned to other clusters (Figure 6F). This shows that the CNN classifier can be trained
457 with a first group of samples and then it can successfully analyze new datasets acquired later on.

458 The comparison between female and male hemocyte compositions revealed that the only clusters
459 significantly different in terms of relative abundance are *Pc*2_CT and *Pc*3_CT (Group I
460 intermediate hemocytes) and *Pc*5_CT (Figure 6F). We found *Pc*5_CT consists of dead cells and
461 doublets and the difference in composition might be explained by variability in sample preparation
462 and data collection. Significantly, prior studies detected no differences between females and males
463 hemocytes composition through manual classification and counting using a classical
464 morphological approach (Accorsi et al., 2013). The unsupervised and high-throughput analysis
465 presented here, in contrast, allowed us to determine that both subpopulations of intermediate cells
466 as defined by Image3C were significantly less abundant in the male animals (*Pc*2_CT: 5% and 1%
467 in female and male, respectively; *Pc*3_CT: 6% and 1% in female and male, respectively) (Figure
468 6F). While the biological significance of this observation is not going to be further investigated in
469 this paper, the discovery highlights the power of Image3C analysis compared to traditional
470 methods for determining and quantitating the composition of cell populations. These experiments
471 demonstrate that Image3C, in combination with the presented convolutional classifier, is capable
472 of analyzing large experimental datasets and identifying significances with small effect sizes.
473 Importantly, Image3C analysis is independent of observer biases and does not require prior
474 knowledge about expected tissue composition or the expected effect of treatment on cell
475 morphology.

476

477 Conclusion

478 We have developed a powerful new method to analyze at single-cell resolution the composition
479 of any cell population obtained from research organisms for which species-specific reagents (such
480 as fluorescently tagged antibodies), biomarkers, single-cell atlases or a high quality genome for a

481 scRNA-Seq approach are not available. We demonstrated that Image3C can identify cell types
482 based on morphology and/or function through *de novo* clustering and highlight important changes
483 in cell type abundance and cell population composition caused by experimental or natural
484 perturbation (sex, treatment, experimental protocol). Furthermore, in combination with the CNN
485 classifier trained on these clusters, we demonstrate that Image3C is capable of unsupervised, bias-
486 free and high-throughput analysis of large experimental datasets making it possible to compare a
487 specific cell type behavior or population composition across multiple experiments. Image3C is
488 extremely versatile and can be applied to any tissue or cell population of interest and is adaptable
489 to a variety of experimental designs. In addition, given the recent advancement in image-based
490 flow cytometry that enables image capturing together with cell sorting (Nitta et al., 2018), a
491 scRNA-Seq approach in combination with the Image3C pipeline would enable simultaneous
492 analysis of both the morphological/phenotypic and genetic properties of a cell population at single
493 cell resolution.
494

495 **Materials and Methods**

496 *Collection of zebrafish whole kidney marrow (WKM)*

497 Twelve-month-old, wild type, female, adult zebrafish were euthanized with cold 500 mg/L MS-
498 222 solution for 5 min. WKM was dissected as previously described (D. Traver et al., 2003) and
499 transferred to 40 µm cell strainer with 1 mL of L-15 media containing 10% water, 10 mM HEPES
500 and 20 U/mL Heparin (L-90). Cells were gently forced through the cell strainer with the plunger
501 of a 3 mL disposable syringe. The strainer was washed once with 1 mL of L-90 and the resulting
502 single cell suspension was centrifuged at 500 rcf at 4 ºC for 5 min. The supernatant was discarded,
503 and the cells were resuspended in 1 mL of L-90 containing 5% fetal calf serum (FCS), 4 mM L-
504 Glutamine, and 10,000 U of both Penicillin and Streptomycin (L-90 media). The cells were
505 counted in a 1:20 dilution on the EC-800 flow cytometer (Sony) using scatter properties.

506

507 *Collection of apple snail hemocytes*

508 Specimens of the apple snail *Pomacea canaliculata* (Mollusca, Gastropoda, Ampullariidae)
509 were maintained and bred in captivity, in a water recirculation system filled with artificial
510 freshwater (2.7 mM $CaCl_2$, 0.8 mM $MgSO_4$, 1.8 mM $NaHCO_3$, 1:5000 Remineralize Balanced
511 Minerals in Liquid Form [Brightwell Aquatics]). The snails were fed twice a week and kept in a
512 10:14 light:dark cycle. Wild type adult snails, 7-9 months old and with a shell size of 45-60 mm
513 were starved for 5 days before the hemolymph collection (Accorsi et al., 2013). If not differently
514 specified, female snails were used for the experiments. The withdrawal was performed applying a
515 pressure on the operculum and dropping the hemolymph directly into an ice-cold tube. The
516 hemolymph collected from different animals was not pooled together. The hemolymph was
517 immediately diluted 1:4 in Bge medium + 10% fetal bovine serum (FBS) and then centrifuged at

518 500 rcf for 5 min. The pellet of cells was resuspended in 100 µl of Bge medium + 10% FBS. The
519 Bge medium (also known as *Biomphalaria glabrata* embryonic cell line medium) is constituted
520 by 22% (v/v) Schneiders's Drosophila Medium, 4.5 g/L Lactalbumin hydrolysate, 1.3 g/L
521 Galactose, 0.02 g/L Gentamycin in MilliQ water, pH 7.0.
522
523 *Experiment 1: Morphology Assay in homeostasis conditions*
524 WKM cells from zebrafish were isolated as described before and plated at 4 x 10^5 cells/well in
525 a 96-well plate in 200 µL of L-90 media and incubated for 3 h at room temperature. Cells were
526 stained with 5 µM Draq5 (Thermo Fisher Scientific) for 10 min and subsequently run on the
527 ImageStream®^{X} Mark II (Amnis Millipore Sigma), where 10,000 nucleated and focused events
528 were recorded for each sample (n=8). Erythrocytes were out-gated to enrich for immune relevant
529 cells and to prevent over-clustering in the subsequent analysis. The latter is due to the fact that fish
530 erythrocytes are nucleated and their biconcave shape results in different morphological feature
531 values only depending on their orientation during image acquisition.
532 The *P. canaliculata* hemocytes were stained with 5 µM Draq5 (Thermo Fisher Scientific) for
533 10 min, moved to ice and subsequently run on the ImageStream®^{X} Mark II (Amnis Millipore
534 Sigma), where 10,000 nucleated and focused events were imaged for each sample (n=5).
535
536 *Experiment 2: Phagocytosis Assay*
537 *Staphylococcus aureus* (Thermo Fisher Scientific) were resuspended in PBS at the final
538 concentration of 100 mg/ml and incubated with 5 uM CellTrace Violet (CTV; Thermo Fisher
539 Scientific) for 20 min. Labelled bacteria were centrifuged and resuspended in PBS for 3 times to
540 remove unbound dye and then stored at -20 ºC as single-use aliquots. Cells, obtained from fish

541 WKM or snail hemolymph and in a single cell suspension, were plated in a 96-well plate at a
542 concentration of 4 x 10^5 cells/well in 200 µL of medium and incubated with 2 x 10^7 CTV-coupled
543 *S. aureus*/well for 3 h at room temperature. As control, the phagocytosis was inhibited incubating
544 the cells + CTV-*S. aureus* mix either on ice (for both species) or with 0.08 mg/mL cytochalasin B
545 (CCB) for zebrafish cells or with 30 mM EDTA and 10 mM HEPES for apple snail cells (Cueto,
546 Rodriguez, Vega, & Castro-Vazquez, 2015; Li et al., 2006). After 2 h and 30 min we added 5 µM
547 dihydrorhodamine-123 (DHR) (Thermo Fisher Scientific) to the cell suspension to stain cells
548 positive for reactive oxygen species (ROS) production. To control for this treatment with DHR,
549 we incubated one aliquot of cells with 10 ng/mL phorbol 12-myristate 13-acetate (PMA) to induce
550 ROS production. At 2 h and 50 min since the beginning of incubation with CTV-*S. aureus*, all the
551 samples were stained with 5 µM Draq5 for 10 min. After 3 h incubation with bacteria, cells were
552 moved and stored on ice and subsequently run on the ImageStream®^{X} Mark II (Amnis Millipore
553 Sigma), where 10,000 nucleated and focused events were imaged for each sample (at least n=4
554 snail and n=6 fish) at a speed of 1,000 images/sec.
555
556 *Data collection on ImageStream®^{X} Mark II*
557 Following cell preparation, data were acquired from each sample on the ImageStream®^{X} Mark
558 II (Amnis Millipore Sigma) at 60x magnification, slow/sensitive flow speed (1,000 images/sec),
559 using 633, 488 and 405 nm laser excitation. Bright field was acquired on channels 1 and 9, DHR
560 (488 nm excitation) on channel 2, CTV-*S. aureus* (405 nm excitation) on channel 7, Draq5 (633
561 nm excitation) on channel 11, and SSC was acquired on channel 6. Single color controls were also
562 acquired for each fluorescent channel to allow for fluorescence spillover correction.
563

564 *Data analysis and* de novo *cluster identification*

565 A schematic representation of the pipeline with software used, format of the exported files and
566 approximation of time required for running the individual steps is provided in Figure S1. Raw
567 images from the ImageStream®^{X} Mark II system (RIF files, a type of modified 16 bit TIFF file)
568 were compensated (spillover and other corrections applied), background was subtracted, and
569 features were calculated using IDEAS 6.2 software (Amnis Millipore, free for download once an
570 Amnis user account is created). The resulting compensated image files (CIF files) were used to
571 quantify features for all cells and samples. Table S1a and S1b report the list of features used for
572 each organism and for each experiment and their description. These per-object feature matrices
573 (DAF files) were then exported from IDEAS into FCS files. Exported FCS files were processed in
574 R (R Core Team, 2014). In order to trim redundant features that contribute noise but little new
575 information, Spearman's correlation values for each pair of features were calculated using all
576 events of a representative sample and one of the features of the pair was trimmed when correlation
577 was ≥ 0.85 for the pair (Figure S2) (Caicedo et al., 2017). The Spearman's correlation of the mean
578 values of remaining features per each sample were then used to identify outliers among sample
579 replicates. Samples with correlation of mean feature values below 0.85 with the set were discarded
580 (Figure S3). Also, while morphological features did not require any transformation or
581 normalization, fluorescence intensity features were transformed using the estimateLogicle() and
582 transform() functions from the R flowCore package (Ellis et al., 2018; Hahne et al., 2009) to
583 improve homoscedasticity (homogeneity of variance) of distributions. DNA intensity features
584 were also normalized to align all 2N and 4N peak positions and to remove intensity drift between
585 samples (Figure S4) using the gaussNorm() function from flowStats package (Hahne,
586 Gopalakrishnan, Khodabakhshi, Wong, & Lee, 2018). The processed data was exported from R

587 (R Core Team, 2014) using writeflowSet() function in flowCore package (Ellis et al., 2018; Hahne
588 et al., 2009) as CSV or FCS files, depending on downstream needs for the file output.
589 These processed data files were then imported into the VorteX clustering environment for X-
590 Shift k-nearest-neighbor clustering (free to install) (Samusik et al., 2016). During the import into
591 VorteX, all features were scaled to 1SD to equalize the contribution of features towards clustering.
592 Clustering was performed in VorteX testing a range of k values (typically from 5 to 150), choosing
593 a final k value using the 'find elbow point for cluster number' function in VorteX and confirming
594 visually that over- or under-clustering did not occur. Force directed layout (FDL) graphs of a subset
595 of cells obtained from each set of samples were also generated in VorteX and cell coordinates in
596 the resultant 2D space were exported along with graphML representation of the FDL graph.
597 Finally, tabular data (CSV files) was exported from VorteX including a master table of every cell
598 event with its cluster assignment and original sample ID, as well as a table of the average feature
599 values for each cluster and counts of cells per cluster and per sample.
600 Clustering results were further analyzed and plotted in R (R Core Team, 2014) by merging all
601 cell events and feature values with cluster assignments and X/Y coordinates for FDL graph. Using
602 this merged data and the graphML file exported from VorteX, new FDL graphs were created for
603 each treatment condition using the igraph package (Csardi & Nepusz, 2006) in R (R Core Team,
604 2014). Statistical analysis of differences in cell counts per cluster by condition were performed
605 using negative binomial regression of cell counts per cluster, plots of statistic results and other
606 results were generated, and CSV files containing cell ID, sample ID, feature values, X/Y
607 coordinates in FDL graph were exported for each sample. The subsequent use of FCS Express Plus
608 version 6 (DeNovo software, free alternative are mentioned later in the text) allowed visualization
609 of cell images using DAF/CIF files by cluster and by customized subsets of the FDL graphs.

610 DAF files were opened in FCS Express Plus software and the "R add parameters"
611 transformation feature with a custom script was used to merge the clustering data saved in the CSV
612 files generated above with both DAF and CIF files (feature values and image sets, respectively).
613 This allowed to visualize image galleries of cells within each cluster and gate by features of interest
614 on 2D plots (more traditional flow cytometry analysis) for exploring the clustering results and
615 identifying clusters and populations of interest. FCS Express Plus is a proprietary software, but
616 similar results can be obtained with IDEAS software where a text file with cluster IDs for each
617 image event can be imported and the cluster information can be matched to the event images.
618 The full complement of R packages used includes flowCore (Ellis et al., 2018; Hahne et al.,
619 2009), flowStats (Hahne et al., 2018), igraph (Csardi & Nepusz, 2006), ggcyto (Jiang, 2015),
620 ggridges (Wilke, 2018), ggplot2 (Wickham, 2016), stringr (Wickham, 2010), hmisc (Harrell &
621 Dupont, 2019) and caret (Kuhn, 2008). The GitHub repository at
622 https://github.com/stowersinstitute/LIBPB-1390-Image3C reports a detailed description of all
623 these processing steps and includes sample scripts and workflow files.
624
625 *Setup and use of a Convolutional Neural Network (CNN) Classifier*
626 Once the clusters were defined with the previously described *de novo* clustering analysis, we
627 used a CNN (LeCun et al., 1989) based on the architecture of DenseNet (Huang et al., 2017) for
628 image classification. Out of all the events captured with the ImageStream®^{X} Mark II system, we
629 selected only single nucleated objects applying gates on area vs aspect ratio plot and Draq5
630 intensity plot to achieve this selection, respectively. For these objects, we exported 16 bit TIFF
631 images (one channel per fluorescence/BrightField image 'color') using IDEAS 6.2 software
632 (Amnis/Millipore, free for download once an Amnis user account is created).

633 Because images from the ImageStream®^{X} Mark II system have non-uniform sizes, each image
634 was cropped or padded to 32x32 pixels using NumPy indexing (Walt, Colbert, & Varoquaux,
635 2011) in a Python script. The CNN consists of three dense blocks that transition from three-channel
636 image input of 32x32x3 to a final size of 4x4x87 with 87 feature maps. A dense block includes
637 three convolution layers, each followed by leaky ReLU activation. The last step of the block is a
638 strided convolution used to down-sample the width and height of the feature maps by a factor of
639 2. The final layer of the CNN flattens the 4x4x87 array into a 1D vector of length 1392 and is fully
640 connected to the output layer that is a 1D vector with a length of the number of classes for
641 prediction. The CNN used softmax cross entropy for the loss function with L2 regularization, and
642 the Adam optimizer (Kingma & Ba, 2015) was used to minimize the loss function. The CNN was
643 implemented in Python using the TensorFlow platform (Abadi et al., 2015) and the SciPy
644 ecosystem (T. Oliphant, 2006; T. E. Oliphant, 2007; Pedregosa et al., 2011).
645 The CNN was used to train a classifier using over 35,000 images of *P. canaliculata* hemolymph
646 cell types in homeostasis condition acquired with the ImageStream®^{X} Mark II system. The event
647 images were randomly split in 80% for training, 10% for testing during training, and 10% for final
648 validation. The first 80% of the images were used together with their cluster IDs obtained by the
649 *de novo* clustering to train the classifier. The learning rate for the Adam optimizer was set to 0.0006
650 with a decaying learning rate starting at 0.001 and decreasing by 1% each step. The training
651 proceeded for 25,000 iterations with a batch size of 256 randomly selected images for each
652 iteration. After each iteration, 10% of the cells of the original *P. canaliculata* dataset was used to
653 test the classifier. The loss and accuracy of the CNN were recorded after every 100 iterations to
654 monitor the performance. The CNN loss was defined by the softmax of the cross entropy (Dahal,
655 2017) between the final output and the one-hot-encoded image labels. To avoid the CNN

656 memorizing the training set, L2 regularization was applied to the weights. The training and test set
657 follow the same accuracy and loss trends over all iterations indicating the training set is not
658 memorized and can generalize to predict the test set.

659 The finally trained classifier was tested on the remaining 10% of images that were completely
660 new for the CNN. The trained model was saved for future use, so new images can be inferred by
661 the network to predict cell types. The inference is very fast because only one forward pass is made
662 through the network and no back propagation occurs. The result of the inference is a vector with
663 length equal to the number of cell type classes. Each element of the vector will be the probability
664 of the cell belonging to the corresponding class; the sum of the vector must be 1. Inferring a
665 complete experiment will provide a probability vector for each image; the list of probability
666 vectors can be saved as CSV file. For new images, the inference results will need to be examined
667 to ensure the predictions are reliable. A large majority of the probability vectors should have a
668 maximum greater than 0.5, and a subset of the images should be visually inspected to verify proper
669 class assignment. The GitHub repository at https://github.com/stowersinstitute/LIBPB-1390-
670 Image3C reports a detailed description of all these processing steps.

671

672 *Statistical Analysis*

673 Negative binomial regression was performed on tables of cell counts per cluster and per sample
674 and plots were generated using R (R Core Team, 2014) with the edgeR package (Robinson,
675 McCarthy, & Smyth, 2010), which was developed for RNAseq analysis, but includes generally
676 applicable and user-friendly wrappers for regression and modeling analysis and plotting of results.
677 For the comparison of cellular hemolymph composition between females and males of *P.*
678 *canaliculata* a one-way ANOVA with subsequent FDR (Benjamini-Hochberg) was used.

679
680 *Animal experiment statement*

681 Research and animal care were approved by the Institutional Animal Care and Use Committee

682 (IACUC) of the Stowers Institute for Medical Research.

683

684 Acknowledgements

685 We acknowledge Hua Li for her assistance on the statistical analysis and we also thank the
686 Laboratory Animal Services and the Aquatics Facility at the Stowers Institute for Medical
687 Research for animal husbandry. We would like to thank Blair Benham-Pyle, Carolyn Brewster and
688 Viraj Doddihal as well as Barbara Milutinović for their critical comments on an earlier version of
689 this manuscript. This work was supported by institutional funding to ACB, CW, ASA and NR.
690 ASA is a Howard Hughes Medical Institute Investigator. NR is further supported by the Edward
691 Mallinckrodt Foundation, NIH Grant R01 GM127872, DP2DP2AG071466 and NSF IOS-
692 1933428 and EDGE award 1923372. RP was supported by a grant from the Deutsche
693 Forschungsgemeinschaft (PE 2807/1-1). AA was supported by the Emerging Models grant from
694 the Society for Developmental Biology (SDB) and the postdoctoral fellowship from the American
695 Association of Anatomists (AAA).

696

697 Author Contributions

698 RP, ACB and AA conceived and designed the study with input from ASA and NR. RP performed
699 *D. rerio* experiments. AA performed *P. canaliculata* experiments. ACB conceived and wrote the
700 Image3C pipeline and associated R-scripts. CW designed and optimized the convolutional neural
701 network. RP, ACB, AA and CW analyzed and interpreted the data. RP, ACB and AA wrote the
702 paper. All authors read and edited the paper.

703

704 Data availability statement

705 All original data underlying this manuscript can be accessed from the Stowers Original Data
706 Repository at http://www.stowers.org/research/publications/libpb-1390. Image3C code and

707 description is freely available at the GitHub repository https://github.com/stowersinstitute/LIBPB-
708 1390-Image3C.
709
710 Conflict of interest
711 The authors declare no conflict of interest.
712
713 References

714 Accorsi, A., Bucci, L., de Eguileor, M., Ottaviani, E., & Malagoli, D. (2013). Comparative analysis
715 of circulating hemocytes of the freshwater snail Pomacea canaliculata. *Fish Shellfish*
716 *Immunol, 34*(5), 1260-1268. doi:10.1016/j.fsi.2013.02.008
717 Baron, C. S., Barve, A., Muraro, M. J., van der Linden, R., Dharmadhikari, G., Lyubimova, A., . .
718 . van Oudenaarden, A. (2019). Cell Type Purification by Single-Cell Transcriptome-
719 Trained Sorting. *Cell, 179*(2), 527-542 e519. doi:10.1016/j.cell.2019.08.006
720 Blasi, T., Hennig, H., Summers, H. D., Theis, F. J., Cerveira, J., Patterson, J. O., . . . Rees, P.
721 (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nat*
722 *Commun, 7*, 10256. doi:10.1038/ncomms10256
723 Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., . . . Carpenter, A. E.
724 (2017). Data-analysis strategies for image-based cell profiling. *Nat Methods, 14*(9), 849-
725 863. doi:10.1038/nmeth.4397
726 Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research.
727 *InterJournal, Complex Systems, 1695*(5), 1-9.
728 Cueto, J. A., Rodriguez, C., Vega, I. A., & Castro-Vazquez, A. (2015). Immune Defenses of the
729 Invasive Apple Snail Pomacea canaliculata (Caenogastropoda, Ampullariidae):
730 Phagocytic Hemocytes in the Circulation and the Kidney. *PLoS One, 10*(4), e0123964.
731 doi:10.1371/journal.pone.0123964
732 Ellis, B., Haaland, P., Hahne, F., Le Meur, N., Gopalakrishnan, N., Spidlen, J., & Jiang, M. (2018).
733 flowCore: flowCore: Basic structures for flow cytometry data. R package version 1.46.1.
734 Eulenberg, P., Kohler, N., Blasi, T., Filby, A., Carpenter, A. E., Rees, P., . . . Wolf, F. A. (2017).
735 Reconstructing cell cycle and disease progression using deep learning. *Nat Commun,*
736 *8*(1), 463. doi:10.1038/s41467-017-00623-3
737 Goolam, M., Scialdone, A., Graham, S. J. L., Macaulay, I. C., Jedrusik, A., Hupalowska, A., . . .
738 Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in
739 4-Cell Mouse Embryos. *Cell, 165*(1), 61-74. doi:10.1016/j.cell.2016.01.047
740 Hahne, F., Gopalakrishnan, N., Khodabakhshi, A. H., Wong, C., & Lee, K. (2018). flowStats:
741 Statistical methods for the analysis of flow cytometry data. R package version 3.38.0.
742 Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., . . . Gentleman, R.
743 (2009). flowCore: a Bioconductor package for high throughput flow cytometry. *BMC*
744 *Bioinformatics, 10*, 106. doi:10.1186/1471-2105-10-106
745 Harrell, F. E., & Dupont, C. (2019). Hmisc: Harrell Miscellaneous. Retrieved from
746 http://biostat.mc.vanderbilt.edu/Hmisc
747 Huang, G., Liu, Z., Maaten, L. v. d., & Weinberger, K. Q. (2017, 21-26 July 2017). *Densely*
748 *Connected Convolutional Networks.* Paper presented at the 2017 IEEE Conference on
749 Computer Vision and Pattern Recognition (CVPR).

Jiang, M. (2015). ggcyto: Visualize Cytometry data with ggplot. R package version 1.8.0.

Kaczorowski, K. J., Shekhar, K., Nkulikiyimfura, D., Dekker, C. L., Maecker, H., Davis, M. M., . . . Brodin, P. (2017). Continuous immunotypes describe human immune variation and predict diverse responses. *Proc Natl Acad Sci U S A, 114*(30), E6097-E6106. doi:10.1073/pnas.1705065114

Kimmel, J. C., Penland, L., Rubinstein, N. D., Hendrickson, D. G., Kelley, D. R., & Rosenthal, A. Z. (2019). Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res, 29*(12), 2088-2103. doi:10.1101/gr.253880.119

Kobayashi, H., Lei, C., Wu, Y., Mao, A., Jiang, Y., Guo, B., . . . Goda, K. (2017). Label-free detection of cellular drug responses by high-throughput bright-field imaging and machine learning. *Sci Rep, 7*(1), 12454. doi:10.1038/s41598-017-12378-4

Kuhn, M. (2008). Building Predictive Models inRUsing thecaretPackage. *Journal of Statistical Software, 28*(5). doi:10.18637/jss.v028.i05

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation, 1*(4), 541-551. doi:10.1162/neco.1989.1.4.541

Lei, C., Kobayashi, H., Wu, Y., Li, M., Isozaki, A., Yasumoto, A., . . . Goda, K. (2018). High-throughput imaging flow cytometry by optofluidic time-stretch microscopy. *Nat Protoc, 13*(7), 1603-1631. doi:10.1038/s41596-018-0008-7

Li, J., Barreda, D. R., Zhang, Y.-A., Boshra, H., Gelman, A. E., LaPatra, S., . . . Sunyer, J. O. (2006). B lymphocytes from early vertebrates have potent phagocytic and microbicidal abilities. *Nat Immunol, 7*(10), 1116-1124. doi:10.1038/ni1389

Lippeveld, M., Knill, C., Ladlow, E., Fuller, A., Michaelis, L. J., Saeys, Y., . . . Peralta, D. (2020). Classification of Human White Blood Cells Using Machine Learning for Stain-Free Imaging Flow Cytometry. *Cytometry A, 97*(3), 308-319. doi:10.1002/cyto.a.23920

Lugo-Villarino, G., Balla, K. M., Stachura, D. L., Banuelos, K., Werneck, M. B. F., & Traver, D. (2010). Identification of dendritic antigen-presenting cells in the zebrafish. *Proceedings of the National Academy of Sciences, 107*(36), 15850-15855. doi:10.1073/pnas.1000494107

Maizels, R. M., & Nussey, D. H. (2013). Into the wild: digging at immunology's evolutionary roots. *Nat Immunol, 14*(9), 879-883. doi:10.1038/ni.2643

Nassar, M., Doan, M., Filby, A., Wolkenhauer, O., Fogg, D. K., Piasecka, J., . . . Hennig, H. (2019). Label-Free Identification of White Blood Cells Using Machine Learning. *Cytometry A, 95*(8), 836-842. doi:10.1002/cyto.a.23794

Nitta, N., Sugimura, T., Isozaki, A., Mikami, H., Hiraki, K., Sakuma, S., . . . Goda, K. (2018). Intelligent Image-Activated Cell Sorting. *Cell, 175*(1), 266-276 e213. doi:10.1016/j.cell.2018.08.028

Page, D. M., Wittamer, V., Bertrand, J. Y., Lewis, K. L., Pratt, D. N., Delgado, N., . . . Traver, D. (2013). An evolutionarily conserved program of B-cell development and activation in zebrafish. *Blood, 122*(8), 1-12. doi:10.1182/blood-2012-12-471029

Palić, D., Andreasen, C. B., Ostojić, J., Tell, R. M., & Roth, J. A. (2007). Zebrafish (Danio rerio) whole kidney assays to measure neutrophil extracellular trap release and degranulation of primary granules. *J Immunol Methods, 319*(1-2), 87-97. doi:10.1016/j.jim.2006.11.003

Pepe-Mooney, B. J., Dill, M. T., Alemany, A., Ordovas-Montanes, J., Matsushita, Y., Rao, A., . . . Camargo, F. D. (2019). Single-Cell Analysis of the Liver Epithelium Reveals Dynamic Heterogeneity and an Essential Role for YAP in Homeostasis and Regeneration. *Cell Stem Cell, 25*(1), 23-38 e28. doi:10.1016/j.stem.2019.04.004

Peuß, R., Box, A. C., Chen, S., Wang, Y., Tsuchiya, D., Persons, J. L., . . . Rohner, N. (2020). Adaptation to low parasite abundance affects immune investment and immunopathological responses of cavefish. *Nat Ecol Evol*. doi:10.1038/s41559-020-1234-2

Philippeos, C., Telerman, S. B., Oules, B., Pisco, A. O., Shaw, T. J., Elgueta, R., . . . Watt, F. M. (2018). Spatial and Single-Cell Transcriptional Profiling Identifies Functionally Distinct Human Dermal Fibroblast Subpopulations. *J Invest Dermatol, 138*(4), 811-825. doi:10.1016/j.jid.2018.01.016

R Core Team. (2014). R: A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Rabinovitch, M. (1995). Professional and non-professional phagocytes: an introduction. *Trends in Cell Biology, 5*(3), 85-87. doi:10.1016/s0962-8924(00)88955-2

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L., & Nolan, G. P. (2016). Automated mapping of phenotype space with single-cell data. *Nat Methods, 13*(6), 493-496. doi:10.1038/nmeth.3863

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., 2nd, Treacy, D., Trombetta, J. J., . . . Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science, 352*(6282), 189-196. doi:10.1126/science.aad0501

Traver, D., Paw, B. H., Poss, K. D., Penberthy, W. T., Lin, S., & Zon, L. I. (2003). Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nat Immunol, 4*(12), 1238-1246. doi:10.1038/ni1007

Traver, D., Paw, B. H., Poss, K. D., Penberthy, W. T., Lin, S., & Zon, L. I. (2003). Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nat Immunol, 4*(12), 1238-1246. doi:10.1038/ni1007

van der Meer, W., Scott, C. S., & de Keijzer, M. H. (2004). Automated flagging influences the inconsistency and bias of band cell and atypical lymphocyte morphological differentials. *Clin Chem Lab Med, 42*(4), 371-377. doi:10.1515/CCLM.2004.066

Walt, S. v. d., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering, 13*(2), 22-30. doi:10.1109/MCSE.2011.37

Walters, D. R. (1970). Hemocytes of saturniid silkworms: Their behaviorin vivo andin vitro in response to diapause, development, and injury. *Journal of Experimental Zoology, 174*(4), 441-450. doi:10.1002/jez.1401740407

Wickham, H. (2010). stringr: modern, consistent string processing. *R Journal, 2*(2), 38-40.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* : Springer-Verlag New York.

Wilke, C. O. (2018). ggridges, Ridgeline Plots in 'ggplot2' (Version 0.5.1). Retrieved from https://github.com/clauswilke/ggridges

Wittamer, V., Bertrand, J. Y., Gutschow, P. W., & Traver, D. (2011). Characterization of the mononuclear phagocyte system in zebrafish. *Blood, 117*(26), 7126-7135. doi:10.1182/blood-2010-11-321448

Yakimov, B. P., Gogoleva, M. A., Semenov, A. N., Rodionov, S. A., Novoselova, M. V., Gayer, A. V., . . . Shirshin, E. A. (2019). Label-free characterization of white blood cells using fluorescence lifetime imaging and flow-cytometry: molecular heterogeneity and erythrophagocytosis [Invited]. *Biomed Opt Express, 10*(8), 4220-4236. doi:10.1364/BOE.10.004220

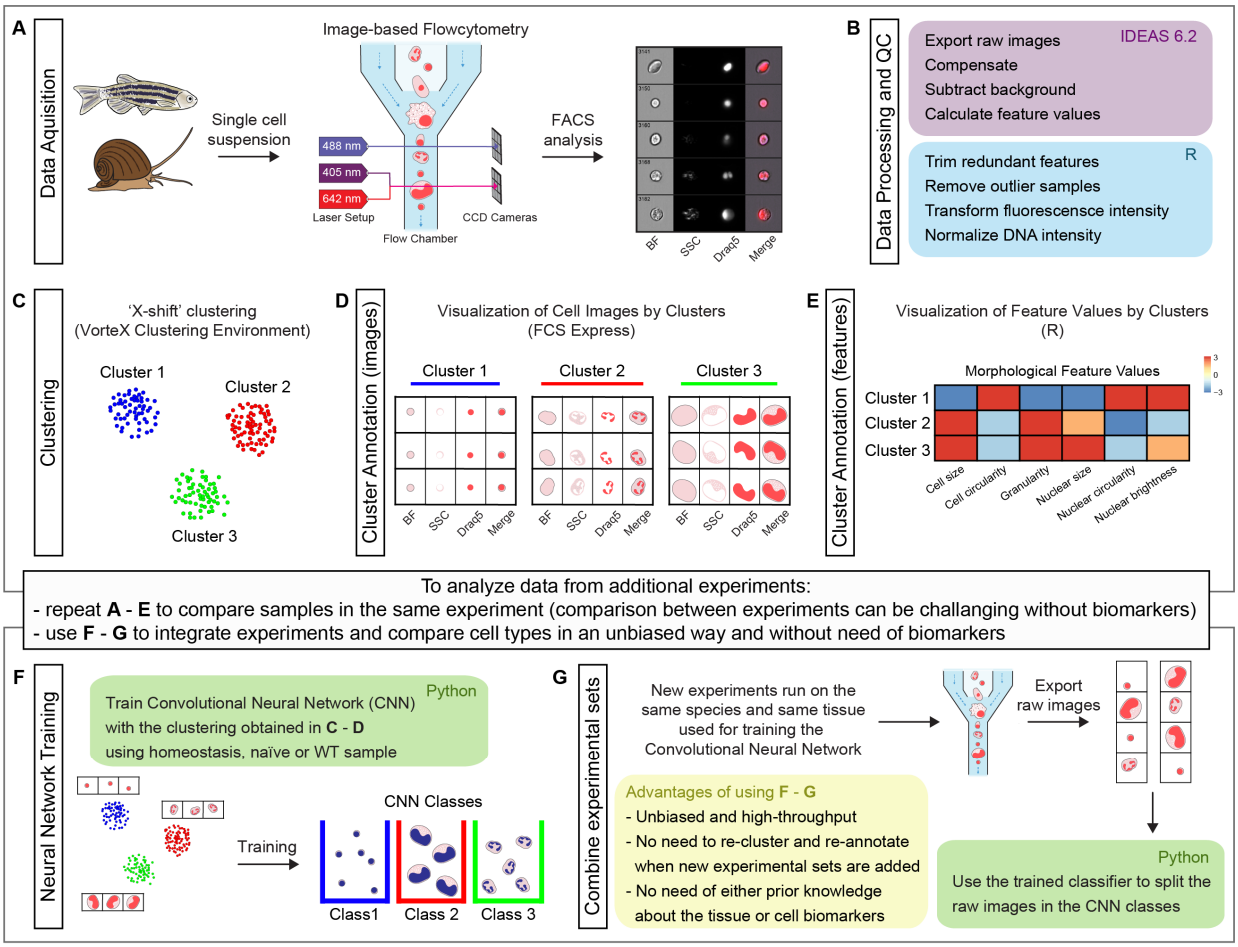


Figure 1: Schematic representation of Image3C workflow, a method for cell clustering based on morphological features

(A) A single cell suspension is prepared for image-based flow cytometric analyses. The cells can be labelled with any reagent working for the species of interest. The signal can highlight specific cell components (*e.g.,* nuclei), metabolic cell states or specific cell functions. The samples are run on the ImageStream®X Mark II and 10,000 nucleated and focused events are saved for each sample as individual raw images.

(B) IDEAS software (Amnis Millipore) is used to open the raw images, compensate for correcting fluorescent spillover, subtract background and quantify values for intrinsic morphological and fluorescent features. R (or R studio) is used to calculate the correlation between features to allow to trim the features that are redundant with others. Samples that are outliers among replicates are also removed prior the final normalization of the fluorescence intensities.

(C) Images are clustered based on morphological and fluorescent feature values and visualized as a Force Directed Layout (FDL) graph where each dot represents one event.

(D) R integration in FCS Express software allows the visualization of cell images by clusters or specifically selected with a gate. This step allows to evaluate the morphological homogeneity of the clusters, determine if the number of clusters is appropriate and explore the phenotype/function of the cells based on visualization of individual channels.

(E) Spearman's correlation plot of feature values by clusters is one of the options available in Image3C for plotting integrated data. This heatmap shows feature similarities and differences between cells belonging to different clusters.

(F) If new experiments are run and new data needs to be analyzed two approaches can be taken. 1. If the goal is comparing samples belonging to the same experiment (*e.g.*, treatment vs its control) the steps described so far from (A) to (E) can be re-applied to the new dataset including a statistical analysis to compare cluster relative abundance. This approach will produce a new set of clusters that will need to be

re-annotated. Compare sets of clusters coming from multiple experiments and multiple rounds of analysis can be challenging without pre-existing knowledge of cell-type, clearly different morphologies or biomarkers that would allow to establish a unique correlation between clusters coming from different FDL graphs. 2. If the goal is integrating experiments and comparing cell type abundance between them, the use of steps (F) and (G) is suggested. A CNN classifier is trained using the images obtained from homeostasis, naïve or WT cells and already organized in clusters in an unbiased way through the first part of our method. This will generate a trained classifier with CNN classes based on FDL clusters.
(G) This classifier is then used for deconvoluting data from new experimental sets and assigning each event to a CNN class with a given probability. This provides high-throughput, unsupervised and unbiased way to compare different experiment sets without the requirement for pre-existing knowledge about the tissue cell types, cell biomarkers or the need to cross-annotate cluster increasing the probability to introduce errors.
The entire pipeline chart and step-by-step technical information, such as software used, time required for processing and exported file format are reported in Figure S1.

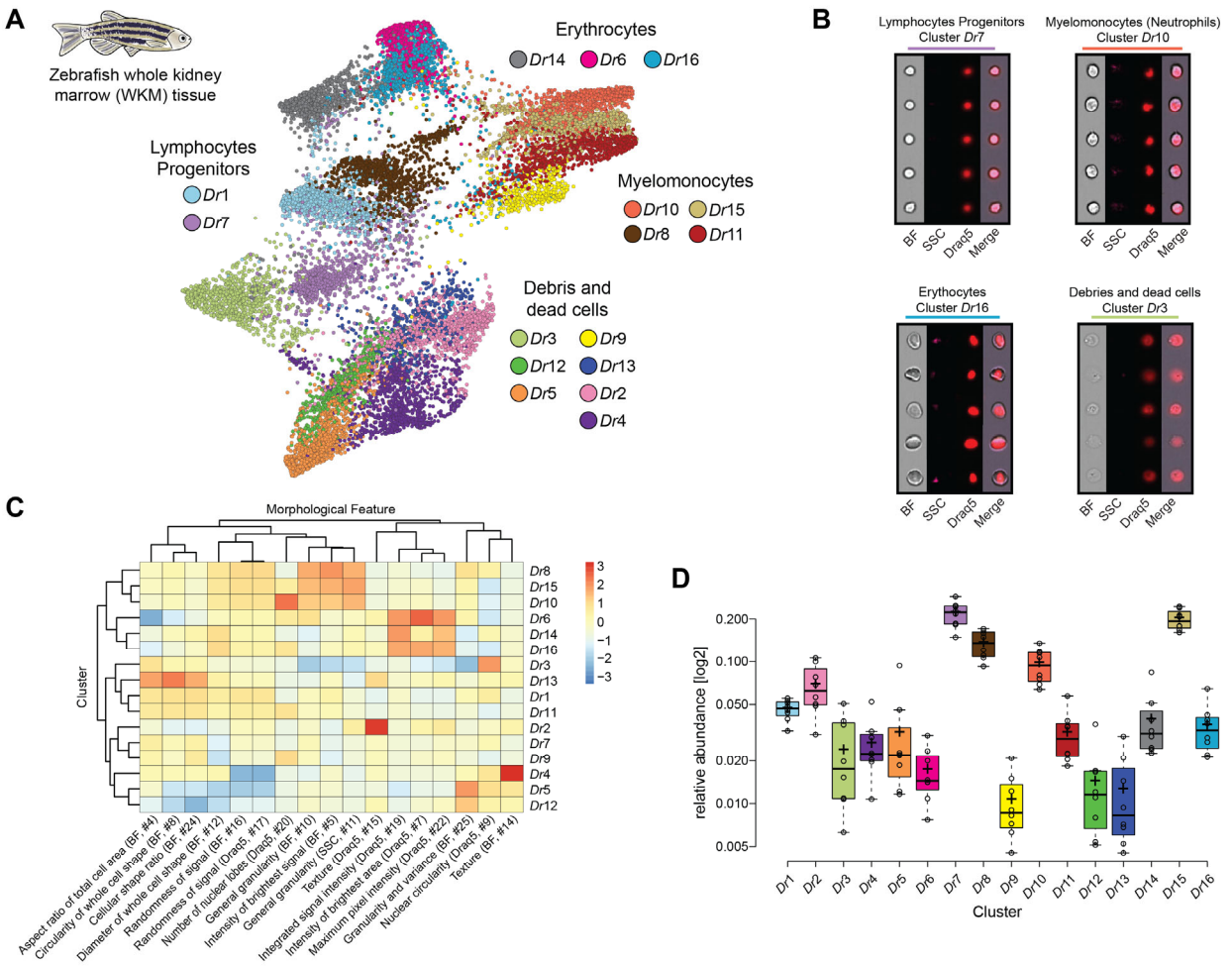


Figure 2: Analysis of cell composition of adult zebrafish WKM

(A) WKM tissue obtained from zebrafish is prepared for image-based flow cytometric analyses and run on the ImageStream®X Mark II (n=8). Standard gating of focused and nucleated events and manual out-gating of most erythrocytes was performed using IDEAS software (Amnis Millipore). The selected images were processed through the pipeline described in Figure 1 and clustered based only on intrinsic morphological and fluorescent feature values. FDL graph visualizes 16 clusters and each color represent a unique cell cluster.

(B) Representative cell images belonging to each cluster are shown to evaluate the homogeneity of the cluster and determine morphology of the cells for cluster annotation (Data File S2). Merge represents the overlay of brightfield (BF), side scatter signal (SSC) and Draq5 (nuclear staining) signal.

(C) Spearman's correlation plot shows the average feature values of the images in each cluster to highlight morphological similarities and differences between events belonging to different clusters, such as cell size or cytoplasm granularity (Table S1a).

(D) Box plot of relative abundance of events within each cluster follows the same color-code used in A.

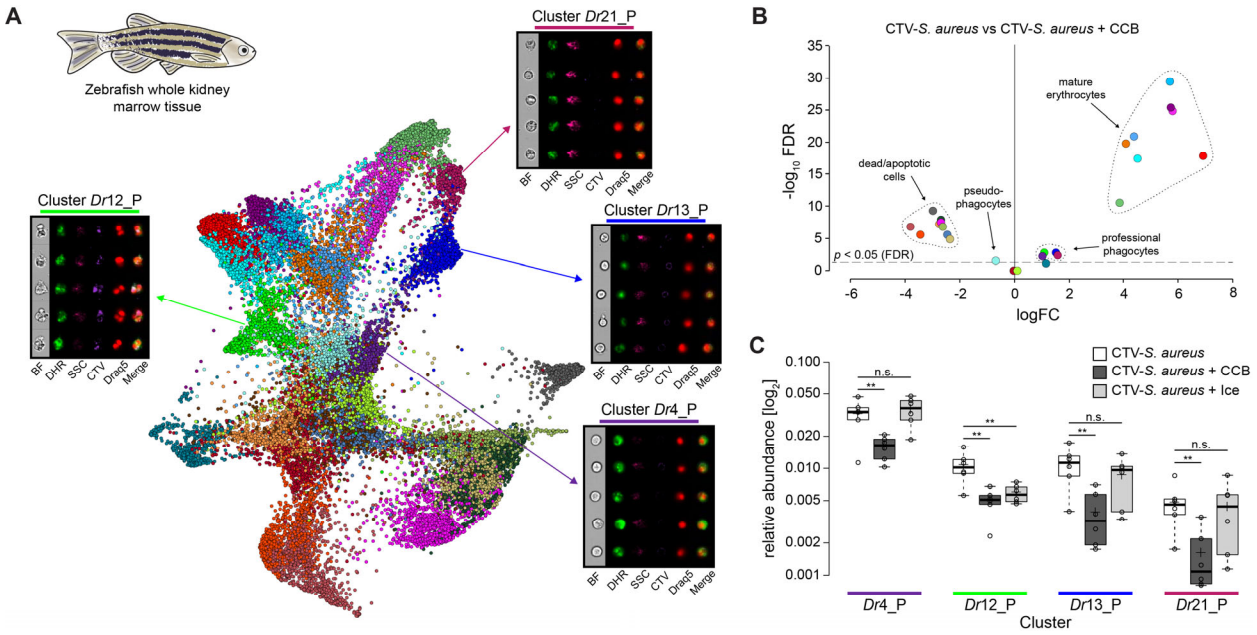


Figure 3: Identification of professional phagocytes in zebrafish WKM

(A) A phagocytosis assay was performed on a cell suspension obtained from zebrafish WKM tissue and the samples were subsequently run on the ImageStream®X Mark II (n=6). FDL graph shows 26 clusters and each color represents a unique cell cluster. Representative cell images belonging to the 4 clusters containing professional phagocytes are shown (Data File S3). Merge represents the overlay of DHR (ROS indicator), CTV (*S. aureus* labeling) and Draq5 (nuclear staining) channels. Table S1b reports the features used for this clustering.

(B) Volcano Plot illustrates comparison of cluster relative abundance between phagocytosis samples (CTV-*S. aureus*) and inhibited-phagocytosis samples (CTV-*S. aureus* + CCB). The log fold change (logFC) is plotted in relation to the FDR (Fold Discovery Rate) corrected p-value ($-\log_{10}$) of each individual cluster calculated with negative binomial regression model (n=6) (Table S2). Dot color follows the same color-code used in A.

(C) Box plot of relative abundances of events within the 4 clusters containing professional phagocytes. Phagocytosis samples (CTV-*S. aureus*), CCB inhibited-phagocytosis samples (CTV-*S. aureus* + CCB) and ice inhibited-phagocytosis samples (CTV-*S. aureus* + Ice) (Figure S5). Statistically significant differences are calculated using the negative binomial regression model between the phagocytosis and the inhibited-phagocytosis samples (Table S2 and S3). ** indicates $p \leq 0.01$ and n.s. indicates not significantly different after FDR (n=6).

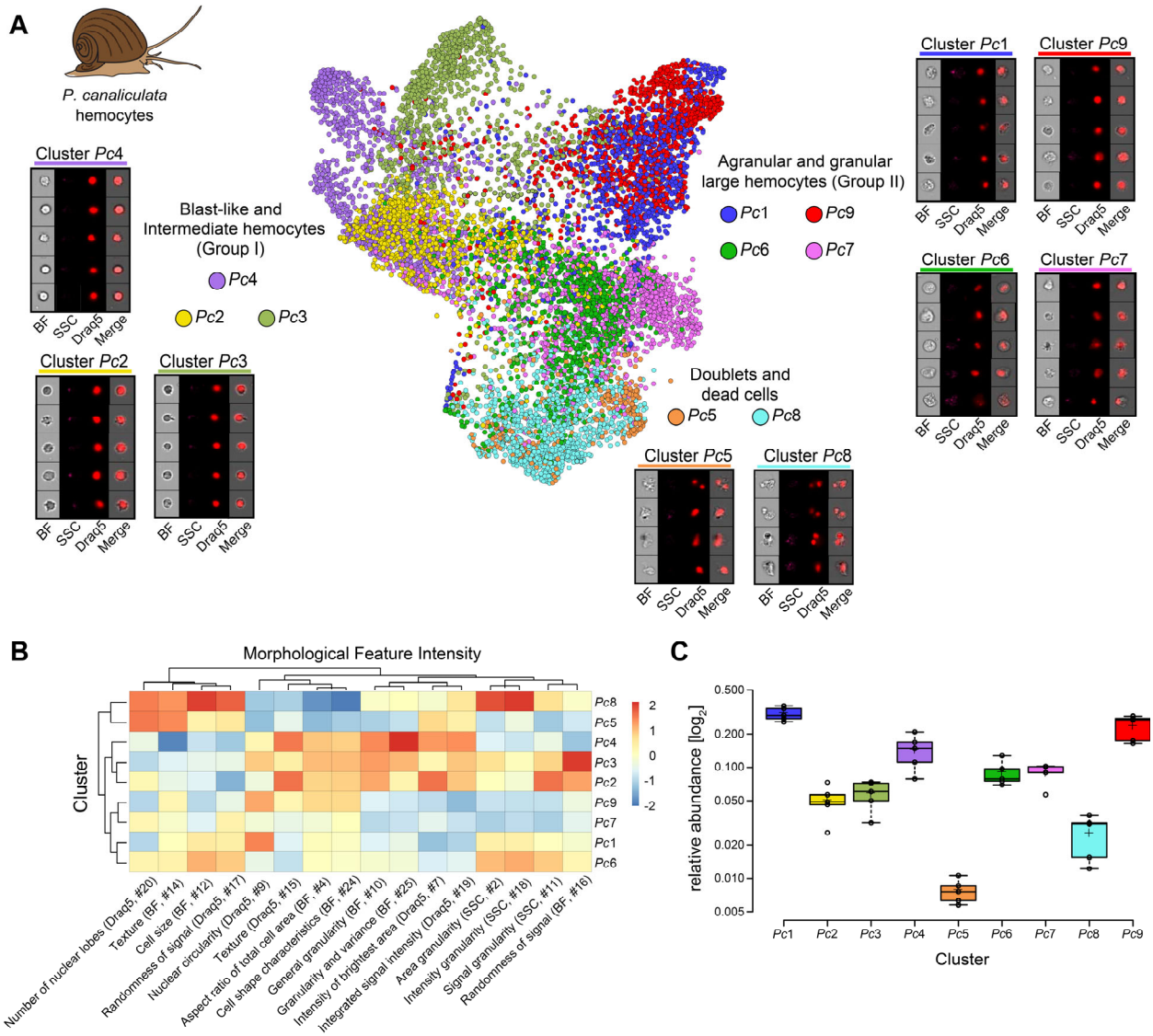


Figure 4: Analysis of *P. canaliculata* hemocyte population using only intrinsic morphological features of the cells

(A) Hemocytes obtained from the apple snail *P. canaliculata* are prepared for image-based flow cytometric analyses and run on the ImageStream®X Mark II (n=5). Standard gating of focused and nucleated events was performed using IDEAS software (Amnis Millipore). The selected images were processed through the pipeline described in Figure 1 and clustered based only on intrinsic morphological and fluorescent feature values. FDL graph is used to visualize the 9 identified clusters and each color represent a unique cell cluster. Representative cell images belonging to each cluster are shown to evaluate the homogeneity of the cluster and determine morphology of the cells for cluster annotation (Data File S4). Merge represents the overlay of brightfield (BF), side scatter signal (SSC) and Draq5 (nuclear staining) signal.

(B) Spearman's correlation plot shows the average feature values of the images in each cluster to highlight morphological similarities and differences between events belonging to different clusters, such as cell size or cytoplasm granularity (Table S1a). Cluster Pc6 is the one among large hemocytes with higher values in features describing cytoplasm granularity (*i.e.*, area granularity #2, intensity granularity #18 and signal granularity #11).

936 (C) Box plot of relative abundance of events within each cluster following the same color-code used in A.
937 Clusters *Pc*5 and *Pc*8, constituted by duplets and dead cells, are those with the lowest number of events,
938 validating the protocol used to prepare these samples.

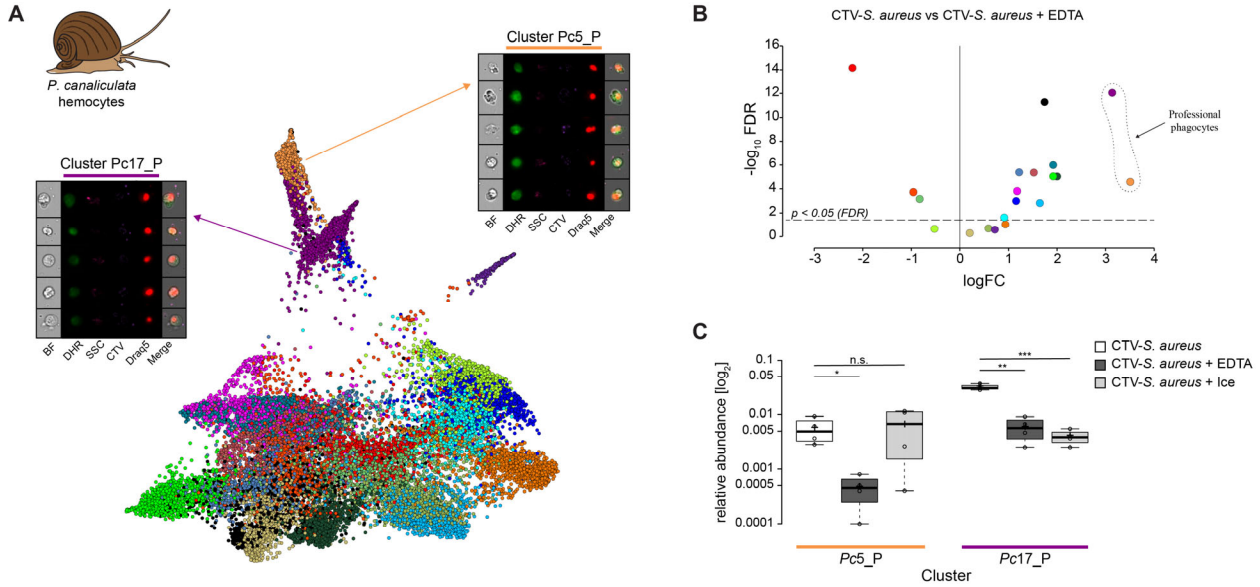
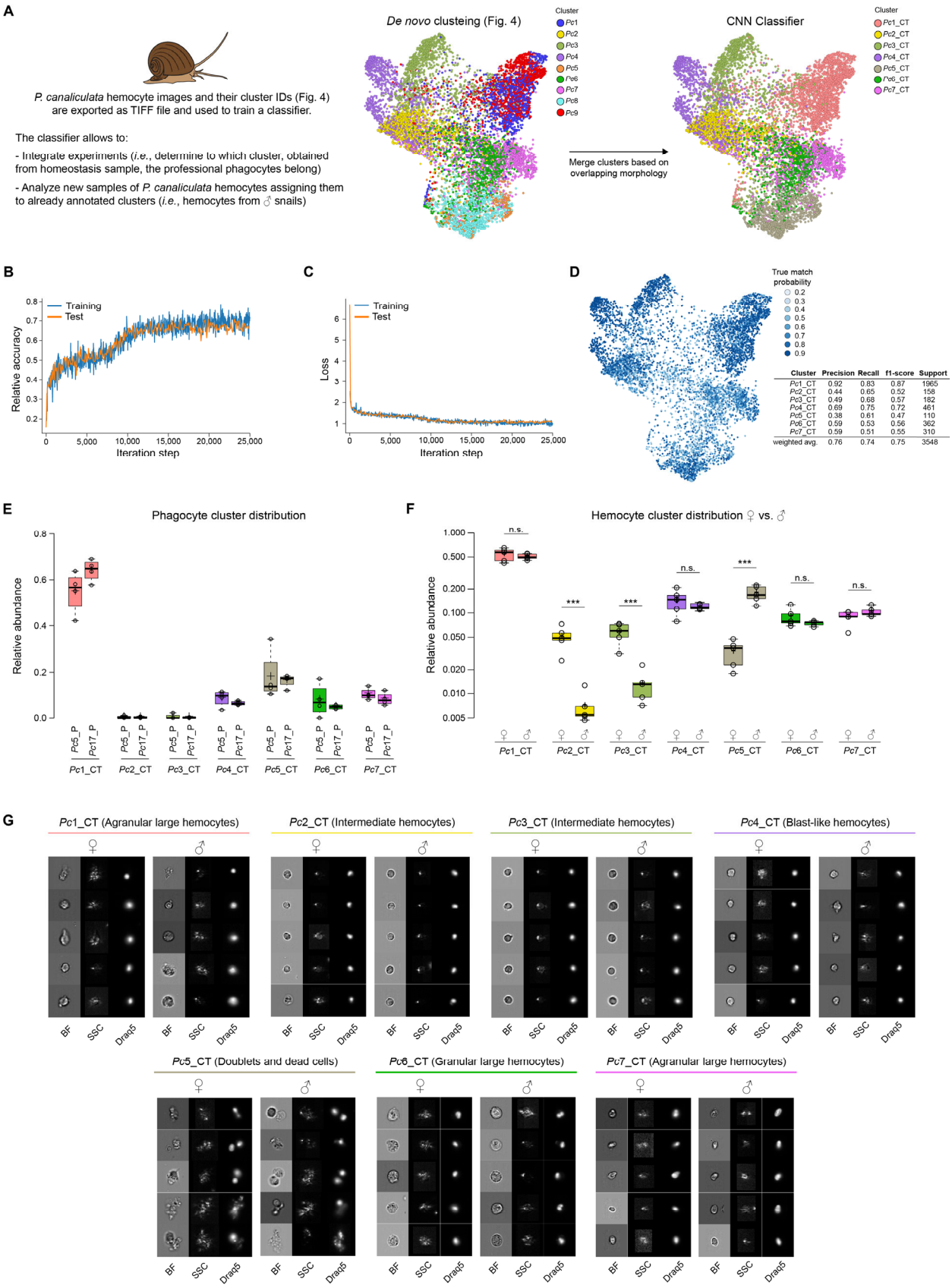


Figure 5: Identification of professional phagocytes among *P. canaliculata* hemocytes

(A) A phagocytosis assay was performed on a cell suspension obtained from apple snail *P. canaliculata* hemolymph and the samples were subsequently run on the ImageStream®X Mark II (n=5). FDL graph shows 20 clusters and each color represents a unique cell cluster. Representative cell images belonging to the 2 clusters containing professional phagocytes are shown (Data File S5). Merge represents the overlay of DHR (ROS indicator), CTV (*S. aureus* labeling) and Draq5 (nuclear staining) channels. Table S1b reports the features used for this clustering.

(B) Volcano Plot illustrates comparison of cluster relative abundance between phagocytosis samples (CTV-*S. aureus*) and inhibited-phagocytosis samples (CTV-*S. aureus* + EDTA). The log fold change (logFC) is plotted in relation to the FDR (Fold Discovery Rate) corrected p-value ($-\log_{10}$) of each individual cluster calculated with negative binomial regression model (n=4) (Suppl. Table S4). Dot color follows the same color-code used in A.

(C) Box plot of relative abundances of events within the 2 clusters containing professional phagocytes. Phagocytosis samples (CTV-*S. aureus*), EDTA inhibited-phagocytosis samples (CTV-*S. aureus* + EDTA) and ice inhibited-phagocytosis samples (CTV-*S. aureus* + Ice) (Suppl. Figure S6). Statistically significant differences are calculated using the negative binomial regression model between the phagocytosis and the inhibited-phagocytosis samples (Suppl. Table S4 and S5). ** indicates $p \leq 0.01$ and n.s. indicates not significantly different after FDR (n=4).


A
P. canaliculata hemocyte images and their cluster IDs (Fig. 4) are exported as TIFF file and used to train a classifier.
The classifier allows to:
- Integrate experiments (i.e., determine to which cluster, obtained from homeostasis sample, the professional phagocytes belong)
- Analyze new samples of P. canaliculata hemocytes assigning them to already annotated clusters (i.e., hemocytes from ♂ snails)
De novo clusteing (Fig. 4)
Cluster
Pc1
Pc2
Pc3
Pc4
Pc5
Pc6
Pc7
Pc8
Pc9
Merge clusters based on overlapping morphology
CNN Classifier
Cluster
Pc1_CT
Pc2_CT
Pc3_CT
Pc4_CT
Pc5_CT
Pc6_CT
Pc7_CT
B
Training
Test
Relative accuracy
Iteration step
C
Training
Test
Loss
Iteration step
D
True match probability
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Cluster Precision Recall f1-score Support
Pc1_CT 0.92 0.83 0.87 1965
Pc2_CT 0.44 0.65 0.52 158
Pc3_CT 0.49 0.68 0.57 182
Pc4_CT 0.69 0.75 0.72 461
Pc5_CT 0.38 0.61 0.47 110
Pc6_CT 0.59 0.53 0.56 362
Pc7_CT 0.59 0.51 0.55 310
weighted avg. 0.76 0.74 0.75 3548
E
Phagocyte cluster distribution
Relative abundance
Pc5_P
Pc17_P
Pc1_CT
Pc2_CT
Pc3_CT
Pc4_CT
Pc5_CT
Pc6_CT
Pc7_CT
F
Hemocyte cluster distribution ♀ vs. ♂
Relative abundance
n.s.
***
♀
♂
Pc1_CT
Pc2_CT
Pc3_CT
Pc4_CT
Pc5_CT
Pc6_CT
Pc7_CT
G
Pc1_CT (Agranular large hemocytes)
Pc2_CT (Intermediate hemocytes)
Pc3_CT (Intermediate hemocytes)
Pc4_CT (Blast-like hemocytes)
Pc5_CT (Doublets and dead cells)
Pc6_CT (Granular large hemocytes)
Pc7_CT (Agranular large hemocytes)
BF
SSC
Draq5

958

959 **Figure 6: Classifier training and use of Convolutional Neural Network for integrating multiple**
960 **experiments**
961 (A) The convolutional neural network (CNN) portion of Image3C allows the unsupervised integration of
962 multiple experiments and the analysis of additional samples assigning new events to already defined and
963 annotated clusters. Images of cells obtained from homeostasis, naïve or WT samples and already clustered
964 *de novo* in an unbiased way by Image3C are exported as TIFF files together with their cluster IDs. *P.*
965 *canaliculata* hemocytes in homeostasis condition obtained from female apple snails (35,000 images) is the
966 dataset used for training the CNN classifier. Before the training, clusters with cells with strongly overlapping
967 morphology were merged (*Pc*1 was merged with *Pc*9 and named *Pc*1_CT; *Pc*5 was merged with *Pc*8 and
968 named *Pc*5_CT). CT stands for Classifier Training.
969 (B) The training was performed on 80% of the exported images and the testing during the training was
970 performed on 10% of the exported images. Relative accuracy was recorded every 100 iterations for 25,000
971 iterations total.
972 (C) The training was performed on 80% of the exported images and the testing during the training was
973 performed on 10% of the exported images. Loss calculation was recorded every 100 iterations for 25,000
974 iterations total and indicates that the training set is not memorized.
975 (D) The True match probability (probability that trained classifier-assigned cluster matches original cluster
976 ID) is given for each cell of the remaining 10% of the original exported images. The detailed precision score
977 for each cluster together with the weighted average (correcting for support) is reported in the table.
978 (E) The CNN classifier allows for integrating experiments, such as phagocytosis assay and morphological
979 assay, with minimal probability of introducing errors because of lack of biomarkers. The plot shows the
980 distribution of the cells belonging to the snail professional phagocytes clusters (*Pc*5_P and *Pc*17_P from
981 Figure 5, phagocytosis assay) among the clusters defined by morphological features using homeostasis
982 conditions.
983 (F) The CNN classifier allows also for the analysis of new samples obtained from the same species and the
984 same tissue used for the training. The new events obtained running male *P. canaliculata* hemocytes, are
985 assigned to clusters that were already annotated during the *de novo* clustering step, allowing for
986 comparison between samples (males vs females) and statistical analysis for cluster relative abundance
987 differences. The box plot shows the comparison of the composition of the hemocyte population between
988 females and males (n=5).
989 (G) Representative cell images for each cluster belonging to female and male hemocytes dataset allows to
990 visually compare the cells used to train the classifier (female) with cells assigned by the classifier and
991 coming from a new set of samples (males). Brightfield (BF), side scatter signal (SSC) and Draq5 (nuclear
992 staining) signal are shown in individual channels.

Supplemental Figures

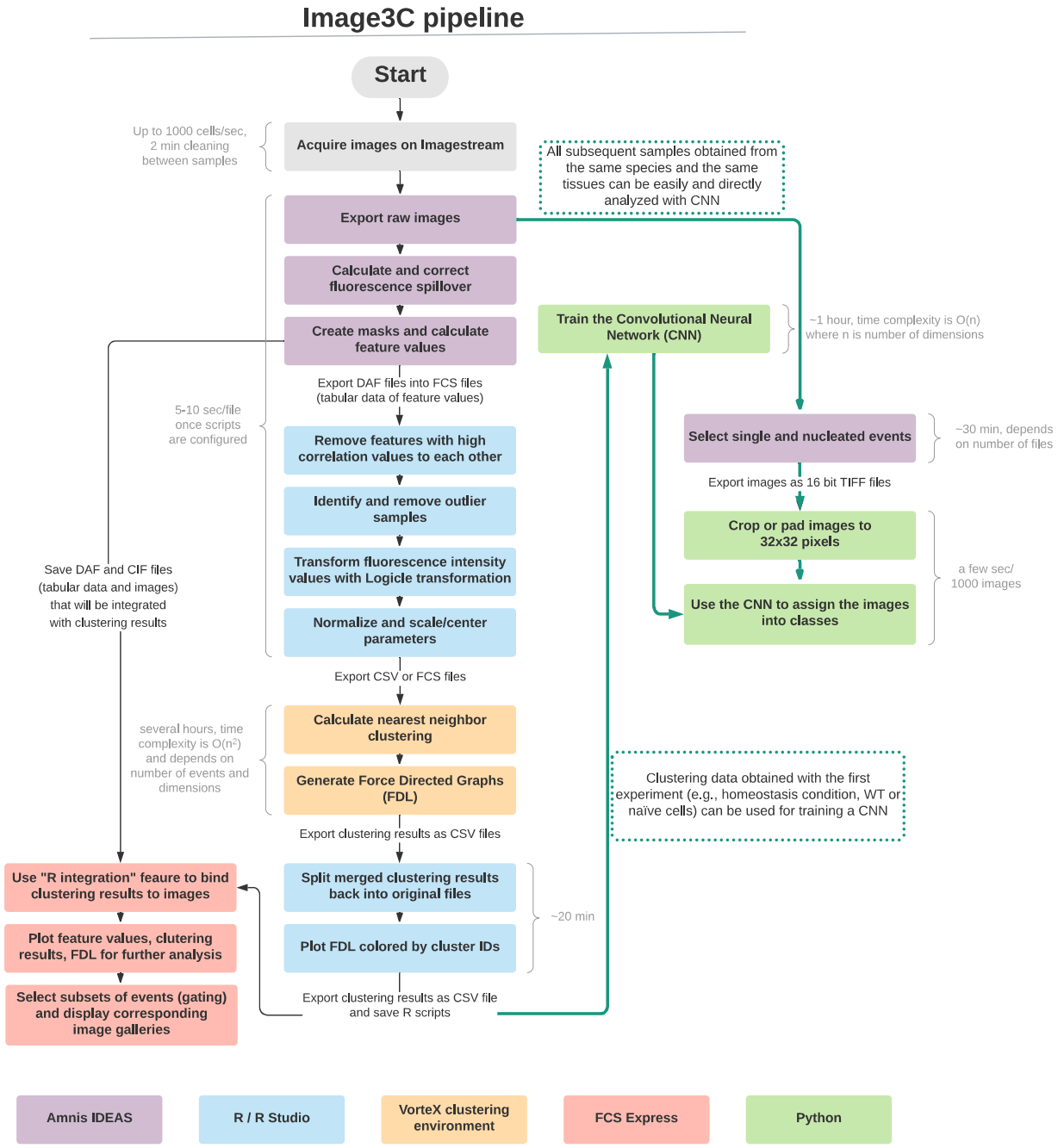


Figure S1: Image3C pipeline with step-by-step technical information
Once the images are collected this pipeline can be followed step-by-step. The software used are color-coded throughout the pipeline, approximation of the time needed to run the individual steps is provided in light gray. The central column and the left side represent the de novo clustering portion of the pipeline, while the right side (green arrows) represents the use of the CNN, being trained with the clusters defined in an unbiased way from the de novo clustering and able to process new samples.

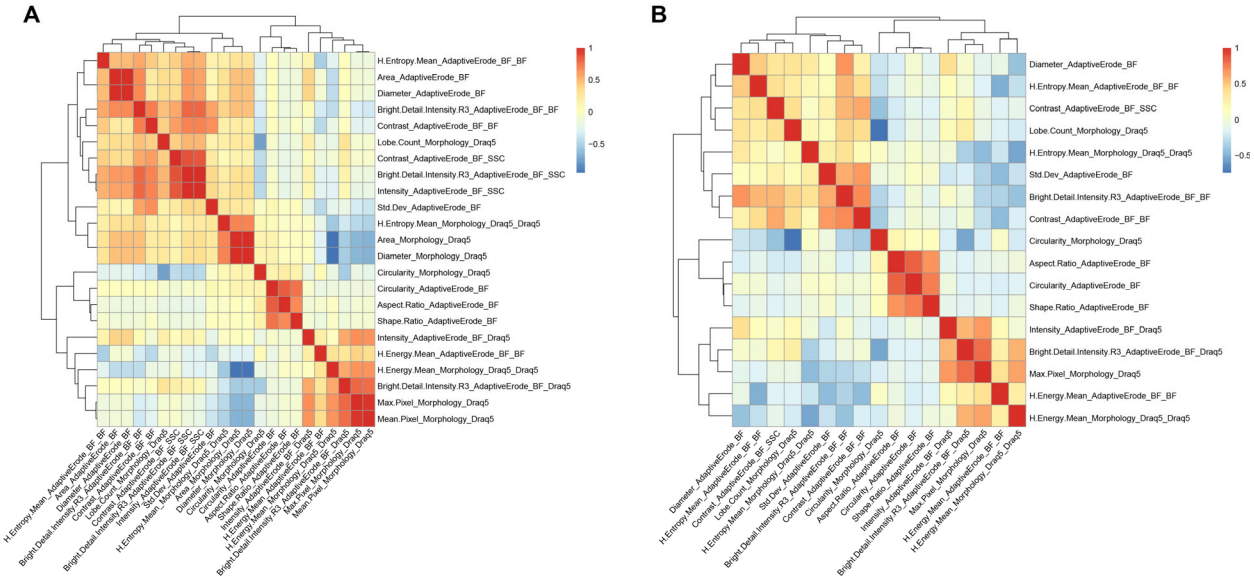


Figure S2: Feature correlation

An example of feature correlation and feature trimming step is shown. The correlation between feature is calculated (a) and the redundant features are removed in order to leave only one representative feature for each group of features with high correlation (b).

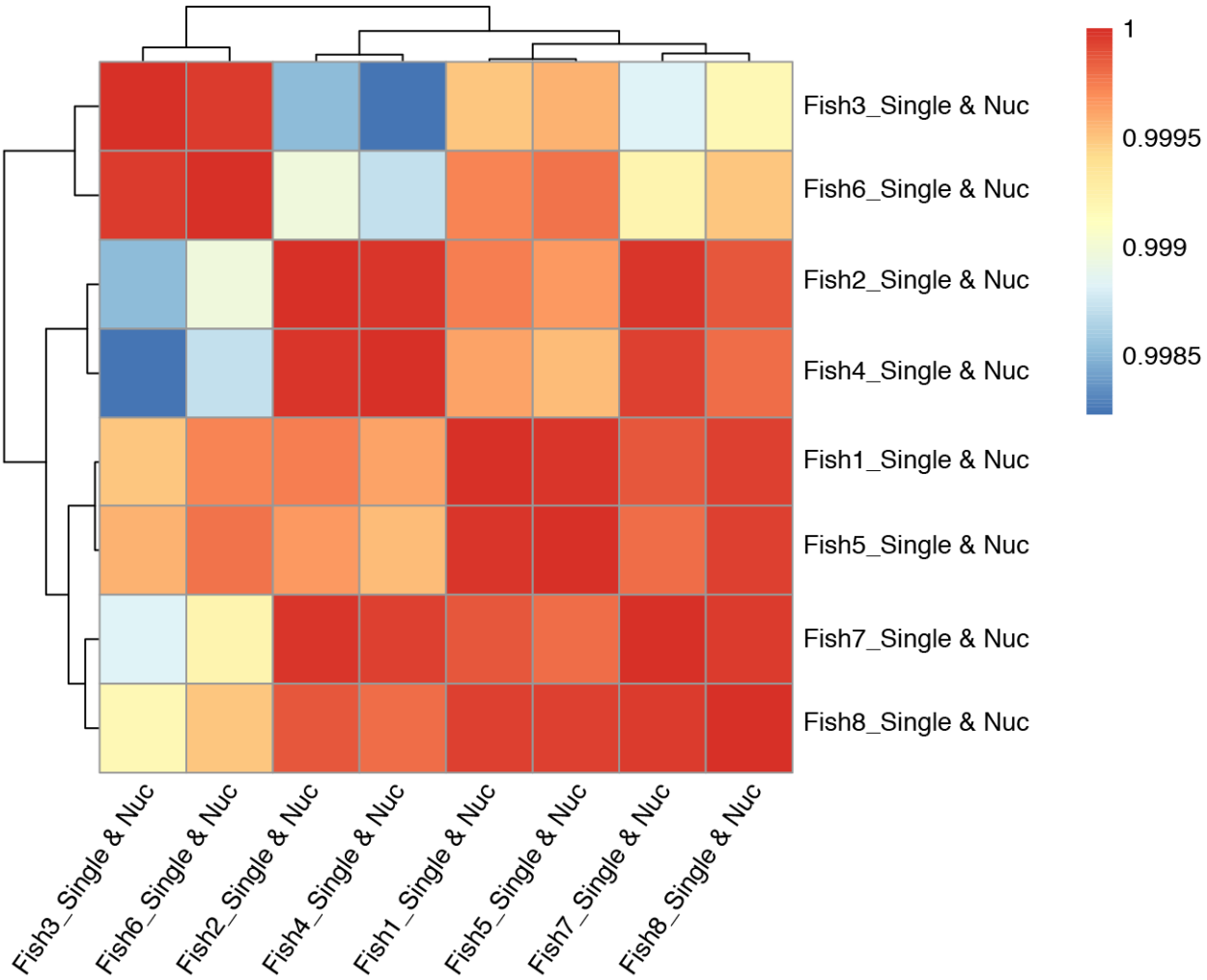


Figure S3: Sample correlation

An example of sample correlation and outlier sample removal step is shown. The correlation between samples is calculated and samples with clear outlier profile are removed from the subsequent analysis.

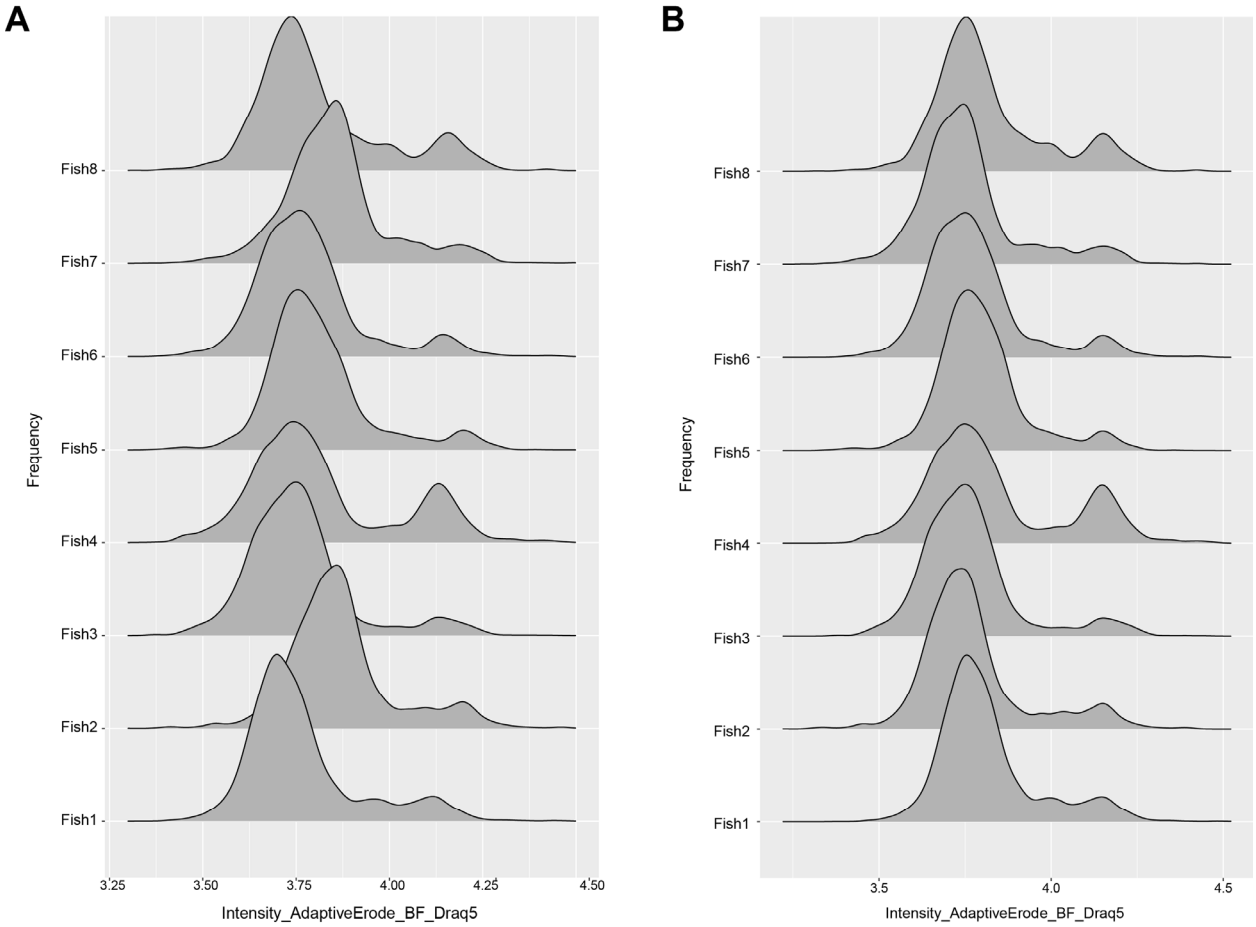
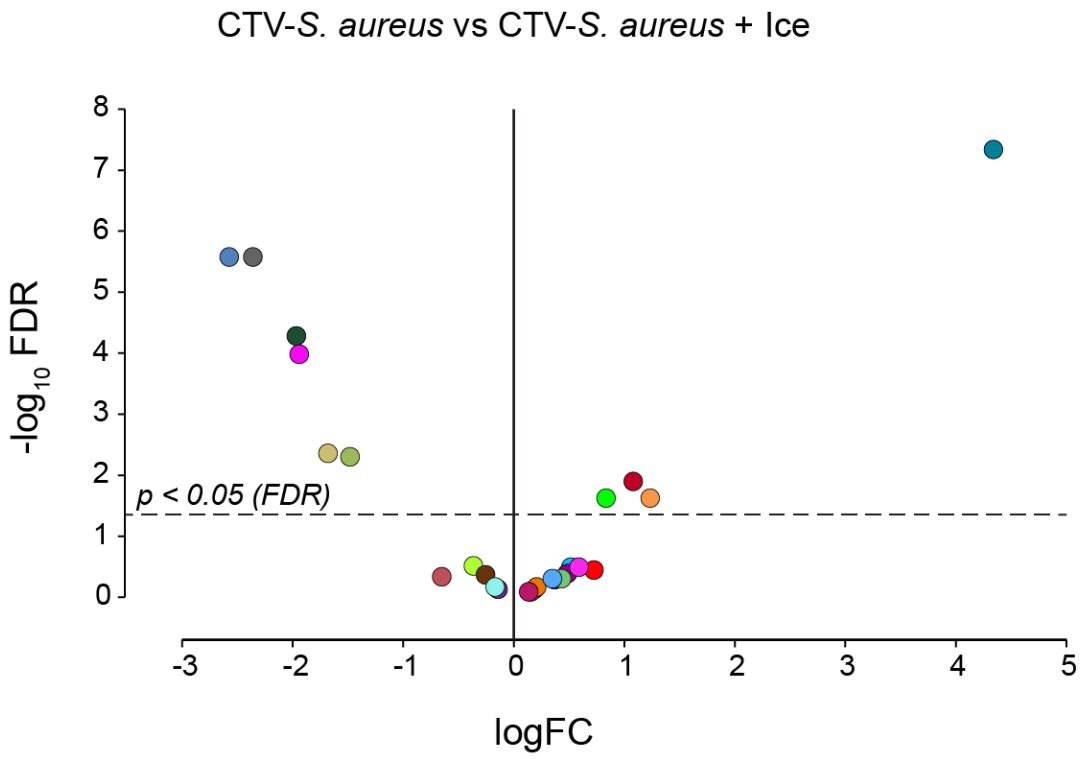


Figure S4: DNA normalization

An example of DNA normalization step is shown. The fluorescent peaks of DNA staining for all the samples of one experiment are shown before (a) and after (b) normalization that aims to align the 2N and 4N peaks across samples.



1019
1020
1021 **Figure S5: Phagocytosis assay on fish WKM with phagocytosis inhibited by ice**
1022 Volcano Plot illustrates comparison between phagocytosis samples (CTV-S. aureus) and inhibited-
1023 phagocytosis samples (CTV-S. aureus + Ice). The log fold change (logFC) is plotted in relation to the FDR
1024 (Fold Discovery Rate) corrected p-value ($-\log_{10}$) of each individual cluster calculated with negative binomial
1025 regression model (n=6) (Table S3). Dot color follows the same color-code used in Figure 3A.

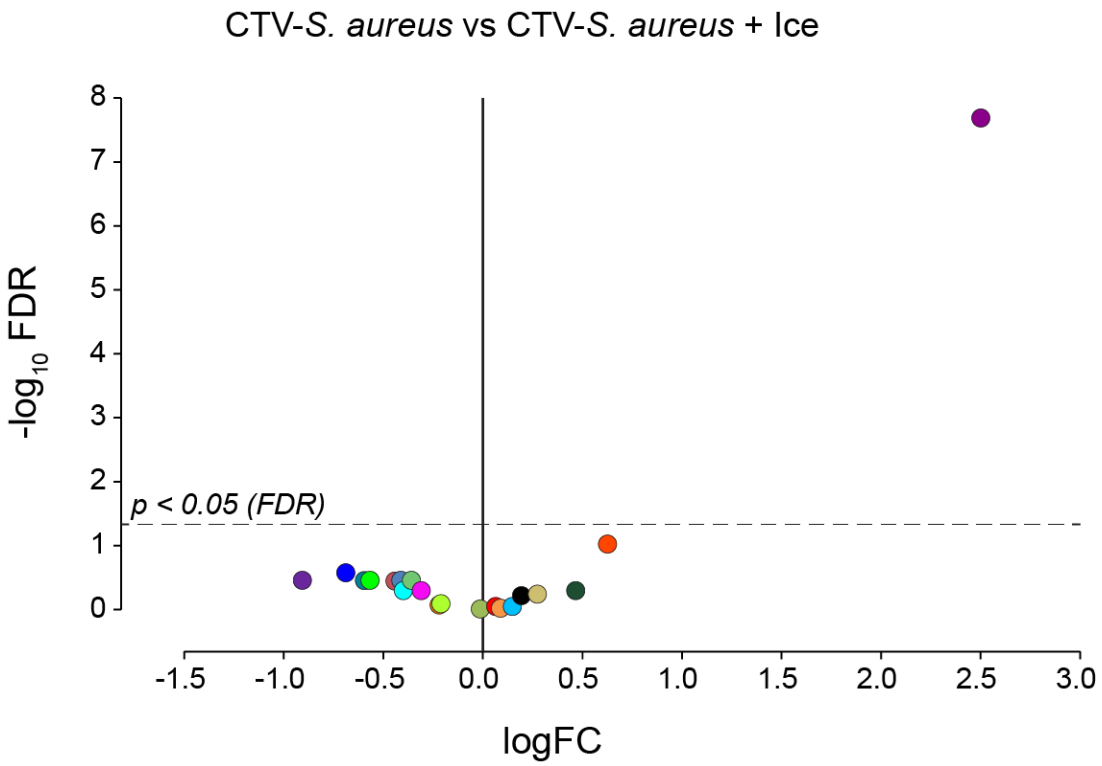


Figure S6: Phagocytosis assay on snail hemocytes with phagocytosis inhibited by ice
Volcano Plot illustrates comparison between phagocytosis samples (CTV-S. aureus) and inhibited-phagocytosis samples (CTV-S. aureus + Ice). The log fold change (logFC) is plotted in relation to the FDR (Fold Discovery Rate) corrected p-value ($-\log_{10}$) of each individual cluster calculated with negative binomial regression model (n=4) (Table S5). Dot color follows the same color-code used in Figure 5A.

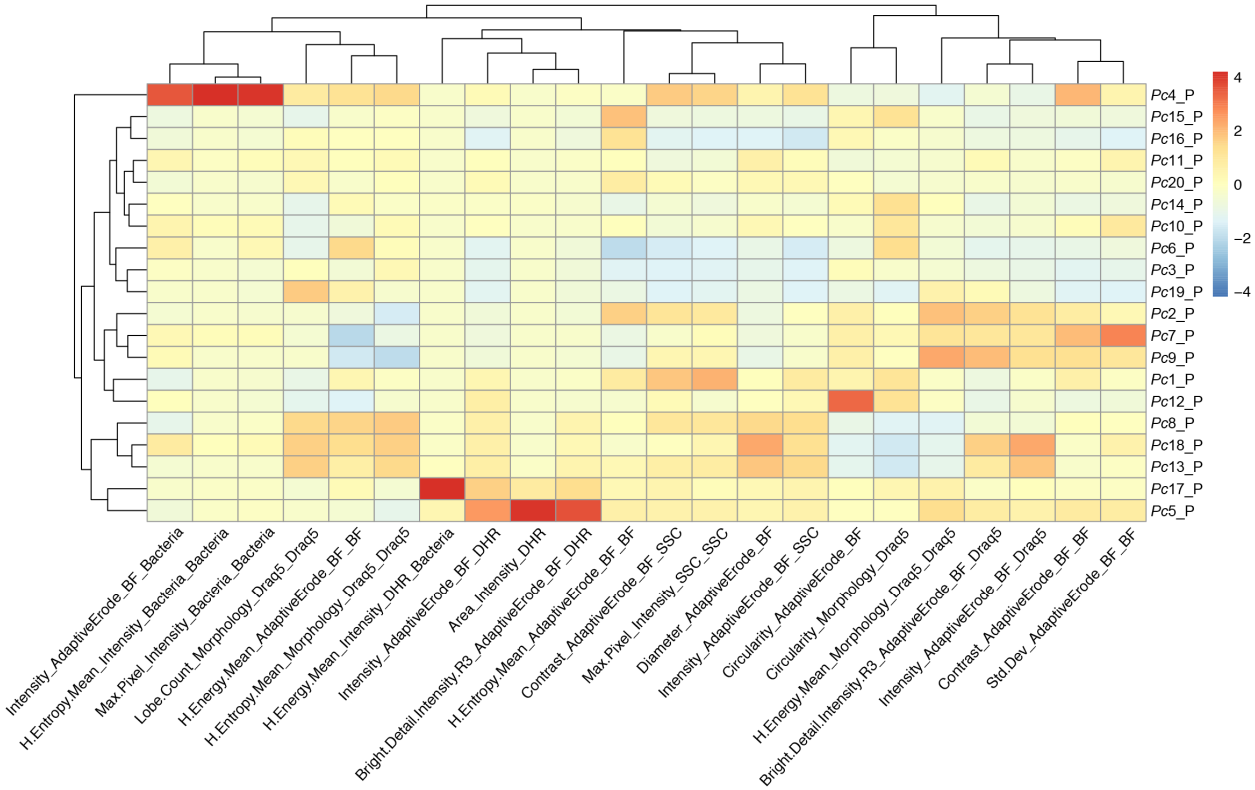


Figure S7: Feature heatmap after phagocytosis assay on snail hemocytes
Spearman's correlation plot shows the average feature values of the images in each cluster to highlight morphological similarities and differences between events belonging to different clusters, such as cell size or cytoplasm granularity (Table S1b). The clusters with higher DHR signal and high bacteria are the one identified as containing professional phagocytes (*Pc*5_P and *Pc*17_P). This is also confirmed by the statistical analysis performed between the phagocytosis samples (CTV-*S. aureus*) and inhibited-phagocytosis samples (CTV-*S. aureus* + EDTA) (Figure 5B and Table S4).

Supplemental Data Files

Data File S1: Cell cluster properties
Cell cluster properties (*e.g.*, cell numbers per cluster, cell feature used for clustering, feature values for each cluster) for zebrafish WKM morphology and phagocytosis assay and for snail hemocyte morphology and phagocytosis assay.

Data File S2: Cell Gallery for zebrafish WKM in homeostasis condition
Representative cell images belonging to each individual cluster identified by Image3C for zebrafish WKM in homeostasis condition are shown. BF is Bright Field, SSC is Side Scatter Signal and Draq5 is nuclear staining. Merge represents the overlay of BF, SSC and Draq5.

Data File S3: Cell Gallery for zebrafish WKM after phagocytosis assay
Representative cell images belonging to each individual cluster identified by Image3C for zebrafish WKM after phagocytosis assay are shown. BF is Bright Field, DHR is a fluorescent ROS indicator, SSC is Side Scatter Signal, Bac is CTV signal (*S. aureus* labeling) and Draq5 is nuclear staining. Merge represents the overlay of DHR, Bac and Draq5.

Data File S4: Cell Gallery for apple snail hemocytes in homeostasis condition
Representative cell images belonging to each individual cluster identified by Image3C for snail hemocytes in homeostasis condition are shown. Ch01 is Bright Field, Ch06 is SSC (Side Scatter Signal) and Ch11 is Draq5 (nuclear staining). Merge represents the overlay of Ch01, Ch06 and Ch11.

Data File S5: Cell Gallery for apple snail hemocytes after phagocytosis assay
Representative cell images belonging to each individual cluster identified by Image3C for snail hemocytes after phagocytosis assay are shown. Ch01 is Bright Field, Ch02 is DHR signal (ROS indicator), Ch06 is SSC (Side Scatter Signal), Ch07 is CTV signal (*S. aureus* labeling) and Ch11 is Draq5 (nuclear staining). Merge represents the overlay of Ch02, Ch06, Ch07 and Ch11.

Supplemental Tables

Table S1a: Names and descriptions of the features quantified by IDEAS software and used for clustering events based on cell morphology in the homeostasis cell composition experiment. BF is Bright Field, CI is Cell Intrinsic, CF is Cell Function.

ID	Feature Name_ ImageMask_Channel	CI or CF	Feature description
1	Area_AdaptiveErode_BF	CI	Cell size
2	Area_Intensity_SSC	CI	Areas of SSC signal above background
3	Area_Morphology_Draq5	CI	Area of DNA signal
4	Aspect.Ratio_AdaptiveErode_BF	CI	Aspect ratio of total cell area
5	Bright.Detail.Intensity.R3_AdaptiveErode_BF_BF	CI	Intensity of brightest staining areas
6	Bright.Detail.Intensity.R3_AdaptiveErode_BF_SSC	CI	Intensity of brightest signal areas
7	Bright.Detail.Intensity.R3_AdaptiveErode_BF_Draq5	CI	Intensity of brightest staining areas
8	Circularity_AdaptiveErode_BF	CI	Circularity of whole cell shape
9	Circularity_Morphology_Draq5	CI	Circularity of nucleus
10	Contrast_AdaptiveErode_BF_BF	CI	Large changes in pixel values - Granularity of signal
11	Contrast_AdaptiveErode_BF_SSC	CI	Large changes in pixel values - Granularity of signal
12	Diameter_AdaptiveErode_BF	CI	Diameter of whole cell shape
13	Diameter_Morphology_Draq5	CI	Diameter of nucleus
14	H.Energy.Mean_AdaptiveErode_BF_BF	CI	Intensity concentration - Texture feature
15	H.Energy.Mean_Morphology_Draq5_Draq5	CI	Intensity concentration - Texture feature
16	H.Entropy.Mean_AdaptiveErode_BF_BF	CI	Intensity concentration and randomness of signal - Texture feature
17	H.Entropy.Mean_Morphology_Draq5_Draq5	CI	Intensity concentration and randomness of signal - Texture feature
18	Intensity_AdaptiveErode_BF_SSC	CI	Integrated intensity of signal within whole cell mask - Cell granularity
19	Intensity_AdaptiveErode_BF_Draq5	CI	Integrated intensity of signal within whole cell mask
20	Lobe.Count_Morphology_Draq5	CI	Number of lobes of nucleus
21	Max.Pixel_Intensity_SSC	CI	Maximum pixel intensity within a whole cell mask - Cell granularity
22	Max.Pixel_Morphology_Draq5	CI	Maximum pixel intensity within a whole cell mask
23	Mean.Pixel_Morphology_Draq5	CI	Mean pixel intensity within a whole cell mask
24	Shape.Ratio_AdaptiveErode_BF	CI	Minimum thickness divided by length - Cell shape
25	Std.Dev_AdaptiveErode_BF	CI	Standard deviation of BF signal - Cell granularity and variance in BF

1081 **Table S1b:** Names and descriptions of the features quantified by IDEAS software and used for
1082 clustering events based on cell morphology and function in the phagocytosis experiment. BF is
1083 Bright Field, CI is Cell Intrinsic, CF is Cell Function.

ID	Feature Name_ ImageMask_Channel	CI or CF	Feature description
1	Area_AdaptiveErode_BF	CI	Cell size
2	Area_Intensity_DHR	CF	Area of DHR staining above background
3	Area_Intensity_Bacteria	CF	Area of CTV staining above background
4	Area_Intensity_SSC	CI	Areas of SSC signal above background
5	Area_Morphology_Draq5	CI	Area of DNA signal
6	Aspect.Ratio_AdaptiveErode_BF	CI	Aspect ratio of total cell area
7	Bright.Detail.Intensity.R3_AdaptiveErode_BF_Bacteria	CF	Intensity of brightest staining areas
8	Bright.Detail.Intensity.R3_AdaptiveErode_BF_BF	CI	Intensity of brightest staining areas
9	Bright.Detail.Intensity.R3_AdaptiveErode_BF_DHR	CF	Intensity of brightest staining areas
10	Bright.Detail.Intensity.R3_AdaptiveErode_BF_Draq5	CI	Intensity of brightest staining areas
11	Bright.Detail.Intensity.R3_AdaptiveErode_BF_SSC	CI	Intensity of brightest signal areas
12	Circularity_AdaptiveErode_BF	CI	Circularity of whole cell shape
13	Circularity_Morphology_Draq5	CI	Circularity of nucleus
14	Contrast_AdaptiveErode_BF_BF	CI	Large changes in pixel values - Granularity of signal
15	Contrast_AdaptiveErode_BF_SSC	CI	Large changes in pixel values - Granularity of signal
16	Diameter_AdaptiveErode_BF	CI	Diameter of whole cell shape
17	Diameter_Morphology_Draq5	CI	Diameter of nucleus
18	H.Energy.Mean_AdaptiveErode_BF_BF	CI	Intensity concentration - Texture feature
19	H.Energy.Mean_Intensity_DHR_DHR	CF	Intensity concentration - Texture feature
20	H.Energy.Mean_Intensity_Bacteria_Bacteria	CF	Intensity concentration - Texture feature
21	H.Energy.Mean_Morphology_Draq5_Draq5	CI	Intensity concentration - Texture feature
22	H.Entropy.Mean_AdaptiveErode_BF_BF	CI	Intensity concentration and randomness of signal - Texture feature
23	H.Entropy.Mean_Intensity_DHR_DHR	CF	Intensity concentration and randomness of signal - Texture feature
24	H.Entropy.Mean_Intensity_Bacteria_Bacteria	CF	Intensity concentration and randomness of signal - Texture feature
25	H.Entropy.Mean_Morphology_Draq5_Draq5	CI	Intensity concentration and randomness of signal - Texture feature
26	Intensity_AdaptiveErode_BF_Bacteria	CF	Integrated intensity of signal within whole cell mask
27	Intensity_AdaptiveErode_BF_DHR	CF	Integrated intensity of signal within whole cell mask
28	Intensity_AdaptiveErode_BF_Draq5	CI	Integrated intensity of signal within whole cell mask
29	Intensity_AdaptiveErode_BF_SSC	CI	Integrated intensity of signal within whole cell mask - Cell granularity

30	Lobe.Count_Morphology_Draq5	CI	Number of lobes of nucleus
31	Max.Pixel_Intensity_Bacteria	CF	Maximum pixel intensity within a whole cell mask
32	Max.Pixel_Intensity_SSC	CF	Maximum pixel intensity within a whole cell mask - Cell granularity
33	Max.Pixel_Morphology_Draq5	CI	Maximum pixel intensity within a whole cell mask
34	Mean.Pixel_Morphology_Draq5	CI	Mean pixel intensity within a whole cell mask
35	Shape.Ratio_AdaptiveErode_BF	CI	Minimum thickness divided by length - Cell shape
36	Std.Dev_AdaptiveErode_BF	CI	Standard deviation of BF signal - Cell granularity and variance in BF

Table S2: Results of negative binomial regression analysis comparing cluster relative abundance between phagocytosis samples (CTV-*S. aureus*) vs phagocytosis inhibited with CCB samples (CTV-*S. aureus* + CCB) in the zebrafish phagocytosis experiment. FC is Fold Change, CPM is Count Per Million, LR is Likelihood Ratio, FDR is Fold Discovery Rate. Relative graph is reported in Figure 3B.

Cluster ID	logFC	logCPM	LR	PValue	FDR
*Dr*1_P	-2.48673	14.76127	24.65067	6.9E-07	1.2E-06
*Dr*2_P	-3.82090	15.11433	30.32912	3.7E-08	7.0E-08
*Dr*3_P	-2.63248	15.10065	30.25504	3.8E-08	7.0E-08
*Dr*5_P	-2.76060	14.21908	33.08750	8.8E-09	1.9E-08
*Dr*6_P	-2.70177	13.37119	36.16033	1.8E-09	4.7E-09
*Dr*7_P	-2.72126	14.24771	34.08437	5.3E-09	1.3E-08
*Dr*8_P	4.48271	14.88169	82.05466	1.3E-19	4.9E-19
*Dr*10_P	-3.45902	14.60211	24.35992	8.0E-07	1.3E-06
*Dr*11_P	6.90476	13.91280	84.08534	4.7E-20	2.1E-19
*Dr*12_P	1.08728	12.83107	11.29997	7.8E-04	1.1E-03
*Dr*13_P	1.51443	12.94985	11.48213	7.0E-04	1.0E-03
*Dr*14_P	-2.99602	11.50543	42.88105	5.8E-11	1.7E-10
*Dr*15_P	-2.38678	12.70026	21.12804	4.3E-06	6.6E-06
*Dr*16_P	5.66338	14.13744	143.18631	5.4E-33	1.4E-31
*Dr*17_P	5.71512	14.80998	122.27043	2.0E-28	2.6E-27
*Dr*19_P	4.07753	14.85917	93.86068	3.4E-22	1.8E-21
*Dr*20_P	3.84792	13.05544	49.27037	2.2E-12	7.3E-12
*Dr*21_P	1.57131	11.87021	9.42118	2.1E-03	2.8E-03
*Dr*23_P	4.37593	16.67274	99.99839	1.5E-23	9.9E-23
*Dr*25_P	5.76977	13.99753	119.32182	8.9E-28	7.7E-27

1093 **Table S3:** Results of negative binomial regression analysis comparing cluster relative abundance
1094 between phagocytosis samples (CTV-*S. aureus*) vs phagocytosis inhibited with ice samples
1095 (CTV-*S. aureus* + Ice) in the zebrafish phagocytosis experiment. FC is Fold Change, CPM is
1096 Count Per Million, LR is Likelihood Ratio, FDR is Fold Discovery Rate. Relative graph is reported
1097 in Figure S5.

Cluster ID	logFC	logCPM	LR	PValue	FDR
*Dr*1_P	-2.57222	14.76127	26.21074	3.1E-07	2.7E-06
*Dr*3_P	-1.47929	15.10065	10.27905	1.3E-03	5.0E-03
*Dr*5_P	1.23557	14.21908	6.95840	8.3E-03	2.4E-02
*Dr*6_P	-1.96681	13.37119	19.93869	8.0E-06	5.2E-05
*Dr*7_P	-1.93868	14.24771	18.19453	2.0E-05	1.0E-04
*Dr*9_P	4.34094	11.68675	36.21393	1.8E-09	4.6E-08
*Dr*12_P	0.83638	12.83107	6.79951	9.1E-03	2.4E-02
*Dr*14_P	-2.35912	11.50543	26.47869	2.7E-07	2.7E-06
*Dr*15_P	-1.67916	12.70026	10.80552	1.0E-03	4.4E-03
*Dr*24_P	1.08190	16.03192	8.34069	3.9E-03	1.3E-02

1098

1099

1100

1101 **Table S4:** Results of negative binomial regression analysis comparing cluster relative abundance
1102 between phagocytosis samples (CTV-*S. aureus*) vs phagocytosis inhibited with EDTA samples
1103 (CTV-*S. aureus* + EDTA) in the apple snail phagocytosis experiment. FC is Fold Change, CPM is
1104 Count Per Million, LR is Likelihood Ratio, FDR is Fold Discovery Rate. Relative graph is reported
1105 in Figure 5B.

Cluster ID	logFC	logCPM	LR	PValue	FDR
*Pc*1_P	1.21972	16.42389	23.86393	1.0E-06	4.1E-06
*Pc*2_P	1.52130	16.31424	23.42745	1.3E-06	4.3E-06
*Pc*5_P	3.50692	11.96025	19.19534	1.2E-05	2.6E-05
*Pc*6_P	2.00062	13.66811	21.45211	3.6E-06	9.1E-06
*Pc*7_P	1.17845	15.71918	15.65951	7.6E-05	1.5E-04
*Pc*8_P	0.91220	14.51336	5.60883	1.8E-02	2.4E-02
*Pc*9_P	1.91957	14.24789	21.73770	3.1E-06	8.9E-06
*Pc*10_P	-0.95771	16.55159	15.15146	9.9E-05	1.8E-04
*Pc*11_P	-2.21453	17.04466	66.60728	3.3E-16	6.6E-15
*Pc*12_P	1.92022	13.48376	27.01612	2.0E-07	1.0E-06
*Pc*13_P	1.15586	13.78276	11.69042	6.3E-04	9.7E-04
*Pc*14_P	1.74206	17.72100	51.24411	8.2E-13	5.4E-12
*Pc*16_P	1.64586	13.23961	10.80603	1.0E-03	1.4E-03
*Pc*17_P	3.13469	13.98527	55.58824	8.9E-14	8.9E-13
*Pc*20_P	-0.82885	16.67206	12.56812	3.9E-04	6.5E-04

1106 **Table S5:** Results of negative binomial regression analysis comparing cluster relative abundance
1107 between phagocytosis samples (CTV-*S. aureus*) vs phagocytosis inhibited with ice samples
1108 (CTV-*S. aureus* + Ice) in the apple snail phagocytosis experiment. FC is Fold Change, CPM is
1109 Count Per Million, LR is Likelihood Ratio, FDR is Fold Discovery Rate. Relative graph is reported
1110 in Figure S6.

Cluster ID	logFC	logCPM	LR	PValue	FDR
*Pc*17_P	2.50037	13.98527	37.29469	1.02E-09	2.03E-08

1111
1112
1113
1114