Title

Label-independent flow cytometry and unsupervised neural network method for de novo clustering of cell populations

Authors

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

[†]These authors contributed equally to this study

[#]These authors share senior authorship

*Author for correspondence: Andrew C. Box (acb@stowers.org)

Affiliation

[1]Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110, United States

[2]Howard Hughes Medical Institute, Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110, United States

[3]Department of Molecular & Integrative Physiology, KU Medical Center, Kansas City, KS 66160, United States.

Abstract

Image-based cell classification has become a common tool to identify phenotypic changes in cells. To date, these approaches are limited to model organisms with species-specific reagents available for cell phenotype identification, clustering and neural network training. Here we present Image3C (Image-Cytometry Cell Classification), a tool that enables cell clustering based on their intrinsic phenotypic features, combining image-based flowcytometry with cell cluster analysis and neural network integration. Using Image3C we recapitulated zebrafish hematopoietic cell lineages and identified cells with specific functions (phagocytes), whose abundance is comparable between treatments. To test Image3C versatility, we performed the same analyses on hemocytes of the snail *Pomacea canaliculata* obtaining results consistent with those collected by classical histochemical approaches. The convolutional neural network, then, uses Image3C clusters and image-based flowcytometry data to analyze large experimental datasets in an unsupervised high-throughput fashion. This tool will allow analysis of cell population composition on any species of interest.

Main text

Modern technologies used to analyze individual cells and subsequently cluster them based on morphology, cell surface protein expression or transcriptome similarities are powerful methods for high-throughput analyses of biological processes at single cell-resolution. Recent advances in image-based cell profiling and single cell RNA-Seq (scRNA-Seq) allow quantification of phenotypic differences in cell populations and comparisons of cell type composition between samples[1]. While studies that use traditional research organisms (*e.g*. mouse, rat, human or fruit fly) benefit from these methods due to the availability of mature genomic platforms and established antibody libraries, the lack of such resources in less traditional organisms prevents extensive use of single-cell based methods to interrogate their biology. In these cases, classical histochemical methods are often used to identify and characterize specific cells, but the quantification analysis of specific cell types can be affected by both observer bias[2] and a dearth of quantitative frameworks for making determination of cell classes.

Automated classification of cells using neural networks has become a promising approach for high-throughput cell analysis[3-7]. Critical for such analysis is the definition of the phenotype that is used to cluster cells. To date, automated clustering and classification techniques required existing knowledge about the organisms or cell type of interest, the availability of cell specific reagents (such as antibodies) or extremely sophisticated equipment not broadly distributed (e.g. single cell sequencing technology)[3-8]. To extend cellular composition analysis to any research organisms without the need for previous knowledge about the cell population of interest or for species-specific reagents at any step of the study, we developed Image3C. Our method analyzes, visualizes and quantifies the composition of cell populations by using cell-intrinsic features and generic, non-species-specific fluorescent probes (*e.g.*, Draq5 or other vital dyes), thus eliminating observer bias

and increasing the analyzed sample size. Image3C is an extremely versatile method that is virtually applicable to any research organism from which dissociated cells can be obtained. By taking advantage of morphology and/or function-related fluorescent probes, Image3C can analyze single cell suspensions derived from any experimental design and identify different constituent cell populations. In addition, we employed a convolutional neural network that uses Image3C defined clusters as training sets and image-based flow cytometry data for unsupervised analysis of cellular composition in large experimental datasets. In summary, Image3C combines modern high-throughput data acquisition through image-based flow cytometry, advanced clustering analysis, statistics to compare the cell composition between different samples and can be used in combination with a neural network component to finely determine changes in the composition of cell population across multiple samples.

The general workflow of Image3C is presented in Fig. 1 using hematopoietic tissue from the zebrafish, *Danio rerio.* We tested whether Image3C can identify homogeneous and biologically meaningful clusters of hematopoietic cells by analyzing only intrinsic morphological and fluorescent features, such as cell and nuclear size, shape, darkfield signal (side scatter, SSC) and texture. Each sample obtained from adult fish was stained and run on the ImageStream®^{X} Mark II (Amnis Millipore Sigma) and individual cell images were collected (Fig. 1a) at a speed of 1,000 images/sec. Feature intensities from both morphological and fluorescent features, such as cell size and nuclear size, were extracted from the cell images using IDEAS software (Amnis Millipore) (Fig. 1a, Table S1 for feature description, Supplemental Methods). The Spearman's correlation values for each pair of features were calculated using all cell events (*i.e.* cell images) of a representative sample and used to trim redundant features[1] (Fig. 1a). The Spearman's correlation of the mean values of remaining features were then used to identify outliers among sample

replicates (Fig. 1a). While morphological features do not require any normalization, fluorescence intensity features often must be transformed using a 'logicle' transformation (R flowCore package)[9-11] to improve homoscedasticity (homogeneity of variance) of distributions. Then, prior to clustering, fluorescent intensity features derived from DNA staining were normalized using the gaussNorm function from the flowStats R package[10-12] to align all 2N and 4N peak positions (Fig. 1a). These feature processing steps must be done independently for each research organism because of the high variability between data and distributions. A final set of feature intensities was used for clustering the events using X-Shift algorithm[13]. Dimensionality reduction and visualization of resultant clusters and events were achieved by generating force directed layout graphs (FDL, Fig. 1b) using a combination of Vortex clustering environment[13] and custom R scripts, respectively (Supplemental Methods). Visualization of the cell images by cluster was done using FCS Express (version 6 Plus) and its integrated R Add Parameters Transformation feature (Fig. 1b, Supplemental Methods). Additionally, cluster feature averages (*i.e.* the mean value of each feature for each cluster) provide a deeper understanding about the morphological features that differ between cells belonging to separate clusters and the cluster distribution can be used to

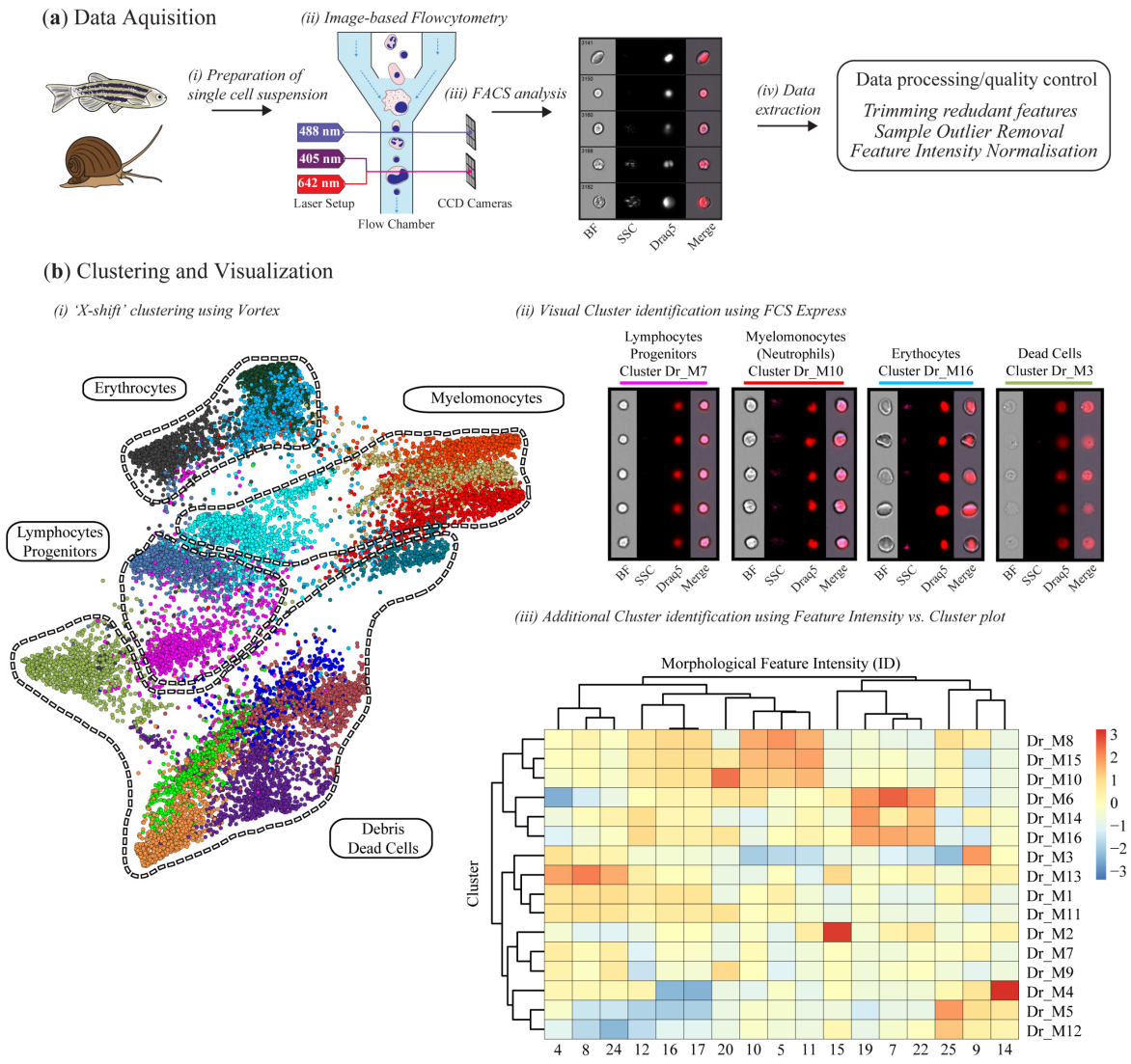

*Fig. 1 | Schematic representation of Image3C using hematopoietic tissue from zebrafish as an example for cell clustering based on morphological features. (**a**) (i) Hematopoietic tissue (or any single suspension of cells of interest) obtained from zebrafish (or any research organism) is prepared for image-based flowcytometric analyses (ii) and run on the ImageStream®^{X} Mark II (n=8). (iii) Standard gating of nucleated events and manual out-gating of most erythrocytes using IDEAS software is followed by (iv) the extraction of intensities for intrinsic morphological and fluorescent features, normalization and quality controls. (**b**) (i) Cell images are clustered based on the intrinsic feature intensities and visualized as a force directed layout (FDL) graph. (ii) R integration in FCS Express software allows the visualization of all the cell images belonging to a specific cluster to evaluate the homogeneity of the cluster and determine phenotype/function of the cells. (iii) In addition to data visualization, Image3C provides a variety of options for integrated data plotting, such as the Spearman's correlation plot of feature intensities per cluster for identification of similarities and differences between cells in different clusters (Table S1 for details).*

derive the most significant contribution to cluster variance from the feature set (Fig. 1b). Finally, statistical analysis to compare cell counts per cluster between potential different treatments is integrated in Image3C and is done using negative binomial regression (Supplemental Methods). As seen in Fig. 1b, Image3C can distinguish between the major classes of hematopoietic cells in

zebrafish (see Data File 1 and 2) that were described using standard flow cytometry sorting and morphological staining approaches[14]. It is noteworthy that this method can clearly identify dead cells and debris (Fig. 1b). The possibility to identify and separate these events from the intact and alive cells allows to optimize experimental conditions and cell treatment protocols in order to minimize cell death and run the subsequent analysis only on the remaining events. In addition, Image3C can identify cells with outstanding morphological features, such as neutrophils from other myelomonocytes (see Fig. 1b).

Next, we sought to determine whether Image3C can be used to detect clusters whose relative abundance significantly changes after specific experimental treatments. We performed a standard phagocytosis assay using hematopoietic cells from zebrafish, which were stained with Draq5 and incubated with CellTrace Violet labeled *Staphylococcus aureus* (CTV-*S. aureus*) and dihydrorhodamine-123 (DHR), a reactive oxygen species that becomes fluorescent if oxidized (Supplemental Methods). The DHR was used as a proxy for cell activation to report oxidative bursting as a consequence of phagocytosis. As control, we inhibited phagocytosis through cytoskeletal impairment by CCB incubation or through incubation at lower temperature (i.e. on ice). Events collected on the ImageStream®^{X} Mark II (Amnis Millipore Sigma) were analyzed with our pipeline and clustered in 26 distinct clusters using intensities of morphological and fluorescent features (see Table S1), such as nuclear staining, *S. aureus* phagocytosis and DHR positivity (Fig. 2a). Professional phagocytes were defined by their ability to take up CTV-*S. aureus* and induce a reactive oxygen species (ROS) response (DHR positive)[15]. To compare between samples incubated with CTV-*S. aureus* and the respective control samples we used the statistical analysis pipeline from Image3C, which is based on a negative binomial regression model (Fig. 2b). In zebrafish, professional phagocytes are mainly granulocytes and monocytic cells and can be

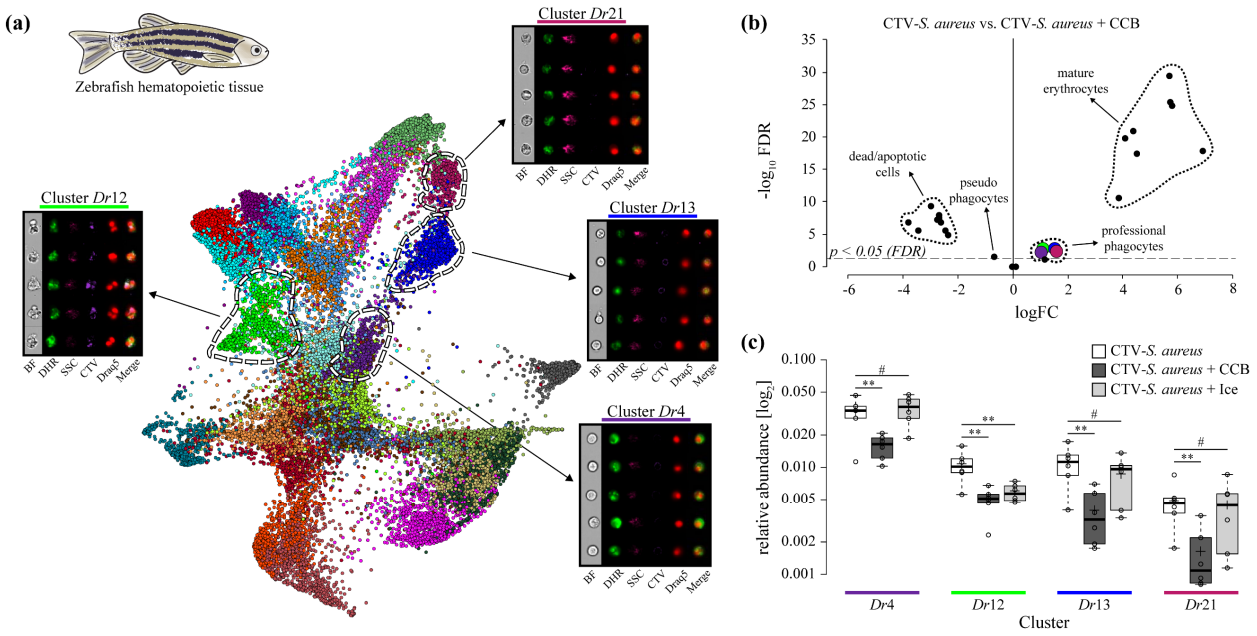

Fig. 2 | Identification of phagocytes in D. rerio *hematopoietic cells using Image3C based on intrinsic feature intensities. (**a**) FDL graph of cluster data, where each color represents a unique cell cluster. Galleries of cluster containing professional phagocytes are shown. Merge represents overlay of DHR, CTV and Draq5 channels. (**b**) Volcano Plot illustrating comparison between treatment sample (hematopoietic cells + CTV-*S. aureus*) and CCB control sample (hematopoietic cells + CTV-*S. aureus* + 0.08 mg/mL CCB). The log fold change (logFC) is plotted in relation to the FDR corrected p-value (-log$_{10}$) of each individual cluster calculated with negative binomial regression model. Clusters containing professional phagocytes are highlighted in the respective color as presented in (a). (**c**) Box plot of relative abundances of cells within cluster containing professional phagocytes in treatment sample (hematopoietic cells + CTV-*S. aureus*), CCB control sample (hematopoietic cells + CTV-*S. aureus* + 0.08 mg/mL CCB) and ice control sample (hematopoietic cells + CTV-*S. aureus* incubated on ice). Statistically significant differences are calculated using the negative binomial regression model between the treatment and the control samples (Supplemental Methods). ** indicates p ≤ 0.01 and # indicates not significantly different after FDR (n=6).*

discriminated from each other based on morphological differences (*i.e.* cell size, granularity and nuclear shape)[16]. By combining the statistical analyses, the visual inspection of the cell galleries (Data File S3) and the intensity of morphological and fluorescent intensities (Data File S2), we identified 4 clusters of professional phagocytes: granulocytes within cluster *Dr*4, *Dr*12 and *Dr*13 and monocytic cells in cluster *Dr*21 (Fig. 2a, 2b). The morphology of cells in cluster *Dr*12 is characteristic of phagocytic neutrophils (Fig. 2a) that become adhesive and produce extracellular traps upon recognition of bacterial antigens[17]. Overall relative abundance of professional phagocytes is 5-10% (Fig. 2c), which is in line with previous studies that estimated the number of

129 professional phagocytes in hematopoietic tissue of adult zebrafish using classical morphological
130 approaches[16].

131 It is interesting to note that CCB selectively affects cell viability based on cell identity (Fig. 2b).
132 We found all erythrocyte containing clusters had a significantly higher cell count in the CTV-*S.*
133 *aureus* samples when compared to the CTV-*S. aureus* + CCB controls (Fig. 2b). Cluster analysis
134 revealed that erythrocytes are almost absent in samples incubated with CCB (Data File S2), while
135 there is a significant increase of dead and apoptotic cells (Fig. 2b, Table S2). Both outcomes are
136 likely due to reduced cell viability of erythrocytes upon CCB incubation. Moreover, we excluded
137 the possibility of higher cell death in the professional phagocytes upon CCB incubation, since we
138 found here pseudo-phagocytes (phagocytes with DHR response but no internalized CTV-*S.*
139 *aureus*) to be significantly more abundant (Fig. 2b, Table S2).

140 Next, we inhibited phagocytosis by incubating the hematopoietic cells on ice (Supplemental
141 Methods) and compared the effectiveness of inhibition with the CCB control (Fig. 2c, Table S3).
142 We found that temperature inhibition of phagocytosis only affects adhesive neutrophils (cluster
143 *Dr*12), probably through the inhibition of adhesion, while CCB effectively blocks phagocytosis in
144 all professional phagocytes in zebrafish hematopoietic tissue (Fig. 2c).

145 To test the versality of Image3C, we repeated the experiments using hemolymph samples from
146 the emerging invertebrate model *Pomacea canaliculata*[18]. For morphological examination of the
147 cellular composition of the hemolymph, we stained the tissue with Draq5 (DNA dye) and run on
148 the ImageStream®^{X} Mark II (Amnis Millipore Sigma) (Supplemental Methods). From the cell
149 images, Image3C analyzed 15 morphological and 10 fluorescent features and identified 9 cell
150 clusters (Fig. 3a). Two of these clusters are constituted by cell doublets, debris and dead cells
151 (clusters *Pc*5 and *Pc*8). (Fig. 3c). Concerning the other clusters, we grouped them into 2 main

categories based on both cell images and previous data[18] (Data File S4). The first category includes small blast-like cells (cluster *Pc*4) and intermediate cells (clusters *Pc*2 and *Pc*3) with high nuclear-cytoplasmic ratio. These cells morphologically resemble the Group I hemocytes previously described using a classical morphological approach [18]. The second category is constituted by larger cells with lower nuclear-cytoplasmic ratio and abundant membrane protrusions (clusters *Pc*1, *Pc*6, *Pc*7 and *Pc*9). Likely, these cells correspond to the previously described Group II hemocytes that include both granular and agranular cells[18]. To identify which of these clusters are enriched with granular cells, the intensities of the morphological features related to cytoplasm texture provided by Image3C were compared between the clusters of this category (Fig. 3b, Data File S4). Cluster *Pc*6 was identified as the one containing the granular hemocytes. The clusters obtained by Image3C, not only were homogeneous and biologically meaningful, but were also consistent with published *P. canaliculata* hemocyte classification obtained by classical morphological methods[18]. Such remarkable consistency has been observed in terms of identified cell morphologies and their relative abundance in the population of circulating hemocytes (Fig. 3c, Data File S4). For example, the relative abundance of the previously reported small blast-like cell is 14.0% a value almost identical to the corresponding cluster *Pc*4 of 13.8%. Similarly, the category of larger hemocytes, or Group II hemocytes represents 80.4% of the circulating cells as measured by traditional morphological methods[18], while clusters *Pc*1, *Pc*6, *Pc*7 and *Pc*9 represent 72.4% of the events analyzed with Image3C. A sub-set of these cells are the granular cells (cluster *Pc*6), which correspond to 7.7% of all hemocytes by classical histological methods[18] and 8.9% by Image3C. The intermediate cells (clusters *Pc*2 and *Pc*3) are less well represented in both approaches, with a

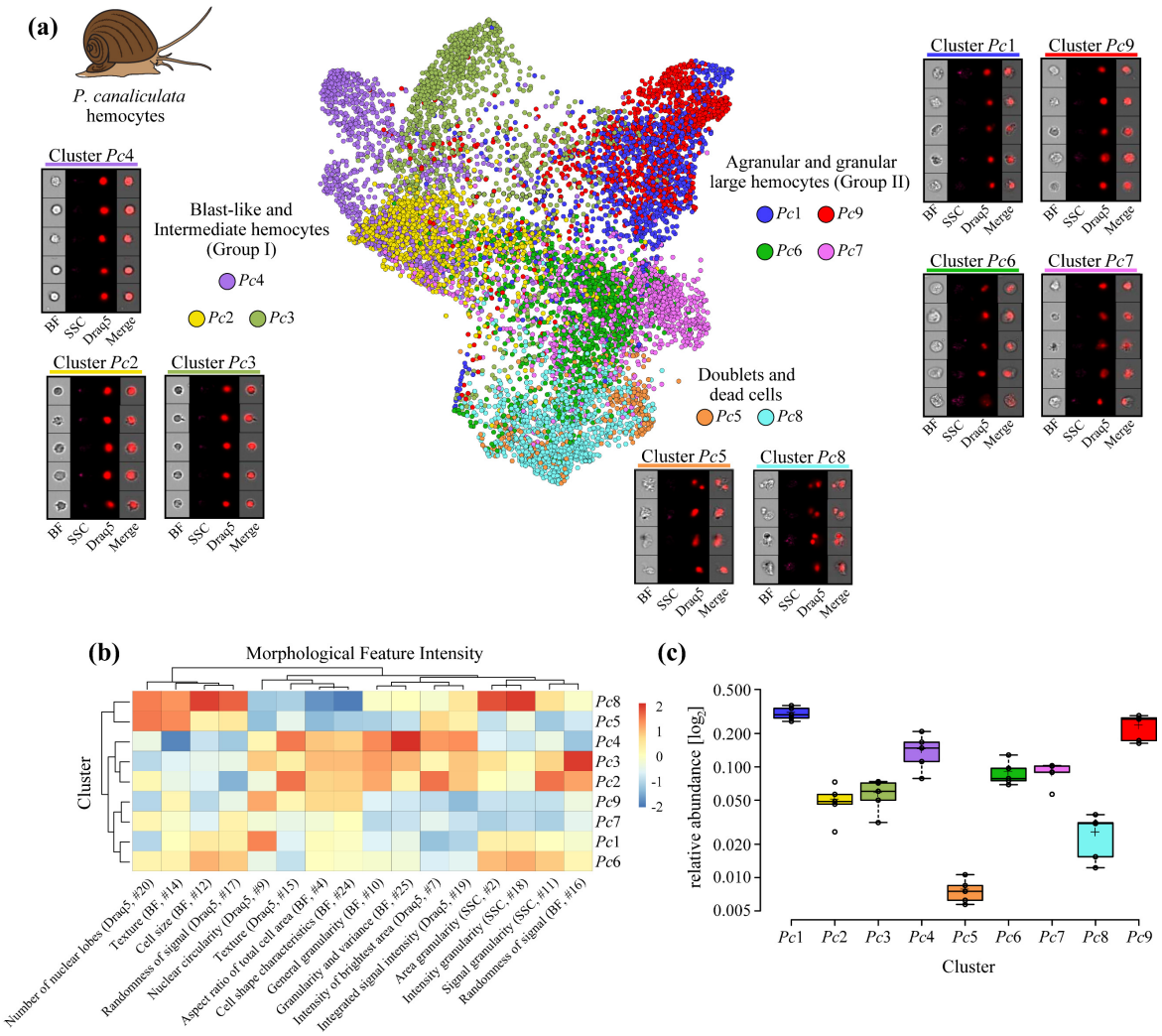

Fig. 3 | Analysis of P. canaliculata *hemocyte population using the Image3C pipeline based only on intrinsic morphological features of the cells. (**a**) FDL graph is used to visualize the 9 identified clusters. Each color represents a unique cell cluster and representative images (galleries) of the cells included in each cluster are shown. Merge represents the overlay of brightfield (BF), side scatter signal (SSC) and Draq5 signal. (**b**) The Spearman's correlation plot of morphological feature intensities per cluster allows the comparison of specific morphological aspects, such as granularity, between cells belonging to different clusters (Table S1 for details). (**c**) Box plot of relative abundance of events within each cluster following the same color-code used in Fig. 2a. Clusters Pc5 and Pc8, constituted by duplets and dead cells, are those with the lowest number of events, validating the protocol used to prepare these samples (n=5).*

173 relative difference in abundance of 5.6% versus 10.6% of the manually and Image3C analyzed
174 events, respectively. However, such difference is likely best explained by the remarkable
175 difference in both, the number of cells and number of features considered for the analyses. Only a
176 few hundred hemocytes were ocularly analyzed based on cell diameter and nuclear-cytoplasmic
177 ratio using traditional histological methods[18], while the automated pipeline used in this study

178 analyzed 10,000 nucleated events for each sample considering 25 cell intrinsic features for each
179 cell. Hence, Image3C represents an unprecedented increase in the accuracy of hemocyte type
180 identification over traditional histological methods.

181 In addition, we performed the same phagocytosis experiment, already done for hematopoietic
182 cells from zebrafish, with hemocytes from *P. canaliculata* (Data File S2, S5, Table S4, S5). Here,
183 we inhibited phagocytosis using either EDTA treatment or low temperature (i.e. incubation on
184 ice). We identified two professional phagocyte clusters (cluster 27430 and 27442, Data File S5),
185 both constituted by large hemocytes (Group II), but with a different DHR signal intensity (ROS
186 response) upon bacteria exposure (cluster 27430 high DHR signal, cluster 27442 low DHR signal,
187 Data File S2 and S5). Similar to the CCB inhibition control in the zebrafish phagocytosis
188 experiment, EDTA is more effective in inhibiting phagocytosis than low temperature since both
189 professional phagocytic clusters (cluster 27430 and 27442) contain significantly higher numbers
190 of cells in the phagocytosis treatment compared to the EDTA control (Table S4). In the ice control
191 sample, however, only cluster 27442 has a significantly higher relative abundance of professional
192 phagocytes compared to the phagocytosis treatment sample (Table S5).

193 The data analysis with Image3C clearly highlighted that the classical phagocytic inhibitors,
194 CCB or EDTA, commonly used in controls for phagocytosis experiments, result in a drastic change
195 of cell morphology, a consequence not easily detectable by other methods and often overlooked.
196 In the present work, these changes significantly modified the overall cell cluster number and
197 distribution, and this must be taken into consideration in any study of morphological features of
198 cells with phagocytosis properties. Furthermore, when determining differences between

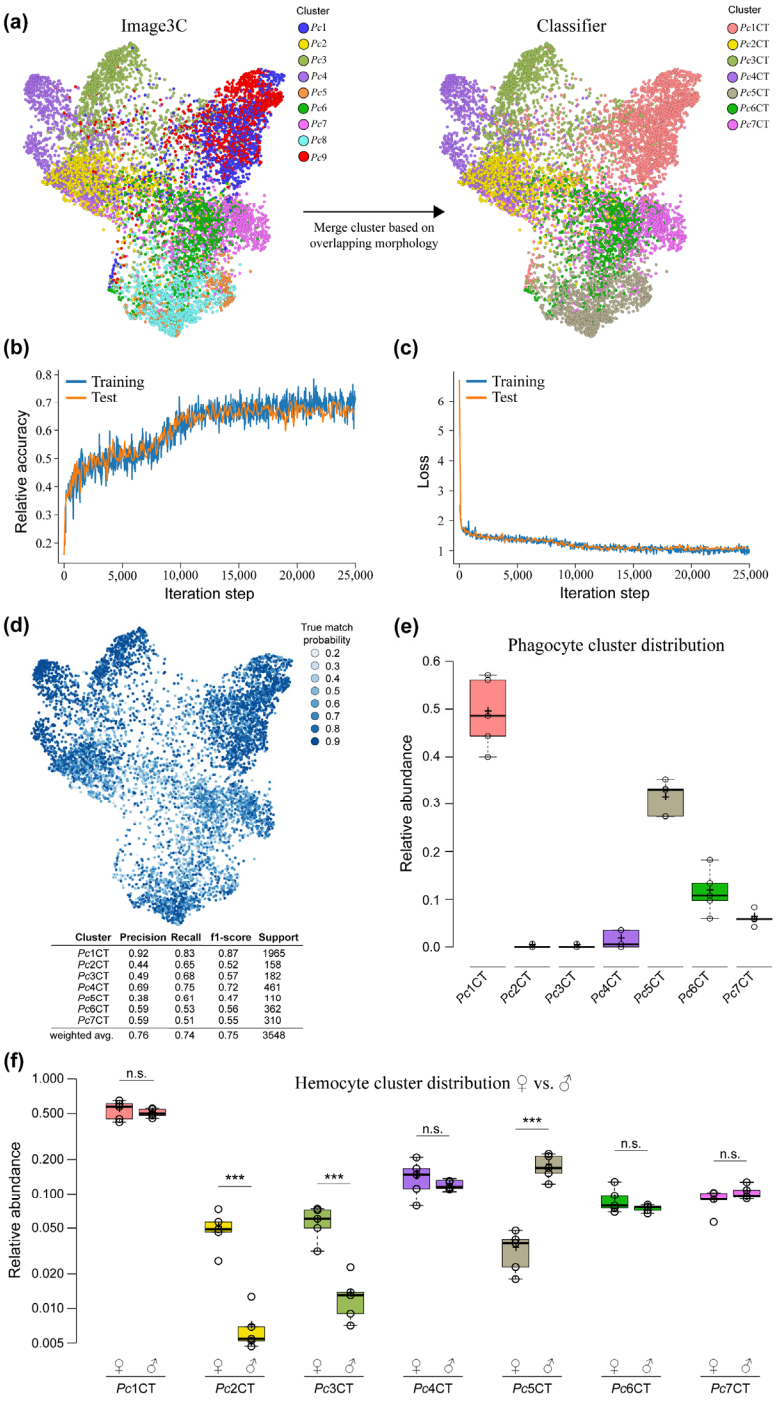

*Fig. 4 | The combination of convolutional neural network with Image3C enables the unsupervised analysis of large experimental datasets. (**a**) Cluster structure from* P. canaliculata *as determined by Image3C was simplified to correct for over clustering (Supplemental Information for details) by combining strongly overlapping clusters (*Pc*1 and* Pc*9 combined to* Pc*1CT;* Pc*5 and* Pc*8 combined to* Pc*5CT). (**b**) Cell images from within resulting clusters were used for neural network training and (**c**) loss calculation for 25,000 iterations. (**d**) The true match probability (probability that trained classifier-assigned cluster matches original Image3C cluster) is given for each cell from the original dataset. The detailed precision score for each cluster together with the weighted average (correcting for support) is given below. (**e**) Distribution of snail phagocytes among the clusters of hemocytes defined by morphological features. (**f**) Comparison of the composition of the hemocyte population between female and male.*

199 experimental treatments, Image3C necessarily combines images and data from all the treatments
200 and re-clusters the cells (Supplemental Methods). Therefore, experiments meant to classify and
201 analyze only innate cell morphologies present in a tissue should be carried out separately from
202 experiments where one or more treatments are likely to significantly affect cell morphology in an

unanticipated manner (e.g. CCB or EDTA incubation). This would prevent treatment effects being conflated with innate morphology differences among unperturbed cell types.

To overcome this potential confounding factor for large scale experiments and allow a direct comparison between same clusters over multiple samples, we designed a convolutional neural network[19] based on the architecture of DenseNet[20] that is able to use Imagestream image files and Image3C cluster information to objectively assign cells to clusters that were previously defined through the Image3C pipeline (Fig. 4). Here, we used the clusters of naïve *P. canaliculata* hemocytes generated by Image3C (Fig. 3a) for setting up the neural network and the first step was to combine Image3C cluster that strongly overlap with one another (Fig. 4a) to correct for clustering and for increase accuracy of the classifier. We used 80% of the cells obtained in the original *P. canaliculata* dataset together with the classifier cluster information to train the classifier with 25,000 iterations (Fig. 4b, c, Supplemental Information for details). After each iteration, 10% of the cells of the original *P. canaliculata* dataset was used to test the classifier (Fig. 4b, c). The relative accuracy for training and testing were determined by scoring numbers of cells whose cluster ID assigned by the classifier matched the cluster ID of the original dataset in relation to the overall cell number used for training and testing, respectively. The network loss was defined by the softmax of the cross entropy[21] between the final output and the one-hot-encoded image labels. Training was performed using the Adam Optimizer[22] with a decaying learning rate starting at 0.001 and decreasing by 1% each step (Fig. 4b, c). To avoid the network memorizing the training set, L2 regularization was applied to the weights. The remaining 10% of the original dataset was used to calculate the precision of the trained classifier (Fig 4d). While clusters with higher support numbers obtained higher precision scores, the weighted average precision score (precision average score across clusters controlling for support numbers) of 0.74 is relatively high considering the

complexity of the phenotype (BF, darkfield and Draq5 images) and comparable to other studies using machine learning for cell classification[5]. The true probability match for each cell (probability for each presented cell given from classifier to match the original Image3C cluster) demonstrates that lower true probability matches occur where cluster strongly overlap (Fig. 4d) potentially giving us information about cell phenotypes that are intermediate between clusters.

To test the efficiency of this pipeline, we extracted the images of the phagocytes obtained with the previous phagocytosis experiment performed on snail hemolymph and determined to which clusters these hemocytes belong through the neural network. We found that 49%, 12% and 6% of the phagocytes belong to cluster *Pc*1CT, *Pc*6CT and *Pc*7CT, respectively (Fig. 4e). These results confirmed the previously published data where the hemocytes able to phagocytize were manually assigned to Group II hemocytes through classical morphological stainings[18]. Only 2% of the phagocytes were clustered in the Group I hemocytes, here represented by cluster *Pc*2CT, *Pc*3CT and *Pc*4CT, while the remaining 31% were assigned by the neural network to the cluster *Pc*5CT, constituted by doublets and dead cells (Fig. 4e). This data can be explained by the fact that *in-vitro* phagocytosis triggers microaggregate formation (hemocyte – hemocyte adhesion) in invertebrate hemocytes that resemble the nodule formation observed *in-vivo*[23].

In an additional test to determine the adaptability of the trained neural network to new datasets, we collected hemocytes from male snails. We stained the cells with Draq5 and recorded BF, SSC and nuclei images from 10,000 cells on the ImageStream®^{X} Mark II (Amnis Millipore Sigma) as described before. We extracted the images of the cells and we used our neural network to determine the relative abundance of hemocytes from males in the 7 clusters used for the training (see Fig 4a). The comparison between female and male hemocyte composition revealed that the only clusters significantly different in terms of relative abundance are *Pc*2CT and *Pc*3CT, defined as Group I

intermediate hemocytes and *Pc*5CT (Fig. 4f). The latter one, comprehending dead cells and doublets, might be explained by the sample preparation and data collection variability, while more interesting is the difference observed in the other two clusters. In the previously published data, no differences were detected through manual classification and counting between females and males hemocytes composition using a classical morphological approach[18]. The unsupervised and high-throughput analysis presented here, instead, allowed us to determine that both subpopulations of intermediate cells defined by the Image3C tool are significantly less represented in the male animals (*Pc*2CT: 5% and 1% in female and male, respectively; *Pc*3CT: 6% and 1%, respectively) (Fig. 4f). While the biological meaning of this difference is not going to be further investigated in this paper, we would like to highlight the power of our tool compared to a more classical approach to determine and analyze the composition of cell population.

These experiments demonstrate that our new tool Image3C in combination with the presented convolutional classifier is capable of analyzing large experimental datasets and identifying significances with small effect sizes independently from observer biases and previous knowledge about the effect of the treatment on the cell morphology.

In summary, we have developed a powerful new method to analyze the composition of any cell population obtained from any research organism of interest at single cell resolution without the need for species-specific reagents such as fluorescently tagged antibodies (multicolor immunophenotyping). We showed how Image3C can cluster cell populations based on morphology and/or function and highlight changes in the cell population composition due to experimental treatments. Furthermore, in combination with the convolutional neural network trained on Image3C clusters, we are capable of unsupervised, bias-free and high-throughput analysis of large experimental datasets with a precise comparison of relative abundance of cells in

the same cluster across different samples. This tool is extremely versatile and can be applied to any cell population of interest and included in any experimental design. In addition, given the recent advancement in image-based flow cytometry that enables image capturing together with cell sorting[24], a scRNA-Seq approach in combination with the Image3C pipeline would enable the simultaneous analysis of both phenotypic and genetic properties of a cell population at single cell resolution. Image3C is freely available from the Github repository[25].

Acknowledgements

We kindly acknowledge Hua Li for her assistance on the statistical analysis and we also thank the Laboratory Animal Services and the Aquatics Facility at the Stowers Institute for Medical Research for animal husbandry. This work was supported by institutional funding to ACB, CW, ASA and NR. ASA is a Howard Hughes Medical Institute Investigator. RP was supported by a grant from the Deutsche Forschungsgemeinschaft (PE 2807/1-1). AA was supported by the Emerging Models grant from the Society for Developmental Biology (SDB) and the postdoctoral fellowship from the American Association of Anatomists (AAA).

Author Contributions

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

Data availability statement

All original data underlying this manuscript can be accessed from the Stowers Original Data Repository at http://www.stowers.org/research/publications/libpb-1390. Image3C code and description is available at https://github.com/stowersinstitute/LIBPB-1390-Image3C.

References

1 Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nature methods* **14**, 849-863, doi:10.1038/nmeth.4397 (2017).

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

3 Eulenberg, P. *et al.* Reconstructing cell cycle and disease progression using deep learning. *Nature communications* **8**, 463, doi:10.1038/s41467-017-00623-3 (2017).

4 Kobayashi, H. *et al.* Label-free detection of cellular drug responses by high-throughput bright-field imaging and machine learning. *Scientific reports* **7**, 12454, doi:10.1038/s41598-017-12378-4 (2017).

5 Blasi, T. *et al.* Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nature communications* **7**, 10256, doi:10.1038/ncomms10256 (2016).

6 Nassar, M. *et al.* Label-Free Identification of White Blood Cells Using Machine Learning. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **95**, 836-842, doi:10.1002/cyto.a.23794 (2019).

7 Lei, C. *et al.* High-throughput imaging flow cytometry by optofluidic time-stretch microscopy. *Nature protocols* **13**, 1603-1631, doi:10.1038/s41596-018-0008-7 (2018).

8 Baron, C. S. *et al.* Cell Type Purification by Single-Cell Transcriptome-Trained Sorting. *Cell* **179**, 527-542 e519, doi:10.1016/j.cell.2019.08.006 (2019).

9 Hahne, F. *et al.* flowCore: a Bioconductor package for high throughput flow cytometry. *BMC bioinformatics* **10**, 106, doi:10.1186/1471-2105-10-106 (2009).

10 Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* **12**, 115-121, doi:10.1038/nmeth.3252 (2015).

11 R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).

12 Hahne, F. *et al.* Per-channel basis normalization methods for flow cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **77**, 121-131, doi:10.1002/cyto.a.20823 (2010).

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

14 Traver, D. *et al.* Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nature Immunology* **4**, 1238-1246, doi:10.1038/ni1007 (2003).

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

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

17 Palic, D., Andreasen, C. B., Ostojic, J., Tell, R. M. & Roth, J. A. Zebrafish (Danio rerio) whole kidney assays to measure neutrophil extracellular trap release and degranulation of primary granules. *Journal of immunological methods* **319**, 87-97, doi:10.1016/j.jim.2006.11.003 (2007).

18 Accorsi, A., Bucci, L., de Eguileor, M., Ottaviani, E. & Malagoli, D. Comparative analysis of circulating hemocytes of the freshwater snail Pomacea canaliculata. *Fish & shellfish immunology* **34**, 1260-1268, doi:10.1016/j.fsi.2013.02.008 (2013).

19 LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1**, 541-551, doi:10.1162/neco.1989.1.4.541 (1989).

20 Huang, G., Liu, Z., Maaten, L. v. d. & Weinberger, K. Q. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2261-2269.

21 Dahal, P. *Softmax cross entropy*, <https://github.com/parasdahal/deepnet> (2017).

22 Kingma, D. P. & Ba, J. L. in *3rd International Conference on Learning Representations.* (eds Yoshua Bengio & Yann LeCun).

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

24 Nitta, N. *et al.* Intelligent Image-Activated Cell Sorting. *Cell* **175**, 266-276 e213, doi:10.1016/j.cell.2018.08.028 (2018).

25 Stowers Institute for Medical Research. *LIBPB-1390-Image3C*, <https://github.com/stowersinstitute/LIBPB-1390-Image3C> (2019).

Online Supplemental Methods

Collection of zebrafish whole kidney marrow (WKM)

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

Collection of apple snail hemocytes

Specimens of the apple snail *Pomacea canaliculata* (Mollusca, Gastropoda, Ampullariidae) were maintained and bred in captivity, in a water recirculation system filled with artificial freshwater (2.7 mM $CaCl_2$, 0.8 mM $MgSO_4$, 1.8 mM $NaHCO_3$, 1:5000 Remineralize Balanced Minerals in Liquid Form [Brightwell Aquatics]). The snails were fed twice a week and kept in a 10:14 light:dark cycle. Wild type adult snails, 7-9 months old and with a shell size of 45-60 mm were starved for 5 days before the hemolymph collection[2]. If not differently specified, female snails were used for the experiments. The withdrawal was performed applying a pressure on the operculum and dropping the hemolymph directly into an ice-cold tube. The hemolymph was not pooled but the cells collected from each animal were individually analyzed. The hemolymph was immediately diluted 1:4 in Bge medium + 10% fetal bovine serum (FBS) and then centrifuged at

500 rcf for 5 min. The pellet of cells was resuspended in 100 µl of Bge medium + 10% FBS. The Bge medium (also known as *Biomphalaria glabrata* embryonic cell line medium) is constituted by 22% (v/v) Schneiders's Drosophila Medium, 4.5 g/L Lactalbumin hydrolysate, 1.3 g/L Galactose, 0.02 g/L Gentamycin in MilliQ water, pH 7.0.

Morphology Assay

The *P. canaliculata* hemocytes were stained with 5 µM Draq5 (Thermo Fisher Scientific) for 10 min, moved to ice and subsequently run one by one on the ImageStream®^{X} Mark II (Amnis Millipore Sigma), where 10,000 nucleated and focused events were recorded for each sample.

D. rerio hematopoietic cells obtained from 8 animals were plated at 4 x 10^5 cells/well in a 96-well plate in 200 µL of medium and incubated for 3 h at room temperature. Cells were stained with 5 µM Draq5 (Thermo Fisher Scientific) for 10 min and subsequently run on the ImageStream®^{X} Mark II (Amnis Millipore Sigma), where 10,000 nucleated and focused events were recorded for each sample. For Image3C analysis, erythrocytes were out-gated to increase number of immune relevant cells and to prevent over clustering. The latter is due to the fact that erythrocytes from fish are nucleated and their biconcave shape result in different morphological feature intensities only depending on their orientation during image acquisition.

Phagocytosis assay

For both animals, cells from a single cell suspension were plated in a 96-well plate at a concentration of 4 x 10^5 cells/well in 200 µL of medium and incubated with 2 x 10^7 CTV-coupled *Staphylococcus aureus*/well (Thermo Fisher Scientific) for 3 h at room temperature. As control for phagocytosis the cells were either incubated with CTV-*S. aureus* on ice or with CTV-*S. aureus* in

the presence of 0.08 mg/mL cytochalasin B (CCB) for zebrafish cells or 30 mM EDTA and 10 mM HEPES for apple snail cells[3]. After 2 h and 30 min we added 5 µM dihydrorhodamine-123 (DHR) (Thermo Fisher Scientific) to the cell suspension to stain cells positive for reactive oxygen species (ROS) production. To control for this treatment with DHR, we incubated the cells with 10 ng/mL phorbol 12-myristate 13-acetate (PMA) to artificially induce ROS production. At 2 h and 50 min since the beginning of incubation with CTV-*S. aureus*, all the samples were stained with 5 µM Draq5 for 10 min. After 3 h incubation with bacteria, cells were moved and stored on ice and subsequently run on the ImageStream®^{X} Mark II (Amnis Millipore Sigma), where 10,000 nucleated and focused events were recorded for each sample.

Data collection on ImageStream®^{X} Mark II

Following cell preparation, data were acquired from each sample on the ImageStream®^{X} Mark II (Amnis Millipore Sigma) at 60x magnification, slow flow speed, using 633, 488 and 405 nm laser excitation. Bright field was acquired on channels 1 and 9. DHR (488 nm excitation) was collected on channel 2, CTV-*S. aureus* (405 nm excitation) on channel 7 and Draq5 (633 nm excitation) on channel 11. SSC was acquired on channel 6.

Data analysis

Raw image data from the ImageStream®^{X} Mark II system was compensated, background was subtracted, and features were calculated using IDEAS 6.2 software (Amnis/Millipore). Feature intensities for all cells and samples were then exported from IDEAS into FCS files for processing in R. See github repository and Table S1 for a full list of features used for each organism and a more detailed description of processing steps. Briefly, exported FCS files were processed in R[4] to

trim redundant features with high correlation values, fluorescence intensity features were transformed using the estimateLogicle() and transform() functions from the flowCore package[5,6], and DNA intensity features were normalized to remove intensity drift between samples using the gaussNorm function from flowStats[7]. The processed data was exported from R[4] using writeflowSet() function in flowCore package[5,6].

Data and clustering results were then imported into the Vortex clustering environment for X-shift k-nearest-neighbor clustering[8]. During the import into Vortex, all features were scaled to 1SD to equalize the contribution of features towards clustering. Clustering was performed in Vortex with a range of k values, typically from 5 to 150, and a final k value chosen using the 'find elbow point for cluster number' function in Vortex and with visual confirmation of the result that over or under-clustering did not occur. Force directed graphs of a subset of cells in each experiment's file set were also generated in Vortex and cell coordinates in the resultant 2d space were exported along with graphml representation of the force directed graph. After clustering and generation of force directed graphs, tabular data was exported from Vortex that included a master table of every cell event and its cluster assignment and original sample ID, as well as a table of the average feature intensities for each cluster and counts of cells per cluster and per sample.

Clustering results were further analyzed and plotted in R[4] by merging all cell events and feature intensities with cluster assignments, and force directed graph X/Y coordinates. Using this merged data and the graphml file exported from Vortex, new force directed graphs were created per treatment condition using the igraph package[9] in R, statistical analysis of differences in cell counts per cluster by condition were performed using negative binomial regression of cell counts per cluster, plots of statistics results and other results generated (see github repository for details), and csv files containing cell and sample ID, feature intensities, X/Y coordinates in force directed and

minimum spanning tree plots were exported for each sample in the experiment set for merging results into daf files in FCS Express Plus version 6 (DeNovo software), which allowed visualization of cell images by cluster and by sub setting of regions within the force directed graphs.

Analysis of daf files was performed in FCS Express by opening daf files and using the "R add parameters" transformation feature to merge the csv files generated above with the daf file feature intensity and image sets. This allowed the generation of image galleries of cells within each cluster and additional analysis in the style of traditional flow cytometry (*i.e.*, gating on 2d plots of features of interest) to explore the clustering results and identify candidate clusters and populations of interest.

The full complement of R packages used includes flowCore[5,6], flowStats[7], igraph[9], ggcyto[10], ggridges[11], ggplot2[12], stringr[13], hmisc[14] and caret[15].

Classifier Setup

We used a convolutional neural network[16] based on the architecture of DenseNet[17] for image classification. Because images from the ImageStream have non-uniform sizes, each image was cropped or padded to 32x32 pixels. The neural network consists of three dense blocks that transition from input three-channel images of 32x32x3 to a final size of 4x4x87 with 87 feature maps. A dense block includes three convolution layers, each followed by leaky relu activation. The output of the dense block is a 2D convolution with a stride of 2 to provide down sampling. The final dense convolutional layer is flattened and fully connected to the output layer that is a 1d vector with a length of the number of classes for prediction. The neural network was implemented in Python using the TensorFlow platform[18] and the SciPy ecosystem[19-21].

Statistics

Negative binomial regression was performed on tables of cell counts per cluster, per sample and plots were generated using R[4] with the edgeR[22] package, which was developed for RNAseq analysis, but includes generally applicable and user-friendly wrappers for regression and modeling analysis and plotting of results. When comparing females and males in Figure 4f to find differences in relative cell abundance in different cluster, a one-way ANOVA was used with subsequent FDR (Benjamini-Hochberg).

Animal experiment statement

Research and animal care were approved by the Institutional Animal Care and Use Committee (IACUC) of the Stowers Institute for Medical Research.

References

1 Traver, D. *et al.* Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nature immunology* **4**, 1238-1246 (2003).

2 Accorsi, A., Bucci, L., de Eguileor, M., Ottaviani, E. & Malagoli, D. Comparative analysis of circulating hemocytes of the freshwater snail Pomacea canaliculata. *Fish & shellfish immunology* **34**, 1260-1268 (2013).

3 Cueto, J. A., Rodriguez, C., Vega, I. A. & Castro-Vazquez, A. Immune Defenses of the Invasive Apple Snail Pomacea canaliculata (Caenogastropoda, Ampullariidae): Phagocytic Hemocytes in the Circulation and the Kidney. *PloS one* **10**, e0123964 (2015).

4 R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).

5 Hahne, F. *et al.* flowCore: a Bioconductor package for high throughput flow cytometry. *BMC bioinformatics* **10**, 106 (2009).

6 flowCore: flowCore: Basic structures for flow cytometry data. R package version 1.46.1. (2018).

7 flowStats: Statistical methods for the analysis of flow cytometry data. R package version 3.38.0. (2018).

8 Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nature methods* **13**, 493-496 (2016).

9 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1-9 (2006).

10 ggcyto: Visualize Cytometry data with ggplot. R package version 1.8.0 (2015).

11 ggridges, Ridgeline Plots in 'ggplot2' v. 0.5.1 (2018).

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

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

14 Hmisc: Harrell Miscellaneous (2019).

15 Kuhn, M. Building Predictive Models inRUsing thecaretPackage. *Journal of Statistical Software* **28** (2008).

16 LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1**, 541-551 (1989).

17 Huang, G., Liu, Z., Maaten, L. v. d. & Weinberger, K. Q. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261-2269.

18 Abadi, M. *et al. TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems*. (2015).

19 Oliphant, T. *A guide to NumPy*. (2006).

20 Oliphant, T. E. Python for Scientific Computing. *Computing in Science & Engineering* **9**, 10-20 (2007).

21 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

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

Supplemental Tables

Table S1: Features used for Clustering. (a) Features used for morphology-based analysis. (b) Features used for functional and morphology-based analysis in phagocytosis experiment.

Table S1 (a)

Feature ID	Feature Name_ImageMask_Channel	Cell Intrinsic (CI) / Cell Function (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 (nuclear staining)
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	Detects large changes in pixel values - can be measure of granularity of signal
11	Contrast_AdaptiveErode_BF_SSC	CI	Detects large changes in pixel values - can be measure of 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	Measure of intensity concentration - texture feature
15	H.Energy.Mean_Morphology_Draq5_Draq5	CI	Measure of intensity concentration - texture feature
16	H.Entropy.Mean_AdaptiveErode_BF_BF	CI	Measure of intensity concentration and randomness of signal - texture feature
17	H.Entropy.Mean_Morphology_Draq5_Draq5	CI	Measure of 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 of stated channel within a whole cell mask - Cell granularity
22	Max.Pixel_Morphology_Draq5	CI	Maximum pixel intensity of stated channel within a whole cell mask
23	Mean.Pixel_Morphology_Draq5	CI	Mean pixel intensity of stated channel within a whole cell mask
24	Shape.Ratio_AdaptiveErode_BF	CI	Minimum thickness divided by length - measure of cell shape characterisic
25	Std.Dev_AdaptiveErode_BF	CI	Standard deviation of BF signal - measure of granularity and variance in BF

Table S1 (b)

Feature ID	Feature Name_ImageMask_Channel	Cell Intrinsic (CI) / Cell Function (CF)	Feature description
1	Area_AdaptiveErode_BF	CI	Cell size
2	Area_Intensity_DHR	CF	Area of DHR staining above background
3	Area_Intensity_Bac	CF	Area of CTV staining above background
4	Area_Intensity_SSC	CI	Areas of SSC signal above background
5	Area_Morphology_DNA	CI	Area of DNA signal (nuclear staining)
6	Aspect.Ratio_AdaptiveErode_BF	CI	Aspect ratio of total cell area
7	Bright.Detail.Intensity.R3_AdaptiveErode_BF_Bac	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_DNA	CI	Circularity of nucleus
14	Contrast_AdaptiveErode_BF_BF	CI	Detects large changes in pixel values - can be measure of granularity of signal
15	Contrast_AdaptiveErode_BF_SSC	CI	Detects large changes in pixel values - can be measure of 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	Measure of intensity concentration - texture feature
19	H.Energy.Mean_Intensity_DHR_DHR	CF	Measure of intensity concentration - texture feature
20	H.Energy.Mean_Intensity_Bac_Bac	CF	Measure of intensity concentration - texture feature
21	H.Energy.Mean_Morphology_Draq5_Draq5	CI	Measure of intensity concentration - texture feature
22	H.Entropy.Mean_AdaptiveErode_BF_BF	CI	Measure of intensity concentration and randomness of signal - texture feature
23	H.Entropy.Mean_Intensity_DHR_DHR	CF	Measure of intensity concentration and randomness of signal - texture feature
24	H.Entropy.Mean_Intensity_Bac_Bac	CF	Measure of intensity concentration and randomness of signal - texture feature
25	H.Entropy.Mean_Morphology_DNA_Draq5	CI	Measure of intensity concentration and randomness of signal - texture feature
26	Intensity_AdaptiveErode_BF_Bac	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_Bac	CF	Maximum pixel intensity of stated channel within a whole cell mask
32	Max.Pixel_Intensity_SSC	CF	Maximum pixel intensity of stated channel within a whole cell mask - Cell granularity
33	Max.Pixel_Morphology_Draq5	CI	Maximum pixel intensity of stated channel within a whole cell mask
34	Mean.Pixel_Morphology_Draq5	CI	Mean pixel intensity of stated channel within a whole cell mask
35	Shape.Ratio_AdaptiveErode_BF	CI	Minimum thickness divided by length - measure of cell shape characterisic
36	Std.Dev_AdaptiveErode_BF	CI	Standard deviation of BF signal - measure of granularity and variance in BF

Table S2: Results of negative binomial regression analysis comparing clusters from zebrafish phagocytosis (Cells + CTV *S. aureus*) with CCB inhibition control (Cells + CTV *S. aureus* + CCB)

Cluster ID	logFC	logCPM	LR	PValue	FDR
Dr1	-2.48673	14.76127	24.65067	6.87E-07	1.19E-06
Dr2	-3.8209	15.11433	30.32912	3.65E-08	7.03E-08
Dr3	-2.63248	15.10065	30.25504	3.79E-08	7.03E-08
Dr5	-2.7606	14.21908	33.0875	8.81E-09	1.91E-08
Dr6	-2.70177	13.37119	36.16033	1.82E-09	4.73E-09
Dr7	-2.72126	14.24771	34.08437	5.28E-09	1.25E-08
Dr8	4.482713	14.88169	82.05466	1.32E-19	4.92E-19
Dr10	-3.45902	14.60211	24.35992	7.99E-07	1.30E-06
Dr11	6.904763	13.9128	84.08534	4.74E-20	2.05E-19
Dr12	1.087279	12.83107	11.29997	0.000775	0.001061
Dr13	1.514425	12.94985	11.48213	0.000703	0.001015
Dr14	-2.99602	11.50543	42.88105	5.82E-11	1.68E-10
Dr15	-2.38678	12.70026	21.12804	4.30E-06	6.57E-06
Dr16	5.663379	14.13744	143.1863	5.35E-33	1.39E-31
Dr17	5.715121	14.80998	122.2704	2.01E-28	2.62E-27
Dr19	4.077533	14.85917	93.86068	3.39E-22	1.76E-21
Dr20	3.847921	13.05544	49.27037	2.23E-12	7.25E-12
Dr21	1.571314	11.87021	9.421178	0.002145	0.002788
Dr23	4.375929	16.67274	99.99839	1.53E-23	9.91E-23
Dr25	5.769772	13.99753	119.3218	8.90E-28	7.72E-27

Table S3: Results of negative binomial regression analysis comparing clusters from zebrafish phagocytosis (Cells + CTV *S. aureus*) with ice inhibition control (Cells + CTV *S. aureus* + ice)

Cluster ID	logFC	logCPM	LR	PValue	FDR
Dr1	-2.57222	14.76127	26.21074	3.06E-07	2.65E-06
Dr3	-1.47929	15.10065	10.27905	0.001345	0.004998
Dr5	1.235565	14.21908	6.9584	0.008343	0.023708
Dr6	-1.96681	13.37119	19.93869	8.00E-06	5.20E-05
Dr7	-1.93868	14.24771	18.19453	1.99E-05	0.000104
Dr9	4.340935	11.68675	36.21393	1.77E-09	4.60E-08
Dr12	0.836382	12.83107	6.799505	0.009118	0.023708
Dr14	-2.35912	11.50543	26.47869	2.66E-07	2.65E-06
Dr15	-1.67916	12.70026	10.80552	0.001012	0.004385
Dr24	1.081904	16.03192	8.340688	0.003877	0.012599

Table S4: Results of negative binomial regression analysis comparing clusters from apple snail phagocytosis (Cells + CTV *S. aureus*) with EDTA inhibition control (Cells + CTV *S. aureus* + EDTA)

Cluster ID	logFC	logCPM	LR	PValue	FDR
27426	1.219719	16.42389	23.86393	1.03E-06	4.14E-06
27427	1.521304	16.31424	23.42745	1.30E-06	4.32E-06
27430	3.506921	11.96025	19.19534	1.18E-05	2.62E-05
27431	2.000616	13.66811	21.45211	3.63E-06	9.07E-06
27432	1.178448	15.71918	15.65951	7.58E-05	0.000152
27433	0.912203	14.51336	5.608834	0.01787	0.023827
27434	1.919568	14.24789	21.7377	3.13E-06	8.93E-06
27435	-0.95771	16.55159	15.15146	9.92E-05	0.00018
27436	-2.21453	17.04466	66.60728	3.31E-16	6.63E-15
27437	1.920223	13.48376	27.01612	2.02E-07	1.01E-06
27438	1.155857	13.78276	11.69042	0.000628	0.000967
27439	1.742058	17.721	51.24411	8.16E-13	5.44E-12
27441	1.645859	13.23961	10.80603	0.001012	0.001445
27442	3.134689	13.98527	55.58824	8.94E-14	8.94E-13
27445	-0.82885	16.67206	12.56812	0.000392	0.000654

Table S5: Results of negative binomial regression analysis comparing clusters from apple snail phagocytosis (Cells + CTV *S. aureus*) with ice inhibition control (Cells + CTV *S. aureus* + ice)

Cluster ID	logFC	logCPM	LR	PValue	FDR
27442	2.500366	13.98527	37.29469	1.02E-09	2.03E-08