Morphological single-cell analysis of peripheral blood mononuclear cells from 390 healthy blood donors with Blood Cell Painting

Authors

Caroline Högel-Starck[1,3], Veera A. Timonen[1,2], Gantugs Atarsaikhan[1,3], Isabel Mogollon[1,*], Minttu Polso[1,*], Antti Hassinen[1], Jarno Honkanen[6], Julius Soini[1], Tanja Ruokoranta[1,5], Toveann Ahlnäs[1], Julianna Juvila[4], Juho J. Miettinen[1], Rodosthenis S. Rodosthenous[1], Mikko Arvas[4], FinnGen, Caroline A. Heckman[1,3], Jukka Partanen[4], Mark Daly[1,7], Aarno Palotie[1,7], Lassi Paavolainen[1,3,†], Vilja Pietiäinen[1,3,†], Esa Pitkänen[1,2,3,†]

Affiliations

1) Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Sciences (HiLIFE), University of Helsinki, Helsinki, Finland

2) Applied Tumor Genomics Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland

3) iCAN Digital Precision Cancer Medicine Flagship, Helsinki, Finland

4) Finnish Red Cross Blood Service Biobank, Vantaa, Finland

5) Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland

6) Translational Immunology Program, Research Programs Unit, University of Helsinki, Helsinki, Finland

7) Broad Institute of MIT and Harvard, Cambridge, MA, USA

* These authors contributed equally to this work.

† These authors contributed equally to this work.

Correspondence: lassi.paavolainen@helsinki.fi, vilja.pietiainen@helsinki.fi, esa.pitkanen@helsinki.fi

Keywords

Cell Painting, peripheral blood mononuclear cell, high content imaging, morphology, phenotypic analysis, healthy blood donors, FinnGen, biobanks, acute myeloid leukemia, cell typing, batch effects, flow cytometry

Abstract

Recent molecular profiling and phenotyping methods combined with machine learning based analyses enable genotype-phenotype discovery at an unprecedented scale. The challenge now lies in unraveling the biological mechanisms underpinning these associations. High content imaging is a cost-effective approach for morphological and

functional profiling of single cells that has provided insight into mechanisms of disease phenotypes, and consequences of genetic and drug perturbations. However, the morphological variability of healthy immune cells – instrumental to understanding disease-specific deviations from the healthy state – is still relatively uncharacterized. To elucidate this variability at scale, we generated high-resolution fluorescent confocal imaging data of peripheral blood mononuclear cell (PBMC) samples from 390 healthy blood donors with the Blood Cell Painting protocol. The protocol, developed here from the popular Cell Painting morphological profiling assay, optimizes for efficiency and throughput, and includes PBMC thawing, plating and fluorescence marker staining of non-adherent blood cells, followed by confocal and widefield imaging with a high content microscope. We assigned cell types based on cellular features with a classifier trained expert annotations, and observed monocytes to be five-fold more frequent in imaging data compared to flow cytometry baseline, with B and T cells being two-fold less frequent. We hypothesize this discrepancy is due to differential adherence between the cell types. We also evaluated three computational methods for correcting batch effects in imaging data, and found Harmony to perform the best, compatible with previous reports. Finally, we performed the Blood Cell Painting protocol on PBMCs in acute myeloid leukemia, and showed the protocol to be able to distinguish between AML FAB subtypes. Our study highlights the utility of high-content imaging with Cell Painting in characterizing and understanding health and disease phenotypes, opening avenues to further studies with integrated imaging and molecular profiling data.

This manuscript is a work in progress, and we anticipate incorporating additional results into subsequent versions.

Background

A thorough biological understanding of the impact of germline variants on health and disease is still lacking despite the recent successful population-scale genome-wide (GWAS) and phenome-wide association studies (PheWAS), leading to the discovery of genotype-phenotype associations (Kurki et al. 2023). An essential connection between genotypes and disease phenotypes lies within cellular phenotypes, which can be analyzed at the single-cell level, for example, with single-cell RNA-sequencing (scRNA-seq) and high-content imaging (HCI). Indeed, single-cell profiling has revealed an unprecedented cellular heterogeneity in healthy and diseased tissues, with implications for the diagnosis and treatment of diseases, through an improved understanding of immunity, for instance (Stubbington et al. 2017). Sequencing-based approaches allow the identification of cell types and subtypes, and mapping of their evolutionary trajectories (La Manno et al. 2018), albeit at a high cost and without the cellular morphological information. HCI allows millions of cells to be routinely profiled in a single experiment providing cellular context and morphological details at a lower

price (Piccinini et al. 2017, Bray et al. 2016, Cimini et al. 2023). Consequently, HCI has been used to reveal how immune cell phenotypes reflect molecular and personal health information, as well as to study the response of leukemic cancer cells to different drugs at single cell level in clinical trials (Severin et al. 2022, Luo, 2022, Severin, 2021, Snijder et al. 2017).

Cell Painting is a popular HCI assay for morphological and functional profiling of single cells and cell organelles via fluorescence microscopy (Gustafsdottir et al. 2013, Bray et al. 2016, Cimini et al. 2023). Phenotypes derived from Cell Painting data have been associated with somatic variants in cancer, transcriptomics, as well as genetic and chemical perturbations (Caicedo et al. 2022, Rohban et al. 2017, Sundaresh et al. 2023). A recent work resulted in a large dataset of Cell Painting and gene expression (L1000 assay) data for >28,000 genetic and chemical perturbations, enabling cross-modality prediction (Haghighi et al. 2022). In a related study, >20,000 single-cell CRISPR knockout experiments were performed in >30 million cells with a Cell Painting readout leading to an atlas capturing biological pathways and protein-protein interactions (Ramezani et al. 2023). Thus far, Cell Painting has been primarily applied to cancer cell lines and human fibroblasts (Bray et al. 2016, Willis et al. 2020), with human immune cells receiving less attention. Nevertheless, morphological profiling with Cell Painting has been used to study immune responses in PBMCs of older adults after viral infection (White, 2020).

In this study, we developed the Blood Cell Painting (BCP) workflow (**Fig. 1**), for high-content fluorescence imaging of PBMCs. Here, we describe the protocol, modified from the original Cell Painting assay (Gustafsdottir et al. 2013, Bray et al. 2016, Cimini et al. 2023). We performed BCP on PBMC samples from 390 healthy blood donors, resulting in a dataset of approximately 50 million imaged cells. The dataset was created with the goal of characterizing and understanding morphological variability of immune cells in healthy individuals. To investigate whether specific immune cell types are preferentially observed in Cell Painting, we performed cell typing on morphological profiles and compared these to cell type frequencies obtained with flow cytometry, and, as a proof-of-concept, used the same methodology for the detection of acute myeloid leukemia (AML) specific cell populations in AML patients' samples. Furthermore, we contribute benchmarking of popular batch effect correction methods applied to HCI data as HCI is vulnerable to batch effects (Chandrasekaran et al. 2023, Arevalo et al. 2023, Way et al. 2023). We hypothesize that large HCI datasets of human PBMCs of healthy individuals as well as patients can be used to gain insight into the blood and immune system traits and, with other omics data, allow the discovery of novel genotype-phenotype associations in health and disease and understanding of the mechanistic underpinnings of these associations.

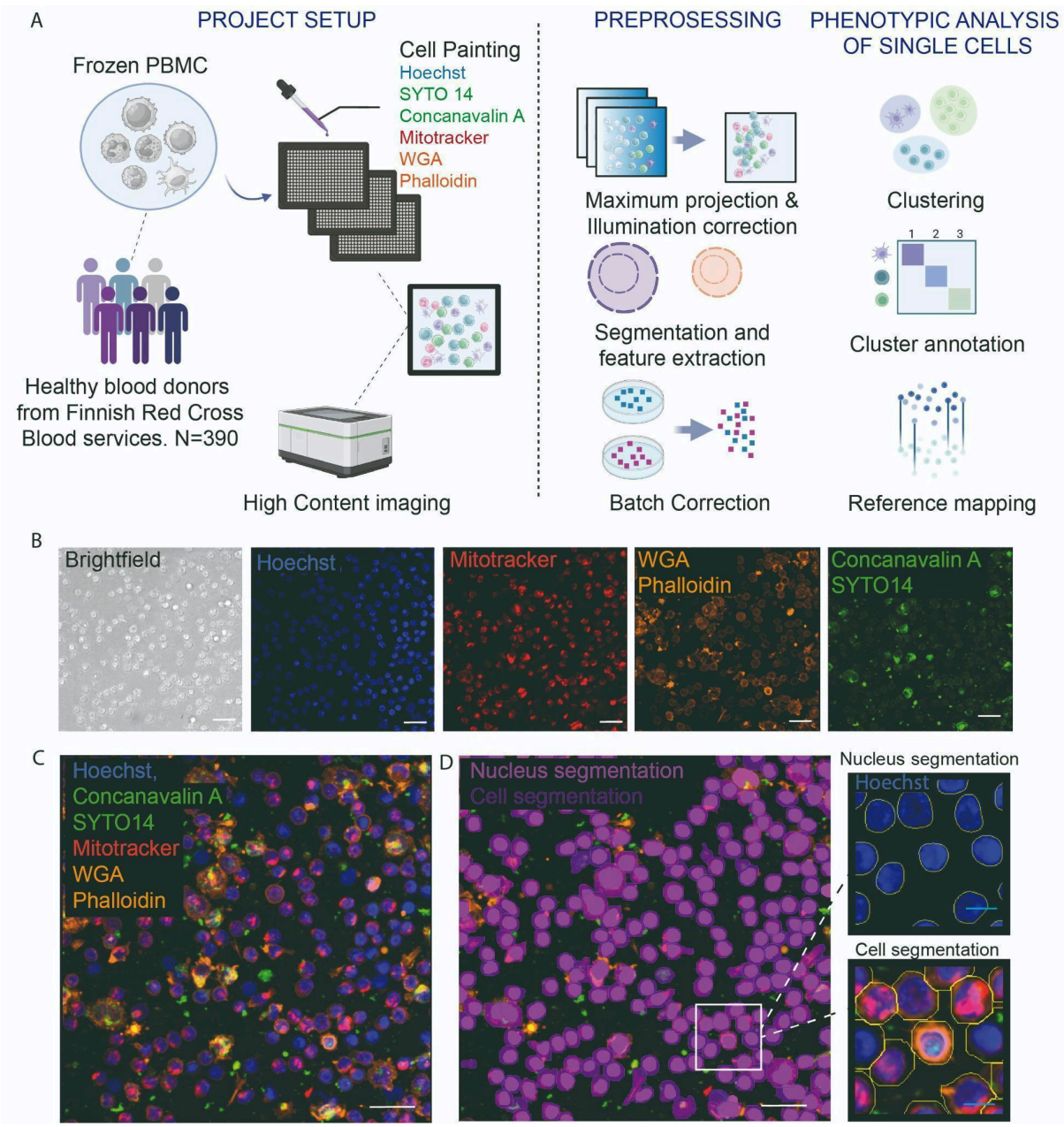

Figure 1. Blood Cell Painting assay and example images. *A) The laboratory workflow for the BCP assay and data analysis. B) Examples of randomly selected images of stained PBMCs using the BCP assay. C) An example channel merged image. D) Visualization of nucleus and cell segmentation carried out in BIAS software. Scale bars: white 20 µm and blue 6 µm.*

Methods

Healthy blood donor sample collection

The healthy blood donor samples profiled in this study originate from a FinnGen project in collaboration with the Finnish Red Cross Blood Service (FRC) Biobank. FRC Blood Service Biobank has participated in the FinnGen study by providing DNA samples from

over 58,000 blood donors. This project aimed to gain insight into the clinical, functional, and metabolic consequences of germline variants enriched in the Finnish population. Collected samples were live frozen PBMCs, plasma and serum samples from blood donors genotyped earlier in FinnGen (Kurki et al. 2023). Full blood count and clinical chemistry analyses were also performed from collected whole blood and serum. All biobank samples from FRC Blood Service Biobank were collected during blood donation from blood donors who met the blood donation health requirements and who had given biobanking consent (Clancy et al. 2024).

After blood donation, collected blood was transported overnight to the production facility in Blood Service and processed within 24 hours. The red blood cells and plasma were separated and collected from the donated blood. The remaining Buffy Coat fraction was passed to isolate PBMCs to the Blood Service Biobank. PBMCs were isolated from the Buffy Coat with Ficoll isogradient centrifugation (Noble & Cutts 1967, Böyum 1968). After the isolation, PBMCs were stored in a freezing medium (5% filtered heat-inactivated human AB serum (Valley Biomedical), 2 mM L-glutamine (Gibco), and 1% penicillin/streptomycin (10,000 units/mL and 10,000 µg/mL respectively) (Gibco) in RPMI 1640 (Gibco) containing 10% DMSO (WAK-chemie medical GMBH)) in cryotubes (1.9ml Tri-coded Tube, FluidX 65-7641, Azenta) and transferred into the automated gas phase liquid nitrogen freezer for long-term storing at -180 C until release and transported to the research laboratory.

Blood Cell Painting protocol – overview of the procedure

Cell culture, staining and fixation of PBMC samples

Steps involved in the modified Cell Painting protocol for PBMC samples are described below, with a more detailed protocol of the methods found in the **Supplementary Text**.

The media used for thawing and culturing the cells was a 5% filtered heat-inactivated human AB serum (Valley Biomedical), 2 mM L-glutamine (Gibco), and 1% penicillin/streptomycin (10,000 units/mL and 10,000 µg/mL respectively) (Gibco) in RPMI 1640 (Gibco). Thawing of the samples was completed according to standard protocols. The cells were washed twice to remove cytotoxic DMSO and resuspended in culture media to contain approximately 50,000 cells/ml. Prior to the seeding of the cells, the wells of 384-well plates (CellCarrier-384 Ultra/PhenoPlate 384-well, Revvity) were coated with Poly-L-lysine hydrobromide (MP Biomedicals) to enhance electrostatic interactions between the negatively charged ions on the cell membrane to the positively charged coated well (Jacobson and Branton 1977).

Each imaged plate contained ten donor samples and a control sample. As control samples, we used sister samples from a large cohort of PBMCs collected from donors during a singular blood donation. Samples collected from a singular donor were used on 11–17 plates as controls. With the aim to image at least 100,000 cells per donor, approximately 50,000 cells per well were added in parallel to five wells on 384-well

imaging plates. One control sample was added on each plate to five wells, aiming for at least 100,000 imaged cells per control. The cells were incubated at 37°C for 1.5 hours prior to adding Mitotracker (see below). After a complete two-hour incubation time at 37°C and 5% CO_2, the plate was fixed with 8% paraformaldehyde.

Fluorescence staining of healthy donor PBMCs

The staining pipeline used was based on an optimized Cell Painting assay (Bray et al., 2016, Cimini et al. 2023). The mitochondria of the living PBMCs were stained with Mitotracker Deep Red (Invitrogen) before fixation of the cells. Wheat-germ agglutinin/Alexa Fluor 555 (WGA) conjugate stock (Invitrogen), Concanavalin A/Alexa Fluor 488 conjugate stock (Invitrogen), Phalloidin/Alexa Fluor 568 conjugate stock (Invitrogen), SYTO14 green fluorescent nucleic acid stain stock (Invitrogen) and Hoechst 33342 stock (Invitrogen) were added after fixation and permeabilization of the cells, described in detail in **Supplementary Text**.

High content imaging with Opera Phenix microscope

The plates were imaged with Opera Phenix High Content Screening System (Revvity, Waltham, MA, USA) using 40x water immersion objective (NA 1.1) and four lasers (405 nm, 488 nm, 561 nm and 640 nm). A total of 64 (8x8) fields-of-view (FOV) with 5% overlap were imaged per well using five predetermined Z focus planes with laser-based autofocusing. The images were captured with two Andor Zyla sCMOS cameras (16-bit, field of view 320x320 μm^2, effective xy resolution 0.33 µm). Both confocal and widefield images were acquired. Detailed description of excitation and emission sources are described in **Supplementary Table 3**.

Image analysis

Biology Image Analysis Software (BIAS) (http://single-celltechnologies.com/bias-2/; Single-Cell Technologies) was used for primary image processing, segmentation, and feature extraction (Mund, 2022). Both confocal and widefield images were processed using the same analysis workflow with the exception that widefield images had one extra channel used in feature extraction (**Supplementary Text**). 3D image stacks were first projected into 2D images using maximum intensity projection for fluorescence images and average intensity projection for brightfield images. Non-uniform illumination was corrected on these images using the CIDRE method (Smith et al. 2015). Next, nuclear and cellular regions were segmented using two pre-trained deep learning models. Nuclei were segmented using the nucleAIzer (Hollandi et al. 2020) model and the cellular regions using the nucleAIzer extension for cytoplasm segmentation (available at: https://zenodo.org/records/6790845). These two segmented regions

were linked as cell objects including one segmented region from both models. Classical intensity, shape, and texture features were extracted from all channels on cell object regions to present cellular features.

To ensure that cellular features from different plates are comparable, features were standardized plate-wise with respect to cellular features of the control sample on the plate. First, for each feature, we calculated mean μ and standard deviation σ for control cells. Then, the donor cellular feature values x were standardized as $z = \frac{(x-\mu)}{\sigma}$.

Batch correction of cellular features

Three methods were employed for the batch correction of cellular features from BIAS software: Harmony (Korsunsky et al. 2019), pyComBat (Behdenna et al. 2023) and least-squares linear regression. Harmony clusters data points into temporary clusters, computes correction factors based on the cluster centroids, and repeats it until satisfactory entropy and diversity in the clusters have been reached. PyComBat is a Python implementation of the batch correction method ComBat (Johnson et al. 2007), which uses parametric and non-parametric empirical Bayes approaches to correct batch effects. We evaluated three potential sources of batch effect, plate, well, and FOV, and their combinations. For linear regression, corrected features were derived by subtracting the observed feature value from the feature value predicted by the correction model.

We used a batch effect score to evaluate the effectiveness of the above methods by measuring whether cells from the same plate are more clustered together than expected, indicative of batch effect. The score is derived by computing, for each cell i in the data, the fraction of cells from the same plate as cell i amongst k nearest cells, and normalizing by the expected fraction. To find the nearest cells, Euclidean distance in the UMAP space of cellular features was used. For each batch correction experiment, let N be the number of cells in the whole dataset, n^p be the number of cells from the plate p, and the number of cells from the same plate in the nearest k neighbors be x_i where i is the cell id. We define the batch effect score as

$$y = \frac{1}{N}\sum_{i}^{N}\frac{z_i^p}{m^p}\ ,\ where\ m^p = \frac{n^p}{N}\ and\ z_i^p = \frac{x_i^p}{k},$$

or the average of the logarithm of observed to expected fractions taken over all cells. The batch effect score for an individual cell is zero if the two ratios match, and increases when there are more cells from the same plate than expected, suggesting the presence of a batch effect. The ability of the score to detect batch effects depends on the choice of

k, such that larger values of k both allow the detection of smaller batch effects and accommodate larger between-plate differences in cell amounts.

Flow cytometry analysis

Cell populations were profiled using high-throughput flow cytometry (HTFC; FIMM High Throughput Biomedicine unit, HiLIFE, University of Helsinki) (**Suppl. Table 4**). The cell type markers used were CD45 (leukocyte), CD11c (dendritic cell), CD3 (T cell), CD8 (cytotoxic T cell), CD19 (B cell), CD56 (NK cell) and CD14 (monocyte/macrophage), as well as non-viable cell exclusion dye DRAQ7. Individual cell types for comparison to Cell Painting features were gated following the protocol described in **Supplementary Figure 1**.

Visualization of BCP cellular features

Cellular features are visualized by reducing feature dimensionality to two with Uniform Manifold Approximation and Projection (UMAP) (Mclnnes et al. 2018). UMAP is a non-linear dimensionality reduction algorithm that aims to preserve the similarities and dissimilarities between data points while reducing the feature space dimensionality. Input cellular features to UMAP were normalized to zero mean and standard deviation of one. The used UMAP parameters were n_neighbors=100 and min_dist=0.01. In order to plot millions of data points, we used Datashader (Datashader v0.16.1., 2024, April 19, https://datashader.org/) library to accurately represent the data distributions by aggregating local data points to avoid overplotting. In UMAP plots, local regions with multiple colors (labels) appear darker than regions with fewer colors.

Acute myeloid leukemia (AML) patient samples and MOLM-13 cell line sample

To test BCP on PBMCs from leukemia, we used six patient-derived AML (AML1-AML6) samples from five different patients (**Suppl. Table 7**). Two of the AML samples were collected from the same patient, with a six-month interval between collections. The AML samples were previously clinically subtyped as acute myeloid leukemia with minimal maturation (M1, n=3), acute myeloid leukemia with maturation (M2, n=1) or acute monocytic leukemia (M5, n=2) in the FAB (French-American-British) classification system (Bennett et al. 1976). The two AML M5 samples are from the same patient at diagnosis (AML5) and refractory (AML6). In addition, a MOLM-13 cell line (ACC 554) acquired from the FIMM HTB unit, originally established from the peripheral blood of a patient at relapse of AML (FAB M5a, Matsuo et al. 1997) was used in the study. All AML patient samples and the MOLM-13 sample were thawed, plated, fixed, stained, imaged and analyzed using the same BCP protocol as for the healthy donor PBMC samples.

Results

Morphological profiling of 48.7 million PBMCs from 390 healthy blood donors with Blood Cell Painting

We developed a modified Cell Painting protocol for PBMCs denoted here as Blood Cell Painting (BCP; **Fig. 1 A**; **Methods**). The work was optimized and designed so that one person could complete a wet lab work for up to four plates during a week (one plate per day), and perform sample staining and microscopy during the last day of the work week. The plates were then imaged over the weekend using automated microscopy, with one plate taking approximately 10 hours to image.

We performed BCP HCI on PBMCs in 400 samples from 390 healthy blood donors. For each sample, we aimed to image at least 100,000 cells dispersed in five wells on the same plate. Altogether, 41 plates were imaged resulting in a dataset of a total of 48.7 million cells in 4.74 million images excluding control samples. We acquired confocal images with four fluorescent channels and widefield images with five fluorescent channels using six fluorescent dyes to label different components of the cell in addition to brightfield images (**Fig. 1 B-D**). Approximately 35 TB of data was generated (530 GB/confocal plate, 320 GB/widefield plate).

Cell typing of PBMCs based on flow cytometry and Blood Cell Painting morphological profiles

We set out to identify cell types of PBMCs in our BCP dataset. For this purpose, we first clustered cells of 20 donors by their morphological profiles and assigned cell types with a supervised learning procedure where two experts annotated cell types for a subset of the cells to train a neural network classifier (**Methods**). Cells were categorized into four classes according to the following criteria (**Fig. 2 A**): 1) monocytes, large cells characterized by a kidney-shaped nucleus (bi-lobed structure). These cells exhibited a bright signal in the ER/Golgi channel (green) and the mitochondrial marker (red). The actin marker (phalloidin) formed a substantial mesh around the nucleus (orange), 2) macrophages, similar to monocytes but with a more rounded nucleus, 3) B and T cells, small cells with a rounded nucleus showing low signal intensity in the ER/Golgi (green), mitochondrial (red) and actin channels (orange), and 4) uncategorised including cells that did not fit the morphologies described above, cell debris and dead cells. The uncategorised group also typically included cells exhibiting only nuclear signals or those cropped on the edges of the image. We then evaluated our annotated training set with K-fold cross-validation with ten folds (in BIAS software). More specifically, the sample data was divided randomly into ten parts, then nine of those parts were used for training and one was reserved for testing. The procedure was repeated ten times, each time reserving a different tenth for testing. This procedure resulted in cells classified as

B and T cells (38% of all cells), monocytes (16%), macrophages (14%) and other cells or debris (32%) (**Fig. 2 B, C**; **Fig. 3 B**; **Suppl. Fig. 2**).

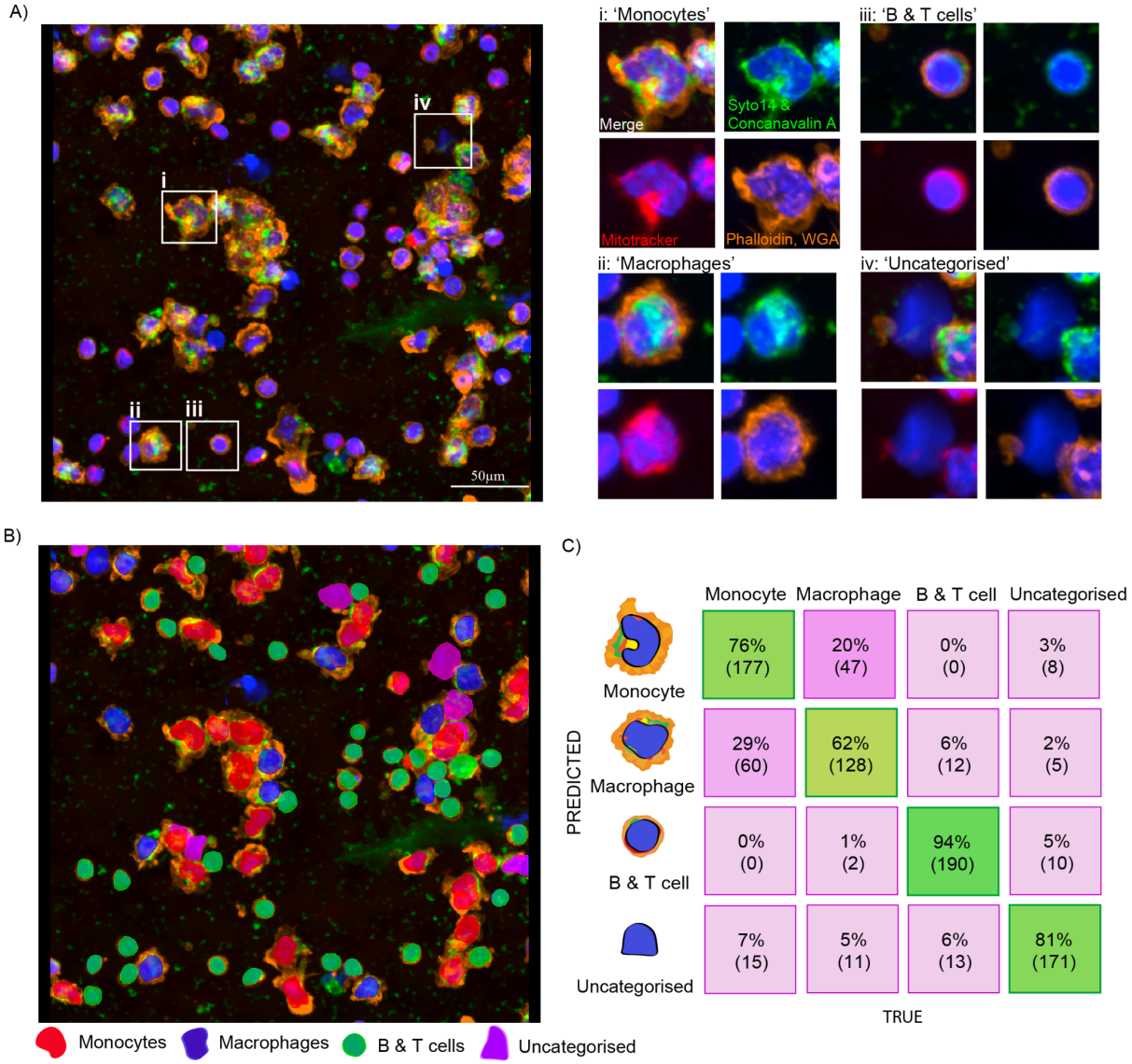

Figure 2. Cell classes based on morphology by expert annotations. *A) Examples of cell classes annotated for the training set. Cells were manually classified into four classes: monocytes (i), macrophages (ii), B & T cells (iii), and uncategorised (iv). B) Single-cell predictions from a representative field-of-view. Classification of cells into four classes is highlighted with the predicted class color: monocytes (red), macrophages (blue), B & T cells (green), uncategorised (magenta). C) Cross-validation confusion matrix obtained for the cell type classifier trained in BIAS software after 10-fold cross-validation. Correct predictions are shown in the diagonal and in green color. Scale bars: 50 um.*

To obtain estimates of cell type frequencies with an established method, we performed flow cytometry on the same 20 samples, and identified cell types by gating (**Methods**; **Suppl. Table 5**). In contrast to cell types assigned in BCP images, flow cytometry showed B and T cells to be two-fold more frequent (87%; 76% CD3+, 11% CD19+),

whereas monocytes were five-fold less frequent (3%) (**Fig. 3 B, D**; **Suppl. Fig. 3**). Of the CD3+ cells, 23% were CD8+ cytotoxic T cells. NK and dendritic cells were rare at 1% and 0.004%, respectively. Both BCP and flow cytometry data formed clusters in a two-dimensional UMAP projection distinguishing the cell types (**Fig. 3 A, C**). Many of the cell types left uncategorized in BCP formed distinct clusters likely corresponding to specific cell types, dead cells or debris. In general, the cell type fractions observed in flow cytometry corresponded to proportions reported in healthy individuals (Sender et al. 2023).

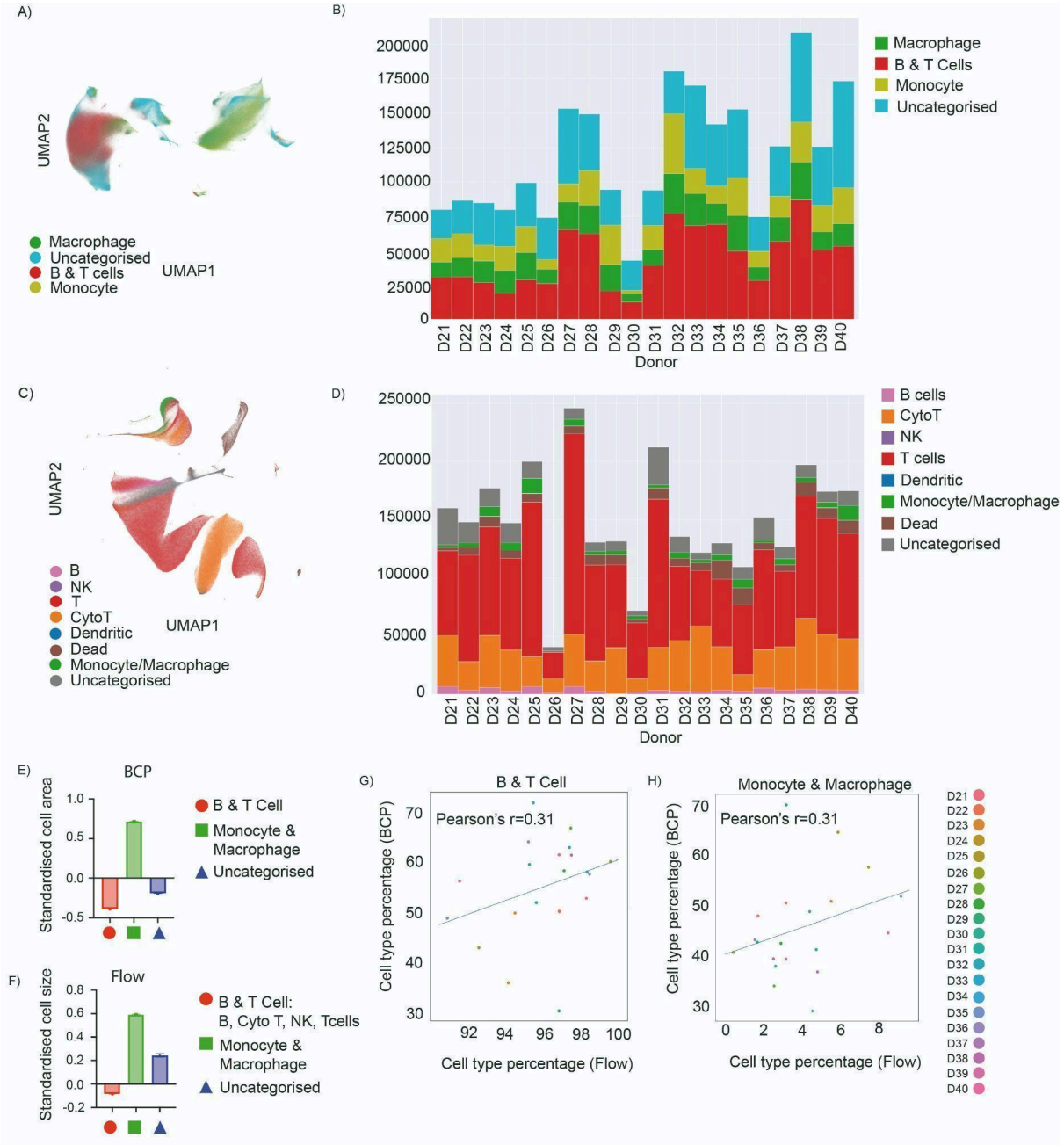

Figure 3. Comparison of BCP and flow cytometry data for 20 healthy blood donors. *A) UMAP representation of the BCP data. B) Frequency of cells by type in the BCP data. C) UMAP representation of the flow cytometry data. D) Frequency of cells by type in the flow cytometry data. E-F) Standard cell areas derived from BCP (E) and flow cytometry (F),*

respectively. Barplots show a mean with a 95% confidence interval. G) B and T cell, and H) macrophage and monocyte frequencies detected in BCP (Y-axis) and flow cytometry (X-axis), with respect to the total number of B, T, macrophage and monocyte cells.

Benchmarking of batch effect correction methods for Blood Cell Painting profiling data

We examined the quality of the whole BCP dataset of 48.7 million cells, which consisted of a median of 119,008 cells per donor and 23,682 cells per well (**Fig. 4 A, B**). Cellular features extracted from the BCP confocal images for the whole dataset of 390 donors revealed a degree of clustering by plate (**Fig. 4 D**), even after normalizing features for controls (**Methods**). This prompted us to examine the data for presence of batch effects by quantifying the degree of unexpected co-occurrence of cells from the same plate (batch effect score; **Methods**). We evaluated the data for batch effects before any batch correction, and after correction with two popular batch correction methods, Harmony and ComBat, as well as with linear regression. As batch factors to be corrected in the data, we used plate, well and FOV separately as well as combinations of these three factors.

We evaluated the performance of each method with a subset of the whole dataset containing 2.69 million cells from 269 donors in 27 plates (10,000 cells from each donor). As expected, the batch effect score was highest in data before any batch correction (**Fig. 4 C**). Harmony outperformed the other methods when both plate and well were used as the correcting factors, decreasing the batch effect score by 65.51% compared to 59.1% of ComBat and 8.4% of linear regression (**Fig. 4 C**; **Suppl. Table 6**). Similarly, two-dimensional representations of cellular features showed minimal clustering by plate in data corrected with Harmony (**Fig. 4 D**). Although we saw improvement when using both the plate and well as factors in Harmony and ComBat, we could not evaluate these methods with all three factors (plate+well+FOV) due to excessive memory requirements.

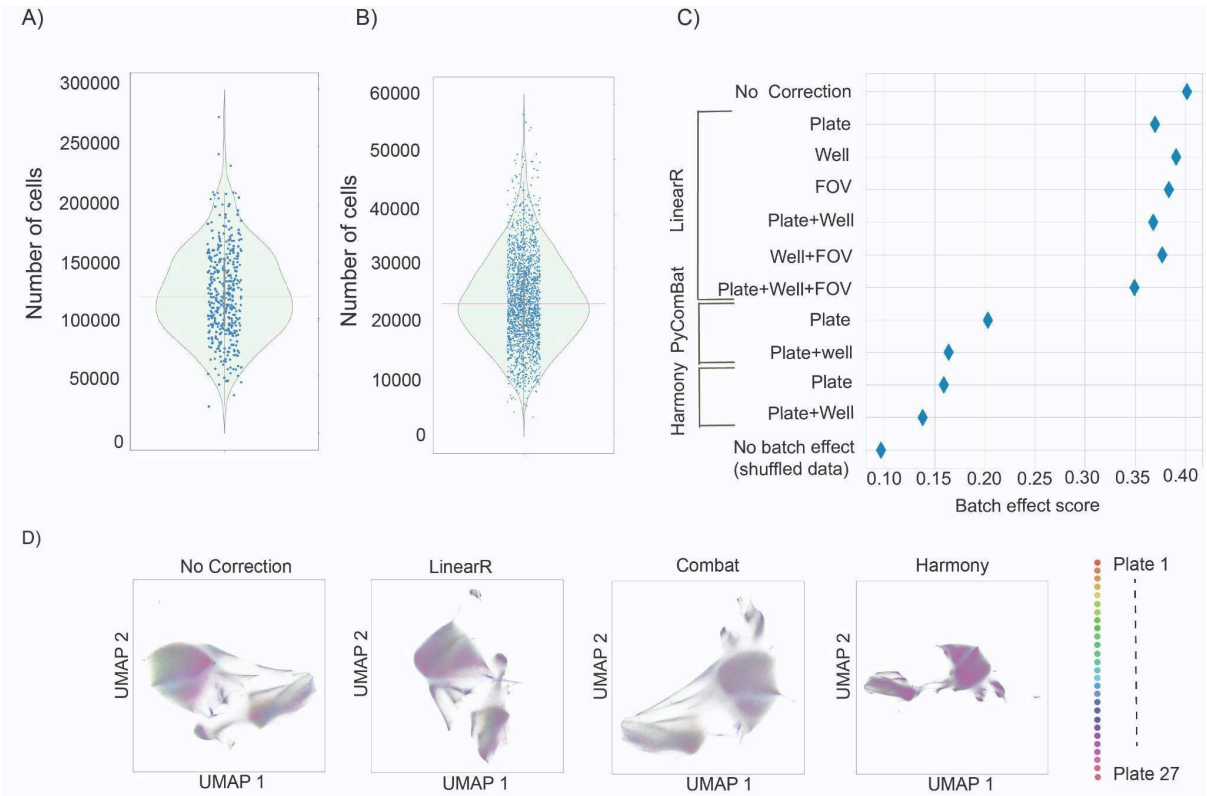

Figure 4. Quality control, and batch effect quantification and correction for BCP data. *Cell count distributions of A) 390 donors and B) 2000 wells (4,870,934 cells in total). C) Quantification of batch effect in uncorrected data, and in data corrected with linear regression (LinearR), ComBat, and Harmony for combinations of three batch factors (Plate, Well, FOV). 95% confidence intervals indicated with bars. D) Two-dimensional UMAP representations of cellular features before and after correction with the three methods. Colors indicate the source plate of each cell (27 plates in total).*

Blood Cell Painting profiling of AML

As a proof-of-concept study to test the generalization of the pipeline developed for healthy PBMCs, we performed the BCP protocol on six AML PBMC patient samples and one AML cell line sample (MOLM-13) (Methods, **Suppl. Table 7**). All samples were plated together with healthy PBMC samples, aiming at least 100,000 single cells dispersed over five wells, with the standard control sample and imaged with both confocal and widefield microscopy.

The cells from the AML samples formed distinct clusters in a two-dimensional representation of cells by their cellular features (**Fig. 5**). We observed co-clustering of cells from AMLs with minimal maturation (FAB M1, AML2-4) or maturation (FAB M2, AML1), with cells from these AMLs being mostly distinct from healthy cells. On the other hand, cells from acute monocytic leukemias (FAB M5; AML5-6, MOLM-13) formed distinct clusters but were part of a supercluster composed of monocytes.

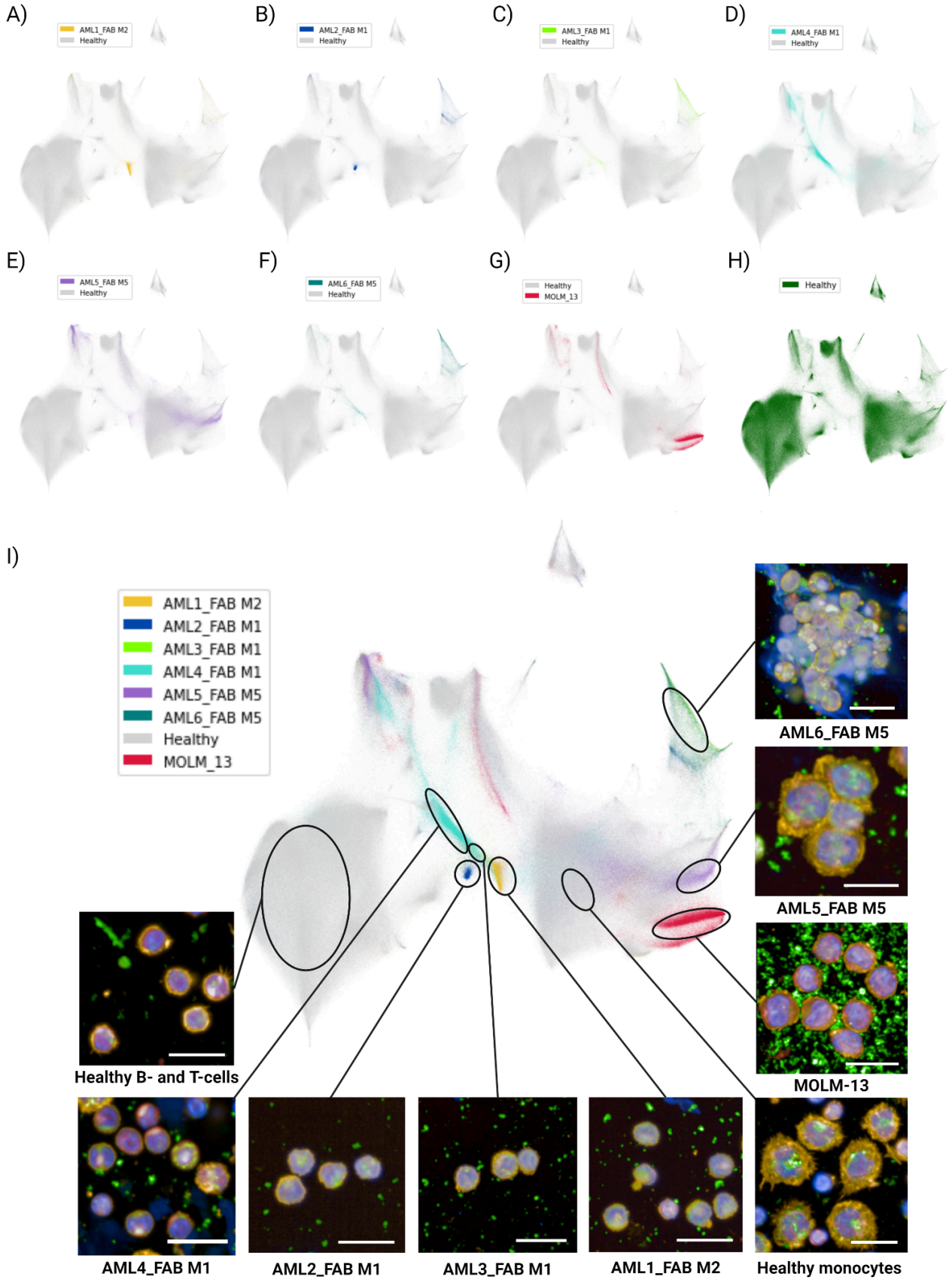
A)
AML1_FAB M2
Healthy
B)
AML2_FAB M1
Healthy
C)
AML3_FAB M1
Healthy
D)
AML4_FAB M1
Healthy
E)
AML5_FAB M5
Healthy
F)
AML6_FAB M5
Healthy
G)
Healthy
MOLM_13
H)
Healthy
I)
AML1_FAB M2
AML2_FAB M1
AML3_FAB M1
AML4_FAB M1
AML5_FAB M5
AML6_FAB M5
Healthy
MOLM_13
AML6_FAB M5
AML5_FAB M5
MOLM-13
Healthy B- and T-cells
AML4_FAB M1
AML2_FAB M1
AML3_FAB M1
AML1_FAB M2
Healthy monocytes

Figure 5. UMAP of BCP cellular features with example images of each sample. *A-H) AML patient samples 1-6 and the MOLM-13 sample compared to healthy PBMCs (H). I) Example images of each patient and cell line samples, as well as images of healthy B and T cells and monocytes. Scale bars: white 20 μm.*

Discussion

In this study, we aimed to characterize the morphological variability of mononuclear cells in healthy blood donors. To achieve this, we developed a modified Cell Painting (Bray et al. 2016) protocol, termed Blood Cell Painting (BCP), specifically for PBMCs, optimizing it for efficiency and throughput (**Fig. 1**). With BCP, we profiled what we believe to be one of the largest Cell Painting datasets of PBMCs to date, compassing 48.7 million imaged cells from 390 healthy individuals. This work adds a novel dimension to the study of the human immune system through blood profiling (Brodin & Davis 2017).

We anticipate that our extensive dataset will serve as a valuable resource for future research to delineate disease phenotypes by capturing the range of normal morphological variability and providing a reference for comparative analyses. For instance, our and similar future datasets could establish a baseline for healthy individuals in Cell Painting-based drug screens of pathologies of the blood, such as leukemias. Towards this goal, our proof-of-concept application of BCP on AML patient PBMC samples successfully differentiated between leukemia FAB subtypes (**Fig. 5**). The results revealed distinct morphological and cell type-specific differences within the PBMC samples, demonstrating BCP's potential for profiling leukemic PBMC samples. Nevertheless, comprehensive benchmarking is still necessary to establish BCP as a reliable tool for AML classification and profiling.

We assigned imaged cells with cell types using a supervised learning procedure based on expert assessment (**Fig. 2**). We found that monocytes and macrophages were five-fold more numerous in BCP data than in flow cytometry likely due to these cells' adhesive properties compared to lymphocytes (**Fig. 3**). Monocytes and macrophages adhere to the bottom of the well, whereas B and T cells, being non-adherent (Rajaraman et al. 1977), are more easily washed away during the multiple washing steps of the protocol. Therefore it is important to coat the wells for imaging with a coating agent to ensure that most lymphocytes adhere to the well bottom, as the monocytes and macrophages naturally do, and are not washed away by the multiple washing steps. Our preliminary work towards cell typing with BCP cellular features could be further developed with a more comprehensive effort to be able to accurately assign labels also to less common cell types such as NK cells and dendritic cells.

HCI has been shown to exhibit technical variation stemming from experimental conditions (Arevalo et al. 2023, Way et al. 2023). Compatible with these earlier studies we found cells imaged on the same plate to display more similar cellular features than expected by chance (**Fig. 4**). This batch effect was present despite standardizing data

against control samples that were shared between plates. We then evaluated the effectiveness of two popular methods for batch correction, Harmony and ComBat, and simple linear regression in our BCP dataset. We found Harmony to outperform the other methods, similar to what was recently shown for Cell Painting (Arevalo et al. 2023) and scRNA-seq data (Tran et al. 2020) in comprehensive benchmarking studies.

HCI data can be used to provide insight into how molecular and health factors shape the functional and phenotypic properties of the human immune system (Severin et al. 2022), and how perturbations, such as by drugs, can affect these properties and blood cell function (Heinemann et al. 2022). For example, Rezaei et al. (2023) have shown Cell Painting to capture mitochondrial stress response upon drug treatment. Similar studies are being performed by the JUMP consortium (Chandrasekaran et al. 2023). By combining phenotypes and mechanisms of action, we will be better able to understand the effects of variants in terms of their resistance and vulnerability to both known drugs as well as investigational compounds. We hypothesize that Cell Painting can be leveraged to discover how the genotype–specific phenotypes can be perturbed by drugs.

Recent developments in deep machine learning allow training performant models on large-scale data, such as massive HCI datasets. We envision our BCP data to facilitate self-supervised training of very large encoder models (Bommasani et al. 2021, Pfaendler et al. 2023) for modeling morphological and functional characteristics of PBMCs, leading to the discovery of novel genotype-phenotype associations and blood cell trait determinants. Another exciting direction is training label-free models for inexpensive morphological profiling (Cross-Zamirski et al. 2022, Timonen et al. 2023) with large BCP datasets. Our results will pave the way for understanding the causes and consequences of hematological phenotypes in disease and health by contributing a large-scale morphological characterization of immune cells in healthy individuals.

Code availability

The code used to analyze the BCP dataset, and dataset details will be available at https://github.com/bioimage-profiling/blood-cell-painting-paper.

Data availability

Information on the availability of the BCP data generated in this study will be updated in the subsequent versions of this manuscript.

Ethical approval

The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa have approved the FinnGen study protocol Nr HUS/990/2017. A formal biobank decision for the research project has been granted by the Finnish Red Cross Blood

Service Biobank. Blood was collected from volunteering blood donors that previously have partaken in a FinnGen study and have signed a biobanking consent form with the Blood Service Biobank (The Finnish Biobank Act 688/2012). The primary cells used in this project were human PBMCs, isolated from venous whole blood, and frozen according to the Blood Service Biobank Biobank guidelines and sent to FIMM for thawing, cell seeding, fixation, staining and imaging.

AML samples were collected in a study approved by an ethical committee of the Helsinki University Hospital (Ethical Committee Statement 303/13/03/01/201, latest amendment 7 dated 15 June 2016; latest HUS study permit HUS/395/2018 dated 13 February 2018; Permit numbers 239/13/03/00/2010, 303/13/03/01/2011).

Author contributions

Blood Cell Painting Assay development: CHS, MP, VP

Blood Cell Painting Assay data modeling: GA, VT, LP, EP

Methodology: CHS, VT, IM, MP, GA, AH, LP, VP, EP

Formal analysis: CHS, VT, GA, IM

Investigation: CHS, VT, GA, IM, MP, JS, LP, VP, EP

Visualization: MP, VT, GA, IM, CHS

Wetlab: CHS, JH, JS, TR, TA

Data curation: VT, GA, CHS, IM, MP

Software: GA, VT, IM

Resources: JP, AH, JJ, JJM, MA, CAH, RR, MD, AP, LP, VP, EP

Imaging development: AH, MP, CHS

Flow cytometry: TR, TA

Conceptualization, Supervision, Funding acquisition: LP, VP, EP

Acknowledgements

This study was supported by funding from the Research Council of Finland (322675 and 328890 to EP; 340273 and 346604 to LP), FinnGen and iCAN Flagship. The Finnish Red Cross Blood Service Biobank personnel are acknowledged for organizing research project-specific sample collection in blood donation activities. The authors wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources. HC imaging and analysis were performed at the FIMM High Content Imaging

and Analysis- HCA unit, and flow cytometry at the FIMM High Throughput Biomedicine-HTB unit, both supported by HiLIFE, University of Helsinki, Biocenter Finland, and Euro-Bioimaging. Prof. Peter Horvath is acknowledged for his support in image analysis and for providing BIAS software (Single Cell Technologies Ltd.) for our use. Prof. Olli Kallioniemi is acknowledged for scientific discussion. BioRender.com was used in creating the figures. We thank all the donors and patients who have participated in this project by donating their samples.

We want to acknowledge the participants and investigators of the FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis AG, and Boehringer Ingelheim International GmbH. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank (https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank). All Finnish Biobanks are members of BBMRI.fi infrastructure (https://www.bbmri-eric.eu/national-nodes/finland/). Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB.

References

Arevalo, J., Su, E., van Dijk, R., Carpenter, A. E., & Singh, S. (2023). Evaluating batch correction methods for image-based cell profiling. *bioRxiv*.

Behdenna, A., Colange, M., Haziza, J., Gema, A., Appé, G., Azencott, C. A., & Nordor, A. (2023). pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC bioinformatics*, *24*(1), 459.

Bennett, J. M., Catovsky, D., Daniel, M. T., Flandrin, G., Galton, D. A., Gralnick, H. R., & Sultan, C. (1976). Proposals for the classification of the acute leukaemias French-American-British (FAB) co-operative group. *British journal of haematology*, *33*(4), 451-458.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boyum, A. (1968). Isolation of mononuclear cells and granulocytes from human blood. *Scand j clin lab invest*, *21*, 77-89.

Bray, M. A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., ... & Carpenter, A. E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, *11*(9), 1757-1774.

Brodin, P., & Davis, M. M. (2017). Human immune system variation. *Nature reviews immunology*, *17*(1), 21-29.

Caicedo, J. C., Arevalo, J., Piccioni, F., Bray, M. A., Hartland, C. L., Wu, X., ... & Singh, S. (2022). Cell Painting predicts impact of lung cancer variants. *Molecular biology of the cell*, *33*(6), ar49.

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., Arevalo, J., Bennion, M., ... & Carpenter, A. E. (2023). JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *Biorxiv*, 2023-03.

Cimini, B. A., Chandrasekaran, S. N., Kost-Alimova, M., Miller, L., Goodale, A., Fritchman, B., ... & Carpenter, A. E. (2023). Optimizing the Cell Painting assay for image-based profiling. *Nature protocols*, *18*(7), 1981-2013.

Clancy, J., Ritari, J., Vaittinen, E., Arvas, M., Tammi, S., FinnGen, ... & Partanen, J. (2024). Blood donor biobank as a resource in personalised biomedical genetic research. *European Journal of Human Genetics*, 1-9.

Cross-Zamirski, J. O., Mouchet, E., Williams, G., Schönlieb, C. B., Turkki, R., & Wang, Y. (2022). Label-free prediction of cell painting from brightfield images. *Scientific reports*, *12*(1), 10001.

Gustafsdottir, S. M., Ljosa, V., Sokolnicki, K. L., Anthony Wilson, J., Walpita, D., Kemp, M. M., ... & Shamji, A. F. (2013). Multiplex cytological profiling assay to measure diverse cellular states. *PloS one*, *8*(12), e80999.

Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E., & Singh, S. (2022). High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature methods*, *19*(12), 1550-1557.

Heinemann, T., Kornauth, C., Severin, Y., Vladimer, G. I., Pemovska, T., Hadzijusufovic, E., ... & Snijder, B. (2022). Deep morphology learning enhances ex vivo drug profiling-based precision medicine. *Blood Cancer Discovery*, *3*(6), 502-515.

Hollandi, R., Szkalisity, A., Toth, T., Tasnadi, E., Molnar, C., Mathe, B., ... & Horvath, P. (2020). nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Systems*, *10*(5), 453-458.

Jacobson, B. S., & Branton, D. (1977). Plasma membrane: rapid isolation and exposure of the cytoplasmic surface by use of positively charged beads. *Science*, *195*(4275), 302-304.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118-127.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., ... & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, *16*(12), 1289-1296.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., ... & Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494-498.

Matsuo, Y., MacLeod, R. A. F., Uphoff, C. C., Drexler, H. G., Nishizaki, C., Katayama, Y., ... & Orita, K. (1997). Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins (11; 9)(q23; p22p23). *Leukemia*, *11*(9), 1469-1477.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Noble, P. B., & Cutts, J. H. (1967). Separation of blood leukocytes by Ficoll gradient. *The Canadian Veterinary Journal*, *8*(5), 110.

Pfaendler, R., Hanimann, J., Lee, S., & Snijder, B. (2023). Self-supervised vision transformers accurately decode cellular state heterogeneity. *Biorxiv*, 2023-01.

Piccinini, F., Balassa, T., Szkalisity, A., Molnar, C., Paavolainen, L., Kujala, K., ... & Horvath, P. (2017). Advanced cell classifier: user-friendly machine-learning-based software for discovering phenotypes in high-content imaging data. *Cell systems*, *4*(6), 651-655.

Rajaraman, R., Fox, R. A., Vethamany, V. G., Fernandez, L. A., & MacSween, J. M. (1977). Adhesion and spreading behaviour of human peripheral blood mononuclear cells (PBMC) in vitro. *Experimental Cell Research*, *107*(1), 179-190.

Ramezani, M., Bauman, J., Singh, A., Weisbart, E., Yong, J., Lozada, M., ... & Neal, J. T. (2023). A genome-wide atlas of human cell morphology. *bioRxiv*.

Rohban, M. H., Singh, S., Wu, X., Berthet, J. B., Bray, M. A., Shrestha, Y., ... & Carpenter, A. E. (2017). Systematic morphological profiling of human gene and allele function via cell painting. eLife 6.

Santos, R., Ursu, O., Gaulton, A. et al.(2017). A comprehensive map of molecular drug targets. Nat. Rev. Drug Discovery, *16*(1), 19-34.

Sender, R., Weiss, Y., Navon, Y., Milo, I., Azulay, N., Keren, L., ... & Milo, R. (2023). The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences*, *120*(44), e2308511120.

Severin, Y., Hale, B. D., Mena, J., Goslings, D., Frey, B. M., & Snijder, B. (2022). Multiplexed high-throughput immune cell imaging reveals molecular health-associated phenotypes. *Science Advances*, *8*(44), eabn5631.

Smith, K., Li, Y., Piccinini, F., Csucs, G., Balazs, C., Bevilacqua, A., & Horvath, P. (2015). CIDRE: an illumination-correction method for optical microscopy. *Nature methods*, *12*(5), 404-406.

Snijder, B., Vladimer, G. I., Krall, N., Miura, K., Schmolke, A. S., Kornauth, C., ... & Superti-Furga, G. (2017). Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. *The Lancet Haematology*, *4*(12), e595-e606.

Solomon, M. J., & Varshavsky, A. (1985). Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences*, *82*(19), 6470-6474.

Stubbington, M. J., Rozenblatt-Rosen, O., Regev, A., & Teichmann, S. A. (2017). Single-cell transcriptomics to explore the immune system in health and disease. *Science*, *358*(6359), 58-63.

Sundaresh, A., Meistermann, D., Lampela, R., Yang, Z., Woldegebriel, R., Ganna, A., ... & Kilpinen, H. (2023). Characterization of cortical neurodevelopment in vitro using gene expression and morphology profiles from single cells. *bioRxiv*, 2023-12.

Timonen, V. A., Kerkelä, E., Impola, U., Penna, L., Partanen, J., Kilpivaara, O., ... & Pitkänen, E. (2023). DeepIFC: virtual fluorescent labeling of blood cells in imaging flow cytometry data with deep learning. *Cytometry Part A*, *103*(10), 807-817.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology*, *21*, 1-32.

Way, G. P., Sailem, H., Shave, S., Kasprowicz, R., & Carragher, N. O. (2023). Evolution and impact of high content imaging. *Slas Discovery*, *28*(7), 292-305.

White, B., Komalo, B., Nicolaisen, L., Donne, M., Marsh, C., DeVay, R. M., ... & Kamens, B. (2020). A multi-phenotype system to discover therapies for age-related dysregulation of the immune response to viral infections. *bioRxiv*, 2020-07.

Willis, C., Nyffeler, J., & Harrill, J. (2020). Phenotypic profiling of reference chemicals across biologically diverse cell types using the cell painting assay. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, *25*(7), 755-769.