Image-based profiling of healthy donor immune cells with Blood Cell Painting reveals novel genotype-phenotype associations

Authors

Caroline Högel-Starck[1,2*], Veera A. Timonen[1,3*], Gantugs Atarsaikhan[1,2], Isabel Mogollon[1,**], Minttu Polso[1,2**], Antti Hassinen[1], Jarno Honkanen[4,6], Cindy David Sarmento[1], Jessica Koski[2,3,9], Julius Soini[1], Lea Urpa[1], Tanja Ruokoranta[1,5], Toveann Ahlnäs[1], Julianna Juvila[4], Juho J. Miettinen[1], Rodosthenis S. Rodosthenous[1], FinnGen, Outi Kilpivaara[2,3,9,10,11], Mikko Arvas[8], Caroline A. Heckman[1,2], Jukka Partanen[8], Mark Daly[1,7], Aarno Palotie[1,7], Hanna M. Ollila[1,12,7,13], Lassi Paavolainen[1,2,†], Vilja Pietiäinen[1,2,†], Esa Pitkänen[1,2,3,†]

Affiliations

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

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

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

8) Finnish Red Cross Blood Service, Research and Development, Helsinki, Finland.

9) Department of Medical and Clinical Genetics, Medicum, Faculty of Medicine, University of Helsinki, Helsinki, Finland

10) HUSLAB Laboratory of Genetics, HUS Diagnostic Center, Helsinki University Hospital, Helsinki, Finland

11) K. Albin Johansson Cancer Research Fellow, Foundation for the Finnish Cancer Institute, Helsinki, Finland

12) Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

13) Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

* These authors contributed equally to this work.

** 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, Blood Cell Painting, peripheral blood mononuclear cell, high content imaging, morphology, phenotypic analysis, healthy blood donors, acute myeloid leukemia, cell typing, flow cytometry, genome-wide association study, health registry data

Abstract

The morphological diversity of blood immune cells of healthy individuals, critical for recognizing disease-related phenotypes, remains largely uncharacterized. To address this gap, we developed Blood Cell Painting (BCP): a high content, high throughput fluorescence imaging assay for peripheral blood mononuclear cells. We generated a BCP Atlas with images of 50 million cells from 390 healthy blood donors, identifying 18 distinct immune cell morphology clusters. A genome-wide association study of BCP-derived imaging-based cellular features revealed 93 significant associations across 30 genetic loci. These loci include genes linked to mast cell function, inflammation, immune signaling, mitochondrial maintenance and circadian immune modulation. We also observed correlations between immune cell morphological features and clinical traits, such as respiratory conditions and healthcare visits related to contraceptive management, potentially reflecting hormonal influences on immune cell phenotypes. As a proof of concept for clinical application, acute myeloid leukemia subtypes were distinguished by BCP. Our study establishes BCP as a versatile method for immune cell profiling to uncover genetic, phenotypic and clinical determinants of immune cell morphology in health and disease.

Background

Despite the success of population-scale genome-wide (GWAS) and phenome-wide association studies (PheWAS)[1], the biological mechanisms linking germline variants to phenotypes remain poorly understood. A crucial intermediate between genotypes and disease phenotypes lies at the level of cellular phenotypes, which can be interrogated at the single-cell level, for example, with single-cell RNA-sequencing (scRNA-seq) and high content imaging (HCI). 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[2].

Sequencing-based approaches allow the identification of cell types and subtypes, and mapping of their evolutionary trajectories[3], but these methods are costly and do not capture rich cellular morphology or spatial architecture. By contrast, HCI allows millions of cells to be routinely profiled and quantified in a single experiment, providing cellular context and morphological details at a lower cost[4,5,6]. Consequently, HCI has been used to reveal how immune cell phenotypes reflect molecular and personal health information and how leukemic cancer cells respond, at single-cell level, to different drugs in clinical trials[7,8].

Cell Painting is a widely adopted HCI assay for morphological and functional profiling of single cells and cell organelles via fluorescence microscopy[5,6]. Cellular phenotypes derived with Cell Painting have been associated with somatic variants in cancer, transcriptomics, as well as genetic and chemical perturbations[9,10,11]. For example, recent large-scale studies have combined Cell Painting with transcriptomics for >28,000 perturbations[12] and performed >20,000 CRISPR perturbation experiments in >30 million cells, resulting in an atlas of biological pathways and interactions[13]. To date, Cell Painting has primarily been applied to study the adherent cancer cell lines and human fibroblasts[5,14], with less focus on human immune cells playing a central role in health and disease.

In this study, we developed the Blood Cell Painting (BCP) workflow (**Fig. 1**) for high content fluorescence microscopy imaging of peripheral blood mononuclear cells (PBMCs), modified from the original Cell Painting assay[5,6]. Using BCP, we profiled PBMC samples from 390 healthy blood donors, yielding 50 million imaged cells. We optimized and benchmarked a popular batch effect correction method Harmony for large-scale BCP data. Finally, we associated BCP profiles to genotypes and health registry phenotypes, revealing novel links to immune cell morphology and function.

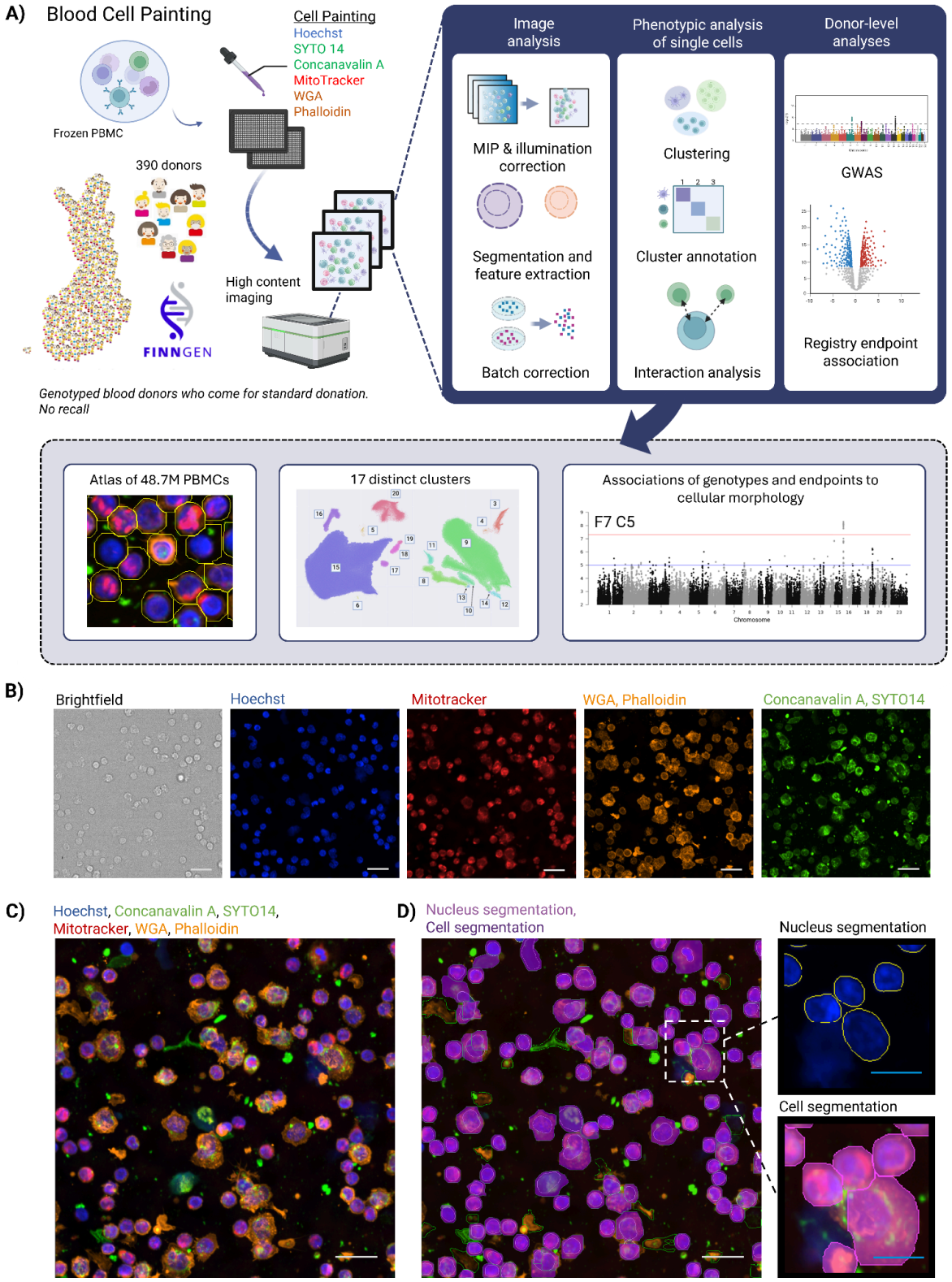

Figure 1. Blood Cell Painting — study design and the BCP assay. *A) Overview of the workflow and study results. Created with Biorender.com. B) Example images of PBMCs stained using the BCP assay. C) Example of merged image of four channels containing stains Hoechst 33342, MitoTracker Deep Red, Wheat-germ agglutinin/Alexa Fluor 555 (WGA), Phalloidin/Alexa Fluor 568, Concanavalin A/Alexa Fluor 488 and SYTO14 green*

fluorescent nucleic acid. D) Visualization of nucleus and cell segmentation. Scale bars: white 20 μm and blue 10 μm.

Results

Morphological profiling and cell typing of 50 million PBMCs from 390 healthy blood donors with Blood Cell Painting

We developed a modified Cell Painting protocol for PBMCs, termed Blood Cell Painting (BCP; **Fig. 1 A**; **Methods**). Using the BCP protocol, we profiled PBMCs in 400 samples from 390 healthy blood donors. For each sample, we aimed to image at least 100,000 cells distributed across five wells on the same plate. Altogether, 41 plates were imaged, yielding a total of 50 million cells in 4.74 million images after excluding control samples. We acquired confocal images with four fluorescent channels and widefield images with five fluorescent channels using six fluorescent dyes (**Fig. 1 A**; **Methods**) to label different components of the cell. In addition, brightfield images were collected in both imaging modalities (**Fig. 1 B-D**). A total of 186 features were extracted from images for downstream analyses.

To identify major PBMC cell types in our BCP dataset, we clustered single-cell morphologies from 20 donors and performed supervised cell type annotation where two experts labeled a subset of the cells to train a neural network classifier (**Methods**). Cells were assigned into four classes: 1) monocytes, 2) macrophages, 3) B and T cells, and 4) uncategorised (**Fig. 2 A**). We then evaluated our annotated training set with ten-fold cross-validation (**Methods**). This procedure resulted in cells classified as B and T cells (38% of all cells), monocytes (16%), macrophages (14%) and uncategorised cells or debris (32%) (**Fig. 2 B, C**; **Suppl. Fig. 3 B**; **Suppl. Fig. 2**).

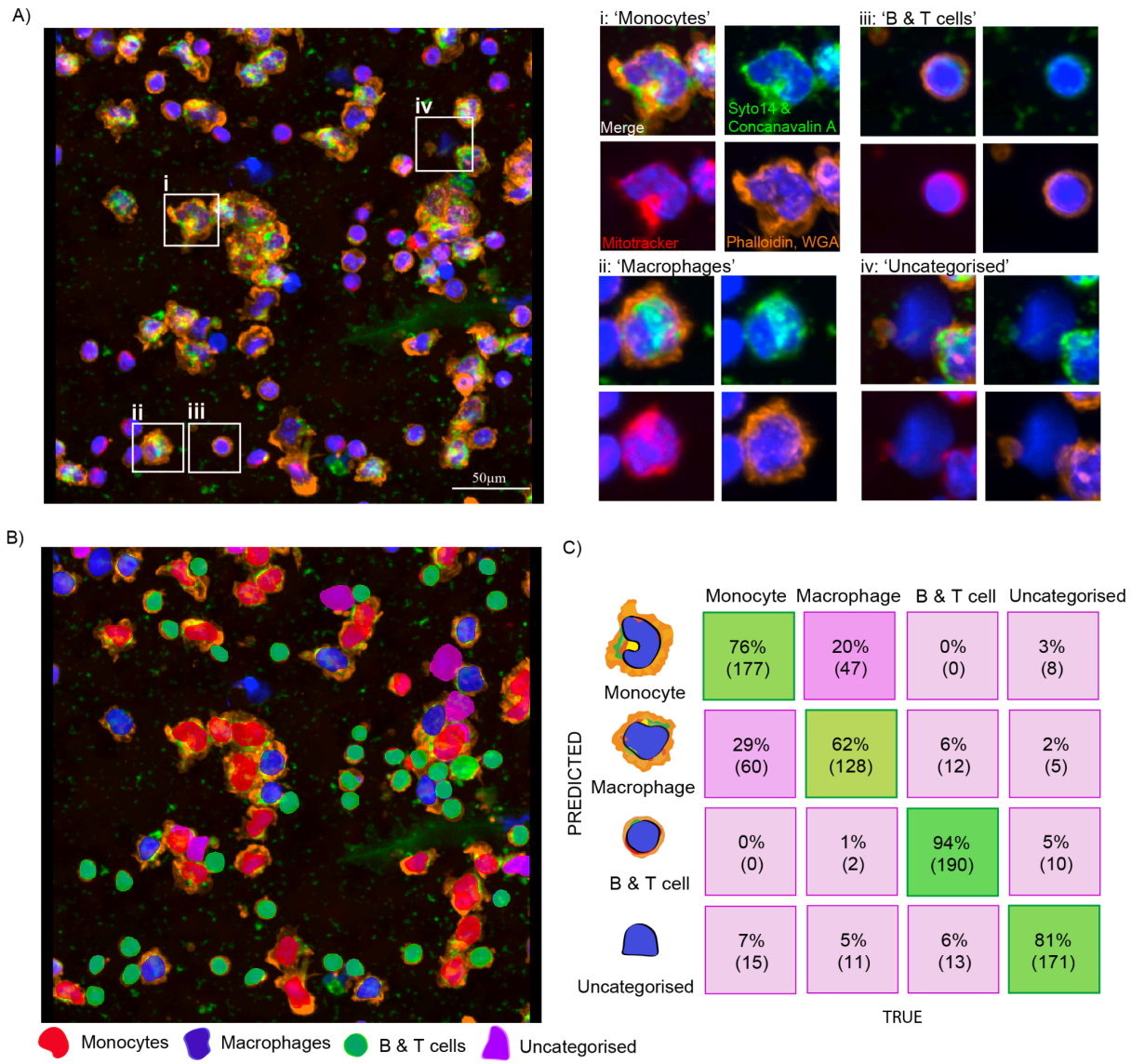

Figure 2. Cell classes based on morphology by expert annotations. *A) Examples of cell classes annotated for the training set. Cells were manually annotated into four classes: monocytes (i), macrophages (ii), B & T cells (iii), and uncategorised (iv). B) Single-cell predictions from a representative 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 µm.*

To obtain independent estimates of cell type frequencies, we performed flow cytometry-based immunophenotyping on the same 20 samples (**Methods**; **Suppl. Table 5**). In contrast to the BCP-based typing, flow cytometry indicated B and T cells to be two-fold as frequent (87%; 76% CD3+, 11% CD19+), and five-fold fewer monocytes (3%) (**Suppl. Fig. 3 B, D**; **Suppl. Fig. 4**). Among T cells (CD3+), 23% were CD8+ cytotoxic T cells. NK and dendritic cells were rare (1% and 0.004%, respectively). Despite these differences, both BCP and flow cytometry data showed clear clustering in

UMAP embeddings (**Suppl. Fig. 3 A, C**). Notably, many of the cell types that remained uncategorized in BCP formed distinct clusters, suggesting they correspond to specific cell types, dead cells or debris (**Suppl. Fig. 5**). In general, the cell type fractions observed in flow cytometry corresponded to proportions reported in healthy individuals with the same methodology[15]. Cell size estimates derived from BCP and flow cytometry (**Suppl. Fig. 3E–F**) reflected known size differences among PBMC subtypes[16,17]. These observations across platforms further support the validity of our cell type assignments using BCP. Finally, we found that cell type proportions in BCP were moderately correlated with flow cytometry for both B & T cells and monocyte/macrophage populations (r=0.31; **Suppl. Fig. 3G,H**), indicating general concordance between the two methods despite biology- or platform-specific biases.

Benchmarking batch effect correction methods for Blood Cell Painting profiling data

We next examined the quality and batch effects in the full BCP dataset of 50 million cells. Each donor contributed a median of 119,008 cells, and each well contained a median of 23,682 cells (**Fig. 3 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. 3 D**), even after normalizing features for controls (**Methods**). This prompted us to quantify batch effects as the degree of unexpected co-occurrence of cells from the same plate (batch effect score; **Methods**). We evaluated the data for batch effects before and after correction using Harmony[18], ComBat[18,19] and linear regression. As batch factors to be corrected in the data, we used plate, well and field-of-view (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). Harmony outperformed the other methods when both plate and well were used as the correcting factors, decreasing the batch effect score by 65.5% compared to 59.1% of ComBat and 8.4% of linear regression (**Fig. 3 C**; **Suppl. Table 6**). Consistently, two-dimensional UMAP representations of cellular features showed minimal clustering by plate in data corrected with Harmony (**Fig. 3 D**). Although we saw improvement when using both 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. Harmony batch-corrected data was used for all subsequent analyses.

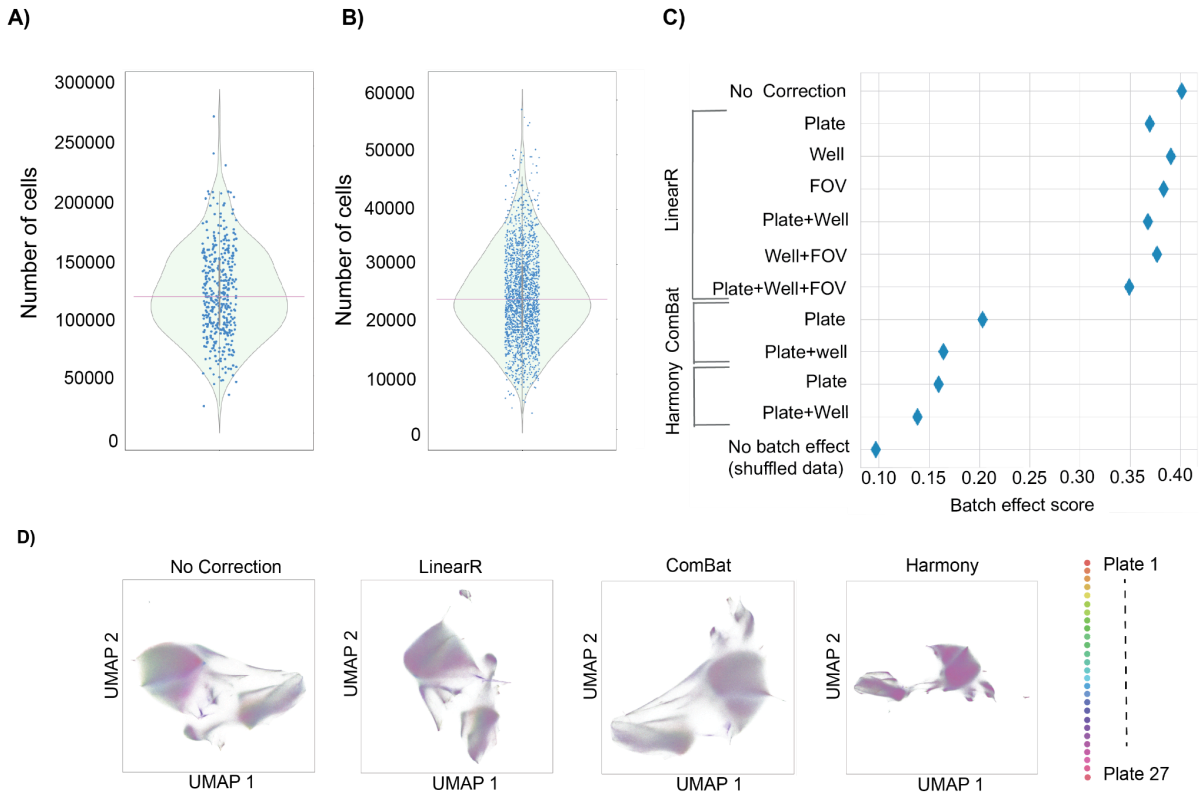

Figure 3. Quality control, and batch effect quantification and correction for BCP data. *Cell count distributions of A) 390 donors and B) 2000 wells (4,870,934 cells in total). C) Quantification of batch effect in uncorrected data, and in data corrected with linear regression (LinearR), ComBat, and Harmony 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 Atlas of peripheral blood mononuclear cells from healthy donors

We combined all 390 donor samples after batch correction to create, to our knowledge, the largest morphological atlas of PBMCs. Cell clustering by morphological profiles yielded 18 distinct clusters encompassing a total of 42.5 million cells (clusters C3-C20; **Fig. 4 A**; **Suppl. Table 8**; **Methods**). Applying our morphology-based cell type classifier to the BCP Atlas revealed the two largest cell clusters to be primarily composed of lymphocyte cells (C15, 58% of all cells), and monocytes and macrophages (C9, 29%), respectively (**Fig. 4 B**). Remaining smaller clusters consisted mostly of uncategorized cells (13%).

Visual examination of BCP images of cells from each cluster confirmed marked morphological differences between clusters (**Fig. 4 C**, **Suppl. Fig. 5**). Several clusters were enriched for cells with C-shaped or otherwise non-spherical nuclei (C3–6, C9, C18, C19). Cluster C11 consisted of cells with elongated shapes, potentially due to interacting cells or active cell division. In cluster C12, we observed larger, possibly damaged or

dying cells with strong actin staining. Intriguingly, cells in cluster C13 showed a striking semicircular pattern on the mitochondrial channel wrapping around a large portion of the nucleus. This pattern may reflect perinuclear mitochondrial aggregation, a localized accumulation of mitochondria around the nucleus[20].

We next examined whether cell-cell interactions could be observed in BCP. For each cell, we quantified the relative enrichment (RE) of neighboring cells of each type within a 22 μm radius (**Methods**). We found that each cell type was significantly enriched in neighbors of the same type (RE>1; $p<1x10^{-15}$, one-way ANOVA; **Suppl. Fig. 6**). In contrast, other cell types were less frequent within the interaction radius (RE<1). We also quantified the RE in each of the 18 cell clusters and observed that monocytes had a RE slightly above 1 for cluster C9 cells (**Suppl. Fig. 5**), which were earlier classified mostly as monocytes and macrophages (**Fig. 4**). Similarly suggestive of self-enrichment, B/T cell interactions were found to be slightly more frequent in cluster C15 (**Suppl. Fig. 5**).

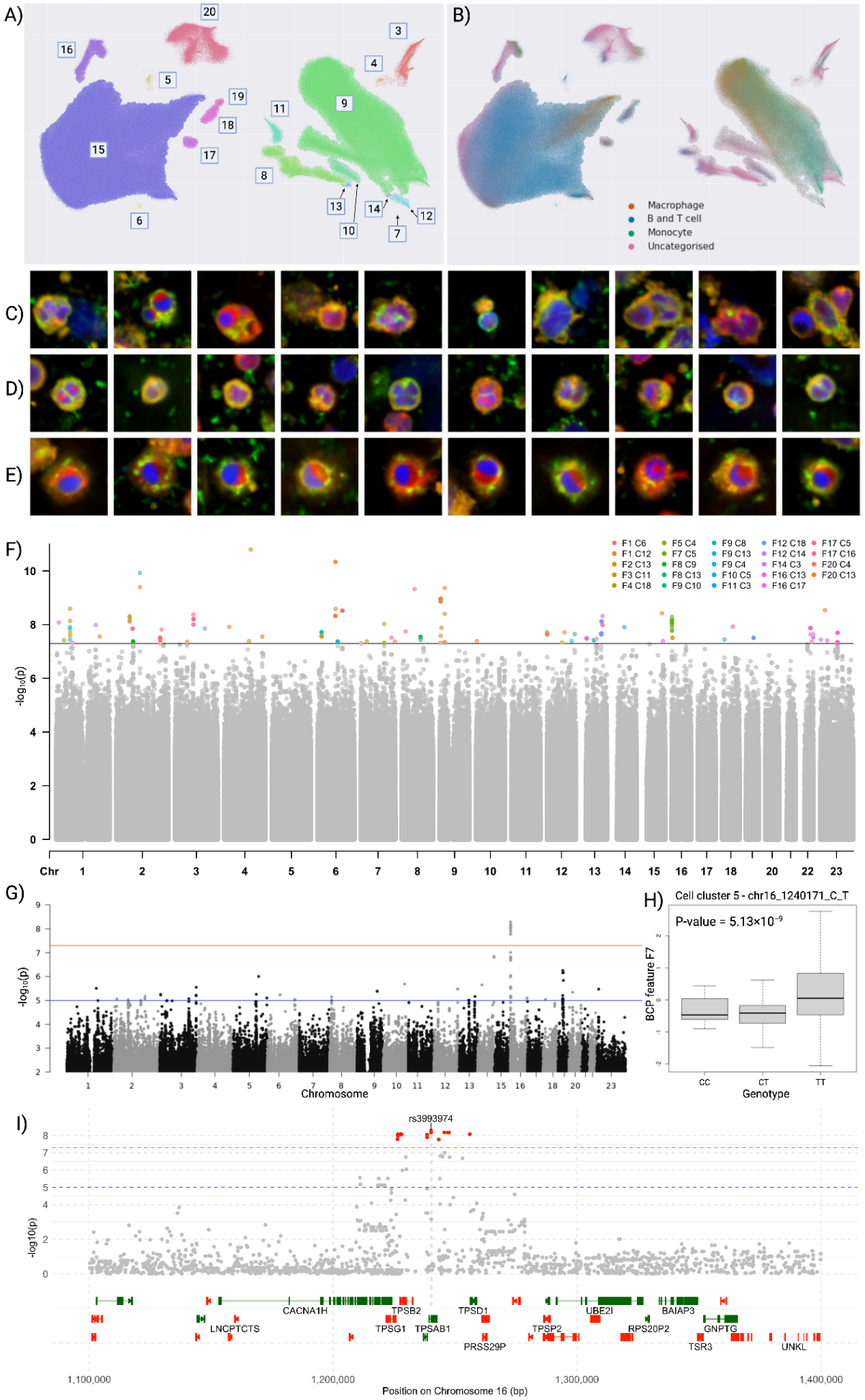
A)
B)
Macrophage
B and T cell
Monocyte
Uncategorised
C)
D)
E)
F)
-log10(p)
Chr
Chromosome
G)
-log10(p)
Chromosome
H)
Cell cluster 5 - chr16_1240171_C_T
P-value = 5.13×10⁻⁹
BCP feature F7
CC
CT
TT
Genotype
I)
rs3993974
-log10(p)
CACNA1H
TPSB2
TPSD1
UBE2I
BAIAP3
LNCPTCTS
TPSG1
TPSAB1
TPSP2
RPS20P2
GNPTG
PRSS29P
TSR3
UNKL
1,100,000
1,200,000
1,300,000
1,400,000
Position on Chromosome 16 (bp)

Figure 4. GWAS reveals distinct cell phenotypes are significantly associated with variants. *UMAP of the BCP Atlas colored by A) cell clusters C3-C20 and B) cell type identified by a supervised classification procedure. Cell image examples from clusters C) C3, D) C5 and E) C13. F) A combination Manhattan plot of all 93 genome-wide significant associations (-log$_{10}$(p)), where all significant associations for BCP features and clusters are highlighted. G) Genotype-phenotype associations for phenotype F7 in cell cluster C5 (-log$_{10}$(p)), H) Association between lead variant rs3993974 and phenotype F7 in cell cluster C5. I) A locus zoom plot at rs3993974.*

Genome-wide association study of Blood Cell Painting features reveals 30 loci influencing immune cell morphology

To discover germline genetic variants influencing immune cell morphology, we performed a GWAS on BCP image-derived features in healthy blood donors. As quantitative traits, we used twenty principal components of donor-averaged BCP features (F1-F20; **Methods**) in 17 cell clusters, capturing major axes of morphological variation, to discover associations of cellular characteristics with germline genetic variants. We analyzed ~10 million variants (BCP donor minor allele frequency, MAF>1.25%, at least ten carriers in donors) in 368 genotyped donors (FinnGen R12)[1]. This yielded 106 genome-wide significant associations (p<5x10^{-8}) (**Suppl. Fig. 7**), of which 93 passed quality control (**Methods**), spanning 30 genomic loci (**Fig. 4 f**, **Suppl. Table 9).** Fine-mapping with SuSiE[21] implicated high-confidence variants in genes related to transcriptional regulation (*C1orf21, ZMYM2*), inflammatory signaling and tissue remodeling (*CHI3L1*), cell cycle and circadian regulation (*CSNK1E*), vesicle trafficking and secretion (*PTPRN2, NCALD*), and extracellular matrix interactions (*PDGFRL*) (**Suppl. Table 12; Methods**).

A prominent association was found in cluster C5 at a locus containing tryptase-encoding genes *TPSB2*, *TPSAB1* and *TPSD1*. The lead variant (rs3993974, GWAS p=5.13x10^{-9}; SuSiE PIP=0.11; **Suppl. Table 10**, **Suppl. Table 12, Fig. 4 h, i**) is strongly associated with increased tryptase levels in plasma (tryptase µg/l, β=0.23, p=4.63x10^{-11}; **Methods**; **Suppl. Table 11**). This region had a strong expression and splicing quantitative trait locus (eQTL, sQTL) signal for tryptase genes (**Suppl. Table 13**). In addition, this region was marked by direct transcription factor binding by CTCF and nine additional transcription factors, as assessed by ChIPseq (**Suppl. Table 14**). Tryptases are involved in asthma and other allergic and inflammatory disorders, and are contained in mast cells[22]. The associated BCP feature (F7) reflected cell length and nuclear stain intensity (**Fig. 4 g, Suppl. Fig. 8 a, e, i; Suppl. Fig. 9).** Cluster C5 also showed associations for nuclear morphology (F17) with rs71330739 (p=4.32×10^{-8}) near cell adhesion and migration related *PARVG* and *PARVB*, and rs62486739 (p=3.07×10^{-8}; PIP=0.44) in *SSBP1*, a housekeeping gene involved in mitochondrial biogenesis (**Suppl. Fig. 8 b, f, j; Suppl. Fig. 9)**. These variants further support the notion that C5 includes rare or immature myeloid populations with distinct nuclear morphologies. While mature mast cells are

not generally present in peripheral blood, we speculate that this cluster may be composed of mast or granulocyte cell precursors.

In cell cluster C3, nucleus solidity, cavity and textural patterns (F14) were associated with rs965618 (p=2.28x10^{-8}; PIP=0.35) (**Suppl. Fig. 8 c, g, k; Suppl. Fig. 9)**. The lead variant is an expression quantitative trait locus (eQTL) across several tissues for *SCO2*, *CPT1B* and *CHKB* that are central regulators of immunity and lipid metabolism across multiple tissues (**Suppl. Table 13**). The locus was also strongly associated with increased mean corpuscular volume (Kanta β=0.02, p=4.0x10^{-27}) and monocyte counts (OpenTargets β=0.01, p=3.99x10^{-8}; **Suppl. Table 11**). In cell cluster C13, which harbored a striking phenotype possibly due to perinuclear mitochondrial aggregation discussed above (**Fig. 4e),** nuclear, mitochondrial and actin characteristics (F8) were associated with rs13251546 with a clear additive effect (**Suppl. Fig. 8 d, h, l; Suppl. Fig. 9)**. The region has previously been associated with visual disorders, varicose veins and smoking status[23].

As an example of variants previously implicated in immune-related conditions and highlighted by the BCP GWAS, we found psoriasis-linked variants (OpenTargets β=-3.6, p=5.00x10^{-15}; **Suppl. Table 11**) at an enhancer site near *CD28* to associate with BCP features (F20; rs115029131, p=3.21x10^{-8}). Many cells in the associating cluster C4 had a non-spherical nucleus, and might thus be monocytes or macrophages (**Suppl. Fig. 5**). Normally CD28 is not expressed in monocytes and macrophages, but can be induced under chronic inflammation[24].

Finally, we performed GWAS for all BCP phenotypes without cell clustering, but no genome-wide significant associations were found. A comparison of our results to previous GWASs[30,31,29] revealed positive correlations in association strengths (r=0.06–0.18, p<0.005) (**Suppl. Fig. 10**, **Suppl. Table 15, Suppl. Table 19**).

Association of BCP cell characteristics to donor age, sex and registry endpoints

We examined how demographic factors and health registry phenotypes relate to BCP morphological characteristics. Donor age and sex were found to associate (FDR<5%; **Methods**) with several features such as nucleus, mitochondria and RNA intensity and texture as well as cell shape (F1, F2, F7, F13, F16, F17, F20; **Suppl. Table 17**) in cell clusters C9 and C11 (**Suppl. Table 16**). C9 is composed mostly of monocytes and macrophages with strong actin intensity (**Suppl. Fig. 5**). In C11 consisting of elongated cells, sex was associated with cell shape, texture and length as well as RNA dye intensity (FDR=3%; F2).

Morphological features of cells in C19 were associated (FDR<10%; F5, F9, F17) with inflammatory and allergic endpoints such as asthma, sinusitis and pneumonia (**Suppl. Table 18**). These cells were also associated with abnormalities of breathing, septum deviation and vocal larynx issues. Nuclei in C19 exhibit a lobular structure, suggesting they may be granulocytes or their precursors that survived PBMC isolation. Morphological characteristics of cells in C8, potentially composed of T cells, were linked

to common rhinitis (F13). Interestingly, healthcare visits related to contraceptive management (ICD-10 Z30) were associated with elongated cell shapes (F16) in C11 (**Suppl. Fig. 5**).

BCP profiling of acute myeloid leukemia PBMC patient samples

As a proof-of-concept study to test the generalization of the pipeline developed for healthy PBMCs, we performed the BCP protocol on PBMC samples of six acute myeloid leukemia (AML) patients and an AML cell line (MOLM-13) (**Methods**, **Suppl. Table 7**). PBMCs from AML patients displayed distinct morphological profiles (**Fig. 5**). In AMLs with minimal maturation (FAB M1, AML 2-3) and acute monocytic leukemias (M5, AML 6), we observed a cell cluster not found in healthy cells nor the other AML samples. AML samples 4 (M1) and 5 (M5) had low levels of clinically malignant cells in the blood, indicated by most of their cells clustering among the healthy cells. Finally, the AML cell line (MOLM-13) forms distinct clusters within a supercluster composed of monocytes.

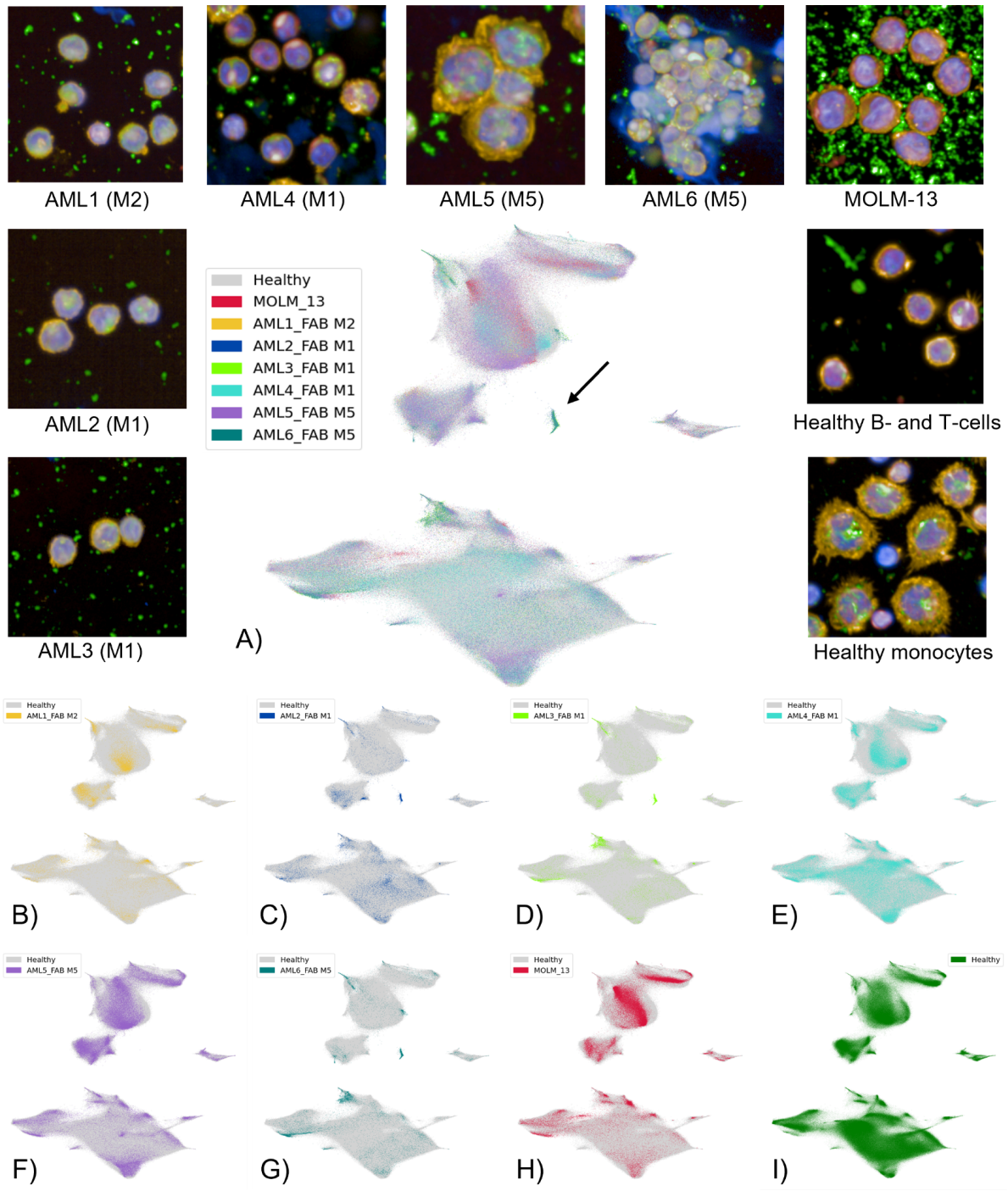

Figure 5. AML samples show a distinct distribution of BCP features compared to healthy donors. *A) UMAP of BCP cellular features from AML patient and cell line samples with health PBMCs. Example images of each patient and cell line samples, as well as images of healthy B and T cells and monocytes. A cell cluster not found in healthy cells is indicated with an arrow. B-G) AML patient samples 1-6 and (H) the MOLM-13 sample compared to (I) healthy PBMCs.*

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[5] protocol termed Blood Cell Painting (BCP), specifically for PBMCs, optimizing it for robustness and throughput. With BCP, we created the largest Cell Painting dataset of PBMCs to date, a BCP Atlas encompassing 50 million imaged cells from 390 healthy individuals. The large number of cells profiled allowed us to capture rare morphological phenotypes, such as perinuclear aggregation of mitochondria[33] in immune cells. This work adds a novel dimension to the study of the human immune system through blood profiling[34].

HCI has been shown to exhibit technical variation stemming from experimental conditions such as plate effects and cell handling[35,36]. In line with earlier studies[35,37], batch correction with Harmony outperformed alternatives, and we adapted it to scale across tens of millions of cells. Additionally, we observed biases in cell type representation, where monocytes and macrophages appeared more common than expected. The observed difference in cell frequency between flow cytometry and the BCP protocol likely stems from two related factors: higher concentration of B/T cells at the edges of the well and the non-adherent nature of these cell types[38], making them more susceptible to loss during washing compared to monocytes. Further optimization of Poly-L-lysine coating of the wells to improve cell adhesion can mitigate this issue. Our preliminary supervised cell type assignments suggest that further refinement, possibly including marker-based sorting, could improve accuracy for rare populations such as NK and dendritic cells.

Driven by the genotyping and health registry data available in FinnGen[1] for our study cohort, we were able to examine the determinants and correlates of immune cell morphological phenotypes. We discovered 30 novel genomic loci associated with morphological cell characteristics. The genetic architecture of immune cell morphology reflected variation in immune cell activation and differentiation, subcellular structural organization, and susceptibility to allergic and inflammatory conditions. As an example, we uncovered genetic variation at tryptase-encoding genes to strongly influence morphology of a distinct cluster of cells. Tryptases, the most abundant serine proteases stored in mast cell granules, are vital for mast cell function and activation, with a key role in allergic reactions and asthma[22]. While mature mast cells are not typically found in peripheral blood, immature mast cells can be present.

We also observed associations between BCP features and genetic variants previously implicated in psoriasis. For instance, a psoriasis-linked variant near *CD28* was associated with cells characterized by non-spherical nuclei. This morphology may reflect, for example, activated or differentiating monocytes and macrophages. These cells are known to undergo nuclear deformation in response to environmental stimuli and immune activation[39]. Systemic inflammation in psoriasis has been associated with

elevated levels of low-density granulocytes (LDGs), a neutrophil subset with segmented or irregular nuclei that co-fractionate with PBMCs[40]. These findings suggest that inflammatory skin disorders may leave a morphological footprint in the peripheral immune system, linking genotype-associated immune dysregulation with altered nuclear architecture.

Previous studies have identified numerous germline variants associated with immune cell traits measured by flow cytometry[25,26,27,28,29,30] and imaging[31], and immunologic endpoints[32]. Although the genome-wide significant variants found in this study were novel, we found variants with $p<10^{-5}$ in our study earlier associated with blood indices or morphological traits. The correlation of our GWAS results with reported variant associations from flow cytometry, bulk blood traits and high content imaging studies provides orthogonal validation for the ability of BCP to capture meaningful aspects of immune cell biology[30,31,29].

Compatible with our genetic findings, BCP morphological features were associated with inflammatory endpoints such as asthma, allergies and sinusitis. We also observed morphological associations in cells with elongated morphology to leukocyte counts and healthcare visits related to contraceptive management. However, a more careful examination of this association is warranted as the endpoint does not distinguish between use of specific contraceptives, and includes any contraceptive-related visits to healthcare. Similarly to previously reported age-related decline of mitochondrial activity[7], we found mitochondrial variability in these cells to correlate with donor age.

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 representing a reference for comparative analyses. Our and similar future datasets could establish a baseline for healthy individuals in Cell Painting-based drug screens of pathologies of the blood[41], such as leukemias. Towards this goal, our proof-of-concept application of BCP on AML patient PBMC samples differentiated between leukemia FAB subtypes. The results revealed distinct morphological and cell type-specific differences within AML samples, demonstrating the potential of BCP for profiling leukemic samples. Nevertheless, comprehensive benchmarking with a larger number of samples is essential for establishing the BCP as a reliable tool for AML classification and profiling towards diagnostics.

We foresee the training of label-free models with large BCP datasets as a promising direction for inexpensive morphological profiling[45,46]. Our publicly available BCP Atlas offers a foundation for label-free models[42,43,44] and large-scale biological discovery.

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 Biobank. The Blood Service Biobank contributed DNA samples from 58,000 blood donors to FinnGen. For a subset of these donors, live-frozen PBMCs, plasma, and serum samples were collected and stored, along with genotype data generated earlier in FinnGen[1]. Full blood count and clinical chemistry analyses were also performed from collected whole blood and serum. All biobank samples from Finnish Red Cross Blood Service Biobank were collected along with normal blood donation from donors who met the blood donation health requirements and who had given biobanking consent[47]. For blood donation, blood donors must be between 18 and 70 years of age, and feeling healthy and well. If a medical condition has been diagnosed, it must be well-controlled and medicated. Certain medications and medical conditions may disqualify potential donors from donating blood either temporarily or permanently.

Donated blood is collected in a standard blood bag system with Citrate Phosphate Dextrose (CPD) as anticoagulant (CompoFlow®, Fresenius Kabi, Cat.No. CQ32250). After blood donation, donated blood was transported overnight to the Blood Service's production facility and processed within 24 hours. The red blood cells and plasma were separated from the donated whole blood by centrifugation and used for preparation of blood products. The remaining Buffy Coat fraction (~55 ml) was passed for isolation of PBMCs to the Blood Service Biobank. PBMCs were isolated from the Buffy Coat with Ficoll isogradient centrifugation[48,49]. 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 (1.9 ml 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

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**.

Cell culture, staining and fixation of PBMC samples

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[50].

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[5,6]. 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 (1.5 µm distance between 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**. The experimental workflow was optimized and designed so that a single person could prepare up to four plates during a week (one plate per day), and perform sample staining during the last day of the work week. Plates were then imaged by automated microscopy over the weekend, with one plate taking approximately 10 hours to image. Approximately 35 TB of images were generated during data collection (530 GB/confocal plate, 320 GB/widefield plate).

Image analysis

Biology Image Analysis Software (BIAS) (https://www.sct.bio/the-bias; Single-Cell Technologies) was used for primary image processing, segmentation, and feature extraction[51]. 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[52]. Next, nuclear and cellular regions were segmented using two pre-trained deep learning models. Nuclei were segmented using the nucleAIzer[53] 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 186 BCP 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 $z = \frac{(x-\mu)}{\sigma}$.

Outlier and edge removal

Out of ~55 million cells imaged, approximately 5 million were from control samples and were excluded from analysis after being used for feature scaling (Methods). Outlier removal was done based on features "NUCLEUS INTENSITY-MEAN Alexa 568" and "NUCLEUS INTENSITY-MEAN Hoechst 33342", where all cells with scaled features with values over 40 were removed from the dataset. These were mostly falsely segmented high-intensity artefacts. In addition, edge cell removal was done based on values of features "centroid_x", "centroid_y" and "CELL ENCLOSING CIRCLE RADIUS", where cells were removed if they were touching the edge of the well to ensure that all analyzed cells fit completely in the field-of-view. These steps resulted in 44.8 million cells for downstream analysis and batch correction.

Batch correction of BCP cellular features

Three methods were employed for the batch correction of cellular features: Harmony[18], pyComBat[19] and least-squares linear regression. Harmony clusters data points into temporary clusters, computes correction factors based on the cluster centroids, and repeats until satisfactory entropy and diversity in the clusters have been reached. PyComBat is a Python implementation of the ComBat[54] batch correction method, which uses parametric and non-parametric empirical Bayes approaches to correct batch effects. We evaluated three potential sources of batch effect: plate, well, and

field-of-view (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

$$y = \frac{1}{N}\sum_{i}^{N} log_2 \frac{z_i^p}{m^p}\ ,\ where\ m^p = \frac{n^p}{N}\ and\ z_i^p = \frac{x_i^p}{k},$$

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. The value of k used for this study was 200.

We were not able to perform batch correction on the quality-controlled set of 44.8 million cells due to insufficient memory (1,490 GB RAM) to run Harmony. Therefore, we first ran Harmony on 35 million cells in a single batch, and then corrected the remaining cells as follows: for each remaining cell, feature value was assigned as the average of three Harmony-corrected cells closest with respect to Euclidean distance to the uncorrected cell in the feature space.

Clustering of cells by BCP cellular features

Clustering of cells by BCP features was done by first performing two-dimensional UMAP (n_neighbors=100, min_dist=0.01) and then applying DBSCAN clustering[55] on the UMAP coordinates. After clustering cells with DBSCAN, all cells not assigned to any cluster were removed as outliers, leaving eight initial clusters. After initial clustering, further subclustering was done based on visual inspection of major clusters that could be clustered further (two largest clusters C1 and C2) (**Fig. 4**). Parameters for subclusters were assessed by testing multiple different values for maximum distance between two points to be considered neighbours (ε) and minimum number of points within ε

(minPts). The parameters for initial clustering were ε=0.08 and 35 minimum samples. The parameters for subclustering of cluster C1 were ε=0.1 and 65 minimum samples, and for cluster C2 ε=0.09 and 60 minimum samples.

Cell typing

To classify single cells into immune subtypes, we used a supervised machine learning pipeline in the BIAS software. We manually annotated ~200 cells per class to train a Support Vector Machine (SVM) classifier, which was refined using an active learning loop. K-means clustering (iterations = 100, max deviance = 0) was applied to assist in stratified annotation and visualization of the feature space.

Cells were annotated to four categories: monocytes, macrophages, B & T cells, and uncategorised cells. The aim was to maintain class balance during annotation to support robust training and quantification. Annotation criteria were based on established morphologies in multiplexed confocal fluorescence images. Cells were assigned into the four categories as follows (**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 that have differentiated from monocytes in wells, 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 segmented objects that did not fit the morphologies described above, for instance cell debris, dead cells, and cells exhibiting only nuclear signals. All extracted intensity, texture, and morphology features were used as inputs without prior feature selection, enabling the model to determine optimal feature weighting.

The classifier was evaluated with ten-fold cross-validation: the dataset was partitioned into ten subsets, each in turn used as a test set while training on the remaining data. Classification performance was assessed using a confusion matrix that confirmed high accuracy for B & T cells and moderate separation between monocytes and macrophages. The final classifier model was then trained with the complete training dataset.

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 were visualized by reducing feature dimensionality to two with Uniform Manifold Approximation and Projection (UMAP)[56]. 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. CMplot version 4.5.1 was used to visualize multiple GWAS in the same plot.

Quantification of cell-cell interactions

To detect and quantify cell-cell interactions, we used the subset of 2,090,581 cells from 20 donors, for which cell typing and clustering data were available. The number of proximal cells, measured as the number of cells present within the interaction radius from the center cell, was computed using the KDTree nearest neighbor algorithm (scikit-learn v.1.1.1). The co-localization of immune cells within a 22-25 μm radius has been associated with specific immune processes and clinical variables, indicating biologically significant interactions between cells[57,58]. Therefore, given the average cell enclosing circle radius (36.33 pixels), we set the interaction distance to 146 pixels, equivalent to 22 μm or 2 cell diameters (0.15 μm/pixel image resolution). We also computed a measure of relative enrichment (RE) to determine if certain cell types were more prevalent in the set of proximal cells than expected. Let R be the total number of cells within the interaction radius r, let x_r^C be the number of cells of type C within the interaction radius r, let N be the total number of cells in the dataset, and n^C be the total number of cells of type C. We defined RE as:

$$RE = \frac{obs+1}{exp+1}, \; where\ obs = \frac{X_r^C}{R} \; and\ exp = \frac{n^C}{N}$$

Principal component analysis of BCP cellular features

Principal component analysis (PCA) was performed for 186 BCP cellular features in order to reduce data complexity and computational burden for subsequent analyses. PCA was done with 20 principal components (PCs) denoted F1–F20 which explained >95% of variance in the dataset. To obtain donor-level features, the average of each PC in each cell cluster was computed for each donor.

Genome-wide association study for cellular BCP features

Genome-wide association study (GWAS) was performed to find associations between BCP cell features and genotypes in FinnGen (R12). We utilized the FinnGen DF12 dataset with 500,348 donors, 2,502 disease phenotypes and over 21 million genotyped variants[1]. We considered 186 BCP features of the four fluorescent channels transformed to 20 PCs (**Suppl. Fig. 9**). GWAS was run for all combinations of the 17 cell clusters and 20 PCs. Cell cluster C7 was not included in the analysis as it contained cells from only 67 donors. The donor-level average value of each PC in each cell cluster was treated as phenotype in GWAS. Altogether 340 GWAS were run (20 phenotypes x 17 clusters). Out of 390 donors imaged, 22 did not have covariate data available or were otherwise excluded from analysis in earlier data collection of FinnGen due to quality control or ancestry[1]. The remaining 368 donors had 10.4 million germline variants present with minor allele frequency (MAF) > 0.0125 in FinnGen, and were tested against altogether 22 million germline variants available in FinnGen.

Rank inverse normalization transformation (RINT) was applied to features with inflation (λ>1.05) or deflation of lambda values (λ<0.98). Regenie model[59] was used to run GWAS and the model was trained with a high-quality LD-pruned genetic relationship matrix (GRM) available in FinnGen[1]. This GRM contains no population outliers or genetic duplicates, and was further pruned to only contain variants for which BCP donors had MAF>0.0125.

As GWAS covariates, age, sex and the first ten genetic ancestry principal components were used. A base model for associations was also investigated with cellular feature values as independent variable and age, sex and ancestry principal components as covariates.

Fine-mapping of associated variants in GWAS

Fine-mapping was done for all 30 genomic regions highlighted as significant in GWAS utilizing SuSIE[21] with the FinnGen R12 pipeline and default parameters (max_region_width=6,000,000; p_threshold=5×10^{-8}; n_causal_variants=10; window_shrink_ratio=0.9; good_cred_r2=0.25).

Comparing BCP GWAS results to previous studies

We combined summary statistics from previous GWASs with our BCP GWAS by variant rsIDs. For each previous study, the Pearson correlation coefficient r was computed between $-\log_{10}(p)$-values reported in the study and obtained from the BCP GWAS for variant associations (**Suppl. Fig. 10**). Confidence intervals for r were estimated using bootstrapping with 1,000 samples.

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 ten-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[60]. 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[61] (FAB M5a) 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. All AML 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.

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

BCP image data generated in this study will be made publicly available in BioImage Archive (accession number pending). Instructions on how to access FinnGen genotype and health registry data are available at https://www.finngen.fi/en/access_results.

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 guidelines and sent to FIMM for thawing, cell seeding, fixation, staining and imaging.

AML samples were collected in a study approved by an ethics committee of the Helsinki University Hospital (Ethics 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, CDS, IM

Investigation: CHS, VT, GA, IM, MP, JS, JK, CDS, LU, HMO, LP, VP, EP

Visualization: MP, VT, GA, IM, CDS, 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, IM

Flow cytometry: TR, TA

Genetic analyses: VT, LU, OK, HMO, EP

Registry data analyses: VT, JK, HMO, EP

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), the Research Fund of the Finnish Red Cross Blood Service, 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. We would like to thank Prof. Matti Pirinen, Pietro Della Briotta Parolo, Vincent Llorens, Dr. Markus Vähä-Koskela, Anna Nylund, and the FinnGen helpdesk for their assistance. Prof. Olli Kallioniemi is acknowledged for scientific discussions. BioRender.com was used in creating 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 Inc. (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., Johnson&Johnson Innovative Medicine Inc., Novartis AG, Boehringer Ingelheim International GmbH and Bayer AG. 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

1 Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., ... & Waring, J. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature, 613(7944), 508-518. https://doi.org/10.1038/s41586-022-05473-8

2 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. 10.1126/science.aan6828

3 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. https://doi.org/10.1038/s41586-018-0414-6

4 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. 10.1016/j.cels.2017.05.012

5 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. https://doi.org/10.1038/nprot.2016.105

6 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. https://doi.org/10.1038/s41596-023-00840-9

7 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. 10.1126/sciadv.abn5631

8 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. 10.1016/S2352-3026(17)30208-9

9 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. https://doi.org/10.1091/mbc.E21-11-0538

10 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. https://doi.org/10.7554/eLife.24060

11 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. https://doi.org/10.1101/2023.12.03.569583

12 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. https://doi.org/10.1101/2021.09.08.459417

13 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. https://doi.org/10.1038/s41592-024-02537-7

14 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. https://doi.org/10.1177/2472555220928004

15 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. https://doi.org/10.1073/pnas.2308511120

16 Abbas, A. K., Lichtman, A. H., & Pillai, S. (2014). Cellular and molecular immunology E-Book: Cellular and molecular immunology E-Book. Elsevier Health Sciences.

17 Palmer, L., Briggs, C., McFadden, S., Zini, G., Burthem, J., Rozenberg, G., ... & Machin, S. J. (2015). ICSH recommendations for the standardization of nomenclature and grading of peripheral blood cell morphological features. International journal of laboratory hematology, 37(3), 287-303. https://doi.org/10.1111/ijlh.12327

18 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. https://doi.org/10.1038/s41592-019-0619-0

19 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. https://doi.org/10.1186/s12859-023-05578-5

20 Al-Mehdi, A. B., Pastukh, V. M., Swiger, B. M., Reed, D. J., Patel, M. R., Bardwell, G. C., ... & Gillespie, M. N. (2012). Perinuclear mitochondrial clustering creates an oxidant-rich nuclear domain required for hypoxia-induced transcription. Science signaling, 5(231), ra47-ra47. https://www.science.org/doi/full/10.1126/scisignal.2002712

21 Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(5), 1273-1300. https://doi.org/10.1111/rssb.12388

22 Payne, V., & Kam, P. C. A. (2004). Mast cell tryptase: a review of its physiology and clinical significance. Anaesthesia, 59(7), 695-703. https://doi.org/10.1111/j.1365-2044.2004.03757.x

23 Khan, N. W., Wissinger, B., Kohl, S., & Sieving, P. A. (2007). CNGB3 achromatopsia with progressive loss of residual cone function and impaired rod-mediated function. Investigative ophthalmology & visual science, 48(8), 3864-3871. https://doi.org/10.1167/iovs.06-1521

24 Estrada-Capetillo, L., Aragoneses-Fenoll, L., Domínguez-Soto, Á., Fuentelsaz-Romero, S., Nieto, C., Simón-Fuentes, M., ... & Puig-Kröger, A. (2021). CD28 is expressed by macrophages with anti-inflammatory potential and limits their T-cell activating capacity. European Journal of Immunology, 51(4), 824-834. https://doi.org/10.1002/eji.202048806

25 Nalls, M. A., Couper, D. J., Tanaka, T., van Rooij, F. J., Chen, M. H., Smith, A. V., ... & Ganesh, S. K. (2011). Multiple loci are associated with white blood cell phenotypes. PLoS genetics, 7(6), e1002113. https://doi.org/10.1371/journal.pgen.1002113

26 Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., ... & Soranzo, N. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell, 167(5), 1415-1429. 10.1016/j.cell.2016.10.042

27 Chen, M. H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., ... & Lettre, G. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell, 182(5), 1198-1213. 10.1016/j.cell.2020.06.045

28 Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., ... & Soranzo, N. (2020). The polygenic and monogenic basis of blood traits and diseases. Cell, 182(5), 1214-1231. 10.1016/j.cell.2020.08.008

29 Akbari, P., Vuckovic, D., Stefanucci, L., Jiang, T., Kundu, K., Kreuzhuber, R., ... & Astle, W. J. (2023). A genome-wide association study of blood cell morphology identifies cellular proteins implicated in disease aetiology. Nature communications, 14(1), 5023. https://doi.org/10.1038/s41467-023-40679-y

30 Deecke, L., Homann, J., Goldeck, D., Luessi, F., Vandebergh, M., Ohlei, O., ... & Lill, C. M. (2024). Genome-wide association study reveals different T cell distributions in peripheral blood of healthy individuals at high genetic risk of type 1 diabetes and long COVID. medRxiv, 2024-02. https://doi.org/10.1101/2024.02.08.24302520

31 Tegtmeyer, M., Arora, J., Asgari, S., Cimini, B. A., Nadig, A., Peirent, E., ... & Raychaudhuri, S. (2024). High-dimensional phenotyping to define the genetic basis of cellular morphology. Nature Communications, 15(1), 347. https://doi.org/10.1038/s41467-023-44045-w

32 Kundu, K., Tardaguila, M., Mann, A. L., Watt, S., Ponstingl, H., Vasquez, L., ... & Soranzo, N. (2022). Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for 12 immune-mediated diseases. Nature genetics, 54(3), 251-262. https://doi.org/10.1038/s41588-022-01025-y

33 Lee, P., Chandel, N. S., & Simon, M. C. (2020). Cellular adaptation to hypoxia through hypoxia inducible factors and beyond. Nature reviews Molecular cell biology, 21(5), 268-283. https://doi.org/10.1038/s41580-020-0227-y

34 Brodin, P., & Davis, M. M. (2017). Human immune system variation. Nature reviews immunology, 17(1), 21-29. https://doi.org/10.1038/nri.2016.125

35 Arevalo, J., Su, E., van Dijk, R., Carpenter, A. E., & Singh, S. (2023). Evaluating batch correction methods for image-based cell profiling. bioRxiv. https://doi.org/10.1101/2023.09.15.558001

36 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. https://doi.org/10.1016/j.slasd.2023.08.009

37 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. https://doi.org/10.1186/s13059-019-1850-9

38 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. https://doi.org/10.1016/0014-4827(77)90399-8

39 Trescos, Y., Tessier, E., Rougeaux, C., Goossens, P. L., & Tournier, J. N. (2015). Micropatterned macrophage analysis reveals global cytoskeleton constraints induced by Bacillus anthracis edema toxin. Infection and Immunity, 83(8), 3114-3125. https://doi.org/10.1128/iai.00479-15

40 Czerwińska, J., & Owczarczyk-Saczonek, A. (2022). The role of the neutrophilic network in the pathogenesis of psoriasis. International journal of molecular sciences, 23(3), 1840. https://doi.org/10.3390/ijms23031840

41 Chandrasekaran, S. N., Cimini, B. A., Goodale, A., Miller, L., Kost-Alimova, M., Jamali, N., ... & Carpenter, A. E. (2024). Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nature Methods, 21(6), 1114-1121. https://doi.org/10.1038/s41592-024-02241-6

42 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. https://doi.org/10.48550/arXiv.2108.07258

43 Pfaendler, R., Hanimann, J., Lee, S., & Snijder, B. (2023). Self-supervised vision transformers accurately decode cellular state heterogeneity. Biorxiv, 2023-01. https://doi.org/10.1101/2023.01.16.524226

44 Atarsaikhan, G., Mogollon, I., Välimäki, K., iCAN, Mirtti, T., Pellinen, T., & Paavolainen, L. (2025). Self-supervised learning enables unbiased patient characterization from multiplexed cancer tissue microscopy images. bioRxiv, 2025-03. https://doi.org/10.1101/2025.03.05.640729

45 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. https://doi.org/10.1038/s41598-022-12914-x

46 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. https://doi.org/10.1002/cyto.a.24770

47 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. https://doi.org/10.1038/s41431-023-01528-0

48 Noble, P. B., & Cutts, J. H. (1967). Separation of blood leukocytes by Ficoll gradient. The Canadian Veterinary Journal, 8(5), 110. https://pmc.ncbi.nlm.nih.gov/articles/PMC1696854/

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

50 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. DOI: 10.1126/science.831278

51 Mund, A., Coscia, F., Kriston, A., Hollandi, R., Kovács, F., Brunner, A. D., ... & Mann, M. (2022). Deep Visual Proteomics defines single-cell identity and heterogeneity. Nature Biotechnology, 40(8), 1231-1240. https://doi.org/10.1038/s41587-022-01302-5

52 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. https://doi.org/10.1038/nmeth.3323

53 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. 10.1016/j.cels.2020.04.003

54 Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118-127. https://doi.org/10.1093/biostatistics/kxj037

55 Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. Journal of Statistical Software, 91, 1-30. 10.18637/jss.v091.i01

56 McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. https://doi.org/10.1038/s43586-024-00363-x

57 Brück, O., Lee, M. H., Turkki, R., Uski, I., Penttilä, P., Paavolainen, L., ... & Kreutzman, A. (2021). Spatial immunoprofiling of the intratumoral and peritumoral tissue of renal cell carcinoma patients. Modern Pathology, 34(12), 2229-2241. https://doi.org/10.1038/s41379-021-00864-0

58 Liarski, V. M., Sibley, A., van Panhuys, N., Ai, J., Chang, A., Kennedy, D., ... & Clark, M. R. (2019). Quantifying in situ adaptive immune cell cognate interactions in humans. Nature immunology, 20(4), 503-513. https://doi.org/10.1038/s41590-019-0315-3

59 Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., ... & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nature genetics, 53(7), 1097-1103. https://doi.org/10.1038/s41588-021-00870-7

60 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. https://doi.org/10.1111/j.1365-2141.1976.tb03563.x

61 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. https://doi.org/10.1038/sj.leu.2400768