ImmCellTyper: an integrated computational pipeline for systematic mining of Mass Cytometry data to assist deep immune profiling

Jing Sun[1*~], Desmond Choy[3*], Nicolas Sompairac[3], Shirin Jamshidi[3], Michele Mishto[1-2], Shahram Kordasti[3-5~]

[1] Centre for Inflammation Biology and Cancer Immunology & Peter Gorer Department of Immunobiology, King's College London, SE1 1UL London, United Kingdom

[2] Research group of Molecular Immunology, Francis Crick Institute, NW1 1AT London, United Kingdom

[3] School of Cancer and Pharmaceutical Sciences, King's College London, London, United Kingdom

[4] Haematology Department, Guy's Hospital, London, United Kingdom

[5] Department of Clinical and Molecular Sciences, Università Politecnica delle Marche, Ancona, Italy

*These authors contribute equally to this research work

~ corresponding authors

Jing Sun: jing.1.sun@kcl.ac.uk

Shahram Kordasti: shahram.kordasti@kcl.ac.uk

Abstract

Mass cytometry, also known as Cytometry by time-of-flight (CyTOF), is a cutting-edge high-dimensional technology for profiling marker expression at the single-cell level. This technology significantly advances clinical research in immune monitoring and the interrogation of immune cell populations. Nevertheless, the vast amount of data generated by CyTOF poses a daunting challenge for analysis. To address this, we describe ImmCellTyper (https://github.com/JingAnyaSun/ImmCellTyper), a novel and robust toolkit designed for CyTOF data analysis. The analytical framework incorporates an in-house developed semi-supervised clustering tool named BinaryClust, which first characterises main cell lineages, followed by in-depth interrogation for population of interest using unsupervised methods. BinaryClust was benchmarked with existing clustering tools and demonstrated superior accuracy and speed across two datasets comprising around 4 million cells, performing as good as manual gating by human experts. Furthermore, this computational pipeline provides a variety of visualization and analytical tools spanning from quality control to differential analysis, which can be tailored to user's specific needs, aiming to provide a one-stop solution for CyTOF data analysis. The general workflow consists of five key steps: 1) Batch effect evaluation and correction, 2) Data quality control and pre-processing, 3) Main cell lineage characterisation and quantification, 4) Extraction and in-depth investigation of cell type of interest; 5) Differential analysis of cell abundance and functional marker expression (supporting multiple study groups). Overall, ImmCellTyper integrates expert's biological knowledge in a semi-supervised fashion to accurately deconvolute well-defined main cell lineages, while also preserving the potential of unsupervised approaches to discover novel cell subsets and providing a user-friendly toolset to remove the analytical barrier for high-dimensional immune profiling.

Introduction

Mass cytometry or cytometry by time-of-flight (CyTOF) is a powerful high-throughput single-cell technology which employs stable elemental isotopes, as the same manner of fluorophores, to detect cellular proteins of interest. This approach successfully tackles the panel multiplex challenges faced by traditional flow cytometry due to spectral overlap and permits simultaneous measurement of over 40 parameters on millions of cells. To date, CyTOF has been widely applied in basic and translational medical research, such as deep immunophenotyping and characterisation of novel refined cell subsets, immune monitoring of cell-adoptive therapy, and dissecting cell subpopulations from heterogeneous tumour samples (Spitzer and Nolan). Nonetheless, the advantages of CyTOF come with the problem of handling high-dimensional dataset. Traditional gating strategy for flow cytometry, while still serving as the 'gold standard' for cell population identification in cytometry data, may not be an optimal option for CyTOF data due to its high-dimensional settings. The high-parametric resolution in CyTOF, aimed at revealing previously undiscovered cell subpopulations, leads to a significant increase in the complexity of gating schemas and hierarchical depth, which makes manual gating extremely labour intensive and time-consuming (Kimball et al.). Therefore, effective computational toolkits and pipelines are entailed for CyTOF data mining and analysis.

Efforts have been directed toward developing means of clustering algorithms to deconvolute a pool of live cell mixture into distinct cell populations, which facilitates CyTOF data analysis. For instance, unsupervised methods, which include flowSOM(Van Gassen et al.), Phenograph(Levine et al.), X-shift(Samusik et al.), spade(Qiu et al.), and DensVM(Becher et al.) etc., often combine with dimension reduction techniques like t-SNE(van der Maaten and Hinton, 2008), UMAP(McInnes and Healy, 2018), PCA etc., and require manual annotation of each cluster based on the marker expression patterns indicated by heatmaps. This approach works well to analyse populations in a data-driven manner, with all files concatenated and analysed all at once. Compared to manual gating, it has advantages in terms of convenience, efficiency, and relative unbiasedness from biological preconception, facilitating the detection of novel cell phenotypes for deep phenotyping. However, the unsupervised approaches are not always ideal and suitable for cytometry data. Mathematical clustering does not

necessarily have biological meaning, leading to occasional inaccuracies. And several benchmarking studies suggested that the accuracy for these unsupervised tools may not be optimal(Liu et al.). Additionally, the technical uncertainty of results produced by different clustering approaches remains a conundrum. Even for the same unsupervised method, the discrepancy among different runs without setting a seed, reduces the reproducibility and may cause confusion, particularly for biologists with limited computational knowledge. Moreover, manual validation is also essential, as biological annotation is the inevitable step to provide biological relevant labels for the clusters. However, this process can be time-consuming and subjective, hindering the automation of pipelines. This problem is particularly pronounced for a large marker panel and samples with high heterogeneity, resulting in a higher number of clusters that need to be annotated.

Advances in artificial intelligence have accelerated the development of alternative clustering methods in a supervised manner for cell type inference. These methods consider the 'ground truth' or prior knowledge about the marker expression of each given cell types to automatically label each cell. Currently, a couple of semi-automatic methods have been developed, such as Linear Discriminant Analysis(LDA)(Abdelaal et al.), DGCyTOF(Cheng et al., 2022), CyAnno(Kaushik et al., 2021), DeepCyTOF(Li et al., 2017), Automated Cell-type Discovery and Classification (ACDC) (Lee et al.) and Semi-supervised Category Identification and Assignment(SCINA) (Zhang et al.) etc. ACDC(Lee *et al.*) and SCINA (Zhang *et al.*) use a matrix of pre-defined markers for each cell type to annotate the clusters showing the same signature. These methods assume that markers are either expressed or not expressed (binary), which limits their ability to distinguish cell subtypes with similar phenotypes, particularly non-canonical cell types that cannot be easily separated linearly. Alternatively, DeepCyTOF and LDA use a maker expression matrix extracted from manually gated cell types as a training dataset to build a machine learning model for cell type prediction. This approach has higher precision and accuracy compared with aforementioned methods(Liu *et al.*). Nonetheless, it can be labour intensive for preparation of the training set manually. Also, these methods are limited in their ability to predict novel cell subsets beyond the pre-gated set of cell types and lack a systematic and comprehensive way to assign the cells which were not identifiable under any of the gated cell types. New solutions have emerged with algorithms such as DGCyTOF(Cheng *et al.*, 2022) and CyAnno(Kaushik *et al.*, 2021), the former

adopts a deep learning classification combined with hierarchical stable-clustering methods and an iteration calibration system to identify known cell types and assign novel subsets; while CyAnno is based on a machine-learning framework which allows the integrative modelling of both 'gated' and 'ungated' cells. Both methods have demonstrated high accuracy in their test datasets, but unfortunately are not widely used by the research community, possibly due to the issue for the hassle of training data preparation, the lack of user-friendliness and implementation challenges for bench researchers.

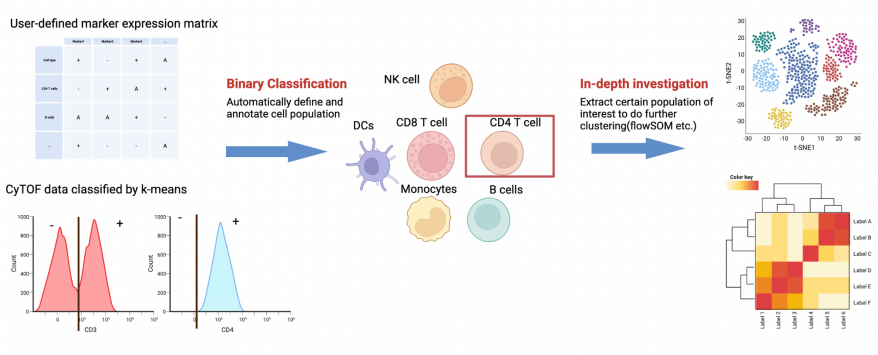
To address the common drawbacks of current semi-supervised and unsupervised clustering algorithms and preserve their strengths in discovering both canonical and non-canonical cell subsets respectively, we propose a strategy implemented in ImmCellTyper for cell classification named BinaryClust. By considering biologists' prior biological knowledge and interpretation for canonical cell clusters in a semi-supervised manner, BinaryClust first automatically characterises the main cell lineages in a fast and accurate way. Subsequently, it extracts specific cell types of interest for further clustering using unsupervised algorithms to identify cell subsets including previously unreported non-conventional population. In addition, this R-implemented pipeline takes advantage of SingleCellExperiment class for data management, providing an easy-to-use and organized systematic workflow of CyTOF data handling. The whole pipeline includes quality control and batch effect correction, which helps to effectively pool datasets from different batches, and ensures the robustness for downstream analysis. Meanwhile, modules like dimension reduction, semi-supervised and unsupervised clustering (flowSOM and Phenograph), interactive data visualisation, and statistical testing for complex study design were also incorporated in this pipeline. Compared with existing integrated computational workflow, such as CATALYST(Nowicka et al., 2017), CapX(Marsh-Wakefield et al.), Cytofkit(Chen et al., 2016) and ImmunoCluster(Opzoomer et al., 2021) which was developed and maintained by our team, this recently developed toolkit advanced further on coherence, functionality and user-friendliness. Overall, this approach has the potential to facilitate and smooth the investigation of CyTOF-based research.

Results

BinaryClust is comparable with manual gating in quantifying the abundances of main cell lineages.

161 The core concept for BinaryClust is depicted in Figure1, where a simple user-designed cell
162 type marker expression matrix is required and serves as a reference to characterise the main
163 cell lineages, with positive markers indicated in '+', negative markers in '-', and irrelevant
164 markers in 'A'. K-means(k=2) will be applied to divide the positive and negative cell
165 population of each marker, then align it to the reference table to infer main cell types. This is
166 followed by the extraction of population-of-interest for downstream clustering using
167 unsupervised methods for subpopulation discovery.
168
169 To assess the performance of the automated cell type classification and prediction, we
170 generated a test CyTOF dataset using peripheral blood mononuclear cell (PBMC) samples
171 from seven patients with myeloproliferative neoplasm (MPN) and two healthy donors,
172 employing a 37-marker deep immunophenotyping panel (Supplementary Table 1). Manual
173 gating was performed by two independent experts which identified seven main cell lineages.
174 These results act as the reference for evaluating the computer-aided methods. The
175 hierarchical sequential gating strategy was explicitly illustrated in Supplementary Figure 1. We
176 evaluated the agreement between manual gating results and BinaryClust results regarding
177 cell frequencies of each population. The mean value was calculated from manual gating
178 results of two experts to compare with BinaryClust-generated results using Pearson
179 correlation analysis. As shown in Figure2A, the two methods exhibit a strong correlation with
180 coefficient(R^2) equals to 1,1,0.99,0.96,0.96,0.99,0.99 for CD4 T cells, CD8 T cells, dendritic
181 cells, NK cells, monocytes, and gamma delta T cells, respectively (all $P < 0.0001$). And most of
182 the data points remain close to the line of equality (red line, $R^2=1$), indicating a high degree of
183 agreement. Meanwhile, the Bland-Altman plots in Figure 2B also suggest no consistent bias of
184 manual gating versus BinaryClust across all the identified cell types. The good performance of
185 BinaryClust was further validated in the influenza dataset published by our group (Alimam et
186 al.), which contains FCS files from 11 individuals with 6 main immune cell types detected, the
187 average F-measure and ARI reached 0.98 and 0.91 respectively (Supplementary Figure2A-B).

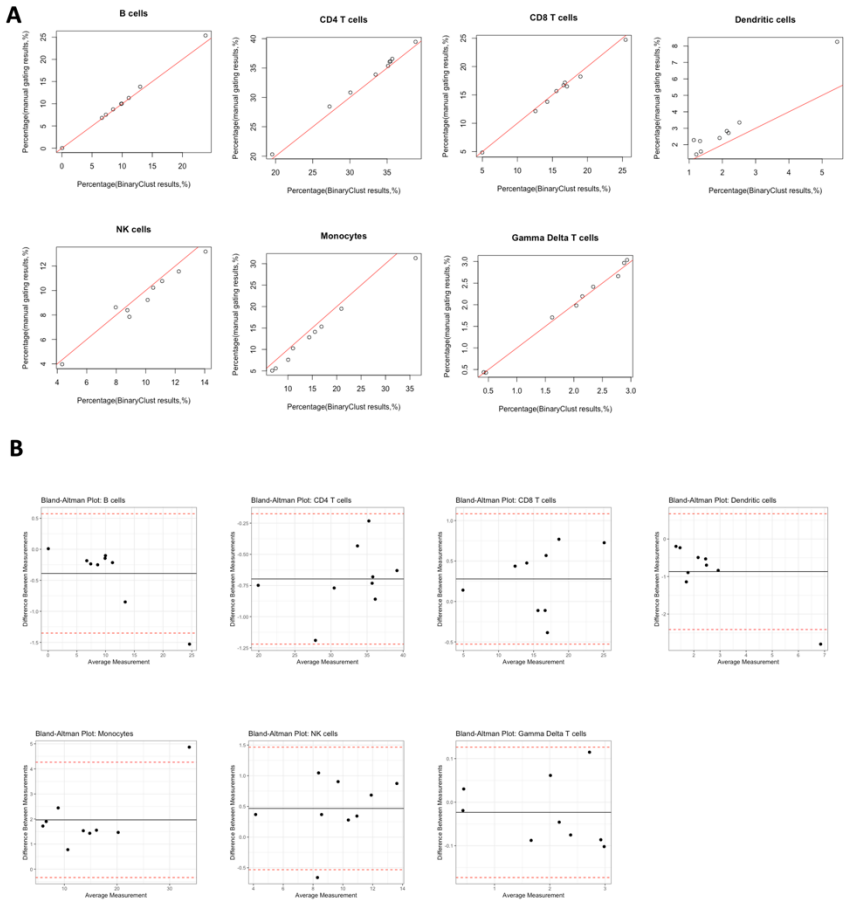
Figure 1



188

189 **Figure 1** Schematic diagram of BinaryClust method. Semi-supervised classification is first performed on
190 selected markers in the user-defined marker expression matrix to classify and annotate major cell
191 types. Population-of-interest can be further extracted and explored using unsupervised clustering
192 methods followed by differential analysis. Figure was generated by BioRender
193 (https://www.biorender.com/).

Figure 2



194

195 **Figure 2** Agreement evaluation comparing manual gating and BinaryClust in MPN cohort(n=9). Manual
196 gating of B cells, CD4 T cells, CD8 T cells, dendritic cells, NK cells, monocytes and gamma delta T cells
197 were performed by two independent experts using Cytobank, and mean values of the population
198 percentages were calculated to compare with BinaryClust results. Each dot represents one patient

sample. (A) Scatter plot showing the correlation between the two methods, with the red line indicating perfect agreement (correlation coefficient = 1); (B) Bland-Altman plots of the two measurement methods among all the cell populations, with the black line suggesting the mean observed difference and red dotted lines indicating limits of agreement (1.96 x standard deviations).

BinaryClust achieves high accuracy and speed compared with flowSOM and LDA

To further evaluate BinaryClust's performance, we compared it with the well-performing unsupervised algorithm flowSOM and supervised classifier Linear discriminant analysis (LDA). FlowSOM was run on the same MPN dataset with k value set as 20, followed by manual annotation and cluster merging to identify the same cell populations as in manual gating and BinaryClust. The cell frequencies derived from BinaryClust, flowSOM, and manual gating (expert1 and expert2) were compared using interaction plot (Figure 3). We observed that BinaryClust remains consistent with manual gating (all P>0.05), whereas flowSOM identified significantly less Gamma Delta T cells and dendritic cells (P<0.001 and P=0.006, respectively) compared with the other three measurements. We also increased the initial k value to 40 to improve accuracy the accuracy by over clustering then merging clusters (ref). We found that the average frequency for Gamma Delta T cells and dendritic cells increased to 1.5% and 4.65% respectively but still remained significantly different (P<0.05) from the other two methods. It is also interesting to see that although the same gating strategy was applied, different experts obtained results with slight variation despite no statistical significance. The same findings were confirmed using boxplot as shown in Supplementary Figure 3.

Figure 3

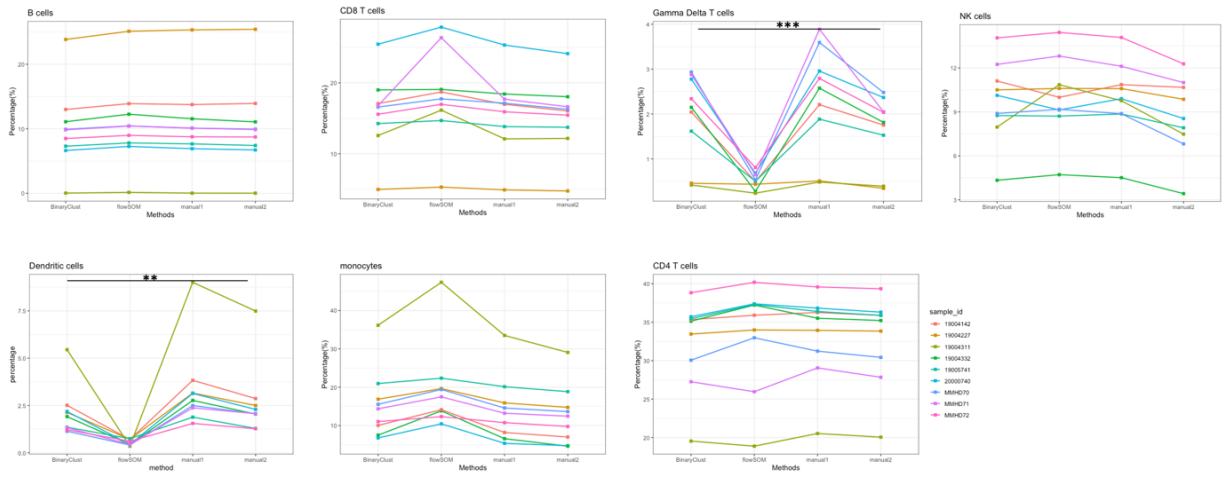


Figure 3 Comparison of manual gating (manual1 and manual2), BinaryClust and flowSOM clustering results in MPN cohort(n=9). Interaction plots showing the individual measurement (percentage) of each study participant with indicated colors by different methods across main cell lineages (B cells, CD8 T cells, Gamma Delta T cells, NK cells, Dendritic cells, monocytes and CD4 T cells).

Subsequently, LDA was also tested on the same dataset. Since this method requires a training dataset from manual gating to build a model, we exported the cell events and cluster assignment from Cytobank, then equally partitioned all cells into training and test dataset. Due to the difference in method implementation, we didn't include LDA for cell abundance quantification comparison with other methods, since only half of the cells were used for prediction. To ensure an equitable comparison and further evaluate the accuracy, F-measure, and Adjusted Rand Index (ARI) were calculated using manual gating cluster IDs as reference. As indicated in Table 1, for the MPN dataset, BinaryClust achieved high F-measure and ARI in all 7 cell types with an average of 0.94 and 0.91 respectively. The performance remains excellent in influenza dataset as well (average F-measure = 0.98, Supplementary Figure 2C). LDA has equivalent prediction accuracy as BinaryClust with an average F-measure of 0.93 and ARI of 0.90. Both supervised methods outperformed flowSOM (average F-measure=0.75, average ARI=0.66), which is in line with previous benchmarking study on evaluating supervised and unsupervised clustering algorithms(Liu *et al.*). Notably, for cell types that constitute a substantial proportion in the pool like CD4 T cells, CD8 T cells, NK cells, monocytes, and B cells, flowSOM can identify them with high precision and sensitivity, while this is not the case for dendritic cells and gamma delta T cells which account for less than 3% of the whole population (F-measure = 0.51 and 0.15, respectively). Even so, the overall F-measure and ARI for flowSOM considering cell proportion reached 0.81 and 0.80.

Table 1 Precision, recall, F-measure, and ARI of indicated clustering methods.

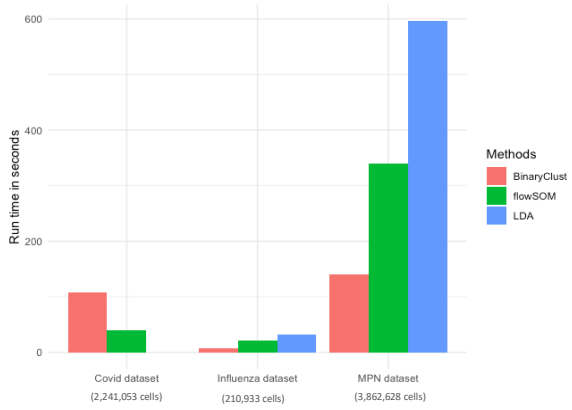
	Gated population	Counts	Cluster Cell Counts	True Positive	Precesion	Recall	F-measure	Average F-measure	ARI	Average ARI
BinaryClust	CD4 T cells	1270041	1232832	1226712	1.00	0.97	0.98	0.94	0.95	0.91
	CD8 T cells	601995	601639	582402	0.97	0.97	0.97		0.95	
	NK cells	389261	380999	376561	0.99	0.97	0.98		0.97	
	Monocytes	576624	625735	568924	0.91	0.99	0.95		0.92	
	B cells	397279	380170	374501	0.99	0.94	0.96		0.95	
	Dendritic cells	135894	86012	84130	0.98	0.62	0.80		0.74	
	TCRgd T cells	87194	73599	73562	1.00	0.84	0.92		0.91	
flowSOM (k=20)	CD4 T cells	1270041	1268929	1226712	0.97	0.97	0.97	0.75	0.90	0.66
	CD8 T cells	601995	692374	571344	0.83	0.95	0.89		0.81	
	NK cells	389261	393646	365403	0.93	0.94	0.93		0.91	
	Monocytes	576624	792447	568284	0.72	0.99	0.85		0.73	
	B cells	397279	405886	390077	0.96	0.98	0.97		0.96	
	Dendritic cells	135894	20832	18254	0.88	0.13	0.51		0.22	
	TCRgd T cells	87194	19720	4967	0.25	0.06	0.15		0.08	
LDA	CD4 T cells	634733	665481	632700	0.95	1.00	0.97	0.93	0.93	0.90
	CD8 T cells	301138	327469	299309	0.91	0.99	0.95		0.92	
	NK cells	194470	202110	187259	0.93	0.96	0.94		0.92	
	Monocytes	288338	301387	282836	0.94	0.98	0.96		0.94	
	B cells	198582	211090	195908	0.93	0.99	0.96		0.94	
	Dendritic cells	68152	90456	61785	0.68	0.91	0.79		0.75	
	TCRgd T cells	43405	43064	39976	0.93	0.92	0.92		0.92	

246 ARI, Adjusted Rand Index; LDA, Linear discriminant analysis

247

248 It has been demonstrated that flowSOM and LDA are among the fastest clustering algorithms
249 without compromising their performance. Here, we compared the speed amongst the three
250 approaches in MPN, Influenza and COVID-19 datasets(Chevrier et al.)which contain 2,231,053,
251 210,933 and 3,862,628 cells respectively. Since LDA requires training data to be executed, we
252 did not run it on COVID-19 dataset due to the absence of manual gating results. As shown in
253 Figure 4, BinaryClust exhibited the highest speed in both MPN and Influenza dataset but fell
254 behind flowSOM slightly in COVID-19 dataset.

Figure 4



255

256 **Figure 4** Comparison of BinaryClust, flowSOM and LDA on speed. Bar chart showing runtime (in
257 seconds) of the three methods in three different datasets.

258

259 **ImmCellTyper pipeline supports interactive data visualisation and comparison among study**
260 **groups.**

261

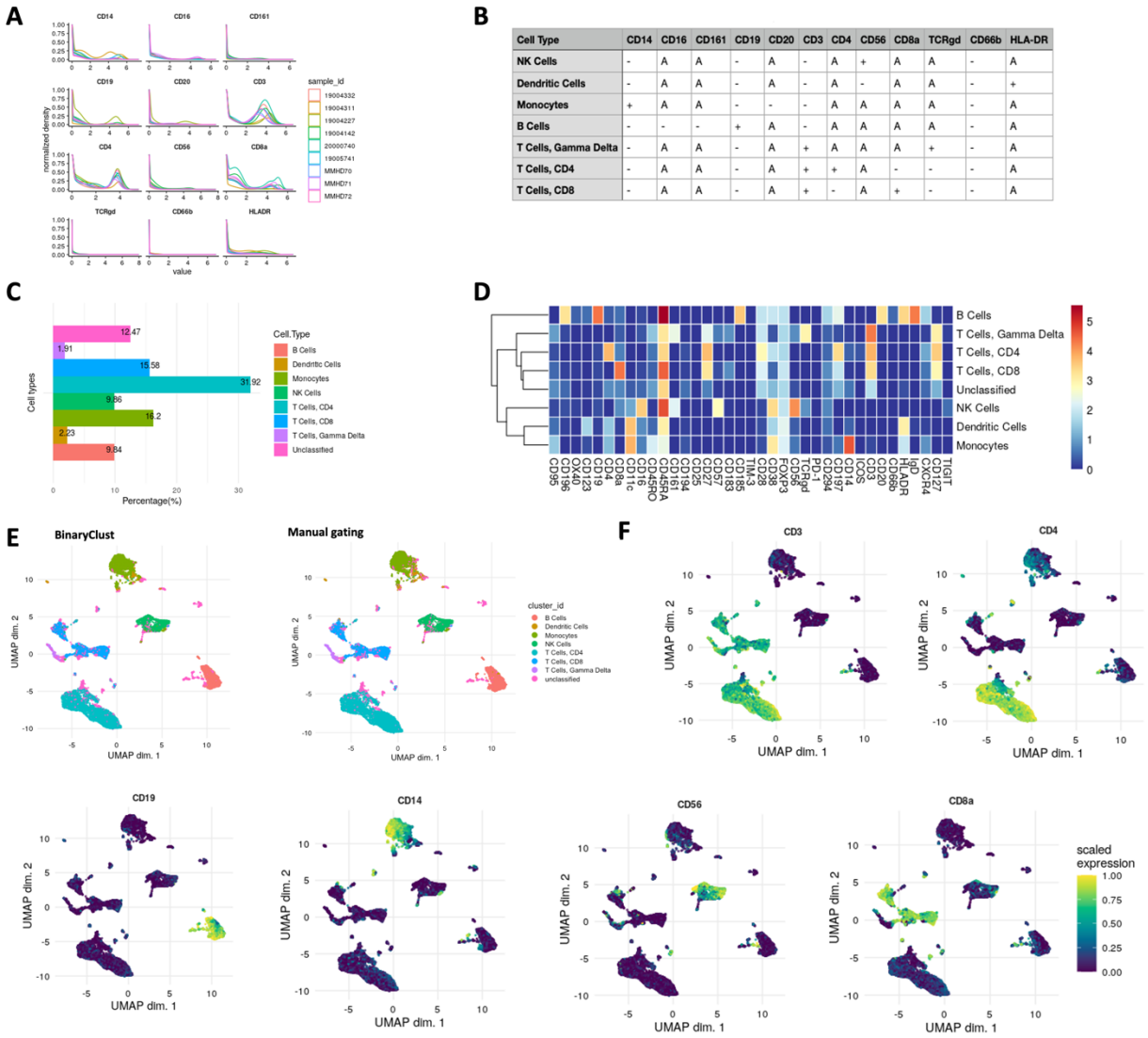
262 BinaryClust is an important component of ImmCellTyper, a comprehensive integrated pipeline
263 designed for systematic CyTOF data mining. Hence, we utilised the visualisation functions to
264 further prove the reliability and robustness of BinaryClust.

265

266 BinaryClust inherently considers CyTOF makers as binary distributed which is in most of the
267 case, but not always. Therefore, it is crucial to check the marker behaviour before running
268 the pipeline. As shown in Figure 5A, all markers selected for the classification matrix (Figure
269 5B) displayed a binary distribution implying the suitability for this pipeline. After clustering
270 (Figure 5C), median marker expression heatmap (Figure 5D) can check the reliability of the

results before proceeding for downstream analysis. In MPN dataset, we projected the cluster assignment resulted from BinaryClust and manual gating to UMAP (Figure 5E), along with the expression of the phenotypic markers (Figure 5F and supplementary Figure 4A). High similarity was observed between BinaryClust and manual gating, whereas slight difference was found on flowSOM results coloured UMAP on islands of CD8 T cells and CD4 T cells (Supplementary Figure 4A). FlowSOM appeared to classify cells that were close on spatial distance into the same cluster, in contrast to prior knowledge-based methods: BinaryClust and manual gating.

Figure 5



Cell Type	CD14	CD16	CD161	CD19	CD20	CD3	CD4	CD56	CD8a	TCRgd	CD66b	HLA-DR
NK Cells	-	A	A	-	A	-	A	+	A	A	-	A
Dendritic Cells	-	A	A	-	A	-	A	-	A	A	-	+
Monocytes	+	A	A	-	-	-	A	A	A	A	-	A
B Cells	-	-	-	+	A	-	A	A	A	A	-	A
T Cells, Gamma Delta	-	A	A	-	A	+	A	A	A	+	-	A
T Cells, CD4	-	A	A	-	A	+	+	A	-	-	-	A
T Cells, CD8	-	A	A	-	A	+	-	A	+	-	-	A

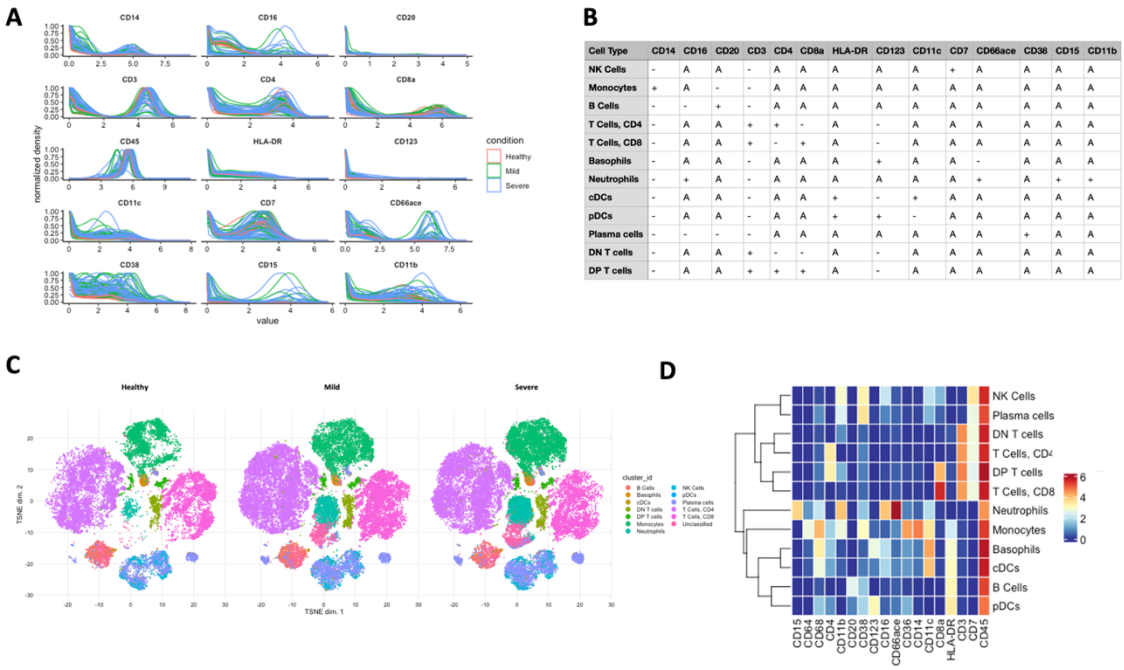
Figure 5 Cell type characterization and visualisation using ImmCellTyper pipeline in MPN dataset(n=9). (A)Intensity distribution of selected phenotypic markers used for BinaryClust classification, coloured by sample_id;(B) Pre-defined expression classification matrix for the MPN dataset, '+' indicates positive, '-' indicates negative and 'A 'suggests 'any';(C) Proportion of the main cell lineages of all cells in the concatenated FCS files after classification; (D) Median marker expression heatmap of

BinaryClust classification results; (E)UMAP plot of random downsample of 2000 cells per patient coloured by main cell types based on BinaryClust classification(left) and manual gating results(right). (F) UMAP plots coloured by normalized expression of indicated markers (CD3, CD4, CD8a, CD20, CD19, CD14, and CD56) across 2000 cells per sample.

Application of ImmCellTyper pipeline to the COVID-19 dataset demonstrates its versatile functionalities for comprehensive data analysis

To showcase the analytical and visualisation functions of ImmCellTyper pipeline, we used the dataset published by Chevrier et al. (Chevrier *et al.*), which described the immune signature of mild and severe COVID-19 patients in comparison with healthy individuals. There are a total of 82 FCS files with a 40-plex marker panel focusing on innate immunity in this dataset. An initial marker expression check was carried out based on the user-defined matrix (Figure 6B) and displayed in Figure 6A. BinaryClust was then performed and identified 12 cell populations as expected. We used t-distributed stochastic neighbour embedding(t-SNE) in this dataset for dimension reduction, coloured by cell types and faceted by disease conditions, which exhibited substantial immune alteration among healthy control, mild and severe COVID-19 patients (Figure 6C). From the heatmap in Figure 6D, we can have an overview of the marker expression of each population, which remains consistent with our initial definition indicated in the expression matrix.

Figure 6



Cell Type	CD14	CD16	CD20	CD3	CD4	CD8a	HLA-DR	CD123	CD11c	CD7	CD66ace	CD38	CD15	CD11b
NK Cells	-	A	A	-	A	A	A	A	A	+	A	A	A	A
Monocytes	+	A	-	-	A	A	A	A	A	A	A	A	A	A
B Cells	-	-	+	-	A	A	A	A	A	A	A	A	A	A
T Cells, CD4	-	A	A	+	+	-	A	-	A	A	A	A	A	A
T Cells, CD8	-	A	A	+	-	+	A	-	A	A	A	A	A	A
Basophils	-	A	A	-	A	A	A	+	A	A	-	A	A	A
Neutrophils	-	+	A	-	A	A	A	A	A	A	+	A	+	+
cDCs	-	A	A	-	A	A	+	-	+	A	A	A	A	A
pDCs	-	A	A	-	A	A	+	+	-	A	A	A	A	A
Plasma cells	-	-	-	-	A	A	A	A	A	A	A	+	A	A
DN T cells	-	A	A	+	-	-	A	-	A	A	A	A	A	A
DP T cells	-	A	A	+	+	+	A	-	A	A	A	A	A	A

305 **Figure 6** Applying ImmCellTyper pipeline on COVID-19 patient dataset(n=82) published by Chevrier
306 et.al. (A) Marker intensity distribution of selected phenotypic markers used for BinaryClust
307 classification, coloured by disease severity(n=22 healthy individuals, 28 mild COVID-19 patients and 38
308 severe COVID-19 patients);(B)Pre-defined marker expression classification matrix used for BinaryClust;
309 (C) T-SNE plots, with 1000 cells per sample, were coloured by the main cell types generated by
310 BinaryClust and faceted by different study groups;(D)The corresponding median marker expression
311 heatmap of BinaryClust results for the COVID-19 dataset.
312
313 The abundances of the identified cell types were quantified for each individual study
314 participant in the format of stacked histogram (Figure 7A) and summarized in boxplot (Figure
315 7B). Compared with healthy volunteers, there are significant immune alterations in COVID-19
316 patients in B cells, Basophils, cDCs, Monocytes, NK cells, CD8 T cells, Neutrophils and pDCs (all
317 $P<0.05$), which depicts a similar trend to the original paper, despite slight discrepancies
318 caused by different statistical methods used. Kruskal Wallis test was conducted for this
319 dataset followed by multiple testing correction (Benjamini-Hochberg procedure) and Dunn's
320 test for post hoc analysis, while Chevirer et al. used Mann-Whitney Wilcoxon test corrected
321 by Holm method. We then selected 9 markers (IL-6, PD-L1, VISTA, IDO, TIM-3, TMEM173,
322 GranzymeB, PPARg, Ki-67) as the state markers which can reflect the functional or
323 proliferative status of the immune cell types. Notably, Granzyme B was observed to be
324 significantly highly expressed in COVID-19 patients versus healthy control across all identified
325 immune cell types(all $P<0.05$), TMEM173 was substantially up-regulated in COVID-19 patients
326 particularly in monocytes, neutrophils, and CD4 T cells; and PD-L1 expression remained low
327 for all cells, indicating the immune system of COVID-19 patients were highly activated without
328 exhaustion regardless of disease severity, which was not explored previously. Since
329 monocytes and neutrophils were of particular interest in the original paper, and the panel
330 also included specific markers for in-depth interrogation, we then extracted the two
331 population and carried out Phenograph(k=60) to investigate the subclusters. As shown in
332 Figure 7D, Phenograph returned 16 and 14 subclusters respectively for monocytes and
333 neutrophils, with heatmap provided in Supplementary Figure 5.
334
335 Reanalysis of the COVID-19 dataset demonstrated the concordance of ImmCellTyper pipeline
336 with original reports with additional findings, as well as its versatile interactive data
337 visualisation functionalities.

Figure 7

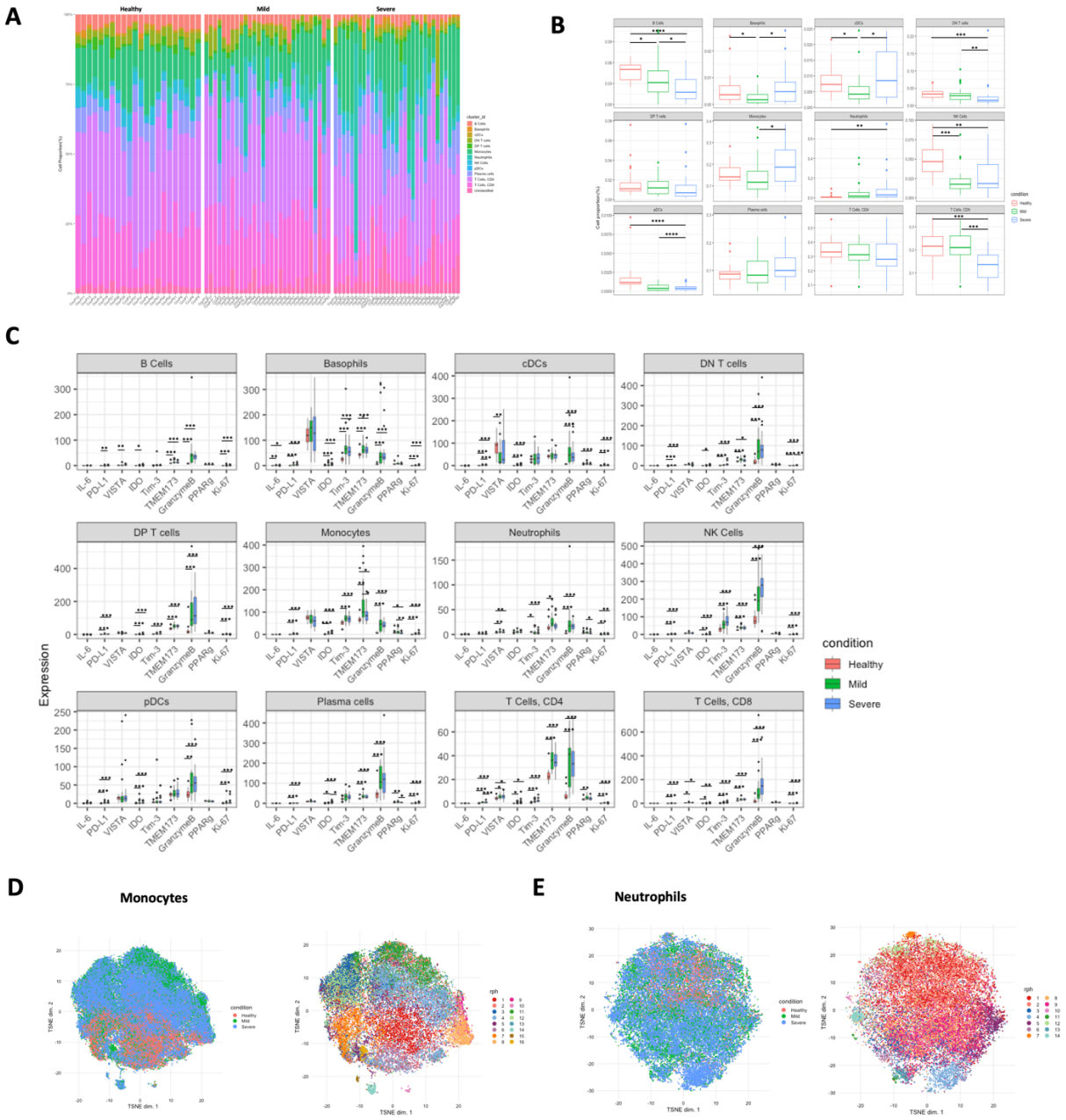


Figure 7 Quantification and statistical analysis comparing the study conditions in COVID-19 dataset(n=82). (A)Stacked histogram of main cell type composition per individual generated by BinaryClust, and grouped by study conditions (healthy, mild and severe); (B)Boxplots representing cell abundance frequencies among the study conditions, faceted by different main cell types;(C) State marker expression intensities with comparison of the study groups across the main cell types; (D)Clusters of monocytes and neutrophils were extracted from the whole cells for downstream interrogation. T-SNE plots with random downsample of 1000 monocyte cells and (E) neutrophils per sample were coloured by study conditions and Phenograph clustering results(k=60), respectively. Statistical significance was marked by asterisk. * P<0.05, **P<0.01, ***P<0.001, ****P<0.0001

Discussion

In this study, we present an analytical pipeline named ImmCellTyper for systematic exploration of CyTOF data. This pipeline addresses a comprehensive range of analytical needs encompassing data quality check, batch effects examination/correction, cell type identification, and downstream differential analysis accompanied by high-quality, publishable data visualisations. Furthermore, ImmCellTyper includes an in-house developed, knowledge-based, semi-supervised classifier BinaryClust with high accuracy and speed, and also integrates the well-performing and state-of-the-art unsupervised algorithms (Phenograph and flowSOM). By adopting the strategy of first obtaining the main cell types and then select specific cell types for in-depth interrogation with higher clustering resolution, ImmCellTyper combines the advantages of both supervised and unsupervised clustering algorithms for discovery of both major populations and refined subpopulations. Designed for ease of use, this pipeline features a clear and user-friendly workflow, and is compatible to the widely used pipeline CATALYST, aiming to provide a one-stop solution for CyTOF users.

For validating the robustness of the automated cell characterization function of ImmCellTyper: BinaryClust, we involved two independent datasets (MPN dataset and Influenza dataset) and compared with existing clustering tools flowSOM and LDA. Previous Benchmarking studies indicated that flowSOM was one of the best-performing unsupervised tools in precision, speed and stability, with F-measure ranges from 0.58~0.90 in various datasets(Liu *et al.*), and was widely used in high-impact research publications(Liu et al.; Liu *et al.*). However, the results of flowSOM vary upon the k values set by the users which causes some uncertainty and impairs reproducibility of the results. In contrast, semi-supervised clustering methods like ACDC, DeepCyTOF(Li *et al.*, 2017) and LDA outperformed the unsupervised methods in the same benchmarking study. The F-measure obtained by ACDC and LDA ranged from 0.78~0.99, which was significantly higher than unsupervised approaches including Accense, PhenoGraph, Xshift, k-means, flowMeans(Aghaeepour et al.), FlowSOM and DEPECHE(Theorell et al.)(F-measure: 0.28~0.93)(Liu *et al.*). In this study, we chose LDA because it has proven to be superior to ACDC and DeepCyTOF in terms of precision and speed(Abdelaal *et al.*; Liu *et al.*). The excellent performance of LDA was demonstrated in MPN dataset of this study with F-measure reached 0.93 for around 4 million cells. BinaryClust was

381 comparable in accuracy(F-measure=0.94) but much faster in speed. Both semi-supervised
382 approaches outperformed flowSOM as expected. Although semi-supervised methods seem to
383 perform better, the mainstream for CyTOF data analysis still relies on unsupervised methods.
384 The reason for this might be that most of the semi-supervised methods rely on manual gating
385 as reference and requires either the manually gated expression matrix or user-defined binary
386 matrix as training data. This process takes extra time and efforts especially for defining
387 subpopulations, and may introduce end-user bias. Additionally, the common restriction for all
388 the supervised methods is that they have limited capability to reveal novel subsets as those
389 are not predefined in the training pool, which poses a critical challenge for the pursuit of
390 novel discoveries. Considering all these strength and limitations, as a semi-supervised
391 method, BinaryClust possesses the advantages of incorporating prior biological knowledge,
392 being reproducible across runs and accurate on cell type identification. Meanwhile, the input
393 requirement is a simple matrix of cell population definition based on marker positivity,
394 avoiding the hassle to manually gate example FCS files. With a deep understanding that
395 certain CyTOF markers are expressed on a continuum rather than binary manner, this
396 automated supervised approach will only be used for classification of main cell lineages. The
397 test datasets in this study contained data from human PBMC samples which represented a
398 heterogenous pool of immune cells, covering the major types of various immune populations,
399 and demonstrated excellent performance of BinaryClust on classifying CD4 T cells, CD8 T cells,
400 gamma delta T cells, NK cells, monocytes, dendritic cells, and B cells. We also recommend
401 checking marker expression distribution before running the pipeline to ensure accurate
402 results. After the above step, this pipeline supports cell population extraction for further
403 dissection into subclusters using unsupervised approaches, which excel in detecting rare and
404 refined subpopulations. Phenograph is a particularly effective tool for this purpose. By
405 applying this strategy, our previous work with BinaryClust facilitated the evaluation of the
406 impact of systematic anti-cancer agents on lymphocyte population in non-small cell lung
407 cancer patients, as well as the characteristics of autologous T cell products after
408 manufacturing (O'Brien Gore et al.).
409
410 Another purpose behind the development of the ImmCellTyper framework is to cater to
411 general analysis needs, spanning from pre-processing to downstream differential analysis.
412 There are existing well-established pipelines like CATALYST(Nowicka *et al.*, 2017),

413 diffcyt(Weber et al., 2019), Cytofkit(Chen *et al.*, 2016), and ImmunoCluster(Opzoomer *et al.*,
414 2021) etc., each of them has its own strength and limitations. One of the challenges existed in
415 CyTOF data pre-processing is effectively integrating multiple batches, given the technical
416 differences arising from experiments and instrumental acquisition, which could result in the
417 separation of clustering across batches and potentially confounding the signal of interest. In
418 ImmCellTyper pipeline, we provide the functionality for batch effects examination and
419 incorporated two well-performing batch correction algorithms CytoNorm (Van Gassen et al.)
420 and CytofRUV(Trussart et al.) into our framework prior to downstream analysis, which
421 ensures the quality of the data and, to the best of our knowledge, makes it the first integrated
422 pipeline with this module. For the downstream analysis, tools like diffcyt and CATALYST group
423 markers into phenotypic and state makers, enabling the detection of differentially abundant
424 cell clusters and differential expression of functional markers within each cell
425 population(Weber *et al.*, 2019). This well-recognised strategy has been extensively applied in
426 various studies. But one limitation is that it directly classifies cells into high-resolution
427 clusters, and cannot automatically merge subclusters into one major population with similar
428 phenotypes(Weber *et al.*, 2019). Our pipeline addresses this problem via a precise and
429 convenient solution: using the semi-supervised classifier (BinaryClust) of ImmCellTyper. The
430 two pipelines (CATALYST and ImmCellTyper) are well compatible, allowing users to leverage
431 their functions together and providing a broader range of analytical options. On the other
432 hand, Cytofkit has the advantages of the graphical user interface(GUI) for non-specialists,
433 integration of a variety of clustering(DensVM(Becher *et al.*), ClusterX, FlowSOM, Phenograph),
434 dimension reduction(PCA, ISOMAP, t-SNE) methods, and inference of the relatedness among
435 cell populations, but it does not support complex study design, group comparison and
436 statistical testing(Chen *et al.*, 2016). Several methods have been developed to fit with the
437 high-dimensional features of CyTOF data for differential analysis including diffcyt,
438 CellCnn(Arvaniti and Claassen), cydar(Lun et al.), citrus(Bruggner et al.), cyEMD(Arend et al.)
439 etc. Nonetheless, none of them is applicable to compare more than two study groups at once
440 which involves multiple testing. Therefore, to accommodate study designs involving multiple
441 groups, ImmCellTyper does statistics via first applying Kruskal Wallis test, followed by
442 Benjamin-Hochberg (BH) procedure for multiple testing correction and post hoc analysis via
443 Dunn's test or pairwise-Wilcoxon test. In cases where the comparison involves only two

groups, Mann-Whitney test will be applied in terms of cell frequency, as the distribution of CyTOF data generally does not fit for normal distribution.

One limitation of this computational pipeline is that its semi-supervised cluster identification may not be well-suited for the direct identification of specific subpopulations which are defined by continuum markers. This is due to the method's reliance on the presumption that makers are binary distributed. In addition, for users with general interest in charactering subclusters of each main cell lineage, it can be laborious to first perform BinaryClust then extract every population for unsupervised clustering. In such scenarios, BinaryClust can be used in parallel with other clustering methods or pipelines like CATALYST. This allows users to quickly obtain additional information about main cell types with the accuracy comparable to manual gating.

In summary, we introduce a novel open-source R-implemented strategy and a versatile toolbox for CyTOF data analysis. The future direction is to automate the data clean-up, compensation, and bead normalisation steps. Additionally, we also aim to implement the whole pipeline into Python, a more accessible programming language for bioinformatics novices and biologists who wish to perform high-dimensional data analysis independently.

Materials and methods

BinaryClust

Most of the CyTOF markers exhibit log-normal or bi-modal distribution with zero inflation after arcsinh transformation. Thus, we employ binary classification using k-means to group cells into negative and positive populations(k=2) for each marker indicated in the user-defined classification matrix. Here, k-means is an unsupervised clustering algorithm to cluster the data based on the Euclidean distance among points, which is calculated by the formula below:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Then assign each data point into a cluster centroid which is denoted by c_i, and dist()is the Euclidean distance:

$$arg \min_{ci \in C}(ci, x)^2$$

By aligning the k-means results with the user-designed classification matrix, cell populations can be subsequently classified and annotated.

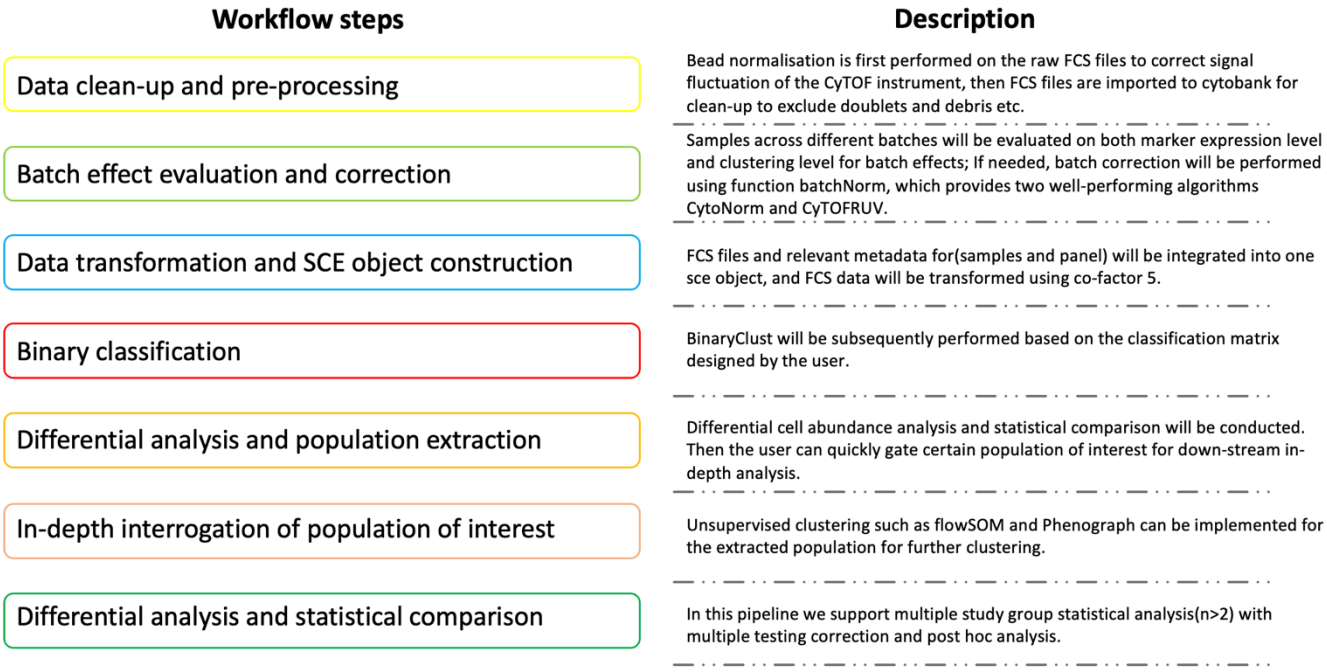
R package ImmCellTyper

This computational pipeline is implemented in the R package ImmCellTyper and publicly available on Github (https://github.com/JingAnyaSun/ImmCellTyper). Instructions for package installation and function usage can be found in the README file on the Github page. It is recommended to first examine the unwanted non-biological variation across batches and perform additional batch normalisation if necessary. Afterwards, import the data into the second part of the pipeline for downstream analysis. Users are required to prepare all the FCS files, sample metadata containing the details and grouping information of each sample, panel metadata with the information of the antibody panel with metal tags used in the experiments, and cell type classification matrix with phenotypic marker expression in a binary manner of each cell lineage. All files need to be in the right format to use the pipeline.

ImmCellTyper workflow overview

ImmCellTyper pipeline is composed of two separate sub-pipelines, as illustrated in Figure 8: (1) Sub-pipeline1: Batch effects evaluation and correction. Batch effects occur when samples were collected and measured at different sites or time points, especially for large-scale studies. It is crucial to remove the unwanted variation which might interfere the true biological differences. Users can systematically examine batch effects on two levels including marker behaviours and clustering results based on the method introduced by Trussart et al. (Trussart *et al.*) If needed, CytoNorm and CytofRUV, which are well-performing correction algorithms, can be used to align the existing batch effects.

504 (2) Sub-pipeline2: semi-supervised classification, differential analysis, and in-depth
505 investigation. When data are cleaned and normalised, they can be imported into the second
506 part of the pipeline, constructed into a SingleCellExperiment (sce) object, and undergo semi-
507 supervised classification to identify the major cell types and test the differential frequencies
508 or state marker expression among study groups. After that, if the users have certain interests
509 of specific cell types and pre-design the panel for that, or in another circumstance, the initial
510 statistics draw the user's attention into a certain cell type, institutively, further clustering
511 using unsupervised tools like flowSOM or Phenograph should be conducted with an increased
512 cluster resolution and deeper investigation for cell subsets, after extracting the cell
513 population of interest. ImmCellTyper pipeline has the same data storage and infrastructure as
514 CATALYST, therefore all the functions in CATALYST can be seamlessly used in ImmCellTyper,
515 tailored to user's analytical needs. We do not elaborate on the basic functions of CATALYST,
516 which can be found in the tutorial vignettes of the
517 package(https://github.com/HelenaLC/CATALYST).



518 **Figure 8** Overall schematic outline of the ImmCellTyper workflow with description for each step.

519

520 ***Batch correction algorithms***

521

522 The batch effect correction algorithms embedded in the function of 'batchNorm' comprises
523 CytoNorm, as described by Van Gassen et al.(Van Gassen *et al.*) and CytofRUV by Trussart et

524 al.(Trussart *et al.*). Both algorithms rely on anchors (reference samples/technical replicates)
525 across batches to perform normalisation. CytoNorm uses flowSOM clustering to first identify
526 clusters prior to a population-specific transformation on the reference samples by computing
527 the quantile values and aligning them with splines, whereas CytofRUV applies remove
528 unwanted variation III (RUV-III) to CyTOF data by estimating and eliminating the non-
529 biological variation of the pseudo-replicates.
530
531 ***Agreement evaluation***
532
533 To assess the agreement among manual gating, BinaryClust classification and flowSOM
534 clustering, we used correlation, interaction plot and Bland-Altman analysis. Correlation
535 evaluates the relationship between two variables which doesn't mean concordance, but if
536 two methods agree, surely, they should be highly correlated. We compared the results of cell
537 population frequency generated by experts and BinaryClust using Pearson correlation,
538 calculated the correlation coefficient and P value, with the line of equality indicating perfect
539 agreement (red solid line, R^2=1) in the plots to help gauge the degree of agreement between
540 the two methods; Bland-Altman plot refers to a dot plot of the difference between two
541 variables(y-axis) against the mean of them(x-axis), as described by J.Martian Bland and
542 Douglas G. Altman in 1986, which represents a graphical magnitude of bias(average of
543 difference) with 95% confidence interval. The math formula for the limit of agreement is as
544 below:
545 $\text{Limits of agreement} = \text{mean difference observed} \pm 1.96 \times \text{standard deviation}$
546
547 We also used Interaction plot to display the interaction effects of the three methods including
548 manual gating, BinaryClust classification and flowSOM clustering on the measurement of cell
549 frequencies to evaluate the agreement.
550
551 ***F-measure***
552
553 F-measure is a method to assess the accuracy of the clustering method compared to gold
554 standard, which is manual gating results in this study. It stands for the harmonic mean of the
555 precision and recall values, which can be calculated using below formula:

$$F - measure = 2x \frac{Precision \times Recall}{Precison + Recall}$$

Here, precision represents the positive predictive value: the proportion of true positive instances divided by the instances classified as positive by the clustering algorithm; recall evaluates the sensitivity, which is the number of true positive events correctly identified by the algorithm among all events that belong to the cluster. F-measure ranges from 0 to 1, where 1 suggests perfect performance and 0 indicates poor precision and recall.

Adjusted Rand Index

Adjusted rand index is a widely used method to measure the similarity between two clustering results. We used it to assess the agreement between test clustering algorithm and the gold standard labels which is derived from manual gating. ARI is defined as the following formula based on a contingency matrix (where n_{ij}, a_i, b_j are values from the contingency table):

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

ARI ranges from -1 to 1, with 1 indicates a perfect match, 0 means not relevant, and -1 indicates complete mismatch between two clustering results.

Differential analysis

In order to perform the differential analysis of cell abundances and state marker expression amongst study groups, we consider the number of study groups into two conditions in function 'StatTest': 1) n=2, Mann-Whitney Wilcoxon analysis will be applied as cell frequency/marker expression doesn't fit normal distribution; 2) n=3, Kruskal- Wallis test will be first performed followed by multiple testing correction using BH procedure and post-hoc analysis(Dunn's test or pairwise Wilcoxon test).

Sample collection and preparation for CyTOF

586

587 PBMC samples from MPN patients were requested and obtained from biobank at Guy's
588 Hospital, under a protocol approved by the KCL Biobank Access Committee (REC18/EE/0025).
589 Healthy volunteers were recruited at Guy's hospital with informed consent and ethical
590 approval by King's College Research Ethics Committee (HR-17/18-5960 MOD-20/21-5960) in
591 accordance with the Declaration of Helsinki. All identifiable information of study participants
592 were securely stored in a trusted research environment managed by members of the team.
593

594 Venous blood samples from healthy volunteers were collected in BD vacutainer EDTA tubes,
595 and PBMCs were isolated and purified using Ficoll-Hypaque density gradient centrifugation. In
596 brief, blood was carefully layered onto Ficoll and spined at 460g for 20 minutes at room
597 temperature (RT) without brake; Upon observing a clear separation of blood components,
598 PBMCs were then carefully isolated and washed three times with RPMI 1640 medium to
599 remove other contaminants. Subsequently, the purified PBMCs were cryopreserved in liquid
600 nitrogen at a density of $1X10^7$ cells per vial.
601

602 **CyTOF Antibody staining**
603

604 Cryopreserved cells were quickly thawed at 37°C in water bath, suspended in pre-warmed
605 RPMI 1640 medium, washed three times, and underwent Fc-blocking using human TruStain
606 FcX (Biolegend) for 10 mins at RT. Given that the antibody panel employs indirect detection of
607 CD95 using anti-CD95-APC and anti-APC-106Cd, cells were initially stained with anti-CD95-APC
608 for 30 minutes in a dark place, followed by two washes with 2ml cell staining buffer (CSB).
609 After the removal of supernatant, cells were resuspended in 300µl CSB and transferred into
610 the dip tube containing lyophilized antibody mix of Maxpar direct immune profiling assay
611 (MDIPA, Standard BioTools). Additional antibodies used to study T cell activation, migration,
612 and exhaustion status (anti-TIM3, anti-PD-1, anti-ICOS, anti-TIGIT, anti-OX40) were also
613 added. Cells were incubated with the antibody mix for 30 mins at RT in compliance with
614 manufacturer's instructions, washed twice and fixed using a freshly prepared 1.6%
615 paraformaldehyde (PFA, Thermo Fisher Scientific) solution for 10 mins at RT. In the end, cells
616 were washed with CSB twice to remove residual PFA and stained with 125nM Cell-ID
617 intercalator-Iridium in 1 ml Maxpar Fix and Perm Buffer (Standard BioTools) overnight at 4°C.

618 The stained cells were frozen down using freezing media (fetal bovine serum containing 10%
619 DMSO) and stored at -80°C freezer before data acquisition.
620
621 On the day for CyTOF acquisition, cryopreserved samples were thawed, washed twice with
622 1ml CSB, followed by additional two washes with 1ml cell acquisition solution (CAS, Standard
623 BioTools), and centrifuged at 800g for 5 minutes. Cells were then filtered through a 40 µm cell
624 strainer to avoid blockage and cell count was determined by Countess automated cell counter
625 (Invitrogen). EQ four-element calibration beads (Standard BioTools) were added to a final
626 concentration of $0.5x10^6$ cells/ml to adjust signal fluctuation of the instrument. CyTOF
627 acquisition was performed on Helios mass cytometer system. For each batch of cell staining
628 and run, technical replicates from the same healthy donor were included to evaluate and
629 correct batch effects.
630
631 The full antibody panel including the metal tag, clone, and supplier is listed in Supplementary
632 Table 1.
633
634 ***CyTOF data pre-processing***
635
636 FCS files were first processed for bead normalisation using CyTOF v7.0 system control
637 software (Standard BioTools) to correct signal drift during acquisition. Subsequently, the files
638 were imported into Cytobank (Cytobank Inc.) for data cleaning, with a detailed procedure
639 illustrated in Supplementary Figure 6. The aim is to remove non-events including debris,
640 doublets, normalisation (EQ) beads and other undesired events like dead cells. The CD45+ cell
641 population was pre-gated using Cytobank for down-stream analysis.
642
643 ***Manual gating***
644
645 Manual gating for major cell populations was performed using Cytobank platform, and
646 gatingML files were exported and converted into gatingSet object to extract labels of each cell
647 via the R packages flowWorkspace, openCyto and CytoML. These data were used as reference
648 for benchmarking computer-aided algorithms.
649

650 ***Dataset availability***

651

652 In this study, we tested ImmCellTyper pipeline on the MPN cohort with 7 MPN patients and 2
653 healthy volunteers, influenza cohort with 11 patients and the COVID-19 cohort with 59
654 COVID-19 patients and 23 healthy volunteers. The FCS files (after clean-up and gating of CD45
655 population) and metadata of the MPN cohort were deposited in Zenodo (doi:
656 10.5281/zenodo.10076940); The influenza cohort was published by our lab(Alimam *et al.*),
657 and the data were stored in Zenodo (doi: 10.5281/zenodo.7982165); The COVID-19 dataset
658 was previously published by Chevrier et al. (Chevrier *et al.*), and the FCS files can be retrieved
659 from Mendeley Data (https://data.mendeley.com/datasets/vyy8ttw7n9/1).

660

661 **Author contributions**

662

663 Jing Sun, Conceptualization, Data curation, Software, Formal analysis, Supervision, Validation,
664 Visualisation, Methodology, Writing – original draft, Writing – review and editing, Resources,
665 Experiments; Desmond Choy, Conceptualization, Software, Supervision, Methodology,
666 Resources, Investigation; Nicolas Sompairac, Technical supervision, Resources, Data curation,
667 Writing – review and editing; Shirin Jamshidi, Technical supervision, Resources, Writing –
668 review and editing; Shahram Kordasti, Conceptualization, Resources, Funding acquisition,
669 Supervision, Writing – review and editing, Project administration; Michele Mishto: Writing –
670 review and editing

671

672 **Competing interests**

673

674 The authors declare that there are no competing interests exist.

675

676 **Acknowledgement**

677

678 This work is supported in part by: Blood Cancer UK [Ref. 22009] and CRUK City of London
679 Centre (CoL) Award [CTRQQR-2021/100004] to MM and SK; JS was supported by the K-CSC
680 scholarship.

681

682 We thank BRC flowCore at Guy's Hospital for technical support, Biobank at King's College
683 London for patient samples; We are thankful for Cynthia Bishop, Katrina Todd, Rosa Andres
684 Ejarque, Elena Torre, Robert Page, Rita Antunes Dos Reis and Mishto lab members at King's
685 College London, Daniel Dancer, Thomas Adejumo at Standard Biotools for technical assistance
686 and sample management, as well as all study participants for their kind support for this
687 research. Special thanks to Eliezer Van Allen and Jihye Park at Dana Farber Cancer Institute to
688 review and comment this work.
689

690 **References**

691 Abdelaal, T., van Unen, V., Höllt, T., Koning, F., Reinders, M.J.T., and Mahfouz, A. Predicting
692 Cell Populations in Single Cell Mass Cytometry Data.
693 Aghaeepour, N., Nikolic R Fau - Hoos, H.H., Hoos Hh Fau - Brinkman, R.R., and Brinkman, R.R.
694 Rapid cell population identification in flow cytometry data.
695 Alimam, S.A.-O., Ann Timms, J., Harrison, C.A.-O.X., Dillon, R., Mare, T., DeLavallade, H., Radia,
696 D.A.-O., Woodley, C., Francis, Y., Sanchez, K., et al. Altered immune response to the annual
697 influenza A vaccine in patients with myeloproliferative neoplasms.
698 Arend, L., Bernett, J., Manz, Q., Klug, M., Lazareva, O., Baumbach, J., Bongiovanni, D., and List,
699 M. A systematic comparison of novel and existing differential analysis methods for CyTOF
700 data. LID - bbab471 [pii] LID - 10.1093/bib/bbab471 [doi].
701 Arvaniti, E.A.-O., and Claassen, M. Sensitive detection of rare disease-associated cell subsets
702 via representation learning.
703 Becher, B.A.-O., Schlitzer, A., Chen, J., Mair, F., Sumatoh, H.R., Teng, K.W., Low, D., Ruedl, C.,
704 Riccardi-Castagnoli, P., Poidinger, M.A.-O., et al. High-dimensional analysis of the murine
705 myeloid cell system.
706 Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., and Nolan, G.P. Automated
707 identification of stratifying signatures in cellular subpopulations.
708 Chen, H., Lau, M.C., Wong, M.T., Newell, E.W., Poidinger, M., and Chen, J. (2016). Cytofkit: A
709 Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. PLoS Comput
710 Biol *12*, e1005112. 10.1371/journal.pcbi.1005112.
711 Cheng, L., Karkhanis, P., Gokbag, B., Liu, Y., and Li, L. (2022). DGCyTOF: Deep learning with
712 graphic cluster visualization to predict cell types of single cell mass cytometry data. PLOS
713 Computational Biology *18*, e1008885. 10.1371/journal.pcbi.1008885.
714 Chevrier, S., Zurbuchen, Y., Cervia, C., Adamo, S., Raeber, M.E., de Souza, N., Sivapatham, S.,
715 Jacobs, A., Bachli, E., Rudiger, A., et al. A distinct innate immune signature marks progression
716 from mild to severe COVID-19.
717 Kaushik, A., Dunham, D., He, Z., Manohar, M., Desai, M., Nadeau, K.C., and Andorf, S. (2021).
718 CyAnno: a semi-automated approach for cell type annotation of mass cytometry datasets.
719 Bioinformatics *37*, 4164-4171. 10.1093/bioinformatics/btab409.
720 Kimball, A.K., Oko, L.M., Bullock, B.A.-O., Nemenoff, R.A., van Dyk, L.A.-O., and Clambey, E.A.-
721 O. A Beginner's Guide to Analyzing and Visualizing Mass Cytometry Data.
722 Lee, H.C., Kosoy, R., Becker, C.E., Dudley, J.T., and Kidd, B.A. Automated cell type discovery
723 and classification through knowledge transfer.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir el, A.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis.

Li, H., Shaham, U., Stanton, K.P., Yao, Y., Montgomery, R.R., and Kluger, Y. (2017). Gating mass cytometry data by deep learning. Bioinformatics *33*, 3423-3430. 10.1093/bioinformatics/btx448.

Liu, P., Liu, S., Fang, Y., Xue, X., Zou, J., Tseng, G., and Konnikova, L. Recent Advances in Computer-Assisted Algorithms for Cell Subtype Identification of Cytometry Data.

Liu, X., Song, W., Wong, B.Y., Zhang, T., Yu, S., Lin, G.N., and Ding, X. A comparison framework and guideline of clustering methods for mass cytometry data.

Lun, A.T.L., Richard, A.C., and Marioni, J.C. Testing for differential abundance in mass cytometry data.

Marsh-Wakefield, F., Kruzins, A., McGuire, H.M., Yang, S., Bryant, C., Fazekas de St Groth, B., Nassif, N., Byrne, S.N., Gibson, J., Brown, C., et al. Mass Cytometry Discovers Two Discrete Subsets of CD39(-)Treg Which Discriminate MGUS From Multiple Myeloma.

McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Nowicka, M., Krieg, C., Crowell, H.L., Weber, L.M., Hartmann, F.J., Guglietta, S., Becher, B., Levesque, M.P., and Robinson, M.D. (2017). CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. F1000Res *6*, 748. 10.12688/f1000research.11622.3.

O'Brien Gore, C., Billman, A., Hunjan, S., Colebrook, J., Choy, D., Li, W., Haynes, J., Wade, J., Hobern, E., McDonald, L., et al. Pre-treatment with systemic agents for advanced NSCLC elicits changes in the phenotype of autologous T cell therapy products.

Opzoomer, J.W., Timms, J.A., Blighe, K., Mourikis, T.P., Chapuis, N., Bekoe, R., Kareemaghay, S., Nocerino, P., Apollonio, B., Ramsay, A.G., et al. (2021). ImmunoCluster provides a computational framework for the nonspecialist to profile high-dimensional cytometry data. Elife *10*. 10.7554/eLife.62915.

Qiu, P., Simonds Ef Fau - Bendall, S.C., Bendall Sc Fau - Gibbs, K.D., Jr., Gibbs Kd Jr Fau - Bruggner, R.V., Bruggner Rv Fau - Linderman, M.D., Linderman Md Fau - Sachs, K., Sachs K Fau - Nolan, G.P., Nolan Gp Fau - Plevritis, S.K., and Plevritis, S.K. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE.

Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L., and Nolan, G.P. Automated mapping of phenotype space with single-cell data.

Spitzer, M.H., and Nolan, G.P. Mass Cytometry: Single Cells, Many Features.

Theorell, A., Bryceson, Y.T., and Theorell, J.A.-O. Determination of essential phenotypic elements of clusters in high-dimensional entities-DEPECHE.

Trussart, M.A.-O., Teh, C.A.-O., Tan, T., Leong, L., Gray, D.A.-O., and Speed, T.A.-O. Removing unwanted variation with CytofRUV to integrate multiple CyTOF datasets. LID - 10.7554/eLife.59630 [doi] LID - e59630.

van der Maaten, L., and Hinton, G. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research *9*, 2579-2605.

Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., and Saeys, Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data.

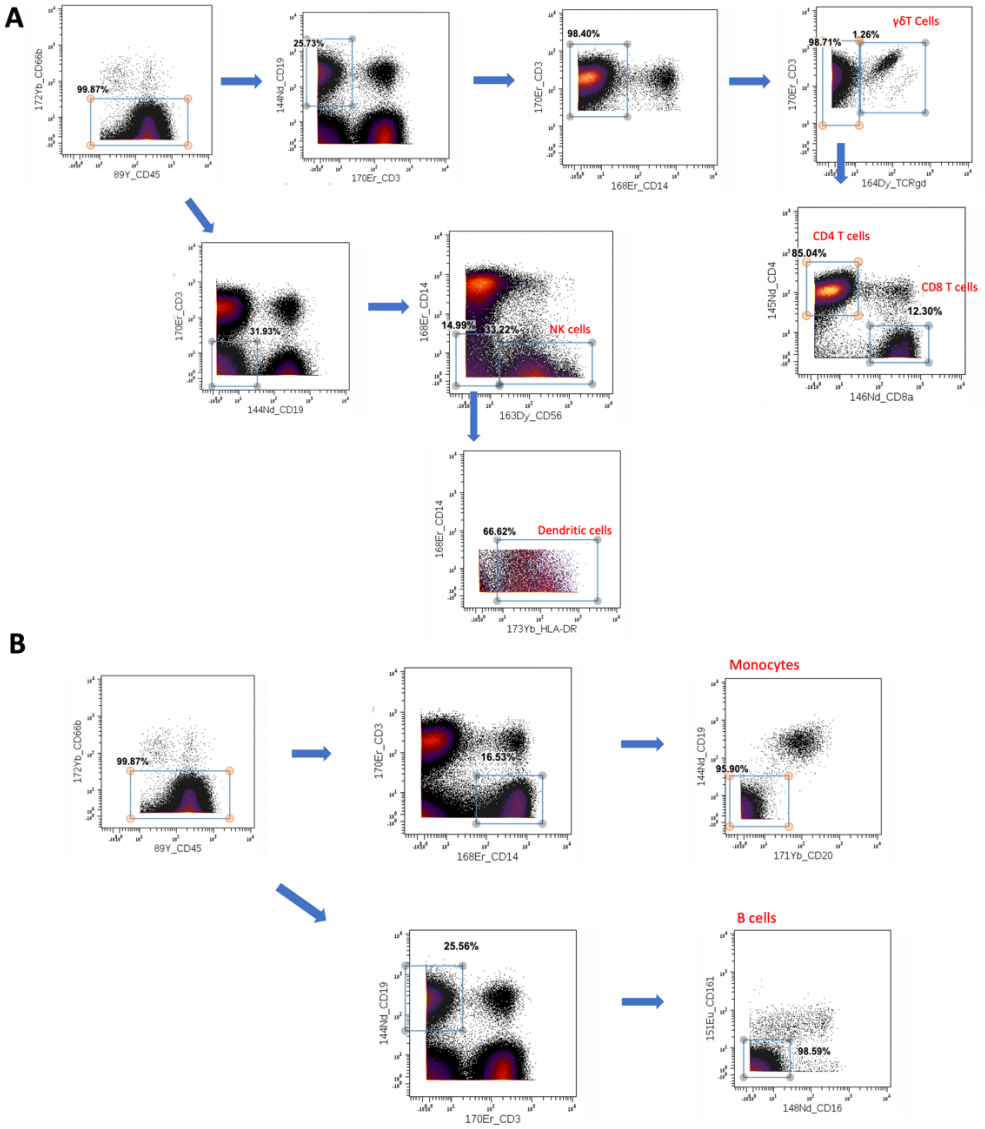
Van Gassen, S.A.-O., Gaudilliere, B., Angst, M.S., Saeys, Y., and Aghaeepour, N. CytoNorm: A Normalization Algorithm for Cytometry Data.

Weber, L.M., Nowicka, M., Soneson, C., and Robinson, M.D. (2019). diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. Communications Biology *2*, 183. 10.1038/s42003-019-0415-5.

Zhang, Z.A.-O., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E.W., Modrusan, Z., et al. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. LID - 10.3390/genes10070531 [doi] LID - 531.

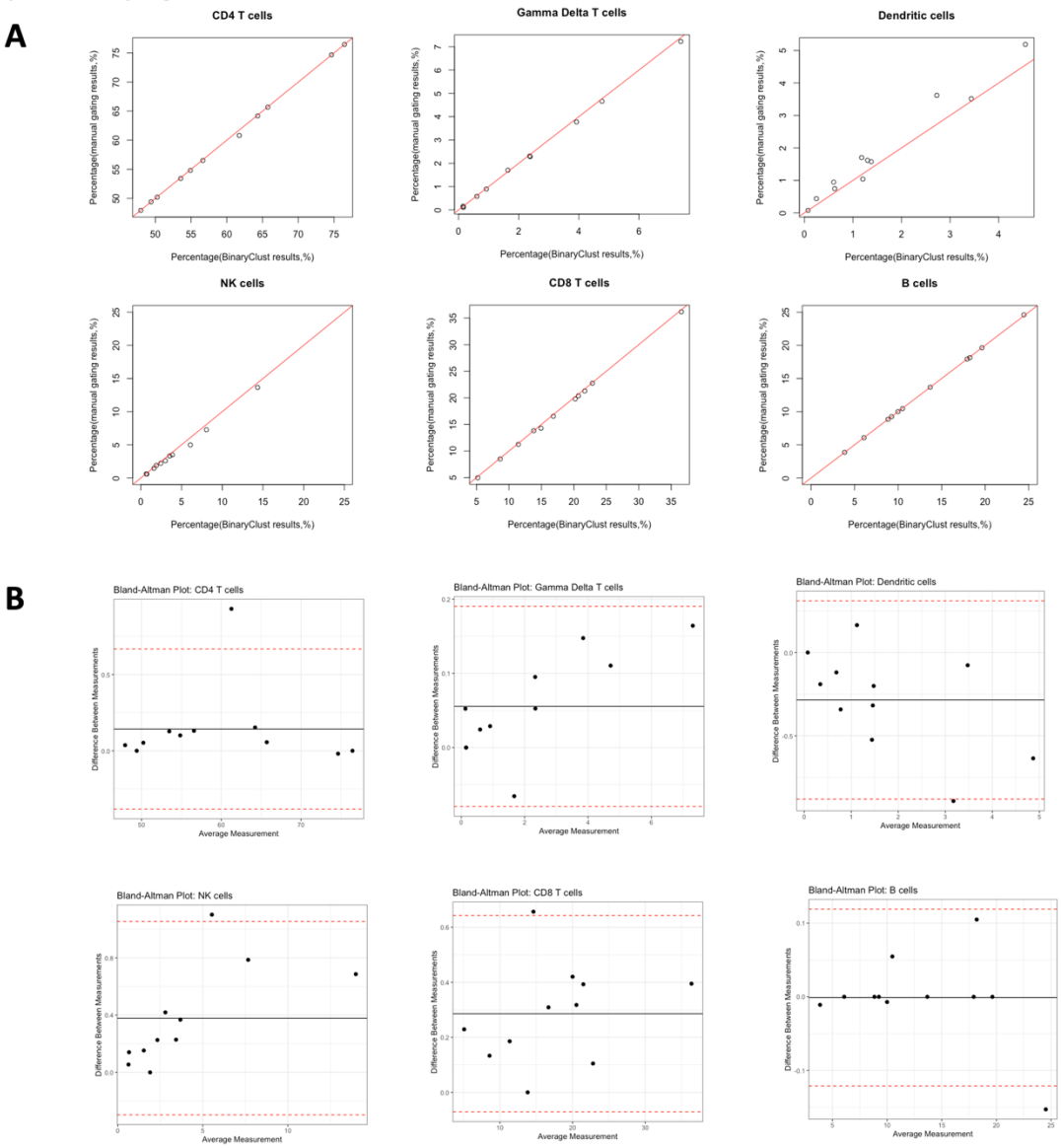
Supplementary Figures

Supplemental Figure 1



Supplementary Figure 1 Manual hierarchical gating strategy for main cell linages from human PBMC samples (MPN dataset, n=9). All .FCS files were cleaned up to remove doublets, normalisation beads and debris (please refer to the methods for standard clean-up procedure), and pre-gated for CD45+ leukocytes. (A) Serial bi-axial scatter plots representing the gating diagram for T cell subsets (CD4 T cells, CD8 T cells and gamma-delta T cells), NK cells and dendritic cells based on the indicated phenotypic markers; (B) Serial bi-axial scatter plot indicating the gating strategy to isolate monocytes and B cells from leukocytes. All manual gating was done using Cytobank platform(https://premium.cytobank.org/cytobank/).

Supplementary Figure 2

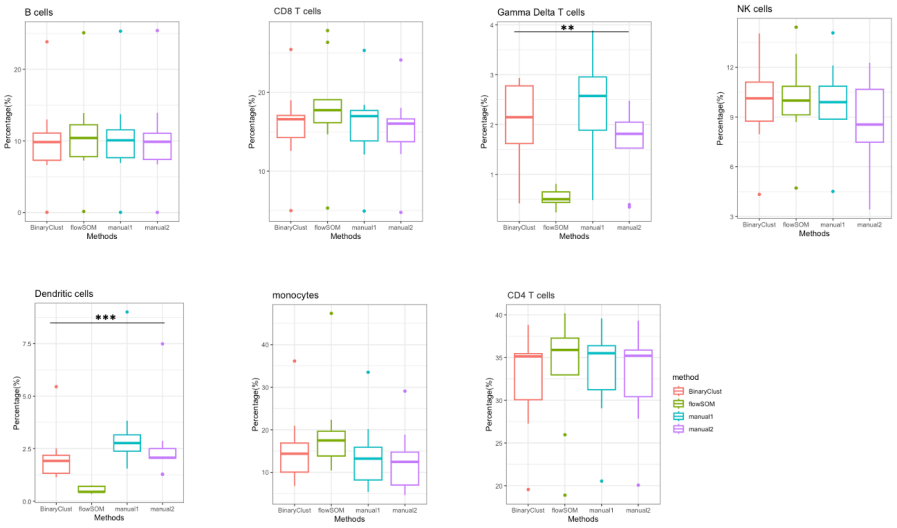


C

	Gated population	Counts	Cluster Cell Counts	True Positive	Precision	Recall	F-measure	Average F-measure	ARI	Average ARI
BinaryClust2	CD4 T cells	109498	109647	109404	1.00	1.00	1.00	0.98	0.95	0.91
	CD8 T cells	55449	56132	55435	0.99	1.00	0.99		0.95	
	NK cells	10759	11506	10748	0.93	1.00	0.97		0.97	
	B cells	16892	16881	16862	1.00	1.00	1.00		0.95	
	Dendritic cells	5090	4426	4417	1.00	0.87	0.93		0.74	
	TCRgd T cells	6848	7026	6837	0.97	1.00	0.99		0.91	

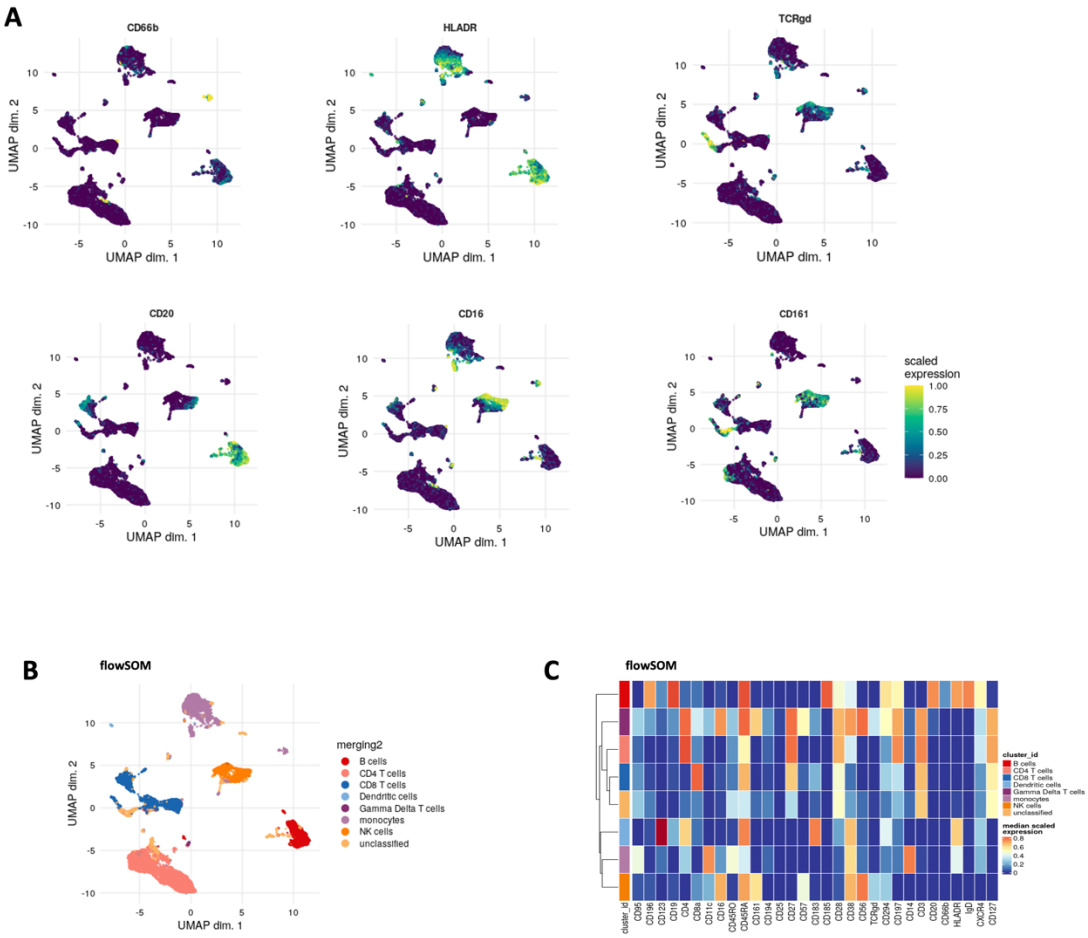
Supplementary Figure 2 Agreement evaluation between ImmCellTyper and manual gating in influenza dataset(n=11). Manual gating was performed using Cytobank and exported manually. (A) Correlation plots between ImmCellTyper results and manual gating results concerning percentages in CD4 T cells, Gamma Delta T cells, dendritic cells, NK cells, CD8 T cells and B cells, with red line indicating perfect agreement (correlation coefficient =1); (B)Bland-Altman plots of the two measurements in the indicated populations, with black line suggesting mean difference between measurements and dotted red line indicating limits of agreement (1.96 x standard deviations). (C) Calculation of precision, recall, F-measure for ImmCellTyper method in comparison to manual gating in the indicated cell populations.

Supplementary Figure 3



Supplementary Figure 3 Boxplots of the indicated cell percentages generated by different methods. Statistical significance was marked by asterisk. * P<0.05, **P<0.01, ***P<0.001.

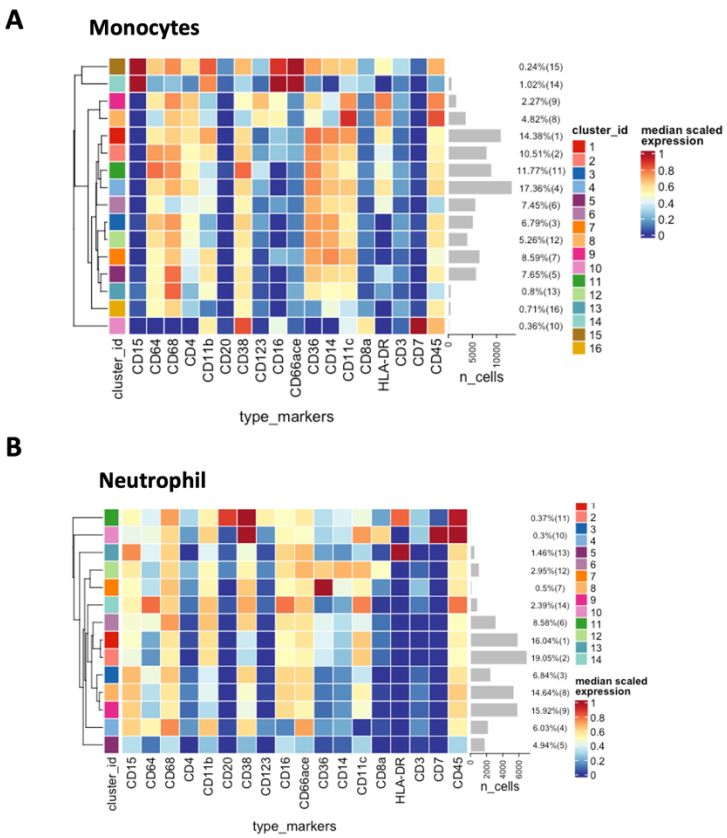
Supplementary Figure 4



Supplementary Figure 4 (A) UMAP plots of normalized expression of indicated markers (CD66b, HLADR, TCRgd, CD20, CD16, CD161) across 2000 cells per sample in MPN dataset;(B) FlowSOM

806 clustering was performed on the same dataset to compare with BinarClust, k=20 was chosen followed
807 by manual annotation of each cluster. UMAP plot was projected with merged flowSOM clusters with
808 biological annotation (downsample 2000 cells per sample) ;(B) The corresponding median marker
809 expression heatmap after flowSOM clustering and annotation.

Supplementary Figure 5

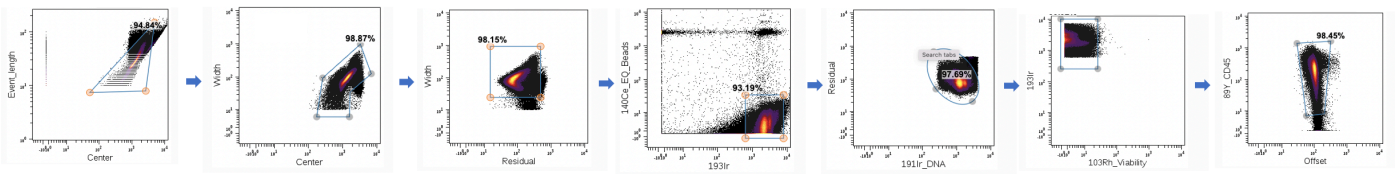


810

811 **Supplementary Figure 5** Marker expression heatmap of (A)monocytes and (B) neutrophils generated
812 by Phenograph clustering.

813

Supplementary Figure 6



814

815 **Supplementary Figure 6** Clean-up procedure of CyTOF data using Cytobank.

816

817

818

819

820 **Supplementary Table 1** CyTOF antibody panel for MPN cohort

Antibody	Clone	Metal tag	Manufacturer
Anti-CD45	HI30	89Y	Standard BioTools
Live/dead Indicator	N/A	103Rh	Standard BioTools
Anti-CD95	DX2	APC	Biolegend
Anti-APC	APC003	106Cd	Standard BioTools
Anti-CD196	G034E3	141Pr	Standard BioTools
Anti-OX40	ACT35	142Nd	Standard BioTools
Anti-CD123	6H6	143Nd	Standard BioTools
Anti-CD19	HIB19	144Nd	Standard BioTools
Anti-CD4	RPA-T4	145Nd	Standard BioTools
Anti-CD8a	RPA-T8	146Nd	Standard BioTools
Anti-CD11c	Bu15	147Sm	Standard BioTools
Anti-CD16	3G8	148Nd	Standard BioTools
Anti-CD45RO	UCHL1	149Sm	Standard BioTools
Anti-CD45RA	HI100	150Nd	Standard BioTools
Anti-CD161	HP-3G10	151Eu	Standard BioTools
Anti-CD194	L291H4	152Sm	Standard BioTools
Anti-CD25	BC96	153Eu	Standard BioTools
Anti-CD27	O323	154Sm	Standard BioTools
Anti-CD57	HCD57	155Gd	Standard BioTools
Anti-CD183	G025H7	156Gd	Standard BioTools
Anti-CD185	J252D4	158Gd	Standard BioTools
Anti-TIM-3	F38-2E2	159Tb	Standard BioTools
Anti-CD28	CD28.2	160Gd	Standard BioTools
Anti-CD38	HB-7	161Dy	Standard BioTools
Anti-CD56	NCAM16.2	163Dy	Standard BioTools
Anti-TCRgd	B1	164Dy	Standard BioTools
Anti-PD-1	EH12.2H7	165Ho	Standard BioTools
Anti-CD294	BM16	166Er	Standard BioTools
Anti-CD197	G043H7	167Er	Standard BioTools
Anti-CD14	63D3	168Er	Standard BioTools
Anti-ICOS	C398.4A	169Tm	Standard BioTools
Anti-CD3	UCHT1	170Er	Standard BioTools
Anti-CD20	2H7	171Yb	Standard BioTools

Anti-CD66b	G10F5	172Yb	Standard BioTools821
Anti-HLA-DR	LN3	173Yb	Standard BioTools
Anti-IgD	IA6-2	174Yb	Standard BioTools
Anti-CXCR4	12G5	175Lu	Standard BioTools
Anti-CD127	A019D5	176Yb	Standard BioTools
Anti-TIGIT	MBSA43	209Bi	Standard BioTools