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 [1]. 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 [2]. 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[3], Phenograph[4], X-shift[5], spade[6], and DensVM[7] etc., often combine with dimension reduction techniques like t-SNE[8], UMAP[9], 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[10]. Additionally, the technical

97 uncertainty of results produced by different clustering approaches remains a conundrum.
98 Even for the same unsupervised method, the discrepancy among different runs without
99 setting a seed, reduces the reproducibility and may cause confusion, particularly for biologists
100 with limited computational knowledge. Moreover, manual validation is also essential, as
101 biological annotation is the inevitable step to provide biological relevant labels for the clusters.
102 However, this process can be time-consuming and subjective, hindering the automation of
103 pipelines. This problem is particularly pronounced for a large marker panel and samples with
104 high heterogeneity, resulting in a higher number of clusters that need to be annotated.
105
106 Advances in artificial intelligence have accelerated the development of alternative clustering
107 methods in a supervised manner for cell type inference. These methods consider the 'ground
108 truth' or prior knowledge about the marker expression of each given cell types to
109 automatically label each cell. Currently, a couple of semi-automatic methods have been
110 developed, such as Linear Discriminant Analysis(LDA)[11], DGCyTOF[12], CyAnno[13], DeepCyTOF[14],
111 Automated Cell-type Discovery and Classification (ACDC) [15]and Semi-supervised Category
112 Identification and Assignment(SCINA) [16] etc. ACDC[15] and SCINA[16] use a matrix of pre-defined
113 markers for each cell type to annotate the clusters showing the same signature. These
114 methods assume that markers are either expressed or not expressed (binary), which limits
115 their ability to distinguish cell subtypes with similar phenotypes, particularly non-canonical
116 cell types that cannot be easily separated linearly. Alternatively, DeepCyTOF and LDA use a
117 marker expression matrix extracted from manually gated cell types as a training dataset to
118 build a machine learning model for cell type prediction. This approach has higher precision
119 and accuracy compared with aforementioned methods[10]. Nonetheless, it can be labour
120 intensive for preparation of the training set manually. Also, these methods are limited in their
121 ability to predict novel cell subsets beyond the pre-gated set of cell types and lack a
122 systematic and comprehensive way to assign the cells which were not identifiable under any
123 of the gated cell types. New solutions have emerged with algorithms such as DGCyTOF[12] and
124 CyAnno[13], the former adopts a deep learning classification combined with hierarchical stable-
125 clustering methods and an iteration calibration system to identify known cell types and assign
126 novel subsets; while CyAnno is based on a machine-learning framework which allows the
127 integrative modelling of both 'gated' and 'ungated' cells. Both methods have demonstrated
128 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[17], CapX[18], Cytofkit[19] and ImmunoCluster[20] 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

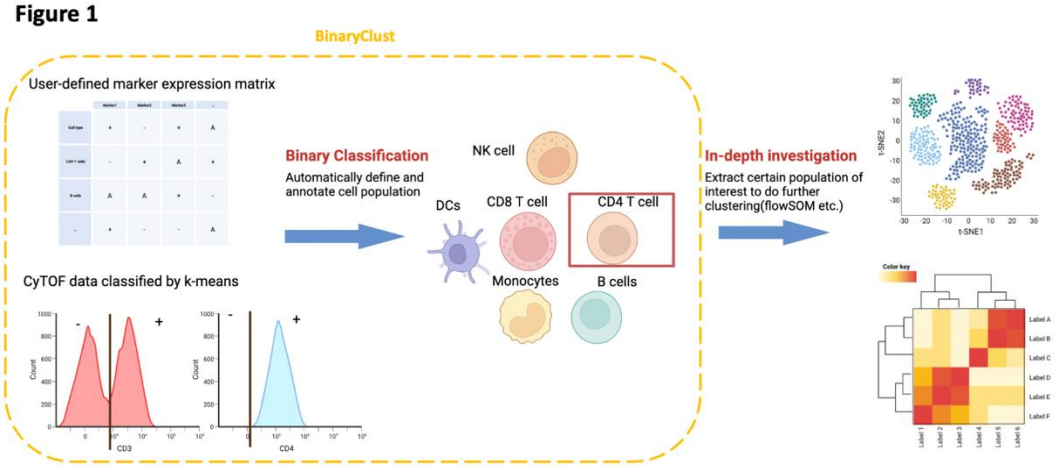
BinaryClust is comparable with manual gating in quantifying the abundances of main cell lineages.

The core concept for BinaryClust is depicted in Figure1, where a simple user-designed cell type marker expression matrix is required and serves as a reference to characterise the main cell lineages, with positive markers indicated in '+', negative markers in '-', and irrelevant markers in 'A'. K-means(k=2) will be applied to divide the positive and negative cell population of each marker, then align it to the reference table to infer main cell types. This is

161 followed by the extraction of population-of-interest for downstream clustering using
162 unsupervised methods for subpopulation discovery.
163
164 To assess the performance of the automated cell type classification and prediction, we
165 generated a test CyTOF dataset using peripheral blood mononuclear cell (PBMC) samples
166 from seven patients with myeloproliferative neoplasm (MPN) and two healthy donors,
167 employing a 37-marker deep immunophenotyping panel (Supplementary Table 1). Manual
168 gating was performed by two independent experts which identified seven main cell lineages.
169 These results act as the reference for evaluating the computer-aided methods. The
170 hierarchical sequential gating strategy was explicitly illustrated in Supplementary Figure 1. We
171 evaluated the agreement between manual gating results and BinaryClust results regarding
172 cell frequencies of each population. The mean value was calculated from manual gating
173 results of two experts to compare with BinaryClust-generated results using Pearson
174 correlation analysis. As shown in Figure2A, the two methods exhibit a strong correlation with
175 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
176 cells, NK cells, monocytes, and gamma delta T cells, respectively (all $P < 0.0001$). And most of
177 the data points remain close to the line of equality (red line, $R^2=1$), indicating a high degree of
178 agreement. Meanwhile, the Bland-Altman plots in Figure 2B also suggest no consistent bias of
179 manual gating versus BinaryClust across all the identified cell types. The good performance of
180 BinaryClust was further validated in the influenza dataset published by our group[21], which
181 contains FCS files from 11 individuals with 6 main immune cell types detected, the average F-
182 measure and ARI reached 0.98 and 0.91 respectively (Supplementary Figure2A-B).

183

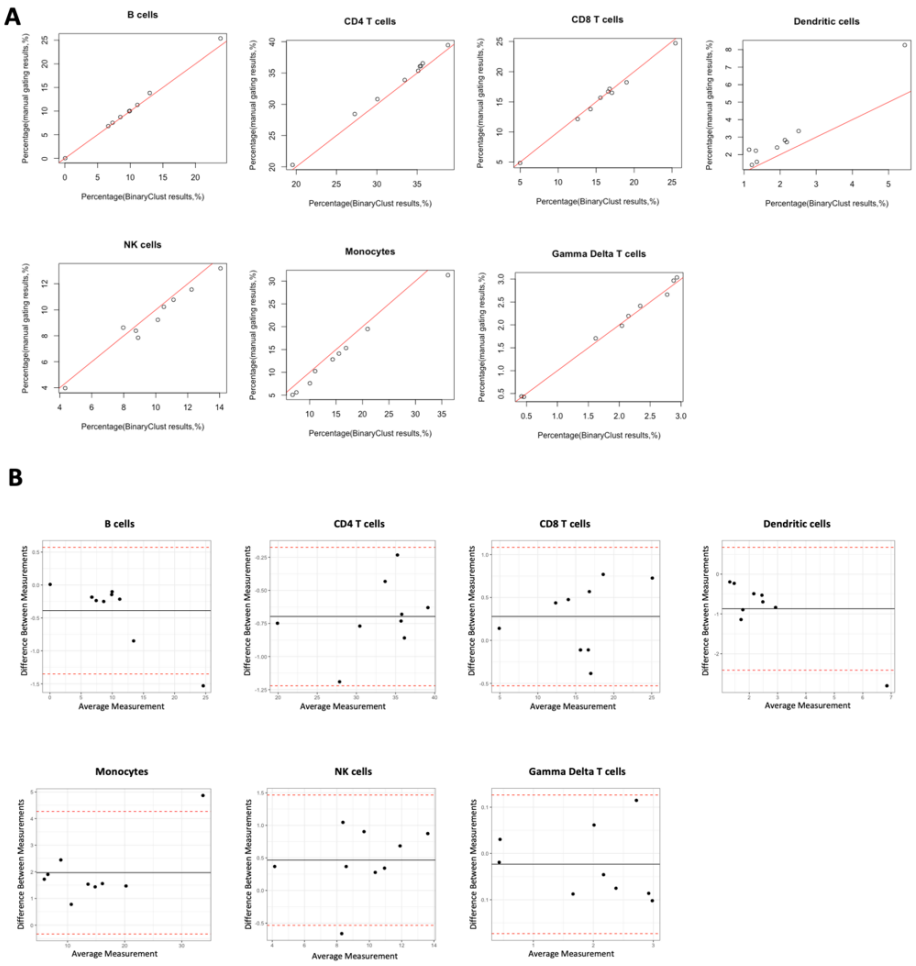
Figure 1



184 **Figure 1** Schematic diagram of BinaryClust framework. Semi-supervised classification is first performed
185 on selected markers in the user-defined marker expression matrix to classify and annotate major cell
186 types. Population-of-interest can be further extracted and explored using unsupervised clustering
187 methods followed by differential analysis. Figure was generated by BioRender
188 (https://www.biorender.com/).
189 **Figure 2** Agreement evaluation comparing manual gating and BinaryClust in MPN cohort(n=9). Manual
190 gating of B cells, CD4 T cells, CD8 T cells, dendritic cells, NK cells, monocytes and gamma delta T cells
191 were performed by two independent experts using Cytobank, and mean values of the population
192 percentages were calculated to compare with BinaryClust results. Each dot represents one patient
193 sample. (A) Scatter plot showing the correlation between the two methods, with the red line
194 indicating perfect agreement (correlation coefficient = 1); (B) Bland-Altman plots of the two
195 measurement methods among all the cell populations, with the black line suggesting the mean
196 observed difference and red dotted lines indicating limits of agreement (1.96 x standard deviations).
197

198 **BinaryClust achieves high accuracy and speed compared with flowSOM and LDA**

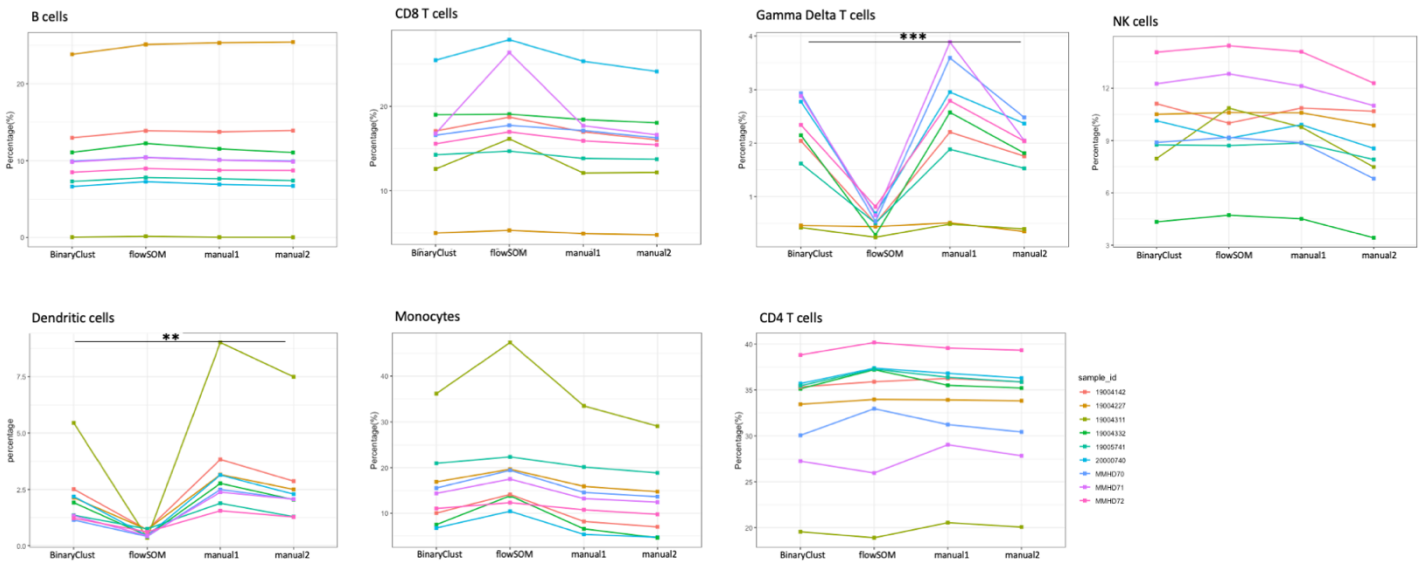
199 **Figure 2**



200 To further evaluate
202 BinaryClust's performance, we compared it with the well-performing unsupervised algorithm

203 flowSOM and supervised classifier Linear discriminant analysis (LDA). FlowSOM was run on
204 the same MPN dataset with k value set as 20, followed by manual annotation and cluster
205 merging to identify the same cell populations as in manual gating and BinaryClust. The cell
206 frequencies derived from BinaryClust, flowSOM, and manual gating (expert1 and expert2)
207 were compared using interaction plot (Figure 3). We observed that BinaryClust remains
208 consistent with manual gating (all P>0.05), whereas flowSOM identified significantly less
209 Gamma Delta T cells and dendritic cells (P<0.001 and P=0.006, respectively) compared with
210 the other three measurements. We also increased the initial k value to 40 to improve
211 accuracy by over clustering then merging clusters [3]. We found that the average frequency for
212 Gamma Delta T cells and dendritic cells increased to 1.5% and 4.65% respectively but still
213 remained significantly different (P<0.05) from the other two methods. It is also interesting to
214 see that although the same gating strategy was applied, different experts obtained results
215 with slight variation despite no statistical significance. The same findings were confirmed
216 using boxplot as shown in Supplementary Figure 3.

Figure 3



217 **Figure 3** Comparison of manual gating (manual1 and manual2), BinaryClust and flowSOM clustering
218 results in MPN cohort(n=9). Interaction plots showing the individual measurement (percentage) of
219 each study participant with indicated colors by different methods across main cell lineages (B cells,
220 CD8 T cells, Gamma Delta T cells, NK cells, Dendritic cells, monocytes and CD4 T cells); ANOVA was
221 used for statistical testing, and significance was marked by asterisk. * P<0.05, **P<0.01, ***P<0.001,
222 ****P<0.0001.
223

224 Subsequently, LDA was also tested on the same dataset. Since this method requires a training
225 dataset from manual gating to build a model, we exported the cell events and cluster
226 assignment from Cytobank, then equally partitioned all cells into training and test dataset.
227 Due to the difference in method implementation, we did not include LDA for cell abundance
228 quantification comparison with other methods, since only half of the cells were used for
229 prediction. To ensure an equitable comparison and further evaluate the accuracy, F-measure,
230 and Adjusted Rand Index (ARI) were calculated using manual gating cluster IDs as reference.
231 As indicated in Table 1, for the MPN dataset, BinaryClust achieved high F-measure and ARI in
232 all 7 cell types with an average of 0.94 and 0.91 respectively. The performance remains
233 excellent in influenza dataset as well (average F-measure = 0.98, Supplementary Figure 2C).
234 LDA has equivalent prediction accuracy as BinaryClust with an average F-measure of 0.93 and
235 ARI of 0.90. Both supervised methods outperformed flowSOM (average F-measure=0.75,
236 average ARI=0.66), which is in line with previous benchmarking study on evaluating
237 supervised and unsupervised clustering algorithms[10]. Notably, for cell types that constitute a
238 substantial proportion in the pool like CD4 T cells, CD8 T cells, NK cells, monocytes, and B cells,
239 flowSOM can identify them with high precision and sensitivity, while this is not the case for
240 dendritic cells and gamma delta T cells which account for less than 3% of the whole
241 population (F-measure = 0.51 and 0.15, respectively). Even so, the overall F-measure and ARI
242 for flowSOM considering cell proportion reached 0.81 and 0.80.

243 **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
Bina Clust2	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	

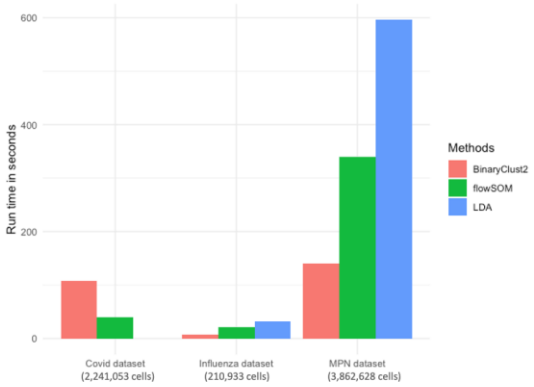
244 ARI, Adjusted Rand Index; LDA, Linear discriminant analysis

245

246 It has been demonstrated that flowSOM and LDA are among the fastest clustering algorithms
247 without compromising their performance. Here, we compared the speed amongst the three
248 approaches in MPN, Influenza and COVID-19 datasets[22]which contain 2,231,053, 210,933 and

249 3,862,628 cells respectively. Since LDA requires training data to be executed, we did not run it
250 on COVID-19 dataset due to the absence of manual gating results. As shown in Figure 4,
251 BinaryClust exhibited the highest speed in both MPN and Influenza dataset but fell behind
252 flowSOM slightly in COVID-19 dataset.

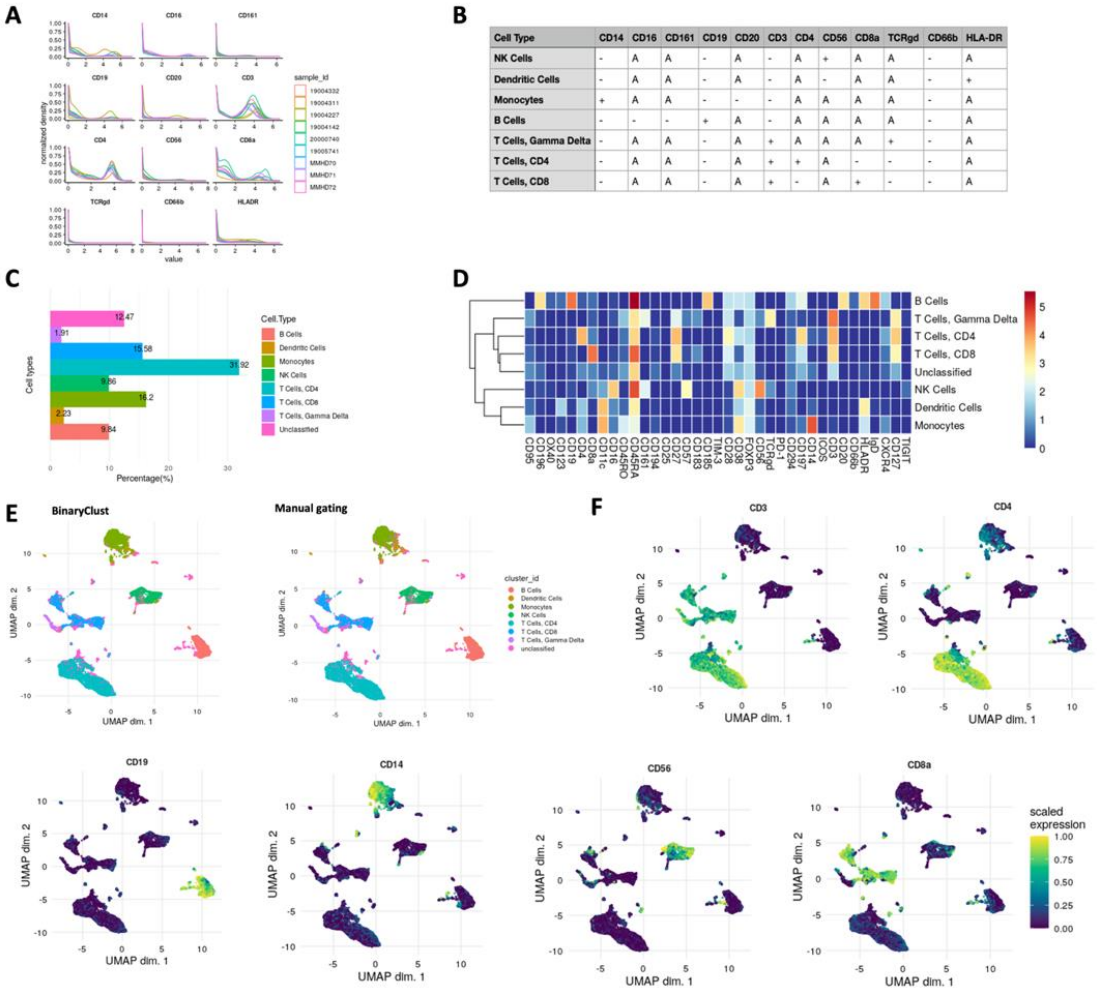
Figure 4



253
254 **Figure 4** Comparison of BinaryClust, flowSOM and LDA on speed. Bar chart showing runtime (in
255 seconds) of the three methods in three different datasets.
256
257 **ImmCellTyper pipeline supports interactive data visualisation and comparison among study**
258 **groups.**
259
260 BinaryClust is an important component of ImmCellTyper, a comprehensive integrated pipeline
261 designed for systematic CyTOF data mining. Hence, we utilised the visualisation functions to
262 further prove the reliability and robustness of BinaryClust.
263
264 BinaryClust inherently considers CyTOF makers as binary distributed which is in most of the
265 case, but not always. Therefore, it is crucial to check the marker behaviour before running
266 the pipeline. As shown in Figure 5A, all markers selected for the classification matrix (Figure
267 5B) displayed a binary distribution implying the suitability for this pipeline. After clustering
268 (Figure 5C), median marker expression heatmap (Figure 5D) can check the reliability of the
269 results before proceeding for downstream analysis. In MPN dataset, we projected the cluster
270 assignment resulted from BinaryClust and manual gating to UMAP (Figure 5E), along with the
271 expression of the phenotypic markers (Figure 5F and supplementary Figure 4A). High
272 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

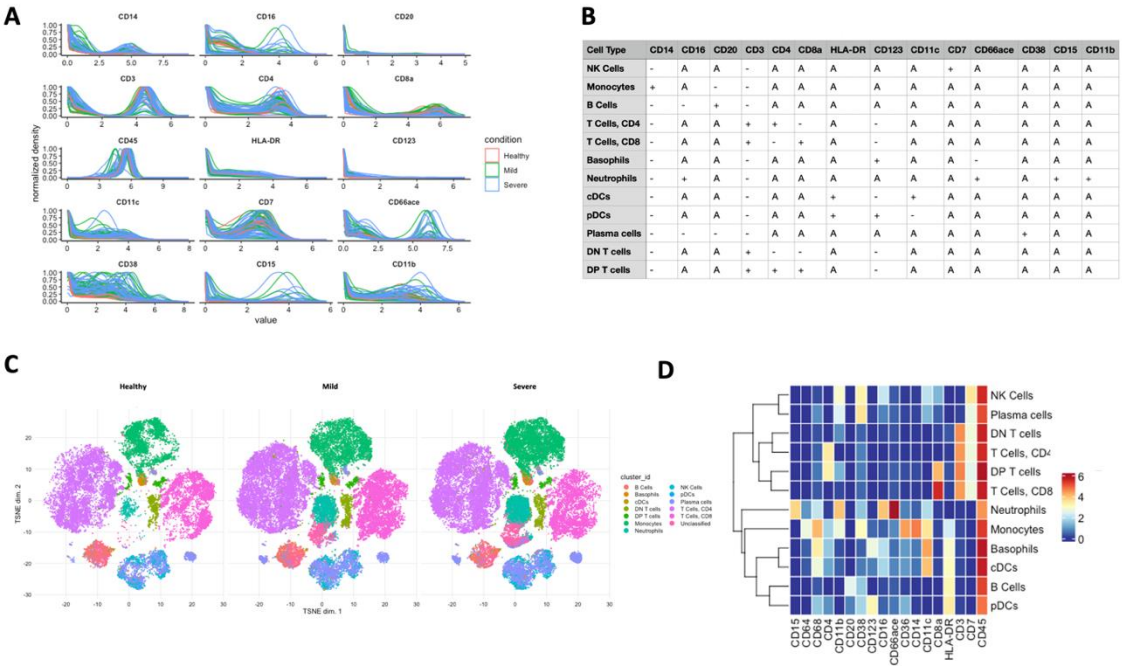


Cell Type	CD14	CD16	CD161	CD19	CD20	CD3	CD4	CD56	CD8a	TCRgd	CD66b	HLA-DR
NK Cells	-	A	A	-	A	-	A	+	A	A	-	A
Dendritic Cells	-	A	A	-	A	-	A	-	A	A	-	+
Monocytes	+	A	A	-	-	-	A	A	A	A	-	A
B Cells	-	-	-	+	A	-	A	A	A	A	-	A
T Cells, Gamma Delta	-	A	A	-	A	+	A	A	A	+	-	A
T Cells, CD4	-	A	A	-	A	+	+	A	-	-	-	A
T Cells,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

289 **Application of ImmCellTyper pipeline to the COVID-19 dataset demonstrates its versatile**
290 **functionalities for comprehensive data analysis**
291
292 To showcase the analytical and visualisation functions of ImmCellTyper pipeline, we used the
293 dataset published by Chevrier et al. [22], which described the immune signature of mild and
294 severe COVID-19 patients in comparison with healthy individuals. There are a total of 82 FCS
295 files with a 40-plex marker panel focusing on innate immunity in this dataset. An initial
296 marker expression check was carried out based on the user-defined matrix (Figure 6B) and
297 displayed in Figure 6A. BinaryClust was then performed and identified 12 cell populations as
298 expected. We used t-distributed stochastic neighbour embedding(t-SNE) in this dataset for
299 dimension reduction, coloured by cell types and faceted by disease conditions, which
300 exhibited substantial immune alteration among healthy control, mild and severe COVID-19
301 patients (Figure 6C). From the heatmap in Figure 6D, we can have an overview of the marker
302 expression of each population, which remains consistent with our initial definition indicated
303 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

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

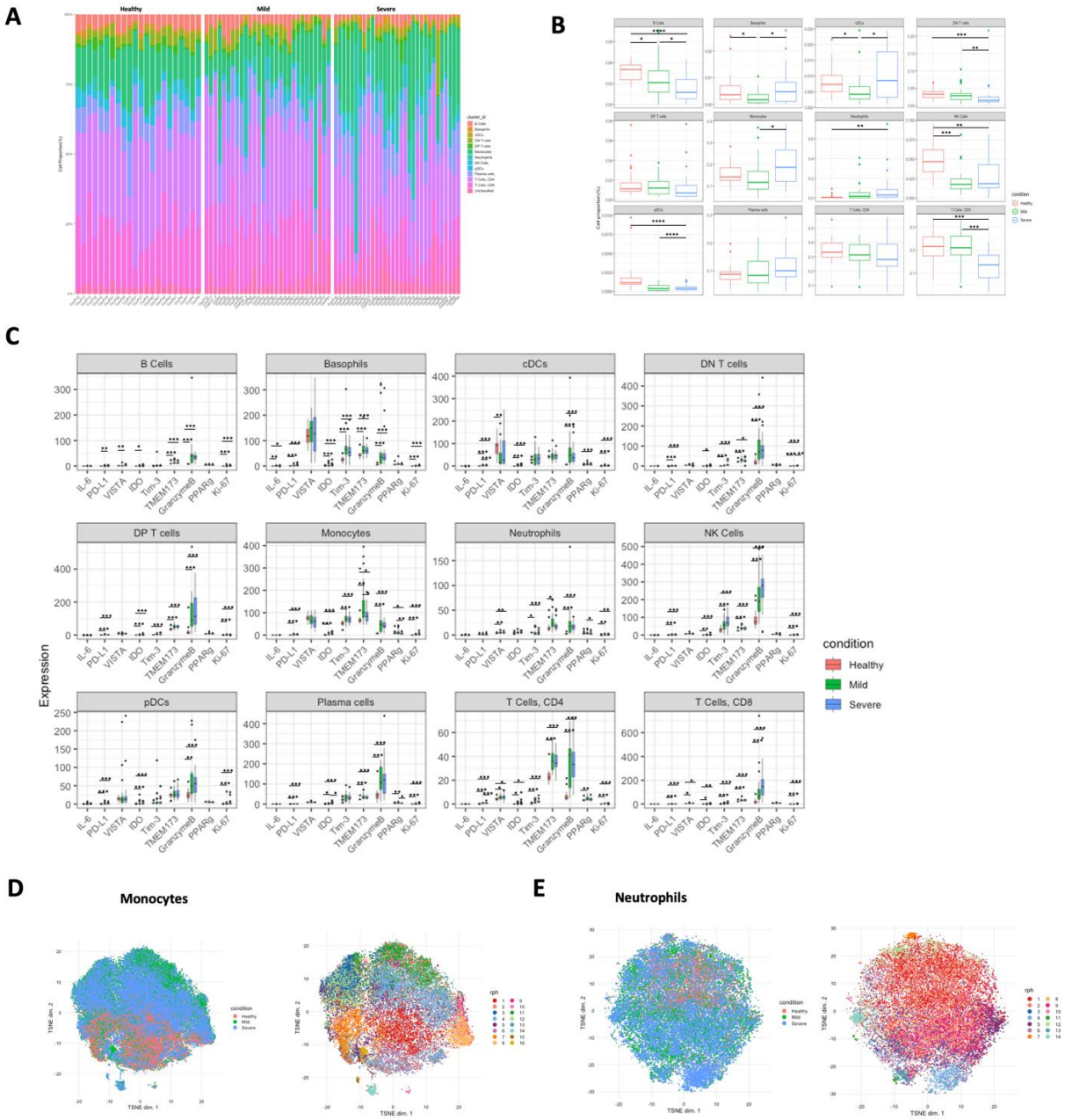


Figure 7 Quantification and statistical analysis comparing the study conditions in COVID-19 dataset(n=82). (A)Stacked histogram of main cell type composition per individual generated by BinaryClust, and grouped by study conditions (healthy, mild and severe); (B)Boxplots representing cell abundance frequencies 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

349 **Discussion**

350

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

364

365 For validating the robustness of the automated cell characterization function of ImmCellTyper:
366 BinaryClust, we involved two independent datasets (MPN dataset and Influenza dataset) and
367 compared with existing clustering tools flowSOM and LDA. Previous Benchmarking studies
368 indicated that flowSOM was one of the best-performing unsupervised tools in precision,
369 speed and stability, with F-measure ranges from 0.58~0.90 in various datasets[10], and was
370 widely used in high-impact research publications[10,23]. However, the results of flowSOM vary
371 upon the k values set by the users which causes some uncertainty and impairs reproducibility
372 of the results. In contrast, semi-supervised clustering methods like ACDC, DeepCyTOF[14] and
373 LDA outperformed the unsupervised methods in the same benchmarking study. The F-
374 measure obtained by ACDC and LDA ranged from 0.78~0.99, which was significantly higher
375 than unsupervised approaches including Accense, PhenoGraph, Xshift, k-means, flowMeans[24],
376 FlowSOM and DEPECHE[25](F-measure: 0.28~0.93)[23]. In this study, we chose LDA because it has
377 proven to be superior to ACDC and DeepCyTOF in terms of precision and speed[10,11]. The
378 excellent performance of LDA was demonstrated in MPN dataset of this study with F-measure
379 reached 0.93 for around 4 million cells. BinaryClust was comparable in accuracy(F-
380 measure=0.94) but much faster in speed. Both semi-supervised approaches outperformed

381 flowSOM as expected. Although semi-supervised methods seem to perform better, the
382 mainstream for CyTOF data analysis still relies on unsupervised methods. The reason for this
383 might be that most of the semi-supervised methods rely on manual gating as reference and
384 requires either the manually gated expression matrix or user-defined binary matrix as training
385 data. This process takes extra time and efforts especially for defining subpopulations, and
386 may introduce end-user bias. Additionally, the common restriction for all the supervised
387 methods is that they have limited capability to reveal novel subsets as those are not
388 predefined in the training pool, which poses a critical challenge for the pursuit of novel
389 discoveries. Considering all these strength and limitations, as a semi-supervised method,
390 BinaryClust possesses the advantages of incorporating prior biological knowledge, being
391 reproducible across runs and accurate on cell type identification. Meanwhile, the input
392 requirement is a simple matrix of cell population definition based on marker positivity,
393 avoiding the hassle to manually gate example FCS files. With a deep understanding that
394 certain CyTOF markers are expressed on a continuum rather than binary manner, this
395 automated supervised approach will only be used for classification of main cell lineages. The
396 test datasets in this study contained data from human PBMC samples which represented a
397 heterogenous pool of immune cells, covering the major types of various immune populations,
398 and demonstrated excellent performance of BinaryClust on classifying CD4 T cells, CD8 T cells,
399 gamma delta T cells, NK cells, monocytes, dendritic cells, and B cells. We also recommend
400 checking marker expression distribution before running the pipeline to ensure accurate
401 results. After the above step, this pipeline supports cell population extraction for further
402 dissection into subclusters using unsupervised approaches, which excel in detecting rare and
403 refined subpopulations. Phenograph is a particularly effective tool for this purpose. By
404 applying this strategy, our previous work with BinaryClust facilitated the evaluation of the
405 impact of systematic anti-cancer agents on lymphocyte population in non-small cell lung
406 cancer patients, as well as the characteristics of autologous T cell products after
407 manufacturing [26].
408
409 Another purpose behind the development of the ImmCellTyper framework is to cater to
410 general analysis needs, spanning from pre-processing to downstream differential analysis.
411 There are existing well-established pipelines like CATALYST[17], diffcyt[27], Cytofkit[19], and
412 ImmunoCluster[20] etc., each of them has its own strength and limitations. One of the

413 challenges existed in CyTOF data pre-processing is effectively integrating multiple batches,
414 given the technical differences arising from experiments and instrumental acquisition, which
415 could result in the separation of clustering across batches and potentially confounding the
416 signal of interest. In ImmCellTyper pipeline, we provide the functionality for batch effects
417 examination and incorporated two well-performing batch correction algorithms CytoNorm[28]
418 and CytofRUV[29] into our framework prior to downstream analysis, which ensures the quality
419 of the data and, to the best of our knowledge, makes it the first integrated pipeline with this
420 module. For the downstream analysis, tools like diffcyt and CATALYST group markers into
421 phenotypic and state makers, enabling the detection of differentially abundant cell clusters
422 and differential expression of functional markers within each cell population[27]. This well-
423 recognised strategy has been extensively applied in various studies. But one limitation is that
424 it directly classifies cells into high-resolution clusters, and cannot automatically merge
425 subclusters into one major population with similar phenotypes[27]. Our pipeline addresses this
426 problem via a precise and convenient solution: using the semi-supervised classifier
427 (BinaryClust) of ImmCellTyper. The two pipelines (CATALYST and ImmCellTyper) are well
428 compatible, allowing users to leverage their functions together and providing a broader range
429 of analytical options. On the other hand, Cytofkit has the advantages of the graphical user
430 interface(GUI) for non-specialists, integration of a variety of clustering(DensVM[7], ClusterX,
431 FlowSOM, Phenograph), dimension reduction(PCA, ISOMAP, t-SNE) methods, and inference of
432 the relatedness among cell populations, but it does not support complex study design, group
433 comparison and statistical testing[19]. Several methods have been developed to fit with the
434 high-dimensional features of CyTOF data for differential analysis including diffcyt, CellCnn[30],
435 cydar[31], citrus[32], cyEMD[33] etc. Nonetheless, none of them is applicable to compare more than
436 two study groups at once which involves multiple testing. Therefore, to accommodate study
437 designs involving multiple groups, ImmCellTyper does statistics via first applying Kruskal
438 Wallis test, followed by Benjamin-Hochberg (BH) procedure for multiple testing correction
439 and post hoc analysis via Dunn's test or pairwise-Wilcoxon test. In cases where the
440 comparison involves only two groups, Mann-Whitney test will be applied in terms of cell
441 frequency, as the distribution of CyTOF data generally does not fit for normal distribution.
442
443 One limitation of this computational pipeline is that its semi-supervised cluster identification
444 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.

475

476 ***R package ImmCellTyper***

477

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

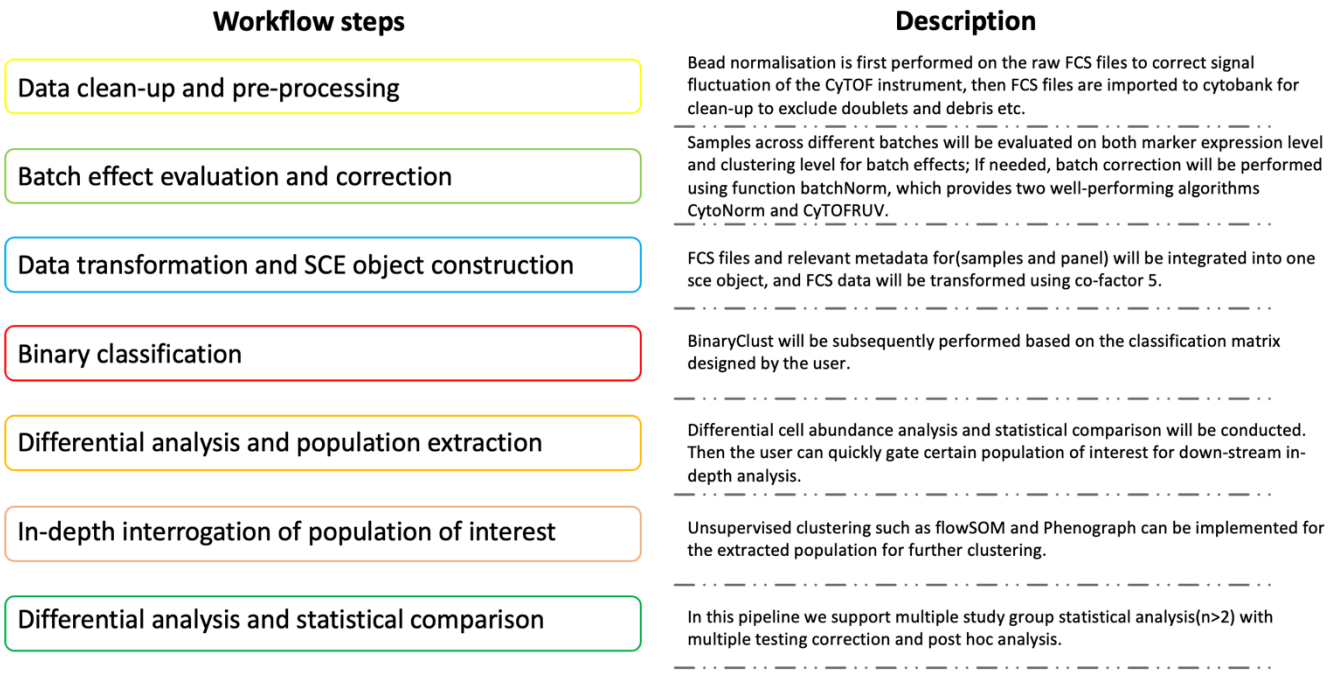
488

489 ***ImmCellTyper workflow overview***

490

491 ImmCellTyper pipeline is composed of seven steps corresponding to two separate sub-
492 pipelines, as described in Figure 8 and Github vignettes
493 (https://github.com/JingAnyaSun/ImmCellTyper/tree/main/vignettes):
494 (1) Sub-pipeline1 (corresponds to Figure 8, workflow step 2): Batch effects evaluation and
495 correction. Batch effects occur when samples were collected and measured at different sites
496 or time points, especially for large-scale studies. It is crucial to remove the unwanted variation
497 which might interfere the true biological differences. After bead normalisation and data
498 clean-up (Figure 8, workflow step 1), which can be performed using third-party platforms,
499 such as CyTOF v7.0 system control software and Cytobank, users can systematically examine
500 batch effects on two levels including marker behaviours and clustering results based on the
501 method introduced by Trussart et al.[29]If needed, CytoNorm and CytofRUV, which are well-
502 performing correction algorithms, can be used to align the existing batch effects.

503 (2) Sub-pipeline2 (corresponds to Figure 8, workflow step 3-7): semi-supervised classification,
504 differential analysis, and in-depth investigation. When data are cleaned and normalised, they
505 can be imported into the second part of the pipeline, constructed into a SingleCellExperiment
506 (SCE) object, and undergo semi-supervised classification to identify the major cell types and
507 test the differential frequencies or state marker expression among study groups. After that, if
508 the users have certain interests of specific cell types and pre-design the panel for that, or in
509 another circumstance, the initial statistics draw the user's attention into a certain cell type,
510 institutively, further clustering using unsupervised tools like flowSOM or Phenograph should
511 be conducted with an increased cluster resolution and deeper investigation for cell subsets,
512 after extracting the cell population of interest. ImmCellTyper pipeline has the same data
513 storage and infrastructure as CATALYST, therefore all the functions in CATALYST can be
514 seamlessly used in ImmCellTyper, tailored to user's analytical needs. We do not elaborate on
515 the basic functions of CATALYST, which can be found in the tutorial vignettes of the
516 package(https://github.com/HelenaLC/CATALYST).



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

518

519 ***Batch correction algorithms***

520

521 The batch effect correction algorithms embedded in the function of 'batchNorm' comprises
522 CytoNorm, as described by Van Gassen et al.[28] and CytofRUV by Trussart et al.[29]. Both

523 algorithms rely on anchors (reference samples/technical replicates) across batches to perform
524 normalisation. CytoNorm uses flowSOM clustering to first identify clusters prior to a
525 population-specific transformation on the reference samples by computing the quantile
526 values and aligning them with splines, whereas CytofRUV applies remove unwanted variation
527 III (RUV-III) to CyTOF data by estimating and eliminating the non-biological variation of the
528 pseudo-replicates.
529
530 ***Agreement evaluation***
531
532 To assess the agreement among manual gating, BinaryClust classification and flowSOM
533 clustering, we used correlation, interaction plot and Bland-Altman analysis. Correlation
534 evaluates the relationship between two variables which doesn't mean concordance, but if
535 two methods agree, surely, they should be highly correlated. We compared the results of cell
536 population frequency generated by experts and BinaryClust using Pearson correlation,
537 calculated the correlation coefficient and P value, with the line of equality indicating perfect
538 agreement (red solid line, R^2=1) in the plots to help gauge the degree of agreement between
539 the two methods; Bland-Altman plot refers to a dot plot of the difference between two
540 variables(y-axis) against the mean of them(x-axis), as described by J.Martian Bland and
541 Douglas G. Altman in 1986, which represents a graphical magnitude of bias(average of
542 difference) with 95% confidence interval. The math formula for the limit of agreement is as
543 below:

$$\text{Limits of agreement} = \text{mean difference observed} \pm 1.96 \times \text{standard deviation}$$

544
545 We also used Interaction plot to display the interaction effects of the three methods including
546 manual gating, BinaryClust classification and flowSOM clustering on the measurement of cell
547 frequencies to evaluate the agreement.
548
549 ***F-measure***
550
551 F-measure is a method to assess the accuracy of the clustering method compared to gold
552 standard, which is manual gating results in this study. It stands for the harmonic mean of the
553 precision and recall values, which can be calculated using below formula:

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

554

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

560

561 ***Adjusted Rand Index***

562

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

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

567
568

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

571

572 ***Differential analysis***

573

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

580

581 ***Sample collection and preparation for CyTOF***

582

583 PBMC samples from MPN patients were requested and obtained from biobank at Guy's
584 Hospital, under a protocol approved by the KCL Biobank Access Committee (REC18/EE/0025).
585 Healthy volunteers were recruited at Guy's hospital with informed consent and ethical
586 approval by King's College Research Ethics Committee (HR-17/18-5960 MOD-20/21-5960) in
587 accordance with the Declaration of Helsinki. All identifiable information of study participants
588 were securely stored in a trusted research environment managed by members of the team.
589
590 Venous blood samples from healthy volunteers were collected in BD vacutainer EDTA tubes,
591 and PBMCs were isolated and purified using Ficoll-Hypaque density gradient centrifugation. In
592 brief, blood was carefully layered onto Ficoll and spined at 460g for 20 minutes at room
593 temperature (RT) without brake; Upon observing a clear separation of blood components,
594 PBMCs were then carefully isolated and washed three times with RPMI 1640 medium to
595 remove other contaminants. Subsequently, the purified PBMCs were cryopreserved in liquid
596 nitrogen at a density of $1X10^7$ cells per vial.
597
598 **CyTOF Antibody staining**
599
600 Cryopreserved cells were quickly thawed at 37°C in water bath, suspended in pre-warmed
601 RPMI 1640 medium, washed three times, and underwent Fc-blocking using human TruStain
602 FcX (Biolegend) for 10 mins at RT. Given that the antibody panel employs indirect detection of
603 CD95 using anti-CD95-APC and anti-APC-106Cd, cells were initially stained with anti-CD95-APC
604 for 30 minutes in a dark place, followed by two washes with 2ml cell staining buffer (CSB).
605 After the removal of supernatant, cells were resuspended in 300µl CSB and transferred into
606 the dip tube containing lyophilized antibody mix of Maxpar direct immune profiling assay
607 (MDIPA, Standard BioTools). Additional antibodies used to study T cell activation, migration,
608 and exhaustion status (anti-TIM3, anti-PD-1, anti-ICOS, anti-TIGIT, anti-OX40) were also added.
609 Cells were incubated with the antibody mix for 30 mins at RT in compliance with
610 manufacturer's instructions, washed twice and fixed using a freshly prepared 1.6%
611 paraformaldehyde (PFA, Thermo Fisher Scientific) solution for 10 mins at RT. In the end, cells
612 were washed with CSB twice to remove residual PFA and stained with 125nM Cell-ID
613 intercalator-Iridium in 1 ml Maxpar Fix and Perm Buffer (Standard BioTools) overnight at 4°C.

614 The stained cells were frozen down using freezing media (fetal bovine serum containing 10%
615 DMSO) and stored at -80°C freezer before data acquisition.
616
617 On the day for CyTOF acquisition, cryopreserved samples were thawed, washed twice with
618 1ml CSB, followed by additional two washes with 1ml cell acquisition solution (CAS, Standard
619 BioTools), and centrifuged at 800g for 5 minutes. Cells were then filtered through a 40 µm cell
620 strainer to avoid blockage and cell count was determined by Countess automated cell counter
621 (Invitrogen). EQ four-element calibration beads (Standard BioTools) were added to a final
622 concentration of $0.5x10^6$ cells/ml to adjust signal fluctuation of the instrument. CyTOF
623 acquisition was performed on Helios mass cytometer system. For each batch of cell staining
624 and run, technical replicates from the same healthy donor were included to evaluate and
625 correct batch effects.
626
627 The full antibody panel including the metal tag, clone, and supplier is listed in Supplementary
628 Table 1.
629
630 ***CyTOF data pre-processing***
631
632 FCS files were first processed for bead normalisation using CyTOF v7.0 system control
633 software (Standard BioTools) to correct signal drift during acquisition. Subsequently, the files
634 were imported into Cytobank (Cytobank Inc.) for data cleaning, with a detailed procedure
635 illustrated in Supplementary Figure 6. The aim is to remove non-events including debris,
636 doublets, normalisation (EQ) beads and other undesired events like dead cells. The CD45+ cell
637 population was pre-gated using Cytobank for down-stream analysis.
638
639 ***Manual gating***
640
641 Manual gating for major cell populations was performed using Cytobank platform, and
642 gatingML files were exported and converted into gatingSet object to extract labels of each cell
643 via the R packages flowWorkspace, openCyto and CytoML. These data were used as reference
644 for benchmarking computer-aided algorithms.
645

646 ***Dataset availability***

647

648 In this study, we tested ImmCellTyper pipeline on the MPN cohort with 7 MPN patients and 2
649 healthy volunteers, influenza cohort with 11 patients and the COVID-19 cohort with 59
650 COVID-19 patients and 23 healthy volunteers. The FCS files (after clean-up and gating of CD45
651 population) and metadata of the MPN cohort were deposited in Zenodo (doi:
652 10.5281/zenodo.10076940); The influenza cohort was published by our lab[21], and the data
653 were stored in Zenodo (doi: 10.5281/zenodo.7982165); The COVID-19 dataset was previously
654 published by Chevrier et al.[22], and the FCS files can be retrieved from Mendeley Data
655 (https://data.mendeley.com/datasets/vyy8ttw7n9/1).

656

657 **Author contributions**

658

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

667

668 **Competing interests**

669

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

671

672 **Acknowledgement**

673

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

677

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

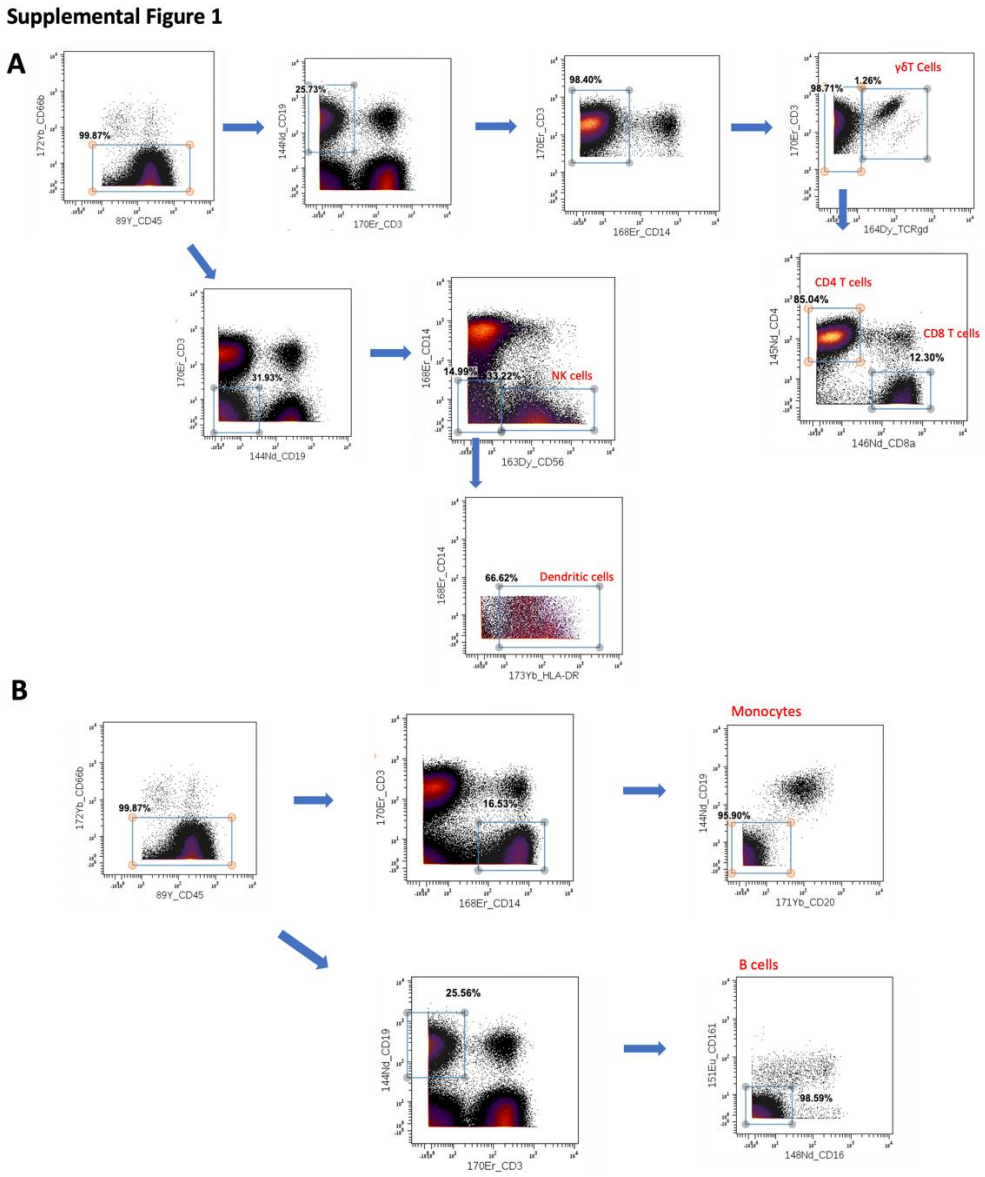
686 **References**

687 1. Spitzer, M.H., and Nolan, G.P. (2016). Mass Cytometry: Single Cells, Many Features.
688 Cell *165*, 780-791. 10.1016/j.cell.2016.04.019.
689 2. Kimball, A.K., Oko, L.M., Bullock, B.L., Nemenoff, R.A., van Dyk, L.F., and Clambey, E.T.
690 (2018). A Beginner's Guide to Analyzing and Visualizing Mass Cytometry Data. J
691 Immunol *200*, 3-22. 10.4049/jimmunol.1701494.
692 3. Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P.,
693 Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for
694 visualization and interpretation of cytometry data. Cytometry A *87*, 636-645.
695 10.1002/cyto.a.22625.
696 4. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir el, A.D., Tadmor, M.D., Litvin,
697 O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic
698 Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*,
699 184-197. 10.1016/j.cell.2015.05.047.
700 5. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L., and Nolan, G.P. (2016). Automated
701 mapping of phenotype space with single-cell data. Nat Methods *13*, 493-496.
702 10.1038/nmeth.3863.
703 6. Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Jr., Bruggner, R.V., Linderman, M.D.,
704 Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from
705 high-dimensional cytometry data with SPADE. Nat Biotechnol *29*, 886-891.
706 10.1038/nbt.1991.
707 7. Becher, B., Schlitzer, A., Chen, J., Mair, F., Sumatoh, H.R., Teng, K.W.W., Low, D., Ruedl,
708 C., Riccardi-Castagnoli, P., Poidinger, M., et al. (2014). High-dimensional analysis of the
709 murine myeloid cell system. Nature Immunology *15*, 1181-1189. 10.1038/ni.3006.
710 8. van der Maaten, L., and Hinton, G. (2008). Viualizing data using t-SNE. Journal of
711 Machine Learning Research *9*, 2579-2605.
712 9. McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and
713 Projection for Dimension Reduction. ArXiv e-prints *1802.03426*.
714 10. Liu, X., Song, W., Wong, B.Y., Zhang, T., Yu, S., Lin, G.N., and Ding, X. (2019). A
715 comparison framework and guideline of clustering methods for mass cytometry data.
716 Genome Biology *20*, 297. 10.1186/s13059-019-1917-7.
717 11. Abdelaal, T., van Unen, V., Höllt, T., Koning, F., Reinders, M.J.T., and Mahfouz, A.
718 (2019). Predicting Cell Populations in Single Cell Mass Cytometry Data. Cytometry A.
719 10.1002/cyto.a.23738.

720 12. Cheng, L., Karkhanis, P., Gokbag, B., Liu, Y., and Li, L. (2022). DGCyTOF: Deep learning
721 with graphic cluster visualization to predict cell types of single cell mass cytometry
722 data. PLOS Computational Biology *18*, e1008885. 10.1371/journal.pcbi.1008885.
723 13. Kaushik, A., Dunham, D., He, Z., Manohar, M., Desai, M., Nadeau, K.C., and Andorf, S.
724 (2021). CyAnno: a semi-automated approach for cell type annotation of mass
725 cytometry datasets. Bioinformatics *37*, 4164-4171. 10.1093/bioinformatics/btab409.
726 14. Li, H., Shaham, U., Stanton, K.P., Yao, Y., Montgomery, R.R., and Kluger, Y. (2017).
727 Gating mass cytometry data by deep learning. Bioinformatics *33*, 3423-3430.
728 10.1093/bioinformatics/btx448.
729 15. Lee, H.C., Kosoy, R., Becker, C.E., Dudley, J.T., and Kidd, B.A. (2017). Automated cell
730 type discovery and classification through knowledge transfer. Bioinformatics *33*, 1689-
731 1695. 10.1093/bioinformatics/btx054.
732 16. Zhang, Z., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski,
733 E.W., Modrusan, Z., et al. (2019). SCINA: A Semi-Supervised Subtyping Algorithm of
734 Single Cells and Bulk Samples. Genes (Basel) *10*. 10.3390/genes10070531.
735 17. Nowicka, M., Krieg, C., Crowell, H.L., Weber, L.M., Hartmann, F.J., Guglietta, S., Becher,
736 B., Levesque, M.P., and Robinson, M.D. (2017). CyTOF workflow: differential discovery
737 in high-throughput high-dimensional cytometry datasets. F1000Res *6*, 748.
738 10.12688/f1000research.11622.3.
739 18. Marsh-Wakefield, F., Kruzins, A., McGuire, H.M., Yang, S., Bryant, C., Fazekas de St
740 Groth, B., Nassif, N., Byrne, S.N., Gibson, J., Brown, C., et al. (2019). Mass Cytometry
741 Discovers Two Discrete Subsets of CD39(-)Treg Which Discriminate MGUS From
742 Multiple Myeloma. Front Immunol *10*, 1596. 10.3389/fimmu.2019.01596.
743 19. Chen, H., Lau, M.C., Wong, M.T., Newell, E.W., Poidinger, M., and Chen, J. (2016).
744 Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis
745 Pipeline. PLoS Comput Biol *12*, e1005112. 10.1371/journal.pcbi.1005112.
746 20. Opzoomer, J.W., Timms, J.A., Blighe, K., Mourikis, T.P., Chapuis, N., Bekoe, R.,
747 Kareemaghay, S., Nocerino, P., Apollonio, B., Ramsay, A.G., et al. (2021).
748 ImmunoCluster provides a computational framework for the nonspecialist to profile
749 high-dimensional cytometry data. Elife *10*. 10.7554/eLife.62915.
750 21. Alimam, S., Ann Timms, J., Harrison, C.N., Dillon, R., Mare, T., DeLavallade, H., Radia,
751 D., Woodley, C., Francis, Y., Sanchez, K., et al. (2021). Altered immune response to the
752 annual influenza A vaccine in patients with myeloproliferative neoplasms. Br J
753 Haematol *193*, 150-154. 10.1111/bjh.17096.
754 22. Chevrier, S., Zurbuchen, Y., Cervia, C., Adamo, S., Raeber, M.E., de Souza, N.,
755 Sivapatham, S., Jacobs, A., Bachli, E., Rudiger, A., et al. (2021). A distinct innate
756 immune signature marks progression from mild to severe COVID-19. Cell Rep Med *2*,
757 100166. 10.1016/j.xcrm.2020.100166.
758 23. Liu, P., Liu, S., Fang, Y., Xue, X., Zou, J., Tseng, G., and Konnikova, L. (2020). Recent
759 Advances in Computer-Assisted Algorithms for Cell Subtype Identification of
760 Cytometry Data. Front Cell Dev Biol *8*, 234. 10.3389/fcell.2020.00234.
761 24. Aghaeepour, N., Nikolic, R., Hoos, H.H., and Brinkman, R.R. (2011). Rapid cell
762 population identification in flow cytometry data. Cytometry A *79*, 6-13.
763 10.1002/cyto.a.21007.
764 25. Theorell, A., Bryceson, Y.T., and Theorell, J. (2019). Determination of essential
765 phenotypic elements of clusters in high-dimensional entities-DEPECHE. PLoS One *14*,
766 e0203247. 10.1371/journal.pone.0203247.

767 26. O'Brien Gore, C., Billman, A., Hunjan, S., Colebrook, J., Choy, D., Li, W., Haynes, J.,
768 Wade, J., Hobern, E., McDonald, L., et al. (2023). Pre-treatment with systemic agents
769 for advanced NSCLC elicits changes in the phenotype of autologous T cell therapy
770 products. Mol Ther Oncolytics *31*, 100749. 10.1016/j.omto.2023.100749.
771 27. Weber, L.M., Nowicka, M., Soneson, C., and Robinson, M.D. (2019). diffcyt: Differential
772 discovery in high-dimensional cytometry via high-resolution clustering.
773 Communications Biology *2*, 183. 10.1038/s42003-019-0415-5.
774 28. Van Gassen, S., Gaudilliere, B., Angst, M.S., Saeys, Y., and Aghaeepour, N. (2020).
775 CytoNorm: A Normalization Algorithm for Cytometry Data. Cytometry A *97*, 268-278.
776 10.1002/cyto.a.23904.
777 29. Trussart, M., Teh, C.E., Tan, T., Leong, L., Gray, D.H., and Speed, T.P. (2020). Removing
778 unwanted variation with CytofRUV to integrate multiple CyTOF datasets. Elife *9*.
779 10.7554/eLife.59630.
780 30. Arvaniti, E., and Claassen, M. (2017). Sensitive detection of rare disease-associated cell
781 subsets via representation learning. Nature Communications *8*, 14825.
782 10.1038/ncomms14825.
783 31. Lun, A.T.L., Richard, A.C., and Marioni, J.C. (2017). Testing for differential abundance in
784 mass cytometry data. Nat Methods *14*, 707-709. 10.1038/nmeth.4295.
785 32. Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., and Nolan, G.P. (2014).
786 Automated identification of stratifying signatures in cellular subpopulations. Proc Natl
787 Acad Sci U S A *111*, E2770-2777. 10.1073/pnas.1408792111.
788 33. Arend, L., Bernett, J., Manz, Q., Klug, M., Lazareva, O., Baumbach, J., Bongiovanni, D.,
789 and List, M. (2022). A systematic comparison of novel and existing differential analysis
790 methods for CyTOF data. Briefings in Bioinformatics *23*, bbab471.
791 10.1093/bib/bbab471.
792

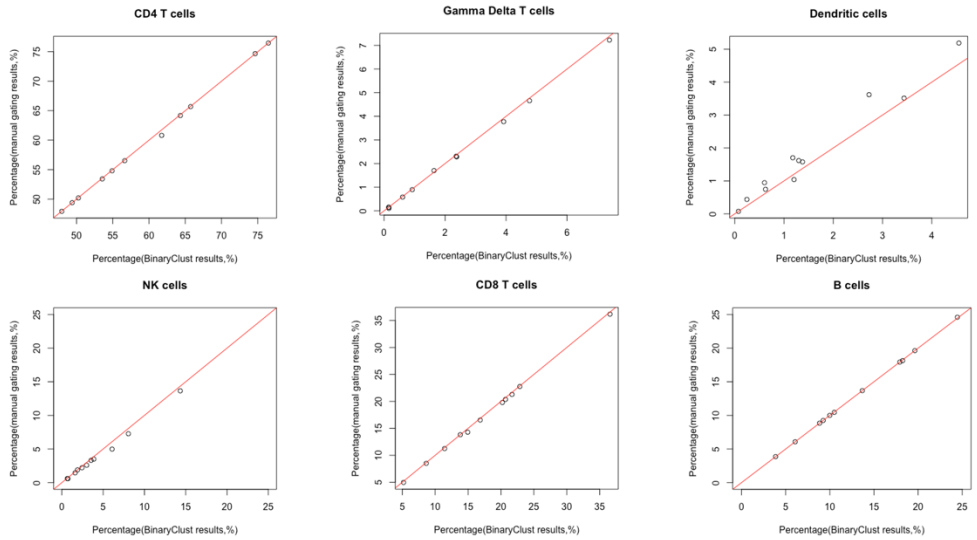
Supplementary Figures



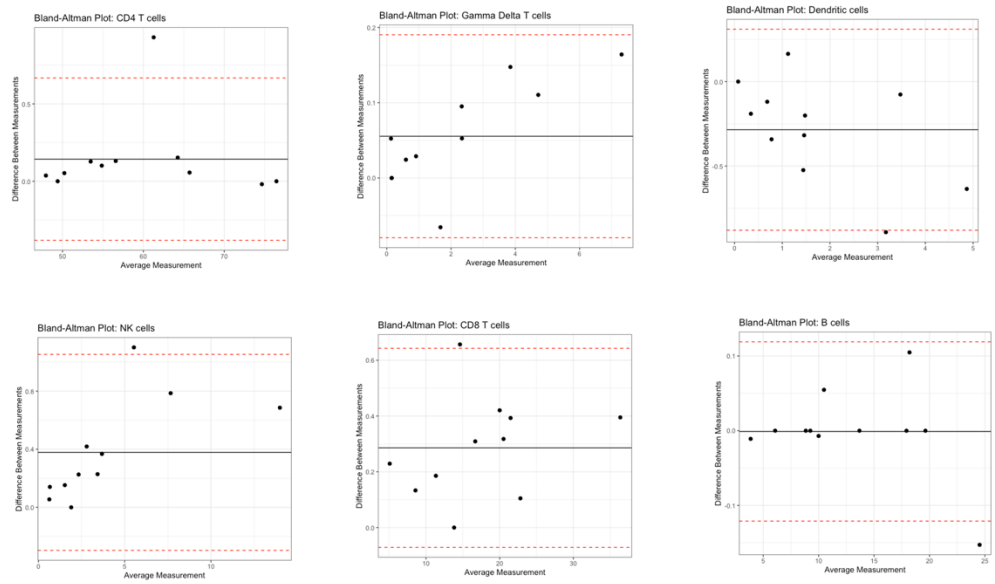
Supplementary Figure 1 Manual hierarchical gating strategy for main cell linages from human PBMC samples (MPN dataset, n=9). All .FCS files were cleaned up to remove doublets, normalisation beads and debris (please refer to the methods for standard clean-up procedure), and pre-gated for CD45+ leukocytes. (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

A



B

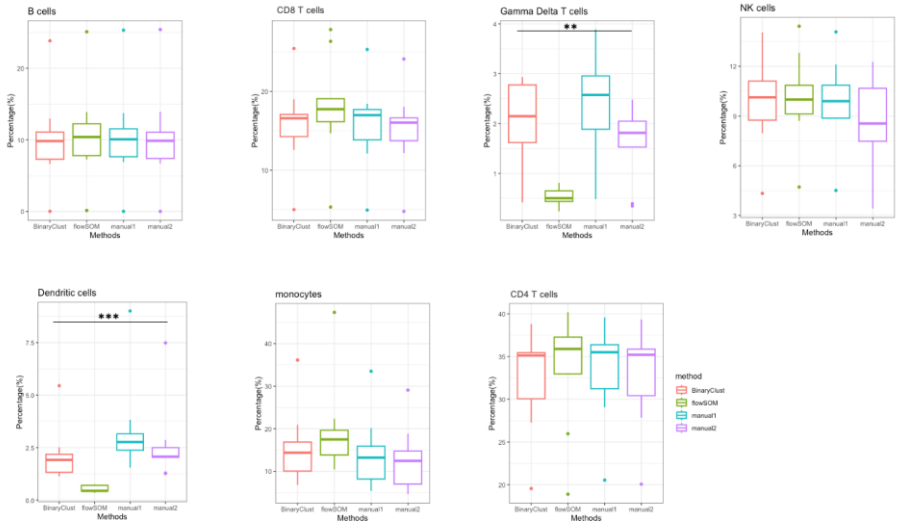


C

	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	

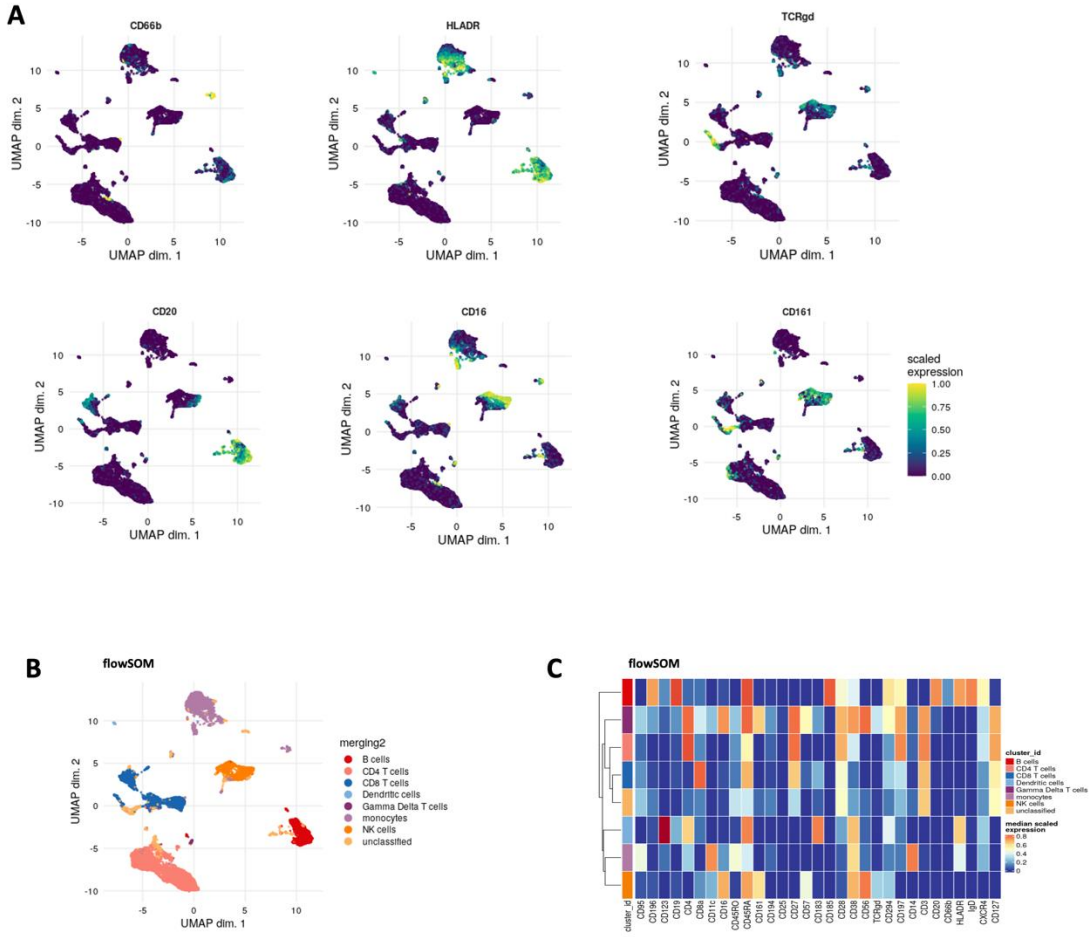
Supplementary Figure 2 Agreement evaluation between ImmCellTyper and manual gating in influenza dataset(n=11). Manual gating was performed using Cytobank and exported manually. (A) Correlation plots between ImmCellTyper results and manual gating results concerning percentages in CD4 T cells, Gamma Delta 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.

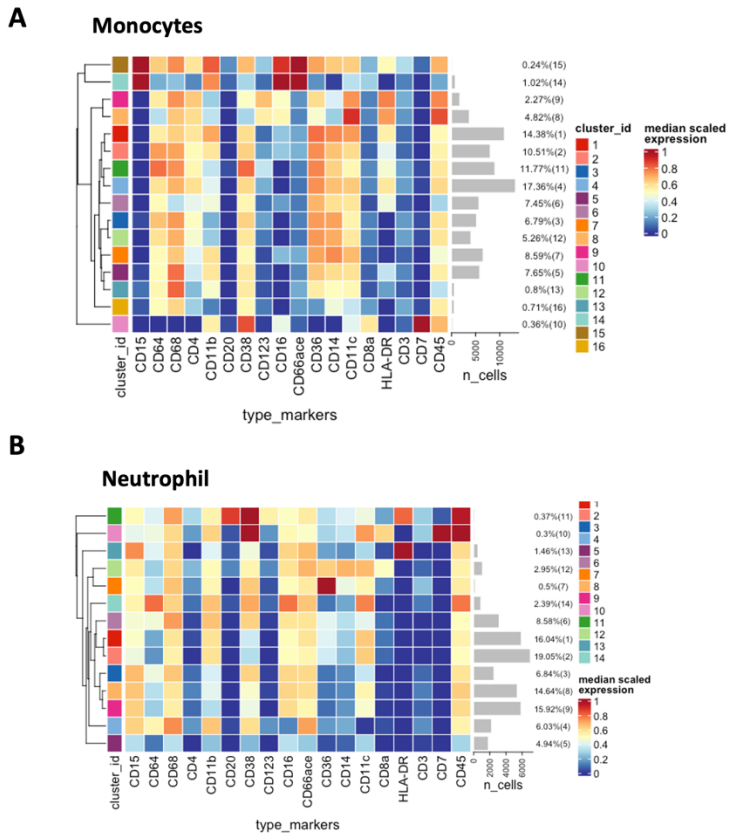
Supplementary Figure 4



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

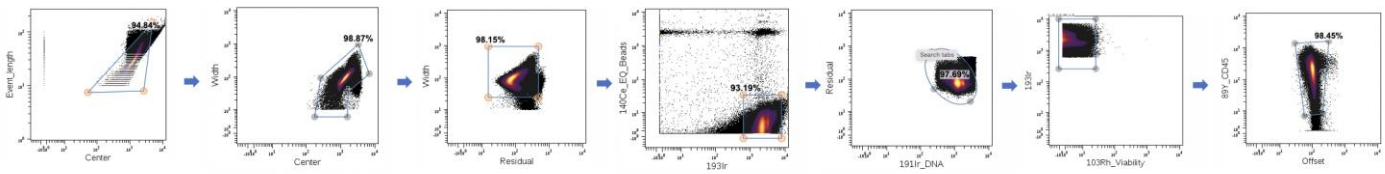
clustering was performed on the same dataset to compare with BinarClust, k=20 was chosen followed by manual annotation of each cluster. UMAP plot was projected with merged flowSOM clusters with biological annotation (downsample 2000 cells per sample) ;(B) The corresponding median marker expression heatmap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

Supplementary Figure 5



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

Supplementary Figure 6



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

835 **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 BioTools836
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