bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Meta-analysis of Cytometry Data Reveals Racial Differences in Immune Cells

Zicheng Hu[1], Chethan Jujjavarapu[1], Jake J. Hughey[2], Sandra Andorf[3], Pier Federico Gherardini[4], Matthew H. Spitzer[5], Patrick Dunn[6], Cristel G Thomas[6], John Campbell[6], Jeff Wiser[6], Garry P. Nolan[4], Sanchita Bhattacharya[1*], Atul J. Butte[1*]

[1] Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA. [2] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA. [3] Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [4] Department of Microbiology and Immunology, Stanford University, Stanford, CA, USA. [5] Department of Microbiology and Immunology, Helen Diller Family Comprehensive Cancer Center, Parker Institute for Cancer Immunotherapy, University of California, San Francisco, San Francisco, California, CA, USA. [6] Northrop Grumman Technology Services Health IT, Rockville, MD, USA. Correspondence should be addressed to A.J.B. (Atul.Butte@ucsf.edu)

* Equal contribution

Abstract:

While meta-analysis has demonstrated increased statistical power and more robust estimations in studies, the application of this commonly accepted methodology to cytometry data has been challenging. Different cytometry studies often involve diverse sets of markers. Moreover, the detected values of the same marker are inconsistent between studies due to different experimental designs and cytometer configurations. As a result, the cell subsets identified by existing auto-gating methods cannot be directly compared across studies. We developed MetaCyto for automated meta-analysis of both flow and mass cytometry (CyTOF) data. By combining clustering methods with a silhouette scanning method, MetaCyto is able to identify common cell subsets across studies, thus enabling meta-analysis. Applying MetaCyto on a set of 10 heterogeneous cytometry studies with a total of 5966 samples allowed us to identify multiple cell populations exhibiting differences in phenotype and abundance across races. Software is released to the public through GitHub (github.com/hzc363/MetaCyto).

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Main text:

Meta-analysis of existing data across different studies offers multiple benefits. The aggregated data allows researchers to test hypotheses with increased statistical power. The involvement of multiple independent studies increases the robustness of conclusions drawn. In addition, the complexity of aggregated data allows researchers to test or generate new hypotheses. These benefits have been shown by many studies in areas such as genomics, cancer biology and clinical research and have led to important new biomedical findings [1–4]. For example, one study showed the correlation between neo-antigen abundance in tumors and patient survival by performing meta-analysis of RNA sequencing data from The Cancer Genome Atlas[5]. In another study, meta-analysis of genome-wide association studies identified novel loci that affect risk of type 1 diabetes[6].

With the recent advances in high-throughput cytometry technologies the immune system can be characterized simultaneously at the single cell level with up to 45 parameters, thus minimizing the technical limitations and allowing capture of invaluable information from immunology studies[7–9]. Open science initiatives make more and more research data accessible, and the availability of shared cytometry data, including data from flow cytometry and mass cytometry (CyTOF), is growing exponentially. Notably, the ImmPort database (www.immport.org), a repository for immunology-related research and clinical trials, provides numerous studies with thousands of cytometry datasets[10]. However, meta-analysis of cytometry datasets remains particularly challenging. Different studies use diverse sets of protein markers and fluorophore/isotope combinations. The detected values of the same marker are inconsistent between studies because of different cytometer configurations or operators. In addition, the high dimensionality of cytometry data, especially mass cytometry (CyTOF) data, makes manual gating based meta-analysis difficult and time consuming.

The major challenge of automated meta-analysis of cytometry data lies in the identification of common cell subsets across studies. Multiple automated gating methods have been proposed to analyze cytometry data from a single experiment. The performance of these

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methods has been extensively tested in several studies[11,12]. However, the results of most auto-gating methods, such as FlowSOM[13], FlowMeans[14] and CITRUS[15], cannot be compared across studies. The cell subsets identified by these methods are usually labeled with anonymous identifiers with no cell-specific annotation, making it impossible to identify the common cell populations across different studies. The heterogeneity of cytometry data across studies also prevents identifying common cell populations based on marker values.

In addition, many clustering methods are sensitive to parameter choices. For example, FlowSOM, FlowMeans and SPADE[16] require users to pre-specify the number of clusters. As a result, extensive parameter tuning and manual inspection is required for every cytometry dataset. In meta-analysis where large numbers of cytometry datasets are involved, this manual step becomes a major technical burden.

Here, we developed MetaCyto to enable automated meta-analysis of cytometry datasets, including both conventional flow and CyTOF cytometry data. MetaCyto employs novel computational approaches to identify common cell subsets across studies in either of two fully automated pipelines: unsupervised analysis and guided analysis. The unsupervised analysis pipeline rigorously annotates and merges cell subsets identified by existing clustering methods, such as FlowSOM and FlowMeans, allowing the cell subsets to be related across studies. The guided analysis pipeline can identify known cell populations across studies based on pre-specified cell definitions, thus allowing for the search of specific cell subsets defined by immunologists. Applying our method in several cytometry datasets, we demonstrated that both pipelines are able to identify cell subsets from cytometry data across studies accurately without any parameter tuning requirements.

We applied MetaCyto to perform a joint analysis of 10 human immunology cytometry datasets contributed by four different institutions[17–20]. Altogether, this analysis spanned 5966 peripheral blood mononuclear cells (PBMC) or whole blood samples from 1374 healthy subjects, and were acquired using either flow cytometry or CyTOF with a diverse set of markers. These 1374 subjects were

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identified as representing 5 different races. While it is well known that characteristics of multiple immune system-related diseases, such as HIV[21], tuberculosis[22] and hepatitis C[23], vary between racial groups, the heterogeneity of the immune system among the human population has made studying these differences difficult, even within the same racial group[24,25]. We hypothesized that a meta-analysis approach could lead to a better understanding of racial differences in the immune system. Using MetaCyto, we not only confirm a known racial difference, but also identified new cell types whose frequencies or characteristics vary between races.

RESULTS

MetaCyto identifies and labels common cell subsets in cytometry data across studies to enable meta-analysis

Our meta-analysis of cytometry data follows four steps: data aggregation, data pre-processing, identification of common cell subsets across studies, and statistical analysis (**Fig. 1a**). The third step, identification of common cell subsets across studies, has been the main technical challenge preventing automated meta-analysis. Therefore, while all four steps are automated and covered in the MetaCyto software system and documented in the online methods, here we primarily focus on describing our identification and relating of common cell subsets across studies.

The unsupervised analysis pipeline in MetaCyto identifies common cell subsets across different studies in a fully automated way. Cytometry data in each study is first clustered using an existing clustering method (**Fig. 1b Top**). FlowSOM[13] was implemented as the default clustering method due to its speed and performance. However, any other clustering method, such as hierarchical clustering or FlowMeans, could be substituted as well. At this stage, clusters are labeled with non-informative labels, such as C1, C2, C3, which cannot be related across studies. For example, C1 in study 1 and C1 in study 2 represent entirely different cell populations.

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A threshold is then chosen to bisect the distribution of each marker into positive and negative regions, needed to label each cluster in a biological meaningful way (**Fig. 1b Middle**). The selection of a threshold is easy when a clear bi-modal distribution is present, but becomes challenging in other cases. We implemented a Silhouette scanning method, which bisects each marker at the threshold maximizing the average silhouette, a widely used way of describing the quality of clusters[26]. We compared Silhouette scanning against 8 other bisection methods and found it to be superior, when compared with manual gating (**Supplementary Fig. 1**).

Clusters are then labeled for each of the markers based on the following rule: if the marker level of 95% of cells in the cluster are above or below the threshold, the cluster will be labeled as positive or negative for the marker, respectively. Otherwise, the cluster will not be labeled for the marker. For example in Fig 1b, both C2 and C1 in study 2 will be labeled as CD8+ CD4-.

Next, clusters with the same labels are merged into a square shaped cluster (**Fig. 1b Bottom**). In cytometry data with higher dimensions, clusters are hyper-rectangles. Following this stage, common cell subsets across studies can be rigorously identified and annotated. For example, the CD4- CD8+ clusters in both study 1 and study 2 correspond to CD8+ T cells.

MetaCyto is also able to identify cell subsets in a guided analysis pipeline using pre-defined cell definitions. After bisecting each marker into positive and negative regions, cells fulfilling the pre-defined cell definitions are identified. For example, the CD3+ CD4+ CD8- (CD4+ T-cells) cell subset corresponds to the cells that fall into the CD3+ region, CD4+ region and CD8- region concurrently (**Fig. 1c**). Notice that both CD45RA+ and CD45RA- populations are included in the cell subset, because the cell definition does not specify the requirement for CD45RA expression. However, researchers could easily alter the cell definition to CD3+ CD4+ CD8- CD45RA+ to find the CD45RA+ cell subset.

The guided analysis pipeline in MetaCyto can accurately identify cell populations using pre-defined cell definitions.

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A successful meta-analysis of cytometry data requires cell populations to be identified accurately from each study. To evaluate if the guided analysis pipeline of MetaCyto can accurately identify cell subsets from a single study, we downloaded a set of PBMC cytometry data (SDY478) from ImmPort, with which the original authors identified 88 cell types. Correspondingly, we specified the 88 cell definitions (**Supplementary Table 2**) based on the author's gating strategy and identified these cell subsets for each cytometry sample using the guided analysis pipeline in MetaCyto. We compared the proportions of all cell subsets estimated by MetaCyto with the original manual gating results and found that MetaCyto estimations are highly consistent with the manual gating result (Fig. **2a-c**). We also compared our estimations to an existing method, flowDensity[27], which is also able to identify pre-defined cell populations. Our results suggest that MetaCyto performs better than flowDensity in quantifying both major and rare populations (**Fig. 2d,e**).

The unsupervised analysis pipeline in MetaCyto can enhance the quality and robustness of several clustering methods

We then tested the performance of the unsupervised analysis pipeline of MetaCyto. In the unsupervised analysis pipeline, cell clusters are first identified by an existing clustering algorithm, and are then merged into hyper-rectangle clusters (**Fig. 1b)**. To learn how such merging affects the quality of clusters, we evaluated the results of two clustering algorithms, FlowSOM[13] and FlowMeans[14], with and without the merging step. Multiple studies have been conducted to evaluate the performance of existing clustering method for cytometry data[11,12]. The most recent (Weber et al.[12]) compared 15 clustering methods and found FlowSOM generally outperformed other methods after manual tuning.

We downloaded an evaluation dataset, West Nile virus dataset (FlowCAP WNV), used by Weber et al[12] and applied FlowSOM. The clustering result is then labeled and merged. Since FlowSOM requires a pre-specified cluster number (K), we did multiple runs with K ranging from 10 to 90. F-measure is used to evaluate the quality of the clusters. We found that the quality of clusters is comparable

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

before and after merging when K equals to 10. However, the performance of FlowSOM drops when K increases. The subsequent merging step prevented FlowSOM performance to deteriorate (**Fig. 2f**). We then looked at the total number of clusters identified before and after merging. As expected, FlowSOM identified the same number of clusters specified by K. However, when running the merging step after FlowSOM, the total number of clusters no longer increases after a certain point (**Fig. 2g**).

The same results were obtained with FlowMeans[14] (**Supplementary Fig. 2**). This suggests that MetaCyto is able to merge small clusters in a biologically meaningful way, preventing over-partitioning of the cell subsets, thus allowing the clustering analysis to be performed without tuning any parameters.

Meta-analysis of cytometry data using MetaCyto provides consistent results between cytometry panels and confirms previous findings

After demonstrating the performance of MetaCyto in analyzing cytometry data from single studies, we next demonstrated the ability of MetaCyto in yielding consistent results from combining multiple studies. We applied MetaCyto to identify cell types whose proportion or protein expression levels are different between age, gender and race groups. We downloaded 10 studies from ImmPort containing cytometry data. These 10 studies had been contributed from four different institutions, where 88 panels containing 74 different markers were used (**Fig. 3** and **Supplementary Table 3**). Altogether, the dataset contains 5966 whole blood or PBMC samples from 1374 healthy subjects and were acquired using either flow cytometry or CyTOF. The subjects are proportionately distributed by gender, with slightly more female than male (**Supplementary Fig. 3a**). The age span ranging from 19 to 90 years, and comes from five different defined racial groups (**Supplementary Fig. 3 b,c**).

We used both unsupervised and guided MetaCyto analysis pipelines in parallel to identify cell subsets. For the latter, we created 24 cell definitions based on well-defined cell types from the Human

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ImmunoPhenotyping Consortium[28], ranging from effector memory T cells to monocytes (**Supplementary Table 4**).

We then calculated summary statistics for each cell type in each sample, including proportion and median fluorescence intensity (MFI) of each marker. The effect size of age, gender and race on the cell type proportions or markers MFI was estimated using a linear regression model (**Supplementary Table 5-10**).

We validated our results in two ways using the effect size of age, previously well characterized in other studies[20,29]. First, we checked if the results from MetaCyto were internally consistent. The 88 panels were randomly divided into two panel sets. The effect size of age estimated separately in the two panel sets were compared with each other. We repeated the procedure 100 times and found that the effect size estimates from two panels sets were highly correlated with each other using both the guided and unsupervised approaches (**Fig. 4a-c**), demonstrating that MetaCyto is able to derive consistent results from different cytometry panels.

As a subsequent validation, we tested whether results obtained with MetaCyto could replicate results from a previous independent study (Carr et al.[29]). We ran the MetaCyto guided analysis pipeline across all 88 panels together, and among the 24 cell types MetaCyto identified, 14 overlapped with the cell types included in the Carr study. We compared the effect size of age on the proportion of these 14 cell types, between MetaCyto on the 88 panels, and the independent results from Carr, et al. We found that results agree well with each other ($r = 0.69$, $p = 0.006$, **Fig. 4d**).

The Carr study also investigated the effect of gender on immune cells and only identified CD4+ T cells to be significantly different between genders. We compared the effect size of gender from MetaCyto with the results in Carr study and found that the 2 sets of results are highly consistent with each other ($r = 0.71$, $p = 0.004$, **Supplementary Fig. 4**). In addition to finding CD4+ T cells, our study identified that the proportion of effector CD8+ T cells, naive B cells, plasmablasts, regulatory T cells and naive CD4+ T cells are also significantly affected by gender (**Supplementary Table 6**), demonstrating the increased power with larger sample sizes in meta-analysis.

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We then tested whether we could re-discover well known differences in cell populations between races using MetaCyto applied to all 88 panels together. It has been known that Asian individuals have fewer CD4+ T cells in blood than White individuals[30]. We found that MetaCyto is able to identify this racial difference consistently across all flow cytometry and CyTOF panels (**Fig. 4e**). Combining the results from all panels allows us to confirm this known racial difference with high confidence ($p = 1.16 \times 10^{-7}$).

Meta-analysis of cytometry data using MetaCyto identifies novel racial differences in immune cells

In addition to confirming the known racial difference, we were able to identify additional immune cell types to be different between races. Since White subjects were most prevalent in the data, we used this group as the baseline and compared Asian and African American individuals against this baseline.

We first tested each of 24 well-defined cell types using the guided analysis pipeline, based on the cell definition from Human ImmunoPhenotyping Consortium (HIPC)[28]. We found that in addition to bulk CD4+ T cells, the proportion of CD4+ central memory T cells is also lower in Asians compare to Whites (p=0.039). The proportion of NK cells, however, is higher in Asians (p=0.039, **Fig. 5a** and **Supplementary Table 7**).

The phenotypes of multiple cell types, as defined by the MFI for various markers, were affected by race as well (**Fig. 5a, b**, **Supplementary Table 7 and 8**). For example, we found that the expression level of CD94 (KLRD1) is higher in NK cells of both Asian (p=0.00027) and African American (p=0.00036) individuals compared to Whites (**Fig. 5a-c**). We further confirmed that this finding was not an artifact of our methodology with manual gating of the cytometry data from panel 1 of SDY420 (**Fig. 5d**).

In some cases, the marker level in all cell populations are affected by race in the same way. For example, the expression of CD28 on numerous T cell subsets is higher in African Americans as compared

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to Whites. In other cases, the marker level difference was cell type specific. CD25 expression was for instance lower on CD8+ T cells, but higher on NK cells in African Americans compared to Whites (**Fig. 5b**). Such cell type specific changes can only be identified using single cell technologies such as cytometry.

Results from the unsupervised analysis identified multiple cell types, other than the 24 types used in the guided analysis, whose abundance were different between races (**Supplementary Table 9 and 10**). As one example, we found that the proportion of a sub-population of CD8+ T cells, the CD3+ CD4- CD45RA+ CD8+ CD85J- cell population, is significantly higher in Asians than in Whites (**Fig. 5 e,f**). A closer look at the forest plot revealed that the association between this population and race was not at a significant level in most studies taken independently. However, by combining the results from multiple studies, we were able to identify this association with high confidence (p=0.0049).

DISCUSSION

In this study, we developed MetaCyto, a computational tool that allows automated gating and automated meta-analysis of both CyTOF and flow cytometry data. MetaCyto is able to find common cell subsets across studies using either an unsupervised or a guided analysis pipeline. Using publicly available datasets, we showed that both methods in MetaCyto outperform existing auto-gating methods without the need of any parameter tuning. Using MetaCyto, we analyzed cytometry data from 10 studies. After confirming known associations, we identified additional cell populations whose abundance or phenotype is different between races.

Most existing auto-gating methods identify cell populations in cytometry data using unsupervised clustering approaches[11,12,31]. Although such approaches are able to identify cell subsets in an unbiased way, they often miss well-defined cell populations, especially for rare populations such as regulatory T cells. MetaCyto's guided analysis pipeline is able to identify cell populations using user-defined cell definitions. For example, regulatory T cells can easily be

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identified using the definition “CD3+ CD4+ Foxp3+”. Such an approach allows researchers to incorporate their domain knowledge into the analysis, making the result more biologically relevant.

In addition to the guided analysis pipeline, MetaCyto also allows researchers to identify cell populations using un-supervised clustering methods. Successful efforts were made by the community to develop efficient clustering methods for flow cytometry data analysis. We built MetaCyto to be fully compatible with existing clustering methods. MetaCyto is able to merge and transform the clusters from existing clustering algorithms in a biologically meaningful way, therefore improving result quality and enabling further meta-analysis of many studies.

Based on the test result, we recommend over-clustering the data first, followed by the merging of the clusters by MetaCyto. Such a strategy not only makes the method tuning free, but also is more computationally efficient than traditional auto-tuning methods, which require running the clustering algorithm multiple times with different parameters.

Applying MetaCyto to cytometry data from 10 human immunology studies allowed us to thoroughly characterize differences in the immune system between races. Other than the previously known differences in CD4+ T cell abundance between Asians and Whites, we identified novel cell populations whose abundance and marker expression levels were significantly different between races. We believe that our findings will not only help us better understand the heterogeneity of the human immune system in the population, but also serve as the starting point for future in-depth studies to reveal the mechanisms behind racial discrepancies in immune-related diseases.

MetaCyto is primarily designed to improve the robustness and interpretability of cytometry analysis. Inevitably, the sensitivity of the method is decreased, especially in the unsupervised analysis pipeline. Although the merging of cell subsets improves the robustness of the clustering result, some small cell populations of biological meaning may be lost. To overcome this limitation, a more sensitive method, such as CITRUS[15] may be applied to data from a

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

single study first. After identifying the cell subsets of interest from the single study, the guided analysis pipeline of MetaCyto can be used to perform meta-analysis on the cell subsets across studies.

METHODS

Data Aggregation

Flow cytometry data and CyTOF data from SDY112, SDY167[19], SDY180[18], SDY311, SDY312, SDY314, SDY315, SDY420[20], SDY478 and SDY736[17] were downloaded from ImmPort web portal. Only fcs files from pre-vaccination blood samples of healthy adults were included in the meta-analysis. Parameters, including antibodies and fluorescence or isotope labels, used in each fcs file were then identified using the *fcsInfoParser* function in MetaCyto. The fcs files were then organized into panels, which are defined as a collection of fcs files from the same study that have the same set of parameters.

Manual gating results for both FlowCAP WNV data (ID number FR-FCM-ZZY3) were downloaded from the FlowRepository link: community.cytobank.org/cytobank/experiments/4329.

All data sets were downloaded between September 1, 2016 and February 1, 2017.

Data Pre-processing

Flow cytometry data from ImmPort were compensated for fluorescence spillovers using the compensation matrix supplied in each fcs file. All data from ImmPort were arcsinh transformed. For flow cytometry data, the formula f(x) = arcsinh (x/150) was used. For CyTOF data, the formula f(x) = arcsinh (x/8) was used. All transformation and compensation were done using the *preprocessing* or *preprocessing.batch* function in MetaCyto.

Cytometry data FlowCAP WNV was transformed and subset to only include protein markers. The pre-processing was doing using the same code provided by the Weber study[12] : github.com/lmweber/cytometry-clustering-comparison

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identifying cell subsets with the guided analysis pipeline in MetaCyto

Cell definitions were created based on the gating strategies provided by authors of SDY 420 and SDY478 or based on the cell definition from the Human ImmunoPhenotyping Consortium[28] . The cell definitions are available in the **supplementary table 1, 2** and **4**. *searchCluster* or *searchClster.batch* function was used to identify the cell subsets corresponding to the cell definitions. The summary statistics, including proportion and MFI of markers of each cell subsets were generated by the same functions.

Identify cell subsets with the unsupervised analysis pipeline in MetaCyto

In **Fig. 2** and **Supplementary Fig. 2**, FlowSOM[13] and FlowMeans[14] were run using the same code provided by the Weber study using different K values. The resulting clusters from FlowSOM and FlowMeans were labeled and merged using the *labelCluster* and *clusterSearch* function in MetaCyto.

In **Fig. 4 and 5**, Pre-processed data from all 10 studies were clustered using the *autoCluster.batch* function in MetaCyto. The summary statistics of the identified cell subsets were calculated using the *searchCluster.batch* function.

Evaluating the performance of clustering result.

In **Supplementary Fig. 1** and **Fig. 2 a-e**, the proportions of each cell type were provided by the authors of SDY420 and SDY478. MetaCyto or flowDensity[27] was used to estimate the proportion of each cell type. The Spearman correlation coefficient between author's result and MetaCyto or flowDensity result was calculated to measure the performance of MetaCyto and flowDensity.

In **Fig. 2f and Supplementary Fig. 2**, the F measure was used to measure the performance of clustering methods. The F measure was calculated as described in the FlowCAP study[33]. Briefly, for each cell population in manual gating and each cell population in auto-gating result, a 2 × 2 contingency table was calculated containing the false positive (FP), true positive (TP), false negative (FN) and true negative (TN). The recall (Re) was calculated as TP/(TP + FN), the precision

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Pr) was calculated a TP/(TP + FP). The F measure was calculated as F = (2 × Pr × Re)/(Pr + Re). For each population in manual gating result, the best F measure and its corresponding recall and precision were used as the F measure of the population. The overall F measure, Recall and Precision was the average of F measure, Recall and Precision of all manual gated populations, weighted by the size of each manual population.

Evaluating the performance of bisection algorithm

In **Supplementary Fig. 1**, Different Methods were tested to bisect the distribution of each marker.

Silhouette scanning method: The range of a marker was divided into 100 intervals using 99 breaks. The distribution was bisected at each break and the corresponding average silhouette[26] was calculated. The break giving rise to the largest average silhouette was used as the cutoff for bisection.

K-means method: based on the values of a single marker, cells were clustered into 2 groups using k means clustering algorithm were k = 2. The cutoff value for bisection was the border between the 2 groups.

Hierarchical clustering method: based on the values of a single marker, cells were grouped into a Hierarchical tree. The tree was then cut into 2 groups at the top level. The cutoff value for bisection was the border between the 2 groups.

First valley method: The distribution of each marker was smoothed using the *smooth.spline* function. The peaks in the distribution were identified using the *.getPeaks* function in flowDensity package[27]. The lowest points between peaks were defined as valleys. The valley with the smallest marker value was used as cutoff for bisection.

Last valley method: The valley with the largest marker value was used as cutoff for bisection.

Median valley method: The valley closest to the median of the marker value was used as cutoff for bisection.

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mean method: The mean of the marker distribution was used as the cutoff.

Median method: the median of the marker value was used as the cutoff

Middle method: the mean of the max and min of the marker values were used as the cutoff.

After markers in SDY420 data were bisected, cells fulfilling the requirement of each cell definition (listed in **Supplementary Table 1**) were identified. For example, for cell definition "CD3+ CD8+ CD4- ", cells falling into the CD3+ region were identified. Similarly, cells falling into CD8+ and CD4- regions were identified. The intersect of the 3 sets of cells were the cells corresponding to the cell definition "CD3+ CD8+ CD4- ". The proportion of cells corresponding to each cell definition was calculated and compared to the proportion provided by the author. The Spearman correlation was used as a measurement of the bisection algorithm.

Statistical Analysis

For the meta-analysis of the 10 human immunology studies from ImmPort, the proportion or MFI of cell subsets was regressed against age, gender and race (Y ~ age + gender + race) in each cytometry panel. The effect size was defined as the regression coefficient divided by the standard deviation of Y. The overall effect size from all cytometry panels was estimated using a random effect model. For data from the Carr study, the proportion of a cell population was regressed against age and gender. Race information was missing in the data, therefore was omitted in the regression. All statistical analysis was performed using the *metaAnalysis* function in MetaCtyo. The p-value was adjusted using the Benjamini-Hochberg[32] correction.

In **Fig. 4 a, b and d**, Pearson correlations are calculated and tested against the null hypothesis (correlation equals zero) using the *cor.test* function in R.

In **Fig. 5f**, Shapiro-Wilk test was performed to check the normality assumption using the *shapiro.test* function in R. F test was performed

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to check the equal variance assumption using *var.test* function in R. A two-sided unpaired Mann-Whitney test is performed to test the difference between two groups using the *wilcox.test* function in R.

SUPPLEMENTARY INFORMATION:

Supplementary Figure 1: Silhouette scanning bisects the distribution of each marker in a biological meaningful way. (**a**) An example illustrating the silhouette scanning. The range of CD8 is divided into 100 intervals using 99 breaks. The distribution is bisected at each break and the corresponding average silhouette is calculated. The break that gives rise to the largest average silhouette is used as the cutoff for bisection. Grey histogram shows the distribution of CD8. Blue dots show the average silhouette at each break. Red line shows the cutoff that maximizes the average silhouette. Black arrows show the position of 3 peaks. (**b-c**) Using different bisection algorithms, each marker in CyTOF data from SDY420 are bisected into positive and negative regions. 24 cell types were identified using the semi-supervised method as described in **Fig. 1c**. The proportion of each cell type in each sample is calculated and compared with manual gating result. (**b**) The Spearman correlation between the estimated proportion and author's proportion are used to measure the performance of each bisection algorithm. (**c**) Scatter plots showing the result generated by using silhouette scanning, k-means clustering and mean as the bisection algorithm. Each dot represents the proportion of a cell type in a sample. Each color represents a cell type. See online methods for a detailed description of the 9 bisection methods tested.

Supplementary Figure 2: MetaCyto is able to improve the quality and robustness of clustering results from FlowMeans. (**a,b**) FlowMeans is used to cluster FlowCAP WNV data with K ranging from 10 to 90 with or without MetaCyto. F measure (**a**) and the number of clusters (**b**) are showed in the bar plots.

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 3: Demographics of subjects included in the meta-anlaysis of 10 human immunology studies. Bar graphs show the distribution of gender (**a**), race (**b**) and age (**c**).

Supplementary Figure 4: Comparison between the effect sizes of gender estimated by MetaCyto using all 88 panels, against the effect size of gender estimated using the data from Carr, et al.

Supplementary Table 1: A list of cell definitions used to identify the 24 cell populations in cytometry data (SDY420) from ImmPort. The cell definitions are created based on the author's gating strategy provided in SDY420.

Supplementary Table 2: A list of cell definitions used to identify the 88 cell populations in cytometry data (SDY478) from ImmPort. The cell definitions are created based on the author's gating strategy provided in SDY478.

Supplementary Table 3: A summary of 10 studies included in the meta-analysis.

Supplementary Table 4: A list of cell definitions used to identify the 24 cell populations in all 10 studies included in the meta-analysis.

Supplementary Table 5: A table summarizing the effect size of age to the proportion and phenotype of 24 cell populations. The effect sizes are estimated from the meta-analysis of 10 human immunology studies using the guided analysis pipeline in MetaCyto.

Supplementary Table 6: A table summarizing the effect size of gender to the proportion and phenotype of 24 cell populations. The effect sizes are estimated from the meta-analysis of 10 human immunology studies using the guided analysis pipeline in MetaCyto.

Supplementary Table 7: A table summarizing the effect size of race to the proportion and phenotype of 24 cell populations when comparing cytometry data of blood from Asian and White subjects. The effect sizes are estimated from the meta-analysis of 10 human immunology studies using the guided analysis pipeline in MetaCyto.

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Table 8: A table summarizing the effect size of race to the proportion and phenotype of 24 cell populations when comparing cytometry data of blood from African American and White subjects. The effect sizes are estimated from the meta-analysis of 10 human immunology studies using the guided analysis pipeline in MetaCyto.

Supplementary Table 9: A table summarizing the effect size of race to the proportion of cell subsets identified by the unsupervised analysis pipeline when comparing cytometry data of blood from Asian and White subjects. The effect sizes are estimated from the meta-analysis of 10 human immunology studies.

Supplementary Table 10: A table summarizing the effect size of race to the proportion of cell subsets identified by the unsupervised analysis pipeline when comparing cytometry data of blood from African American and White subjects. The effect sizes are estimated from the meta-analysis of 10 human immunology studies.

ACKNOWLEDGMENTS

We thank Marina Sirota, Dvir Aran, Henry Schaefer, Elizabeth Thomson, Kelly Zalocusky and Matthew Kan for helpful discussion.

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN272201200028C). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COMPETING FINANCIAL INTERESTS

The authors declare no conflict of interest

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

AUTHOR CONTRIBUTIONS

Z.H., A.J.B. and S.B. conceived the study and developed the method. Z.H. and C.J. designed the overall structure of MetaCyto and performed the meta-analysis of cytometry data from ImmPort. J.H., M.S., P.F.G., S.A., P.D., C.T., J.W., G.N. gave valuable input and suggestions for analysis. H.Z. and C.J. wrote the manuscript and made figures with input from co-authors.

REFERENCES:

1. Sutton, A. J., Abrams, K. R., Jones, D. R. & Sheldon, T. A. Methods for Meta-analysis in Medical Research Contents Preface Acknowledgements Part A: Meta-Analysis Methodology: The Basics.
2. Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.* **10,** R65 (2008).
3. Kodama, K. *et al.* Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 7049–54 (2012).
4. Boulé, N. G., Haddad, E., Kenny, G. P., Wells, G. A. & Sigal, R. J. Effects of Exercise on Glycemic Control and Body Mass in Type 2 Diabetes Mellitus. *JAMA* **286,** 1218 (2001).
5. Brown, S. D. *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.* **24,** 743–50 (2014).
6. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41,** 703–707 (2009).
7. Perfetto, S. P., Chattopadhyay, P. K. & Roederer, M. Innovation: Seventeen-colour flow cytometry: unravelling the immune system. *Nat. Rev. Immunol.* **4,** 648–655 (2004).
8. Shapiro, H. M. Multistation multiparameter flow cytometry: A critical review and rationale. *Cytometry* **3,** 227–243 (1983).
9. Bandura, D. R. *et al.* Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Anal.*

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chem. **81,** 6813–6822 (2009).
10. Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* **58,** 234–239
11. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10,** 228–238 (2013).
12. Weber, L. M. & Robinson, M. D. Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. *bioRxiv* 47613 (2016). doi:10.1101/047613
13. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom. Part A* **87,** 636–645 (2015).
14. Aghaeepour, N., Nikolic, R., Hoos, H. H. & Brinkman, R. R. Rapid cell population identification in flow cytometry data. *Cytom. Part A* **79A,** 6–13 (2011).
15. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci.* **111,** E2770–E2777 (2014).
16. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29,** 886–891 (2011).
17. Wertheimer, A. M. *et al.* Aging and Cytomegalovirus Infection Differentially and Jointly Affect Distinct Circulating T Cell Subsets in Humans. *J. Immunol.* **192,** 2143–2155 (2014).
18. Obermoser, G. *et al.* Systems Scale Interactive Exploration Reveals Quantitative and Qualitative Differences in Response to Influenza and Pneumococcal Vaccines. *Immunity* **38,** 831–844 (2013).
19. Ledgerwood, J. E. *et al.* Influenza virus h5 DNA vaccination is immunogenic by intramuscular and intradermal routes in humans. *Clin. Vaccine Immunol.* **19,** 1792–7 (2012).
20. Whiting, C. C. *et al.* Large-Scale and Comprehensive Immune Profiling and Functional Analysis of Normal Human Aging. *PLoS One* **10,** e0133627 (2015).
21. Achhra, A. C. *et al.* Difference in absolute CD4+ count according to CD4 percentage between Asian and Caucasian HIV-infected patients. *J. AIDS Clin. Res.* **1,** 1–4 (2010).
22. Coussens, A. K. *et al.* Ethnic Variation in Inflammatory Profile in Tuberculosis. *PLoS Pathog.* **9,** e1003468 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

23. Golden-Mason, L., Klarquist, J., Wahed, A. S. & Rosen, H. R. Cutting edge: programmed death-1 expression is increased on immunocytes in chronic hepatitis C virus and predicts failure of response to antiviral therapy: race-dependent differences. *J. Immunol.* **180,** 3637–41 (2008).
24. Li, Y. *et al.* Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nat. Med.* **22,** 952–60 (2016).
25. Brodin, P. *et al.* Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* **160,** 37–47 (2015).
26. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
27. Malek, M. *et al.* flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics* **31,** 606–607 (2015).
28. Finak, G. *et al.* Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. *Sci. Rep.* **6,** 20686 (2016).
29. Carr, E. J. *et al.* The cellular composition of the human immune system is shaped by age and cohabitation. *Nat. Immunol.* **17,** 461–468 (2016).
30. Howard, R. R., Fasano, C. S., Frey, L. & Miller, C. H. Reference intervals of CD3, CD4, CD8, CD4/CD8, and absolute CD4 values in asian and non-asian populations. *Cytometry* **26,** 231–232 (1996).
31. Saeys, Y., Gassen, S. Van & Lambrecht, B. N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **16,** 449–462 (2016).
32. Author, T., Benjamini, Y., Hochberg, Y. & Benjaminit, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Source J. R. Stat. Soc. Ser. B J. R. Stat. Soc. Ser. BMethodological) J. R. Stat. Soc. B* **57,** 289–300 (1995).
33. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10,** 228–238 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends:

Figure 1: MetaCyto identifies and labels common cell subsets in cytometry data across studies. (**a**) Schematic illustration of the 4 steps MetaCyto uses to perform meta-analysis of cytometry data. (**b**) Schematic illustration of the unsupervised analysis pipeline in MetaCyto. Top: Cytometry data from different studies are first clustered using a clustering method, such as FlowSOM. Middle: Each marker is bisected into positive and negative regions using the silhouette scanning method. Each identified clusters is labeled based on their position relative to this threshold. Bottom: Clusters with the same label are merged together into rectangles or hyper-rectangles. (**c**) An example illustrating the guided analysis pipeline in MetaCyto. Each marker in the data is bisected into positive and negative regions using the silhouette scanning method. The CD3+ CD4+ CD8- cluster corresponds to cells that fall into CD3+ region, CD4+ region and CD8- region at the same time. Red histograms show the distribution of markers in CD3+ CD4+ CD8- subset. Grey histograms show the distribution of markers of all cells.

Figure 2: Both guided and unsupervised analysis pipelines in MetaCyto accurately identify cell populations. (**a-c**) Scatter plots showing the comparison between proportions of cell types estimated by the guided analysis pipeline in MetaCyto and proportions provided by the authors of SDY478. All cell populations (**a**), CD16- monocytes (**b**), and effector memory CD8+ T cells (**c**) are included in the plots. Each dot represents the proportion of a cell type in a sample. Each color represents a cell type. (**d**) Scatter plots showing the comparison between flowDensity and manual gating. All cell populations are included. (**e**) The 88 cell types are broken down into rare and major populations based on their mean proportion in the manual gating results. The cell types whose mean proportions are less than 2 percent are defined as rare population, the rest cell types are defined as major populations. Spearman correlation between MetaCyto or flowDensity's results and manual gating results are calculated to measure the performance. (**f,g**) FlowSOM is used to cluster the West Niles Virus dataset (FlowCAP WNV) with K ranging from 10 to 90

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with or without the merge step in MetaCyto unsupervised analysis pipeline. F measure (**f**) and the number of clusters (**g**) are showed in the bar plots. See **Supplementary Fig. 2** for a comparison between MetaCyto and FlowMeans.

Figure 3: Data from 10 human immunology studies includes highly heterogeneous cytometry panels. Eighty-eight panels with diverse sets of markers were used in these 10 studies, with the panels represented vertically. The specific markers used are represented horizontally. A red square in each grid element indicates that particular marker was used in a study panel.

Figure 4: Meta-analysis of cytometry data using MetaCyto provides consistent results between cytometry panels and confirms previous findings. (**a**) The 88 cytometry panels included in the meta-analysis were randomly split into 2 panel sets. Twenty-four specific cell types were identified by the guided analysis pipeline. The effect size of age on the proportion of the cell types was estimated in the two panel sets independently and then compared with each other. (b) Eighty-two cell subsets were identified by the unsupervised analysis pipeline in two independent panel sets and compared. (**c**) The procedure of (**a**) and (**b**) were repeated 100 times, the mean and standard deviation of the 100 correlation between two panel sets are shown in the bar plot. (**d**) Comparison between the effect sizes of gender estimated by MetaCyto using all 88 panels, against the effect size of age estimated using the data from Carr, et al. (**e**) Forest plot showing the effect size of race (Asian compared to White) on the proportion of CD4+ T cells in the blood. The effect sizes are estimated within each panel first, and are combined using a random effect model. Panels in blue represent flow cytometry data. Panels in black represent CyTOF data. In (**a**,**b**,**d**), r represents the Pearson correlation; p represents the p value of r not equal to 0. In (**e**), p represents the p value of the overall effect size not equal to 0.

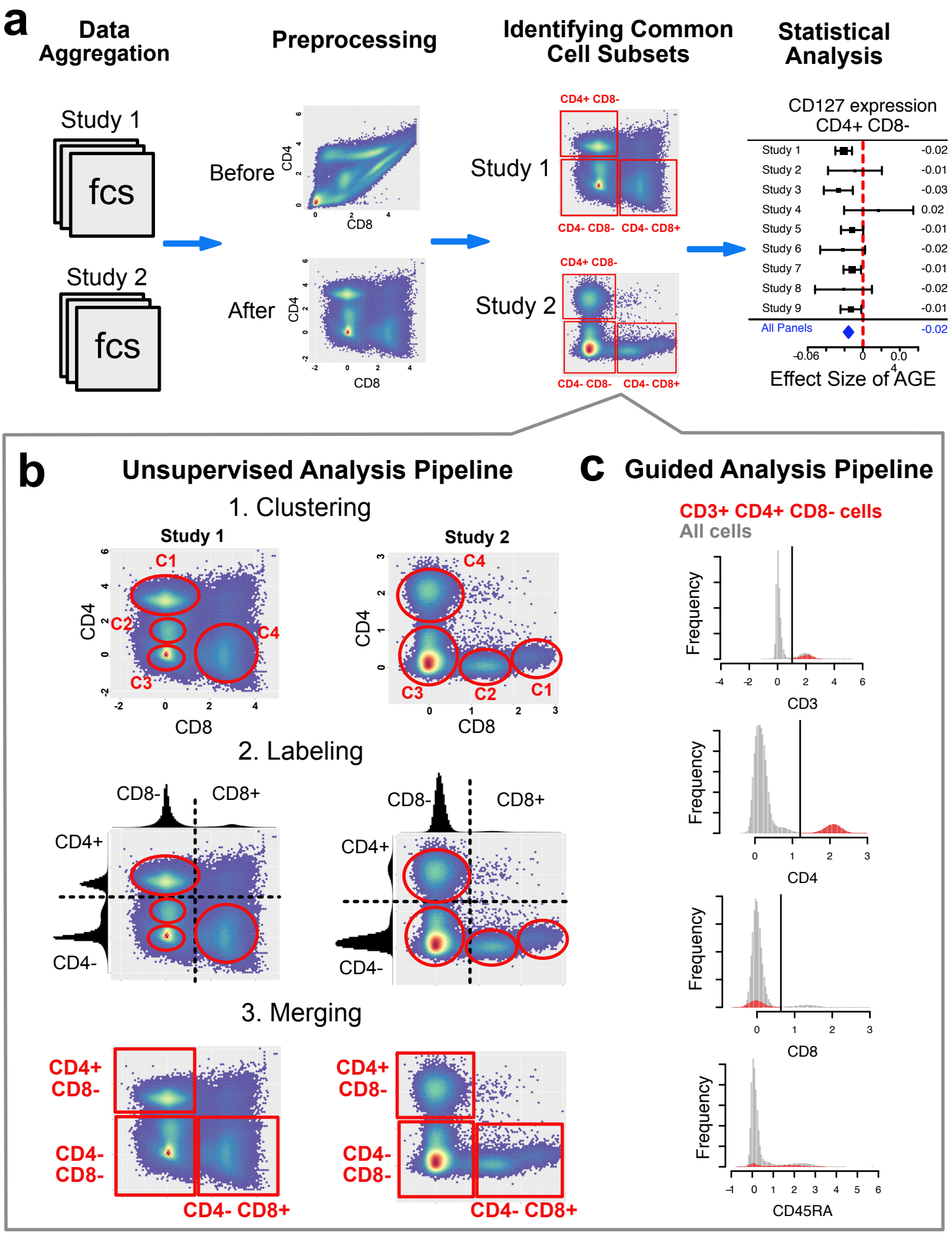
Figure 5: Meta-analysis of cytometry data using MetaCyto identifies multiple racial differences in immune cells. (**a**) Heat map showing the effect size of race (Asian vs. White) on the proportion and the expression of 7 selected markers in 13 selected cell types. Color scale represents the value of the effect size (red means higher in

bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Asian than White, blue means lower). The area of circle represents the absolute value of the effect size. Red squares indicate significant differences. (**b**) Heat map showing the effect size of race (African American vs. White) on the proportion and expression of 5 selected markers in 9 selected cell types. (**c**) Forest plot showing the effect size of race (Asian vs. White) on the MFI of CD94 on NK cells. (**d**) Cytometry data from panel 1 of SDY420 is manually gated to identify NK cells. Representative histogram of CD94 expression on NK cells from Asian, African American and White subjects were shown. (**e**) Forest plot showing the effect size of race (Asian vs. White) on the proportion of a novel cell subset (CD3+CD4-CD45RA+CD8+CD85J-) identified by the unsupervised analysis pipeline. (**f**) The proportion of the CD3+CD4-CD45RA+CD8+CD85J- cell subset in PBMC from Asian and White subjects in panel 2 of SDY312. The p values in **a-c** are calculated using random effect models, adjusted using Benjamini-Hochberg correction. p value in **f** is calculated from unpaired Mann-Whitney test without correction. See **Supplementary Table 5-8** for a complete list of differences in immune cells between races.

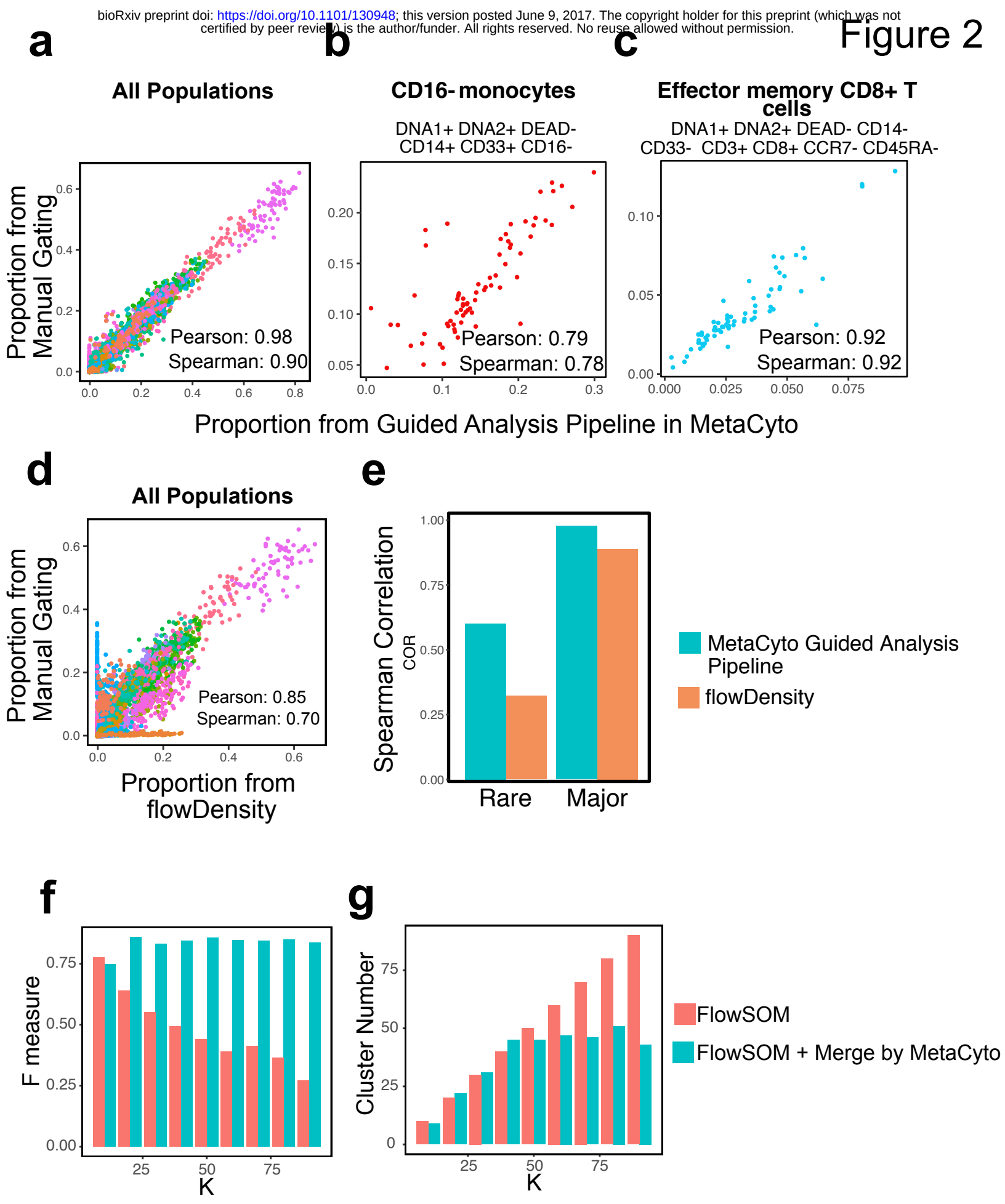
bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1



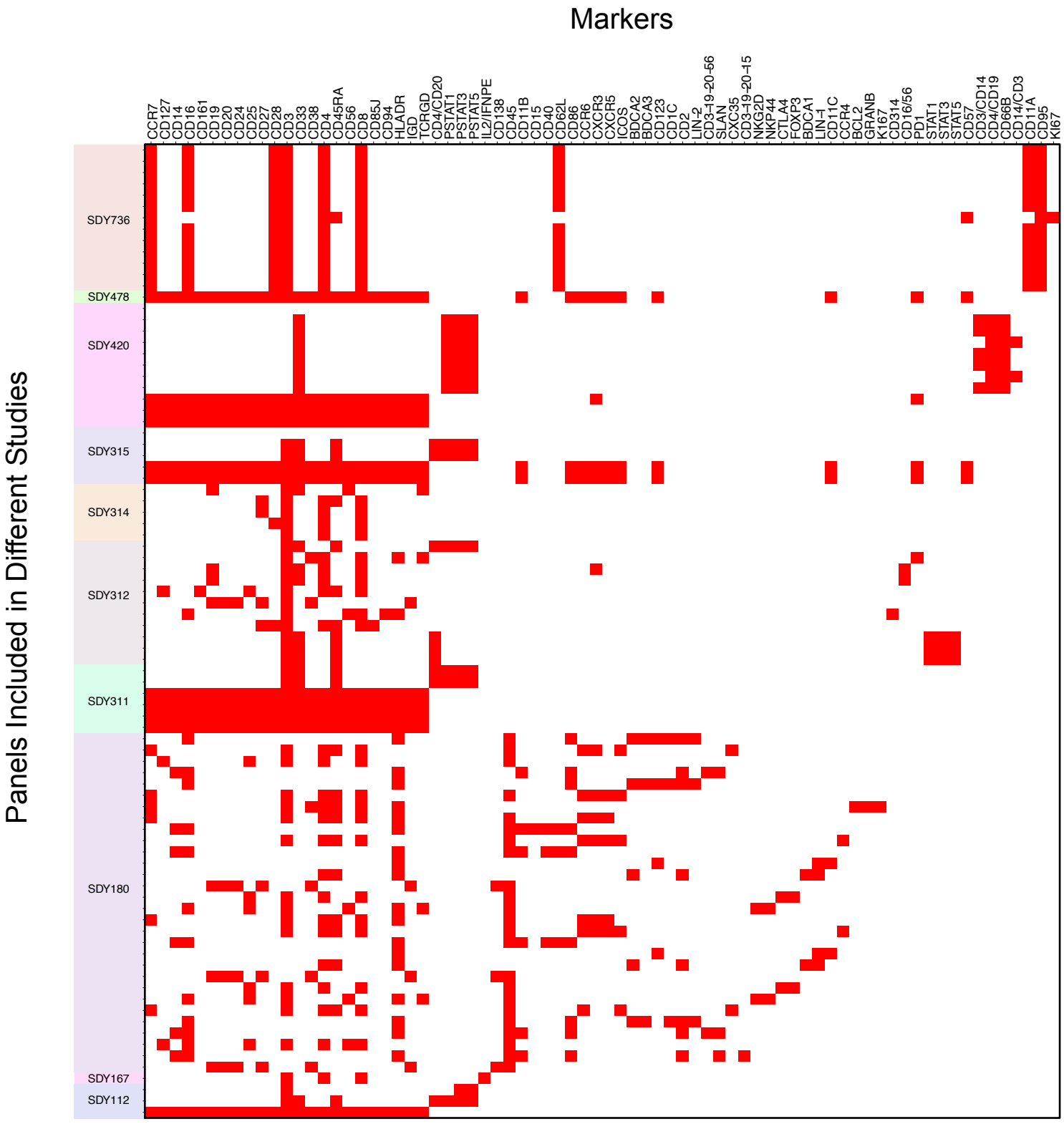
bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2



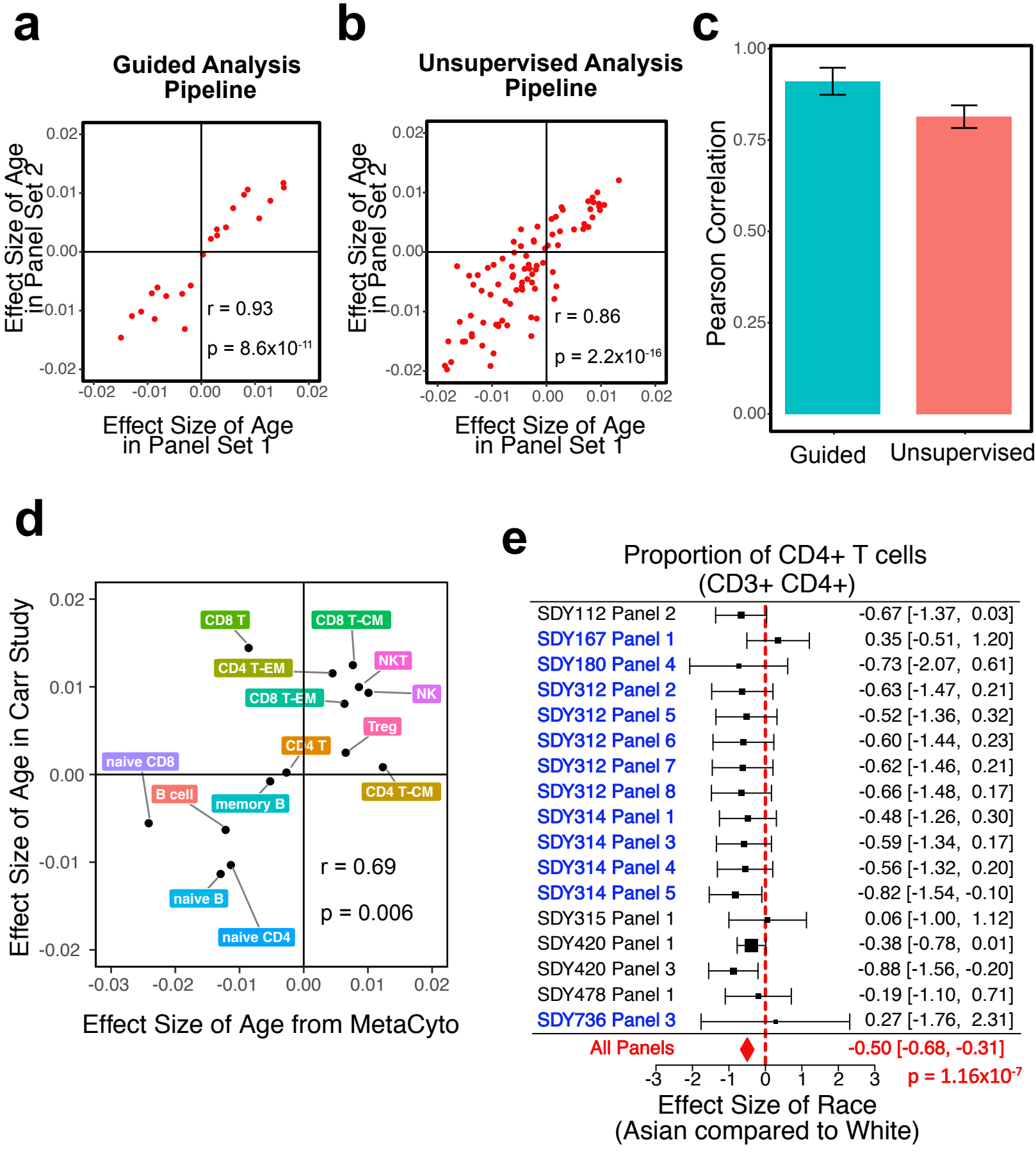
bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3



bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/130948; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5