High-throughput single-cell quantification of hundreds of proteins using conventional flow cytometry and machine learning

Etienne Becht1, Daniel Tolstrup2, Charles-Antoine Dutertre3,4,5, Peter A. Morawski6, Daniel J. Campbell6,7, Florent Ginhoux3,5,8, Evan W. Newell1†, Raphael Gottardo1†, Mark B. Headley2,7†

Modern immunologic research increasingly requires high-dimensional analyses to understand the complex milieu of cell types that comprise the tissue microenvironments of disease. To achieve this, we developed Infinity Flow combining hundreds of overlapping flow cytometry panels using machine learning to enable the simultaneous analysis of the coexpression patterns of hundreds of surface-expressed proteins across millions of individual cells. In this study, we demonstrate that this approach allows the comprehensive analysis of the cellular constituency of the steady-state murine lung and the identification of previously unknown cellular heterogeneity in the lungs of melanoma metastasis–bearing mice. We show that by using supervised machine learning, Infinity Flow enhances the accuracy and depth of clustering or dimensionality reduction algorithms. Infinity Flow is a highly scalable, low-cost, and accessible solution to single-cell proteomics in complex tissues.

INTRODUCTION

One of the cornerstones of immunology has been the detailed phenotypic classification of the cellular populations of the immune system. These efforts have largely been guided by identifying heterogeneity in surface protein expression across cell types. However, the breadth of this classification has been hindered by technical requirements allowing the analysis of only small subsets of markers per sample and analytical tools that are generally manual and low throughput. Over the past decade, methods that enable deeper interrogation of cellular heterogeneity in complex tissues and systems have provided a better understanding of the mechanistic underpinnings of disease. The key to this idea is being able to simultaneously measure large numbers of parameters on individual cells. This increased dimensionality facilitates the understanding of the unique characteristics of each individual cell and how cells interact within a given system. Modern flow cytometric approaches exemplify this by using panels of multiple fluorochrome-conjugated (conventional flow cytometry) or metal-conjugated (mass cytometry) antibodies to measure protein expression profiles of individual cells with high cell throughput to capture and analyze both common and rare cell populations. With current instrumentation, fluorescence and mass cytometric approaches are unfortunately limited to 40 or fewer parameters. However, at least 371 cluster of differentiation (CD) markers are currently recognized (1), not including relevant surface proteins without CD designation. This disparity indicates that we are vastly limited in the comprehensiveness of these assays. In addition, robust implementation of panels approaching 40 parameters requires extensive experience and optimization, as well as dedicated and uncommon instrumentation. Recently, several approaches have been published using oligonucleotide-conjugated antibodies in combination with single-cell RNA sequencing technology to simultaneously measure single-cell transcriptomes and hundreds of surface proteins (2, 3). Unfortunately, these techniques are still limited in terms of cellular throughput (tens of thousands of single cells), remain expensive (requiring deep sequencing and large antibody panels), and often require complicated experimental protocols. Hence, there is a need for methods enabling concordant analysis of the expression patterns of high numbers of proteins at a single-cell level while retaining both accessibility, low cost, and high cell throughput.

Parallel to these experimental developments, computational techniques have also matured and found new applications in biology. Numerous studies have shown that independent single-cell experiments can be combined into a single augmented dataset, suggesting that multiple and partially overlapping datasets can be leveraged to obtain a single integrated and informative data matrix (4–9). Machine learning techniques have also substantially improved in the last decade, with well-known applications to fields such as genomics, computer vision, or speech recognition (10, 11). Although deep learning (using deep neural networks) is one of the most popular techniques at the moment, machine learning encompasses a large number of tools including regularized regression (12), support vector machines (SVMs) (13), or ensemble of decision trees (14–16), to name a few. A common feature of these algorithms is their ability to model nonlinear relationships. Nonlinearity corresponds to a relationship between variables that cannot be accurately modeled by a straight line (or its high-dimensional equivalent). In the context of single-cell analyses, the relationship between expression levels of a set of surface proteins is a case of a nonlinear system. This nonlinear system is often implicitly modeled as a Boolean formula, for instance, “CD45 expression and CD19 expression implies CD79b expression.” Immunologists have extensively published these limited nonlinear relationships between marker expression and cell phenotypes in the
past decades. The actual system is, however, more subtle, as cell surface proteins are not categorical but continuous variables, depend on the biological context, and can be multimodal.

In this article, we demonstrate the capabilities of Infinity Flow, a low-cost methodology and computational toolset for high-dimensional single-cell analysis of complex cell suspensions using standard flow cytometry instrumentation, off-the-shelf reagents, and scalable antibody panels. Infinity Flow uses machine learning to systematize imputation of cell surface protein expression at the single-cell level from experimental flow cytometry data. We leverage massively parallel cytometry (MPC) experiments, which sparsely measure hundreds of exploratory antibodies and exhaustively measure a small number of well-characterized and informative so-called Backbone antibodies. Machine learning is then used to impute expression levels of each exploratory marker for each single cell based on nonlinear functions of the measured Backbone markers. Infinity Flow enables the concurrent analysis of coexpression patterns of hundreds of surface proteins across millions of cells at single-cell resolution at a fraction of a cent per cell (U.S. dollars). This approach can be implemented on any existing flow cytometry platform with off-the-shelf reagents, rendering it highly accessible. We previously illustrated a proof-of-concept approach to delineate conventional dendritic cells 2 (cDC2) and circulating monocytes (17) or to resolve the developmental trajectory of the neutrophil lineage during hematopoiesis (18). In this article, we present the complete and expanded methodology including an open-source R package infinityFlow submitted to the Bioconductor repository, optimized machine learning algorithms with benchmarks, single-cell background correction, and high-dimensional computational analyses of millions of single cells from the murine lung at steady state and during metastatic seeding. Using our approach, we show that nonlinear regression is superior to linear regression for machine learning–based prediction of flow cytometric data. Further, we highlight the ability for Infinity Flow to augment a standard flow cytometric panel and enable more robust population and protein expression characterization. We use this approach to comprehensively define the cell populations of the murine lung, only using a 15-color flow cytometry panel. Last, we identify two populations of tumor-ingesting macrophages during early metastatic seeding of the lung by melanoma cells that can be separated by PD-L1, MHCII (major histocompatibility complex II), and CD11c expression.

RESULTS

The Infinity Flow pipeline

Infinity Flow processes data from low-cost (table S1) commercially available plate-based antibody screening panels (which we term MPC), using machine learning to achieve the simultaneous analysis of hundreds of surface markers on millions of single cells isolated from complex tissues. Similar to conventional flow cytometry, MPC experiments start by staining a sample with a cocktail of fluorescently labeled antibodies, here termed the Backbone panel (Fig. 1A). Critical to the MPC approach, the Backbone panel should be designed with at least one empty fluorescence channel [usually phycoerythrin (PE) or allophycocyanin (APC)] and optimized to minimize spectral overlap into this open channel. The sample is then aliquoted across w wells (typically w ≈ 300) (Fig. 1B) (17, 19–24). Each well contains a distinct antibody clone conjugated to the fluorophore not used for the Backbone. We term the set of well-specific markers the Infinity panel. After completion of the staining step, each well contains a fraction of the total sample, and all cells within are stained with all Backbone markers in addition to a single Infinity marker (Fig. 1C). Each well is then acquired as an individual flow cytometry sample using a conventional flow cytometer. Commercial kits exist from various vendors with a range of 240 to 371 individual antibodies; these kits can be further augmented with researcher-selected antibodies as needed.

Each well-specific dataset (stored in its own FCS file), resulting from an MPC experiment, contains empirical measurements of forward scatter (FSC) and side scatter (SSC) parameters, every Backbone marker, and empirical measurement of one of the Infinity panel markers. The data matrix resulting from all wells can be viewed as a dense matrix of n events × b Backbone markers jointly measured with a sparse n × w matrix of Infinity markers (Fig. 1D). Following standard quality control of the data (Methods), the data are analyzed using nonlinear multivariate regression to recast this disjointed data structure into a single cohesive expression matrix of all markers across all cells. To achieve this, we train, for every well, a machine learning model that predicts the expression of the Infinity panel marker on a continuous scale from the measured FSC, SSC, and Backbone marker intensities. Once trained, these models are applied across the whole dataset to estimate the intensity of each of the w Infinity antibodies across the n events, resulting in an n × w dense Infinity matrix of imputed intensities (Fig. 1E).

This computational workflow is illustrated on cells isolated via collagenase digestion of nonperfused whole mouse lungs and stained using a standard 14-color immunoprofiling Backbone panel and a BioLegend Murine LEGENDScreen MPC kit. The lung, even under homeostatic conditions, contains an exceptionally diverse cellular milieu, composed of common and rare, immune and nonimmune cells, providing an ideal testing ground for our high-dimensional approach. For this simplified example, we subsampled a total of 1000 mouse lung cells from 10 wells each containing an antibody to a distinct CD molecule. We used hierarchical clustering to highlight structure within the Backbone (fluorescence and scatter) (Fig. 1F). This structure correlates with distinct expression patterns in the sparse Infinity marker measurements (Fig. 1G). The goal of Infinity Flow is to model these correlations in a data-driven manner using machine learning. These models then impute the expression of each Infinity panel marker across every cell in the dataset (Fig. 1H). The dense, continuous, and single-cell data format of Infinity Flow’s output enables easy visualization and exploration of any combination of co-expression patterns across both the Backbone and Infinity panels (Fig. 1I). Infinity Flow’s output is notably compatible with standard flow cytometry analysis software [e.g., FlowJo or flowCore (25)] and can be used as input for further downstream analyses such as clustering, dimensionality reduction, or automated gating.

Nonlinear regression models accurately impute cytometry data

To maximize the quality of the predictions of the Infinity Flow algorithm, we quantitatively assessed its performance across Infinity markers and machine learning algorithms on the mouse lung cell dataset. Protein expression on cells is a continuous variable. However, standard flow cytometry analysis generally involves a progressive gating scheme whereby expression levels are discretized into two or more bins (e.g., low expression versus high expression), with potentially imbalanced frequencies (e.g., many negative events and few
positive events). To account for this imbalance and pseudo-discrete format, we decided to use the area under the receiver operating characteristic (ROC) curve (AUC) on held-out data as a performance metric as opposed to continuous metrics such as the $R^2$ coefficient (Methods), focusing our effort on the magnitude and frequency of imputed expression.

The Infinity Flow R package supports a variety of linear and nonlinear regression–based machine learning models. Given the expected nonlinear relationships of protein expressions, we hypothesized that nonlinear models would outperform linear models in this setting. We therefore compared the performance of the nonlinear algorithms (12, 13, 16, 26) to linear L1-penalized models (12). Consistent with our hypothesis, nonlinear methods significantly outperformed linear models (Fig. 2A). Overall, each nonlinear regression method performed well, with a median AUC between 0.88 and 0.89 (table S2), while the linear model performed much worse (median AUC of 0.71). Gradient boosted trees implemented by the XGBoost library (16) performed slightly better in this benchmark (Fig. 2A and fig. S1).

True staining performance can vary widely across the range of empirical measurements. This can be due to a lack of the marker in
the sample [e.g., red blood cell (RBC) lysis during sample preparation precludes detectable expression of the RBC-specific marker Ter-119], sensitivity to tissue preparation method (e.g., CD115 is heavily cleaved by many collagenase enzymes), staining approach (many chemokine receptors require staining at 37°C for robust signal), or simple reagent quality as not all antibodies perform equally well. Thus, evaluation of performance must account for this empirical variability and algorithm performance. Consistently, performance metrics for Infinity markers fell within a range of AUC from 0.72 to 0.98. We highlight 12 markers sampled across the whole range of performances in Fig. 2B. As expected, markers that resulted in high AUC were typically multimodal and either highly or commonly expressed. Exhaustive manual examination showed that 155 of 252 phenotypic markers (61.5%, excluding isotype controls and autofluorescence measurements) yielded meaningful imputed signal.

For a limited set of seven markers consisting mostly of T cell receptor (TCR) spectratyping antibodies (e.g., TCR Vβ7 and TCR Vβ9), the models were unable to accurately impute expression levels despite the presence of a positive cell population (fig. S2). This reflects an expected limitation of the approach: The expression of an Infinity marker requires a unique signature in the Backbone data to be accurately imputed. A TCR Vβ7+ T cell, however, resembles any other T cell with our Backbone panel. This expected negative result underscores the importance of the Backbone panel design in the Infinity Flow approach. Further, it highlights that future refinement of the choice of Backbone antibodies and tuning methodology to a given sample type may enhance the overall breadth of the assay.

Markers restricted to rare cell populations, in general, gave lower AUC values, although still well above an AUC of 0.5 (corresponding to a random guess) (Fig. 2B). To estimate a minimal number of...
positive events allowing satisfactory performance when predicting a rarely expressed marker, we studied the influence of the number of events during training on the performance of these models. We trained each algorithm for each marker on a varying number of training events (from 1000 to 50,000) and tested accuracy on held-out data. For each algorithm, performance increased with training set size and plateaued after 10,000 events. Widely expressed markers such as CD48 or CD98 were mostly unaffected by this restriction of the training set size. Performance was, however, affected for CD200R3, a marker specifically expressed on basophils, which are exceedingly rare in the analyzed dataset (0.82% of the total live single cells). At very low cell numbers (1000 sampled event per well or, on average, 8 basophils and 992 nonbasophils during training), XGBoost and SVM still resulted in high AUC (≥0.8), suggesting that these methods are able to accurately predict the expression of a marker specific for a rare cell population even when training data are limited (Fig. 2C and fig. S3).

The performance benchmarks described above are performed per well. However, technical variability (batch effects) can further affect prediction accuracy across wells. The goal of the Infinity Flow method is to enable cross-well imputation of marker expression with high accuracy. Thus, to test for the ability of these models to generalize to data independent from a specific well, we compared the cross-well predictions with within-well predictions of the models by rotating out Backbone markers one at a time and using the remaining Backbone markers to predict the intensity of the left-out marker. The impact of cross-well prediction on performance was limited (median ΔR^2 from −0.015 to −0.012 across algorithms; fig. S4A). XGBoost was more accurate than other nonlinear models for virtually any marker–algorithm pair (fig. S4B).

Last, we found that the total runtime of XGBoost was the shortest of all benchmarked methods (Fig. 2D). Because of its accuracy at both high and low cell numbers and speed, the XGBoost implementation of gradient boosted trees is the default imputation method offered in the Infinity Flow package. All tested nonlinear algorithms provided accurate and robust imputed data for most markers, highlighting the robustness of the Infinity Flow approach to the choice of multivariate regression framework used.

**Infinity Flow enables cell-level background correction in MPC assays**

One of the limitations of flow cytometry data is that the signal on nonexpressing cells still follows a nonrandom pattern, mostly due to autofluorescence and unspecific antibody binding, even with standard Fc receptor blocking approaches. Hence, our derived predictions include a significant and underdispersed contribution from nonspecific signal. Using Infinity Flow–imputed data, we were able to enhance the signal of dimly expressed antibodies by performing cell-wise correction of the nonspecific background fluorescence signal: Because Infinity Flow jointly imputes the expression of each Infinity marker alongside the “expression” of isotype-matched control antibodies across every cell, we could mathematically model this background signal (Methods). The residuals of these models, akin to per-cell background subtracted expression values, were used as background-corrected data (fig. S5A). These background-corrected data enabled clear identification of marker-expressing cells for some dimly expressed markers, such as CD279 (PD-1) (fig. S5B), and removed spurious correlation patterns within nonexpressing cells (fig. S5, C and D). Of note, this approach could also find applications in conventional flow cytometry, whenever isotype controls are included in the antibody panel, as per-cell background correction is easier to interpret than traditional methods based on manual gating.

**Infinity Flow enables the comprehensive annotation of cellular populations in complex samples**

Having established the accuracy and robustness of the Infinity Flow approach, we next sought to use it in a real-world experimental context. Our laboratories have extensive experience using flow cytometry to profile the cellular environments of tissues such as lung (27). In the past, comprehensive analysis of such a tissue, inclusive of diverse immune and nonimmune populations, would have required several multiparameter flow cytometry panels run in sequence and would still have failed to classify many cell types. Further, comprehensive phenotyping of these cells would have been exceptionally labor-intensive by this standard approach. We thus evaluated whether the Infinity Flow approach could improve upon a conventional multicolor flow cytometry panel and offer new insights into the cellular constituency of complex tissues and disease states.

We first applied Infinity Flow to the complete mouse lung cell dataset (partly shown in Fig. 1F). Mouse lung cells were isolated from the unperfused lungs of C57BL/6 mice at the steady state and stained using a 14-color Backbone panel consisting of 11 antibodies in individual channels, 2 multiplexed channels (containing various lineage markers in combination), live/dead stain, and 2 light scatter parameters (table S3). The Infinity Panel consisted of 252 PE-conjugated antibodies and 14 matched isotype and autofluorescence controls (table S4). A total of 28,715,415 live single-cell events were acquired. We used 50% of the measured events to train the machine learning models, and output imputed data for w = 266 antibodies across n = 10,000 × 266 = 2,660,000 single cells distinct from those used for training the models (fig. S6).

Our standard gating scheme on the Backbone for a panel of this nature allowed us to account for only 37.7% of live single lung cells (Fig. 3, A and B). Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction (28, 29) of the Backbone data, however, revealed additional information pertaining to cell populations within the Backbone measurements and not captured using this standard approach (Fig. 3B). Most of the unclassified events were either nonimmune (CD45+ or positive for the highly multiplexed lineage channel composed of a cocktail of eight antibodies specific to T cells, B cells, natural killer (NK) cells, neutrophils, alveolar macrophages, and eosinophils (table S3). Infinity Flow deconvolved this multiplexed lineage channel into its constituent markers, enabling straightforward and accurate identification of the individual cell types, which comprise these events (fig. S7). These results highlight the ability of this approach to access and use the information contained within these highly multiplexed channels.

To enhance our ability to delineate cell types within the murine lung, we applied Phenograph clustering (30) to the Backbone measurements. Centroids of the Infinity Flow–imputed expression intensities across a set of 39 known cell type defining markers across clusters allowed identification of almost every cell cluster (97% of live single cells, 32 of 33 clusters) (fig. S8). Only a single cluster (cluster 18) that was negative for virtually every lineage marker was left unclassified (fig. S8). In addition, combining Infinity Flow–imputed data with Phenograph clustering allowed us to identify and classify rare cell populations of CD200R3+ FcεR1+ basophils and of innate lymphoid cells (ILCs). These rare cell populations represented...
Fig. 3. Infinity Flow enables near-exhaustive phenotyping of lung cells. (A) Manual gating of the Backbone data from cells isolated from the lungs of C57BL/6 mice and (B) UMAP dimensionality reduction of the Backbone data, colored by cell phenotypes from manual gating. (C) Median Infinity Flow predictions of 155 Infinity markers showing staining across cell clusters defined by the Phenograph algorithm on the Backbone data. (D) Projection of the cell clusters on the UMAP embedding of the Backbone data.
respectively only 0.82 and 1.08% of the total live single cells (fig. S9). This analysis thus resulted in a near-exhaustive phenotypic classification of the mouse lung at the single-cell level, associated with extensive protein expression profiles for each cell population (Fig. 3C). From the Infinity Flow–predicted expression data, we selected CD31, CD41, Clec12a, and CD200R3 as markers that differentiated basophils and eosinophils within FcεR1+ cells. Staining with these markers in an independent flow cytometry experiment validated the predicted coexpression patterns for these markers (fig. S10). In addition, this approach enabled classification of several major components of the nonimmune constituency of the lung, an unexpected benefit that could be enhanced with Backbone panel design to enable deep simultaneous classification of both immune and nonimmune cell types in the lung (Fig. 3, C and D).

In addition to broad cellular characterization, Infinity Flow can also be used to analyze a particular cell population in depth. We performed MPC on human peripheral blood mononuclear cells from a single donor and processed the resulting dataset with the Infinity Flow computational pipeline. This allowed us to identify markers that were heterogeneously expressed within CD8 T cells in the blood, such as CD55, CD95, CD279 (PD-1), and CX3CR1 (fig. S10). In this experiment, we stained four fractions of this sample using a CD45

---

**Fig. 4. Infinity Flow increases the signal-to-noise ratio of MPC datasets.** (A) Side-by-side Phenograph clustering and UMAP embedding of the Backbone data (left) and the Infinity Flow–augmented dataset (right). (B) Distribution of markers B cell subtype 1 and B cell subtype 2 in the B cell Backbone cluster B20 (gray) and the two Infinity Flow–augmented B cell clusters IF3 and IF5. (C) Distribution of markers of naïve and previously activated CD4 T cells in the CD4 T cell Backbone cluster B7 (gray) and the two Infinity Flow–augmented CD4 T cell clusters IF10 and IF14. (D) Distribution of markers of NK and T cells in the mixed T CD8 and NK cell Backbone cluster B16 (gray), or the Infinity Flow–augmented (E) T cell cluster IF11 or (G) NK cell cluster IF20.
barcoding technique within the backbone. The staining-predicted co-expression patterns were consistent across subsamples (fig. S10). These results show that Infinity Flow can be performed to analyze a cell population in depth and is compatible with barcoding techniques.

**Infinity Flow increases the signal-to-noise ratio of MPC datasets**

One of the exciting applications of this method is to better understand and profile the heterogeneity in cell types within a complex tissue. Clustering and nonlinear dimensionality reduction algorithms are becoming increasingly common methods to define cell subsets within single-cell data. However, these approaches are often unsupervised and are therefore limited in their ability to distinguish real patterns from noise in data. Because Infinity Flow uses supervised learning methods, we reasoned that it could enhance our ability to identify biologically relevant cell populations in Infinity Flow–processed datasets compared to the analysis of the Backbone data alone. To compare the richness of Infinity Flow’s output matrix to the initial Backbone matrix, we reanalyzed the murine lung dataset using UMAP dimensionality reduction and Phenograph clustering (Methods), now including both the Backbone and imputed measurements as input (Fig. 4A).

Overall, applying UMAP dimensionality reduction and Phenograph clustering on the Infinity Flow–augmented data leads to more accurate and informative Phenograph clustering and UMAP embedding. Notably, it separated B cells into two distinct Phenograph clusters that were well separated in the UMAP embedding (Fig. 4, A and B, and movie S1). Imputed marker expression suggested that Infinity Flow was able to resolve the recently described B1 (glycosylated CD43+ IgD+ CD21/35+ CD272+) and B2 (glycosylated CD43+ IgD+ CD21/35+ CD272-) subsets of B cells in this dataset (31, 32). Similarly, Phenograph clustering of the Infinity Flow–augmented data separated CD4 T cells into two distinct clusters, corresponding to naïve (CD44+ CD38+ CD62L+ CD45RB+) and previously activated (CD44+ CD38+ CD62L+ CD45RB-) CD4 T cells (Fig. 4, A and C). Phenograph clustering of the Backbone data also created a cluster overlapping the boundary between T cells and NK cells (Fig. 4, A, D, and E) containing both CD3+ NK1− CD49b+ and CD3− NK− 1.1− CD49b− cells, while Phenograph clustering of the Infinity Flow–augmented data segregated these two cell populations more accurately into two distinct T cell and NK cell clusters (Fig. 4, F and G). In addition, the Infinity Flow–augmented UMAP positioned dendritic cell and monocyte clusters in closer proximity, in accordance with their ontogeny (movie S1). These examples illustrate how the single-cell resolution and high number of parameters of Infinity Flow’s imputed data allow a more accurate and granular delineation of the cell types present.

**Infinity Flow identifies heterogeneity within tumor-ingesting macrophages during metastatic seeding of the lung**

A key promise of high-dimensional cellular profiling is the ability to better explore and understand the complexity of cellular phenotypes in disease states. We previously identified a unique axis in the early immune response to pulmonary metastasis wherein lung macrophages encounter and ingest tumor material in the metastatic niche (27). This earlier analysis was, however, limited to a relatively narrow immune-profiling panel (table S3), which, by itself, lacked the capacity to finely delineate myeloid subsets within this early disease state. To assess Infinity Flow’s ability to profile and detect subtly distinct cell populations within the complex environment of early lung metastasis, we stained cells from the lungs of mice implanted with metastatic ZsGreen-transfected B16 melanoma cells 24 hours prior. In this system, the B16 melanoma cells express the green fluorescent protein, ZsGreen, enabling identification of phagocytes that have ingested tumor cells (Fig. 5A). Our prior studies identified a population of macrophages that critically support metastatic growth. 

![Fig. 5. Infinity Flow identifies heterogeneity within tumor-ingesting macrophages during metastatic seeding of the lung.](image)

(A) Outline of the experimental setting. (B) Color-coded expression of ZsGreen (Backbone) and MHCII, CD64, and CD26 (imputed) on a UMAP embedding of myeloid cells. (C) ZsGreen+ events from two Phenograph clusters of macrophages. (D) Bar plot representing the AUC of every imputed marker for ZsGreen+ cells from the two macrophages clusters. (E) Color-coded densities of single cells for pairs of markers, overlayed with contours of the two macrophages clusters. (F) Median fluorescence intensities of MHCII and CD11c for PD−L1− and PD−L1+ macrophages in an independent validation cytometry experiment (n = 8). (G) Representative contour plot of PD− L1− and PD− L1+ macrophages on a CD11c versus MHCII plot.
A portion of these macrophages ingest tumor material very early in the process of metastatic seeding. Consistently with our prior analysis, most of the tumor-ingesting (ZsGreen⁺) immune cells within the myeloid compartment phenotypically fit the profile of macrophages, expressing high levels of CD64, low levels of CD26, and, in this context, high levels of MHCII (Fig. 5B). Phenogroup clustering of Infinity Flow’s output revealed two distinct clusters of ZsGreen⁺ macrophages (clusters 3 and 17; Fig. 5C and fig. S11). This heterogeneity was not appreciated in our previous analysis using conventional gating (27). We next sought markers distinguishing these two clusters of tumor-ingesting macrophages. A variety of markers showed distinct expression patterns between the two clusters, and we selected three among the ones that best separated the two clusters (Fig. 5D). We found that a combination of CD11c, MHCII, and PD-L1 effectively defined each cluster, with cluster 3 being low and cluster 17 being high for each of these markers (Fig. 5E). To validate these findings, we replicated this same stain via conventional flow cytometry in a separate cohort of mice. We were able to confirm the presence of both PD – L1high and PD – L1low subsets of metastasis-infiltrating macrophages in all mice (n = 8), with a phenotype consistent with those observed in the Infinity Flow dataset (Fig. 5, F and G). Further exploration is required to ascertain the functional relevance of these two distinct populations of metastasis-infiltrating macrophages, but these validated findings demonstrate the potential for Infinity Flow in interrogating the cellular constituencies of disease states and deriving new experimental hypotheses.

DISCUSSION

In this article, we described Infinity Flow, a combined experimental and computational workflow enabling the analysis of the expression and coexpression patterns of hundreds of surface proteins across millions of single cells. The Infinity Flow computational pipeline is available as an open-source R package and can run on standard FCS files generated from MPC experiments. Others have published large-scale antibody screens using MPC experiments, but the analysis of these datasets has been limited to manual or automated analysis of the Backbone data (19–24). In contrast, Infinity Flow uses supervised machine learning (specifically multivariate nonlinear regression algorithms) to aggregate expression data from all the well-specific FCS files into a single dense matrix of imputed expression. To achieve this, we leverage the fact that the scatter and Backbone markers are measured for every cell in every well and thus allow detection of nonlinear correlations between the scatter characteristics and expression levels of the Backbone markers and the Infinity marker for that well. This information is then used to model and impute expression values for all markers on every cell in the study. Our study suggests that nonlinear machine learning models vastly outperform linear models for this task.

By imputing the expression levels of hundreds of proteins across millions of single cells, Infinity Flow enables broad phenotypic characterization of many cell subsets at the level of a whole organ. The depth of this characterization is limited by the complexity of the Backbone antibody panel. For this approach to work, it is essential that Backbone markers do inform upon the expression of the Infinity marker. We notably highlighted that the Backbone panel we used to analyze the mouse lung at steady state was unable to accurately predict expression patterns of specific TCR antibodies (e.g., Vb7 and Va3.2) despite clear staining of discrete T cell subpopulations in the Infinity antibody measurements (fig. S2). Extra Backbone antibodies could allow more depth by using antibodies specifically targeting rare cell subsets (e.g., to resolve Tγδ cells from Tαβ cells). The prediction performance of the models could help inform on which exploratory markers could best complement a given backbone or which backbone antibodies are redundant with the others and could be discarded. To produce most of the datasets analyzed in this article, we used a relatively modest LSRFortessa cytometer. State-of-the-art instruments such as spectral flow cytometers allow the resolution of an increased number of cell subsets while still increasing the overall number of parameters that can be analyzed many-fold over what is currently possible on even the most advanced of these instruments. Our results highlight the ability of Infinity Flow to deconvolve highly multiplexed acquisition channels. This was notably made possible by designing a multiplexed channel with distinct modes within positive cells. By optimizing the multiplexing of more channels, one could likely achieve high cellular resolution while using only a limited number of channels and therefore successfully use the Infinity Flow pipeline on even the most modest of cytometers. The design of the backbone panel, while critical, does not differ much from the design of a traditional antibody panel: The goal remains to maximize the cellular resolution, with the added constraint that spectral overlap with the variable channel should be minimized. The broad characterization of the mouse lung achieved here, however, highlights that Infinity Flow is well suited to deeply characterize the cellular composition of complex tissues at the protein level. In this article, we only analyzed single-sample datasets. Conventional antibody-based barcoding of multiple samples mixed in equal cellular proportions should, however, enable multisample Infinity Flow. An alternative approach to MPC is the use of multiple variable antibodies per well, as recently performed by Glass et al. (33) using mass cytometry. The computational pipeline introduced here straightforwardly extends to this type of dataset. This method and our Infinity Flow method represent complementary approaches to high-dimensional flow cytometric analysis of tissues, allowing researchers to select the method of choice based on their access to technology. The Glass et al. (33) approach is suited best to mass cytometry–based technology, as multiple complex panels need to be simultaneously optimized. In mass cytometry, spectral overlap across channels is limited. This is not the case for flow cytometry, which largely complicates the simultaneous measurement of multiple variable antibodies. Infinity Flow requires a single backbone panel that, in our experience, can be used across an array of tissues and contexts. Further, the approaches differ in throughput, as Infinity Flow can easily allow sampling on the order of tens of millions of cells, while mass cytometry–based approaches are more limited in cell throughput, although both methods offer higher throughput than sequencing–based approaches.

The critical importance of the Backbone panel for imputation highlights that any pattern captured by Infinity Flow has an underlying signature in the Backbone data. As the Backbone is an information bottleneck in MPC assays, one could intuitively expect that performing dimensionality reduction or clustering on the Infinity Flow–augmented data matrix would yield similar results to performing these analyses on the Backbone data directly. We found that the augmented data matrix enhanced our ability to identify cellular heterogeneity within individual cell clusters. We believe a key element enabling this is the use of supervised machine learning models that are capable of highlighting subtle patterns in the Backbone data as relevant for the prediction of a given Infinity antibody, a process
comparable to feature engineering for machine learning. Similarly, we have found that inclusion of FSC and SSC in the analysis has a tangible benefit to the accuracy of the algorithm as well as providing better population discrimination using UMAP. For example, the high SSC measurement characteristic of eosinophils or the variation in FSC between large and small cells in a sample can improve performance and should thus be included in analysis using Infinity Flow. These results highlight the complexity of inferences, which can be derived from multicolor flow cytometry data when treated as continuous, with interrogation of marker-marker relationships in an unbiased and nonlinear fashion.

Infinity Flow provides a flexible and robust platform for high-dimensional flow cytometry analysis; however, the method remains subject to cardinal elements of quality flow cytometry experimental design. As highlighted above, the Background panel design is crucial to the performance of the approach. This incorporates not only including appropriate markers for the scientific question and cell populations to be interrogated but also equally importantly following standard guidelines for high-quality flow cytometric panel design. These include consideration of fluorophore and marker combinations (e.g., pairing bright fluorophores with dimly expressed markers), minimizing the use of fluorophores with high spectral overlap and/or spillover spread for markers that will stain the same cells, proper antibody titration, and fluorescence compensation. Infinity Flow is relatively robust to minor issues in fluorescence compensation as these will generally be consistent across wells, and the use of a common fluorophore for Infinity Markers minimizes cross-marker inconsistency in compensation. The exception to this is considerations around spectral overlap from the Infinity Marker fluorophore into adjacent channels. While the backbone panel staining should be similar, if not identical, across wells facilitating establishment of proper cytometer settings and compensation values, the staining pattern and brightness across each well for the Infinity markers vary markedly across the entire assay. As most Infinity Flow assays use PE in the Infinity Channel, it is thus quite important that utilization of fluorophores that receive significant spillover spread from the PE channel is minimized. This will aid in not only the accuracy and consistency of compensation across wells but also the impact of spread error in discriminating between signal and noise. Fluorophores that donate spillover spread into the PE channel should also be used judiciously; however, as these patterns will be consistent across wells, it is less confounding. These features must be empirically determined by the researcher, as spillover spread characteristics are highly dependent on the instrument being used. In-depth discussion of these features of flow cytometric experimental design has recently been discussed in great detail (34), which provides an excellent resource for proper understanding and implementation of these principles. Regarding interpretation of Infinity Flow data, it is important that consideration be given to the limits of the approach. Infinity Flow yields accurate imputation of marker expression on a per-cell basis; however, patterns identified should always be empirically validated in subsequent experiments. Care must also be taken in interpreting population frequency. Proportions of individual populations in the output will generally reflect the true nature of the sample. However, particularly for rare populations, per-well sampling bias can lead to imprecise estimates of population size.

Compared to other high-dimensional single-cell protein expression approaches such as oligonucleotide-tagged antibodies (2,35) or mass cytometry (36), Infinity Flow is affordable, allows many cells to be profiled because of the high cellular throughput of conventional flow cytometry, and does not require dedicated equipment or expertise other than those used for conventional flow cytometry. The increased cellular throughput notably allows profiling of both abundant and rare cell populations in a single experiment without pre-enrichment. Issues that prevent high parameterization with other methods, such as fluorescence overlap and availability of unique fluorophores (conventional flow cytometry), availability of unique metals (mass cytometry), or available sequencing space (CITE-seq or Abseq), are therefore not a hindrance to the Infinity Flow method. There is no theoretical limit to the number of surface proteins that can be assessed. All staining of Infinity markers is done in parallel wells, using antibodies conjugated to the same common fluorophore (PE or APC, in general). The primary factors restricting the total number of parameters that can be stained are the availability of high-quality fluorescently conjugated antibodies and enough single cells to divide the initial sample across as many wells as there are Infinity markers. In general, between $10^5$ and $10^6$ cells are required per well to account for cell loss during staining while still enabling detection of rare populations. We previously demonstrated the ability for Infinity Flow to profile pre-enriched rare human circulating cDC2s (17) and here show that it can inform upon the cellular composition and phenotypes of a whole organ, including rare cell populations, without pre-enrichment. As Infinity Flow is highly scalable, samples with a relatively small number of cells can still be interrogated by limiting the number of Infinity markers to a smaller subset of interest. Conversely, the studies presented here make use of commercially available kits with several hundred preconjugated antibodies. These kits can, however, be readily augmented by adding additional wells with a custom set of Infinity antibodies best suited to the specific scientific question.

Given the wealth of information provided by the Infinity Flow pipeline and its affordability and compatibility with standard flow cytometers, we anticipate that it will be widely adopted.

METHODS

Data generation

Preparation of single-cell lung suspensions

For all experiments, single-cell digests were prepared from lung as previously described (27). Briefly, lungs were collected from fifteen 8-week-old male C57BL/6 mice following euthanasia by overdose with 2.5% Avertin. For steady-state lung analysis, mice were previously unmanipulated. For analysis of metastasis-bearing lungs, all mice were injected 24 hours prior with $5 \times 10^5$ ZsGreen-expressing B16-F10 melanoma tumor cells via the tail vein, as previously described (27). Following harvest, lungs were suspended in 5 ml of Dulbecco’s modified Eagle’s medium (GIBCO) with a combination of digestive enzymes [Liberase (0.26 U/ml; Sigma-Aldrich) and deoxyribonuclease I (0.25 mg/ml; Sigma-Aldrich)]. Samples were placed in C-Tubes (Miltenyi) and briefly processed with the GentleMACS Octomacs Tissue Dissociator (Miltenyi) using the standard lung processing protocol built into the instrument. Samples were then incubated at 37°C for 30 min with shaking at 150 rpm and processed a second time via GentleMACS. Samples were filtered around spectral overlap from the Infinity Marker fluorophore into adjacent channels. While the backbone panel staining should be similar, if not identical, across wells facilitating establishment of proper cytometer settings and compensation values, the staining pattern and brightness across each well for the Infinity markers vary markedly across the entire assay. As most Infinity Flow assays use PE in the Infinity Channel, it is thus quite important that utilization of fluorophores that receive significant spillover spread from the PE channel is minimized. This will aid in not only the accuracy and consistency of compensation across wells but also the impact of spread error in discriminating between signal and noise. Fluorophores that donate spillover spread into the PE channel should also be used judiciously; however, as these patterns will be consistent across wells, it is less confounding. These features must be empirically determined by the researcher, as spillover spread characteristics are highly dependent on the instrument being used. In-depth discussion of these features of flow cytometric experimental design has recently been discussed in great detail (34), which provides an excellent resource for proper understanding and implementation of these principles. Regarding interpretation of Infinity Flow data, it is important that consideration be given to the limits of the approach. Infinity Flow yields accurate imputation of marker expression on a per-cell basis; however, patterns identified should always be empirically validated in subsequent experiments. Care must also be taken in interpreting population frequency. Proportions of individual populations in the output will generally reflect the true nature of the sample. However, particularly for rare populations, per-well sampling bias can lead to imprecise estimates of population size.

Compared to other high-dimensional single-cell protein expression approaches such as oligonucleotide-tagged antibodies (2,35) or mass cytometry (36), Infinity Flow is affordable, allows many cells...
resuspended at $20 \times 10^6$ cells/ml in PBS + Zombie NIR Live/Dead below. Briefly, single-cell suspension of lung cells was washed and centrifuged at 300 $g$ for 5 min. Cells were resuspended at $20 \times 10^6$ cells/ml in Cell Staining Buffer, and nonspecific staining was then performed using BD Amine Reactive Beads (BD Biosciences) stained with Backbone panel antibodies. Compensation for Zombie NIR was performed independently. Singlet and live cell gating was then performed consistently for each FCS file, and these preprocessed data (well-compensated, live singlets) were exported into individual FCS files for downstream input into the Infinity Flow pipeline as described below. For basophil validation experiment (fig. S10), $2 \times 10^6$ lung cells were stained with the Basophil Validation Panel (table S3) and data were collected on a BD FACSymphony A5 cytometer.

**Computational analyses**

**The infinityFlow R package**

Infinity Flow is implemented as an R package, infinityFlow (https://www.github.com/ebcht/infinityFlow). A flow chart summarizing the key steps of the pipeline is shown in fig. S12. The input data consist of a set of FCS files, one per MPC well. These files can be uniformly downsampled in the event that computational resources are limited. The input data are then transformed using logicle transformations (37) with parameters estimated as in the flowCore R package (25). To harmonize data across input wells, each Backbone channel is linearly transformed to have 0 mean and unit variance (Z score) within each well. Nonlinear multivariate regression models are then trained and used to impute Infinity antibodies’ expression intensities as described below. UMAP nonlinear dimensionality reduction is run on the concatenated Backbone data matrix. The output matrix contains the Backbone measurements, imputed Infinity antibody expression values, UMAP coordinates, and well identifiers for each event. Another similar output matrix is produced using background-corrected signal, as described below. For each of these output matrices, a concatenated FCS file is output, as well as one FCS file per input well. Last, a UMAP plot color-coded by the measured or imputed expression of each Backbone or Infinity marker is produced, similarly to fig. S6.

**Regression models**

The InfinityFlow R package currently supports four classes of nonlinear multivariate regression models. These four methods are SVMs (13) implemented in the e1071 R library, gradient boosted trees implemented in the XGBoost R library (16), neural networks using the tensorflow and keras R packages (26), as well as polynomial models of arbitrary degrees, with or without L1-regularization (12), implemented in the glmnet R library. One regression model per algorithm and Infinity antibody is fit. The data are first randomly split into a training set (50% of the events) and a test set to evaluate performance (50% of the events). Last, the size of the output is chosen by the user and drawn from the test set for downstream analyses. For all models, the target variable is the intensity of the Infinity antibody, while the predictor variables are intensities of Backbone antibodies.

**Regression models’ hyperparameters**

The regression models’ hyperparameters can be chosen by the user. For convenience, we provide default values for each regression model that provided the best performance on a subset of the data on which we benchmarked these hyperparameters. These hyperparameter values were used throughout the manuscript. For SVM, these arguments’ values are nu-regression with a radial basis function kernel.
and nu = 0.5. For XGBoost, we used nrounds = 1500 and eta = 0.03. For neural networks, we used a fully connected neural network with one hidden layer of the same size as the input layer, rectified linear unit as activation function for nodes in the hidden and input layers, a minibatch size of 128, stochastic gradient descent for optimization with a learning rate of 0.005, mean-squared error as a loss function, 20% of the training events as a validation set, and up to 1000 epochs during training, with an early-termination criteria of 20 rounds without reaching a new minimum of the loss function in the validation set. Adding additional hidden layers or using tanh as activation function did not substantially affect performance of the NN models, while training with more examples to accommodate events used for performance evaluation and early stopping had a marginal positive impact on performance that remained lower than the one of XGBoost models (fig. S13). For LASSO models, we used either first- or third-degree polynomials, with an amount of L1-penalty automatically chosen using 10-fold cross-validation.

### Isotype-specific background staining correction

A massively parallel flow cytometry kit used here contains isotype control antibodies for each of the antibody isotypes present. Because Infinity Flow imputes every Infinity antibody, it produces imputed expression levels for both a given Infinity antibody A and its corresponding isotype control \( I \). To correct for unspecific staining, we fit the following linear model:

\[
I = \beta_0 + A \beta_1 + \epsilon,
\]

where the residuals \( \epsilon \) are used as background-corrected intensities.

### Dimensionality reduction

For dimensionality reduction, we used the UMAP algorithm (29) using its \( R \) implementation in the uwot package. UMAP was run with parameters \( n_{neighbors} = 15,\ min\_dist = 0.2, \) metric = “euclidean,” and \( n\_epochs = 2000 \). To run UMAP on imputed data, we first removed every events whose PE intensity was higher than 1/32 of the cytometer’s manufacturer-reported maximum linear range, as non-linear effects were causing compensation issues for these events (fig. S14), and applied background correction to the imputed intensities. To plot color-coded markers’ intensities on UMAP embeddings, we truncated the imputed intensities data vectors to their lower and upper 5/1000 quantiles.

### Performance metrics

To assess the performance of the regression models, we first bina-

rized each Infinity antibody’s PE intensity using manual gating. We used this binary vector and the continuous vector of imputed expres-
sion levels for both a given Infinity antibody \( A \) and its corresponding isotype control \( I \). To correct for unspecific staining, we fit the following linear model:

\[
T = \beta_0 + A \beta_1 + \epsilon,
\]

We used paired one-sided Student’s \( t \) tests to compare the performance of XGBoost to the other nonlinear algorithms benchmarked.
23. A. Uezumi, M. Nakatani, M. Ikemoto-Uezumi, N. Yamamoto, M. Morita, A. Yamaguchi,
25. F. Hahne, M. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain,
26. F. Chollet, J. J. Allaire,
30. J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E. A. D. Amir, M. D. Tadmor, O. Litvin,
31. N. Baumgarth, B-1 cell heterogeneity and the regulation of natural and antigen-induced
marker changes in human mesoderm development.

Cell Stem Cell
A. E. Corcoran, R. Walker, I. Douagi, F. Lanner, P. J. Rugg-Gunn, Comprehensive cell
analytics pipeline.
Front. Immunol.
14, 513–517 (2016).
38. D. J. Hand, R. J. Till, A simple generalisation of the area under the ROC curve for multiple
class classification problems.
Mach. Learn.
7, 44447 (2017).
39. S. C. Bendall, P. Qiú, E. A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner,
40. M. K. Quinn, H. Radbruch, T. R. D. J. Radstake, S. Rahim, H. P. Rahn, B. Rajwa,
41. G. Ravičandran, Y. Z. A. Rehbin, D. Recktenwald, D. Reimer, C. Reis e Sousa,
42. E. B. M. Remmerswaal, L. Richter, G. L. Rico, A. Röddel, A. M. Rieger, J. P. Robinson,
43. C. M. Schmidt-Weber, K. Suttner, A combined
44. A. K. Simon, K. M. Sitnik, S. Zannini, D. W. A. Spiegel, S. A. Spiegel, A. S. Spiegel,
45. J. G. Borger, B. P. Nolan, Data-driven phenotypic dissection of AML reveals progenitor-like
cells that correlate with prognosis. Cell 162, 184–197 (2015).
46. N. Baumgarth, B-1 cell heterogeneity and the regulation of natural and antigen-induced
igm production.
Front. Immunol.
7, 324 (2016).
47. K. M. Haas, B-1 lymphocytes in mice and non-human primates. Ann. N. Y. Acad. Sci.
1362, 98–109 (2015).
48. D. R. Glass, A. G. Tsai, J. P. Oliveria, F. J. Hartmann, S. C. Kimmey, A. A. Calderon, L. Borges,
49. M. C. Glass, J. C. Weg, M. M. Davis, S. C. Bendall, An integrated multi-omic single-cell
atlas of human b cell identity.
Mol. Cell. Proteomics
14, 513–517 (2016).
50. D. J. Hand, R. J. Till, A simple generalisation of the area under the ROC curve for multiple
class classification problems.
Mach. Learn.
7, 44447 (2017).
51. S. C. Bendall, P. Qiú, E. A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner,
52. M. K. Quinn, H. Radbruch, T. R. D. J. Radstake, S. Rahim, H. P. Rahn, B. Rajwa,
53. G. Ravičandran, Y. Z. A. Rehbin, D. Recktenwald, D. Reimer, C. Reis e Sousa,
54. E. B. M. Remmerswaal, L. Richter, G. L. Rico, A. Röddel, A. M. Rieger, J. P. Robinson,
55. C. M. Schmidt-Weber, K. Suttner, A combined
44. A. K. Simon, K. M. Sitnik, S. Zannini, D. W. A. Spiegel, S. A. Spiegel, A. S. Spiegel,
56. J. G. Borger, B. P. Nolan, Data-driven phenotypic dissection of AML reveals progenitor-like
cells that correlate with prognosis. Cell 162, 184–197 (2015).
57. N. Baumgarth, B-1 cell heterogeneity and the regulation of natural and antigen-induced
igm production.
Front. Immunol.
7, 324 (2016).
58. K. M. Haas, B-1 lymphocytes in mice and non-human primates. Ann. N. Y. Acad. Sci.
1362, 98–109 (2015).
59. D. R. Glass, A. G. Tsai, J. P. Oliveria, F. J. Hartmann, S. C. Kimmey, A. A. Calderon, L. Borges,
60. M. C. Glass, J. C. Weg, M. M. Davis, S. C. Bendall, An integrated multi-omic single-cell
atlas of human b cell identity.
Mol. Cell. Proteomics
14, 513–517 (2016).
61. D. J. Hand, R. J. Till, A simple generalisation of the area under the ROC curve for multiple
class classification problems.
Mach. Learn.
7, 44447 (2017).
62. S. C. Bendall, P. Qiú, E. A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner,
63. M. K. Quinn, H. Radbruch, T. R. D. J. Radstake, S. Rahim, H. P. Rahn, B. Rajwa,
64. G. Ravičandran, Y. Z. A. Rehbin, D. Recktenwald, D. Reimer, C. Reis e Sousa,
65. E. B. M. Remmerswaal, L. Richter, G. L. Rico, A. Röddel, A. M. Rieger, J. P. Robinson,
66. C. M. Schmidt-Weber, K. Suttner, A combined
44. A. K. Simon, K. M. Sitnik, S. Zannini, D. W. A. Spiegel, S. A. Spiegel, A. S. Spiegel,
67. J. G. Borger, B. P. Nolan, Data-driven phenotypic dissection of AML reveals progenitor-like
cells that correlate with prognosis. Cell 162, 184–197 (2015).
68. N. Baumgarth, B-1 cell heterogeneity and the regulation of natural and antigen-induced
igm production.
Front. Immunol.
7, 324 (2016).
69. K. M. Haas, B-1 lymphocytes in mice and non-human primates. Ann. N. Y. Acad. Sci.
1362, 98–109 (2015).
guided development of the Infinity Flow pipeline, designed experiments, provided critical analysis of data, and participated in figure preparation and manuscript drafting. R.G. guided development of the Infinity Flow pipeline, designed experiments, provided critical analysis of data, and participated in figure preparation and manuscript drafting. M.B.H. guided development of the Infinity Flow Pipeline, designed and performed experiments, performed data analysis, produced figures, and wrote the main manuscript. Competing interests: E.W.N. is a co-founder, advisor, and shareholder of ImmunoScape Pte. Ltd. and an advisor for Neogene Therapeutics and NanoString Technologies. R.G. declares ownership in CellSpace Biosciences. R.G. declares ownership in Ozette Technologies. The authors declare no other competing interests. Data and materials availability: The input, raw predictions and background-corrected predictions for the murine lung dataset at steady state are available at https://flowrepository.org/id/FR-FCM-Z2LP. The development version of the Infinity Flow software is available at https://github.com/ebecht/infinityFlow, and is submitted to the Bioconductor repository. All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 7 December 2020
Accepted 14 July 2021
Published 22 September 2021
10.1126/sciadv.abg0505

Citation: E. Becht, D. Tolstrup, C.-A. Dutertre, P. A. Morawski, D. J. Campbell, F. Ginhoux, E. W. Newell, R. Gottardo, M. B. Headley, High-throughput single-cell quantification of hundreds of proteins using conventional flow cytometry and machine learning. Sci. Adv. 7, eabg0505 (2021).