A new data-driven cell population discovery and annotation method for single-cell data, FAUST, reveals correlates of clinical response to cancer immunotherapy

Evan Greene[1,2]* egreene@fredhutch.org
Greg Finak[1,2] gfinak@fredhutch.org
Leonard A. D'Amico[1,4] ldamico@fredhutch.org
Nina Bhardwaj[7] nina.bhardwaj@mssm.edu
Candice D. Church[5] cdchurch@uw.edu
Chihiro Morishima[5] chihiro@u.washington.edu
Nirasha Ramchurren[1,4] nramchur@fredhutch.org
Janis M. Taube[6] jtaube1@jhmi.edu
Paul T. Nghiem[3,5] pnghiem@uw.edu
Martin A. Cheever[3,4] mcheever@fredhutch.org
Steven P. Fling[1,4] sfling@fredhutch.org
Raphael Gottardo[1,2]* rgottard@fredhutch.org

[1] Vaccine and Infectious Disease Division, [2] Biostatistics Bioinformatics and Epidemiology Division
[3] Clinical Research Division, [4] Cancer Immunotherapy Trials Network
Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[5] Division of Dermatology, Department of Medicine University of Washington, Seattle, WA, USA
[6] Bloomberg Kimmel Institute for Cancer Immunotherapy and the Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[7] Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai New York, NY, USA

Abstract

We introduce a non-parametric method for unbiased cell population discovery in single-cell flow and mass cytometry that annotates cell populations with biologically interpretable phenotypes through a new procedure called Full Annotation Using Shape-constrained Trees (FAUST). We used FAUST to discover novel (and validate known) cell populations associated with treatment outcome across three cancer immunotherapy clinical trials. In a Merkel cell carcinoma anti-PD-1 trial, we detected a PD-1 expressing CD8+ T cell population – undetected both by manual gating and existing computational discovery approaches – in blood at baseline that was associated with outcome and correlated with PD-1 IHC and T cell clonality in the tumor. We also validated a previously reported cellular correlate in a melanoma trial, and detected it *de novo* in two independent trials. We show that FAUST's phenotypic annotations enable cross-study data integration and multivariate analysis in the presence of heterogeneous data and diverse immunophenotyping staining panels, demonstrating FAUST is a powerful method for unbiased discovery in single-cell data.

1 Introduction

Cytometry is used throughout the biological sciences to interrogate the state of an individual's immune system at a single-cell level. Modern instruments can measure approximately thirty (via fluorescence) or forty (via mass) protein markers per individual cell [1] and increasing throughput can quantify millions of cells per sample. In typical clinical trials, multiple biological samples are measured per subject in a longitudinal design. Consequently, a single clinical trial can produce hundreds of high-dimensional samples that together contain measurements on millions of cells.

To analyze these data, cell sub-populations of interest must be identified within each sample. The manual process of identifying cell sub-populations is called "gating". An investigator gates a single sample by sequentially inspecting bi-variate scatter plots of protein expression and grouping cells with similar expression profiles together. Each sample is gated according to the same scheme,

*Corresponding author

and samples are usually compared on the basis of the frequencies of cells found within each cell sub-population.

Manual gating introduces the potential for bias into cytometry data analysis [1, 2]. One source of bias is the choice of gating strategy, since it is fixed in advance and is only one of many possible strategies to identify a cell phenotype. A different strategy can lead to different gate placements and consequently different cell counts. A more serious source of bias arises from the fact that manual gating only identifies cell populations deemed important *a-priori* by the investigator. Since the number of possible populations grows exponentially with the number of measured protein markers, manual identification cannot be used to perform unbiased discovery and analysis on high-dimensional cytometry data: there are too many combinations of markers for a single person to consider.

Researchers have developed numerous computational methods over the last decade to address manual gating's limitations [3, 4]. Many such methods [4–7] have helped scientists interrogate the immune system in a variety of clinical settings [8, 9]. Despite these successes, computational approaches to gating face significant challenges of their own when applied to large experimental datasets. Similar to manual gating, methods often require that investigators either bound or specify the number of clusters (i.e., cell sub-populations) in a sample [5, 10], or know the relevant clusters in advance [11]. This information is generally not available in the discovery context. One recommended solution is to partition a dataset into a very large number of clusters in order to capture its main structure [12]. However, as observed in [13], when methods make strong assumptions about the distribution of protein measurements [14, 15], the structure captured by over-partitioning can reflect a method's parametric assumptions rather than biological signal.

Another challenge for many methods is that biologically equivalent clusters are given different, uninformative labels when samples are analyzed independently. In such cases, methods must provide a mechanism to match clusters across samples. One matching approach is to define a metric on the space of protein measurements to enable the quantification of cluster similarities across samples [16, 17]. However, as the dimensionality of the data increases, choosing an appropriate metric becomes more difficult due to sparsity [12]. A different approach is to concatenate experimental samples together and then cluster the combined data [6, 18, 19]. This approach can mask biological signal in the presence of batch effects or large sample-to-sample

variation in protein expression. It also introduces the risk that a method will fail to identify small-but-biologically-interesting clusters, since computational limitations lead many methods to recommend sub-sampling cells from each sample before combining the samples for analysis [7].

In order to address these issues we have developed a non-parametric gating method for cytometry experiments named Full Annotation Using Shaped-constrained Trees (FAUST, Figure 1). FAUST defines cell sub-populations as modes of the joint-distribution of protein expression within each sample. Direct non-parametric estimation of the joint distribution is often computationally infeasible for cytometry data due to its dimensionality and throughput [20]. FAUST instead selects a subset of consistently well-separated protein markers using a novel *depth score*, bounds a standardized set of phenotypic regions containing modes of interest for the selected markers alone, and annotates those regions relative to data-derived annotation boundaries. By standardization, we mean that the number of regions is fixed across samples, but the location of the boundaries of those regions can vary from sample to sample. Consequently, FAUST clusters are annotated with biologically interpretable labels and each represents a cell sub-population with a homogeneous phenotype.

FAUST's standardization of phenotypic regions provides a common solution to three major challenges posed by sample- and batch-heterogeneity in cytometry experiments: cluster discovery, cluster matching, and cluster labeling. Since each discovered cluster is merely a collection of cells falling within a phenotypic region, FAUST can accommodate significant sample-to-sample heterogeneity. Similarly, since each region (and therefore each cluster) is assigned exactly one phenotypic label, the labels can be used to match clusters across samples and interpret the cell type of each cluster. An additional benefit of matching regions by phenotypic labels is robustness to sparsity since cell counts within a region can vary by orders of magnitude across samples. Here we apply the unbiased FAUST procedure to analyze data generated from three cancer immunotherapy clinical trials and demonstrate how our approach can be used to discover candidate biomarkers associated with outcome and perform cross-study analyses in the presence of heterogeneous marker panels.

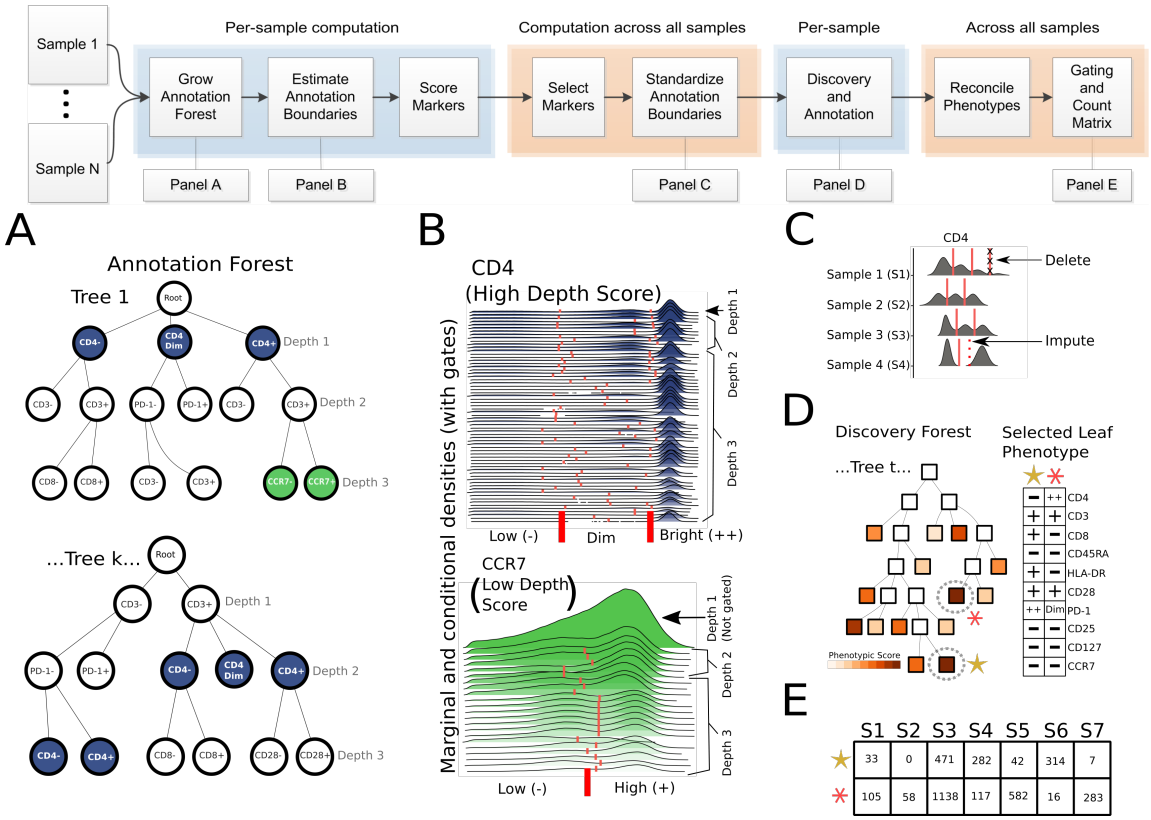

Figure 1: **Overview of FAUST.** FAUST estimates annotation boundaries for an experimental unit. An experimental unit is user defined and can be a sample, stimulation condition, subject, batch, or site. This schematic overview of FAUST assumes the experimental unit is an individual sample stained with a panel of cell markers as detected by cytometry. A) To estimate annotation boundaries, FAUST grows an exhaustive forest of 1-dimensional, depth-3 gating strategies, constrained by shape: if, prior to depth-3, the cells in a node of the gating strategy have unimodal expression along all markers, the gating strategy along that path terminates. B) Annotation boundaries are estimated for markers within an experimental unit by averaging over gates drawn for that marker over the entire annotation forest. A "depth score" (Methods 4.4) is derived for each marker and it quantifies how well-gated the marker is in each experimental unit. The distribution of scores across experimental units is used to determine whether a marker should be included in the discovery process and to determine the number of annotation boundaries a marker should receive. C) This procedure ensures that FAUST selects a standard set of markers for discovery and annotation as well as a standard number of annotation boundaries per selected marker. D) For each experimental unit, FAUST then relaxes the depth-3 constraint and conducts a search of 1-dimensional gating strategies in order to discover and select phenotypes present in the experimental unit. Each discovered phenotype is given a score that quantifies the homogeneity of cells in an experimental unit with that phenotype; high-scoring phenotypes are then selected for annotation (Methods 4.8). Each selected phenotype is annotated using all selected markers from step C), regardless of the specific gating strategy that led to the phenotype's discovery. E) FAUST returns an annotated count matrix with counts of cells in each phenotypic region discovered and selected in step D) that also survives down-selection by frequency of occurrence across experimental units.

2 Results

2.1 FAUST identifies baseline CD8+ T cells in blood that associate with outcome in CITN-09, a Merkel cell carcinoma anti-PD-1 trial

We used FAUST to perform cell sub-population discovery in cytometry data generated from peripheral blood mononuclear cells (PBMCs) isolated from patients with Merkel cell carcinoma (MCC) receiving pembrolizumab on the Cancer Immunotherapy Trials Network (CITN) phase 2 clinical trial CITN-09 [21], with the goal of identifying baseline correlates of response to treatment (NCT02267603, see supplementary table S1). We analyzed 78 longitudinal samples stained with immunophenotyping panels to identify T cell subsets within whole blood (Methods 4.10). FAUST selected 10 markers for discovery and subsequently annotated 402 discovered cell sub-populations using these markers, corresponding to 94.8% of cells in the median sample. Of these, 238 had phenotypes that included a "CD3+" annotation. Since the panel was designed to investigate T cells, only these CD3+ sub-populations were used for downstream correlates analysis.

Following [22], we used binomial generalized linear mixed models (GLMMs) to test each sub-population for differential abundance at the baseline time point (prior to receiving anti-PD-1 therapy) between responders and non-responders in 27 subjects (equation (4.5) specifies the model). We defined *responders* as subjects that exhibited either a complete (CR) or partial (PR) response (per RECIST1.1 [23]), and *non-responders* as subjects exhibiting progressive (PD) or stable (SD) disease. At an FDR-adjusted 5% level [24], four sub-populations were associated with response to therapy. Two had a CD28+ HLA-DR+ CD8+ annotation, with PD-1 dim (FDR-adjusted p-value: 0.022) or PD-1 bright (FDR-adjusted p-value: 0.030), respectively. The third had an HLA-DR- CD28+ CD4 bright PD-1 dim annotation (FDR-adjusted p-value: 0.022), while the fourth had an HLA-DR- CD28- CD4 bright PD-1 dim annotation (FDR-adjusted p-value: 0.027). The observed CD28+ phenotypes agree with published findings highlighting the importance of CD28 expression in CD8+ T cells in anti-PD1 immunotherapy [25, 26]. Effect sizes with 95% confidence intervals for the correlates are reported in Supplementary Table A.6. Three of the four correlates were annotated CD45RA- and CCR7-, indicating they represented effector-memory T cells. The complete phenotypes are described in Figure 2.

We inspected the primary flow cytometry data to confirm that the discovered population phe-

notypes matched the underlying protein expression. By plotting cluster densities against samples (Figure 2A), we observed that the FAUST annotations accurately described the observed cellular phenotypes in these sub-populations. We also visualized these data using UMAP embeddings [27] with "qualitative" parameter settings [28] (Figure 2B,C). We observed FAUST clusters were not typically separated into disjoint "islands" in the UMAP embedding (Figure 2C), and that single UMAP "islands" contained significant variation in expression of some of the measured protein markers (Figure 2B). Taken together, these observations demonstrate that visualizations derived via dimensionality reduction (here, UMAP) do not necessarily reflect all variation measured in the underlying protein data, and that any method that solely relies on UMAP for population discovery would likely miss these sub-populations.

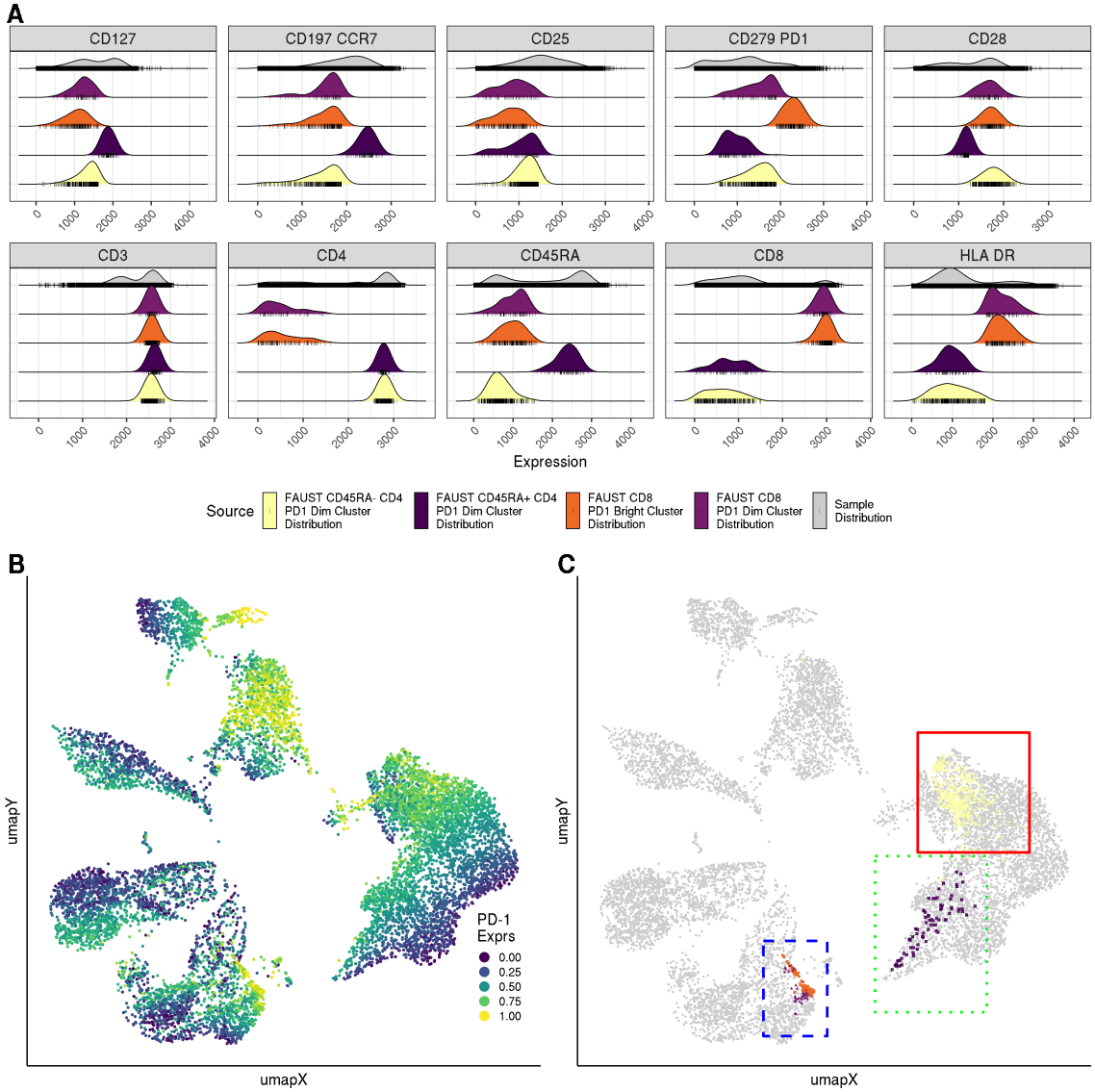

Figure 2: **FAUST annotations reflect underlying protein expression not captured by dimensionality reduction.** A) In a baseline responder's sample, the densities of per-marker fluorescence intensity for cells in the four correlates (different colors) as well as the entire collection of live lymphocytes in the sample (gray) are compared. Cells used in density calculations are marked by tick marks and demonstrate that differences in cluster annotations reflect strict expression differences in the underlying data. B) A UMAP embedding computed from the same sample as panel A using the ten stated protein markers. All cells in the sample were used to compute the embedding. The embedding is colored by the relative intensity of observed PD-1 expression, windsorized at the 1st and 99th percentile, and scaled to the unit interval. A random subset of 10,000 cells is displayed from 233,736 cells in the sample together with the complete set of 61 CD8+ PD-1 dim cells, 176 CD8+ PD-1 bright cells, 450 CD45RA- CD4 bright PD-1 dim cells, and 76 CD45RA+ CD4 bright PD-1 dim cells. C) The same UMAP embedding highlighting the location of the cells from the four discovered sub-populations. FAUST annotations are listed in depth-score order (Methods 4.4), from highest depth score to lowest. The sub-populations are annotated by FAUST as: CD4 bright CD3+ CD8- CD45RA- HLA-DR- CD28+ PD-1 dim CD25- CD127- CCR7- (yellow cells in solid red box); CD4- CD3+ CD8+ CD45RA- HLA-DR+ CD28+ PD-1 bright CD25- CD127- CCR7- (orange cells in dashed blue box); CD4- CD3+ CD8+ CD45RA- HLA-DR+ CD28+ PD-1 dim CD25- CD127- CCR7- (purple cells in dashed blue box); CD4 bright CD3+ CD8- CD45RA+ HLA-DR- CD28- PD-1 dim CD25- CD127+ CCR7+ (dark blue in dotted green box).

Reports that CD8 T cells co-expressing HLA-DR and CD28 can exhibit anti-viral properties [29], as well as reports of CD28 dependent rescue of exhausted CD8 T cells by anti-PD1 therapies in mice [26], led us to investigate the association between the abundance of the therapeutic-response-associated sub-populations discovered by FAUST and tumor viral status of each subject, as MCC is a viral-associated malignancy. We adapted the differential abundance GLMM to test for an interaction between response to therapy and tumor viral status in the four cell sub-populations discovered and annotated by FAUST. This interaction was statistically significant for both CD8+ correlates. The observed interaction p-value of 0.026 for the CD8+ PD-1 dim correlate (Figure 3A) suggested that these T cells may be particularly relevant in subjects with virus-positive tumors. In order to further investigate the relevance of these T cells measured in blood, we examined published data on PD-1 immunohistochemistry (IHC) staining in tumor biopsies from the same patients (described in [30]). Importantly, the in-tumor PD-1 measurement is a known outcome correlate in MCC [30]. Limited overlap between the assays resulted in only five subjects where both flow cytometry and tumor biopsy anti-PD-1 IHC staining were available, and only four of these were virus-positive. Nonetheless, the frequencies of the CD8+ PD-1 dim T cells were strongly correlated ($\rho = 0.945$) with the PD-1 total IHC measurements within the four virus-positive subjects (Figure 3B).

We also examined published TCR clonality data generated from patient tumor samples, described in [31]. Ten subjects passing clonality QC were common to the two datasets, six of which were virus positive. Frequencies of the FAUST populations within these six subjects were strongly correlated ($\rho = 0.952$) with the measurement of productive clonality (Figure 3C). Normalizing the correlate cell counts by the total number of CD3+ annotated FAUST sub-populations (i.e., total T cells, the recommended normalization constant for T cell clonality) instead of total lymphocyte count produced an observed correlation of $\rho = 0.972$ (Supplementary Figure S1). Together, these results led us to hypothesize that the CD8+ T cell correlate discovered by FAUST in blood represents a circulating population of tumor-associated virus-specific T cells that are also detectable in the tumor and whose presence in the tumor is known to correlate with outcome. Due to the small sample size, this hypothesis must be confirmed on an independent, larger set of patient samples. However, our results demonstrate that FAUST discovers and annotates cell sub-populations that are immunologically plausible, suggest a testable hypothesis for follow-up

experimentation, and potentially have clinical utility.

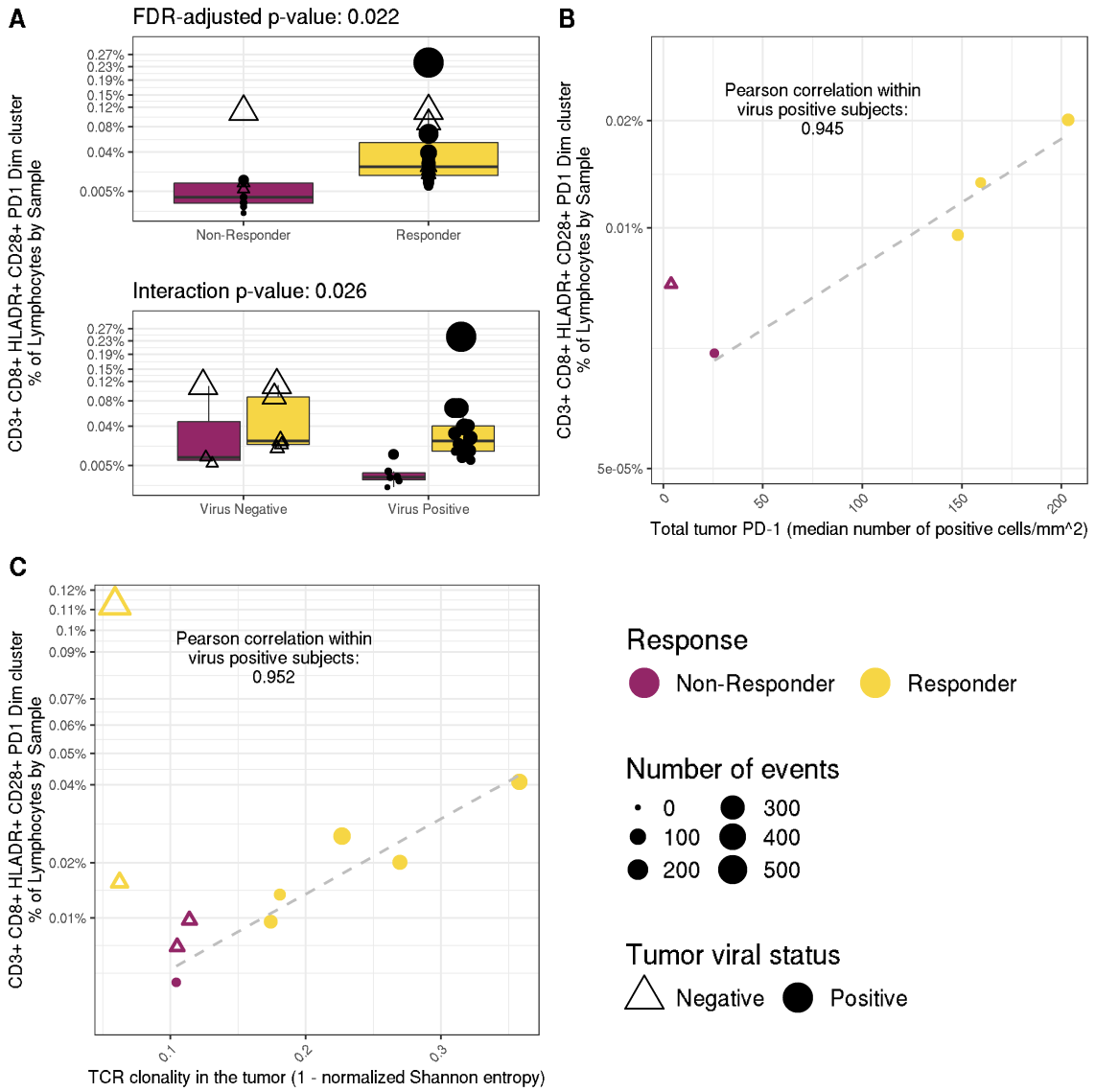

Figure 3: **A CD8+ PD-1 dim CD28+ HLA-DR+ T cell sub-population discovered and annotated by FAUST is associated with outcome in virus positive subjects and with independent measurements of PD-1 and T cell clonality in the tumor.** A) Boxplots of the abundance of the CD8+ PD-1 dim CD28+ HLA-DR+ T cell outcome correlate discovered by FAUST, stratified by subjects' response to therapy (FDR adjusted p-value contrasting all responders (n = 18) vs all non-responders (n = 9): 0.022) and their viral status (unadjusted p-value of interaction: 0.026). B) The abundance of the CD8+ PD-1 dim CD28+ HLA-DR+ T cell correlate among virus positive subjects against total PD-1 expression measured by IHC from tumor biopsies as described in [30], with observed correlation in virus positive subjects (n=4) of 0.942. C) The abundance of the CD8+ PD-1 dim CD28+ HLA-DR+ T cell correlate among virus positive subjects plotted against productive clonality (1- normalized entropy) from tumor samples as described in [31], with observed correlation in virus positive subjects (n=6) of 0.959. Supplementary Figure S2 displays the remaining correlates.

2.2 FAUST sub-populations capture underlying biological and technical signals in longitudinal studies

Consistently identifying and annotating cell populations that are missing across a subset of samples is a significant challenge in computational cytometry analysis [32]. To demonstrate how FAUST's phenotypic standardization can address this issue we examined the longitudinal profiles of specific cell sub-populations in the MCC anti-PD-1 trial for which we expected longitudinal changes in the abundance of these populations due to known technical effects. In the MCC anti-PD-1 trial, we examined all CD8+ T cells with the PD-1-bright phenotype. The temporal abundance of these cells is shown in (Figure 4A) and reveals that these cells are not detectable in most samples after subjects have received pembrolizumab therapy, presumably from pembrolizumab blocking the detecting antibody. This is consistent with the expected behavior of anti-PD-1 as observed in other trials run within the CITN (data not shown).

We also analyzed flow cytometry data from a second CITN trial: CITN-07 (NCT02129075, see supplementary table S1 for trial data), a randomized phase II trial studying immune responses against a DEC-205/NY-ESO-1 fusion protein (CDX-1401) and a neoantigen-based melanoma vaccine plus poly-ICLC when delivered with or without recombinant FLT3 ligand (CDX-301) in treating patient with stage IIB to stage IV melanoma. The cytometry data consisted of fresh whole blood stained for myeloid cell phenotyping (Methods 4.12). Here, FAUST discovered and annotated 132 cell sub-populations using 10 markers (selected by depth-score), assigning phenotypic labels to 93.2% of cells in the median sample.

In the FLT3-Ligand + therapeutic Vx trial we expected to observe expansion of dendritic cells in response to FLT3-L stimulation [33]. Examination of the longitudinal profile of clusters with phenotypic annotations consistent with dendritic cells (Figure 4B) revealed dynamic expansion and contraction of the total DC compartment in the FLT3-L stimulated cohort but not in the unstimulated-by-FLT3-L-pre-treatment cohort. The expansion peaked at day 8 after FLT3-L simulation in cycles 1 and 2. This dynamic is consistent with observations from manual gating of the DC population [34], the expected biological effect of FLT3-L [33], and the timing of FLT3 administration.

These results demonstrate that FAUST is able to detect, annotate, and correctly assign abundance to cell sub-populations, including those that are missing in some samples. The longitudinal

behavior of PD-1 bright T cell populations in the MCC anti-PD-1 trial and the dendritic cells in the FLT3 ligand + CDX-1401 trial are consistent with manual gating of cytometry data and serve as an internal validation of the methodology.

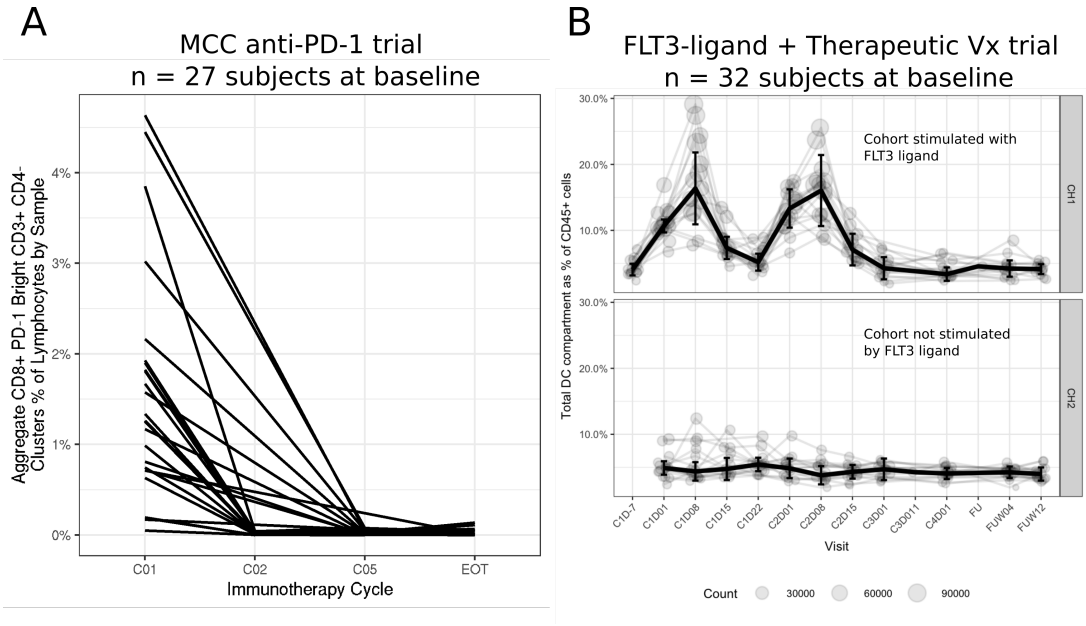

Figure 4: **The longitudinal profiles of aggregated FAUST cell populations in a pembrolizumab therapy trial and a FLT3-L + CDX-1401 trial.** A) The aggregated frequency of all CD8+ PD-1-bright T-cell populations found by FAUST across all time points. B) The longitudinal profiles of all cell sub-populations with phenotypes consistent with the DC compartment: CD19-, CD3-, CD56-, HLA-DR+, CD14- CD16- and CD11C+/-. Light colored lines show individual subjects. The dark line shows the median across subjects over time. Error bars show the 95% confidence intervals of median estimate at each time point. Cohort 1 (n=16 subjects), cohort 2 (n=16 subjects).

2.3 FAUST identifies phenotypically similar myeloid sub-populations associated with clinical response across multiple cancer immunotherapy trials

Both the MCC anti-PD-1 and FLT3-L + therapeutic Vx trials had cytometry datasets stained with a myeloid phenotyping panel. We selected two additional myeloid phenotyping datasets (one CyTOF discovery and one FACS validation assay) from a previously-published anti-PD-1 trial in metastatic melanoma [8]. We will refer to these as the melanoma anti-PD-1 FACS and melanoma anti-PD-1 CyTOF datasets. In each study, a different staining panel was used to interrogate the myeloid compartment. Details of the FAUST analysis of these data are provided in Methods 4.

A principal finding of the published analysis of the melanoma anti-PD-1 trial was that the

frequency of CD14+ CD16- HLA-DR^{hi} cells was associated with response to therapy. In all four datasets FAUST identified cell sub-populations associated with clinical outcome at baseline (FDR-adjusted 5% level, using binomial GLMMs to test for differential abundance) whose phenotype was consistent with the previously-published CD14+ CD16- HLA-DR^{hi} phenotype (Figure 5A-D). Complete phenotypes, effect sizes and confidence intervals for the myeloid baseline predictors discovered in the MCC anti-PD-1 myeloid phenotyping data are in Supplementary Table S2; those discovered in the FLT3-L + therapeutic Vx trial are in Supplementary Table S3. These results demonstrate the power of our approach to detect candidate biomarkers in a robust manner across different platforms, staining panels, and experimental designs.

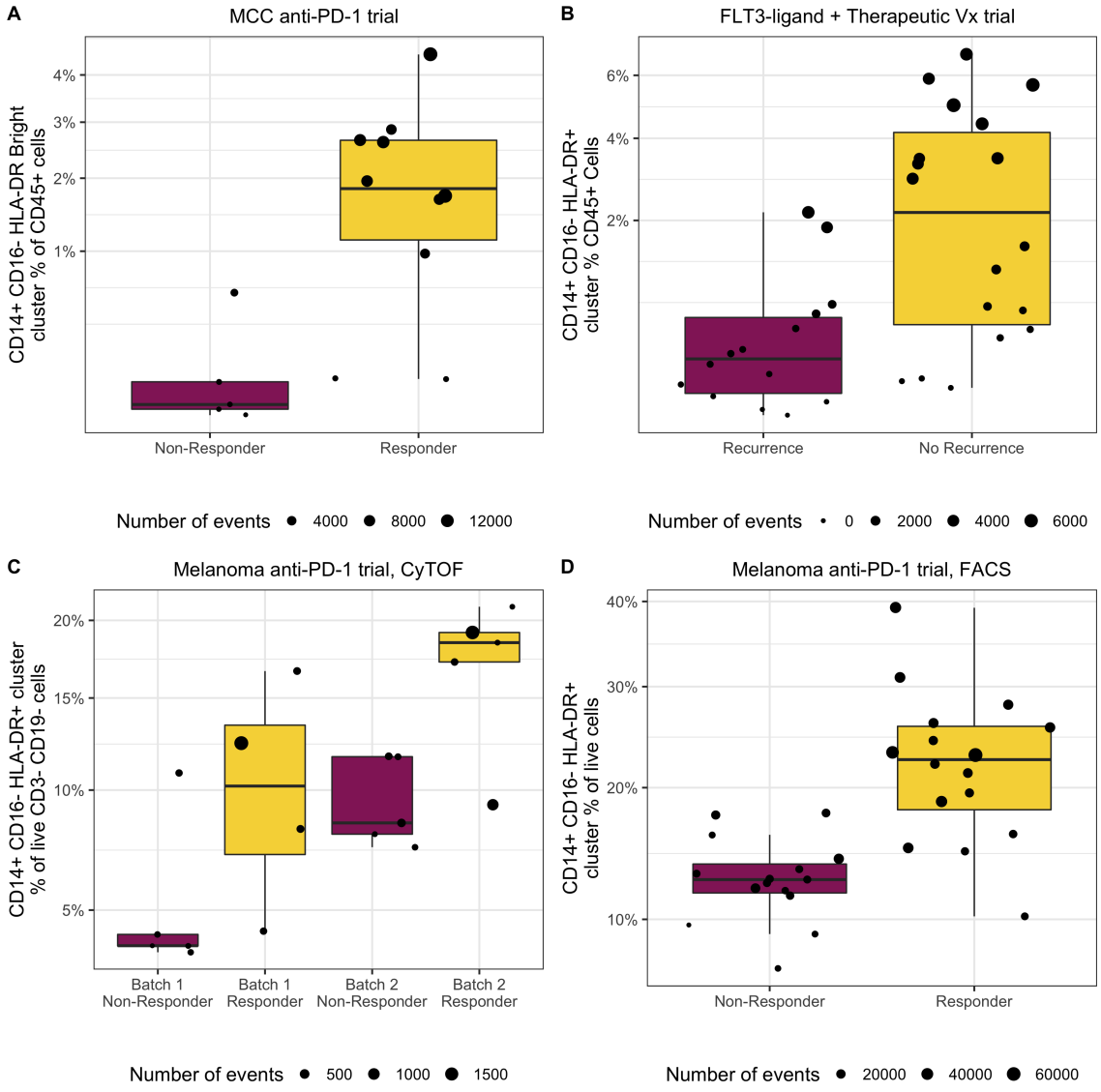

Figure 5: **FAUST consistently discovers CD14+CD16-HLADR+ monocytes associated with outcome at baseline across immunotherapy trials.** A) The baseline outcome-associated sub-population discovered by FAUST in the MCC anti-PD-1 trial myeloid data (n=15, 10 Responders, 5 Non-Responders). The full FAUST annotation for the sub-population was CD33 bright CD16- CD15- HLA-DR bright CD14+ CD3- CD11B+ CD20- CD19- CD56- CD11C+. B) The baseline outcome-associated sub-population discovered by FAUST in the FLT3-L therapeutic Vx trial myeloid data (n=32, 18 No Recurrence, 14 Recurrence). The full FAUST annotation for the sub-population was CD8- CD3- HLA-DR+ CD4- CD19- CD14+ CD11C+ CD123- CD16- CD56-. C) The baseline outcome-associated sub-population found by FAUST from the re-analysis of the Krieg CyTOF panel 03 (stratified by batch) (n=19, 10 Responder, 9 Non-Responder). The full FAUST annotation for the sub-population was CD16- CD14+ CD11B+ CD11C+ ICAM1+ CD62L- CD33+ PDL1+ CD7- CD56- HLA-DR+. D) The baseline outcome-associated sub-population found by FAUST from the re-analysis of the Krieg FACS validation data (n=31, 16 Responder, 15 Non-Responder). The full FAUST annotation for the sub-population was CD3- CD4+ HLA-DR+ CD19- CD14+ CD11B+ CD56- CD16- CD45RO+.

2.4 FAUST enables cross-study comparisons between different marker panels

FAUST annotations make it possible to test hypotheses involving prior biological knowledge of hierarchical relationships among cell types. By jointly modeling those annotated populations related through biological hierarchy, we are able to account for their dependence structure when conducting secondary tests of interests. This is analogous to the techniques used to perform gene set enrichment analysis in gene expression data [35]. We contrast this approach against aggregating (i.e., summing) cell sub-population counts on the basis of their common annotations to derive ancestral populations that resemble those obtained by manual gating, which we hypothesized can obscure interesting signals in the data.

To demonstrate this we tested each of four different myeloid sub-compartments for association with outcome at baseline in each of the three trials which used heterogeneous marker panels. We used the FAUST annotations to define membership in the myeloid compartment (described below), excluding the Krieg CyTOF dataset since 10 of 19 baseline samples had fewer than 1500 total cells. All FAUST sub-populations that were annotated as lineage negative (CD3-, CD56-, CD19-) and expressing HLA-DR (either dim or bright) were selected as part of the myeloid compartment. We further defined myeloid sub-compartments in terms of a sub-population's CD14 and CD16 expression, with CD14- CD16- cells defined as dendritic cells, and other combinations as double-positive, CD14+, or CD16+ monocytes, respectively.

We fit two models to each dataset. First, a multivariate model of all candidate cell sub-populations was fit (Methods 4.15), and the cell sub-populations' model coefficients were aggregated over each sub-compartment to test for increased abundance in responders vs. non-responders at baseline. This model represents the cell population analog of gene set enrichment analysis. Second, a univariate model was fit to cell counts derived by summing over each myeloid sub-compartment (Methods 4.16), producing a single coefficient to test for increased abundance in responders vs. non-responders at baseline. This represents the modeling approach one would undertake if the myeloid sub-compartments were defined using a manual gating strategy. One-sided 99% confidence intervals (Bonferroni-adjusted 95% CIs) were computed for all tests.

Using the aggregate model, we only observed significantly increased abundance of the CD14+CD16- sub-compartment among responders (Figure 6A) in the melanoma anti-PD-1 trial FACS dataset, a finding consistent with the authors' validation analysis [8]. We did not observe

significantly increased abundance in either CITN trial dataset using the aggregate model. However, using the multivariate model, we observed significantly increased abundance in the CD14+CD16- monocyte sub-compartment across all datasets (Figure 6A).

These results suggested that sub-populations defined by manual gating may not exhibit significant differential abundance when they don't capture all the heterogeneity in a cell population measured in the dataset. To test this, we used the binomial model (Methods 4.12) to model all cell population counts derived in 32 baseline samples from the CD45+ sub-populations defined by manual gating in CITN-07 (Supplementary Table S7), and did not detect an association between the CD14, mDC, or pDC sub-populations and non-recurrence at the FDR-adjusted 5% level. Similarly, we did not identify statistically significant correlates of outcome in the MCC anti-PD-1 trial when we fit our binomial model (Methods 4.10) to counts derived from populations identified by the manual gating strategy in the 27 baseline T cell samples (Supplementary Table S5).

In contrast, the multivariate model also detected a significant association between outcome and increased abundance in the CD14-CD16- dendritic cell sub-compartment (Figure 6B) in the two CITN trials, consistent with our analysis of baseline predictors in those trials. We did not detect such an association in the DC sub-compartment in the Melanoma anti-PD1 trial. Since both the CITN trials used fresh blood samples for analysis while the latter used frozen PBMC samples [8], we hypothesize the observed differences in modeling outcomes is due to cryopreservation status, a hypothesis supported by studies [36, 37] that examine the differential effect of cryopreservation on monocytes and DCs, respectively. This multivariate modeling approach demonstrates how FAUST can enable cross-study data integration and analysis even in the presence of heterogeneous staining panels.

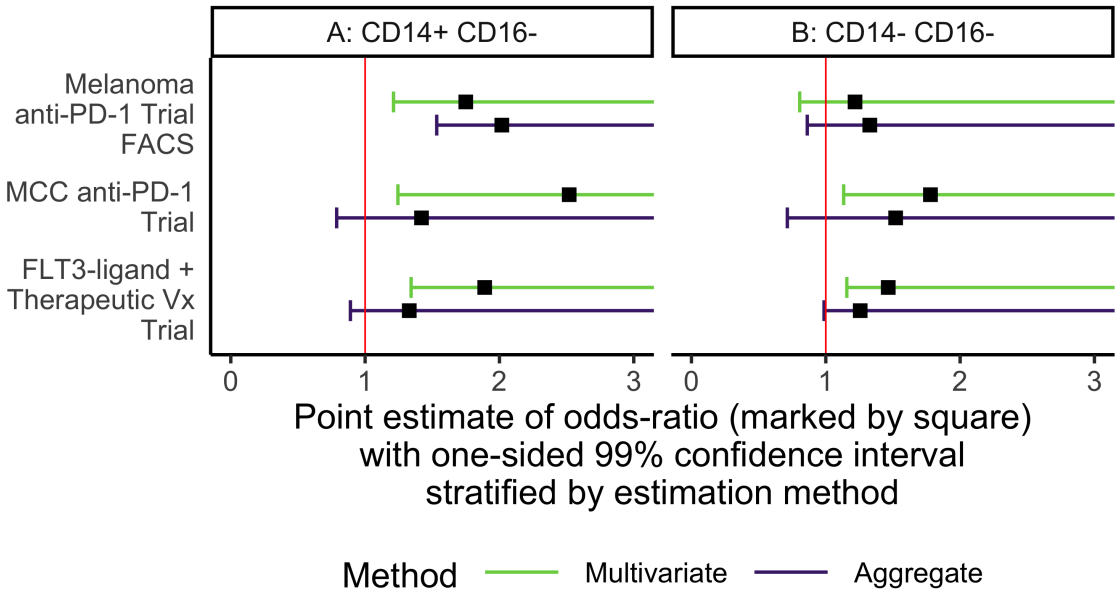

Figure 6: **Standardized annotation of clusters enables cross-study meta-analysis of datasets stained with disparate marker panels.** Differential abundance between responders and non-responders across the different sub-compartments was tested by aggregating model coefficients (analogous to meta-analysis over cell sub-populations in a sub-compartment) from a multivariate GLMM and by univariate modeling of aggregated cell counts. One-sided, 99% (Bonferroni-adjusted) confidence intervals for increased abundance in responders vs. non-responders are displayed for each sub-compartment in each dataset. In all modeling scenarios, when the whisker of a forest-plot line crosses the vertical red-line at 1, this indicates the increased odds in the responders vs non-responders are not statistically significant at the Bonferonni-adjusted level. A) Cells in the CD14+ CD16- HLA-DR+ sub-compartment were found to be significantly more abundant in responders than non-responders in all datasets tested using the multivariate modeling approach. In the univariate modeling of aggregate cell counts, the CD14+ CD16- HLA-DR+ sub-compartment was only significant in the melanoma anti-PD-1 FACS dataset, consistent with the authors' published findings. The x-axis can be interpreted as the odds increase in the probability of observing more cells in the responders than the non-responders in the compartment. B) Cells in the CD14- CD16- HLA-DR+ sub-compartment were found to be significantly more abundant in responders than non-responders in the two CITN datasets tested using the multivariate modeling approach. We hypothesize that the observed difference between the CITN trials and the melanoma anti-PD-1 trial is explained by cryopreservation in the latter trial, since it has been reported that cryopreservation affects the relative abundance of pDCs and mDCs [37], but does not affect monocyte function [36]. See supplementary Information A.12 for results from the other compartments.

2.5 FAUST is robust to different data generating processes

We re-analyzed the MCC anti-PD-1 T cell dataset described in section 2.1 with the clustering methods *densityCut* [38], *FlowSOM* [5], *Phenograph* [39], and FAUST. For all non-FAUST methods, we set tuning parameters to the settings reported in [4] when possible. Among all compared methods, FAUST was the only method to discover baseline T cell subsets associated with response to therapy at the FDR-adjusted 5% level (Supplementary Section A.7).

We also conducted a simulation study that simulated the discovery process in cytometry data analysis by inducing a differentially abundant population associated with a simulated response to therapy, in datasets generated from a variety of mixture models (Supplementary Section A.13). We compared FAUST to FlowSOM in this study since FlowSOM is computationally efficient, is recommended in the review [4], and is used in the *diffCyt* method [18]. Across simulation settings, we found that FAUST consistently performed the discovery task well, while FlowSOM's discovery performance was adversely affected by departures from normality combined with simulated batch effects and nuisance variables.This study confirms our empirical finding that FAUST robustly detects signals in data that are not found by other discovery methods.

3 Discussion

We applied FAUST to five datasets (CyTOF and flow) from three independent immunotherapy trials. Across these trials, FAUST discovered cell sub-populations and labeled them with annotations that are generally consistent with previous manual gating of the cytometry data (when aggregated by appropriate annotation) as well as with the known biological context, strongly supporting this novel unbiased approach.

We found FAUST discovered cell populations associated with clinical outcome in the analyzed datasets that are missed by other methods. Notably, manual gating did not identify statistically significant correlates of outcome in the MCC anti-PD-1 baseline T cell data. The multivariate analyses (Section 2.4) found that only some fully-annotated sub-populations exhibit differential abundance (captured by the individual model coefficients), differences that can be obscured when the cell counts are aggregated for a single test. Since aggregation produces clusters that are similar to those obtained by standard manual gating, the aggregate models suggest two

ways that manual analysis can fail to uncover signal present in a dataset. Manual analysis may not gate out sub-populations that differ between conditions (due to bias), or may incompletely describe the heterogeneity of protein expression in the gated cell populations. Both occurred in the CITN datasets: in the MCC anti-PD-1 trial, neither HLA-DR+CD14+CD16- monocytes nor HLA-DR+CD14-CD16- DCs were manually gated (Supplementary Table S6), since the manual gating strategy was designed to interrogate MDSCs in the subjects. In the FLT3-L trial, CD14+, mDCs, and pDCs were manually gated, but a differential signal at the FDR-adjusted 5% level was not detected at baseline.

In contrast, the unbiased approach taken by FAUST leads it to conduct an exhaustive search of (methodologically constrained) gating strategies in order to estimate the location of annotation boundaries for markers in each sample. When FAUST is applied to a heterogeneous population of cells (e.g. live lymphocytes which contain T cells, B cells, monocytes, etc.), this means FAUST uses information from many different cell types to estimate sample-specific annotation boundaries for each marker it selects. FAUST goes on to use these boundaries to annotate the sub-populations it subsequently discovers in each sample. In consequence, FAUST produces annotations that describe the protein expression of each discovered sub-population relative to the starting population of cells in the sample, and differs in kind from the standard paradigm of following a path in a single gating strategy to arrive at a phenotype. We hypothesize it is these methodological characteristics – as well as its pervasive use of non-parametric statistical methods – that explain FAUST's discovery performance relative to manual gating on the analyzed datasets.

The sub-populations discovered by FAUST are consistent with their immunological context and recent literature. The PD-1 dim CD28+ T cell sub-population identified in the MCC anti-PD-1 trial may represent virus specific T cells as evidenced by their correlation with T cell clonality measurements from the tumor biopsy (Figure 3C). This further accords with literature that highlights the role of CD28 in anti-PD-1 immunotherapy, which reports CD28 signaling disrupted by PD-1 impairs T cell function [25]. It has also been reported that, following PD-1 blockade, CD28 is necessary for CD8 T cell proliferation [26]. The sub-populations are also consistent with reports that certain PD-1^{int} $CD8^{+}$ T cells are responsible for viral control in mice [40] after PD-1 blockade [41]. Taken together with our findings, the PD-1 dim CD28+ T cell sub-population may have prognostic value in MCC subjects with virus-positive tumors, though we emphasize this

hypothesis requires further validation in an independent cohort. Supporting this assertion is the surprisingly strong correlation between the T cell frequency and anti-PD-1 IHC measured from the tumor where the latter is a known prognostic marker. Although this evidence is tempered by small sample size, its strength warrants further investigation. The consistent detection of myeloid sub-populations with a CD14+CD16-HLA-DR+ phenotype across four different datasets from three independent trials spanning different cancer types and therapies strongly suggests that FAUST is detecting real biological signals in the analyzed datasets.

As with any computational method, FAUST has tuning parameters that need to be adjusted to analyze real experimental data. These parameters are described in section 4.9, and in our view are uniquely interpretable among computational methods since they affect how FAUST processes 1-dimensional density estimates. Our results demonstrate that FAUST can consistently detect immunologically-plausible candidate biomarkers from measurements made in blood using a simple, well-understood assay. Many large experimental flow cytometry datasets already exist, and FAUST has the potential for the productive re-analysis and meta-analysis of such data.

4 Methods

4.1 FAUST method: underlying statistical model

FAUST assumes the following criteria are met in a cytometry experiment consisting of n experimental units E_i, $1 \leq i \leq n$.

Assumption 1. *Each sample in the cytometry experiment has been compensated (as needed) as well as pre-gated to remove debris and dead cells.*

If pre-gating has not been performed by an investigator, computational methods [42, 43] can be used before applying FAUST to cytometry data in order to guarantee this assumption is met.

Assumption 2. *In each sample, measurements on the live cells are made using a common set of p transformed protein markers.*

Let n_i denote the number of events in the ith experimental unit. FAUST supposes each event $E_{i,j}$ in an experimental unit E_i, of dimension p (the number of markers), arises as a sample from a

finite mixture model

$$E_{i,j} \sim \sum_{m=1}^{M} \omega_m \cdot f_{m,i}(\mathbf{x}) \ , \tag{4.1}$$

for $1 \leq j \leq n_i$, with $M \in \mathbb{N}$, $0 \leq \omega_m \leq 1$ and $\sum_{m=1}^{M} \omega_m = 1$ for all $1 \leq i \leq n$. FAUST assumes the mixture components $f_{m,i}$ of an experimental unit in (4.1) belong to the class of densities on the space of protein measurements

$$\mathcal{F}_i \equiv \left\{ f_{m,i} \,\middle|\, \exists\, \lambda_{m,i} \in \mathbb{R}^k,\ \sigma_{m,i} \in \mathbb{R} \text{ such that } \frac{f_{m,i} + \lambda_{m,i}}{\sigma_{m,i}} \in \mathcal{F}\ \forall\, 1 \leq m \leq M \right\} \tag{4.2}$$

for each experimental unit i, with the common class $\mathcal{F}$ is defined as

$$\mathcal{F} \equiv \{ f_m \,|\, f \text{ is unimodal along all margins} \} \ . \tag{4.3}$$

(4.2) expresses the fundamental modeling assumption: each mixture (4.1) that generates an experimental unit consists of a common set of densities (4.3), with unit-specific changes to location (the translations $\lambda_{m,i}$) and scale (the scalar multiples $\sigma_{m,i}$) of the component densities. These unit-specific modifications represent technical and biological effects. We emphasize that we only assume marginal unimodality for the f in (4.3), but make no assumptions about the joint-distribution of these densities.

4.2 FAUST method: overview

FAUST is designed to perform independent approximate modal clustering of each mixture (4.1) in each experimental unit. Its approximation strategy is to use 1-dimensional densities to grow an exhaustive forest of gating strategies (section 4.3), from which it estimates a standardized set of annotation boundaries for all markers in a mixture, which exhibit 1-dimensional multimodality either marginally or across a large number of conditional 1-dimensional density estimates. Annotation boundaries are estimated (section 4.5) by taking a weighted average of marginal and conditional 1-dimensional antimodes for a marker that FAUST selects, using a score (section 4.4) that quantifies if the marker has persistent multimodality in the experimental unit. FAUST also uses the distribution of the depth score across units to select a subset of markers to use for cluster

discovery and annotation (section 4.6).

FAUST defines a cluster as a subset of events in an experimental unit that fall inside either a conical or hyper-rectangular region bounded by the Cartesian product of the standardized set of annotation boundaries. FAUST discovers cluster phenotypes by growing a forest of partition trees for each experimental unit (trees are grown at random, following a strategy related to growing the annotation forest), and locating a sub-collection of homogeneous leaf nodes in the forest relative to the standardized phenotypic boundaries (section 4.8). FAUST collects a list of phenotypes discovered in each experimental unit and counts how often each phenotype appears across the set of lists. If a phenotype exceeds a user-specified filtering threshold, FAUST will annotate that cluster in each experimental unit relative to the standardized annotation boundaries. Intuitively, each annotation is a pointer to a modal region of each experimental unit's mixture distribution. FAUST concludes by deriving a count matrix, with each row corresponding to a sample in the experiment, each column an annotated cluster, and each entry the cell count corresponding to the annotated cluster in the sample.

4.3 FAUST method: growing the annotation forest

For all markers in a sample, all cells for each marker are tested for unimodality using the dip test [44]. The hypothesis of unimodality is rejected for any marker that has dip test p-values below 0.25. All markers which are deemed multimodal according to this dip criterion are then used to start gating strategies. Gate locations for each strategy are determined using the taut string density estimator [45]. The location of each gate is the mid-point of any anti-modal component of the taut string. Since the taut string makes no assumptions about the number of modes present in a density, in principle this approach can lead to estimating an arbitrary number of gates in a given strategy. In practice, we only pursue strategies containing 4 or fewer gates under the assumption that marker expression of 5 expression categories does not reflect biological signal.

Once the initial set of gates are computed for a given marker, events are divided into sub-collections relative to the gates for that marker and the procedure recurses and repeats along each sub-collection. Algorithm 1 gives an overview of the procedure. A gating strategy terminates when it meets any of the following stopping conditions. First, once a strategy involves any three combinations of markers, it terminates. This is because the space of gating strategies grows

factorially with the number of markers. Due to this growth rate, nodes in the forest are penalized factorially relative to their depth in the gating strategy when we subsequently compute the depth score. Second, if at any point in a strategy FAUST fails to reject the null hypothesis of unimodality for all tested markers, the strategy terminates regardless of depth. Finally, a gating strategy terminates along a branch if all nodes along the branch contain too few cells. The algorithm displayed here assumes event measurements are distinct in the cytometry dataset, and all nodes in the forest contain in excess of 500 events. For details of how FAUST breaks ties and deals with nodes containing between 25 and 500 events, we refer the reader to [46].

Algorithm 1 Grow Annotation Forest

```
function growAnnotationForest(currentCells, currentDepth, activeMarkers)
    if (length(currentCells) < 500) or (currentDepth > 3) then
        return strategy                    ▷ Gating strategy stops due to depth, event constraints.
    else
        currentDepth ← currentDepth + 1
        multimodalList ← empty list
        for markerIndex ∈ (columns(expressionMatrix) ∩ activeMarkers) do
            if pValue(dipTest(expressionMatrix[currentCells,markerIndex])) < 0.25 then
                append(multimodalList, markerIndex)
        if length(multimodalList) == 0 then
            return strategy                ▷ Gating strategy stops due to shape constraint.
        else
            for markerIndex in multimodalList do
                boundaryList ← empty list
                Compute taut string density estimate of expressionMatrix[currentCells,markerIndex]
                boundaryList ← mid-points of antimodal components of taut string
                remainingMarkers ← activeMarkers \ markerIndex
                for i in [1,length(boundaryList)] do
                    lg ← boundaryList[(i-1)]
                    ug ← boundaryList[i]
                    newCells ← rows of expressionMatrix[currentCells,markerIndex] between lg and ug
                    growAnnotationForest(newCells, currentDepth,remainingMarkers)
```

4.4 FAUST method: depth score computation

Suppose there are $p > 1$ active markers in a sample. To compute the depth score for any of the p markers, the annotation forest is first examined to determine the following quantities: d_1, the number of times different markers are gated in the root population; d_2, the number of times children of the root are gated; and d_3 the number of times grandchildren of the root are gated. For

$i \in \{1, 2, 3\}$ define

$$\delta_i \equiv \frac{1}{d_i} \ .$$

For $1 \leq m \leq p$, let

$$\mathcal{N}_m \equiv \{N_{m,1}, N_{m,2}, \ldots, N_{m,n}\}$$

be the set of all n parent nodes in the annotation forest for which the null hypothesis of unimodality is rejected for marker m. For a parent node $1 \leq j \leq n$, let 1_R denote the indicator function that is 1 when $N_{m,j}$ is the root population. Similarly, let 1_C denote an indicator of a child of the root, and 1_G a grandchild of the root. Define the scoring function

$$Q(N_{m,j}) \equiv (1-\alpha_R)1_R(N_{m,j}) + (1-\alpha_R)(1-\alpha_C)1_C(N_{m,j}) + (1-\alpha_R)(1-\alpha_C)(1-\alpha_G)1_G(N_{m,j}) \ ,$$

where, abusing notation, we let

$$\alpha_R \equiv \alpha_R(N_{m,j}) \equiv \text{the dip test p-value in the root population of the gating strategy that led to } N_{m,j} \ .$$

We allow α_C and α_G to be defined similarly. The function Q can be interpreted as a measure of the quality of the gating strategy that led to node $N_{m,j}$. In the case of a grandchild node that had clear modal separation along all markers in the strategy, $Q(N_{m,j}) \approx 1$, while a grandchild node that had p-values of 0.25 at each ancestral node, $Q(N_{m,j}) \approx 27/64 = 0.75^3$.

Let $\mathcal{P}_m$ be the population size for marker m in the root population. Next define

$$P(N_{m,j}) \equiv \frac{\text{\# of cells in node } N_{m,j}}{\mathcal{P}_m} \ .$$

Finally, define

$$D(N_{m,j}) \equiv \delta_1 \cdot 1_R(N_{m,j}) + \delta_2 \cdot 1_C(N_{m,j}) + \delta_3 \cdot 1_G(N_{m,j}) \ .$$

The depth score is defined to be the normalized sum

$$DS(\mathcal{N}_m) \equiv \frac{\sum_{i=1}^{n} Q(N_{m,i}) \cdot P(N_{m,i}) \cdot D(N_{m,j})}{\max_{1 \leq q \leq p} DS(\mathcal{N}_q)} \equiv \frac{\sum_{i=1}^{n} \omega(N_{m,i})}{\max_{1 \leq q \leq p} DS(\mathcal{N}_q)} \ . \tag{4.4}$$

The depth score maps $\mathcal{N}_m$ into $[0, 1]$, with at least one marker in a gated sample achieving

the maximal score of 1. This is taken as a measure of separation quality: the best scoring marker according to the depth score is taken to be the best separated marker in that sample at the root population, and conditionally along all other gating strategies. Normalizing to the unit interval allows depth scores to be compared across experimental units for given markers. By using the factorial weights δ_i, the depth score also explains why FAUST only explores gating strategies involving, at most, combinations of three markers in its scoring and marker selection phase. Adding more combinations of markers induces a factorial increase in computational cost. But any marker that enters a gating strategy at depth 4 (or beyond) will be dominated in depth score by those markers initially gated in the annotation forest at or near the root population. Consequently, after normalization in experiments with a large number of markers, such markers have depth score an ϵ above zero, and are effectively never selected by FAUST for discovery and annotation. Hence the restriction to 3-marker gating strategies.

4.5 FAUST method: annotation boundary estimation

The depth score is also used to estimate annotation boundaries. Recalling FAUST only explores gating strategies with 4 or fewer annotation boundaries, FAUST partitions the set

$$\mathcal{N}_m = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4 \ .$$

Define

$$\mathcal{G}_1 \equiv \{N_{m,i} \in \mathcal{N}_m \,|\, N_{m,i} \text{ has a single gate determined by the taut string}\} \ .$$

$\mathcal{G}_2, \mathcal{G}_3$, and $\mathcal{G}_4$ are defined similarly. In other words each $\mathcal{G}_i$ is the subset of nodes in the annotation forest for marker m i gates. Recalling (4.4), we can partition the score sum

$$\sum_{i=1}^{n} \omega(N_{m,i}) = \sum_{j=1}^{4} \sum_{N \in \mathcal{G}_j} \omega(N) \ .$$

FAUST selects the number of annotation boundaries for the marker m by choosing the set $\mathcal{G}_j$ with the maximal sum $\sum_{N \in \mathcal{G}_j} \omega(N)$. Letting $g_1(N_{m,j})$ denote the smallest gate location estimated by the taut string in node $N_{m,j}$ (which is the only gate location if FAUST selects $\mathcal{G}_1$), FAUST estimates

the phenotypic boundary locations for the marker by taking the weighted average

$$\frac{\sum_{N \in \mathcal{G}_j} \omega(N) g_1(N)}{\sum_{N \in \mathcal{G}_j} \omega(N)} .$$

In the event FAUST selects $\mathcal{G}_j, j > 1$ (i.e., multiple annotation boundaries), similar weighted averages are taken for $g_2(N_{m,j})$, etc.

4.6 FAUST method: marker selection

Markers are selected by comparing the user-selected, empirical depth score quantile 4.9.4 across experimental units to a user-selected threshold value 4.9.5. All markers whose empirical quantile exceeds the threshold are used for discovery and annotation.

4.7 FAUST method: boundary standardization

FAUST standardizes the number of annotation boundaries for each marker by majority vote. The most frequently occurring number of annotation boundaries across experimental units is chosen as the *standard* number. This behavior can be over-ridden via the preference list tuning parameter (see 4.9.6) in order to incorporate prior biological information into FAUST.

Next, for a given marker, FAUST selects the set of samples where the number of annotation boundaries for that marker matches the standard. Then, by rank, FAUST computes the median absolute deviation of the location of each phenotypic boundary across experimental units. We refer to these median boundary locations as the *standard boundaries*.

FAUST enforces standardization of annotation boundaries for non-conforming experimental units by imputation or deletion . Imputation in an experimental unit occurs when FAUST estimates fewer boundaries than the standard. In this case, each boundary in the non-conforming unit is matched to one of the standards by distance. Unmatched standards are used to impute the missing boundaries. Similar distance computations are done in the case of deletion, but FAUST deletes boundaries that are farthest from the standards. For both imputation and deletion, if multiple boundaries match the same standard, then the boundary minimizing the distance is kept, and the other boundaries are deleted. Should this result in standards that don't map to any boundaries, then those unmatched standards are used to impute the missing boundaries.

4.8 FAUST method: phenotype discovery and cluster annotation

For each experimental unit, FAUST constructs a forest of partition trees (randomly sampled) and annotates selected leaves from this forest relative to the standardized annotation boundaries. Partition tree construction is similar to tree construction for the annotation forest (4.3), but they are not depth-constrained: a tree continues to grow following the previously described strategy until each leaf is unimodal according to the dip test [44] or contains fewer than 25 cells. Consequently, a single partition tree defines a clustering of an experimental unit. Clusterings from the forest of partition trees are combined into a single clustering in the following manner. To ensure cells are not assigned to multiple clusters, a subset of leaves of the partition forest are selected by scoring leaves according to shape criteria, and then selecting a subset of leaves across partition trees that share no cells to maximize their total shape score. Only the selected leaves are given phenotypic annotations. FAUST keeps a list of discovered phenotypes for each experimental unit, and concludes by returning exact counts of cells in each sample whose phenotypes exceed a user-specified occurrence frequency threshold. For more details of the scoring and selection procedure, we refer the reader to [46].

4.9 FAUST method: tuning parameters

We describe the key tuning parameters of FAUST.

4.9.1 Starting cell population

The name of the population in the manual gating strategy where FAUST conducts discovery and annotation.

4.9.2 Active markers

A list of all markers in the experiment that can possibly be used for discovery and annotation in the starting cell population. FAUST will only compute the depth score for markers in this initial set.

4.9.3 Marker boundary matrix

A $2 \times n$ matrix of lower and upper protein expression bounds. By default, it is set for $-\inf$ and $\inf$ for all markers in a flow experiment. When the manual gating strategy does not remove all debris or doublets from the starting cell population, samples can appear to have clusters of events along at very low or very high expression values for some markers. By setting boundaries for those markers to exclude these doublet or debris clusters, FAUST treats all events below the lower and above the upper bounds as default low or high, respectively. These events are not dropped from the experiment. However, they are ignored when testing for multimodality and subsequent density estimation. In the case of mass cytometry experiments, the default lower boundary is set to 0 for all markers in an experiment in order to accommodate the zero-inflation common to mass cytometry data. The number of events in a marker that fall between the lower and upper marker boundaries in the starting cell population define the *effective sample size* for that marker.

4.9.4 Depth-score selection quantile

The empirical quantile of a marker's depth-score across all experimental units that is used to compare against a user-selected depth-score threshold. By default, this parameter is set to the median.

4.9.5 Depth-score selection threshold

A value in $[0, 1]$ used to select a subset of markers to be used in discovery and annotation based on their empirical depth score selection quantile. By default, this parameter is set to 0.01.

4.9.6 Supervised Boundary Estimation List

Allows the user to modify FAUST's default gate standardization methodology for each marker. This parameter is one way to incorporate prior (biological) knowledge in the FAUST procedure: if a marker is known to have a certain range of expression, such as low-dim-bright, this can be used to encourage or force FAUST to estimate the corresponding number of annotation boundaries from the data. Similarly, if FMO controls have been collected for a marker, this parameter can be used to set the phenotypic boundary according to the controls.

4.9.7 Phenotype Occurrence Threshold

An integer value (set to 1 by default) used to include or exclude discovered phenotypes in the final count matrix returned by FAUST. If a phenotype appears at least Phenotype Occurrence Threshold times across experimental units, it is included in the final counts matrix. By default, all discovered phenotypes are included. Phenotypes exceeding the threshold are assumed to be biological signal while those that fall below it are assumed to be sample- or batch-specific effects. A consequence of this assumption is that all cells in a sample associated with any phenotype falling below the threshold are re-annotated with a common non-informative label indicating those phenotypes ought not be analyzed due to their rarity.

4.10 CITN-09 T cell Panel Analysis

The CITN-09 T cell staining panel is described in the supplementary information A.10.1. FAUST tuning parameter settings (above) for this dataset are described in supplementary section A.9.2.

Between one and four samples were collected from 27 patients with stage IV and unresectable stage IIIB Merkel Cell Carcinoma and [21, 47] spanning the course of treatment. All 27 patients had samples collected at baseline (cycle C01, before initiation of anti-PD-1 therapy); 16 at cycle C02 (3 weeks post-treatment of the second cycle of therapy); 22 at cycle C05 (12 weeks post-treatment of the fifth cycle of therapy); and 13 at end of trial (EOT, patient specific). 18 of 27 subjects responded to therapy (CR/PR) for an observed response rate of 67%. Each sample was pre-gated to remove debris and identify live lymphocytes. Let $c_{i,k}$ denote the number of events in FAUST cluster k for sample i. Let n_i denote the number of events in the ith subject's baseline sample. Similar to [22], we assume $c_{i,k} \sim \text{Binomial}(n_i, \mu_{i,k})$. Our model is

$$\text{logit}^{-1}(\mu_{i,k}) = \beta_0 + \beta_1 \cdot \text{Responder} + \xi_{i,k} \,, \tag{4.5}$$

where Responder is an indicator variable equal to 1 when the subject exhibits complete or partial response to therapy, and 0 otherwise, and each $\xi_{i,k} \sim N(0, \sigma^2_{i,k})$ is a subject-level random effect. The R package **lme4** was used to fit all GLMMs [48].

4.11 CITN-09 Myeloid Panel

The CITN-09 Myeloid staining panel is described in supplementary information A.10.2. FAUST tuning parameter settings are described in supplementary information A.9.3. This dataset consisted of 69 samples stained to investigate myeloid cells. An initial screen comparing the ratio of the number of events in the singlet gate to the number of events in the root population led us to remove 14 samples from analysis due to low quality. We ran FAUST on the remaining 55 samples which consisted of 16 samples collected at cycle C01, before initiation of anti-PD-1 therapy; 15 at cycle C02; 15 at cycle C05; and 9 at EOT. Of the 16 baseline samples, 1 was coded as inevaluable "NE". This sample was removed from downstream statistical analysis. 10 of the 15 subjects with baseline samples available responded to therapy (PR/CR), for an observed response rate of 67%. Discovery and annotation was run at the individual sample level using cells in the "45+" node of the manual gating strategy. FAUST selected 11 markers: CD33, CD16, CD15, HLA-DR, CD14, CD3, CD11B, CD20, CD19, CD56, CD11C. FAUST annotated 102 cell sub-populations in terms of these markers, labeling 92.9% of the cells in the median sample. The statistical model used here is identical to (4.5), with counts are now derived from the 15 baseline samples.

4.12 CITN-07 Phenotyping Panel Analysis

We ran FAUST on this dataset comprising of a total of 358 longitudinal samples from 35 subjects in two cohorts (Cohort 1: with FLT-3 pre-treatment and Cohort 2: without pre-treatment), with between 4 and 12 samples per subject over four cycles of therapy and at end of trial. Subjects were given FLT-3 ligand seven days prior to the start of the first two of four treatment cycles. FLT-3 ligand was given to promote the expansion of myeloid and dendritic cell compartments in order to investigate whether expansion improved response to therapy. FAUST was configured to perform cell population discovery and annotation per sample in order to account for biological and technical heterogeneity. Debris, dead cells and non-lymphocytes were excluded by pre-gating. The CITN-07 Phenotyping staining panel is described in supplementary information A.10.3. FAUST tuning parameter settings are described in supplementary information A.9.1. FAUST discovered 132 cell populations. We tested each discovered cell population at the cohort-specific baseline for association with recurrence of disease (14 subjects had disease recur, 18 did not have disease recur). We analyzed the baseline counts using a model similar to (4.5). Here, the model was adjusted for

subject-to-subject variability using a random effect, while cohort status, recurrence, and NYESO-1 staining of the tumor by immunohistochemistry (measured as positive, negative, or undetermined) were modeled as population effects.

4.13 Krieg et al. CyTOF Analysis

The markers used for the Krieg et al. [8] CyTOF panel are described in supplementary information A.10.4. FAUST tuning parameter settings are described in supplementary information A.9.4. We used FAUST to discover and annotate cell populations in the mass cytometry datasets stained to investigate myeloid cells. Following [8], we removed samples with fewer than 50 cells from our analysis, leaving 19 samples (from 19 subjects) at baseline for downstream statistical analysis. 10 of the 19 samples at baseline were from subjects that went on to exhibit response to therapy. To account for batch effects and small sample sizes, all samples within a batch were concatenated and processed by FAUST. FAUST selected 11 markers for discovery and annotation: CD16, CD14, CD11b, CD11c, CD33, ICAM1, CD62L, PD-L1, CD7, CD56, and HLA-DR and annotated 64 cell sub-populations in terms of these markers, labeling 72.9% of cells in the median sample.

Our baseline model was similar to (4.5), but was modified by

$$\text{logit}^{-1}(\mu_{i,k}) = \beta_0 + \beta_1 \cdot \text{Responder} + \xi_{i,k} + \eta_{i,j} \ ,$$

where $j \in \{1, 2\}$, and $\eta_{i,j} \sim N(0, \sigma_j^2)$ is a random effect included to model the batch effects.

4.14 Krieg et al. FACS Analysis

The Krieg et al. [8] FACS staining panel is described in supplementary information A.10.5. FAUST tuning parameter settings are described in supplementary information A.9.5. We used FAUST to process 31 baseline flow cytometry samples from responders and non-responders to therapy (16 responders, 15 non-responders). FAUST was run at the individual sample level on live cells from the manual gating strategy used by [8]. QC and review of the manual gating strategy let us to make manual adjustments to the "Lymphocytes" gate of 7 samples in this dataset. An example of this gate adjustment is shown in the supplementary information (S4) FAUST selected 9 markers for discovery and annotation: CD3, CD4, HLA-DR, CD19, CD14, CD11b, CD56, CD16, and CD45RO.

FAUST annotated 40 cell sub-populations in terms of these markers, labeling 94.4% of cells in the median sample. The statistical model used here is identical to (4.5), with $c_{i,k}$ now denoting the 40 clusters in the FACS data, and n_i refers to the baseline FACS sample counts.

4.15 Compartment multivariate analysis

All FAUST clusters annotated as CD3-, CD56-, and CD19- and included in the univariate analysis were included in the multivariate analysis. Within this set, sub-populations annotated as HLA-DR- were further excluded. This defined the Myeloid compartment for CITN-07, CITN-09, and the Krieg et al. FACS data [8]. Let k^* denote the number of FAUST clusters within a given study. Let n denote the number of subjects at baseline, and $N = n \cdot k^*$. For $1 \leq i \leq N$, $1 \leq j \leq k^*$ our statistical model is

$$\text{logit}^{-1}(\mu_{i,j}) = \beta_0 + \beta_R \cdot \text{Responder}_i + \sum_{j=1}^{k^*} \left(\beta_{c,j} \cdot \text{Cluster}_{i,j} + \beta_{i,j} \cdot \text{Cluster}_{i,j} \cdot \text{Responder}_i\right) + \xi_i \ , \tag{4.6}$$

where $\text{Cluster}_{i,j}$ is an indicator variable that is 1 when observation i is from cluster j and 0 otherwise, Responder_i is an indicator variable when observation i is taken from a responding subject, and $\eta_i \sim N(0, \sigma_i^2)$ is an observation-level random effect. To test for differential abundance across a compartment, we test for positivity of linear combination of the coefficients $\beta_{i,j}$ in (4.6). For example to test for differential abundance across an entire compartment, we test

$$H_0 \ : \ \beta_R + \frac{1}{k^*} \cdot \sum_{j=1}^{k^*} \beta_{i,j} \leq 0 \ ,$$
$$H_1 \ : \ \beta_R + \frac{1}{k^*} \cdot \sum_{j=1}^{k^*} \beta_{i,j} > 0 \ .$$

4.16 Compartment aggregate analysis

For the aggregate analysis, compartment definitions are the same as presented in section 4.15. Counts are derived by summing across FAUST clusters within each compartment. The model(4.5) is then used to test each derived compartment for differential abundance.

4.17 Code availability

FAUST is available as an R package at https://github.com/RGLab/FAUST.

4.18 Author contributions

E.G., G.F., R.G. designed the FAUST method as well as statistical methods for analyzing FAUST cell populations. E.G., G.F. implemented FAUST and conducted data analyses using FAUST. E.G., G.F., R.G. contributed to design of data analysis plans, data analysis interpretation, and wrote the manuscript. L.A.D., C.D.C., C.M., N.R., J.M.T., P.T.N., M.A.C., S.P.F. contributed to the design of CITN-09 data analysis plans and data analysis interpretation. L.A.D., N.B., N.R., M.A.C., S.P.F. contributed to to the design of CITN-07 data analysis plans and data analysis interpretation. All authors discussed results and commented on the manuscript. All authors approve of this manuscript.

5 Acknowledgments

The authors gratefully acknowledge the clinical trials patients and their families. The authors thank Dr. Suzanne Topalian for helpful discussions and critical review of the manuscript. This work was supported by [1P01CA22551701] to P.T.N.; [UM1CA15496708] to M.A.C.; [R01GM118417] to G.F.; [K24-CA139052] to C.C., P.T.N.; [1U01CA154967] to M.A.C., S.P.F. from the National Cancer Institute; [P30-CA015704] to S.P.F., M.A.C., and P.T.N. from the NIH/NCI Cancer Center Support Grant in Seattle.

References

[1] Yvan Saeys, Sofie Van Gassen, and Bart N Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449, 2016.

[2] Greg Finak, Marc Langweiler, Maria Jaimes, Mehrnoush Malek, Jafar Taghiyar, Yael Korin, Khadir Raddassi, Lesley Devine, Gerlinde Obermoser, Marcin L Pekalski, Nikolas Pontikos, Alain Diaz, Susanne Heck, Federica Villanova, Nadia Terrazzini, Florian Kern, Yu Qian,

Rick Stanton, Kui Wang, Aaron Brandes, John Ramey, Nima Aghaeepour, Tim Mosmann, Richard H Scheuermann, Elaine Reed, Karolina Palucka, Virginia Pascual, Bonnie B Blomberg, Frank Nestle, Robert B Nussenblatt, Ryan Remy Brinkman, Raphael Gottardo, Holden Maecker, and J Philip McCoy. Standardizing flow cytometry immunophenotyping analysis from the human ImmunoPhenotyping consortium. *Sci. Rep.*, 6:20686, February 2016.

[3] Nima Aghaeepour, Greg Finak, The FlowCAP Consortium, The DREAM Consortium, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228, 2013.

[4] Lukas M Weber and Mark D Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096, 2016.

[5] Sofie Van Gassen, Britt Callebaut, Mary J Van Helden, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.

[6] Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature communications*, 8:14825, 2017.

[7] Robert V Bruggner, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014.

[8] Carsten Krieg, Malgorzata Nowicka, Silvia Guglietta, Sabrina Schindler, Felix J Hartmann, Lukas M Weber, Reinhard Dummer, Mark D Robinson, Mitchell P Levesque, and Burkhard Becher. High-dimensional single-cell analysis predicts response to anti-pd-1 immunotherapy. *Nature medicine*, 24(2):144, 2018.

[9] Joseph A. Fraietta, Simon F. Lacey, Elena J. Orlando, Iulian Pruteanu-Malinici, Mercy Gohil, Stefan Lundh, Alina C. Boesteanu, Yan Wang, Roddy S. OâĂŹConnor, Wei-Ting Hwang, Edward Pequignot, David E. Ambrose, Changfeng Zhang, Nicholas Wilcox, Felipe Bedoya, Corin Dorfmeier, Fang Chen, Lifeng Tian, Harit Parakandi, Minnal Gupta, Regina M. Young,

F. Brad Johnson, Irina Kulikovskaya, Li Liu, Jun Xu, Sadik H. Kassim, Megan M. Davis, Bruce L. Levine, Noelle V. Frey, Donald L. Siegel, Alexander C. Huang, E. John Wherry, Hans Bitter, Jennifer L. Brogdon, David L. Porter, Carl H. June, and J. Joseph Melenhorst. Determinants of response and resistance to cd19 chimeric antigen receptor (car) t cell therapy of chronic lymphocytic leukemia. *Nature medicine*, 24(5):563, 2018.

[10] Nima Aghaeepour, Radina Nikolic, Holger H Hoos, and Ryan R Brinkman. Rapid cell population identification in flow cytometry data. *Cytometry Part A*, 79(1):6–13, 2011.

[11] Markus Lux, Ryan Remy Brinkman, Cedric Chauve, Adam Laing, Anna Lorenc, Lucie Abeler-Dörner, Barbara Hammer, and Jonathan Wren. flowlearn: Fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinformatics*, 1:9, 2018.

[12] Yvan Saeys, Sofie Van Gassen, and Bart Lambrecht. Response to Orlova et al.âĂIJscience not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets". *Nature Reviews Immunology*, 18(1):78, 2018.

[13] Guenther Walther, Noah Zimmerman, Wayne Moore, David Parks, Stephen Meehan, Ilana Belitskaya, Jinhui Pan, and Leonore Herzenberg. Automatic clustering of flow cytometry data with density-based merging. *Advances in bioinformatics*, 2009, 2009.

[14] Kenneth Lo, Ryan Remy Brinkman, and Raphael Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 73(4):321–332, 2008.

[15] Daniel Commenges, Chariff Alkhassim, Raphael Gottardo, Boris Hejblum, and Rodolphe Thiebaut. cytometree: A binary tree algorithm for automatic gating in cytometry analysis. *bioRxiv*, page 335554, 2018.

[16] Darya Y Orlova, Noah Zimmerman, Stephen Meehan, Connor Meehan, Jeffrey Waters, Eliver EB Ghosn, Alexander Filatenkov, Gleb A Kolyagin, Yael Gernez, Shanel Tsuda, Wayne Moore, Richard B. Moss, Leonore A. Herzenberg, and Guenther Walther. Earth moverâĂŹs distance (emd): a true metric for comparing biomarker expression levels in cell populations. *PloS one*, 11(3):e0151859, 2016.

[17] Darya Y Orlova, Stephen Meehan, David Parks, Wayne A Moore, Connor Meehan, Qian Zhao, Eliver EB Ghosn, Leonore A Herzenberg, and Guenther Walther. Qfmatch: multidimensional flow and mass cytometry samples alignment. *Scientific reports*, 8(1):3291, 2018.

[18] Lukas M. Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Nature Communications Biology*, 2019.

[19] Zicheng Hu, Chethan Jujjavarapu, Jacob J Hughey, Sandra Andorf, Hao-Chih Lee, Pier Federico Gherardini, Matthew H Spitzer, Cristel G Thomas, John Campbell, Patrick Dunn, Jeff Wiser, Brian A. Kidd, Joel T. Dudley, Garry P. Nolan, Sanchita Bhattacharya, and Atul J. Butte. Metacyto: A tool for automated meta-analysis of mass and flow cytometry data. *Cell Reports*, 24(5):1377–1388, 2018.

[20] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[21] Paul Nghiem, Shailender Bhatia, Evan J. Lipson, William H. Sharfman, Ragini R. Kudchadkar, Andrew S. Brohl, Phillip A. Friedlander, Adil Daud, Harriet M. Kluger, Sunil A. Reddy, Brian C. Boulmay, Adam I. Riker, Melissa A. Burgess, Brent A. Hanks, Thomas Olencki, Kim Margolin, Lisa M. Lundgren, Abha Soni, Nirasha Ramchurren, Candice Church, Song Y. Park, Michi M. Shinohara, Bob Salim, Janis M. Taube, Steven R. Bird, Nageatte Ibrahim, Steven P. Fling, Blanca Homet Moreno, Elad Sharon, Martin A. Cheever, and Suzanne L. Topalian. Durable tumor regression and overall survival in patients with advanced merkel cell carcinoma receiving pembrolizumab as first-line therapy. *Journal of Clinical Oncology*, 37(9):693–702, 2019. PMID: 30726175.

[22] Malgorzata Nowicka, Carsten Krieg, Lukas M Weber, Felix J Hartmann, Silvia Guglietta, Burkhard Becher, Mitchell P Levesque, and Mark D Robinson. Cytof workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 2017.

[23] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe,

and J. Verweij. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, 45(2):228 – 247, 2009. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers.

[24] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[25] Enfu Hui, Jeanne Cheung, Jing Zhu, Xiaolei Su, Marcus J. Taylor, Heidi A. Wallweber, Dibyendu K. Sasmal, Jun Huang, Jeong M. Kim, Ira Mellman, and Ronald D. Vale. T cell costimulatory receptor cd28 is a primary target for pd-1–mediated inhibition. *Science*, 355(6332):1428–1433, 2017.

[26] Alice O. Kamphorst, Andreas Wieland, Tahseen Nasti, Shu Yang, Ruan Zhang, Daniel L. Barber, Bogumila T. Konieczny, Candace Z. Daugherty, Lydia Koenig, Ke Yu, Gabriel L. Sica, Arlene H. Sharpe, Gordon J. Freeman, Bruce R. Blazar, Laurence A. Turka, Taofeek K. Owonikoko, Rathi N. Pillai, Suresh S. Ramalingam, Koichi Araki, and Rafi Ahmed. Rescue of exhausted cd8 t cells by pd-1–targeted therapies is cd28-dependent. *Science*, 355(6332):1423–1427, 2017.

[27] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[28] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38, 2019.

[29] Alan L Landay, Carl E Mackewicz, and Jay A Levy. An activated cd8+ t cell phenotype correlates with anti-hiv activity and asymptomatic clinical status. *Clinical immunology and immunopathology*, 69(1):106–116, 1993.

[30] Nicolas A. Giraldo, Peter Nguyen, Elizabeth L. Engle, Genevieve J. Kaunitz, Tricia R. Cottrell, Sneha Berry, Benjamin Green, Abha Soni, Jonathan D. Cuda, Julie E. Stein, Joel C. Sunshine, Farah Succaria, Haiying Xu, Aleksandra Ogurtsova, Ludmila Danilova, Candice D. Church, Natalie J. Miller, Steve Fling, Lisa Lundgren, Nirasha Ramchurren, Jennifer H. Yearley, Evan J.

Lipson, Mac Cheever, Robert A. Anders, Paul T. Nghiem, Suzanne L. Topalian, and Janis M. Taube. Multidimensional, quantitative assessment of pd-1/pd-l1 expression in patients with merkel cell carcinoma and association with response to pembrolizumab. *Journal for immunotherapy of cancer*, 6(1):99, 2018.

[31] Natalie J. Miller, Candice D. Church, Steven P. Fling, Rima Kulikauskas, Nirasha Ramchurren, Michi M. Shinohara, Harriet M. Kluger, Shailender Bhatia, Lisa Lundgren, Martin A. Cheever, Suzanne L. Topalian, and Paul Nghiem. Merkel cell polyomavirus-specific immune responses in patients with merkel cell carcinoma receiving anti-pd-1 therapy. *Journal for immunotherapy of cancer*, 6(1):131, 2018.

[32] Rossella Melchiotti, Filipe Gracio, Shahram Kordasti, Alan K Todd, and Emanuele de Rinaldis. Cluster stability in the analysis of mass cytometry data. *Cytometry A*, 91(1):73–84, January 2017.

[33] L Fong, Y Hou, A Rivas, C Benike, A Yuen, G A Fisher, M M Davis, and E G Engleman. Altered peptide ligand vaccination with flt3 ligand expanded dendritic cells for tumor immunotherapy. *Proc. Natl. Acad. Sci. U. S. A.*, 98(15):8809–8814, July 2001.

[34] Nina Bhardwaj, Anna C. Pavlick, Marc S. Ernstoff, Brent Allen Hanks, Mark R. Albertini, Jason John Luke, Michael Jay Yellin, Tibor Keler, Thomas A. Davis, Andrea Crocker, Laura Vitale, Chihiro Morishima, Philip Adam Friedlander, Martin A. Cheever, and Steven Fling. A phase ii randomized study of cdx-1401, a dendritic cell targeting ny-eso-1 vaccine, in patients with malignant melanoma pre-treated with recombinant cdx-301, a recombinant human flt3 ligand. *Journal of Clinical Oncology*, 34(15):9589, 2016.

[35] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, 40(17):e133, September 2012.

[36] Evangelia Pardali, Timo Schmitz, Andreas Borgscheiper, Janette Iking, Lars Stegger, and Johannes Waltenberger. Cryopreservation of primary human monocytes does not negatively affect their functionality or their ability to be labelled with radionuclides: basis for molecular imaging and cell therapy. *EJNMMI research*, 6(1):77, 2016.

[37] Jeroen H Gerrits, Petros Athanassopoulos, Lenard MB Vaessen, Mariska Klepper, Willem

Weimar, and Nicole M van Besouw. Peripheral blood manipulation significantly affects the result of dendritic cell monitoring. *Transplant immunology*, 17(3):169–177, 2007.

[38] Jiarui Ding, Sohrab Shah, and Anne Condon. densitycut: An efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics*, 32(17):2567–2576, 2016.

[39] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana PeâĂŹer, and Garry P. Nolan. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184 – 197, 2015.

[40] Ran He, Shiyue Hou, Cheng Liu, Anli Zhang, Qiang Bai, Miao Han, Yu Yang, Gang Wei, Ting Shen, Xinxin Yang, Lifan Xu, Xiangyu Chen, Yaxing Hao, Pengcheng Wang, Chuhong Zhu, Juanjuan Ou, Houjie Liang, Ting Ni, Xiaoyan Zhang, Xinyuan Zhou, Kai Deng, Yaokai Chen, Yadong Luo, Jianqing Xu, Hai Qi, Yuzhang Wu, and Lilin Ye. Follicular cxcr5-expressing cd8+ t cells curtail chronic viral infection. *Nature*, 537(7620):412, 2016.

[41] Jolanda Brummelman, Emilia M.C. Mazza, Giorgia Alvisi, Federico S. Colombo, Andrea Grilli, Joanna Mikulak, Domenico Mavilio, Marco Alloisio, Francesco Ferrari, Egesta Lopci, Pierluigi Novellis, Giulia Veronesi, and Enrico Lugli. High-dimensional single cell analysis identifies stem-like cytotoxic cd8+ t cells infiltrating human tumors. *Journal of Experimental Medicine*, 215(10):2520–2535, 2018.

[42] Greg Finak, Jacob Frelinger, Wenxin Jiang, Evan W Newell, John Ramey, Mark M Davis, Spyros A Kalams, Stephen C De Rosa, and Raphael Gottardo. Opencyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS computational biology*, 10(8):e1003806, 2014.

[43] Greg Finak and Mike Jiang. Flowworkspace: Infrastructure for representing and interacting with the gated cytometry. *R package version*, 3(3), 2011.

[44] John A Hartigan and PM Hartigan. The Dip Test of Unimodality. *The Annals of Statistics*, pages 70–84, 1985.

[45] P Laurie Davies and Arne Kovac. Densities, spectral densities and modality. *The Annals of Statistics*, 32(3):1093–1136, 2004.

[46] Evan Greene, Greg Finak, and Raphael Gottardo. Selective clustering annotated using modes of projections. *arXiv preprint arXiv:1807.10328*, 2018.

[47] Paul T. Nghiem, Shailender Bhatia, Evan J. Lipson, Ragini R. Kudchadkar, Natalie J. Miller, Lakshmanan Annamalai, Sneha Berry, Elliot K. Chartash, Adil Daud, Steven P. Fling, Philip A. Friedlander, Harriet M. Kluger, Holbrook E. Kohrt, Lisa Lundgren, Kim Margolin, Alan Mitchell, Thomas Olencki, Drew M. Pardoll, Sunil A. Reddy, Erica M. Shantha, William H. Sharfman, Elad Sharon, Lynn R. Shemanski, Michi M. Shinohara, Joel C. Sunshine, Janis M. Taube, John A. Thompson, Steven M. Townson, Jennifer H. Yearley, Suzanne L. Topalian, and Martin A. Cheever. Pd-1 blockade with pembrolizumab in advanced merkel-cell carcinoma. *New England Journal of Medicine*, 374(26):2542–2552, 2016. PMID: 27093365.

[48] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[49] Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci. *Handbook of cluster analysis*. CRC Press, 2015.