A Supplementary materials

A.1 Summary of CITN clinical trials

	CITN-07	CITN-09
ClinicalTrials.gov Identifier*	NCT02129075	NCT02267603
Study Type*	Interventional (Clinical Trial)	Interventional (Clinical Trial)
Intervention Model*	Parallel Assignment	Single Group Assignment
Masking*	None (Open Label)	None (Open Label)
Enrollment*	100 participants (Estimated, recorded 6/11/2019)	50 participants (Actual, recorded 6/11/2019)
Primary Purpose*	Treatment	Treatment
Actual Study Start Date*	April 9, 2014	November 25, 2014
Primary Completion Date*	February 2, 2020 (Estimated, recorded 6/11/2019)	February 6, 2018 (Actual, recorded 6/11/2019
Total number of longitudinal samples analyzed by FAUST	358	78 (T cell pane), 55 (Myeloid panel)
Total number of baseline samples (pre-treatment)	32	27 (T cell panel), 15 (Myeloid panel)
Phase*	Phase 2	Phase 2
Intervention/Treatment*	Biological: DEC-205/NY-ESO-1 Fusion Protein CDX-1401 Other: Laboratory Biomarker Analysis Biological: Neoantigen-based Melanoma-Poly-ICLC Vaccine Other: Pharmacological Study Biological: Recombinant Flt3 Ligand	Biological: Pembrolizumab Other: Laboratory Biomarker Analysis
Condition/Disease*	Cutaneous Melanoma Mucosal Melanoma NY-ESO-1 Positive Tumor Cells Present Ocular Melanoma Stage IIB Cutaneous Melanoma AJCC v6 and v7 Stage IIC Cutaneous Melanoma AJCC v6 and v7 Stage III Cutaneous Melanoma AJCC v7 Stage IIIA Cutaneous Melanoma AJCC v7 Stage IIIB Cutaneous Melanoma AJCC v7 Stage IIIC Cutaneous Melanoma AJCC v7 Stage IV Cutaneous Melanoma AJCC v6 and v7	Recurrent Merkel Cell Carcinoma Stage III Merkel Cell Carcinoma AJCC v7 Stage IIIA Merkel Cell Carcinoma AJCC v7 Stage IIIB Merkel Cell Carcinoma AJCC v7 Stage IV Merkel Cell Carcinoma AJCC v7

Table S1: Data listed in all rows with * taken from https://clinicaltrials.gov on June 11, 2019.

A.2 Baseline predictors MCC anti-PD-1 trial myeloid phenotyping panel

Population	Effect Size	Lower 2.5%	Upper 97.5%	BonferroniP
CD33 Bright CD16- CD15- HLADR Bright CD14- CD3- CD11B- CD20- CD19- CD56- CD11C+	2.611	1.073	4.162	0.041
CD33 Bright CD16- CD15- HLADR Bright CD14- CD3- CD11B+ CD20- CD19- CD56- CD11C+	2.776	1.316	4.274	0.009
CD33 Bright CD16- CD15- HLADR Bright CD14+ CD3- CD11B+ CD20- CD19- CD56- CD11C+	3.558	1.837	5.284	0.002
CD33 Bright CD16- CD15+ HLADR Bright CD14+ CD3- CD11B+ CD20- CD19- CD56- CD11C+	3.689	1.859	5.823	0.007
CD33 Bright CD16+ CD15+ HLADR Bright CD14+ CD3- CD11B+ CD20- CD19- CD56- CD11C+	4.754	2.353	7.578	0.011

Table S2: All statistically significant (bonferroni adjusted significance threshold of 5%) from the MCC anti-PD-1 trial.

A.3 Baseline predictors FLT3-L + therapeutic Vx trial.

The complete set of baseline predictors from the FLT3-L + therapeutic Vx trial are listed in Table S3. The top populations, by magnitude, were CD14+CD16- monocyte populations.

Population	estimate	lower	upper	std.error	statistic	p.value	adjusted.p.value
CD8-CD3-HLA DRbright CD4-CD19-CD14+CD11C+CD123-CD16-CD56-	2.44	1.16	3.73	0.66	3.73	9.70e-05	0.01
CD8-CD3-HLA DRdimCD4-CD19-CD14+CD11C+CD123-CD16-CD56-	2.26	1.22	3.31	0.54	4.23	1.17e-05	0.00
CD8dim CD3-HLA DRbright CD4-CD19-CD14+CD11C+CD123-CD16-CD56-	2.14	0.98	3.29	0.59	3.62	1.46e-04	0.02
CD8dim CD3-HLA DRdim CD4-CD19-CD14-CD11C-CD123-CD16+CD56+	1.91	0.93	2.89	0.50	3.82	6.67e-05	0.01
CD8dim CD3-HLA DR-CD4-CD19-CD14-CD11C-CD123-CD16+CD56+	1.60	0.73	2.47	0.44	3.59	1.64e-04	0.02
CD8-CD3-HLA DRdim CD4-CD19-CD14-CD11C-CD123-CD16+CD56+	1.60	0.69	2.50	0.46	3.47	2.65e-04	0.03
CD8bright CD3+HLA DRbright CD4-CD19-CD14-CD11C-CD123-CD16-CD56-	1.50	0.71	2.29	0.40	3.72	1.01e-04	0.01
CD8dim CD3-HLA DRdim CD4-CD19-CD14-CD11C-CD123-CD16+CD56-	1.23	0.75	1.70	0.24	5.07	2.01e-07	0.00
CD8-CD3-HLA DRbright CD4-CD19-CD14-CD11C-CD123-CD16-CD56-	1.22	0.60	1.83	0.32	3.85	5.90e-05	0.01
CD8-CD3-HLA DRdimCD4-CD19-CD14-CD11C-CD123-CD16+CD56-	1.12	0.68	1.56	0.22	5.02	2.64e-07	0.00
CD8dim CD3-HLA DR-CD4-CD19-CD14-CD11C-CD123-CD16+CD56-	0.97	0.58	1.35	0.20	4.92	4.34e-07	0.00
CD8-CD3-HLA DR-CD4-CD19-CD14-CD11C-CD123-CD16+CD56-	0.92	0.53	1.30	0.20	4.67	1.49e-06	0.00
CD8-CD3-HLA DRdim CD4-CD19-CD14-CD11C-CD123-CD16-CD56-	0.75	0.31	1.19	0.22	3.38	3.64e-04	0.05

Table S3: All statistically significant (bonferroni adjusted significance threshold of 5%) from the FLT3-L + therapeutic Vx trial.

A.4 Alternative normalization of PD-1 dim T cell biomarker in CITN-09 correlates with clinical outcome in CITN-09

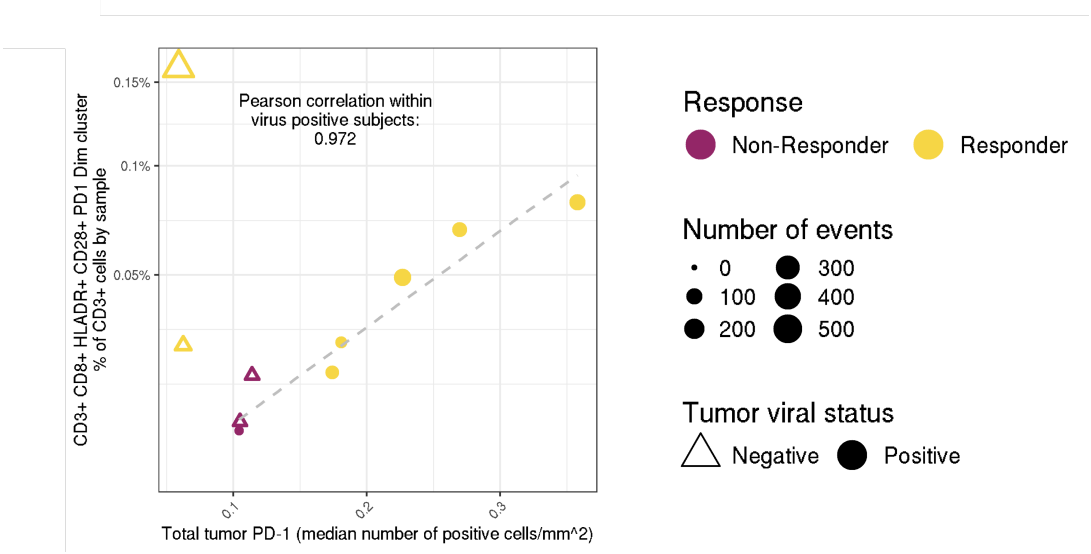

Figure S1: This figure is the same FAUST sub-population reported in figure 3 – CD4- CD3+ CD8+ CD45RA- HLA-DR+ CD28+ PD-1 dim CD25- CD127- CCR7- – normalized by the total count of CD3+ FAUST sub-populations by sample in panel C, displayed here.

A.5 Additional sub-populations associated with with clinical outcome at baseline in CITN-09

The figures in this sub-section are the remaining sub-populations discovered and annotated by FAUST that are associated with clinical outcome at the FDR-adjusted 5% level.

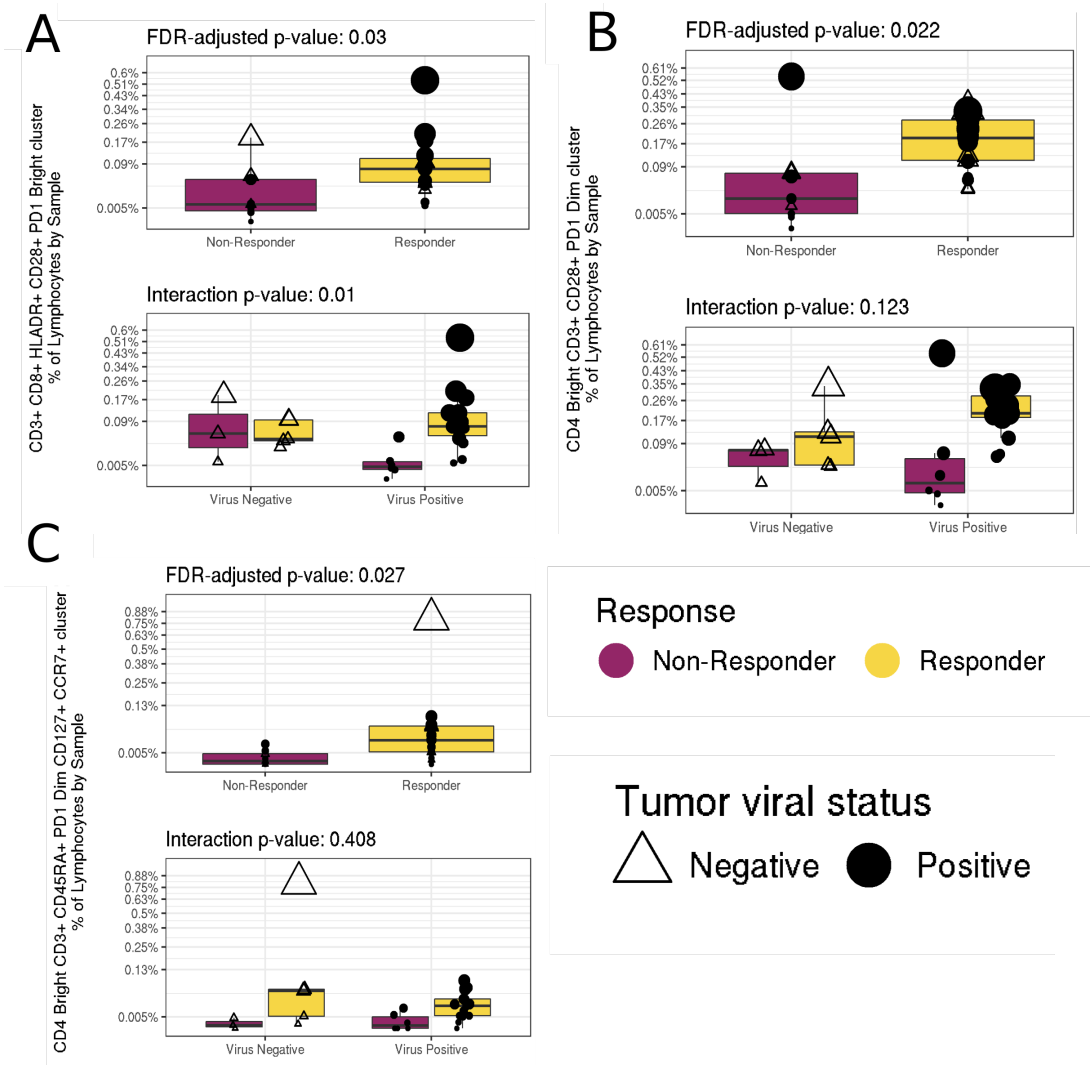

Figure S2: Panel A: The FAUST sub-population annotated CD4- CD3+ CD8+ CD45RA- HLA-DR+ CD28+ PD-1 bright CD25- CD127- CCR7- that is associated with clinical outcome at the FDR-adjusted 5% level. Panel B: The FAUST sub-population annotated CD4 bright CD3+ CD8- CD45RA- HLA-DR- CD28+ PD-1 dim CD25- CD127- CCR7- that is associated with clinical outcome at the FDR-adjusted 5% level. Panel C: The FAUST sub-population annotated CD4 bright CD3+ CD8- CD45RA+ HLA-DR- CD28- PD-1 dim CD25- CD127+ CCR7+ that is associated with clinical outcome at the FDR-adjusted 5% level.

A.6 Effect Sizes and Confidence Intervals in CITN-09

Population	Effect Size	Lower 2.5%	Upper 97.5%
CD4- CD3+ CD8+ CD45RA- HLA-DR+ CD28+ PD-1+ CD25- CD127- CCR7-	1.844	0.784	2.955
CD4- CD3+ CD8+ CD45RA- HLA-DR+ CD28+ PD-1 dim CD25- CD127- CCR7-	1.896	0.898	2.981
CD4 bright CD3+ CD8- CD45RA- HLA-DR- CD28+ PD-1 dim CD25- CD127- CCR7-	1.907	0.929	2.941
CD4 bright CD3+ CD8- CD45RA+ HLA-DR- CD28- PD-1 dim CD25- CD127+ CCR7+	2.999	1.387	4.837

A.7 Alternative analysis of CITN-09 T cell panel

As described in the main text, 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. We ran each method on live lymphocytes from all 78 experimental samples as well as from the 27 baseline samples alone, transforming samples using both the biexponential as well as the hyperbolic arcsine. For all non-FAUST methods, samples were combined before clustering in all scenarios, and we set tuning parameters to the settings reported in [4] when possible. After testing for differential abundance between responders and non-responders using counts derived from each method's clusterings, three clusters defined by densityCut were significantly associated with response to therapy at the FDR-adjusted 0.20 level (Supplementary Table S4), but none of these represented T cells (Supplementary Figure S3). No other clusters produced by densityCut, FlowSOM, or Phenograph were associated with response to therapy at this level of significance. On the other hand, FAUST repeatedly found that CD28+ HLA-DR+ PD-1 expressing effector-memory CD8 T cells as well as CD28+ PD-1 expressing CD4 T cells were associated with response to therapy at baseline across all tested conditions at the FDR-adjusted 5% level.

Method	Num Clusters	Transformation	Input Data	Best FDR	Second FDR
DensityCut	2599	Biexp	Baseline	0.09	1.00
DensityCut	2570	Asinh	Baseline	0.83	0.83
DensityCut	6389	Biexp	All	0.00	0.22
DensityCut	6166	Asinh	All	0.00	1.00
FlowSOM	100	Asinh	Baseline	0.62	0.62
FlowSOM	400	Asinh	Baseline	0.53	0.53
FlowSOM	100	Biexp	Baseline	0.49	0.71
FlowSOM	400	Biexp	Baseline	0.58	0.58
FlowSOM	100	Asinh	All	0.64	0.64
FlowSOM	400	Asinh	All	0.43	0.43
FlowSOM	100	Biexp	All	0.65	0.65
FlowSOM	400	Biexp	All	0.29	0.29
Phenograph	46	Asinh	Baseline	0.35	0.35
Phenograph	45	Biexp	Baseline	0.36	0.36
Phenograph	47	Asinh	All	0.31	0.31
Phenograph	49	Biexp	All	0.34	0.34
FAUST	267	Biexp	Baseline	0.03	0.03
FAUST	290	Asinh	Baseline	0.00	0.02
FAUST	238	Biexp	All	0.02	0.02
FAUST	239	Asinh	All	0.00	0.00

Table S4: Results of applying the clustering methods DensityCut [38], FAUST, FlowSOM [5], and Phenograph [39] to flow cytometry data stained to investigate T cell activity from the MCC anti-PD-1 trial. Tuning parameters for FlowSOM and Phenograph (including the number of clusters for FlowSOM) were set to the parameter settings reported in [4]. There, the supporting table "cytoa23030-sup-0001-suppinfo.xlsx" reports that Phenograph is run with k = 30 neighbors and using the Euclidean metric. "FlowSOM pre" is reported as running with 100 and 400 clusters with "transform=FALSE" and "scale=FALSE". Flow cytometry data are reported as being transformed by the hyperbolic arcsine transformation with cofactor 120. Here, we transform data using both the biexponential transformation (used by CITN on the "Biexp" rows) as well as the hyperbolic arcsine with cofactor 120 (The "Asinh" rows). DensityCut was run totally unsupervised: the K parameter is set to its default value $\log_2(N)$. Samples were concatenated before analysis by each method except FAUST, which was run at the sample level for all analyses. Rows with "Input data" listing "Baseline" only combine patient samples from the baseline time point, while those list All" have samples from all time points combined prior to analysis. The reported FAUST number of clusters is the number of clusters with "CD3+" annotations. The tuning parameters for FAUST in the "Asinh" runs and the baseline "Biexp" run were taken from the FAUST "Biexp" all run, which is reported in the paper. The channel bounds matrix was transformed to the "Asinh" runs by computing the empirical quantiles of the concatenated biexponentially transformed data corresponding to the bounds reported in supplementary section A.9.2, and then computing those quantiles on the transformed concatenated data. Similarly, the baseline phenotypic filtering threshold was scaled from the setting of 5 for the 78 all sample runs to the setting of 2 for the 27 baseline sample runs.

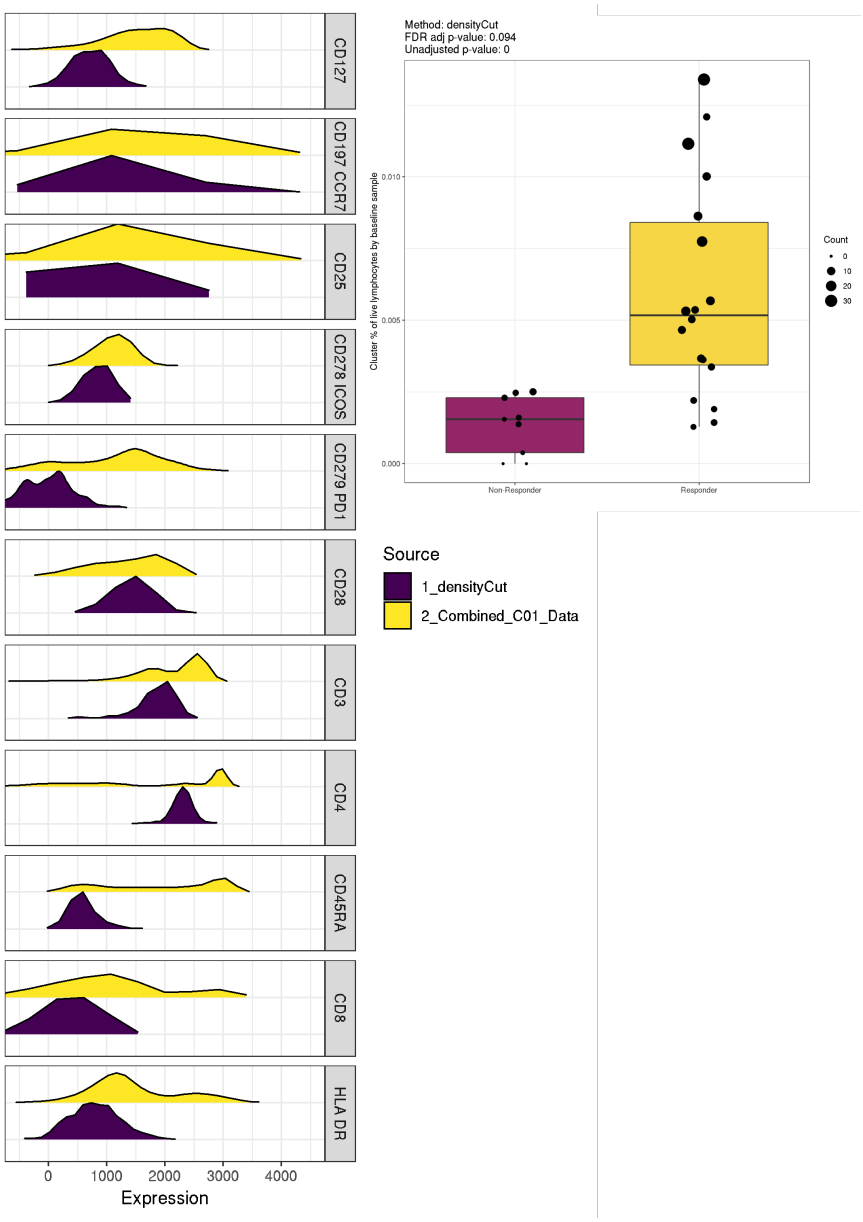

Figure S3: This figure shows the statistically significant correlate determined by running DensityCut on the baseline CITN-09 samples. DensityCut was installed using the command devtools::install_bitbucket("jerry00/densitycut_dev") in R 3.5.0, to install version 0.0.1 of the package. DensityCut was run unsupervised: the K parameter is set to its default value $\log_2(N)$. Two of the correlating densityCut clusters contained 2 and 20 cells in total across baseline samples, and were only measured in 1 and 2 of the 27 baseline subjects, respectively. We thus viewed the observed correlation for these clusters as artifactual. Plots of the third cluster's expression relative to the baseline samples (displayed in this figure) indicated that the cluster was CD3-, and so is not a T cell subset.

A.8 Gating strategy modification examples

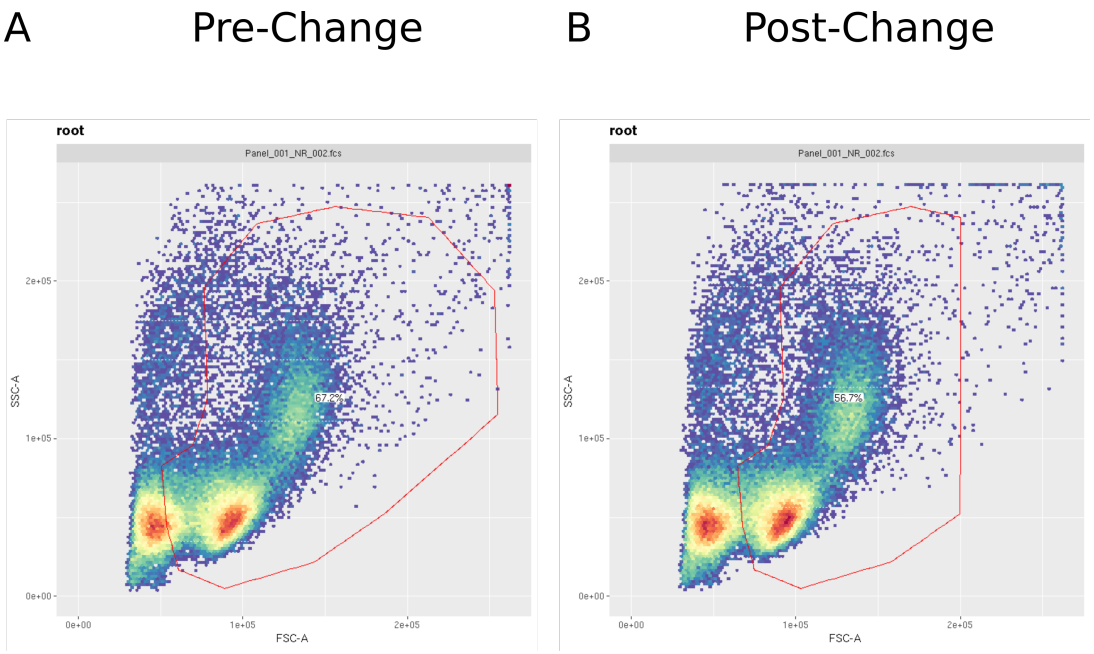

Figure S4: An example of modification to the manual gating strategy of the Krieg et al. FACS data. Panel A shows the initial manual gating strategy for the Lymphocytes of a sample. Panel B shows the same sample with the modified gate.

A.9 FAUST tuning parameter settings for data analysis

A.9.1 CITN-07 Phenotyping panel

The marker boundary matrix for CITN-07. Specific values were set by inspecting histograms of the individual markers across individual samples.

	Low	High
CD123	1500.00	3000.00
CD4	100.00	2500.00
CD14	1.00	3000.00
CD11C	1.00	3000.00
CD56	1.00	3000.00
CD8	1.00	3000.00
CD16	1000.00	3000.00
CD3	1.00	3000.00
CD122	1000.00	3000.00
CD19	1501.00	3000.00
HLA DR	1.00	3500.00

The selection quantile was set to 0.05. The selection threshold was set to $1e-7$. The supervised list was used to encourage PD-1 to have two annotation boundaries estimated. The phenotype occurrence number was set to 70.

A.9.2 CITN-09 T cell panel

The marker boundary matrix. Non-zero values were set by inspecting histograms of the individual samples.

	Low	High
CD278 ICOS	-500.00	3000.00
CD3	-500.00	3000.00
CD127	-500.00	3000.00
CD197 CCR7	500.00	2250.00
CD279 PD1	-250.00	3500.00
CD8	500.00	3000.00
CD4	500.00	3000.00
CD28	-500.00	3000.00
CD25	500.00	2250.00
HLA DR	-500.00	3000.00
CD45RA	1000.00	3000.00

The selection quantile was set to 0.05. The selection threshold was set to The supervised list was used to encourage PD-1 to have two annotation boundaries estimated. The phenotype occurance number was set to 5.

838 ### A.9.3 CITN-09 Myeloid panel

839 The marker boundary matrix. Non-zero values were set by inspecting histograms of the individual
840 samples.

	Low	High
CD11B	1000.00	Inf
CD20	1000.00	Inf
CD14	1000.00	Inf
CD11C	1000.00	Inf
CD56	2000.00	Inf
CD33	-1000.00	Inf
CD16	-Inf	Inf
CD3	1000.00	Inf
CD15	1000.00	Inf
CD19	1750.00	Inf
HLA DR	-1000.00	3750.00

841 The selection quantile was set to 0.50. The selection threshold was set to 0.05. The supervised
842 list was used to encourage both CD33 and HLA-DR to have two annotation boundaries estimated.
843 The phenotype occurance number was set to 14.

A.9.4 Krieg et al. CyTOF

The marker boundary matrix. Upper bounds were set to the 99^{th} quantiles across the dataset. Lower boundes were set to zero by default. Non-zero values were set by concatenating experimental samples together by batch, and inspecting histograms of the concatenated batches.

	Low	High
209Bi_CD11b	0.00	5.55
162Dy_CD11c	0.00	5.28
163Dy_CD7	0.00	4.85
166Er_CD209	1.00	3.00
167Er_CD38	0.00	4.10
151Eu_CD123	0.00	3.89
153Eu_CD62L	0.00	4.57
152Gd_CD66b	2.00	4.00
154Gd_ICAM-1	0.00	4.95
155Gd_CD1c	0.00	2.30
156Gd_CD86	0.00	3.88
160Gd_CD14	0.00	4.33
165Ho_CD16	0.00	4.53
175Lu_PD-L1	0.00	3.73
146Nd_CD64	0.00	2.95
147Sm_CD303	2.00	4.00
148Sm_CD34	1.00	3.00
149Sm_CD141	0.00	3.96
150Sm_CD61	0.00	4.29
169Tm_CD33	0.00	3.68
89Y_CD45	0.00	5.25
173Yb_CD56	0.00	3.43
174Yb_HLA-DR	0.00	5.89

The selection quantile was set to 1: after concatenating by batch, there were two experimental units in this dataset.

The selection threshold was set to 0.05.

The supervised list was set to use only the lower-estimated gate for CD33 after selection and standardization: the upper gate was deemed an artifact of concatenation by inspection. The phenotype occurance number was kept at the default value of 1.

A.9.5 Krieg et al. FACS

The marker boundary matrix. Non-zero values were set by inspecting histograms of the individual samples.

	CD3	CD4	CD11b	CD33	HLA-DR	CD56	CD45RO	CD11c	CD16	CD14	CD19
Low	-Inf	-20.00	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	85.00	-Inf	-Inf
High	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

The selection quantile was set to 0. The selection threshold was set to 0.01. The supervised list was not used. The phenotype occurance number was set to 13.

A.10 Staining panels used in FAUST analyses

Staining panels from the experiments used in FAUST analyses are provided here.

A.10.1 CITN-09 T cell Staining Panel

	name	desc
$P1	FSC-A	
$P2	FSC-H	
$P3	SSC-A	
$P4	SSC-H	
$P5	<PE-A>	CD278 ICOS
$P6	<FITC-A>	CD3
$P7	<BV 421-A>	CD127
$P8	<Alexa Fluor 700-A>	CD197 CCR7
$P9	<PE-Cy7-A>	CD279 PD-1
$P10	<PerCP-Cy5-5-A>	CD8
$P11	<APC-Cy7-A>	CD4
$P12	<ECD-A>	CD28
$P13	<APC-A>	CD25
$P14	PE-Cy5-A	
$P15	<AmCyan-A>	CD45
$P16	<BV 605-A>	HLA DR
$P17	<BV 650-A>	CD45RA
$P18	Time	

A.10.2 CITN-09 Myeloid Staining Panel

	name	desc
$P1	FSC-A	
$P2	FSC-H	
$P3	SSC-A	
$P4	SSC-H	
$P5	<PE-A>	CD11B
$P6	<FITC-A>	CD20
$P7	<BV 421-A>	CD14
$P8	<Alexa Fluor 700-A>	CD11C
$P9	<PE-Cy7-A>	CD56
$P10	<PerCP-Cy5-5-A>	CD33
$P11	<APC-Cy7-A>	CD16
$P12	<ECD-A>	CD3
$P13	<APC-A>	CD15
$P14	<PE-Cy5-A>	CD19
$P15	<AmCyan-A>	CD45
$P16	<BV 605-A>	HLA DR
$P17	BV 650-A	
$P18	Time	

A.10.3 CITN-07 Phenotyping Staining Panel

	name	desc
$P1	FSC-A	
$P2	FSC-H	
$P3	SSC-A	
$P4	SSC-H	
$P5	<PE-A>	CD123
$P6	<FITC-A>	CD4
$P7	<BV 421-A>	CD14
$P8	<Alexa Fluor 700-A>	CD11C
$P9	<PE-Cy7-A>	CD56
$P10	<PerCP-Cy5-5-A>	CD8
$P11	<APC-Cy7-A>	CD16
$P12	<ECD-A>	CD3
$P13	<APC-A>	CD122
$P14	<PE-Cy5-A>	CD19
$P15	<AmCyan-A>	CD45
$P16	<BV 605-A>	HLA DR
$P17	BV 650-A	
$P18	Time	

A.10.4 Krieg et al. Myeloid CyTOF Panel

	name	desc
$P1	Bi209Di	209Bi_CD11b
$P2	Dy162Di	162Dy_CD11c
$P3	Dy163Di	163Dy_CD7
$P4	Er166Di	166Er_CD209
$P5	Er167Di	167Er_CD38
$P6	Eu151Di	151Eu_CD123
$P7	Eu153Di	153Eu_CD62L
$P8	Gd152Di	152Gd_CD66b
$P9	Gd154Di	154Gd_ICAM-1
$P10	Gd155Di	155Gd_CD1c
$P11	Gd156Di	156Gd_CD86
$P12	Gd160Di	160Gd_CD14
$P13	Ho165Di	165Ho_CD16
$P16	Lu175Di	175Lu_PD-L1
$P18	Nd146Di	146Nd_CD64
$P22	Sm147Di	147Sm_CD303
$P23	Sm148Di	148Sm_CD34
$P24	Sm149Di	149Sm_CD141
$P25	Sm150Di	150Sm_CD61
$P26	Tm169Di	169Tm_CD33
$P27	Y89Di	89Y_CD45
$P29	Yb173Di	173Yb_CD56
$P30	Yb174Di	174Yb_HLA-DR

A.10.5 Krieg et al. FACS Panel

	name	desc
$P1	FSC-A	
$P2	FSC-H	
$P3	FSC-W	
$P4	SSC-A	
$P5	SSC-H	
$P6	SSC-W	
$P7	Comp-Brilliant Violet 785-A	CD3
$P8	Comp-Brilliant Violet 711-A	CD4
$P9	Comp-Brilliant Violet 421-A	CD11b
$P10	Comp-PerCP-Cy5-5-A	CD33
$P11	Comp-FITC-A	HLA-DR
$P12	Comp-PE-Cy7-A	CD56
$P13	Comp-PE-Texas Red-A	CD45RO
$P14	Comp-APC-Cy7-A	NIR
$P15	Comp-Alexa Fluor 700-A	CD11c
$P16	Comp-APC-A	CD16
$P17	Comp-PE-A	CD14
$P18	Comp-Brilliant Violet 605-A	CD19
$P19	Time	

A.11 Manual gating strategies

Manual gating strategies for analyzed trials are included here.

A.11.1 CITN-09 T cell Manual Gating Strategy

1	root
2	/Singlets
3	/Singlets/45
4	/Singlets/45/Lymphocytes
5	/Singlets/45/Lymphocytes/CD3
6	/Singlets/45/Lymphocytes/CD3/4
7	/Singlets/45/Lymphocytes/CD3/4/CD25+
8	/Singlets/45/Lymphocytes/CD3/4/CD25+CD45RA+CCR7+
9	/Singlets/45/Lymphocytes/CD3/4/CD25+CD45RA+CCR7-
10	/Singlets/45/Lymphocytes/CD3/4/CD25+CD45RA-CCR7+
11	/Singlets/45/Lymphocytes/CD3/4/CD25+CD45RA-CCR7-
12	/Singlets/45/Lymphocytes/CD3/4/CD25-CD45RA+CCR7+
13	/Singlets/45/Lymphocytes/CD3/4/CD25-CD45RA+CCR7-
14	/Singlets/45/Lymphocytes/CD3/4/CD25-CD45RA-CCR7+
15	/Singlets/45/Lymphocytes/CD3/4/CD25-CD45RA-CCR7-
16	/Singlets/45/Lymphocytes/CD3/4/CD28+
17	/Singlets/45/Lymphocytes/CD3/4/CD28+CD45RA+CCR7+
18	/Singlets/45/Lymphocytes/CD3/4/CD28+CD45RA+CCR7-
19	/Singlets/45/Lymphocytes/CD3/4/CD28+CD45RA-CCR7+
20	/Singlets/45/Lymphocytes/CD3/4/CD28+CD45RA-CCR7-
21	/Singlets/45/Lymphocytes/CD3/4/28-CD45RA+CCR7+
22	/Singlets/45/Lymphocytes/CD3/4/28-CD45RA+CCR7-
23	/Singlets/45/Lymphocytes/CD3/4/28-CD45RA-CCR7+
24	/Singlets/45/Lymphocytes/CD3/4/28-CD45RA-CCR7-
25	/Singlets/45/Lymphocytes/CD3/4/CD45RA+
26	/Singlets/45/Lymphocytes/CD3/4/278+

27	/Singlets/45/Lymphocytes/CD3/4/CCR7+
28	/Singlets/45/Lymphocytes/CD3/4/HLADR+
29	/Singlets/45/Lymphocytes/CD3/4/PD1+
30	/Singlets/45/Lymphocytes/CD3/4/CD45RA+ICOS+CCR7+
31	/Singlets/45/Lymphocytes/CD3/4/CD45RA+ICOS+CCR7-
32	/Singlets/45/Lymphocytes/CD3/4/CD45RA+ICOS-CCR7+
33	/Singlets/45/Lymphocytes/CD3/4/CD45RA+ICOS-CCR7-
34	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7+
35	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7+HLADR+
36	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7+HLADR-
37	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7+PD1+
38	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7+PD1-
39	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7-
40	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7-HLADR+
41	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7-HLADR-
42	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7-PD1+
43	/Singlets/45/Lymphocytes/CD3/4/CD45RA+CCR7-PD1-
44	/Singlets/45/Lymphocytes/CD3/4/CD45RA-ICOS+CCR7+
45	/Singlets/45/Lymphocytes/CD3/4/CD45RA-ICOS+CCR7-
46	/Singlets/45/Lymphocytes/CD3/4/CD45RA-ICOS-CCR7+
47	/Singlets/45/Lymphocytes/CD3/4/CD45RA-ICOS-CCR7-
48	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7+
49	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7+HLADR+
50	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7+HLADR-
51	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7+PD1+
52	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7+PD1-
53	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7-
54	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7-HLADR+
55	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7-HLADR-
56	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7-PD1+

57	/Singlets/45/Lymphocytes/CD3/4/CD45RA-CCR7-PD1-
58	/Singlets/45/Lymphocytes/CD3/4/CD45RA-
59	/Singlets/45/Lymphocytes/CD3/4/treg
60	/Singlets/45/Lymphocytes/CD3/8
61	/Singlets/45/Lymphocytes/CD3/8/CD25+
62	/Singlets/45/Lymphocytes/CD3/8/CD25+CD45RA+CCR7+
63	/Singlets/45/Lymphocytes/CD3/8/CD25+CD45RA+CCR7-
64	/Singlets/45/Lymphocytes/CD3/8/CD25+CD45RA-CCR7+
65	/Singlets/45/Lymphocytes/CD3/8/CD25+CD45RA-CCR7-
66	/Singlets/45/Lymphocytes/CD3/8/CD25-CD45RA+CCR7+
67	/Singlets/45/Lymphocytes/CD3/8/CD25-CD45RA+CCR7-
68	/Singlets/45/Lymphocytes/CD3/8/CD25-CD45RA-CCR7+
69	/Singlets/45/Lymphocytes/CD3/8/CD25-CD45RA-CCR7-
70	/Singlets/45/Lymphocytes/CD3/8/CD28+
71	/Singlets/45/Lymphocytes/CD3/8/CD28+CD45RA+CCR7+
72	/Singlets/45/Lymphocytes/CD3/8/CD28+CD45RA+CCR7-
73	/Singlets/45/Lymphocytes/CD3/8/CD28+CD45RA-CCR7+
74	/Singlets/45/Lymphocytes/CD3/8/CD28+CD45RA-CCR7-
75	/Singlets/45/Lymphocytes/CD3/8/28-CD45RA+CCR7+
76	/Singlets/45/Lymphocytes/CD3/8/28-CD45RA+CCR7-
77	/Singlets/45/Lymphocytes/CD3/8/28-CD45RA-CCR7+
78	/Singlets/45/Lymphocytes/CD3/8/28-CD45RA-CCR7-
79	/Singlets/45/Lymphocytes/CD3/8/CD45RA+
80	/Singlets/45/Lymphocytes/CD3/8/278+
81	/Singlets/45/Lymphocytes/CD3/8/CCR7+
82	/Singlets/45/Lymphocytes/CD3/8/HLADR+
83	/Singlets/45/Lymphocytes/CD3/8/PD1+
84	/Singlets/45/Lymphocytes/CD3/8/CD45RA+ICOS+CCR7+
85	/Singlets/45/Lymphocytes/CD3/8/CD45RA+ICOS+CCR7-
86	/Singlets/45/Lymphocytes/CD3/8/CD45RA+ICOS-CCR7+

87	/Singlets/45/Lymphocytes/CD3/8/CD45RA+ICOS-CCR7-
88	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7+
89	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7+HLADR+
90	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7+HLADR-
91	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7+PD1+
92	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7+PD1-
93	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7-
94	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7-HLADR+
95	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7-HLADR-
96	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7-PD1+
97	/Singlets/45/Lymphocytes/CD3/8/CD45RA+CCR7-PD1-
98	/Singlets/45/Lymphocytes/CD3/8/CD45RA-ICOS+CCR7+
99	/Singlets/45/Lymphocytes/CD3/8/CD45RA-ICOS+CCR7-
100	/Singlets/45/Lymphocytes/CD3/8/CD45RA-ICOS-CCR7+
101	/Singlets/45/Lymphocytes/CD3/8/CD45RA-ICOS-CCR7-
102	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7+
103	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7+HLADR+
104	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7+HLADR-
105	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7+PD1+
106	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7+PD1-
107	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7-
108	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7-HLADR+
109	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7-HLADR-
110	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7-PD1+
111	/Singlets/45/Lymphocytes/CD3/8/CD45RA-CCR7-PD1-
112	/Singlets/45/Lymphocytes/CD3/8/CD45RA-

Table S5: Manual gating strategy applied to CITN-09 T cell panel.

A.11.2 CITN-09 Myeloid Manual Gating Strategy

1	root
2	/Singlets
3	/Singlets/45+
4	/Singlets/45+/CD3-CD19-
5	/Singlets/45+/CD3-CD19-/CD20-
6	/Singlets/45+/CD3-CD19-/CD20-/CD56-
7	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-
8	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-
9	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14+
10	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14+/Q1: CD33-, CD11B+
11	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14+/Q2: CD33+, CD11B+ (m-MDSC)
12	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14+/Q3: CD33+, CD11B-
13	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14+/Q4: CD33-, CD11B-
14	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15+
15	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15+/Q1: CD33-, CD11B+
16	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15+/Q2: CD33+, CD11B+ (PMN-MDSC)
17	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15+/Q3: CD33+, CD11B-
18	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15+/Q4: CD33-, CD11B-
19	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15-
20	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15-/Q1: CD33-, CD11B+

21	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15-/Q2: CD33+, CD11B+ (e-MDSC)
22	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15-/Q3: CD33+, CD11B-
23	/Singlets/45+/CD3-CD19-/CD20-/CD56-/HLADR-/CD16-/CD14-CD15-/Q4: CD33-, CD11B-
24	/Singlets/45+/CD14+
25	/Singlets/45+/CD14-CD15+
26	/Singlets/45+/CD14-CD15-

Table S6: Manual gating strategy applied to CITN-09 Myeloid panel.

A.11.3 CITN-07 Phenotyping Manual Gating Strategy

1	root
2	/Beads
3	/Non-beads
4	/Non-beads/Singlets
5	/Non-beads/Singlets/45+
6	/Non-beads/Singlets/45+/14+
7	/Non-beads/Singlets/45+/14-
8	/Non-beads/Singlets/45+/14-/3-19-
9	/Non-beads/Singlets/45+/14-/3-19-/56-16-
10	/Non-beads/Singlets/45+/14-/3-19-/56-16-/Basophils
11	/Non-beads/Singlets/45+/14-/3-19-/56-16-/Basophils/HLA DR hi
12	/Non-beads/Singlets/45+/14-/3-19-/56-16-/Basophils/HLA DR med
13	/Non-beads/Singlets/45+/14-/3-19-/56-16-/Basophils/HLA DR neg
14	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+
15	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/mDC
16	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/mDC/HLADRhi
17	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/mDC/HLADRmed
18	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/mDC/HLA DR hi
19	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/mDC/HLA DR med
20	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/mDC/HLA DR neg
21	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/pDC
22	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/pDC/HLA DR hi
23	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/pDC/HLA DR med
24	/Non-beads/Singlets/45+/14-/3-19-/56-16-/HLADR+/pDC/HLA DR neg
25	/Non-beads/Singlets/45+/Lymphocytes
26	/Non-beads/Singlets/45+/Lymphocytes/3+
27	/Non-beads/Singlets/45+/Lymphocytes/3+/4&8
28	/Non-beads/Singlets/45+/Lymphocytes/3+/4&8/56+
29	/Non-beads/Singlets/45+/Lymphocytes/3+/4&8/122+

30	/Non-beads/Singlets/45+/Lymphocytes/3+/4&8++
31	/Non-beads/Singlets/45+/Lymphocytes/3+/4+
32	/Non-beads/Singlets/45+/Lymphocytes/3+/4+/HLADR+
33	/Non-beads/Singlets/45+/Lymphocytes/3+/8+
34	/Non-beads/Singlets/45+/Lymphocytes/3+/8+/HLADR+
35	/Non-beads/Singlets/45+/Lymphocytes/3-19-
36	/Non-beads/Singlets/45+/Lymphocytes/3-19-/16+56-
37	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56+
38	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56+/122+
39	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56+16-
40	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56-16-
41	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56-16-/122+
42	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56-16-/HLA DR hi
43	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56-16-/HLA DR med
44	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56-16-/HLA DR neg
45	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56B
46	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56B/HLA DR hi
47	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56B/HLA DR med
48	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56B/HLA DR neg
49	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56B16-
50	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56D
51	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56D/HLA DR hi
52	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56D/HLA DR med
53	/Non-beads/Singlets/45+/Lymphocytes/3-19-/56D/HLA DR neg
54	/Non-beads/Singlets/45+/Lymphocytes/19+
55	/Non-beads/Singlets/45+/Lymphocytes/19+/B CELLS
56	/Non-beads/Singlets/45+/Lymphocytes/19+/B CELLS/HLA DR hi
57	/Non-beads/Singlets/45+/Lymphocytes/19+/B CELLS/HLA DR med
58	/Non-beads/Singlets/45+/Lymphocytes/19+/B CELLS/HLA DR neg

Table S7: Manual gating strategy applied to CITN-07 phenotyping panel.

A.12 Myeloid compartment analysis

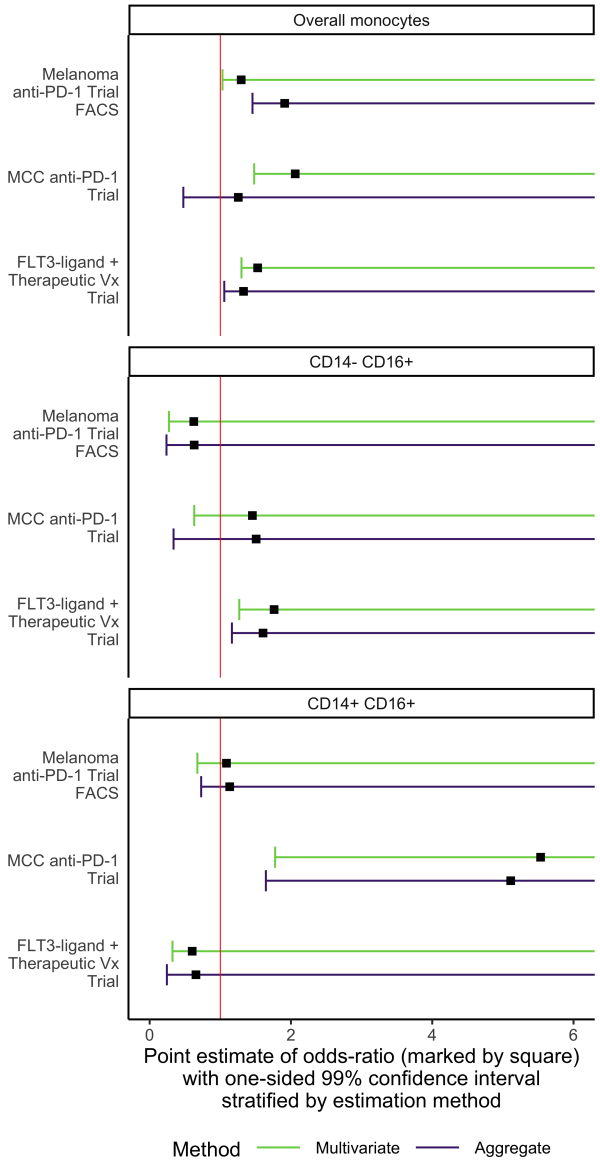

Figure S5: This figure contains results for the remaining compartments from the multivariate and aggregate myeloid compartment analysis described in section 2.4. Here we see the multivariate modeling also reveals evidence of increased abundance in responders across the entire myeloid compartment.

A.13 Simulation study

A.14 Summary of simulation results

To better understand the performance of FAUST relative to other methods, we conducted simulation studies that generated data from a variety of mixture models. Since FAUST assumes that each experimental unit is sampled from a finite mixture model (see Methods 4), all datasets generated in the study were designed to be compatible with the statistical assumptions underpinning FAUST. Components of the mixture for each experimental unit are assumed to arise from a common class of densities, with batch effects and other sources of experimental heterogeneity modeled as unit-specific changes in location and scale of the underlying mixture components. The mixture components represent cell sub-populations within a unit.

The study generated datasets from a variety of mixture models incorporating different combinations of assumptions, detailed in the following sub-sections. The study begins by simulating data from multivariate Gaussian distributions (producing datasets which are favorable to many existing methods) and progressively simulates data that more closely represents flow cytometry and CyTOF datasets. In the study, we compare FAUST to FlowSOM since, as noted in the main text, FlowSOM is computationally efficient, is recommended in the review [4], and is used in the recent *diffCyt* method [18]. Each simulated mixture component (representing a cell sub-population) is partially parameterized by a mean vector and is given a phenotypic label that describes the phenotype of the component. By treating these phenotypic labels as ground-truth, we are able to measure how well the count matrix produced by FAUST agrees with the simulated count matrix, matching discovered and simulated cell populations based on their phenotypes. FAUST is run completely unsupervised across all simulation settings.

Our results demonstrate that FAUST's discovery and annotation strategy does not severely over partition the data under a variety of generative regimes (supplementary figure S11). Results also show that the cell counts derived from FAUST's discovered clusters strongly correlate with the underlying true counts across all simulation settings. We observe a median correlation of 0.91 between FAUST and the simulated truth, when cluster counts are correlated between FAUST clusters and the ground truth using only cluster annotations to perform the comparison (Supplementary Figure S10).

The simulated datasets always include a sub-population that is differentially abundant between

50% of the subjects. Our results show that when we simulate a causal relationship of varying strength between this differential sub-population and a simulated response to therapy, FAUST discovers the differential sub-population, annotates it correctly, and often identifies that the differential population is associated with response to therapy (Supplementary Figures S12, S13, S14).

In the present simulation, FlowSOM clusters are tested for differential abundance under the same causal regimes as FAUST. Our results show that FlowSOM's ability to detect the causal association is adversely affected when the simulation departs from multivariate normality or when the simulated data contains 50 true clusters and batch effects and/or nuisance variables, even when FlowSOM is provided with the true number of clusters as a tuning parameter. As noted in the main text, this study confirms our empirical finding that FAUST robustly detects signals in data that are not found by other discovery methods.

A.14.1 Simulation Goals

The purpose of this simulation study is to assess the performance of the FAUST algorithm, both as a clustering tool and as a discovery tool. datasets are simulated from mixture models following the assumptions of section 4.1. The simulation measures how well FAUST recovers the underlying mixture under a variety of parametric scenarios. The simulation also measures how well FAUST is able to detect a sub-population, elevated in half the samples, that is required to have causal relationship (of varying strength) with a subject's response to therapy. We compare the performance of FAUST to the performance of the FlowSOM clustering algorithm [5].

A.14.2 Baseline simulation description

The basic simulation generates an experimental data collection containing 100 independent samples of 10-dimensional data from a Gaussian mixture model with 10 components. A probability vector

$$\mathbf{p} \sim \text{Dirichlet}\left(\alpha \equiv (1, 1, \ldots, 1)\right) \tag{A.7}$$

of dimension equal to the number of mixture components is generated. In a given simulation iteration, sampling from the Dirichlet continues until all elements are greater or equal to 0.001.

There are four tuning parameters that modify this baseline setting. We will first give a complete

description of how the simulation study works at baseline, and then will describe how the tuning parameters modify the baseline study.

In the basic setting, the size of each of the 100 samples is $n_j = \max(5000, s)$, $1 \leq j \leq 100$, where $s \sim T(\mu = 10000, \nu = 3)$ is a sample from a non-central T distribution with 3 degrees of freedom and non-centrality parameter 10000. Each sample is meant to represent a sample taken from a subject in an immunology study and then interrogated via flow cytometry.

Before generating the samples, a fixed collection of mean vectors μ_c, $1 \leq c \leq 10$ is determined for the ten Gaussian mixture components that is used across all simulated samples. Each of the ten entries of μ_c are randomly selected from the columns of table S8, and represent whether or not the measured variable exhibits a signal. When an entry of μ_c is from the "No Signal" row of table S8, the corresponding variable is labeled "-". Similarly, when an entry of μ_c is from the "Signal" row of table S8, the corresponding variable is labeled "+". An an example, the annotation "V1- V2- V3+ V4- V5+ V6- V7- V8- V9+ V10-" indicates the mean vector μ_c of the mixture component contains 0 for V1, V2, V4, V6, V7, V8, and V10, while it is 7 for V3, 6 for V5, and 4 for V9. Each mean vector is associated with an element of the probability vector (A.7). Covariance matrices Σ_c are always constrained to have variances between 1 and 2, but otherwise are randomly generated sample-by-sample and component-by-component.

Table S8: Possible mean vector entries for the ten simulation variables.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
No Signal	0	0	0	0	0	0	0	0	0	0
Signal	8	8	7	7	6	6	5	5	4	4

Each simulation iteration, 50 of the 100 samples are randomly selected to have a mixture component elevated. Without loss of generality, suppose (A.7) is in sorted order, so that the first entry p_1 is the largest value, the tenth entry p_{10} is the smallest value, and intermediate entries correspond to their order statistics. In the non-elevated samples, the mean-vector μ_c associated with the smallest element of the probability vector (A.7), p_{10}, is identified as the cluster component to elevate. In the samples randomly selected for elevation, the probability vector (A.7) is modified

as follows. The numerical value $p_{\text{target}} \equiv p_7$ is fixed. Next, the intermediate probability vector

$$\begin{aligned} \mathbf{p_{int}} &\equiv \left(p_1 + \frac{p_{10}}{9}, p_2 + \frac{p_{10}}{9}, \ldots, p_9 + \frac{p_{10}}{9}, 0\right) \\ &\equiv (q_1, q_2, \ldots, q_9, 0) \end{aligned} \tag{A.8}$$

is generated. Then (A.8) is modified so that

$$\begin{aligned} \mathbf{p_{elevated}} &\equiv \left(q_1 - q_1 \cdot p_{\text{target}}, q_2 - q_2 \cdot p_{\text{target}}, \ldots, q_9 - q_9 \cdot p_{\text{target}}, p_{\text{target}}\right) \\ &\equiv (r_1, r_2, \ldots, r_9, r_{10}) \, . \end{aligned} \tag{A.9}$$

The transformation from (A.7) to (A.9) causes the identified population to be, on average, the 7^{th} largest mixture component in half the samples, and the smallest mixture component in the other half.

A sample of size n_j with $1 \leq j \leq 100$ is generated by first determining the relative size of each mixture component within the sample. When the sample is selected as having the elevated population, the size of mixture components is determined by taking a sample from a multinomial distribution with n_j trials and cell probabilities determined by (A.9). Otherwise, the size of mixture components is determined by taking a sample from a multinomial distribution with n_j trials and cell probabilities determined by (A.7). In both cases, the resulting multinomial vector is then used to sample multivariate Gaussian samples of the corresponding size, with mean vectors $\mu_c + e_{c,j}$ and covariance matrices Σ_c, for $1 \leq c \leq 10$. The vector $e_{c,j} = (e_{c,j,1}, \ldots, e_{c,j,10})$ is determined by taking a 10 independent samples $\epsilon_{c,j,k} \sim N(0, 1/2)$, $1 \leq k \leq 10$, and then rounding $e_{c,j,k} = round(\epsilon_{c,j,k})$ to the nearest integer. The vector $e_{c,j}$ models sample-specific perturbations (corresponding to subject-level effects) without modifying (with high probability) the semantic interpretation of the annotations corresponding to μ_c. A visualization of the baseline experiment is provided in figure S6.

Once the experimental data is generated, it is processed by FAUST in a completely unsupervised setting. FAUST is set to use individual samples as the experimental unit. All simulated variables are taken as admissible and the channel boundaries are set to the entire real line for all markers. The depth score selection threshold is set to 0.01, the depth score selection quantile is set to the median, and the phenotype occurence number is set to 25. The 100 samples are also concatenated and

clustered by the FlowSOM algorithm in two different ways. First, following the recommendation of [49], the FlowSOM grid is set to $1 \times$ Number of mixture components to simulate one best case scenario: an oracle provides FlowSOM with the true number of clusters. Second, similar to the approach of [18], FlowSOM overpartitions the data by setting the grid to 5×5 (assuming 25 clusters when in truth there are 10).

To test how well each of the three methods discover sub-populations associated with differential abundance, a binary response is generated for each sample in the experiment. For samples where the identified population is elevated, a probability of response $p_{response}$ is varied from 0.50 to 0.80 in increments of 0.05. Each elevated sample is then associated with a response status by sampling from a Bernoulli$(p_{response})$. Similarly, samples where the identified population is not elevated are given a probability of response $q_{response} \equiv 1 - p_{response}$. Each non-elevated sample is then associated with a response status by sampling from a Bernoulli$(q_{response})$.

Once samples are associated with a binary outcome, the clusters produced by each of the three approaches are tested for differential abundance following the strategy described in section A equation (4.5). P-values are adjusted for FDR (q-values) using the method [24]. In the event FAUST discovers the elevated population by exact annotation, the associate q-value is recorded. For FlowSOM, the "best" q-value is defined as follows. Both the cluster containing the largest number of observations from the elevated population in terms of absolute counts, and the cluster containing proportionally the most observation from the elevated population are identified. The minimum q-value from the two clusters (when different) is recorded for both the oracle FlowSOM and overpartitioned FlowSOM clusterings.

We repeat this modeling procedure 50 times for each setting of $p_{response}$. The median q-values across each of the 50 iterations is recorded in a single simulation iteration. We then repeat the entire experimental simulation 50 times, and report the median of median q-values across those 50 simulation runs. In addition, we compute F-measures of the clusterings, along with several other measures of the quality of the FAUST clusterings. We will describe these measurements in the coming figures. Before doing so, we will provide details about simulation tuning parameters.

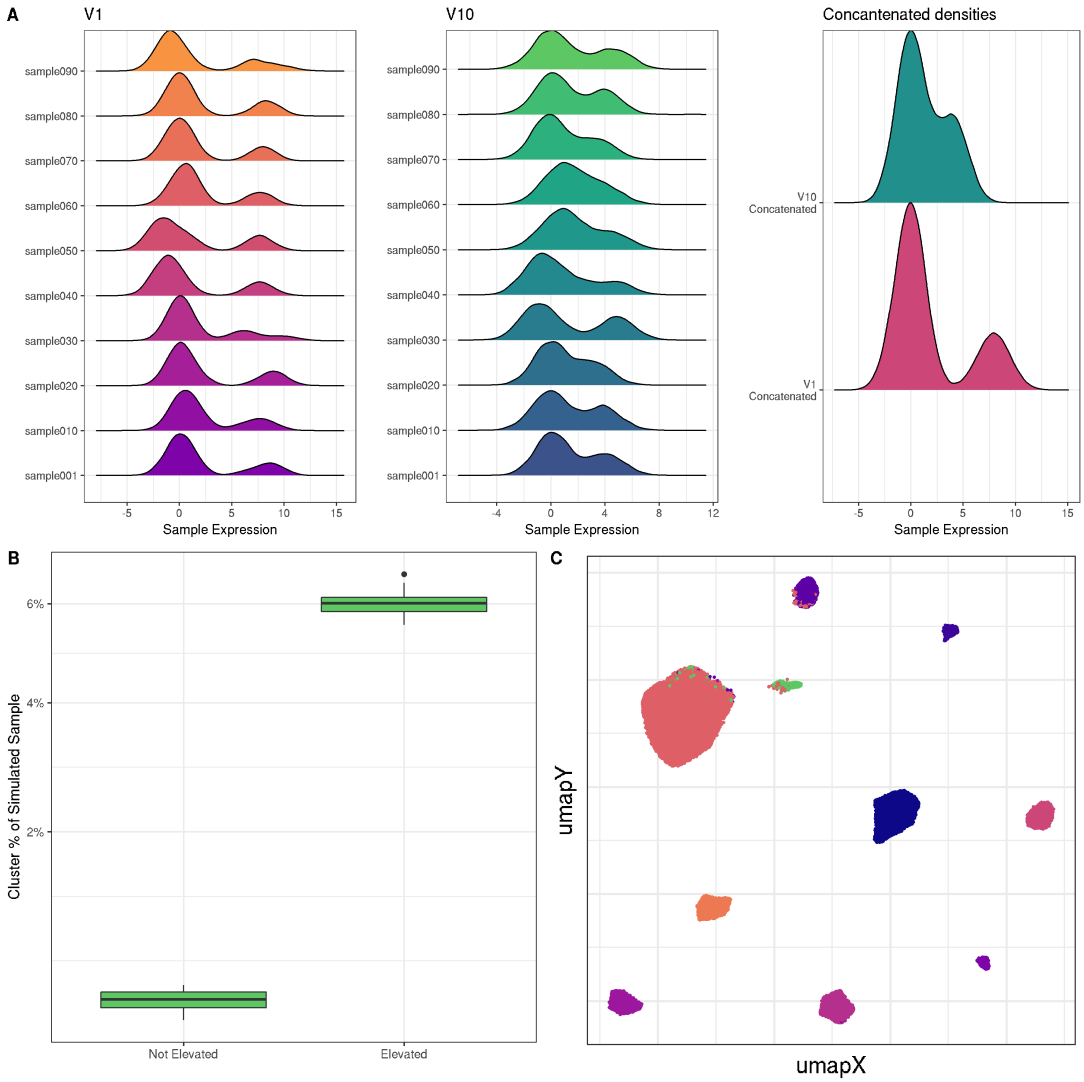

Figure S6: Visual summary of baseline simulation. Panel A shows the best separated variable (V1), worst separated (V10), and their concatenation across 10 samples. Panel B shows the elevated population across the entire 100 sample experiment. Panel C shows the umap generated from the 10 concatenated samples.

A.14.3 Simulation tuning parameters

The first simulation parameter we vary is the underlying number of mixture components: we set this parameter to 25 components and 50 components, in addition to the baseline of 10. While the sample sizes are random, we do not change the underlying sampling scheme, which introduces rarer and rarer populations appear across simulations as the number of mixture components increase. In both the 25 and 50 component setting, the probability vector (A.7) is expanded

accordingly; α continues to be set to 1 for each component. In all cases, sampling from the Dirichlet continues until all elements are greater or equal to 0.001. In the 25 component setting, the elevated population has p_{target} set to p_{18}; in the 50 component setting, p_{target} is set to p_{35}.

The second simulation parameter we vary is used to add a batch effect to the simulation. The batch effect is modeled as a translation of the underlying mean vector. Batches are modeled as groups of 10 samples. After the initial 10 samples are generated, the mean vectors of the Gaussian mixtue components (sampled from (S8)) are translated by a constant vector $\lambda_1 = (1/3, 1/3, \ldots, 1/3)$. After the next 10 samples are generated, the translate increases to the constant vector $\lambda_2 = (2/3, 2/3, \ldots, 2/3)$. This continues in groups of 10 until the final 10 samples are translated by $\lambda_9 = (9/3, 9/3, \ldots, 9/3)$. Figure S7 illustrates an example of a simulated experiment with 50 mixture components and the batch effect parameter turned on.

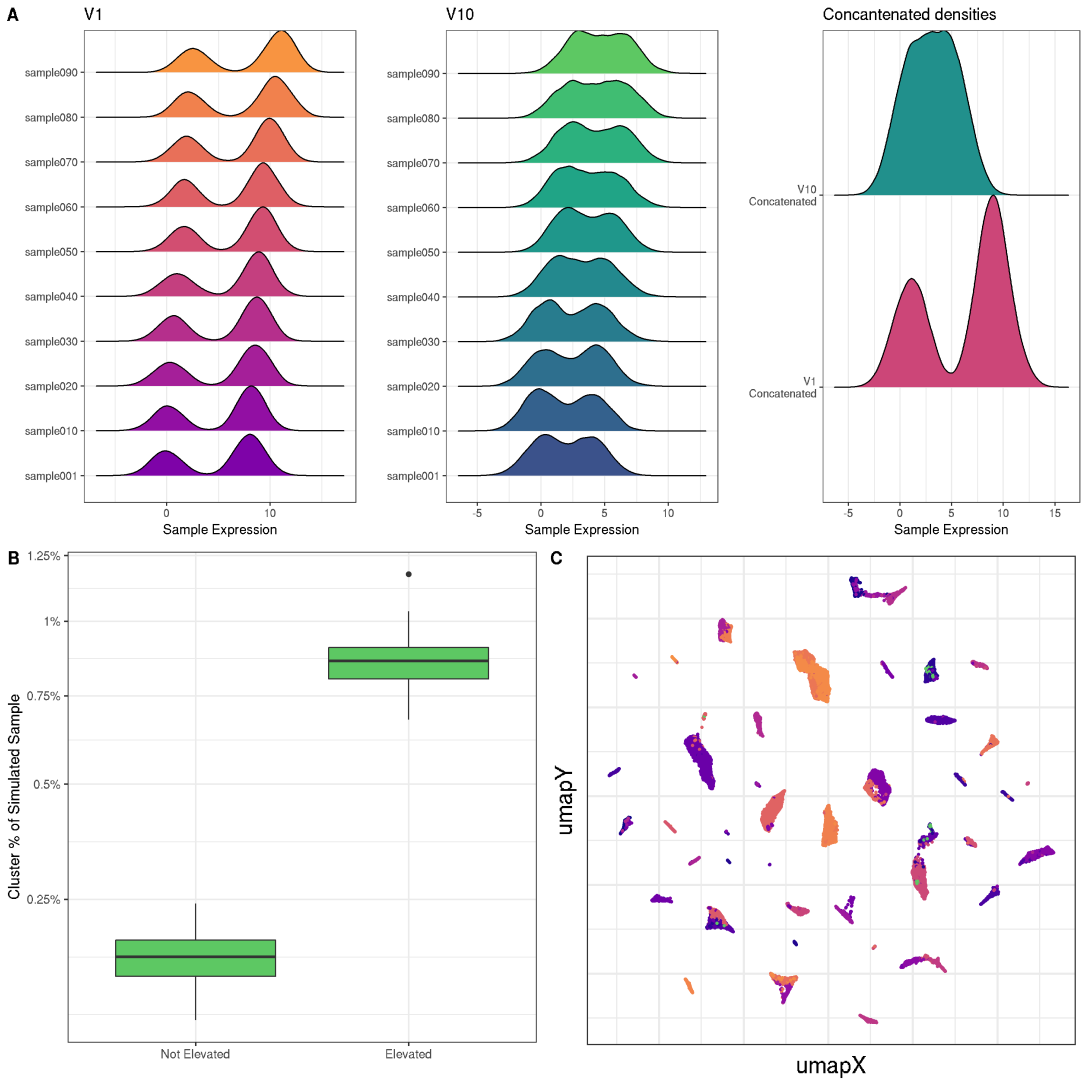

Figure S7: Visual summary of simulation modified from baseline with 50 mixture components and a batch effect turned on. Panel A shows the best separated variable (V1), worst separated (V10), and their concatenation across 10 samples. Panel B shows the elevated population across the entire 100 sample experiment. Panel C shows the umap generated from the 10 concatenated samples.

The third simulation parameter controls whether or not we add nuisance variables to the simulation. This parameter is meant to generate data under the scenario that several markers in the panel are uninformative because of staining issues. When this parameter is turned on the following occurs. Each time a sample of size n_j is generate, an independent sample of size n_j it taken from a Multivariate Gaussian distribution centered at $\mu_{\text{nuisance}} = (5, 5, 5, 5, 5)$, and Σ_{nuisance}

constrained to have variances between 1 and 2 but otherwise random. The independent Gaussian sample is then adjoined to the mixture of size n_j, producing a simulated dataset in 15 dimensions. Since nuisance variables are independently generated, they do not affect the mixture structure of a given simulation; consequently, Consequently, the underlying annotations of observations by their cluster component mean vector are not changed when the nuisance variables are added to the simulation.

The final simulation parameter is used to investigate departures from normality. We explored two possible settings: after generating each sample, the data are transformed coordinate-by-coordinate through the square map $f(x) = x^2$ or the gamma map $g(x) = \Gamma(1 + |(|x/4))$. The square map was used to investigate a mild departure from Normality, while we used the gamma map to transform the mixture into data that looked similar to CyTOF. Under the gamma map, we modify the space possible Gaussian mean vectors (S8) to those determined by table (S9). Figure S8 illustrates an example of a simulated experiment with 25 mixture components, both the batch effect parameter and nuisance variable parameters turned on, and data are transformed by the Gamma map.

Table S9: Possible mean vector entries for the ten simulation variables when data subsequently transformed by the map $g(x) = \Gamma(1 + |(|x/4))$.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
No Signal	0	0	0	0	0	0	0	0	0	0
Signal	8	8	8	7	7	7	7	6	6	6

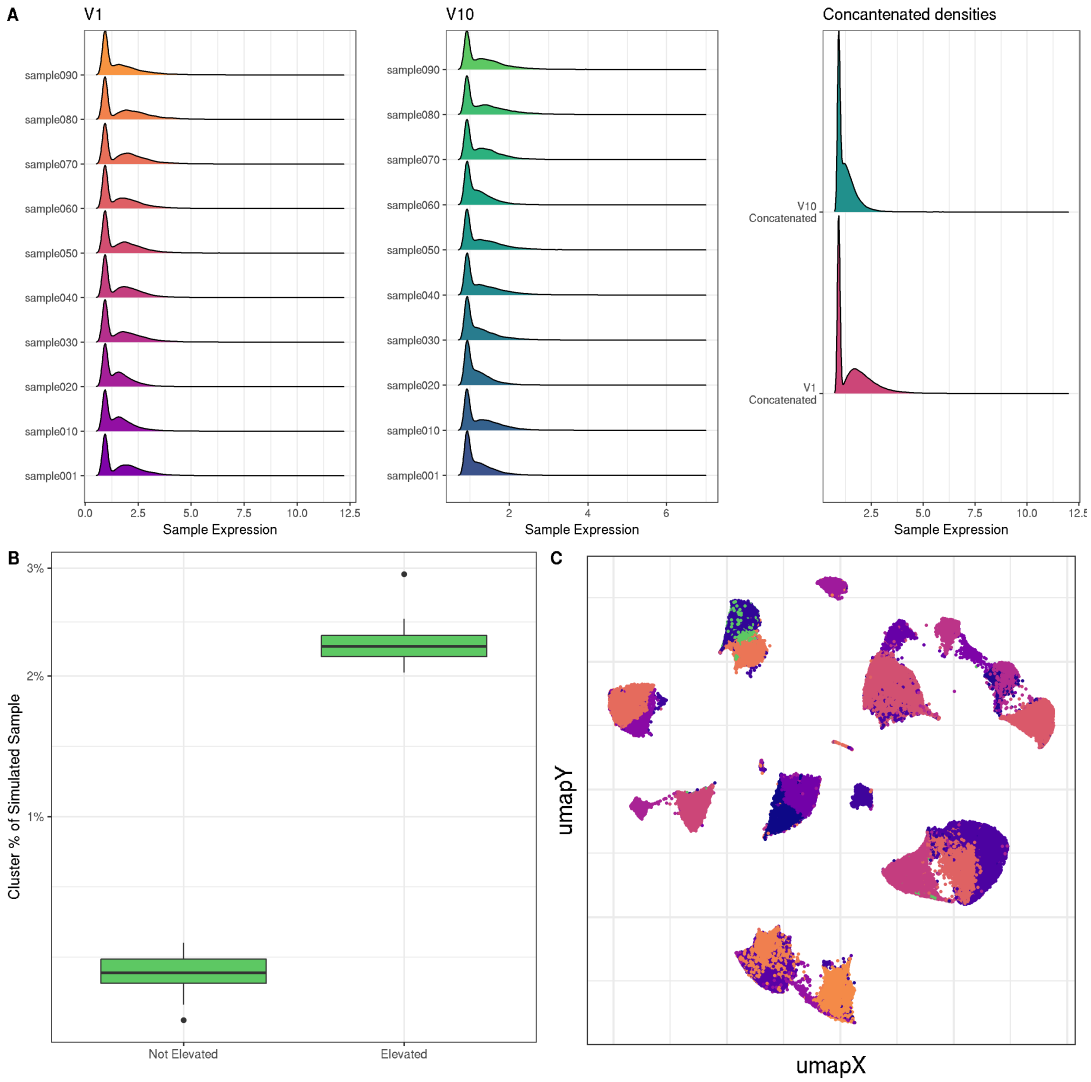

Figure S8: Visual summary of simulation modified from baseline with 25 mixture components, batch effect turned on, nuisance variable turned on, and data transformed coordinate-wise by the map $g(x) = \Gamma(1 + |(|x/4))$ after generation. Panel A shows the best separated variable (V1), worst separated (V10), and their concatenation across 10 samples. Panel B shows the elevated population across the entire 100 sample experiment. Panel C shows the umap generated from the 10 concatenated samples.

A.14.4 Simulation results

By adjusting the tuning parameters described in supplementary section A.14.3, we explore 36 distinct scenarios *in silico*. Each simulation setting is run 50 times, with three exceptions which we now report. The scenario of 25 Clusters with no batch effect but with nuisance variables

transformed by the gamma map completed 34 iterations. The scenario of 25 Clusters with batch effect but with no nuisance variables transformed by the gamma map completed 37 iterations. The scenario of 50 Clusters with batch effect and with nuisance untransformed (the identity map) completed 35 iterations. Based on their log files, these three scenarios did not complete 50 iterations in 7 days of compute time due to generating experiments in which the regression modeling took unusually long to fit to each cluster.

This simulation study shows that departures from multivariate-normality as well as batch-effects combined with large numbers of clusters impair FlowSOM's ability to define clusters that correlate with outcome. FAUST, on the other hand, performed robustly across simulation settings since its key methodological assumption is that some subset of the measured markers in a cytometry dataset are marginally separated into modal groups. In samples that both contain heterogeneous cell populations (such as live lymphocytes) and are stained by a large marker panel, we have empirically seen this is assumption is always met. Plots of the observed expression data show the MCC anti-PD1 dataset has non-Gaussian characteristics, and also has sample-to-sample variation which is common in many cytometry experiments. Hence, the non-Gaussian nature of the MCC anti-PD1 trial data combined with sample-to-sample variation both contribute to the discovery differences observed between FlowSOM and FAUST.

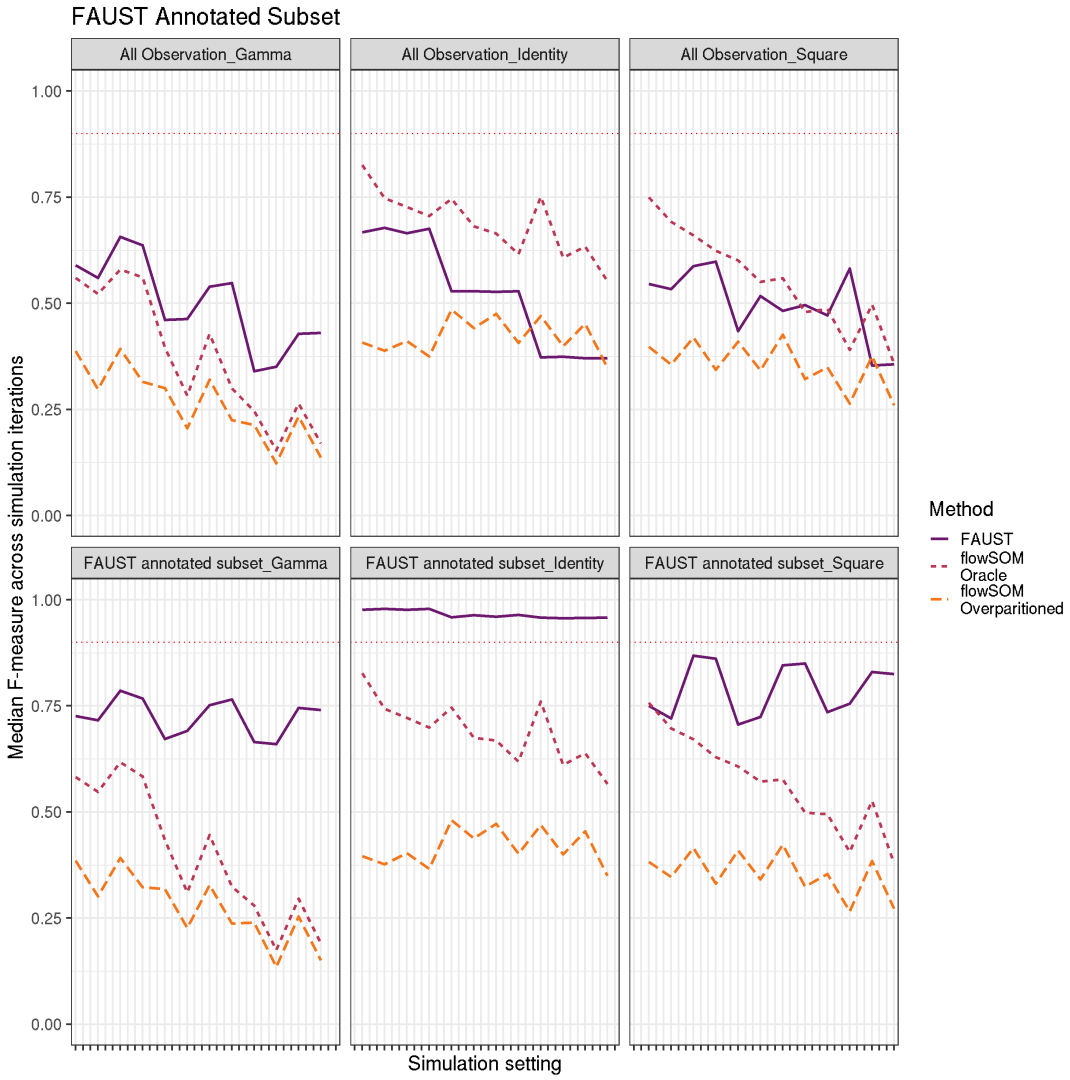

Figure S9: The median F-measure across the 50 simulation iterations (with 3 exceptions). Results are stratified by transformation type: $h(x) = x$ (identity map); $f(x) = x^2$ (square map); $g(x) = \Gamma(1 + |(|x/4))$ (gamma map). F-measures are computed between each method's clustering and the entire simulated dataset (row 1). F-measures are also computed between each method and the subset of observations that FAUST annotates (row 2). The figures show FAUST improves markedly (in terms of F-measure) on the set of labeled observations it labels, while the F-measure of FlowSOM with an oracle and FlowSOM overpartitioned perform similarly on the two sets. This figure provides a demonstration of the difficulty of comparing FAUST clusterings to computational methods in current use: classic measures of clustering performance, such as the F-measure, do not directly account for the biological information present in FAUST annotations. We have observed similiar trends in other clustering metrics, such as the adjusted rand index (data not shown). When the annotated subset is compared to associated subset of the ground truth, FAUST's performance improves markedly in terms of F-measure, while FlowSOM shows no noticeable improvement on the subset.

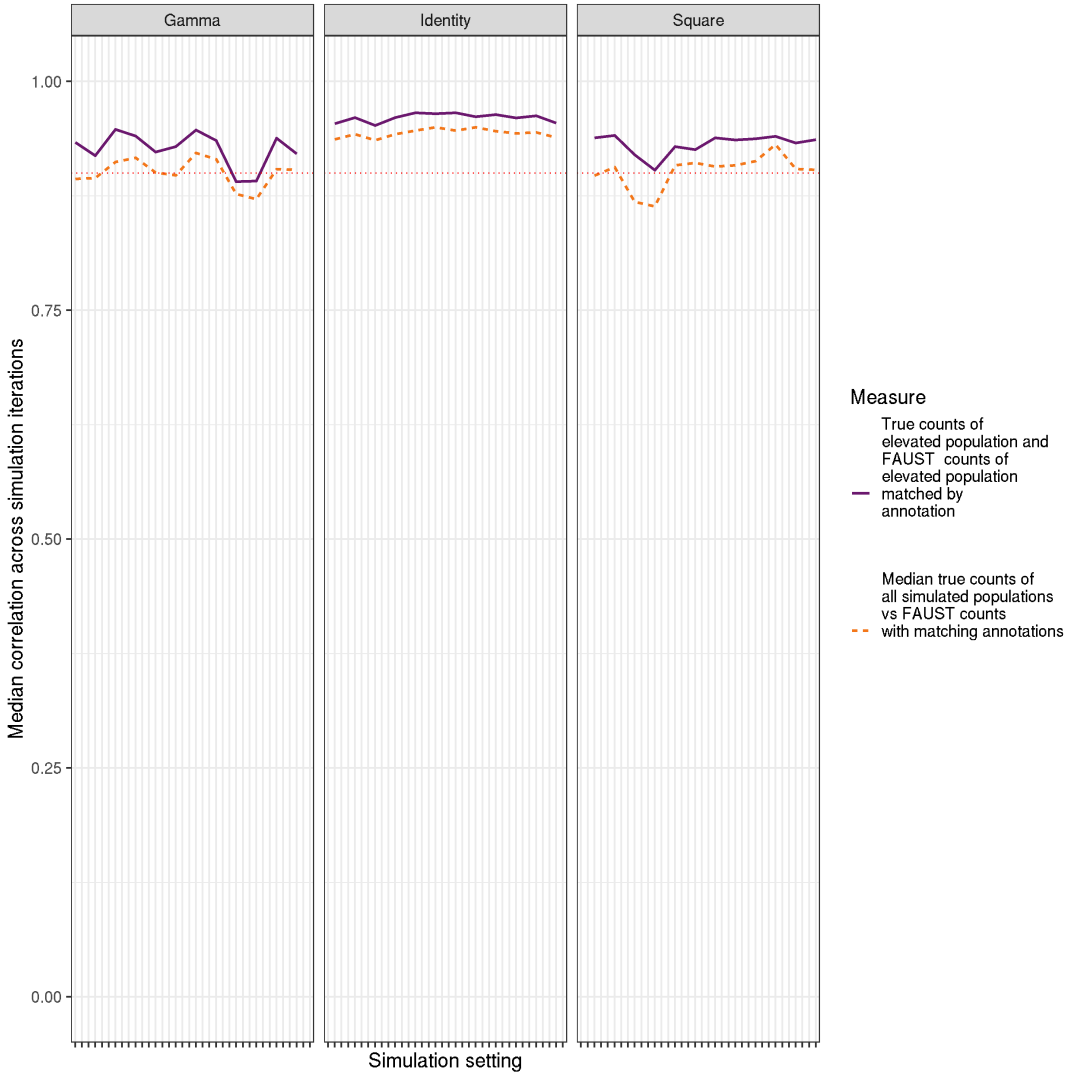

Figure S10: Dashed line: median correlation across 50 simulation between all FAUST clusters and all simulated true populations Solid line: median correlation across 50 simulations between FAUST cluster with differential abundant population and simulated differentially abundant cluster. Correlations are determined only using annotations.

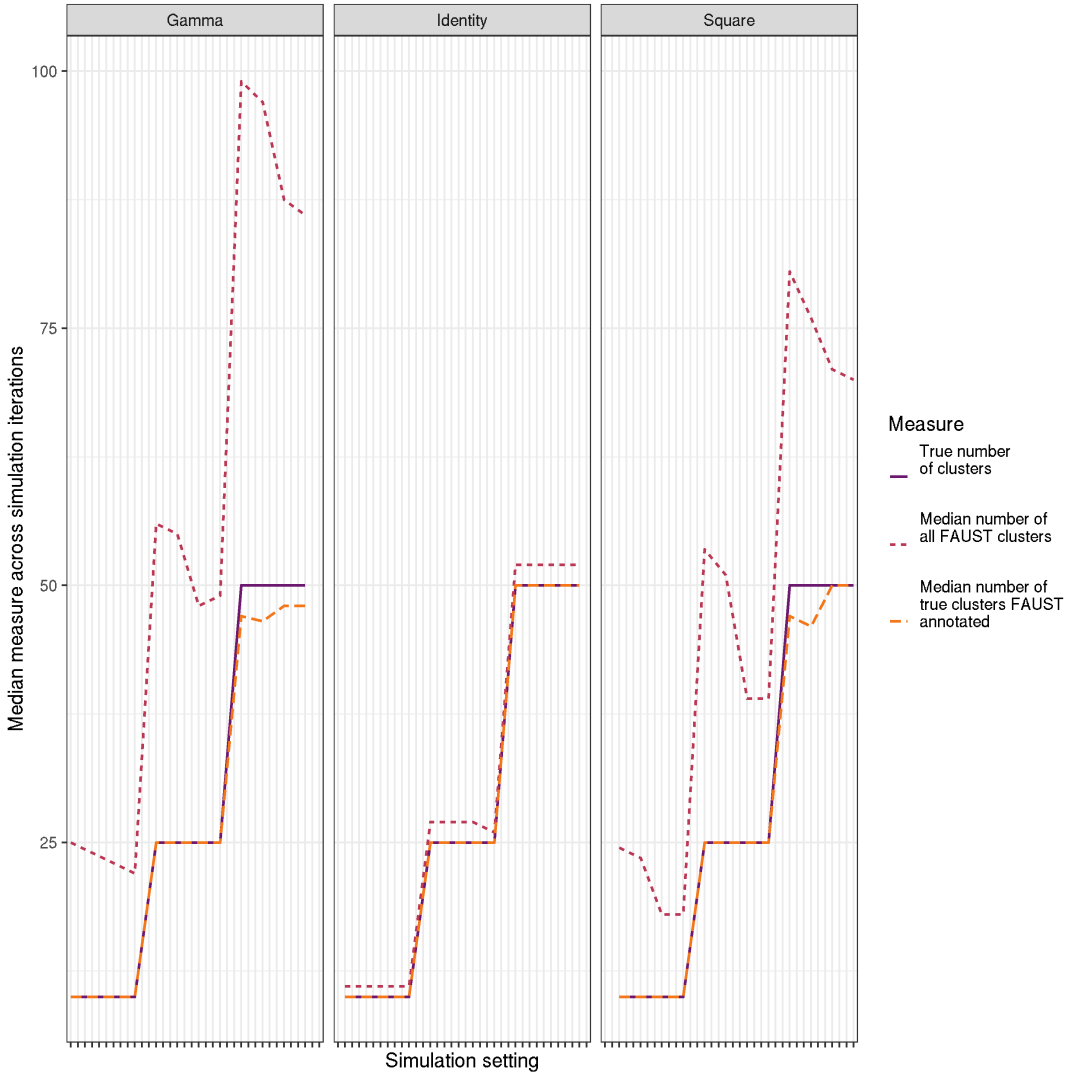

Figure S11: The true number of clusters by simulation setting is the solid purple line. The dashed orange line shows the median number of clusters matching the true annotations produced by FAUST across simulation settings. The dot-dashed red line show the median number of total annotated clusters produced by FAUST across simulation settings.

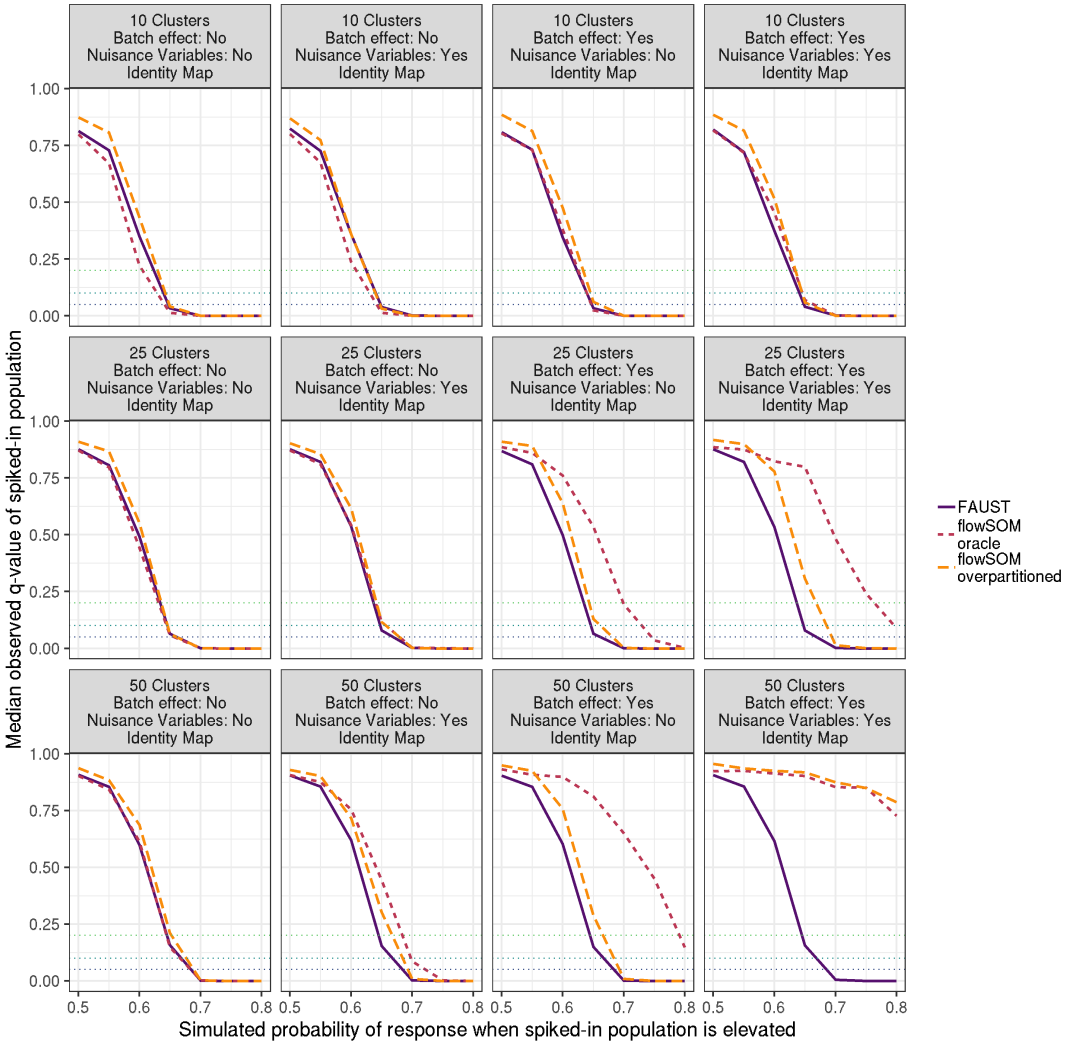

Figure S12: In each simulation, a differentially abundant sub-population is always simulated: 50 subjects have increased abundance relative to the other 50 subjects. For subjects with increased abundance, a stochastic response to therapy is then 50 times, with the response rate for subjects with increased abundance varying along the x-axis. Median FDR-adjusted p-value of FAUST cluster annotated with the differentially abundant population, and median FDR-adjusted for FlowSOM clusters identified as the true clusters are reported across 50 iterations. This plot reports performance when data are generated from a multivariate normal mixture, with different simulation settings.

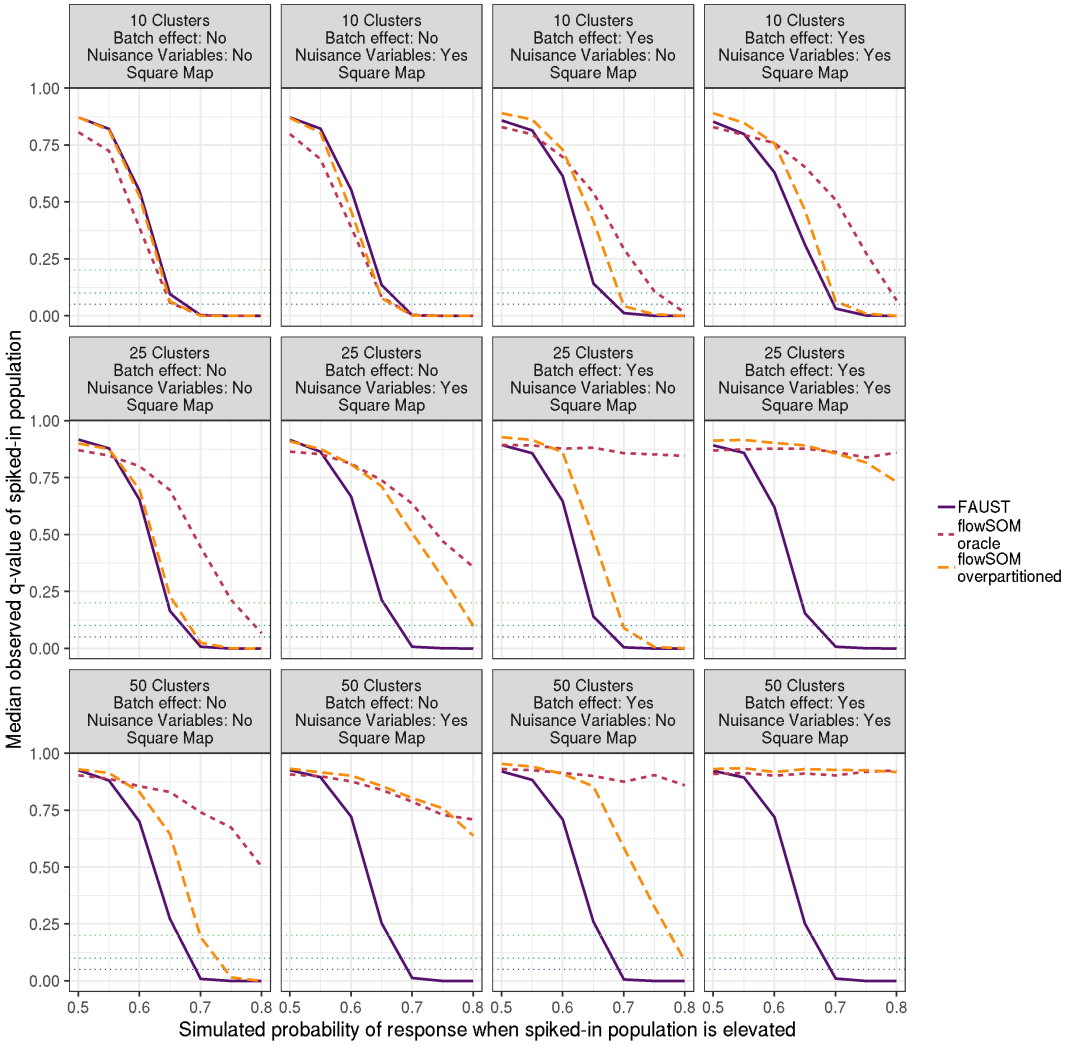

Figure S13: In each simulation, a differentially abundant sub-population is always simulated: 50 subjects have increased abundance relative to the other 50 subjects. For subjects with increased abundance, a stochastic response to therapy is then 50 times, with the response rate for subjects with increased abundance varying along the x-axis. Median FDR-adjusted p-value of FAUST cluster annotated with the differentially abundant population, and median FDR-adjusted for FlowSOM clusters identified as the true clusters are reported across 50 iterations. This plot reports performance when data are transformed by the coordinate map $f(x) = x^2$, with different simulation settings.

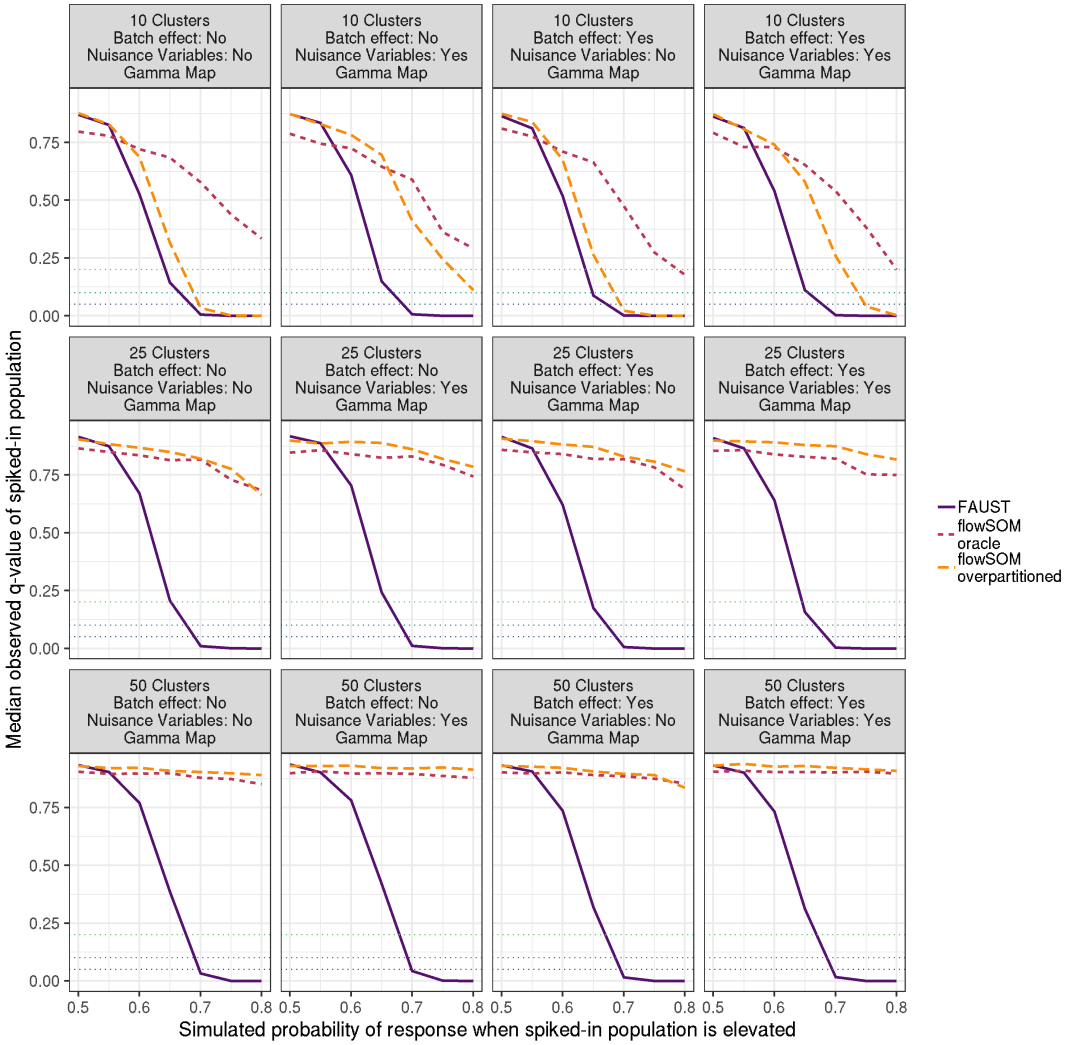

Figure S14: In each simulation, a differentially abundant sub-population is always simulated: 50 subjects have increased abundance relative to the other 50 subjects. For subjects with increased abundance, a stochastic response to therapy is then 50 times, with the response rate for subjects with increased abundance varying along the x-axis. Median FDR-adjusted p-value of FAUST cluster annotated with the differentially abundant population, and median FDR-adjusted for FlowSOM clusters identified as the true clusters are reported across 50 iterations. This plot reports performance when data are transformed by the coordinate map $g(x) = \Gamma(1 + |x/4|)$, with different simulation settings