A Systems-Level Analysis of Common Variable Immunodeficiency

University of California, Los Angeles
Department of Microbiology, Immunology, and Molecular Genetics
Department Honors Thesis

Humza Ali Khan
Butte Laboratory

Table of Contents

I. Abstract

Common variable immunodeficiency (CVID) is a disorder arising from a host of monogenic lesions and is characterized by defective antibody production. Due to its variable etiology, the manifestations of CVID are diverse—ranging from susceptibility to infection to autoimmunity to exaggerated states of inflammation. Patients with autoimmunity require aggressive treatment, yet seldom receive it promptly. The field has yet to devise a method to predict which patients will present with autoimmunity and require additional treatment. We aim to integrate genetic analyses of CVID patients with functional analyses of cell signaling to group patients into autoimmune or non-autoimmune categories. To elucidate the phenotypic and signaling signatures in patients with CVID, we conducted stimulation assays with phospho-protein mass cytometry and higher-dimensional data analytics. Significantly, our mass cytometry panels identify all known circulating immune cell subsets. We created a novel two-component Gaussian mixture model approach to model the populations of responding and non-responding cells upon stimulation. Our initial statistical analyses demonstrated that eosinophils from CVID patients had defective gain-of-function responses to TLR1/2 stimulation. Furthermore, we found that CD16lo monocytes had inappropriate gain-of-function responses to stimulation with PMA/Ionomycin. We also found higher PD-1 expression in the effector CD8 T cells of patients. Our approach will lead to better characterization of signaling in CVID and will potentially allow better classification of patients with this disorder. We also expect that a better understanding of the signaling defects in the circulating immune cells of CVID patients will lead to new therapeutic approaches.

II. Introduction and Literature Review

Primary immunodeficiencies (PIDs) are diseases caused by genetic lesions that impair immune system function. PIDs vary widely in their genetic etiologies and, therefore, in their cellular deficiencies and clinical manifestations (Picard et al, 2018). Genomic sequencing has become indispensable in diagnosing these disorders. When a patient is suspected to have a PID, their DNA is often sent for sequencing of genes relevant to immune function (Mafucci et al, 2016). For example, Invitae, a medical genetics corporation, has a clinically approved PID panel that interrogates 207 PID-linked genes. The "hits" found by PID sequencing panels provide a roadmap for a rational investigation of the genetic lesions that could conceivably manifest as a PID.

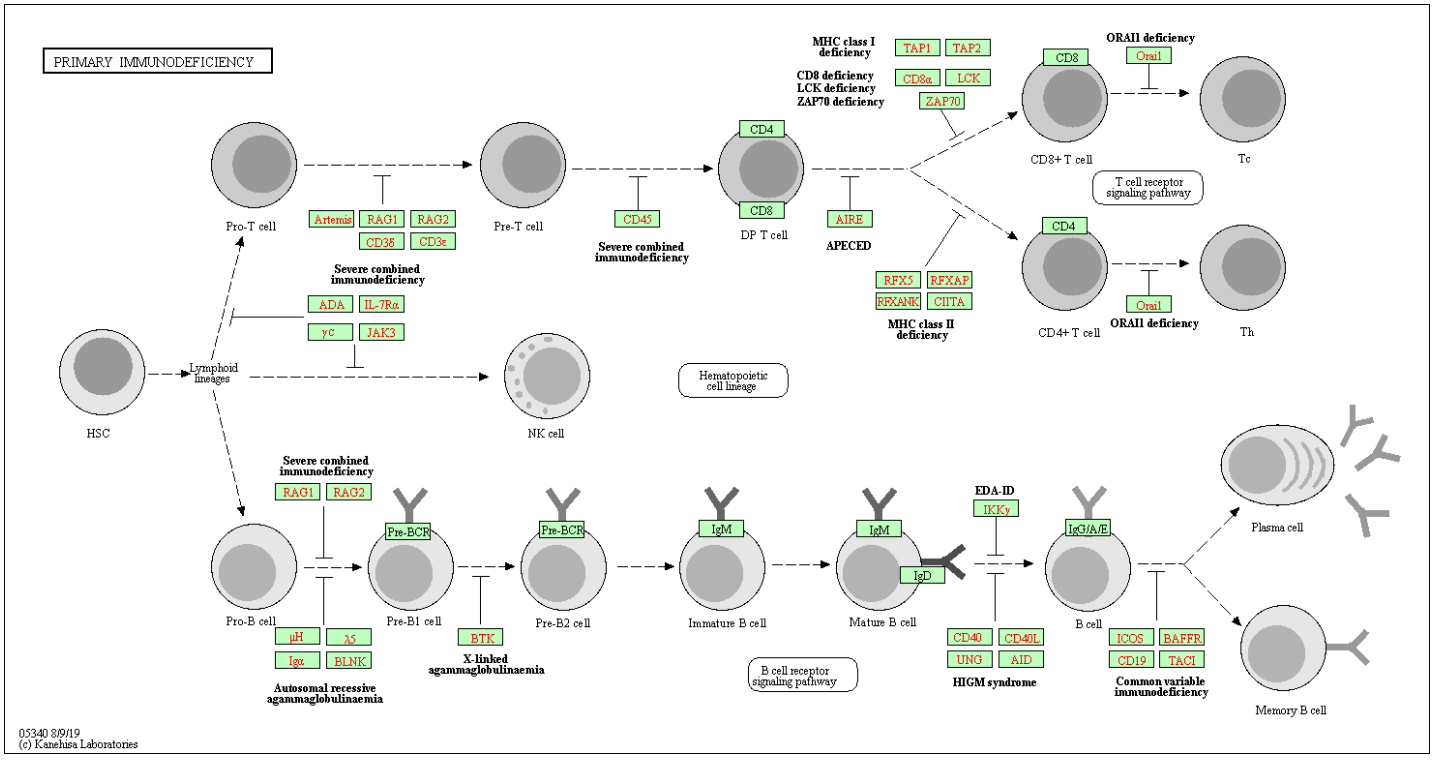


Figure 1. A map of immune cell differentiation labeled with potential mutations and the resulting immunodeficiencies. Obtained from Kanehisa and Goto, 2000.

However, it has been found that individuals with the same genetic variants may have very different presentations (Toubiana et al, 2018). Furthermore, healthy individuals may possess known pathological variants without manifesting any clinical phenotype at all (MacArthur et al, 2010). Additionally, sequencing cannot predict whether an undocumented mutation will be loss-of-function, gain-of-function, or without phenotype at all. While *in silico* methodologies to predict putative pathological variants exist, forecasting downstream protein function from intronic or cryptic variants remains a challenge (Schwarz et al, 2014). Therefore, we cannot consider sequencing the solution to all diagnostic queries in the realm of clinical immunology. In a post-genomic age, it is absolutely necessary to broaden our understanding of PIDs by integrating genomic strategies with studies of cellular function.

Common variable immunodeficiency is a poorly understood PID with multiple genetic etiologies that result in a decrease in immunoglobulin production and a predisposition towards infection. Patients have low levels of IgG and IgA and lack protective immune responses to vaccines; they often present with chronic sinusitis, bronchitis, and miscellaneous infections (Bonilla et al, 2016). The disease burden of CVID is relatively high for a PID, with one in 25,000 people suffering from it. Diagnostically, CVID is only considered when other explanations of hypogammaglobulinemia (ie, X-linked agammaglobulinemia) are ruled out. Diagnosing CVID essentially pools a variety of patients into one diagnostic "box." The heterogeneity of CVID may likely represent a group of disorders that are difficult to diagnose.

At least one of the following:

- Increased susceptibility to infection
- Autoimmune manifestations
- Granulomatous disease
- Unexplained polyclonal lymphoproliferation
- Affected family member with antibody deficiency

AND marked decrease of IgG and marked decrease of IgA with or without low IgM levels (measured at least twice; <2 SD of the normal levels for their age);

AND at least one of the following:

- Poor antibody response to vaccines (and/or absent isohemagglutinins); i.e., absence of protective levels despite vaccination where defined
- Low switched memory B cells (<70% of age-related normal value)

AND secondary causes of hypogammaglobulinemia have been excluded (see separate list)

AND diagnosis is established after the fourth year of life (but symptoms may be present before)

AND no evidence of profound T-cell deficiency, defined as two out of the following (y = year of life):

- CD4 numbers/microliter: 2–6 y < 300, 6–12 y < 250, >12 y < 200
- % Naive CD4: 2–6 y < 25%, 6–16 y < 20%, >16 y < 10%
- T-cell proliferation absent

http://esid.org/Working-Parties/Registry/Diagnosis-criteria.

Table 1. The 2013 ESID criteria for the diagnosis of CVID. Obtained from Ameratunga et al, 2014.

Clinical manifestations of CVID are highly diverse, spanning the gamut of immunodeficiency to autoimmunity to asymptomaticity (Cunningham-Rundles 2012). In cases of autoimmunity, patients often suffer from hemolytic anemia and thrombocytopenia. The immune dysregulation in CVID is especially difficult to treat and the lifespans of patients suffering from autoimmunity are much shorter than those without such complications. Furthermore, a patient may suffer from autoimmune episodes and experience drastic decreases in quality of life before being treated for autoimmunity.

Recent advances in exome sequencing and SNP genotyping have improved our understanding of the pathogenesis of CVID; however, there is still no methodology to predict the disease course that patients will experience (Aggarwal et al, 2020). Therefore, the field requires a method to classify patients into autoimmune or non-autoimmune categories to aggressively treat the more severe cases of autoimmunity. We aim to do this by integrating analyses of patient exomes with analyses of cell function. Such a multi-modal approach has never been used to predict the clinical manifestation of CVID. In this project, *we aim to enhance the diagnosis and treatment of CVID* by integrating genetics and functional characterizations of cellular activity. A long-term goal of this study is to *create a predictive model for CVID disease course*.

Phospho-flow cytometry is an indispensable method for analyzing signaling responses in immune cells (Krutzik et al, 2011). This method measures signal transduction in response to ligation of a cell-surface receptor by quantifying phosphorylation of downstream signaling intermediates. Traditional, fluorescence-based flow cytometry has long been a hallmark method for the functional analysis of PIDs (Takashima et al, 2017). However, these methods are limited in scope due to spectral overlap between fluorescent channels which leads to analytical maxima of ~10-15 markers. Using this approach, only a select few signaling pathways in a limited number of cell types can be analyzed in a single tube.

A novel methodology to overcome this limitation is Cytometry by Time-of-Flight (hereby referred to as mass cytometry or CyTOF). CyTOF utilizes metal-tagged antibodies and time of flight mass spectrometry to allow for the characterization of over 35 cellular surface markers and phospho-proteins with single-cell resolution (Figure 2). Since the metal-tagged antibodies are 99%+ isotopically pure, there is minimal overlap between metal channels. Thus, mass cytometry mitigates a baseline concern of flow cytometry—spectral overlap. Limitations to

CyTOF include run-time throughput, expense, and sensitivity to cell surface proteins (Simoni et al, 2018). Still, CyTOF has immense power in its ability to characterize many cell types and produce large amounts of data from single experiments.

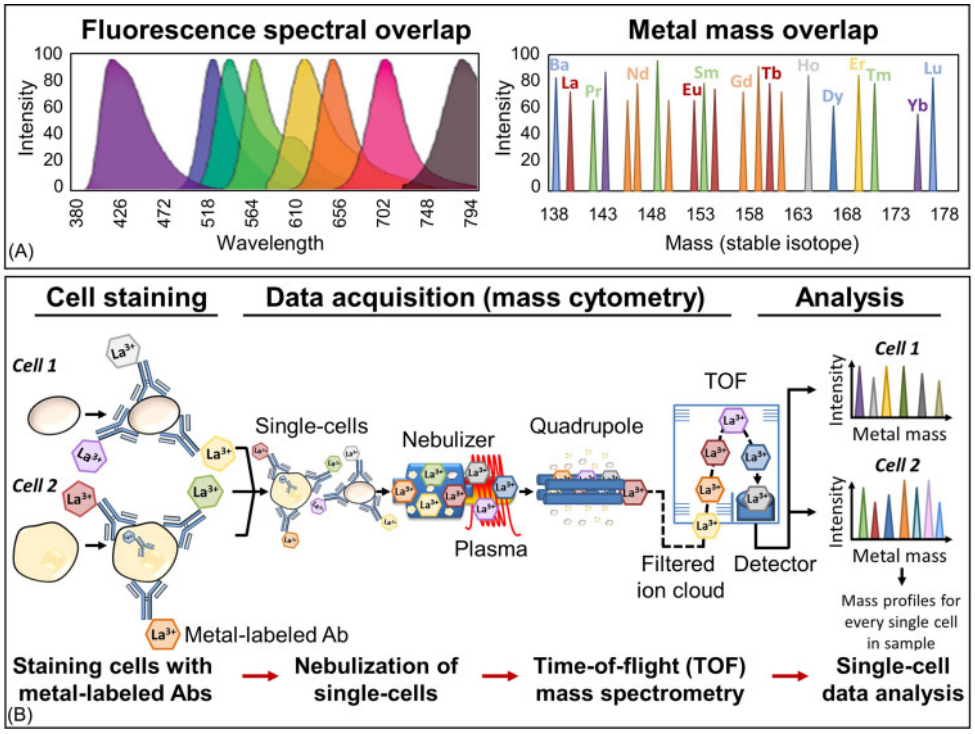


Figure 2. CyTOF workflow and technology. The metals utilized in mass cytometry have much less overlap than the fluorescence overlap seen in flow cytometry, resulting in less compensation. Obtained from Blair et al, 2019.

Using this approach, one can conceivably analyze multiple cell signaling pathways in all circulating immune cells and their respective subsets. Notably, CyTOF creates enormous data sets and requires massive computational power to glean insights. This dilemma is ongoing and often solved with the introduction of dimensionality reduction algorithms for manageable data analysis (Nowicka et al, 2017). In this paper we aim to employ widely used single-cell data analysis tools (ie, PCA, t-SNE). These techniques will be used for clustering based on signaling

events and will provide an exploratory function to discover statistically significant defects. Furthermore, we hope to innovate novel analysis tools for investigating cellular signaling by CyTOF.

In a proof-of-concept paper, our laboratory used CyTOF to identify signaling abnormalities in one patient with a gain-of-function STAT1 mutation and one patient with STAT3 deficiency manifesting in hyper-IgE syndrome (HIES) (Choi et al, 2016). Interestingly, the paper identifies a potential compensatory response by identifying baseline signaling dysregulation in both of these patients; the patient with GOF STAT1 had increased basal phosphorylation of STAT3 in certain T cells. Similarly, the patient with STAT3 deficiency had high basal STAT1 phosphorylation in many cell types. In addition, both patients had cell types with defective responses to cytokines. For example, the GOF STAT1 patient had regulatory T cells with a significantly higher response to IL-10 in STAT1, even though STAT3 is the primary STAT protein downstream of the IL-10 receptor. This collection of signaling findings in monogenic PIDs led us to believe that the diverse clinical manifestation of CVID could be better

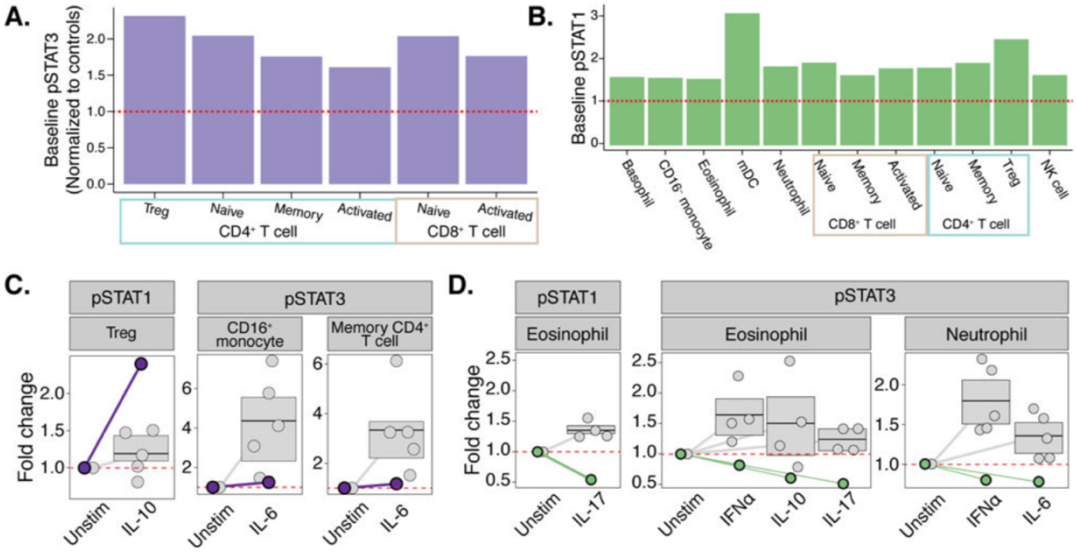


Figure 3. Previous work in the Butte laboratory finding baseline signaling abnormalities in **A)** GOF STAT1 and **B)** STAT3 deficiency patients. Stimulation responses were also aberrant in **C)** GOF STAT1 and **D)** STAT3 deficiency. Obtained from Choi et al, 2017.

explained and stratified by pairing exhaustive cellular signaling analyses with exome sequencing.

By measuring protein phosphorylation across all circulating immune cell types, we can identify individual aberrant signaling pathways within specific cellular subsets. Furthermore, these aberrancies will, hopefully, correlate to clinical phenotype and allow for a description of CVID that considers cellular function in addition to antibody levels and B cell maturity. In the future, understanding signaling deficiencies in patients may serve to elucidate novel targets for biologics in treating patients or easier diagnosis of more aggressive CVID. For example, if we find that a subset of patients who have an abnormal cell signaling signature are more susceptible to autoimmunity, aberrant signaling could be targeted by inhibitors and the patients could be closely monitored for autoimmune manifestations. One aim of this project is to *analyze abnormal signaling in CVID* to explain clinical manifestations of the disease.

Additionally, our experiments possess the capability to study previously understudied cells (T cells, NK cells, etc.) in CVID patients. These cells may contain useful information on patients' clinical phenotypes and have been neglected in many studies of CVID. By studying cell counts, surface protein levels, and signaling in all circulating immune cells, we will add to the depth of knowledge of CVID, which has classically been considered a B cell disorder. Indeed, previous work has shown defects in VDJ recombination in T cells in CVID (Ramesh et al, 2017). There may be many more defects in CVID beyond those in B and T cells, as well.

Additionally, we will perform exome sequencing on all patients. Exome sequencing will identify defects in the coding regions of patients' genomes. Currently, the genetic etiology of CVID is largely unexplainable; in fact, only about 10% of patients have a verifiable genetic explanation for their disorder. For the majority with a genetic explanation, genes relevant to B cell development (*TACI*, *CD20*, *BAFF*) are often mutated. By integrating genomics with

signaling analysis and cellular frequencies, we aim to see if there is *a correlation between certain exome defects and cellular/clinical phenotype*.

CVID is split into subclasses utilizing the EuroClass system, which is dependent on B cell phenotypes (Wehr et al. 2007). Our work will integrate exome sequencing, signaling, and prevalence of cellular subsets to better understand CVID. By using this three-pronged approach, we will create an enormous data set that comprehensively paints a picture of CVID patient phenotype. Likely, we will be able to group patients based on genetics, signaling, and cell phenotypes. Another aim of our study is to *correlate genomics, phenotype, and signaling patterns to understand the subtypes of CVID*.

Patients whom Dr. Butte has organized through his clinic and other connections will be studied—a first-of-its-kind large scale CyTOF analysis on this cohort of patients. The data collected through CyTOF will complement that exome sequencing and will allow us to assess how the immune cells of patients respond to different stimuli—highlighting aberrant signaling signatures. This will allow for an individual, precision-based approach in understanding the clinical symptoms of this disorder.

III. Experimental Design

We chose to use CyTOF to investigate the phenotype and signaling of CVID patient blood cells due to its ability to interrogate a wide range of cellular parameters, thus allowing for characterizations of cellular signaling across many cell types. By using common immune cell markers, our panel allowed for the recognition of all cells present in the blood from both the myeloid and lymphoid lineages (Table 2). Importantly, this represents *all circulating immune cell subsets*.

Cell	Markers
Basophils	HLADR-, CD123+
CD16hi Monocytes	CD66b-, CD14+, CD11c+, CD16+
CD16lo Monocytes	CD66b-, CD14+, CD11c+, CD16-
Central Memory CD4 T Cell	CD66b-, CD14-, CD3+, CD4+, CD27+, CD45RA-
Central Memory CD8 T Cell	CD66b-, CD14-, CD3+, CD8+, CD27+, CD45RA-
Effector CD4 T Cell	CD66b-, CD14-, CD3+, CD4+, CD27-, CD45RA+
Effector CD8 T Cell	CD66b-, CD14-, CD3+, CD8+, CD27-, CD45RA+
Effector Memory CD4 T Cell	CD66b-, CD14-, CD3+, CD4+, CD27-, CD45RA-
Effector Memory CD8 T Cell	CD66b-, CD14-, CD3+, CD8+, CD27-, CD45RA-
Eosinophils	CD66b+, CD16-
IgD+ Memory B Cells	CD66b-, CD14-, CD3-, CD19/CD20+, IgD+, CD27+
mDCs	CD66b-, CD14-, CD3-, CD19/20-, CD56-, CD123-, CD11c+
Naive B Cells	CD66b-, CD14-, CD3-, CD19/CD20+, IgD+, CD27-
Naive CD4 T Cell	CD66b-, CD14-, CD3+, CD4+, CD27+, CD45RA-
Naive CD8 T Cell	CD66b-, CD14-, CD3+, CD8+, CD27+, CD45RA-
Neutrophils	CD66b+, CD16+
NKs	CD66b-, CD14-, CD3-, CD19/20-, CD56+
NKT Cell	CD66b-, CD14-, CD3+, CD19/20-, CD56+
pDCs	CD66b-, CD14-, CD3-, CD19/20-, CD56-, CD123+, CD11c-
Plasmablasts	CD66b-, CD14-, CD3-, CD20+, IgD-, CD19lo, CD38+, CD27+
Regulatory T Cell	CD66b-, CD14-, CD3+, CD4+, FoxP3+
Switched B Cells	CD66b-, CD14-, CD3-, CD19/CD20+, IgD-, CD27-
Switched Memory B Cells	CD66b-, CD14-, CD3-, CD19/CD20+, IgD-, CD27+

Table 2. All cell types recognized by the phospho-signaling panel and the phenotyping panel

Our phospho-protein assay allowed us to functionally characterize cellular immune responses to many relevant stimuli (Table 3). For example, interferon and interleukin stimulation allowed for the assessment of signaling via the JAK/STATs pathways and the mTOR/AKT pathways, while Toll-like receptor (TLR) stimulation assessed the MyD88/IRF7 pathway. Furthermore, our panel also contained markers for proliferation (Ki67) and apoptosis (Cleaved Caspase 3). Interrogating all of these signaling pathways in all circulating immune cells provided a "birds-eye view" of immune responses in CVID patients.

Stimulus	Receptor	Intracellular Target
Unperturbed	NA	NA
R848	TLR1/2	pAKT, pP38, pERK1/2
PMA/Ionomycin	NA	pS6, pAKT, pCREB
PAM3CSK4	TLR7/8	pIRF7
LPS	TLR4	IkBa, pERK1/2, pP38, IkBa
IL6	IL6R	pSTAT1, pSTAT3, IkBa
IL21	IL21R	pSTAT1, pSTAT3, pSTAT5, pAKT, pERK1/2
IL2	IL2R	pSTAT1, pSTAT3, pSTAT5, pAKT, pERK1/2
IL10	IL10R	pSTAT3
IFNγ	IFNGR	pSTAT1, pP38, pERK1/2
IFNα	IFNAR	pSTAT1, pP38, pERK1/2

Table 3. All stimuli, receptors, and intracellular targets in the phospho-signaling panel.

In designing the CyTOF protocol, maintaining high cell counts and adequate cellular staining proved to be two of the biggest hurdles to a successful assay. High cell count (200,000 cells/tube) was maintained by washing blood, which was necessary for removal of IgM in patient sera and then reconstituting the washed blood in one half of the original volume. The protocol was optimized for cell staining by adding surface antibodies after lysis of RBCs, as lysing and fixing samples before staining significantly reduced signal intensity during CyTOF. Furthermore, 2 mM $CaCl_2$ was added to emulate physiological conditions necessary for cell signaling. Staining buffer was phosphate-buffered saline (PBS) with 1% bovine serum albumin (BSA) added to block non-specific binding of antibodies.

After washing the blood, Fc receptors were blocked for 10 minutes to further prevent binding of antibodies by their constant regions. Samples were then stained for 30 minutes at room temperature. They were stimulated with one of ten stimuli or PBS for fifteen minutes at 37 °C. While certain signals may peak at different time points, we used a fifteen-minute stimulation since it is the industry standard. Samples were lysed and fixed for ten minutes. After this process, they were either left at -4 °C after lysis of RBCs or processed altogether. In the phenotyping panel, FoxP3 permeabilization buffer was used for thirty minutes at RT. In the phospho-protein panel, permeabilization was done with methanol on ice for twenty minutes. Intracellular staining on both panels was done for thirty minutes. Samples were washed and incubated with fixation/nuclear permeabilization solution and 125 nM Iridium DNA intercalator. After overnight incubation and washing, cytometry was run on a Helios mass cytometer.

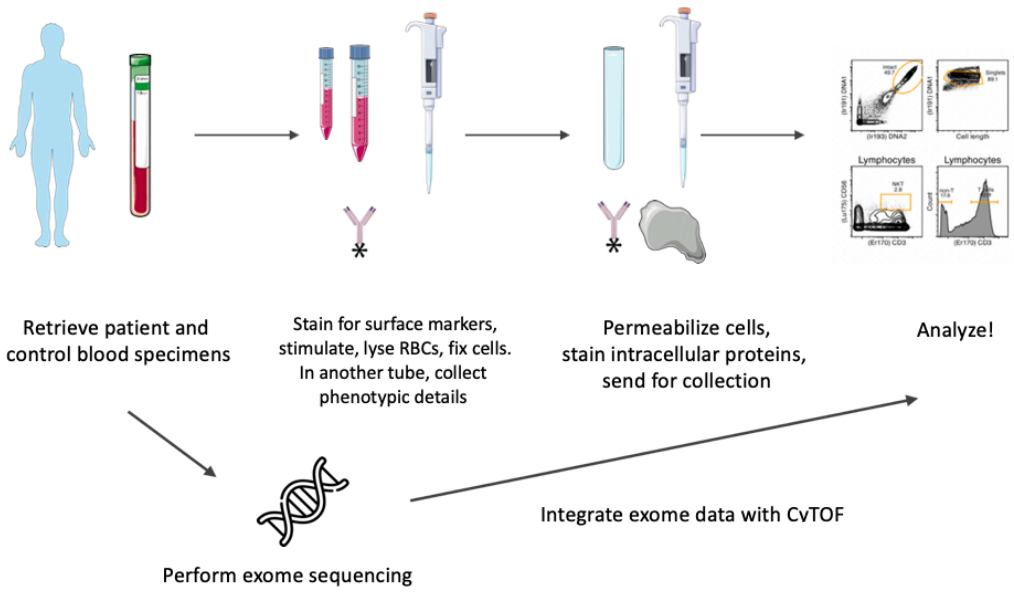


Figure 4. Schematic describing the CyTOF and exome workflow utilized in this paper.

FCS files were loaded into R for analysis. Data was first processed using the Logicle transformation, a hyperbolic sine transformation of data that has become standard in flow/mass cytometry analysis (Parks et al, 2006). A semi-automated gating scheme was then applied to all samples from a batch using CytoRSuite (Hammill 2019). After initial gating, manual gate editing was performed to ensure correct gating of rare cell populations such as basophils, mDC/pDCs, and plasmablasts. After gating, phospho-protein and surface marker data was loaded into a data frame in R and saved as a CSV.

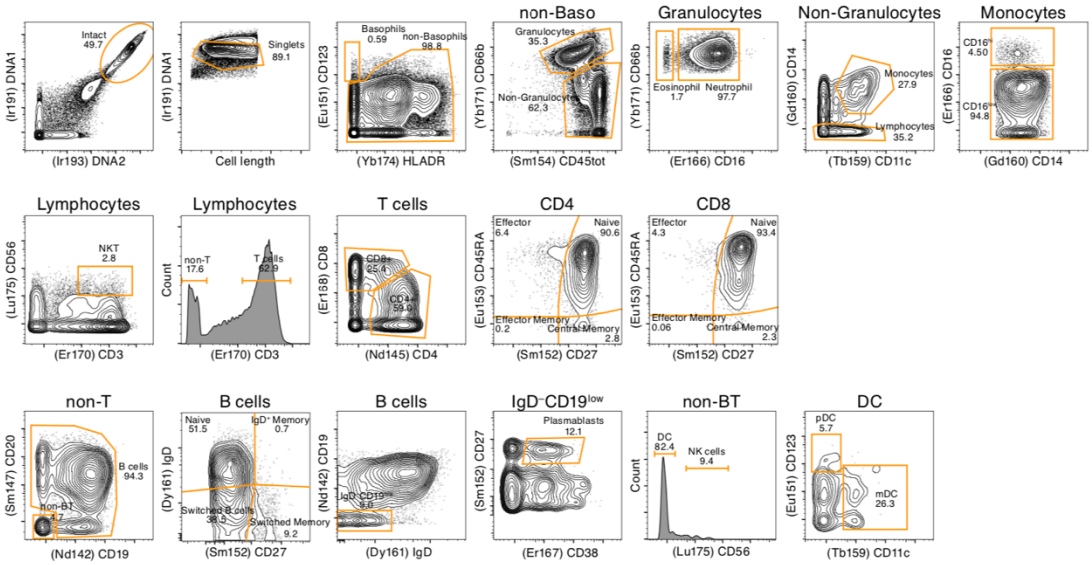


Figure 5. Representative gating scheme utilized in this study. Obtained from Choi et al, 2017.

We used bootstrapping and permutation testing for statistical comparisons. This approach mitigates the influence of outliers and does not require data to be normally distributed. Upon optimization of the GMM pipeline, our data will be re-assessed and validated. Furthermore, our analysis process blinded us to control vs. patient samples, which mitigated bias in gating.

We utilized two-component Gaussian mixture models (GMMs) to cluster cells into either "responding" or "non-responding" groups. Mixture models are powerful probabilistic models that are meant to detect subpopulations in a given distribution. Our GMMs describe both the means of the "non-responding" and "responding" groups and also describe the proportion of cells in either cluster (Figure 6). By creating a pipeline to identify "responding" and "non-responding" groups, we created an automated, unbiased approach to analyze CyTOF signaling data.

In particular, we created a set of GMM functions that constrained the means of stimulated peaks to a specified percent difference from control peaks. This was done to catch shifts in cell response that may represent three or more clusters instead of two and to ensure statistical reliability of our models. Furthermore, our algorithm restrained the GMM models to have positive variance and mixing coefficients (percentage of data covered by one Gaussian component of the model) for both components.

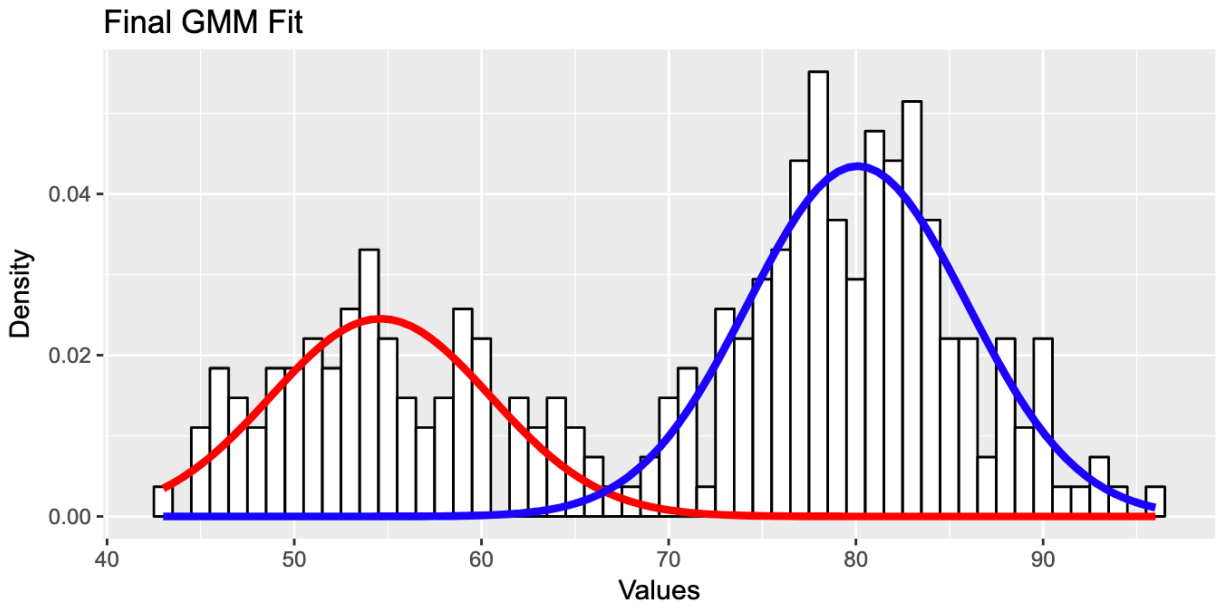


Figure 6. Example of a two-component Gaussian Mixture model fit to bimodal data. Obtained from Chan, 2016.

IV. Results

Our phenotyping panel of cell surface markers permitted us to study cell frequencies within control and patient samples. Total switched B cell counts, ascertained by measuring all IgD- B cells, were depleted in CVID patients, while IgM+, CD38+ B cells (IgM+ plasmablasts) were increased in CVID patients, compared to healthy controls (Figure 7A). Furthermore, CD21 expression on IgD+ memory and switched non-memory B cells was significantly reduced in CVID patients as well (Figure 7B). Notably, in addition to the aberrant predicted B cell phenotyping, we also found higher PD-1 expression in the effector CD8+ T cells of patients (Figure 8).

By analyzing stimulation assays with phospho-protein mass cytometry and higher-dimensional data analytics, we aimed to elucidate signaling deficiencies in patients with CVID. To deal with higher-dimensional data, we performed Principal Components Analysis (PCA) on signaling values in all cell types and found that patient eosinophils separated from their control counterparts (Figure 9A). Upon further interrogation, we found defective gain-of-function responses of pP38, pSTAT3, and Cleaved Caspase-3 to TLR1/2 stimulation in eosinophils (Figure 9B). We were curious about aberrant STAT responses in other cell types and found similar erroneous gain-of-function responses in CD16lo Monocytes in response to PMA/Ionomycin (Figure 10).

By creating GMMs on our data sets, we were able to collect subpopulation means from all of our stimulation condition-intracellular marker-subject-cell type combinations (Figure 11A). Importantly, this figure illustrates only a subset of the cell types that we analyzed and only the unperturbed, healthy control pSTAT5 mean values for those cell types. The full dataset is stored as shown (Figure 11B).

V. Discussion

Our experimental system has replicated findings from seminal studies on CVID (Warnatz et al, 2006). In particular, we show depletion of total switched B cells in CVID patients, a hallmark of the disease. Additionally, we found a decrease in CD27+ memory B cells (data not shown). These findings mirror previous seminal studies on CVID; they show inappropriately naïve B cell development in CVID patients. Utilizing CyTOF will expand on the relatively limited field of CVID signaling research.

In addition to confirming prior findings, our work has also identified novel B cell defects in CVID. For example, a significantly higher amount of IgM+ plasmablasts were found in CVID patients. IgM is the initial antibody isotype secreted upon infection; this result indicates that CVID patients have impaired humoral responses with decreased isotype switching. Furthermore, CD21 expression on IgD+ memory and switched non-memory B cells is decreased. CD21 facilitates B cell responses to complement-bound antigens (Tedder et al, 1997). With decreased CD21 expression, CVID patients may have impaired B cell receptor signaling, which will, in turn, contribute to deficient antibody responses. This finding further explains the lack of B cell signaling, response, and isotype switching in CVID patients.

Our novel results finding higher PD-1 expression on effector CD8+ T cells leads us to posit that humoral and cytotoxic immunity are jointly implicated in the pathophysiology of CVID. PD-1 is critical for maintaining self-tolerance and is also a marker of exhausted T cells (Dai et al, 2014). Our data shows a bimodal distribution of PD-1 expression on effector CD8+ T cells within CVID patients. Perhaps, upon further analysis, it will be found that CVID patients with autoimmunity are in the group of patients with increased amounts of exhausted T cells due to their continuous signaling in response to self-antigen. It is also quite possible that increases in

PD-1 on the T cells of some patients result in defective cytotoxic immunity in these patients and further contribute to their predisposition to infection. In future work, we will determine if there is a correlation between PD-1 expression and clinical manifestation.

It has also been found that eosinophils play a role in autoimmunity (Diny et al, 2017). Abnormal cellular signaling in this cell population suggests a previously unidentified role for innate immune cells, in particular eosinophils, to contribute to the pathology of CVID. Further work must be done to better understand the precise role of eosinophils in this disorder.

We have created a novel methodology for analyzing cellular signal responses with constraints. By modeling signaling parameters in higher-dimensional CyTOF data, we aim to utilize escape the pitfalls of many statistical methods associated with higher-dimensional data (Ronan et al, 2016). Our procedure allows for the conventional statistical analysis of CyTOF data, something not afforded by many algorithms currently in use.

The methodology utilized in this paper described a feasible procedure for gating and extracting CyTOF data, performing model fitting, and then utilizing this data for analysis. Our algorithm can be generalized and utilized in CyTOF or high-dimensional flow cytometry data of any order. In particular, this system will be useful for extraction of useful parameters such as subpopulation means, variances, and proportions from large data sets. We aim eventually to refine this model and release it as an open-source R package.

In our methodology, we first created an unrestrained model on our unperturbed samples, both healthy and CVID, and applied the parameters from the unrestrained model as initial values for our stimulated data. Then, we set constraints on the ability of our restrained model to deviate from the unrestrained means. This allows us to ensure that both components of our models do

not become aberrantly high under stimulation conditions. Hopefully, this will be an added safeguard against erroneously modeling our data.

Having run models with different constraint values (10%, 25%, 50%, 100%, 200%), what is left is to assess which constraint values both accurately represent the subpopulations of our data while also serving the quality control function of flagging distributions which may be tri-modal or more. The constraints serve as an important role in the automation and control of our models as they are applied to every cell type-signaling pathway-subject-stimulation condition. In the future, this work will serve as a model for phospho-signaling mass cytometry and simplify the analysis of CyTOF data. This project, therefore, will be two-fold; we aim to both improve the analysis of mass cytometry data generally and to use improved algorithms to better understand CVID.

VI. Materials and Methods

Human Blood Collection

All human blood was obtained through protocols approved by the institutional review board. Written informed consent was obtained from all donors. Peripheral blood samples were collected from 14 healthy donors and 14 patients with CVID.

CyTOF Antibodies

We utilized both purchased and in-house conjugated antibodies through the UCLA Flow Cytometry Core (Table 4).

Staining and Processing

Heparinized blood samples were washed with PBS+1% BSA to remove IgM from serum. Without this step, serum IgM would be bound by anti-IgM CyTOF antibodies and cellular signal intensity for this marker would be incredibly weak. Blood was then supplemented with 2 mM $CaCl_2$, incubated with FcX (BioLegend, 422302) for 10 minutes at RT, stained with appropriate surface antibodies for 30 minutes.

Phosphorylation Panel

After surface staining, samples were washed again, stimulated with IL-2 (Peprotech, 200-02), IL-6 (Peprotech, 200-06), IL-10 (Peprotech, 200-10), IL-21 (Peprotech, 200-21), IFN-α (Cell Signaling Technology, 8927SC), IFN-γ (Peprotech, 300-02), R848 (Invivogen, vac-r848), PAM3CSK4 (Invivogen, tlrl-pms), LPS (Sigma-Aldrich, L4391), PMA/Ionomycin (Sigma-Aldrich, P1585, I3909), or PBS as control at 37 degrees for 15 minutes. Red blood cells were then lysed using Lyse/Fix Buffer (BD Biosciences, 558049). Cells were washed, put on ice, and

Intracellular Panel			
Metal	Target	Clone	Manufacturer
89Y	CD45	HI30	BioL
115In	*Ki-67*	SolA15	eBio
141Pr	IgM	HMH-88	BioL
143Nd	CD127 (IL-7Ra)	A0195D5	BioL
144Nd	*pAkt*	D9E	CST
145Nd	CD4	RPA-T4	BioL
146Nd	IgD	IA6-2	BioL
147Sm	*pStat5*	47	CST
148Nd	CD16	3G8	BioL
149Sm	CD25	M-A251	BD
150Nd	CD20	2H7	BioL
151Eu	CD123/IL-3R	6H6	DVS
152Sm	CD66b	80H3	DVS
153Eu	*p-Stat1*	4a	DVS
154Sm	HLA-DR	L243	BioL
155Gd	CD45RA	HI100	DVS
156Gd	*p-p38*	D3F9	CST
158Gd	*pSTAT3*	4/P-STAT3	CST
159Tb	CD141	1A4	BD
160Gd	CD11c	Bu15	BioL
162Dy	FoxP3	259D/C7	DVS
163Dy	CD56 (NCAM)	NCAM16.2	BD
164Dy	*IκBα*	L35A5	DVS
165Ho	*pIRF7*	k47-671	BD
166Er	*cCaspase3*	D3F9	CST
167Er	CD27	O323	BioL
168Er	CD8	SK1	BioL
169Tm	CD19	HIB19	BioL
170Er	CD3	UCHT1	BioL
171Yb	*pERK1/2*	D13.14.4E	CST
172Yb	CD38	HIT2	DVS
173Yb	CD14	M5E2	BioL
174Yb	CD21	Bu32	BioL
175Lu	*pS6*	N7-548	DVS
176Yb	*pCREB*	87G3	CST
209Bi	CD11b	ICRF44	DVS

Phenotyping Panel			
Metal	Target	Clone	Manufacturer
89Y	CD45	HI30	BioL
115In	CD57	HCD57	BioL
141Pr	IgM	HMH-88	BioL
142Nd	CD19	HIB19	BioL
143Nd	CD25	M-A251	BD
144Nd	CD8	SK1	BioL
145Nd	CD163	GHI/61	DVS
146Nd	IgD	IA6-2	BioL
147Sm	CD20	2H7	DVS
148Nd	CD16	3G8	BioL
149Sm	CD66b	G10F5	BioL
150Nd	CD161	DX12	BD
151Eu	CD123 (IL-3R)	6H6	DVS
152Sm	TCRgd	11F2	DVS
153Eu	CD185 (CXCR5)	RF8B2	DVS
154Sm	CD197 (CCR7)	G043H7	BioL
155Gd	CD279 (PD-1)	EH12.2H	BioL
158Gd	CD33	WM53	DVS
159Tb	CD11c	Bu15	BioL
160Gd	CD14	M5E2	DVS
161Dy	CD21	BL13	BioL
162Dy	FoxP3	259D/C7	DVS
163Dy	CD56 (NCAM)	NCAM16.2	BD
164Dy	CD196 (CCR6)	11A9	BioL
165Ho	CD61	VI-PL2	DVS
166Er	CD4	SK3	BioL
167Er	CD27	O323	BioL
168Er	CD278 (ICOS)	C398.4A	DVS
169Tm	CD45RA	HI100	BioL
170Er	CD3	UCHT1	BioL
172Yb	CD38	HIT2	DVS
173Yb	CD94	HP-3D9	BD
174Yb	HLA-DR	L243	BioL
175Lu	CD194 (CCR4)	205410	DVS
176Yb	CD127 (IL-7Ra)	A0195D5	DVS
209Bi	CD11b	ICRF44	DVS

Table 4. A) Intracellular panel with antibodies for cell signaling interrogation highlighted in pink **B)** Phenotyping panel.

then permeabilized with methanol. Samples were then stained for intracellular phospho-proteins and incubated with Fix/Perm Buffer (Fluidigm, 201067) and 125 nM Iridium DNA intercalator (Fluidigm, 201192B) and washed with PBS+1% BSA and distilled water (Invitrogen, 10977023) before being run.

Phenotyping Panel

After surface staining, red blood cells were then lysed using Lyse/Fix Buffer. Cells were washed, put on ice, and then permeabilized with FoxP3 Perm Buffer (BioLegend, 421402). Samples were then stained for FoxP3, incubated with Fix/Perm and 125 nM Iridium DNA Intercalator, and washed with PBS+1% BSA and distilled water before being run.

Exome Sequencing

DNA from whole blood was extracted and then sent to Macrogen for whole-exome sequencing.

Statistical Analysis and Code

We utilized permutation testing in R to generate reference distributions of test statistics. By performing 20,000-200,000 permutations and referencing our observed parameters to a distribution of the test statistic, we mitigated the influence of outliers in our data. We calculated two-sided p-values through the permutationTest2 function in R (R Core Team, 2014). Bootstrapped means and calculated 95% confidence intervals are shown in all relevant figures. The CytoRSuite package in R was utilized for gating (Hammil, 2019). Manual gates were drawn with the same cellular populations analyzed from date to date. Values of zero were removed from the data and are common statistical noise in all CyTOF datasets.

At first, the Mixtools package in R was utilized (Benaglia et al. 2017). After trial and error, a new set of GMM functions was made with the framework provided by Fong Chun Chan, a bioinformatician at Achilles Therapeutics (2017). These functions are highly manipulatable in

the constraints that they impress on our models; in particular, they allow for restraints on deviation in the means of a new model from a given mean initialization pair.

VII. Future Experiments

While our experiments show signaling differentials between healthy controls and CVID patients, further work to validate our conclusions must be done. To do so, we have isolated peripheral blood mononuclear cells (PBMCs) from all patients and will reaffirm our work through flow cytometry stimulation assays. This would involve re-stimulation with the cytokine/TLR agonist of interest and interrogation of the defective signaling pathway. If interesting patterns show up in CVID signaling response, or in a subset of patients, then the relevant PBMCs can be utilized for miscellaneous other assays such as Western Blots. This defective signaling will then be connected to patient clinical presentation. Our immediate hurdle, however, is the analysis of our CyTOF data.

Means from all cell type-signaling pathway-stimulation condition-patient combinations have been extracted. Even with computational aberrancies and lack of some low-yield cells (Plasmablasts, pDCs), over 20,000 individual models have been created from each constraint level of our data. In sum, having run these models limiting deviation from unperturbed means at the levels of 10%, 25%, 50%, 100%, and 200%, we have *over 100,000 models to assess*.

We aim to quickly and efficiently access the quality of these models using goodness-of-fit (GoF) tests such as Pearson's Chi-Square Test, the Kolmogorov-Smirnov Test, and the Cramer-Von-Mises Test; these tests provide statistics to assess how well the empirical distributions from our CyTOF data is represented by our two-component Gaussian models. We will also look at how well models fit with different log-likelihood differences. Based on trial and qualitative observation of which tests best discern improper models from well-fitting ones, a standard test will be selected for generalizable quality control of our GMMs.

Still, it will conceivably be difficult to assure that all of our models are decent fits for the empirical data collected. To address this, we will create an easy plotting function in R to gauge our fits qualitatively. Scripts are currently being written to perform GoF quality control and graphing of our data along with the models that have been fit to them. This will allow for quick verification of model fits. Specifically, we are writing scripts to plot many fits (50+) on one page with a green, red, or orange outline dependent on goodness-of-fit and if the model is pushed to the edge of the constraints. If the model is pushed to the edge of the constrained values, it must be evaluated to ensure the validity of the model.

As an aside, our calculations still take over 2 hours to apply constrained GMMs over all of our stimulation data. Therefore, it may be necessary to utilize the 'Rcpp' package and write R functions in C++ to expedite calculations by virtue of memory allocation (Eddelbuettel and Francois 2011).

Once our models all verifiably describe our empirical data well, we will then perform batch effect correction. Batch effects are statistical noise that exists in high throughput experiments due to differences in handling, staining, or miscellaneous other non-biological factors and represent no biological effect. Batch effects will be corrected by calculating the bootstrapped mean of the "non-responding" healthy control unperturbed means and then calculating adjustment factors to align every control subject to have the same baseline values. These same adjustment factors will be applied to patient phospho-signaling values from the same date.

After quality control of our models using GoF tests and correcting for batch effects, we will perform statistical analysis on the fold-change differences in signaling responses to stimulus between healthy controls and patients. will require gauging the 95% confidence interval of

means on healthy control responses to stimulus and then individually comparing patient means to the standards computed from the controls. Similar methodology was utilized in a previous paper analyzing cell signaling in PID patients (Choi et al, 2016). To expand on this approach, we will leverage the mixing coefficients provided by GMMs. By considering the mixing percentage of the Gaussian components of our models, we can both analyze the mean response to stimulus as well as the proportion of cells responding. This technique permits studying not only defects in the ability to signal fully but also defects in the ability of a subgroup of cells to signal.

Our data science approach further requires the use of dimensionality reduction. We will utilize PCA to interrogate our higher dimensional data and to locate defective signaling axes in CVID patients. Furthermore, when enough patient data is available, we aim to use machine learning classifiers such as Support Vector Machines (SVM) to utilize cellular signaling and genetics as a methodology for predicting which patients will suffer from autoimmunity and which patients will suffer from solely immunodeficiency. Other work has done this to computationally diagnose CVID in comparison to other Primary Antibody Deficiencies (PADs); however, to our knowledge, no work has looked to stratify the phenotype of CVID patients utilize machine learning (Emmaneel et al, 2019).

Lastly, we aim to incorporate exome sequencing in our analyses and correlate genomics, signaling measurements, and cell population frequency to clinical phenotype in hopes of better understanding this incredibly complicated condition. Potential difficulties in incorporating genetics into our approach include the possibility of patient mutations in two genes encoding for two components of a signaling pathway; without *a priori* knowledge of how each mutation affects the function of its protein product (gain-of-function/loss-of-function/no effect), it will be

difficult to definitively assess the relevance of a genetic result in our panel. However, we hope that our signaling work will help elucidate any biological effect in these situations.

VIII. Acknowledgments

First and foremost, I'd like to thank the patients involved in this study and their families. The work done here would not be possible without you; hopefully, this project will provide some insight into the complexities of your conditions.

Alexis Stephens aided tremendously in organizing clinical samples; Miriam Guemes and Alejandro Garcia of the UCLA Flow Core graciously provided reagents and ran cytometry.

This work was funded by the Jeffrey Modell Foundation and the Association of Biomedical Research.

Furthermore, this work used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Institute for Digital Research and Education's Research Technology Group.

Manish, I've been lucky to have your guidance since day one—guidance in my life and especially in science. Thank you for believing in me, teaching me, and giving an undergrad their own project that they could flounder in (a lot) but also learn plenty from.

Thank you, Tim, for staying in late to mentor me. Thanks also, for all of the memes.

I'd like to further thank the rest of the Butte lab, who have made my undergraduate academic experience so enriching and fun. Genuinely, I couldn't have done it without such a fantastic group of scientists and friends to guide me and help me out.

Additionally, I couldn't have gotten so obsessed with this work without the support of a fantastic department. Joy Ahn and Juana Escobar have been amazing SAOs. Professors Galic, Bensinger, Su, and Bradley have provided me with some of the best conversations of my time at UCLA. Their fantastic advice and support throughout my research in undergrad have been irreplaceable.

My parents, Sabiha and Shakir Khan, have done everything in their power to empower my interests in science. From 23andMe genotyping for my birthday to supporting me throughout college, they've been understanding and kind to a degree I cannot describe. I cannot thank you two enough.

To some of the best, most supportive friends: Jibran, Jinwon, Alexandra, Sapna, and my roommates, thank you for always reminding me of what's important.

I'd, of course, like to mention my siblings, grandparents, and the other family and friends without whose support this would have been impossible. I could write a full thesis trying to name all of the people who have impacted me positively and guided me to this point. In lieu of that, I will simply say that I love you all. May this thesis serve to prove that I didn't just spend three years partying in college.

IX. References

Aggarwal, Vaishali, et al. "Recent advances in elucidating the genetics of common variable immunodeficiency." *Genes & Diseases* 7.1 (2020): 26-37.

Ameratunga, Rohan, et al. "Comparison of diagnostic criteria for common variable immunodeficiency disorder." *Frontiers in immunology* 5 (2014): 415.

Blair, Thomas A., Andrew L. Frelinger III, and Alan D. Michelson. "Flow cytometry." *Platelets*. Academic Press, 2019. 627-651.

Bonilla, Francisco A., et al. "International Consensus Document (ICON): common variable immunodeficiency disorders." *The Journal of Allergy and Clinical Immunology: In Practice* 4.1 (2016): 38-59.

Choi, Jeff, et al. "Systems approach to uncover signaling networks in primary immunodeficiency diseases." *Journal of Allergy and Clinical Immunology* 140.3 (2017): 881-884.

Cunningham-Rundles, Charlotte. "The many faces of common variable immunodeficiency." *ASH Education Program Book* 2012.1 (2012): 301-305.

Dai, Suya, et al. "The PD-1/PD-Ls pathway and autoimmune diseases." *Cellular immunology* 290.1 (2014): 72-79.

Diny, Nicola L., Noel R. Rose, and Daniela Čiháková. "Eosinophils in autoimmune diseases." *Frontiers in immunology* 8 (2017): 484.

Eddelbuettel, D and Francois, R. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software*, 40(8), 1-18 (2011). http://www.jstatsoft.org/v40/i08/.

Emmaneel, Annelies, et al. "A computational pipeline for the diagnosis of CVID patients." *Frontiers in immunology* 10 (2019): 2009.

Hammill, Dillon. "CytoRSuite: Compensation, Gating & Visualisation Toolkit for Analysis of Flow Cytometry Data." (2019). https://github.com/DillonHammill/CytoRSuite.

Kanehisa, Minoru, and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic acids research* 28.1 (2000): 27-30.

Krutzik, Peter O., et al. "Phospho flow cytometry methods for the analysis of kinase signaling in cell lines and primary human blood samples." *Flow cytometry protocols*. Humana Press (2011): 179-202.

MacArthur, Daniel G., and Chris Tyler-Smith. "Loss-of-function variants in the genomes of healthy humans." *Human molecular genetics* 19.R2 (2010): R125-R130.

Maffucci, Patrick, et al. "Genetic diagnosis using whole exome sequencing in common variable immunodeficiency." *Frontiers in immunology* 7 (2016): 220.

Nowicka, Malgorzata, et al. "CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets." *F1000Research* 6 (2017).

Parks, David R., Mario Roederer, and Wayne A. Moore. "A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data." *Cytometry Part A: The Journal of the International Society for Analytical Cytology* 69.6 (2006): 541-551.

Picard, Capucine, et al. "International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity." *Journal of clinical immunology* 38.1 (2018): 96-128.

Ramesh, Manish, et al. "Clonal and constricted T cell repertoire in Common Variable Immune Deficiency." *Clinical Immunology* 178 (2017): 1-9.

Ronan, Tom, Zhijie Qi, and Kristen M. Naegle. "Avoiding common pitfalls when clustering biological data." *Science signaling* 9.432 (2016): re6-re6.

Schwarz, Jana Marie, et al. "MutationTaster2: mutation prediction for the deep-sequencing age." *Nature methods* 11.4 (2014): 361-362.

Takashima, Takehiro, et al. "Multicolor flow cytometry for the diagnosis of primary immunodeficiency diseases." *Journal of clinical immunology* 37.5 (2017): 486-495.

Tedder, Thomas F., Makoto Inaoki, and Shinichi Sato. "The CD19–CD21 complex regulates signal transduction thresholds governing humoral immunity and autoimmunity." *Immunity* 6.2 (1997): 107-118.

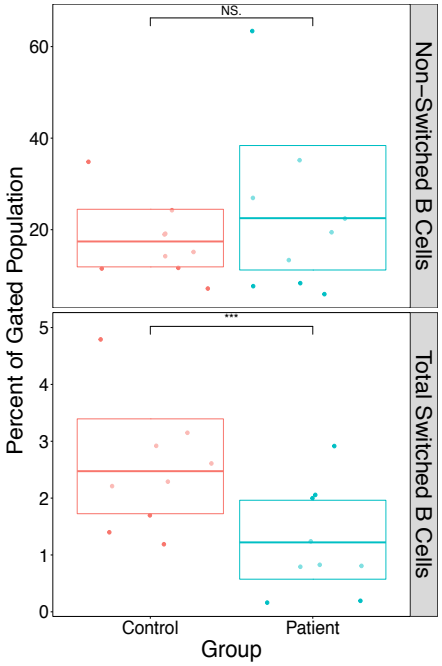
Toubiana, Julie, et al. "Heterozygous STAT1 gain-of-function mutations underlie an unexpectedly broad clinical phenotype." *Blood, The Journal of the American Society of Hematology* 127.25 (2016): 3154-3164.

Warnatz, Klaus, et al. "Severe deficiency of switched memory B cells (CD27+ IgM– IgD–) in subgroups of patients with common variable immunodeficiency: a new approach to classify a heterogeneous disease." *Blood, The Journal of the American Society of Hematology* 99.5 (2002): 1544-1551.

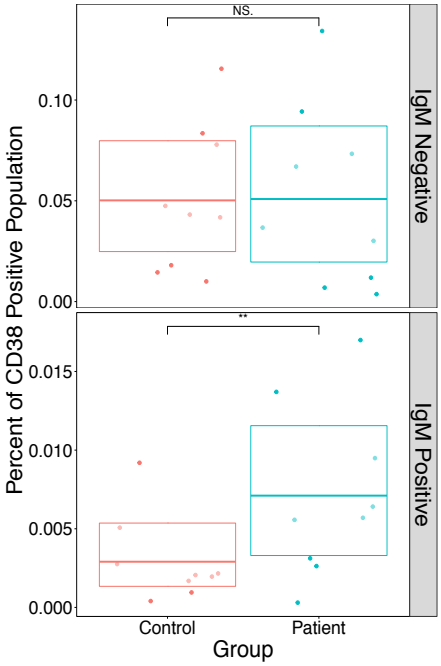
Wehr, Claudia, et al. "The EUROclass trial: defining subgroups in common variable immunodeficiency." *Blood, The Journal of the American Society of Hematology* 111.1 (2008): 77-85.

X. Figures

7A.



7B.



7C.

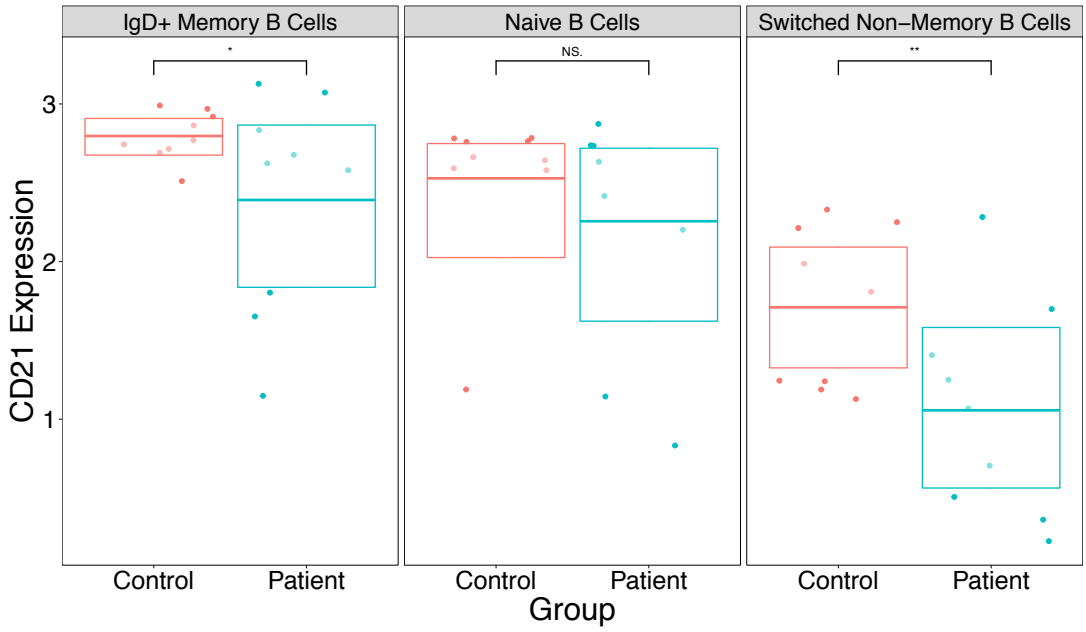


Figure 7. A) Patients with CVID do not have significantly more non-switched B Cells but do have significantly less switched B Cells. Switched vs. Non-switched was determined by the presence of IgM on the cell surface. **B)** There is a significant increase in the amount of IgM+ plasmablasts in CVID patients. **C)** CD21 Expression is decreased on IgD+ Memory B Cells and Switched Non-Memory B Cells.

8.

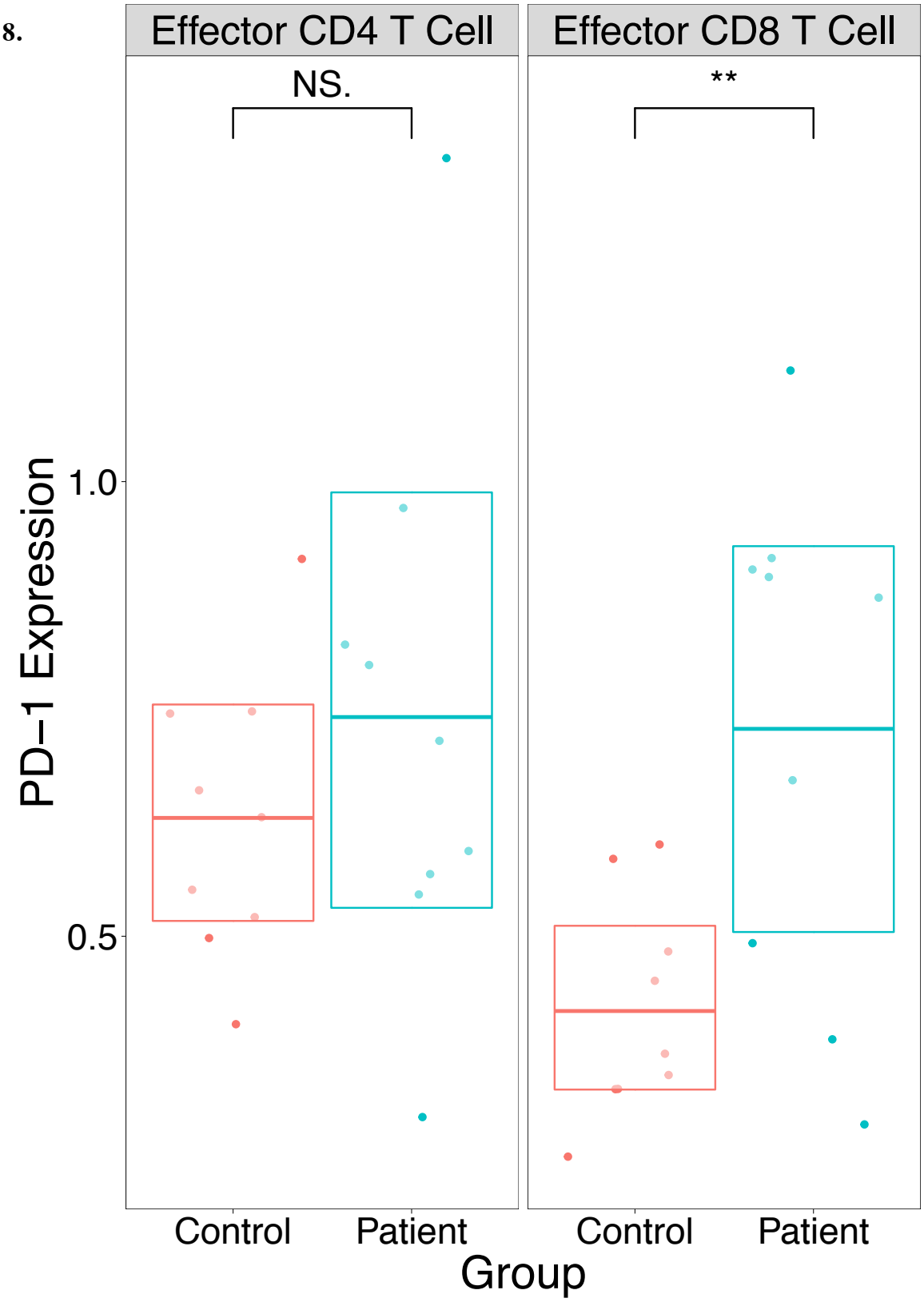
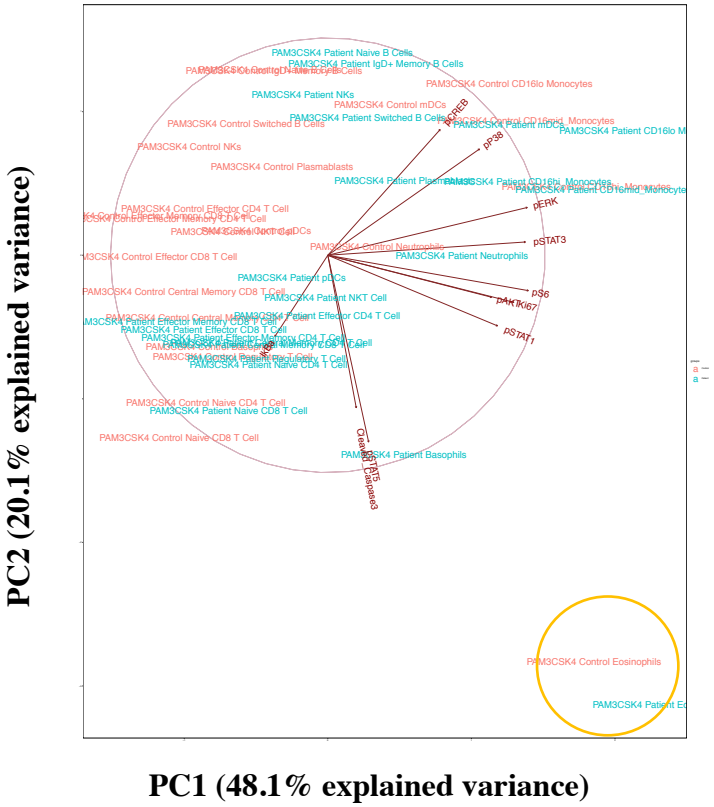


Figure 8. PD-1 Expression on Effector CD8+ T Cells is increased, while this overexpression is not seen on other T cell subsets.

9A.



9B.

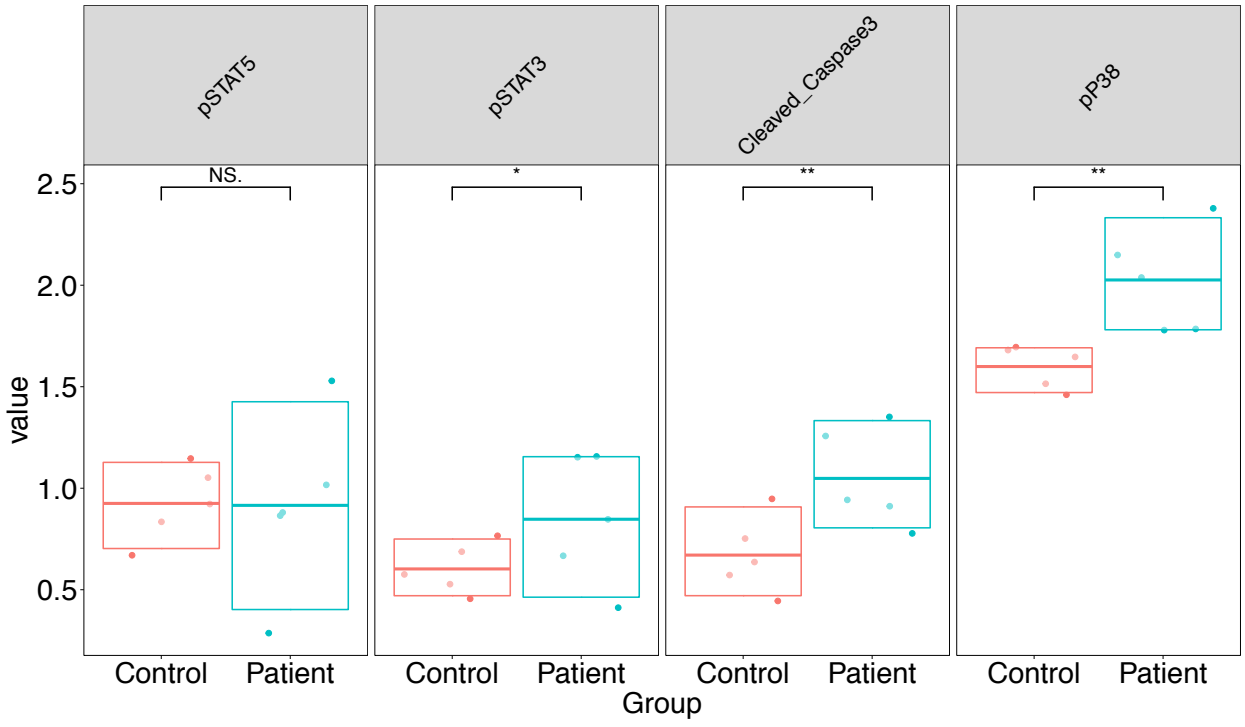


Figure 9. A) Principal Component Analysis (PCA) showing separation of Eosinophil populations from other cell subtypes, thus prompting further analysis. **B)** Eosinophil responses to TLR1/2 stimulation results in defective pP38, pSTAT3, and cleaved Caspase3 response.

10.

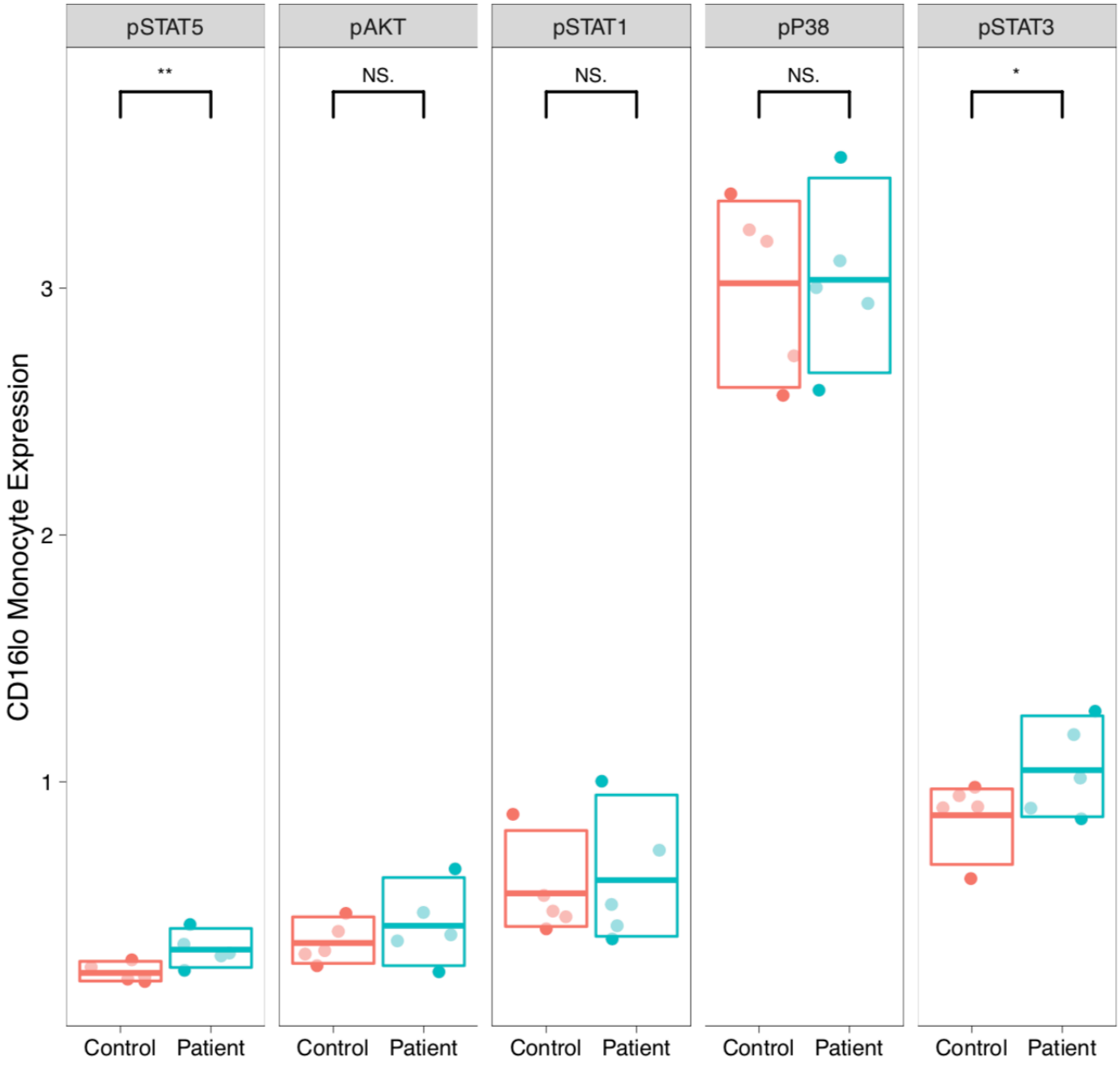
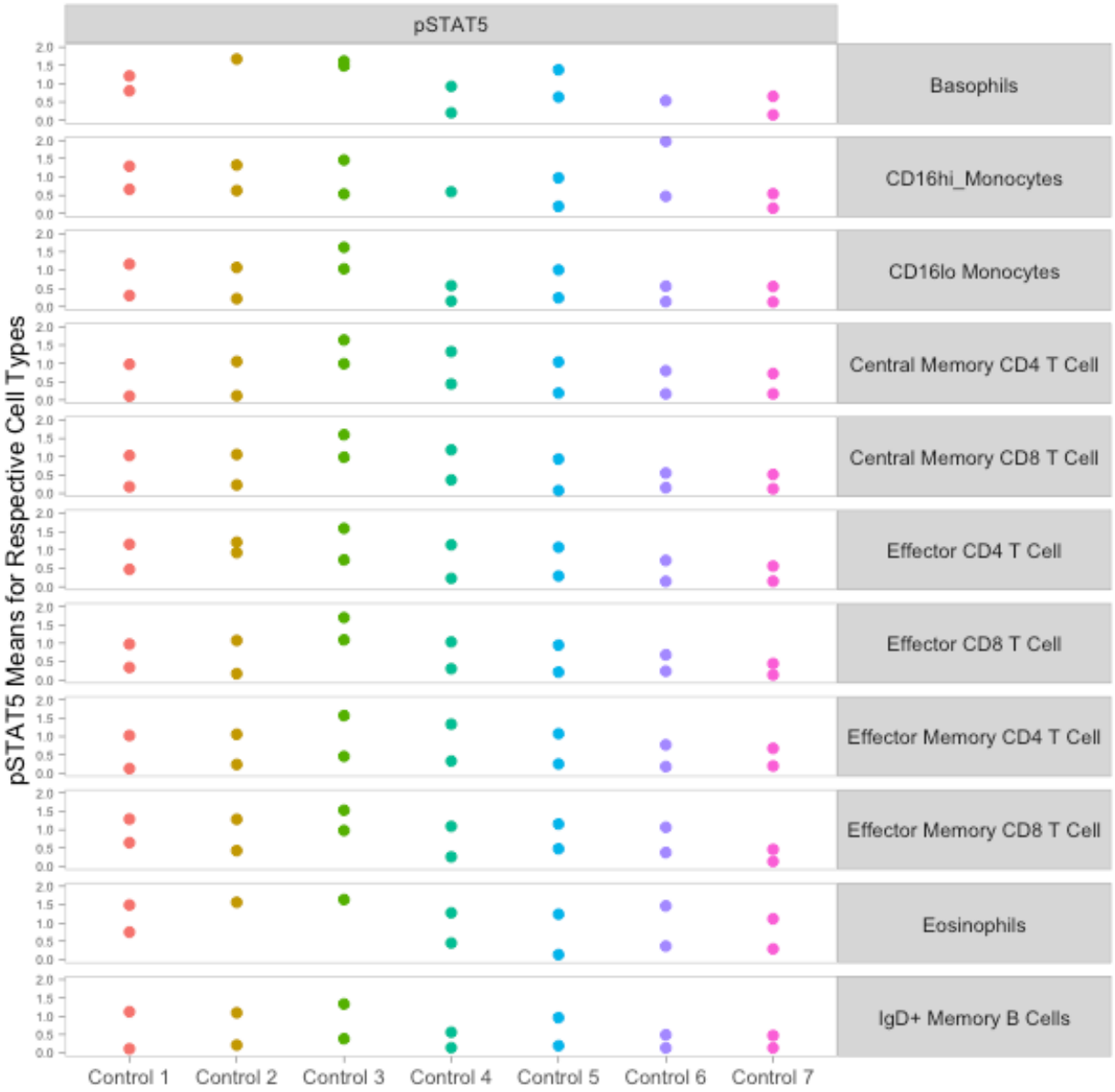


Figure 10. Defective pSTAT5 and pSTAT3 responses in CD16lo Monocytes in response to PMA/Ionomycin.

11A.



11B.

```
> GMMStim
# A tibble: 37,516 x 9
      mu    var lambda  iter loglik.diff Subject   Cell       Stim  State
   <dbl>  <dbl>  <dbl> <int>        <dbl> <chr>     <fct>      <fct> <fct>
 1 1.38  0.192   0.695     2    0         Control 1 Basophils IFNa  pSTAT5
 2 0.843 0.191   0.305     2    0         Control 1 Basophils IFNa  pSTAT5
 3 1.59  0.261   0.518     4    0.956     Control 2 Basophils IFNa  pSTAT5
 4 2.21  0.147   0.482     4    0.956     Control 2 Basophils IFNa  pSTAT5
 5 1.71  0.372   0.753     3  148.        Control 3 Basophils IFNa  pSTAT5
 6 1.79  0.123   0.247     3  148.        Control 3 Basophils IFNa  pSTAT5
 7 0.229 0.0217  0.106    44    0.000977  Control 4 Basophils IFNa  pSTAT5
 8 1.02  0.130   0.894    44    0.000977  Control 4 Basophils IFNa  pSTAT5
 9 1.52  0.0882  0.820    13    0.0270    Control 5 Basophils IFNa  pSTAT5
10 0.704 0.135   0.180    13    0.0270    Control 5 Basophils IFNa  pSTAT5
# … with 37,506 more rows
```

Figure 11. A) Gaussian Mixture Model means from unperturbed samples of healthy controls in a set of the cell types analyzed **B)** The R data frame with final mean, variance, and mixing coefficient (lambda) values for all stimulation condition-intracellular marker-subject-cell type combinations.