A Molecular network approach reveals shared cellular and molecular signatures between chronic fatigue syndrome and other fatiguing illnesses

Authors:

Phillip H. Comella [1,2,3], Edgar Gonzalez-Kozlova [1], Roman Kosoy [1], Alexander W. Charney [1,2,4,5], Irene Font Peradejordi [1,2,6], Shreya Chandrasekar [1,2,6], Scott R. Tyler [1], Wenhui Wang[1], Bojan Losic [1], Jun Zhu [1,2], Gabriel E. Hoffman [1], Seunghee Kim-Schulze [7], Jingjing Qi [7], Manishkumar Patel [7], Andrew Kasarskis [1,2,8], Mayte Suarez-Farinas [1,2], Zeynep H. Gümüş [1,2], Carmen Argmann [1,2], Miriam Merad [7,9,10,11], Christian Becker [12], Noam D. Beckmann [1,2], Eric E. Schadt [1,2,13]

Author Affiliations:

1. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
2. Icahn Institute of Data Science and Genomics Technology, New York, NY 10029
3. Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
4. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
5. Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
6. Cornell Tech at Cornell University, New York, NY, 10044, USA
7. Human Immune Monitoring Center, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
8. Department of Population Health Science and Policy at the Icahn School of Medicine at Mount Sinai
9. Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
10. Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
11. Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
12. Westchester Medical Center, Valhalla, NY 10595 USA
13. Sema4, a Mount Sinai venture, Stamford CT, 06902, USA

Corresponding authors: Phillip Comella (phillip.comella@icahn.mssm.edu), Eric E. Schadt (eric.schadt@mssm.edu)

Competing interest statement: The external patient data used to build the Bayesian network were partially funded as part of research alliance between Janssen Biotech and The Icahn School of Medicine at Mount Sinai.

Intro

The molecular mechanisms of chronic fatigue syndrome (CFS, or Myalgic encephalomyelitis), a disease defined by extreme, long-term fatigue, remain largely uncharacterized, and presently no molecular diagnostic test and no specific treatments exist to diagnose and treat CFS patients. While CFS has historically had an estimated prevalence of 0.1-0.5% [1], concerns of a "long hauler" version of Coronavirus disease 2019 (COVID-19) that symptomatically overlaps CFS to a significant degree **(Supplemental Table-1)** and appears to occur in 10% of COVID-19 patients[2], has raised concerns of a larger spike in CFS [3]. Here, we established molecular signatures of CFS and a corresponding network-based disease context from RNA-sequencing data generated on whole blood and FACs sorted specific peripheral blood mononuclear cells (PBMCs) isolated from CFS cases and non-CFS controls. The immune cell type specific molecular signatures of CFS we identified, overlapped molecular signatures from other fatiguing illnesses, demonstrating a common molecular etiology. Further, after constructing a probabilistic causal model of the CFS gene expression data, we identified master regulator genes modulating network states associated with CFS, suggesting potential therapeutic targets for CFS.

Main Text

Considerable controversy exists as to whether CFS has one or many causes [4, 5] and whether the resulting symptoms are somatic or psychosomatic [6]. Previous studies have proposed viral infections as a possible cause of CFS [7, 8], but to date, no single (or set) of virus(es) has been reported to be causally associated to the syndrome. Other hypotheses have been brought forth, such as immune system abnormalities[9], NK and T cell dysfunction[10, 11], cytokine dysregulations [12, 13], endocrinologic and metabolic abnormalities[14, 15], and miRNA associations with severity[16]. With the lack of understanding of CFS, there is a scarcity of potential therapies for those diagnosed with this syndrome. Preliminary data suggests beneficial effects of B cell depletion therapy via anti-CD20 monoclonal antibodies (rituximab) in a subset of CFS patients [17], as well as low-dose naltrexone [18], but no treatment modality to date has consistently and conclusively been shown to be effective.

To provide some insight into the molecular processes that underlie CFS, we carried out a study on 15 patients diagnosed with CFS and 15 age, sex, and BMI matched controls **(Supplemental Table-2).** CFS was formally diagnosed using the Fukuda Criteria **(Supplemental Table-4)**[19], Canadian Consensus Guidelines [20], and updated international consensus criteria, but excluding any related conditions such as major depressive disorder, Collagen-vascular diseases (CVD), neuromuscular diseases (NMD), and significant cardiac or pulmonary comorbidities **(Online Method)**. All participants were administered a moderate-intensity cardiopulmonary exercise test (CPET), with whole blood draws occurring immediately before the CPET and then 24, 48, and 72 hours post-CPET (**Fig. 1a**). In addition to the CPET, a clinical workup was performed on each participant, which included an EKG, standard metabolic panel blood test, height and weight, resting heart rate and blood pressure, and completion of a health assessment survey **(Online Method, Supplemental Table-3)**. Whole blood samples were cell sorted into B cells, Granulocytes, Monocytes, Natural Killer (NK) Cells, and T cells within 4 hours of collection and processed for RNA-seq. Differential Expression (DE) analyses showed no statistically significant (FDR<0.05) gene expression differences between any time points in either cases or controls for any of the cell types assessed. Relaxing the FDR threshold to <0.1 resulted in signatures detected in a few of the timepoints (CFS Bcell timepoints 1-2; CFS NK timepoints 1-4, 2-4, 3-4; CFS Mono timepoints 1-4; CFS Tempus timepoints 1-2, 2-4; Control Tempus timepoints 3-4). Furthermore, no statistically significant difference between time points was observed in data collected from the Modified Fatigue Impact Scale (MFIS) questionnaire, Karnofsky performance scores, or clinical workups, although patients did report increased physical fatigue (0-10 rating) between timepoint 1 and timepoints 2 and 3 **(Supplemental Table-3)**, consistent with previous findings that CPET does not appear to strongly effect molecular differences in CFS [21, 22] to a significant degree. These results suggest the need for a more intense exercise protocol and/or much larger sample sizes to assess whether post-exertional malaise may have a more modest impact on molecular states. Given the lack of a CPET-associated time signature, for all further analyses presented here, the time series data are treated as biological replicates.

We applied a comprehensive RNA-seq data analysis pipeline (**Fig. 1b-e**.) to analyze whole blood, as well as specific immune cell types isolated from whole blood of all participants. The first part of our analysis involved passing the RNA-seq data through a viral/clonal detection pipeline further detailed in the methods section (**Fig. 1b**).To visualize virome-wide variance differences between cases and controls, we performed principal component analysis (PCA) on whole blood **(Fig. 2a)** . The first two PCs represented ~30% of the total variance, revealing a separation between cases and controls across PC2 (~10% of the total variance), suggesting differences in total viral loads between groups. To further characterize these differences, we tested for differences in the viral load distributions between cases and controls using a Wilcoxon rank (**Fig. 2b**) and Kolmogorov–Smirnov (**Fig. 2c**) tests. Both tests were significant at the p < 0.0001 threshold.

Given the important roles T and B cells play in immune cell function, we investigated the clonality of each cell type by deconvolving the V[D]J genes using the MixCR algorithm [23]. We observed greater read support per clone or a less diverse population of T and B Cell clones in cases compared to controls (p < 0.0001) **(Fig. 2d)**, suggesting a dysregulation of these cell type populations in cases compared to controls.

To establish molecular signatures for CFS (**Fig. 1c, Supplemental Table-5**), we ran both DE (**Fig. 3b)** and Machine Learning Feature Selection (MLFS, methods) analyses (**Fig. 3a)**, leveraging both linear (DE) and non-linear (MLFS, random forest) approaches to identify gene expression features associated with case/control status. The non-linear MLFS strategy was used to cast the broadest net for identifying genes considered as significantly associated with case/control status (Information Gained Score FDR<0.05 in at least 1 timepoint per cell type), while the linear DE strategy was used to determine directionality of the gene expression changes between cases and controls for those genes identified by MLFS procedure (**Fig. 3c**).

Our MLFS strategy **(Supplemental Fig.1, Online Method)** resulted in 10 unique classifiers built per cell type, per timepoint, that were then tested against 10 unique hold-out test sets (**Fig. 3a)**. Classifiers built from B cell, T cell, Granulocyte, and NK cell signatures all had a mean receiver operating characteristic area under the curve (ROC AUC) > 0.80 (compared to an ROC AUC of 0.50 that would be expected by chance), with B cell signatures being the most predictive in terms of the mean ROC AUC (0.87 for B cells). Models built from whole blood had a ROC AUC = 0.63, suggesting that whole blood may be too noisy to detect a meaningful disease signature. The ability of our CFS classifiers to show predictability in this study is indicative of molecular differences between CFS cases and controls. These MLFS genes, along with their DE detected expression directionality, will be referred to as the CFS signatures for the remainder of this paper (**Fig. 3c)**.

To assess the similarity of the DE signatures across the different cell types, we computed the Jaccard index for all signature pairs (depicted as a heatmap in **Fig. 3d**). These results show that the signatures are largely unique across cell types, but that genes that are up (down) regulated within each cell type, cluster together across the cell types. **Fig. 3e** shows the top 2 gene ontology (GO) terms enriched in each direction of the cell-type specific CFS signatures, highlighting some of the biological processes (e.g. immune, metabolic and transcriptional dysregulation) that may be disrupted in CFS cases compared to controls. At the level of DE signatures, there is little overlap between the pathways that are the most significantly enriched across the cell types profiled.

To better characterize the subnetworks and pathways that may be disrupted in CFS, we constructed 6 co-expression networks to organize the gene expression data into coherent subnetworks of co-regulated genes (modules) for each of the 6 cell types (including whole blood) that were profiled in our study. While MLFS and DE aid in identifying the best features for distinguishing cases from controls, the co-expression network structure allows us to organize all of the gene expression traits into co-regulated groups of genes (modules) that place the individually identified MLFS/DE features into a more biologically relevant context. From the co-expression network analysis, we identified 119 co-expression modules spanning the 6 co-expression networks across all cell types (**Fig 1d, Supplemental Table-6**). Interestingly, 56 of the 119 modules were significantly enriched for the CFS signatures (**Fig 4a**).

We examined the top 5 modules enriched for the CFS signature as ranked by FDR adjusted p value: 1) Module 4 (M4) identified in the NK cell network (denoted as NKM4; FDR=1.39e-66, logFC=

1.71); 2) BcellM7 (FDR=2.15e-57, logFC=0.675); 3) MonoM6 (FDR=2.33e-53, logFC=0.658); 4) NKM13 (FDR=2.88e-50, logFC=1.002); 5) NKM9 (FDR=3.10e-50, logFC=0.826), for enrichment of published signatures (**Online Methods, Supplemental Table-7, 13**) for other fatiguing illnesses or diseases involving a strong inflammatory component such as Multisystem Inflammatory Syndrome in Children (MIS-C), Kawasaki Disease (KD), Macrophage Activation Syndrome (MAS), Neonatal Onset multisystem inflammatory (NOMID), Lyme disease, active Influenza (IAV), active COVID-19, early recovery stage after COVID-19, Mixed Connective Tissue Disease (MCTD), Sjögren's Syndrome (SJS), Systemic Lupus Erythematosus (SLE), Systemic Sclerosis (SSC), Undifferentiated Connective Tissue Disease (UCTD), Primary Antiphospholipid Syndrome (PAPS) and Rheumatoid Arthritis (RA). The top CFS modules are enriched for many of these external disease signatures, such as MIS-C, Lyme disease, and COVID-19, but notably are not enriched for any of the autoimmune signatures (**Fig. 4b**). Similarities between CFS, COVID-19, and Chronic Lyme Disease have been suggested at a clinical level [3, 24, 25] and our data further suggests a shared molecular etiology among these diseases.

To further characterize these CFS modules, we searched for biological processes and pathways that were also enriched in them (**Fig. 4c, Supplemental Tables 8-10**). NKM4 was the module most strongly associated with CFS (logOR=1.71, FDR=1.39e-66) and was also enriched for the "recovering COVID-19" signature (Wen Proliferating T-cells, logOR=2.02, FDR= 1.07E-15), highlighting a molecular link that may help explain why so many recovered COVID-19 patients seem to experience CFS-like symptoms [2]. The top biological pathway associated with NKM4 was MAPK Cascade (logFC=4.691, FDR=1.72e-4), consistent with *in vitro* findings of MAPK dysregulation in NK cells of CFS [26] and COVID-19 [27]. Lyme disease signatures were enriched for the greatest number of top CFS modules, with 4 of the top 5 CFS-enriched modules also enriched for the chronic Lyme disease signature, suggesting a potential molecular link between CFS and chronic Lyme disease as has been previously clinically described [28]. Top biological pathways associated with these 4 modules include MAPK Cascade (logFC=4.691, NKM4 FDR=1.72e-4), KRAS Signaling Up (MonoM6 logFC=3.049, FDR=2.00e-4), Neutrophil Degranulation (BcellM7 logFC=3.668, FDR=2.21e-26), and Platelet Alpha Granule (NKM13 logFC=16.495, FDR=1.84e-11). These annotated module similarities appear to correlate with *in vivo* findings of MAPK dysregulation in NK cells of Lyme disease [29] and CFS [26], coagulation dysregulation in Lyme disease [30] and CFS [31], as well as neutrophil dysregulation in Lyme disease [32] and CFS [33] patients. MIS-C and Kawasaki Disease appear to share common molecular processes with CFS through the BcellM7 and NKM13 modules. Top biological pathways associated with these modules include Neutrophil Degranulation (BcellM7 logFC=3.668, FDR=2.21e-26) and Platelet Alpha Granule (NKM13 logFC=16.495, FDR=1.84e-11). These annotated module similarities align with findings of neutrophil dysregulation in Kawasaki Disease [34], MIS-C [35], and CFS [33] along with coagulation dysregulation in Kawasaki Disease [36], MIS-C [37], and CFS [31]. **Supplemental Table-7** shows the enrichment of all disease signatures across all of the CFS enriched modules.

While co-expression networks well organize the expression data into modules of coregulated genes, such networks to not predict regulatory relationships among the genes. To explore more dominant regulatory relationships in the top CFS co-expression modules, we constructed a probabilistic causal Bayesian network (BN) to aid in the identification of master regulators, or key driver genes (KDs), for each of the top CFS-associated modules (**Fig. 1e, Supplemental Table-1**2). Given our small sample size and to achieve adequate power in constructing a reliable BN model, we chose to use gene expression data from a large external gene expression dataset (Mount Sinai Crohn's and Colitis Registry, MSCCR) consisting of 209 healthy individuals from which whole blood samples had been collected. We hypothesized that while gene regulatory relationships may be modulated at different levels in health individuals and those with CFS or the matched controls, the relationships themselves would be largely conserved as has been shown for other diseases [38-40]. We constructed the BN from a CFS-centered set of genes, including genes from the CFS and related disease signatures and co-expression modules significantly enriched for these signatures (13,332 genes in total) (**Fig. 1e, Supplemental Table-11**).

KDs for the 5 top CFS-associated modules were identified from our BN using Key Driver analysis (KDA)[39-42]. The KD analysis predicted 904 KDs within the network, where changes in any of the KDs are predicted to change the molecular states of genes enriched in one or more of the CFS-associated modules. **Fig. 5b-c** highlights KD properties of those KDs that were both global KDs as well as local KDs for CFS modules / signatures. Here we highlight the 11 KDs that were predicted to modulate at least 3 CFS modules/ signatures **(Fig. 5b)**: MXD1, STX3, DYSF, LYN, MLL2, NCOA2, PTPRE, REPS2, RP11-701P16.2, TECPR2, and TUBB1. These KD genes have been shown to be associated with other diseases such as Familial hemophagocytic lymphohistiocytosis 5 (FHL5) [43], chronic myeloid leukemia [44], Dysferlinopathy and inflammatory myopathy [45], Lupus [46], B cell lymphoma [47] as well as immunological aspects of Kabuki syndrome [48], Kaposi's sarcoma-associated herpesvirus [49], modulator of macrophage activation and inflammatory diseases [50], prostate cancer [51], spastic paraparesis [52], and abnormal platelet physiology [53].

To provide experimental support for the KD genes identified from the network, we computed the probability of being loss-of-function intolerant (pLI)[54] across all of the KD genes and compared those scores to non-KDs. The pLI results showed that our KDs predicted to modulate CFS-associated network modules have an increased pLI when compared to the non-KD genes (median pLI for non-KDs = 0.02, median pLI for KDs = 0.68, one-sided Wilcoxon test p value = 1.80e-48), supporting that the KDs are biologically important to the network **(Supplemental Fig. 2, Online Method)**. The combination of high pLI scores and known dysregulation leading to disease states support these CFS KDs as master regulators of vital biological processes associated with CFS. Many of these KDs have been previously shown to be KDs of immunological dysregulation diseases but to the best of our knowledge, have not previously been associated with CFS.

In conclusion, we present an unbiased data driven, network-based approach that identified molecular signatures of CFS and implicates a number of highly coherent co-expression modules to CFS. For the top 5 modules, the complementary analysis shown here point to a common underlying biology that shares immune and metabolic dysregulation also present in other clinically similar diseases such as Lyme, MIS-C, Kawasaki, and recovering COVID-19. Moreover, the top KDs we identified as regulators of these CFS-associated modules are biased towards higher pLI scores, indicating that loss of function mutations in these genes cannot be well tolerated, confirming their critical importance to normal system function. These top KDs we identified for CFS offer interesting points of therapeutic intervention to explore, with the most promising being MXD1, STX3, DYSF, LYN, MLL2, NCOA2, PTPRE, REPS2, RP11-701P16.2, TECPR2, and TUBB1. To help facilitate continued CFS community research we are also providing an interactive website containing the signatures, modules, KDs, and Bayesian network which can be found here: https://irenefp.github.io/bcellm7.html

Figure Captions:

Fig. 1: Study Analysis Workflow.
A: RNA-seq read count data were generated on whole blood and FACs-sorted immune cell samples from CFS cases and controls. B: RNA-seq count data were passed through a viral-clonal detection pipeline. C: RNA-seq count data were passed through our MLFS and DE pipelines, generating predictive signatures of disease. D: Co-expression network construction organized genes into modules, which were annotated for biological pathways and other disease signatures. G: A union of modules with enrichment for CFS and CFS signatures were used with whole blood gene expression in an independent cohort to build a regulatory network where key drivers of disease were identified.

Fig. 2: Viral and Clonal analysis detects dysregulation in CFS.
A: Principal Component Analysis (PCA) of viral load estimated from the whole blood RNAseq data between patients and controls. B: Wilcoxon rank test of the viral mapping mean between patients and controls. C: Kolmogorov–Smirnov distribution test of the viral mapping mean between patients and controls. D: Clonal read support of T and B Cell clones between patients and controls.

Fig. 3: Machine Learning Feature Selection (MLFS) identifies predictive signatures of CFS.
A: Classifier models built from CFS signatures show predictive ROC AUC performance on hold-out test sets across the different cell types (y-axis and color). B: DE analysis for CFS vs HC. X-axis represents the number of DE genes; bars are colored by direction of expression. Y-axis represents the different p value cutoff <0.05 of either Nominal or FDR adjusted p-value. C: CFS signatures were established using MLFS for significance and DE for expression directionality. Signatures are colored per cell type. D: Jaccard index showing signature similarity. E. Top GO enrichment table for signatures, colored per cell type.

Fig. 4: Co-expression network analysis identifies modules of genes dysregulated in CFS and other disease signatures.
A: Table of all co-expression modules significantly associated with CFS signature, colored per cell type. X-axis represents either -log(FDR pvalue) or Odds Ratio (OR) of enrichment for CFS signature. The most significantly associated modules are highlighted with red boxes. B: Enrichment heatmap of top CFS modules and literature signatures from other disease. Y-axis represents disease signatures and are colored and grouped by the category of disease the signature falls into. C: Top 3 functional annotations of top CFS modules. This figure only shows those top modules with significant functional annotations.

Fig. 5: Bayesian Network and Key Driver Analysis identifies regulators of CFS.
A: A 2D representation of the Bayesian network. B and C illustrate key drivers of the network. B: shows the frequency in which a gene is considered a KD. C: shows the DE logFC of the KD gene in the cell specific signatures.

Supplementary Tables:
1-Covid-CFSCommonSymptoms
2-StudySummary
3-ClinicalWorkupQuestionnaire
4-Fukuda
5-Signautures
6-Modules
7-ModuleDiseaseEnrichment
8-ModuleGOAnnotations

9-ModuleMSigHallmarkAnnotations
10-ModuleMSigC7Annotations
11-BayesianNetwork
12-KeyDrivers
13-SignatureRefenences

ACKNOWLEDGEMENTS

ZHG acknowledges funding from grant UL1 TR001863/TR/NCATS NIH HHS/United States.

Methods

-Cohort Generation

The study consists of 15 patients and 15 age, sex, race, and BMI matched control participants who were recruited at the Mount Sinai Hospital. Informed consent was obtained from either the participant or their legally authorized representative, and all study-related activities were conducted under the approval and oversight of Mount Sinai's Institutional Review Board (IRB). Experiment sequence: (1) Clinical evaluation and enrollment; (2) Symptom questionnaire **(Supplemental Tables 3-4)**; (3) Blood draw prior to Cardiopulmonary Exercise Testing (CPET); (4) Standardized CPET and Symptom questionnaire immediately after CPET; (5) Symptom questionnaire and blood draws 24, 48, and 72 hours post-CPET.

Inclusion and Exclusion Criteria:

Inclusion:

- Patients met diagnostic criteria according to:
- Fukuda et al. [19] AND
- The Canadian Consensus Guidelines [20] AND
- The updated international consensus criteria [55]
- 18-57 years of age
- Willing to undergo phlebotomies, one at baseline, one pre-CPET, one 24 hours post CPET, one 48 hours post CPET and one 72 hours post CPET
- Willing and physically able to participate in cardiopulmonary exercise testing, i.e., exclusion of any absolute contraindications

Exclusion:

- Diagnosis of major depressive disorder prior to diagnosis of ME/CFS
- Significant cardiac or pulmonary comorbidities
- Collagen-vascular diseases (CVD) and neuromuscular diseases (NMD)
- Anemia with hemoglobin levels of <10mg/dl
- Active malignancy (but not history thereof)
- Patients unwilling to undergo a wash-out of all immunomodulatory medications for a two-week period prior to cardiopulmonary exercise testing
- Pregnancy
- Any active acute infectious or acute inflammatory state (i.e., viral upper respiratory tract infection, cellulitis, urinary tract infection, sun burn, gastroenteritis, acute diarrheal illness etc.), any current or recent systemic antibiotic therapy (within one month prior to cardiopulmonary exercise testing)
- Inability to perform basic cycle or treadmill ergometer cardiopulmonary exercise testing
- Presence of any absolute contraindications for cardiopulmonary exercise testing:
- Acute myocardial infarction
- Unstable angina
- Uncontrolled arrhythmias causing symptoms or hemodynamic compromise
- Syncope
- Active endocarditis
- Acute myocarditis or pericarditis
- Symptomatic severe aortic stenosis

- Uncontrolled heart failure
- Acute pulmonary embolus or pulmonary infarction
- Thrombosis of lower extremities
- Suspected dissecting aneurysm
- Uncontrolled asthma
- Pulmonary edema
- Room air hypoxemia
- Respiratory failure
- Mental impairment leading to inability to cooperate with testing

-Presence of relative contraindications to cardiopulmonary exercise testing (based on physician assessment at time of testing):

- Severe untreated arterial hypertension
- High degree AV blockade
- Hypertrophic cardiomyopathy
- Significant pulmonary hypertension

Control participants were matched by race, age, sex and BMI. For the matching process, the variables age and BMI were categorized and matched by category rather than continuous value. The age categories were: 18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48-52 and 53-57. The BMI categories were: <19, 19-22, 23-26, 27-30, 31-34, 35-38, 39-42, 43-46 and >46 **(Supplemental Table-2)**. A standardized Control Subject Telephone Screening (CSTS) form was used to determine eligibility of control participants based on matching needs for the study.

-CPET:

Patients were asked to refrain from physical exercise greater than 5 min of walking or the equivalent thereof for 2 days before and 3 days after the CPET (pilot phase) or until the post-exercise time point is reached (either 24, 48 or 72 hours post exercise; main phase). All CPET testing was performed between 8am and 11am. Visual analogue scale (VAS) symptom assessments via modified Fatigue Impact Scale (MFIS) were performed during enrollment, immediately prior, immediately after, as well as 24hrs, 48 hrs and 72hrs after exercise (see Modified Fatigue Impact Scale (MFIS) Sheet in Appendix). Also, on a simplified VAS, global parameters of (1) mental fatigue, (2) physical fatigue and (3) overall body pain were assessed at enrollment, immediately prior to CPET, every 5 minutes during the CPET, immediately after CPET, as well as 24hrs, 48hrs, and 72hrs after CPET (see simplified VAS assessment sheet in Appendix). Patients underwent cycle or treadmill ergometry, with primary option being cycle ergometry due to better signal quality, but with treadmill ergometry being an alternative for patients unable to perform bicycle ergometry. Patients during the first 5 minutes were asked to gradually increase their work rate until reaching 70% of age-predicted maximal heart rate [56], at which point this target heart rate was maintained. Ratings of perceived exertion (RPE) were obtained on a VAS from 1-10 every 5 minutes while undergoing testing (5, 10, 15 and 20min timepoints). Blood pressure measurements and lactic acid measurements (Lactate Pro Portable Analyzer (DKD, Japan) were performed every 5 minutes while exercising. The maximum duration of exercise testing was 25 minutes.

-Peripheral blood processing:

Peripheral blood was collected into Na Heparin vacutainers for immune cell isolation and Tempus RNA vacutainer for total cellular RNA isolation. All blood were processed within 3 hours of phlebotomy. Blood collected in Na Heparin tubes was subjected to a modified Ficoll gradient centrifugation to isolate both peripheral mononuclear cells (PBMCs) and granulocytes. Briefly, following the Ficoll centrifugation, two different cellular fractions were collected, the unique buffy coat layer (PBMCs) and granulocyte on top of the red blood cell layer [57]. PBMCs were washed twice to remove the platelet and used for subsequent four immune cell subset isolations, T cell, B cell, NK cells and monocytes using the Stem cell EasySep system (STEMCELL Technologies Inc.) following the manufacturer's recommendations. The number, viability and the purity of the isolated immune cell subsets were determined by cell counters and flow cytometry as per standard procedures. Only the cells meeting the Quality Control measures were then lysed in RLT buffer with 1% beta-mercapto-ethanol and kept frozen at -80C until commencement of batch sample RNA isolation. The resulting immune cell lysate fractions were: Whole blood cells (from Tempus vacutainer), Granulocytes, Monocytes, B cells, T cells, Natural Killer cells. RNA was extracted using RNeasy kits (Qiagen) and treated with RNAse free DNase-I per manufacturer's instruction (Qiagen) to remove the any contaminated DNA. Quality and Quantity of the RNA were analyzed by Bioanalyzer (Agilent Technologies Inc.) and Nanodrop (ThermoFisher SCIENTIFIC). RNAs were kept frozen stored at -80C until RNAseq library synthesis.

RNA-Seq Library preparation:

About one microgram of total RNA was used for the preparation of the seq library using TruSeq mRNA Seq Kit supplied by Illumina (Cat # 1 FC-122-1001). The protocol followed was as per manufacturer's instruction. Briefly, mRNA was isolated from total RNA using oligo dT on magnetic beads. The mRNA was then be fragmented in the presence of divalent cations at 940C. The fragmented RNA was converted into double stranded cDNA. After polishing the ends of the cDNA, adenine base was added at the 3' ends following which Illumina supplied specific adaptors were ligated. The adaptor ligated DNA was amplified by 15 cycle PCR. The PCR DNA was purified on Ampure beads (Beckman Coulter, part# 63882) to get the final seq library ready for sequencing. The insert size and DNA concentration of the seq library was determined on Agilent Bioanalyzer. Each RNA seq library was layered on one of the eight lanes of the Illumina flow cell at appropriate concentration and bridge amplified to get around 350 million raw reads. The cDNA reads on the flow cell will then sequenced on HiSeq2500 using 100 bp single end recipe. Five barcoded samples were pooled to sequence in one lane.

-Sequencing:

The RNA-Seq libraries were sequenced on the Illumina HiSeq 2500 platform using 100 bp single end protocol following manufacture's recommended procedure. Base calling from Images and fluorescence intensities of the reads were done in situ on the HiSeq 2500 computer using Illumina software. Various QC parameters such as intensities of individual bases, visual and graphic focus quality of the images were monitored periodically to assess the quality of ongoing run. Sequence quality was monitored in terms of colored graphic representation of Q30 values (which is a measure of errors per thousand base), and error rates at 35 and 75 cycles of sequencing were observed to assess the quality of ongoing run. The seq data generated was simultaneously transferred (in real time manner) to high performance computer cluster and processed for Picard based QC analysis. The Seq data passing all QC parameters was used for further analysis.

The raw RNA-seq reads were converted to featureCounts after being aligned to HG19 reference using STAR and the RAPiD [58]pipeline (v2.0.0). Reads not mapped to human HG19 were later analyzed with the Viral Load pipeline described below in these methods. FeatureCounts data were separated into their respective cell types and converted to countspermillion (cpm) with edgeR [59]and filtered so that only genes with >1 cpm in 10% of the data were kept (Bcell=13,262 genes, Granulocytes=10,612 genes, Monocyctes=11,947 genes, NK cells=13,391 genes, Tcells=13,347 genes, Whole Blood=13,415 genes). Gene counts were normalized using the limma [60] functions calcNormFactors(method="TMM") and voom() per cell type.

We used principal component analysis prcomp() [61] and variancePartition [62] package to explore and visualize the gene count expression variance from covariates. Due to the small samples size and the age, sex, BMI, matched study design, no covariates were adjusted for during normalization.

-Differential Expression

The limma package [60] was used to perform differential expression (DE) analysis performed on the RNA-seq normalized counts data per cell type. The function duplicateCorrelation() was used to mitigate multiple measurements from the separate timepoints, followed by lmfit(), contrasts.fit(), and eBayes(). DE analysis revealed no significant (FDR<0.05) genes were found to be differentially expressed among any timepoint pairs within CFS and control status. Due to this result, timepoint data was treated as biological replicates for all further analysis. DE analysis did reveal differentially expressed genes between CFS and control status within each cell type.

-Machine Learning Pipeline

We used machine learning classifiers to unbiasedly reduce dimensionality in the RNA-seq normalized counts data to genes most informative to accurately classify CFS and control samples, known as machine learning feature selection (MLFS) [63] **(Supplemental Fig. 1, Supplemental Table-5)**. This pipeline was performed on each cell type and type point separately to reduce data leakage in the training/testing splits. MLFS was performed in 5-Cross Test Folds, where the data were randomly split into training (24 samples) and testing (6 samples), and stratified by class as to not generate imbalance [64]. The training data were further split in a 5-fold cross validation resulting in trainingII and validation. Random forest classifiers (python sklearn package [65, 66]) were fit on each of the 5 trainingII splits, where the number of trees in the forest is equal to the number of gene features in the data set, resulting in large random forests that gives most, if not all, genes a feature importance (information gained) score [67]. A Bonferroni adjusted p-value was fit to the information gained scores and genes with adjusted pvalue <0.05 were considered significant [68]. This was done for each of the 5 trainingII subsets, results in 5 lists of significant features per cell type, per timepoint. The union of the 5 lists of significant features were pushed to the next step where a new random forest classifier was then fit on the entire training data samples, but with only the significant features found in the previous step. This new classifier was then tested against the testing data and classifier performance was recorded **(Fig. 3a)**. The final result is 5 distinct classifier models, built from each of the 5-Cross Test Folds, along with feature importance ranked lists of all gene features in the dataset. By setting up our machine learning pipeline this way, we used classifiers to find gene features that consistently separate patients from controls, regardless of random effects of splitting the dataset. This was done independently for each timepoint and each cell type so that at no time could an individual be in both the training and testing during the same run through.

-Weighted Gene Coexpression Network Analysis (WGCNA)

WGCNA co-expression networks were built upon each of the different cell types independently, treating biological replicates as independent samples, resulting in 6 co-expression networks **(Supplemental Table-6)** [69]. Network building parameters included softpower thresholds of 14 in the adjacency() function, dynamic tree module cutting, and no samples were excluded from sample trees. To identify modules enriched for CFS disease signal, we calculated enrichment statistics using Fisher's Exact Test (fisher.test() [61]), and corrected for multi-testing following Benjamini-Hochberg procedure [70] (p.adjust () [61])of the MLFS signature genes on each of the co-expression networks **(Supplemental Table-7)**.

-Pathway Enrichment Analyses

Pathway enrichment for GO [71, 72], C7, and Hallmark [73] pathways (N=10,192, 4872, and 50 pathways, respectively) was performed on the MLFS CFS signatures and all co-expression modules **(Supplemental Tables 8-10).** Enrichment was performed separately for each cell type modules and signatures as each cell type had different dataset background (Bcell=13,262 genes, Granulocytes=10,612 genes, Monocyctes=11,947 genes, NK cells=13,391 genes, Tcells=13,347 genes, Whole Blood=13,415 genes) and Benjamini-Hochberg multiple testing correction was performed for each cell type separately as well using p.adjust() function in R [61]. GO enrichment was performed using the R packages goseq [74], topGO [75] and org.Hs.eg.db [76] while C7 and Hallmark enrichment was performed using the R packages HTSanalyzeR [77], GSEABase [78], and GAGE [79].

-B cell and T cell Clonal Detection / Viral Load

To recover information on the clonal composition of T and B-cells, we used the software MIXCR with default parameters over RNA-seq fastq files [23]. Briefly, the tool performs an alignments of sequencing reads to reference V, D, J and C genes of T- or B- cell receptors and filters sequences based on read quality. Next, the assembly of clonotypes is performed using the previous alignments including calibration and correction based on upstream and downstream regions of the target genes in order to extract specific CDR3 regions. Finally, the alignments, clonotypes, and molecular identifiers are exported into a tab delimited text file. The outputs are carefully reviewed and a threshold of at least 10 reads per clonotype is set for further downstream analysis. Visualization and figures are done with ggplot [80] and cowplot [81]packages available in R statistical software [61].

To identify reads that correspond to possible viral species, we used the non-human or non-mapping reads from the main alignment used in this study. First, these reads were assembled into contigs using the aligner inchworm from Trinity Suite [82] stopping before the chrysalis step. Next, we filtered out the alignment for contigs smaller than 200 base pairs. Also, we further refine the contigs by consolidating contig clusters with a similarity of 95% or above by using the software cd-hit [83] with the parameters "-c 0.95 -n 10 -T 2 -p 1 -g 1". Next, we perform alignment with diamond suite [84] against the NCBI viral genome assemblies with the following parameters: "-f 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qframe salltitles -k 1". Finally, the matching alignments were combined with the identifiers per contig in the R statistical software [61] and used in downstream analysis with the limma [60]/edgeR [59, 85] linear modeling and ggplot2 [80]/cowplot [81]packages.

-Bayesian Causal Network

209 healthy whole blood RNA-seq samples (residual counts matrix after adjusting for demographics of age, gender and genetic PC's, and technical covariates RIN, processing batch and ribosomal RNA rate) were acquired from the Mount Sinai Crohn's and Colitis Registry [86] (MSCCR) to be used in the Bayesian Network **(Supplemental Table-11)**. The Bayesian Network was built using RIMBANET [87-90]. Seeding gene list was generated with the union of MLFS signature genes and genes found in all co-expression modules enriched for MLFS signatures, resulting in a seeding gene list of 13,332 genes. Reducing the network search space to the seeding gene list is important for increasing the the fit of the network. MSCCR healthy RNA-seq blood sample expression, with their associated eQTLs were used as priors to build the network [86].

-Interactive Network Visualization

For intuitive visual exploration of the generated dense and complex networks, we have developed a web-based interactive CFS network exploration portal, which builds from our previous work on interactive network visualization [91, 92]. The portal provides a customized exploration environment for the network and other associated meta-data. Briefly, we have implemented the 3D capability by building with Three.js, which is a cross-browser JavaScript library that incorporates the WebGL (Web Graphics Library) API. We have calculated the network layout using the 3d-force-graph, which is a library for Three.js. In addition, we have utilized client-side Javascript libraries d3.js and jQuery in real time. For the styling of web interface elements, we have primarily utilized custom HTML and CSS. The portal utilizes only standard libraries and does not require the use of any additional plug-ins, which enables it to run on all modern web browsers. Furthermore, users can easily define filters on clinical variables (modules, key drivers). The portal is an open-access freely available resource for the dissemination of all our network analysis results to the scientific community. The portal can be accessed at: https://irenefp.github.io/bcellm7.html.

-Key Driver Analysis

Key Driver Analysis was performed using KDA [93] **(Supplemental Table-12)**. For local key drivers, this package defines a background sub-network by looking for a neighborhood K-step away from each node in the target gene list in the network. Stemming from each node in this sub-network, it assesses the enrichment in its k-step (k varies from 1 to K) downstream neighborhood for the target gene list. In this analysis, we used K=6. Target gene lists used for KDA included MLFS signature genes and enriched co-expression module genes. This package defines global drivers or hub genes as genes with number of neighbors exceeding that of the mean of gene neighbors + 2 standard deviations of gene neighbors of all genes in the network.

-pLI

We mapped probability of being LoF intolerant (pLI) score [54] to all genes in the Bayesian Network **(Supplemental Fig. 2)**. To test for differences between KD and non KD genes, we used a one-sided Wilcoxon test.

-Curated External Disease Signatures

Curated external disease signatures from the literature were used to gain further insight into shared disease etiology in our co-expression network. Further signature information can be found in **Supplemental Table-13**. The following signatures were used in our analyses:

- (Nguyen, Whole Blood) CFS vs HC [94]
- (Kaushik, PBMCs) CFS vs HC [95]
- (Gow, PBMCs) CFS vs HC [96]
- (Beckmann, Whole Blood) MIS-C vs HC [97]
- (Wright, Whole Blood) KD vs HC[98]
- (Canna, Whole Blood) MAS vs HC, (Canna, Whole Blood) NOMID vs HC [99]

-Lyme vs HC, (Bouquet, PBMCs) Treatment Lyme vs HC, (Bouquet, PBMCs) 6m Post Lyme vs HC [100]

- (Blanco-Melo, NHBE cells) IAV vs Mock [101]
- (Ramilo, PBMC) IAV vs Bacterial Infection [102]
- (Thair, Whole Blood) COVID-19 vs HC [103]
- (Xiong, PBMCs) COVID-19 vs HC [104]
- (Wen, ASCs) ERS COVID-19 vs HC, (Wen, DCs) ERS COVID-19 vs HC, (Wen, Proliferating Tcells) ERS COVID-19 vs HC [105]
- (Barturen, Whole Blood) HC MCTD vs, (Barturen, Whole Blood) SJS vs HC, (Barturen, Whole Blood) SLE vs HC, (Barturen, Whole Blood) SSC vs HC, (Barturen, Whole Blood) UCTD vs HC, (Barturen, Whole Blood) PAPS vs HC, Barturen, Whole Blood) RA vs HC [106]

-Miscellaneous Statistical/Computational Analyses

R analysis in this paper were performed using R version 4.0.1. R package libraries include:
data.table (v1.13.0)
edgeR (v3.30.3)
limma (v3.44.3)

variancePartition (v1.19.4)
WGCNA (v1.69)
goseq (v1.40.0)
topGO (v2.40.0)
org.Hs.eg.db (v3.11.4)
HTSanalyzeR (v2.34.1)
GSEABase (v1.50.1)
GAGE (v2.38.3)
ggplot (v3.3.2)
cowplot (v1.1.0)
RIMBANET
KDA

Python analysis for the MLFS was performed using python version 3.7.3. Python packages include:

numpy (v1.16.4)
pandas (v0.24.2)
scikit-learn (v0.21.2)

REFERENCES

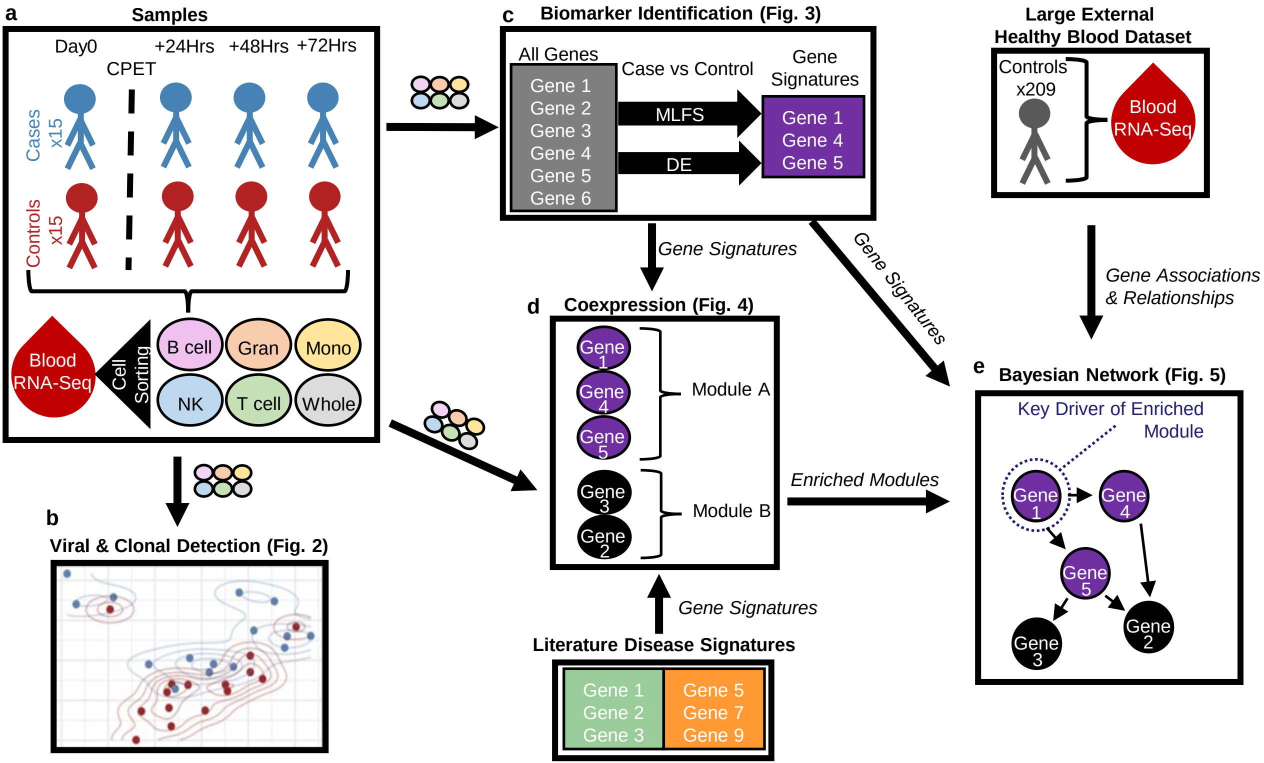
1. Rowe, P.C., et al., *Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Diagnosis and Management in Young People: A Primer.* Front Pediatr, 2017. **5**: p. 121.
2. Greenhalgh, T., et al., *Management of post-acute covid-19 in primary care.* BMJ, 2020. **370**: p. m3026.
3. Rubin, R., *As Their Numbers Grow, COVID-19 "Long Haulers" Stump Experts.* JAMA, 2020. **324**(14): p. 1381-1383.
4. CDC. [Website] 2018; Available from: https://www.cdc.gov/me-cfs/about/possible-causes.html.
5. Tom Whipple, O.M., *Scientists trade insults over myalgic encephalomyelitis (ME) study.* The Times, 2017.
6. Geraghty, K., et al., *The 'cognitive behavioural model' of chronic fatigue syndrome: Critique of a flawed model.* Health Psychol Open, 2019. **6**(1): p. 2055102919838907.
7. Kogelnik, A.M., et al., *Use of valganciclovir in patients with elevated antibody titers against Human Herpesvirus-6 (HHV-6) and Epstein-Barr Virus (EBV) who were experiencing central nervous system dysfunction including long-standing fatigue.* Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology, 2006. **37 Suppl 1**: p. S33-8.
8. Hickie, I., et al., *Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: prospective cohort study.* BMJ, 2006. **333**(7568): p. 575.
9. Landay, A.L., et al., *Chronic fatigue syndrome: clinical condition associated with immune activation.* Lancet, 1991. **338**(8769): p. 707-12.
10. Brenu, E.W., et al., *Natural killer cells in patients with severe chronic fatigue syndrome.* Auto Immun Highlights, 2013. **4**(3): p. 69-80.
11. Rivas, J.L., et al., *Association of T and NK Cell Phenotype With the Diagnosis of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS).* Front Immunol, 2018. **9**: p. 1028.
12. Yang, T., et al., *The clinical value of cytokines in chronic fatigue syndrome.* J Transl Med, 2019. **17**(1): p. 213.
13. Lidbury, B.A., et al., *Rethinking ME/CFS Diagnostic Reference Intervals via Machine Learning, and the Utility of Activin B for Defining Symptom Severity.* Diagnostics (Basel), 2019. **9**(3).
14. Grans, H., et al., *Reduced levels of oestrogen receptor beta mRNA in Swedish patients with chronic fatigue syndrome.* Journal of clinical pathology, 2007. **60**(2): p. 195-8.
15. Maes, M., I. Mihaylova, and J.C. Leunis, *In chronic fatigue syndrome, the decreased levels of omega-3 poly-unsaturated fatty acids are related to lowered serum zinc and defects in T cell activation.* Neuro endocrinology letters, 2005. **26**(6): p. 745-51.
16. Nepotchatykh, E., et al., *Profile of circulating microRNAs in myalgic encephalomyelitis and their relation to symptom severity, and disease pathophysiology.* Sci Rep, 2020. **10**(1): p. 19620.
17. Fluge, O., et al., *B-Lymphocyte Depletion in Myalgic Encephalopathy/ Chronic Fatigue Syndrome. An Open-Label Phase II Study with Rituximab Maintenance Treatment.* PLoS One, 2015. **10**(7): p. e0129898.
18. Bolton, M.J., B.P. Chapman, and H. Van Marwijk, *Low-dose naltrexone as a treatment for chronic fatigue syndrome.* BMJ Case Rep, 2020. **13**(1).
19. Fukuda, K., et al., *The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group.* Annals of internal medicine, 1994. **121**(12): p. 953-9.

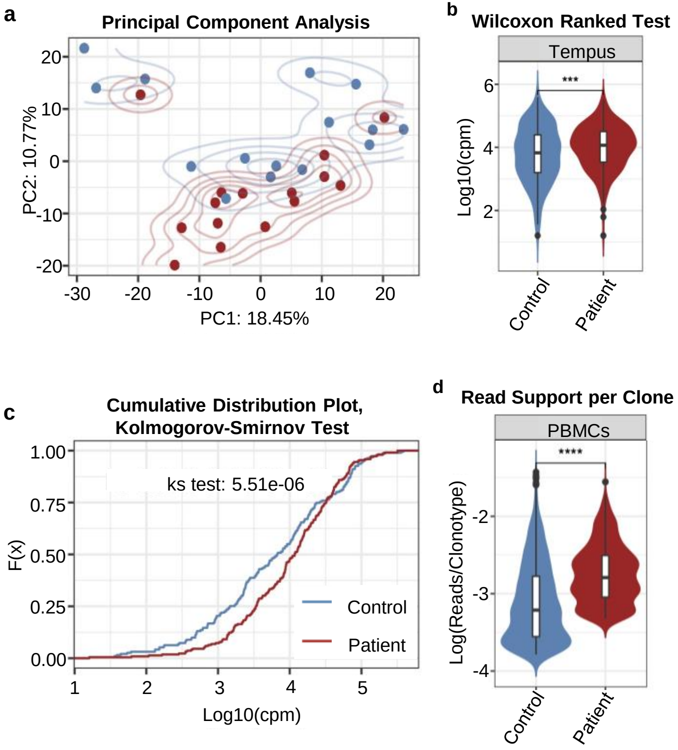
20. Carruthers, B.M., *Definitions and aetiology of myalgic encephalomyelitis: how the Canadian consensus clinical definition of myalgic encephalomyelitis works.* J Clin Pathol, 2007. **60**(2): p. 117-9.
21. Keech, A., et al., *Gene Expression in Response to Exercise in Patients with Chronic Fatigue Syndrome: A Pilot Study.* Front Physiol, 2016. **7**: p. 421.
22. Bouquet, J., et al., *Whole blood human transcriptome and virome analysis of ME/CFS patients experiencing post-exertional malaise following cardiopulmonary exercise testing.* PLoS One, 2019. **14**(3): p. e0212193.
23. Bolotin, D.A., et al., *MiXCR: software for comprehensive adaptive immunity profiling.* Nat Methods, 2015. **12**(5): p. 380-1.
24. *Long COVID: let patients help define long-lasting COVID symptoms.* Nature, 2020. **586**(7828): p. 170.
25. Patrick, D.M., et al., *Lyme Disease Diagnosed by Alternative Methods: A Phenotype Similar to That of Chronic Fatigue Syndrome.* Clin Infect Dis, 2015. **61**(7): p. 1084-91.
26. Huth, T.K., D. Staines, and S. Marshall-Gradisnik, *ERK1/2, MEK1/2 and p38 downstream signalling molecules impaired in CD56 dim CD16+ and CD56 bright CD16 dim/- natural killer cells in Chronic Fatigue Syndrome/Myalgic Encephalomyelitis patients.* J Transl Med, 2016. **14**: p. 97.
27. Bouhaddou, M., et al., *The Global Phosphorylation Landscape of SARS-CoV-2 Infection.* Cell, 2020. **182**(3): p. 685-712 e19.
28. Begon, E., *[Lyme arthritis, Lyme carditis and other presentations potentially associated to Lyme disease].* Med Mal Infect, 2007. **37**(7-8): p. 422-34.
29. Hawley, K., et al., *Macrophage p38 mitogen-activated protein kinase activity regulates invariant natural killer T-cell responses during Borrelia burgdorferi infection.* J Infect Dis, 2012. **206**(2): p. 283-91.
30. Di Domenico, E.G., et al., *The Emerging Role of Microbial Biofilm in Lyme Neuroborreliosis.* Front Neurol, 2018. **9**: p. 1048.
31. Berg, D., et al., *Chronic fatigue syndrome and/or fibromyalgia as a variation of antiphospholipid antibody syndrome: an explanatory model and approach to laboratory diagnosis.* Blood Coagul Fibrinolysis, 1999. **10**(7): p. 435-8.
32. Javid, A., et al., *Hyperglycemia Impairs Neutrophil-Mediated Bacterial Clearance in Mice Infected with the Lyme Disease Pathogen.* PLoS One, 2016. **11**(6): p. e0158019.
33. Kennedy, G., et al., *Increased neutrophil apoptosis in chronic fatigue syndrome.* J Clin Pathol, 2004. **57**(8): p. 891-3.
34. Biezeveld, M.H., et al., *Sustained activation of neutrophils in the course of Kawasaki disease: an association with matrix metalloproteinases.* Clin Exp Immunol, 2005. **141**(1): p. 183-8.
35. Gruber, C.N., et al., *Mapping Systemic Inflammation and Antibody Responses in Multisystem Inflammatory Syndrome in Children (MIS-C).* Cell, 2020.
36. Burns, J.C., et al., *Coagulopathy and platelet activation in Kawasaki syndrome: identification of patients at high risk for development of coronary artery aneurysms.* J Pediatr, 1984. **105**(2): p. 206-11.
37. Hennon, T.R., et al., *COVID-19 associated Multisystem Inflammatory Syndrome in Children (MIS-C) guidelines; a Western New York approach.* Prog Pediatr Cardiol, 2020: p. 101232.
38. Cohain, A.T., et al., *An integrative multiomic network model links lipid metabolism to glucose regulation in coronary artery disease.* Nat Commun, 2021. **12**(1): p. 547.
39. Peters, L.A., et al., *A functional genomics predictive network model identifies regulators of inflammatory bowel disease.* Nat Genet, 2017. **49**(10): p. 1437-1449.
40. Beckmann, N.D., et al., *Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease.* Nat Commun, 2020. **11**(1): p. 3942.

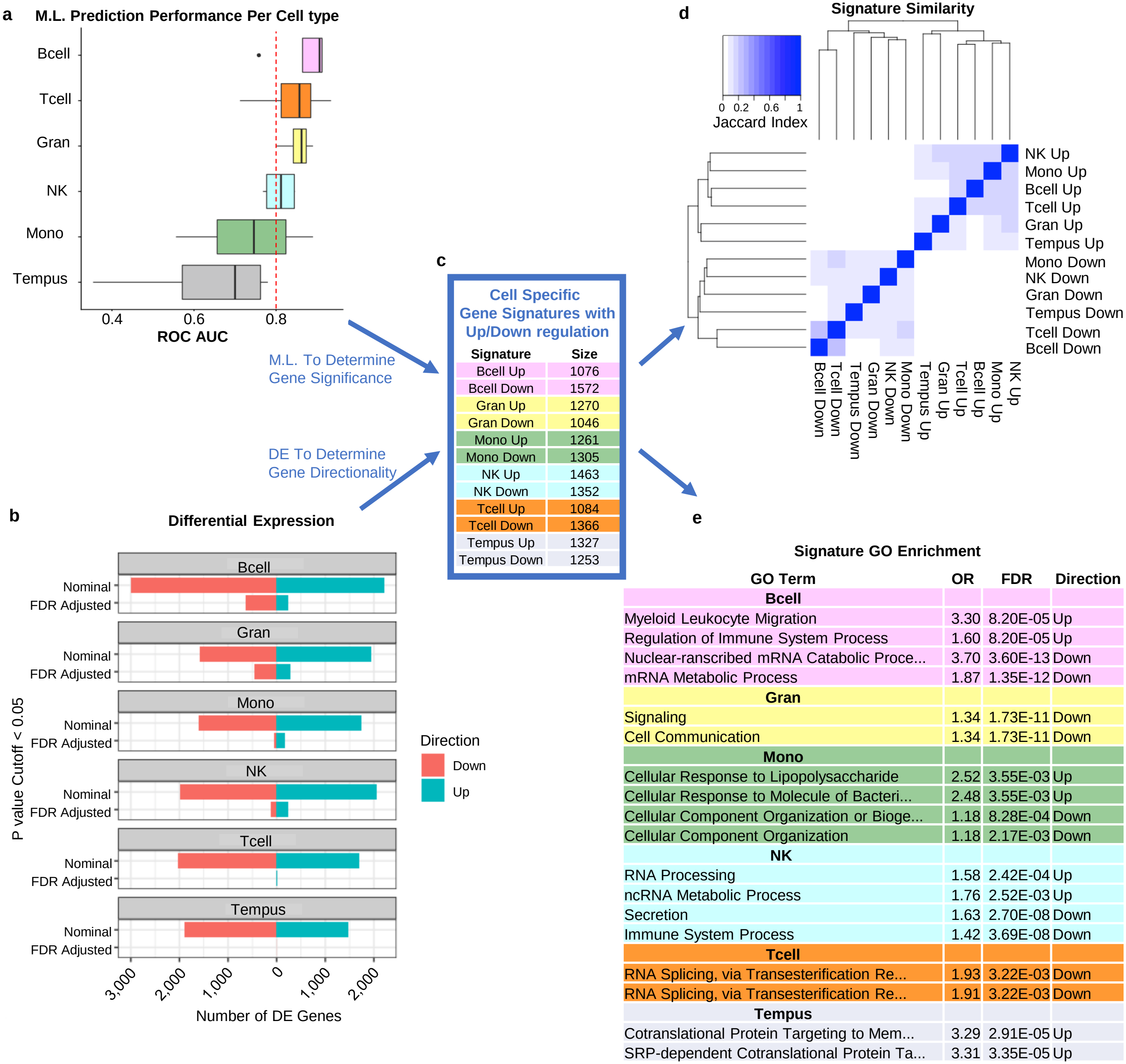
41. Zhang, B., et al., *Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.* Cell, 2013. **153**(3): p. 707-20.
42. Wang, M., et al., *Transformative Network Modeling of Multi-omics Data Reveals Detailed Circuits, Key Regulators, and Potential Therapeutics for Alzheimer's Disease.* Neuron, 2020.
43. !!! INVALID CITATION !!! [33].
44. !!! INVALID CITATION !!! [34].
45. Tang, J., et al., *A novel mutation in the DYSF gene in a patient with a presumed inflammatory myopathy.* Neuropathology, 2018.
46. Ban, T., et al., *Lyn Kinase Suppresses the Transcriptional Activity of IRF5 in the TLR-MyD88 Pathway to Restrain the Development of Autoimmunity.* Immunity, 2016. **45**(2): p. 319-32.
47. Zhang, J., et al., *Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis.* Nat Med, 2015. **21**(10): p. 1190-8.
48. Lin, J.L., et al., *Immunologic assessment and KMT2D mutation detection in Kabuki syndrome.* Clin Genet, 2015. **88**(3): p. 255-60.
49. Wei, X., et al., *NCOA2 promotes lytic reactivation of Kaposi's sarcoma-associated herpesvirus by enhancing the expression of the master switch protein RTA.* PLoS Pathog, 2019. **15**(11): p. e1008160.
50. Han, X., et al., *LncRNA PTPRE-AS1 modulates M2 macrophage activation and inflammatory diseases by epigenetic promotion of PTPRE.* Sci Adv, 2019. **5**(12): p. eaax9230.
51. Oosterhoff, J.K., et al., *REPS2/POB1 is downregulated during human prostate cancer progression and inhibits growth factor signalling in prostate cancer cells.* Oncogene, 2003. **22**(19): p. 2920-5.
52. Oz-Levi, D., et al., *Mutation in TECPR2 reveals a role for autophagy in hereditary spastic paraparesis.* Am J Hum Genet, 2012. **91**(6): p. 1065-72.
53. Stoupa, A., et al., *TUBB1 mutations cause thyroid dysgenesis associated with abnormal platelet physiology.* EMBO Mol Med, 2018. **10**(12).
54. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
55. Carruthers, B.M., et al., *Myalgic encephalomyelitis: International Consensus Criteria.* Journal of internal medicine, 2011. **270**(4): p. 327-38.
56. Zhu, N., et al., *Longitudinal examination of age-predicted symptom-limited exercise maximum HR.* Medicine and science in sports and exercise, 2010. **42**(8): p. 1519-27.
57. Fuss, I.J., et al., *Isolation of whole mononuclear cells from peripheral blood and cord blood.* Curr Protoc Immunol, 2009. **Chapter 7**: p. Unit7 1.
58. Shah, H.e.a., *RAPiD—an agile and dependable RNA-Seq framework.* ASHG2015, 2015. **PgmNr 1856**.
59. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
60. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
61. Team, R.C., *R: A language and environment for statistical computing.* . 2017, R Foundation for Statistical Computing.
62. Hoffman, G.E. and E.E. Schadt, *variancePartition: interpreting drivers of variation in complex gene expression studies.* BMC Bioinformatics, 2016. **17**(1): p. 483.
63. James, G., et al., *An introduction to statistical learning : with applications in R*. Springer texts in statistics,. 2013, New York: Springer. xvi, 426 pages.

64. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. 2nd ed. Springer series in statistics,. 2009, New York, NY: Springer. xxii, 745 p.
65. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
66. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
67. Kullback, S. and R.A. Leibler, *On Information and Sufficiency.* Annals of Mathematical Statistics, 1951. **22**(1): p. 79-86.
68. Dubitzky, W., et al., *Encyclopedia of systems biology*. 2013, New York: Springer Reference. 4 volumes (xlvii, 2366 pages).
69. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
70. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Stat Med, 1990. **9**(7): p. 811-8.
71. al, A.e., *Gene ontology: tool for the unification of biology.* Nat Genet, 2000. **25**.
72. The Gene Ontology, C., *The Gene Ontology Resource: 20 years and still GOing strong.* Nucleic Acids Res, 2019. **47**(D1): p. D330-D338.
73. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
74. Young, M.D., et al., *Gene ontology analysis for RNA-seq: accounting for selection bias.* Genome Biol, 2010. **11**(2): p. R14.
75. Alexa A, R.J., *topGO: Enrichment Analysis for Gene Ontology*. 2020.
76. M, C., *org.Hs.eg.db: Genome wide annotation for Human*.
77. Wang, X., et al., *HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens.* Bioinformatics, 2011. **27**(6): p. 879-80.
78. Morgan M, F.S.a.G.R., *GSEABase: Gene set enrichment data structures and methods*.
79. Luo, W., et al., *GAGE: generally applicable gene set enrichment for pathway analysis.* BMC Bioinformatics, 2009. **10**: p. 161.
80. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis.* Ggplot2: Elegant Graphics for Data Analysis, 2009: p. 1-212.
81. Wilke, C.O., *cowplot – Streamlined plot theme and plot annotations for ggplot2*. 2020.
82. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
83. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-9.
84. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
85. McCarthy, D.J., Y. Chen, and G.K. Smyth, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.* Nucleic Acids Res, 2012. **40**(10): p. 4288-97.
86. Suarez-Farinas, M., et al., *Intestinal inflammation modulates the expression of ACE2 and TMPRSS2 and potentially overlaps with the pathogenesis of SARS-CoV-2 related disease.* Gastroenterology, 2020.
87. Zhu, J., et al., *An integrative genomics approach to the reconstruction of gene networks in segregating populations.* Cytogenet Genome Res, 2004. **105**(2-4): p. 363-74.
88. Zhu, J., et al., *Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation.* PLoS Biol, 2012. **10**(4): p. e1001301.
89. Zhu, J., et al., *Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations.* PLoS Comput Biol, 2007. **3**(4): p. e69.

90. Zhu, J., et al., *Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.* Nat Genet, 2008. **40**(7): p. 854-61.
91. Liluashvili, V., et al., *iCAVE: an open source tool for visualizing biomolecular networks in 3D, stereoscopic 3D and immersive 3D.* Gigascience, 2017. **6**(8): p. 1-13.
92. Kalayci, S. and Z.H. Gumus, *Exploring Biological Networks in 3D, Stereoscopic 3D, and Immersive 3D with iCAVE.* Curr Protoc Bioinformatics, 2018. **61**(1): p. 8 27 1-8 27 26.
93. Zhang, B. and J. Zhu, *Identification of Key Causal Regulators in Gene Networks.* World Congress on Engineering - Wce 2013, Vol Ii, 2013: p. 1309-1312.
94. Nguyen, C.B., et al., *Whole blood gene expression in adolescent chronic fatigue syndrome: an exploratory cross-sectional study suggesting altered B cell differentiation and survival.* J Transl Med, 2017. **15**(1): p. 102.
95. Kaushik, N., et al., *Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome.* J Clin Pathol, 2005. **58**(8): p. 826-32.
96. Gow, J.W., et al., *A gene signature for post-infectious chronic fatigue syndrome.* BMC Med Genomics, 2009. **2**: p. 38.
97. Beckmann, N.D., et al., *Cytotoxic lymphocytes are dysregulated in multisystem inflammatory syndrome in children.* medRxiv, 2020.
98. Wright, V.J., et al., *Diagnosis of Kawasaki Disease Using a Minimal Whole-Blood Gene Expression Signature.* JAMA Pediatr, 2018. **172**(10): p. e182293.
99. Canna, S.W., et al., *An activating NLRC4 inflammasome mutation causes autoinflammation with recurrent macrophage activation syndrome.* Nat Genet, 2014. **46**(10): p. 1140-6.
100. Bouquet, J., et al., *Longitudinal Transcriptome Analysis Reveals a Sustained Differential Gene Expression Signature in Patients Treated for Acute Lyme Disease.* mBio, 2016. **7**(1): p. e00100-16.
101. Blanco-Melo, D., et al., *Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19.* Cell, 2020. **181**(5): p. 1036-1045 e9.
102. Ramilo, O., et al., *Gene expression patterns in blood leukocytes discriminate patients with acute infections.* Blood, 2007. **109**(5): p. 2066-77.
103. Thair, S.A., et al., *Transcriptomic Similarities and Differences in Host Response between SARS-CoV-2 and Other Viral Infections.* medRxiv, 2020: p. 2020.06.18.20131326.
104. Xiong, Y., et al., *Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients.* Emerg Microbes Infect, 2020. **9**(1): p. 761-770.
105. Wen, W., et al., *Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing.* Cell Discov, 2020. **6**: p. 31.
106. Barturen, G., et al., *Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases.* medRxiv, 2020: p. 2020.02.21.20021618.


a
Samples
Day0
CPET
+24Hrs
+48Hrs
+72Hrs
Cases x15
Controls x15
Blood RNA-Seq
Cell Sorting
B cell
Gran
Mono
NK
T cell
Whole
b
Viral & Clonal Detection (Fig. 2)
c
Biomarker Identification (Fig. 3)
All Genes
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Case vs Control
MLFS
DE
Gene Signatures
Gene 1
Gene 4
Gene 5
Large External Healthy Blood Dataset
Controls x209
Blood RNA-Seq
Gene Signatures
Gene Signatures
Gene Associations & Relationships
d
Coexpression (Fig. 4)
Gene 1
Gene 4
Gene 5
Module A
Gene 3
Gene 2
Module B
Enriched Modules
Gene Signatures
Literature Disease Signatures
Gene 1
Gene 2
Gene 3
Gene 5
Gene 7
Gene 9
e
Bayesian Network (Fig. 5)
Key Driver of Enriched Module
Gene 1
Gene 4
Gene 5
Gene 3
Gene 2


a
Principal Component Analysis
PC2: 10.77%
20
10
0
-10
-20
-30
-20
-10
0
10
20
PC1: 18.45%
b
Wilcoxon Ranked Test
Tempus
***
Log10(cpm)
6
4
2
Control
Patient
c
Cumulative Distribution Plot,
Kolmogorov-Smirnov Test
ks test: 5.51e-06
F(x)
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Log10(cpm)
Control
Patient
d
Read Support per Clone
PBMCs
****
Log(Reads/Clonotype)
-2
-3
-4
Control
Patient

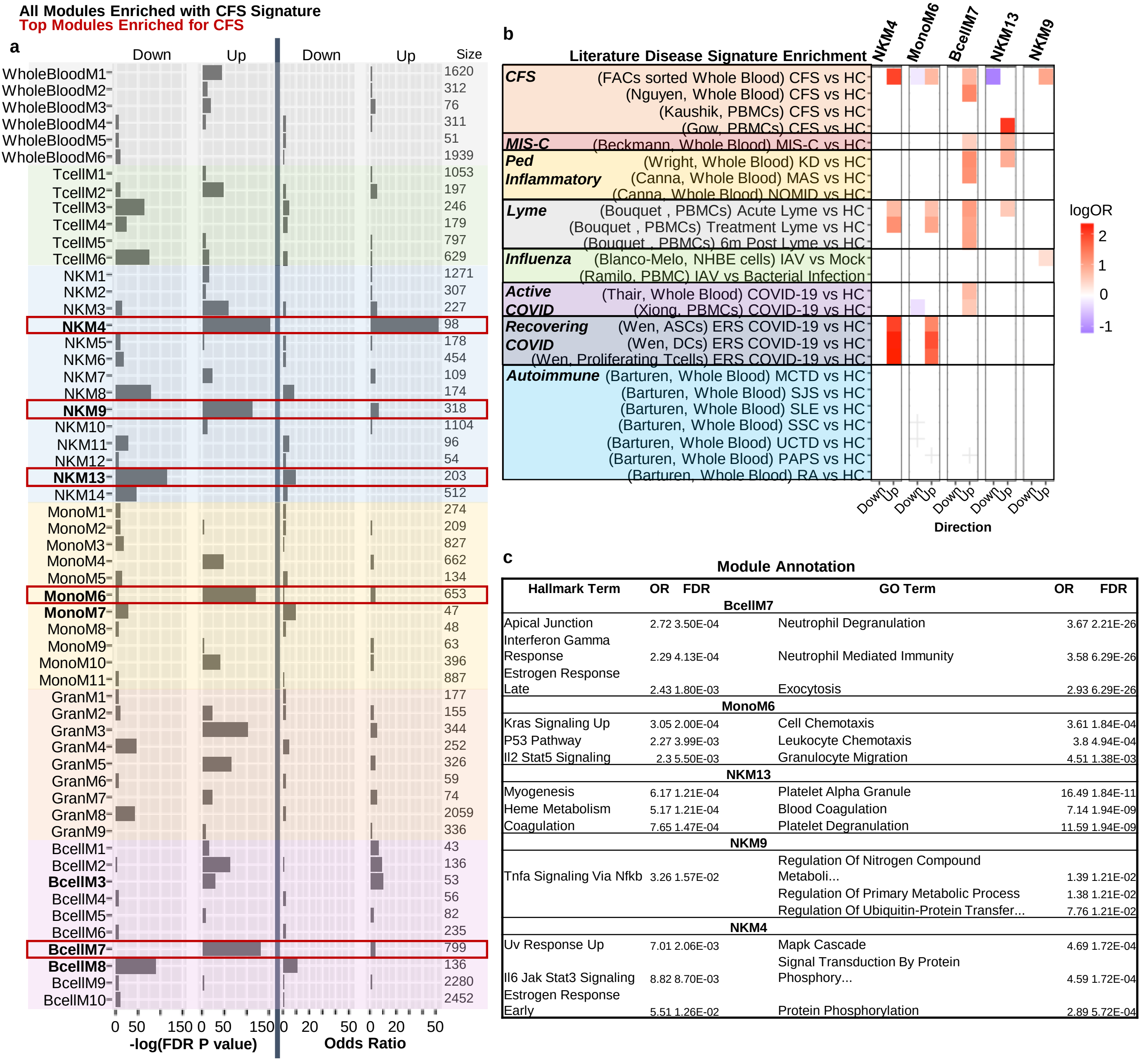


Cell Specific Gene Signatures with Up/Down regulation

Signature	Size
Bcell Up	1076
Bcell Down	1572
Gran Up	1270
Gran Down	1046
Mono Up	1261
Mono Down	1305
NK Up	1463
NK Down	1352
Tcell Up	1084
Tcell Down	1366
Tempus Up	1327
Tempus Down	1253

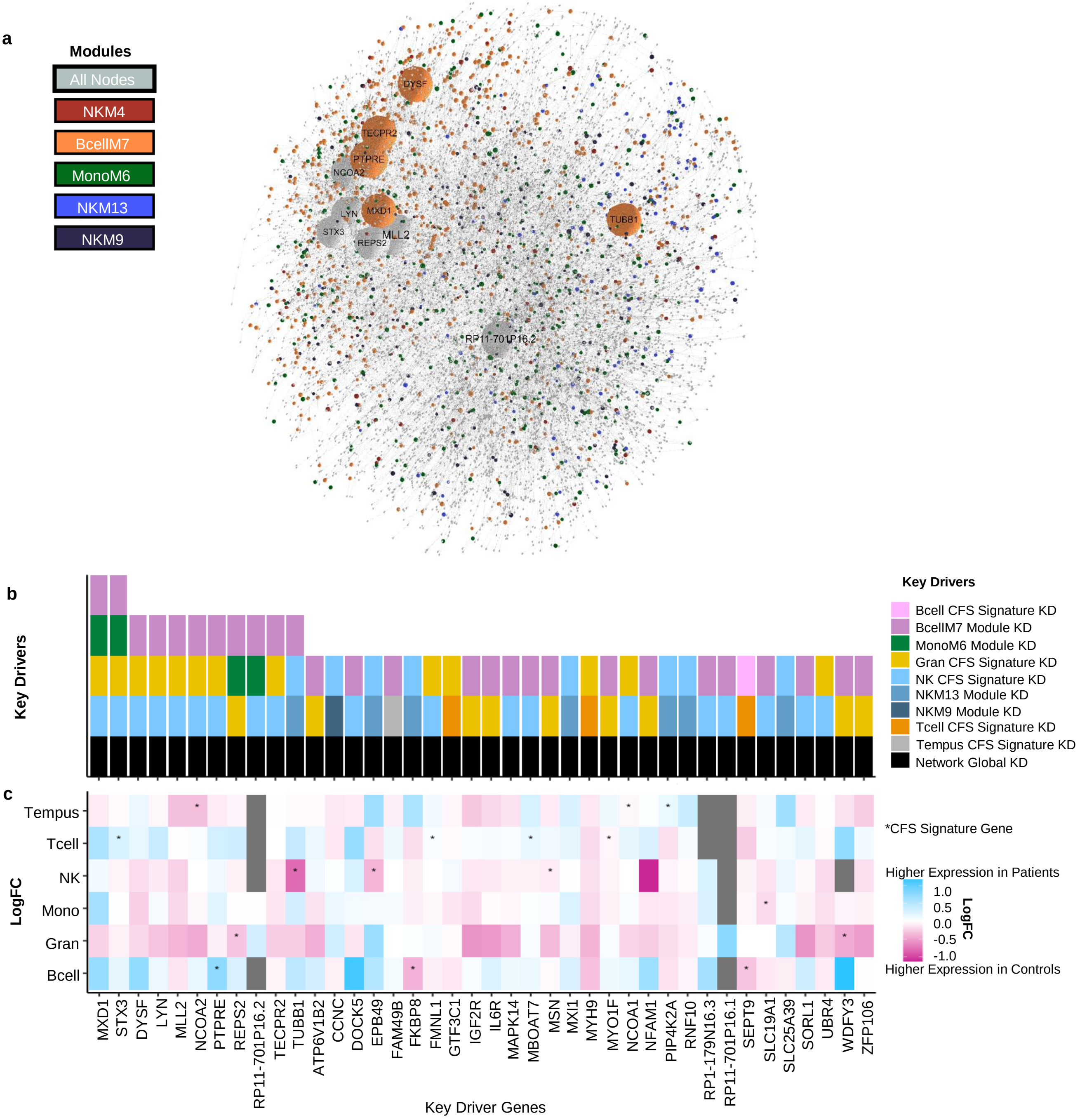
Signature GO Enrichment

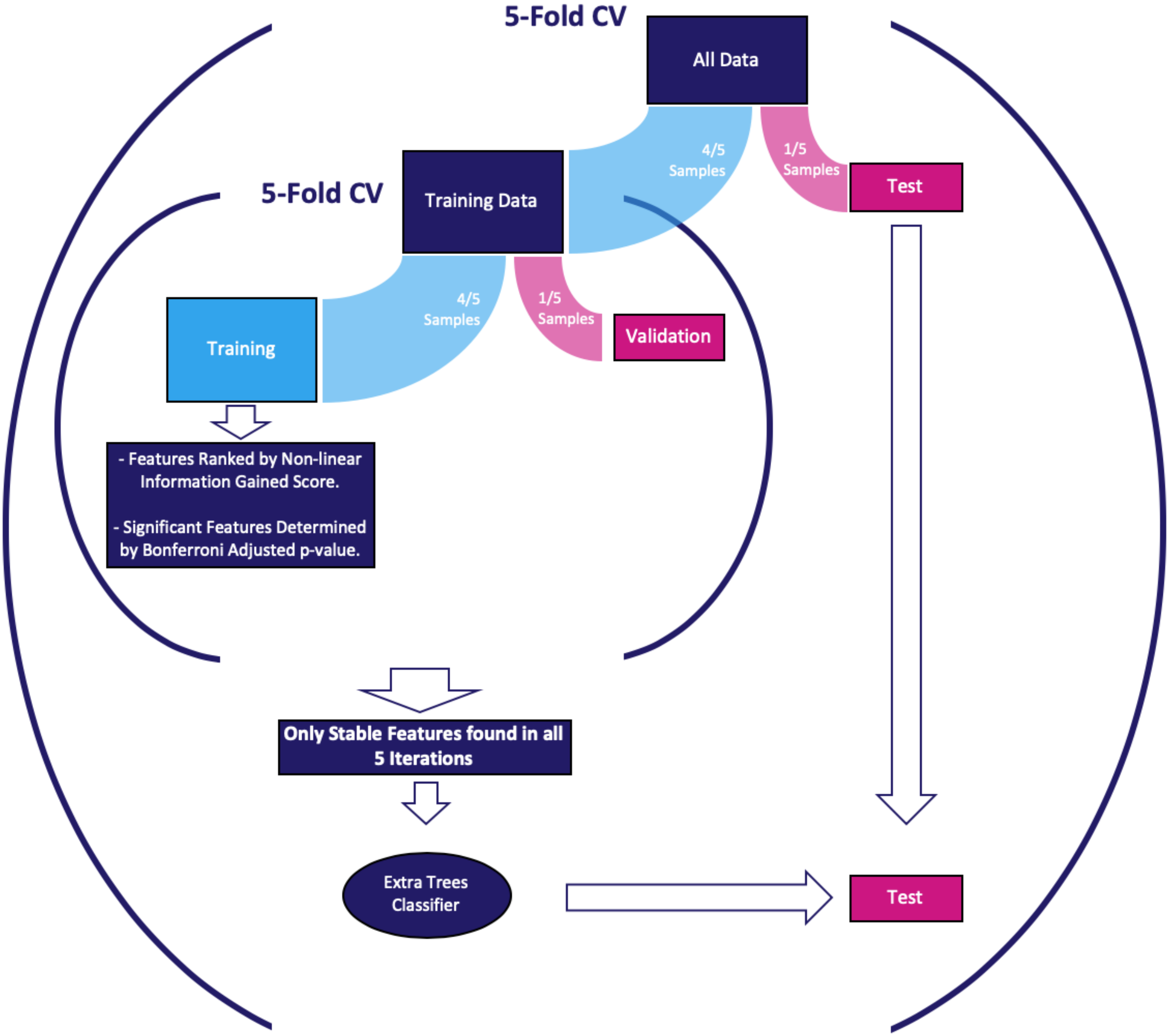
GO Term	OR	FDR	Direction
Bcell			
Myeloid Leukocyte Migration	3.30	8.20E-05	Up
Regulation of Immune System Process	1.60	8.20E-05	Up
Nuclear-ranscribed mRNA Catabolic Proce...	3.70	3.60E-13	Down
mRNA Metabolic Process	1.87	1.35E-12	Down
Gran			
Signaling	1.34	1.73E-11	Down
Cell Communication	1.34	1.73E-11	Down
Mono			
Cellular Response to Lipopolysaccharide	2.52	3.55E-03	Up
Cellular Response to Molecule of Bacteri...	2.48	3.55E-03	Up
Cellular Component Organization or Bioge...	1.18	8.28E-04	Down
Cellular Component Organization	1.18	2.17E-03	Down
NK			
RNA Processing	1.58	2.42E-04	Up
ncRNA Metabolic Process	1.76	2.52E-03	Up
Secretion	1.63	2.70E-08	Down
Immune System Process	1.42	3.69E-08	Down
Tcell			
RNA Splicing, via Transesterification Re...	1.93	3.22E-03	Down
RNA Splicing, via Transesterification Re...	1.91	3.22E-03	Down
Tempus			
Cotranslational Protein Targeting to Mem...	3.29	2.91E-05	Up
SRP-dependent Cotranslational Protein Ta...	3.31	3.35E-05	Up

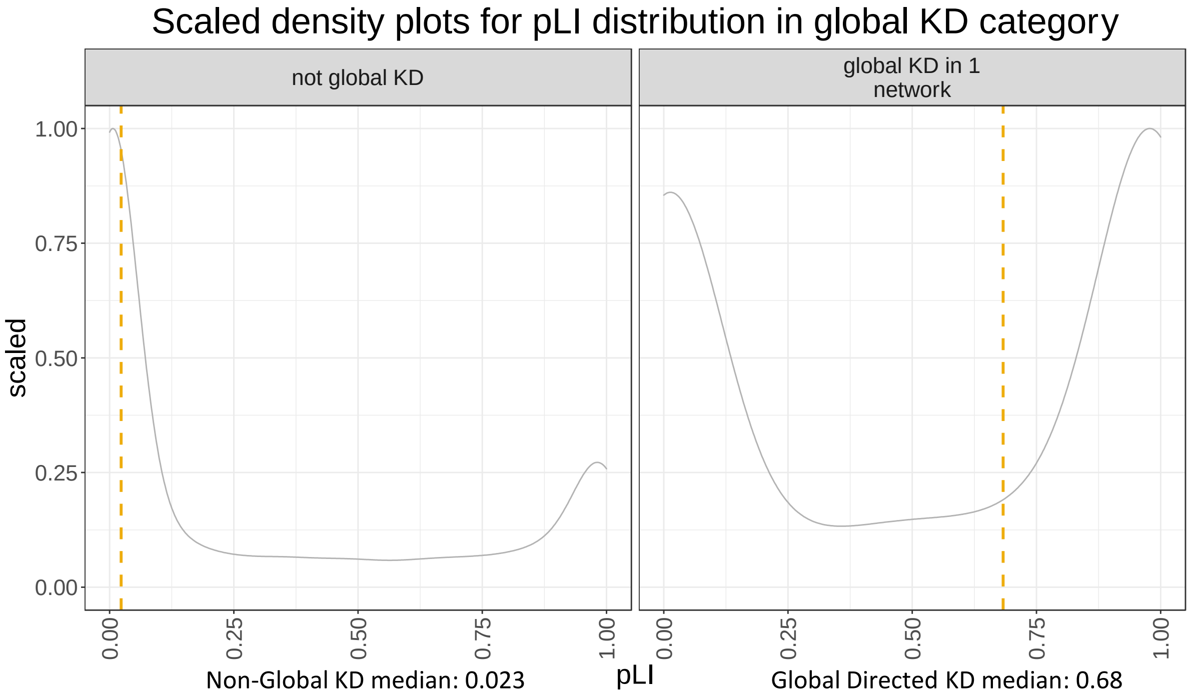


Module Annotation

Hallmark Term	OR	FDR	GO Term	OR	FDR
			BcellM7		
Apical Junction	2.72	3.50E-04	Neutrophil Degranulation	3.67	2.21E-26
Interferon Gamma Response	2.29	4.13E-04	Neutrophil Mediated Immunity	3.58	6.29E-26
Estrogen Response Late	2.43	1.80E-03	Exocytosis	2.93	6.29E-26
			MonoM6		
Kras Signaling Up	3.05	2.00E-04	Cell Chemotaxis	3.61	1.84E-04
P53 Pathway	2.27	3.99E-03	Leukocyte Chemotaxis	3.8	4.94E-04
Il2 Stat5 Signaling	2.3	5.50E-03	Granulocyte Migration	4.51	1.38E-03
			NKM13		
Myogenesis	6.17	1.21E-04	Platelet Alpha Granule	16.49	1.84E-11
Heme Metabolism	5.17	1.21E-04	Blood Coagulation	7.14	1.94E-09
Coagulation	7.65	1.47E-04	Platelet Degranulation	11.59	1.94E-09
			NKM9		
Tnfa Signaling Via Nfkb	3.26	1.57E-02	Regulation Of Nitrogen Compound Metaboli...	1.39	1.21E-02
			Regulation Of Primary Metabolic Process	1.38	1.21E-02
			Regulation Of Ubiquitin-Protein Transfer...	7.76	1.21E-02
			NKM4		
Uv Response Up	7.01	2.06E-03	Mapk Cascade	4.69	1.72E-04
Il6 Jak Stat3 Signaling	8.82	8.70E-03	Signal Transduction By Protein Phosphory...	4.59	1.72E-04
Estrogen Response Early	5.51	1.26E-02	Protein Phosphorylation	2.89	5.72E-04


a
Modules
All Nodes
NKM4
BcellM7
MonoM6
NKM13
NKM9
DYSF
TECPR2
PTPRE
NCOA2
LYN
MXD1
STX3
REPS2
MLL2
TUBB1
RP11-701P16.2
b
Key Drivers
Key Drivers
Bcell CFS Signature KD
BcellM7 Module KD
MonoM6 Module KD
Gran CFS Signature KD
NK CFS Signature KD
NKM13 Module KD
NKM9 Module KD
Tcell CFS Signature KD
Tempus CFS Signature KD
Network Global KD
c
LogFC
Tempus
Tcell
NK
Mono
Gran
Bcell
*CFS Signature Gene
Higher Expression in Patients
1.0
0.5
0.0
-0.5
-1.0
LogFC
Higher Expression in Controls
MXD1
STX3
DYSF
LYN
MLL2
NCOA2
PTPRE
REPS2
RP11-701P16.2
TECPR2
TUBB1
ATP6V1B2
CCNC
DOCK5
EPB49
FAM49B
FKBP8
FMNL1
GTF3C1
IGF2R
IL6R
MAPK14
MBOAT7
MSN
MXI1
MYH9
MYO1F
NCOA1
NFAM1
PIP4K2A
RNF10
RP1-179N16.3
RP11-701P16.1
SEPT9
SLC19A1
SLC25A39
SORL1
UBR4
WDFY3
ZFP106
Key Driver Genes


5-Fold CV
All Data
4/5 Samples
1/5 Samples
Test
5-Fold CV
Training Data
4/5 Samples
1/5 Samples
Validation
Training
- Features Ranked by Non-linear Information Gained Score.
- Significant Features Determined by Bonferroni Adjusted p-value.
Only Stable Features found in all 5 Iterations
Extra Trees Classifier
Test



One-Sided Wilcoxon Test p value = 1.80e-48