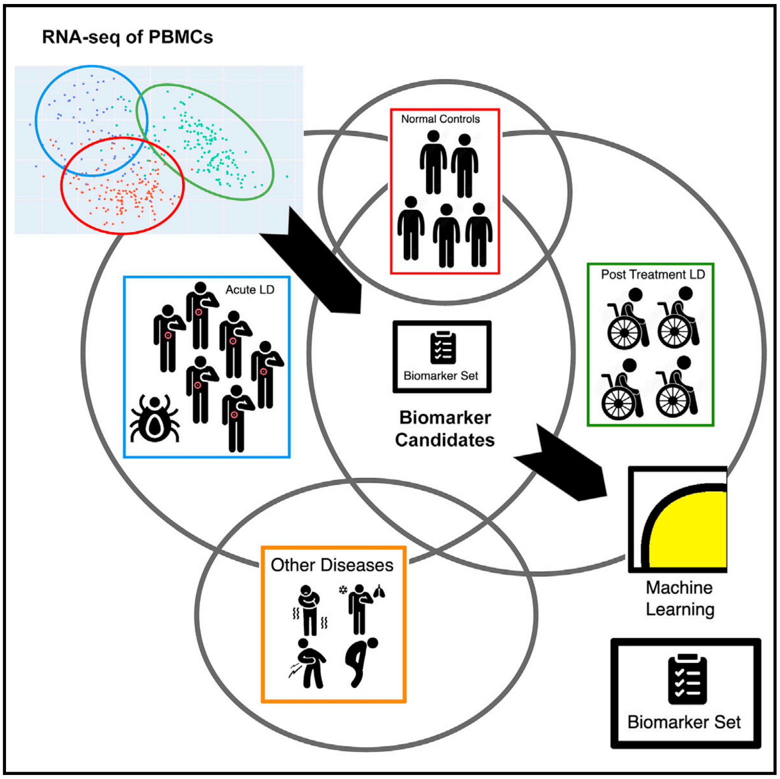
Cell Reports
Medicine

Article

Gene set predictor for post-treatment Lyme disease

Graphical abstract



Highlights

- RNA-seq of PBMCs from 152 individuals with post-treatment Lyme disease
- Differential expression between acute, post-treatment, and uninfected
- Machine learning to identify most relevant genes
- 35 genes identified as potentially useful biomarker set

Authors

Daniel J.B. Clarke, Alison W. Rebman, Jinshui Fan, Mark J. Soloski, John N. Aucott, Avi Ma'ayan

Correspondence

jaucott2@jhmi.edu (J.N.A.),
avi.maayan@mssm.edu (A.M.)

In brief

Clarke et al. analyzed the gene-expression levels from the blood of 152 individuals diagnosed with post-treatment Lyme disease. The results suggest a unique immune response that could be used to further understand the mechanisms of the disease, and the 35 most informative genes can be applied to improve diagnosis.

Clarke et al., 2022, Cell Reports Medicine *3*, 100816
November 15, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.xcrm.2022.100816

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Gene set predictor for post-treatment Lyme disease

Daniel J.B. Clarke,[1] Alison W. Rebman,[2] Jinshui Fan,[2] Mark J. Soloski,[2] John N. Aucott,[2,*] and Avi Ma'ayan[1,3,*]

[1]Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

[2]Lyme Disease Research Center, Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3]Lead contact

*Correspondence: jaucott2@jhmi.edu (J.N.A.), avi.maayan@mssm.edu (A.M.)

https://doi.org/10.1016/j.xcrm.2022.100816

SUMMARY

Lyme disease (LD) is tick-borne disease whose post-treatment sequelae are not well understood. For this study, we enrolled 152 individuals with symptoms of post-treatment LD (PTLD) to profile their peripheral blood mononuclear cells (PBMCs) with RNA sequencing (RNA-seq). Combined with RNA-seq data from 72 individuals with acute LD and 44 uninfected controls, we investigated differences in differential gene expression. We observe that most individuals with PTLD have an inflammatory signature that is distinguished from the acute LD group. By distilling gene sets from this study with gene sets from other sources, we identify a subset of genes that are highly expressed in the cohorts but are not already established as biomarkers for inflammatory response or other viral or bacterial infections. We further reduce this gene set by feature importance to establish an mRNA biomarker set capable of distinguishing healthy individuals from those with acute LD or PTLD as a candidate for translation into an LD diagnostic.

INTRODUCTION

Approximately 30,000 diagnosed cases of Lyme disease (LD) are reported to the CDC each year. However, the actual estimated burden is ~476,000 cases, carrying a yearly healthcare cost of ~$1 billion in the US.[1] Testing and diagnosis of the earliest stages of LD have proven to be difficult or unreliable.[2] The universally accepted diagnostic test for LD is a positive enzyme-linked immunosorbent assay (ELISA) followed by a positive western blot for immunoglobulin M (IgM) and IgG, referred to as the two-tier test (TTT). In addition, there is a recently introduced modified TTT (MTTT) and a test for antibodies reactive to the VlsE1 antigen.[3] The TTT test has a sensitivity of 17%–43% during the early stage of infection.[2] In the absence of a laboratory diagnostic tool, the diagnosis of early LD is reliant on clinical demonstration of the erythema migrans (EM) skin lesion that occasionally does not present or is not observed. This can lead affected individuals to progress to early disseminated or late-stage disease, which can have more difficult-to-treat symptoms, before the disease is diagnosed and treated with antibiotics. Antibiotic treatment includes a dosing regimen of doxycycline, amoxicillin, ceftriaxone, or cefotaxime, dependent on patient age and displayed symptoms.[4] Even when the disease is clearly diagnosed and properly treated, about 10%–20% of affected individuals do not respond completely and develop prolonged symptoms, a condition termed post-treatment LD (PTLD).[5] According to the proposed case definition put forth in the 2006 guidelines of the Infectious Diseases Society of America (IDSA), PTLD is characterized by a previously documented case of LD infection, completion of appropriate antibiotics, and symptoms 6 months after completion of antibiotic treatment of fatigue, bodily pain, and/or cognitive difficulties that impact day-to-day life.[4] In 2020, the IDSA updated these guidelines; however, this proposed research case definition for PTLD was removed.[6] PTLD has been controversial in the medical community due its non-characteristic symptoms and the current inability to identify the causes of the persistent symptoms and their subsequent resolution. While prior studies have shown altered biology in individuals with PTLD, there are no biomarkers to diagnose the condition.[7]

Several studies have examined the gene-expression profile of cells and tissues from individuals with acute LD.[8–12] In the studies examining peripheral blood mononuclear cells (PBMCs), all three studies identified strong differential gene expression (DGE) signatures during acute disease that were differentiated from healthy controls, and these signatures were dominated by the expression of numerous immune-related genes.[8–10] In two of these three studies,[8,9] when gene expression was examined at later time points after antibiotic treatment, the DGE signature can be distinguished from healthy controls up to 1 year post-infection, even though symptoms had largely resolved in many cases. The underlying mechanisms that drive this sustained gene expression are not clear. However, in a third study, it was observed that by 6 months after antibiotic treatment, the gene-expression signatures of Lyme and heathy control cases were indistinguishable.[10] Presently, the reason for this discrepancy is not clear but may relate to the size and/or composition of the cohorts. Importantly, all three studies were able to identify LD gene signatures that may be of value in the diagnosis and staging of human LD. The study examining gene expression

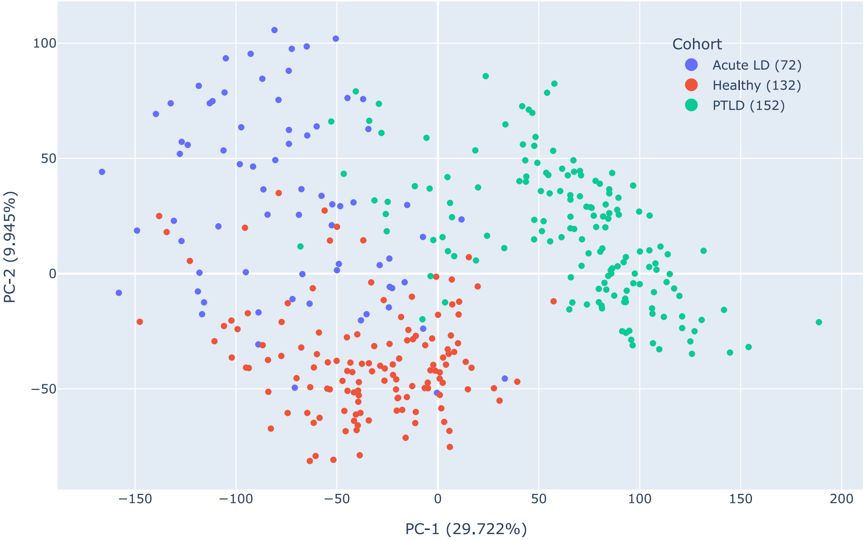


Figure 1. PCA of normalized RNA-seq expression vectors labeled by cohort

Each sample represents one individual from one visit. There were no biological or technical replicates, but some of the acute samples were from the same person at multiple visits. Acute LD refers to the cohort of individuals at their initial visit; healthy refers to non-Lyme-exposed healthy control participant samples taken at several time points; and PTLD refers to the cohort of individuals diagnosed with PTLD.

within whole EM tissue identified several immune related genes, including cytokines, chemokines, Toll-like receptors (TLRs), antimicrobial peptides, interferon-inducible genes, and genes associated with monocytoid cell activation.[11] More recently, single-cell transcriptome analysis of cells recovered from EM lesions clearly identified antigen-driven clonal expansion of novel B cell subsets, as well as the presence of activated T cells and myeloid subsets.[12] Collectively, these studies demonstrate that infection with *B. burgdorferi* is accompanied by activation of cellular elements of the innate and adaptive immune response that can be identified locally in tissue and in blood.

To further our understanding of the molecular mechanisms that may contribute to PTLD symptoms and to identify reliable biomarkers for the diagnosis of PTLD, we examined the transcriptional profiles of PBMCs isolated from 152 individuals with PTLD and compared these profiles with individuals with acute LD and uninfected healthy control participants. Visualization of these cohorts was performed by examining the projection of the RNA sequencing (RNA-seq) profiles into lower dimensions. In addition, differential gene-expression analysis followed by enrichment analysis was employed to identify upstream regulatory mechanisms and disease phenotypes distinctly associated with LD and PTLD. Next, we further analyzed the differentially expressed genes (DEGs) to identify genes that may serve as an mRNA biomarker to confirm LD at the early stages of the disease as well as to assist in distinguishing between completely convalescent individuals and those with PTLD.

RESULTS

RNA-seq profiles were collected from PBMCs isolated from 152 individuals with PTLD. These individuals were compared with previously published RNA-seq profiles from 72 individuals with acute LD (acute cohort) and 44 healthy controls also followed over time (Table S1).[9] Principal-component analysis (PCA) of the dimensionality-reduced profiles suggests that most individuals with PTLD have an expression signature that is different from the healthy controls and profiled individuals with acute LD (Figure 1). A smaller number of individuals with PTLD show expression profiles that are comparable to those in the acute cluster.

To explore the characteristics of the individuals with PTLD, we compared them with the healthy control and the acute LD groups. Differential expression analysis was followed by enrichment analysis (see STAR Methods). When comparing the individuals with PTLD with healthy controls, 1,213 genes were identified as significantly upregulated in PTLD and 803 were identified as significantly downregulated (limma-voom, Benjamini-Hochberg [BH]-adjusted p value < 0.01). The enrichment results are presented as bar charts along with links to the reports in Enrichr (Figures 2A and 2B). The upregulated genes are enriched for immune response genes and upregulation of the cell cycle. Specifically, MSigDB Hallmark sets are enriched for G2-M checkpoint and E2F transcription factor targets (Fisher's exact test, $p < 0.000005$, $q < 0.0001$) and response to herpes simplex virus 1 infection (KEGG pathways, $p < 5.3e-12$, $q < 1.54e-9$). Interestingly, a significant number of DEGs are enriched for cilium components (Gene Ontology [GO: 0005929], $p < 0.0006$, $q < 0.1$) and cilium-related disorders (primary ciliary dyskinesia, $p < 0.00008$). Downregulated genes, when comparing individuals with PTLD with healthy controls, are enriched for Wnt pathway components (Wiki Pathways, Wnt signaling in kidney disease WP4150, $p < 0.00001$, $q < 0.003$) and spinal cord specification genes (Tissue Protein Expression from Human Proteome Map, adult spinal cord, $p < 0.00009$, $q < 0.0027$).

When comparing the 152 gene expression profiles from individuals with PTLD with the 72 individuals with acute LD, we observed similar patterns (Figures 2C and 2D). Specifically, 817 genes are significantly upregulated in the individuals with PTLD, and these genes are enriched for activation of the cell cycle compared with the individuals with acute LD, for example, "Metaphase/Anaphase Phase Transition" from the Elsevier Pathway Collection ($p < 0.000006$, $q < 0.0022$). Interestingly, the same gene sets that are upregulated when comparing the PTLD group with healthy controls are downregulated when comparing PTLD with acute LD (KEGG, herpes simplex virus 1 infection, $p < 1.5e-32$, $q < 4.39e-30$) (Figure 3).

After processing these data to identify DEGs, we analyzed and visualized the results with a Super-Venn diagram (Figure 4). We

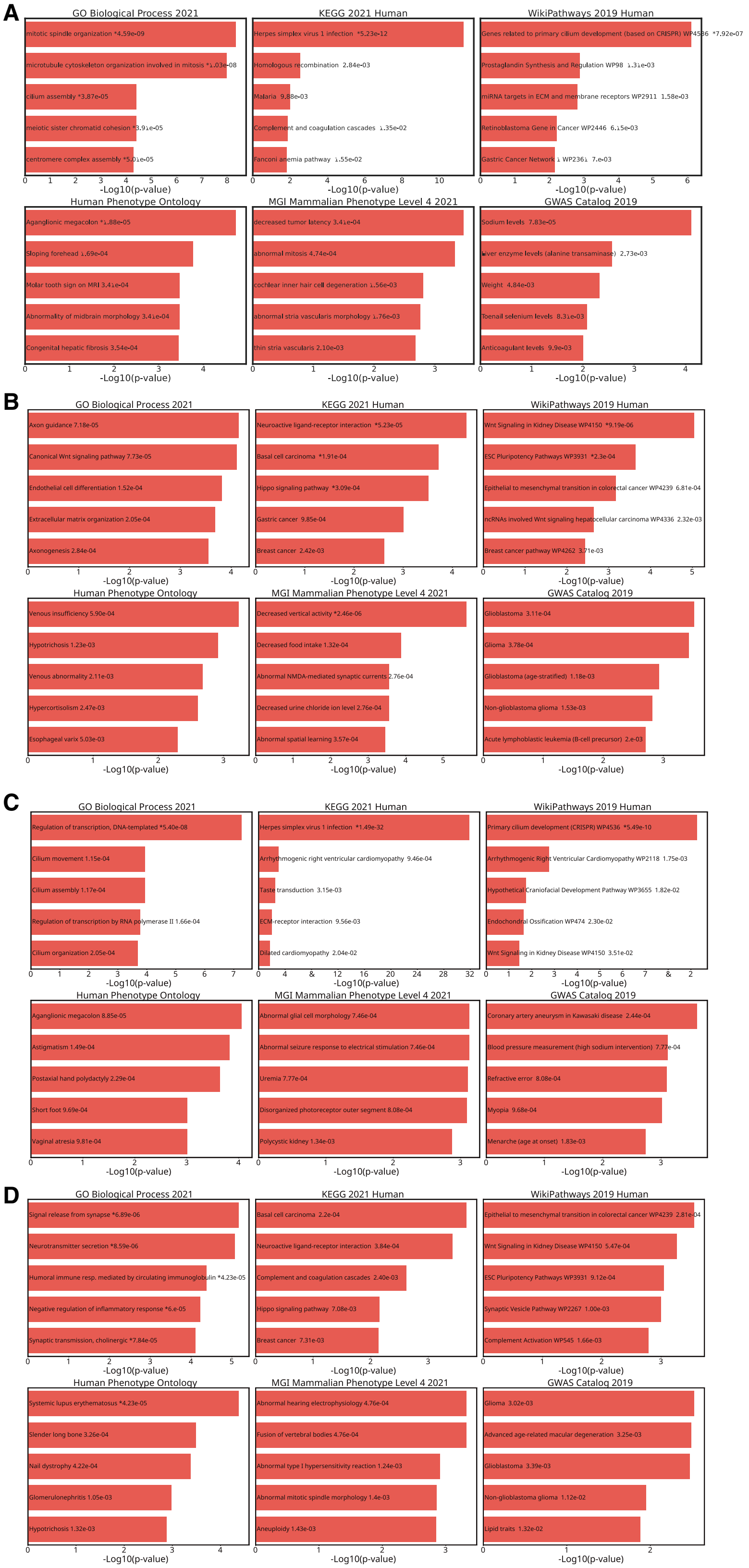


Figure 2. Enrichment analysis of DEGs

(A) Enrichment results from Enrichr for significantly DEGs in controls versus PTLD up (up in PTLD). Full results across more libraries are available from https://maayanlab.cloud/Enrichr/enrich?dataset = 5c8c3715899dbaa6ce7aa17d3fe0e77d.

(B) Enrichment results from Enrichr for significantly DEGs in controls versus PTLD down (down in PTLD). Full results across more libraries are available from https://maayanlab.cloud/Enrichr/enrich?dataset=4a58c5ae103e3fa93861d231a9718f54.

(C) Enrichment results from Enrichr for significantly DEGs in acute LD versus PTLD up (up in PTLD). Full results across more libraries are available from https://maayanlab.cloud/Enrichr/enrich?dataset=1954a8136b6aa8c0b73b1cff30ad5280.

(D) Enrichment results from Enrichr for significantly DEGs in acute LD versus PTLD down (down in PTLD). Full results across more libraries are available from https://maayanlab.cloud/Enrichr/enrich?dataset=00e156d32ab62844391abaf9e3b0b823.

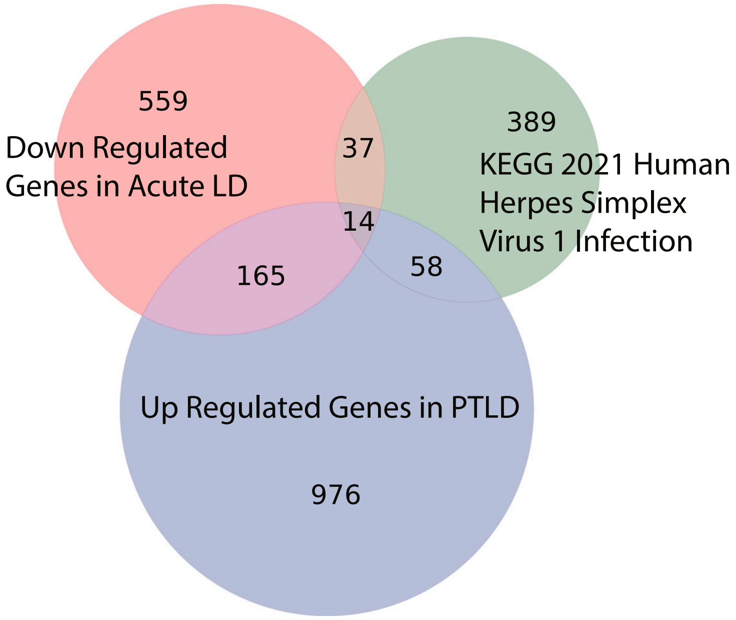

559
Down Regulated Genes in Acute LD
37
389
KEGG 2021 Human Herpes Simplex Virus 1 Infection
14
165
58
Up Regulated Genes in PTLD
976

Figure 3. Differentially expressed genes in acute LD and PTLD compared with the KEGG term herpes simplex virus 1 infection

compared the DEGs from the acute LD and PTLD cohorts with gene sets from other infectious diseases (Table S2). The additional gene sets were extracted from Enrichr.[13] Such gene sets include gene sets extracted from published studies of human PBMCs from subjects virally infected with influenza, HIV, and COVID-19. These gene sets were filtered by genes that are also differentially expressed in acute LD or PTLD. Bacterial infection response genes derived from Enrichr gene sets are those corresponding to studies that profiled human PBMCs from individuals infected with pathogens such as *Mycobacterium tuberculosis* and *Neisseria gonorrhoeae*. In addition, gene sets related to *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Escherichia coli*, and bacterial sepsis extracted from studies that profiled human cells infected with those bacteria were included. Also, gene sets from DisGeNET,[14] and those corresponding to diseases caused specifically by *Spirochetes*, filtered by genes that are also differentially expressed in acute LD or PTLD were included (Table S2).

Next, processing the Super-Venn diagram for overlapping sets, we extracted the DEGs from the acute LD and the PTLD groups that were specifically not present in any viral or bacterial infection gene sets. The resulting gene set contained 41 shared LD up genes, 134 shared LD down genes, and 175 genes that appear to be affected in opposite directions between acute LD and PTLD. The set of 350 candidate genes was further reduced using permutation importance with the goal of achieving an optimal tradeoff between performance and gene set size. Permutation importance was calculated with logistic regression classifiers trained on a randomly selected 50% of the samples to predict the class labels of the remaining 50%. Furthermore, the permutation importance was performed using four separate binary classification tasks: (1) distinguishing acute LD and PTLD together from healthy controls; (2) distinguishing acute LD from healthy controls; (3) distinguishing PTLD from healthy controls; and (4) distinguishing between acute LD and PTLD.

The top k genes from the gene permutation importance test can be used to train a classifier that performs increasingly well as k becomes larger. We found that at around 35 genes, where a split in the permutation importance distribution occurs, we only lose 0.04 area under the receiver operating characteristic (AUROC) curve for the acute LD versus healthy control classifier when compared with using all 350 genes. At fewer than 35 genes, performance begins to degrade rapidly. With 35 genes (Table S3), the classifier performs at 98% accuracy, determining whether a sample is from an individual with LD or a healthy control (Figure 5). Since the 35 selected biomarker genes are specific to LD and are not known to be associated with immune activation functions, we attempted to explore enriched terms associated with these 35 genes (Figure 6). Interestingly, the genes CACNB4 and ALDH7A1 are highly enriched for epilepsy based on OMIM[15] (Fisher's exact test, $p < 0.0047$, $q < 0.0093$), and the genes CACNB4, ALDH7A1, SLC4A10, and SCN3A are enriched for proteins involved in epilepsy based on the Elsevier Pathway Collection database ($p < 0.0013$, $q < 0.093$). Neuropsychiatric symptoms have been reported in individuals with PTLD and in individuals with Lyme encephalopathy[16,17] and this analysis suggests a molecular underpinning of such a phenotype. Three of the genes from the set of 35 biomarkers, namely AHNAK2, APBA1, and SHANK1, encode genes with a PDZ domain that is statistically over-represented ($p < 0.0015$, $q < 0.034$). These genes may form complexes with the several ion channels that are over-represented in the list of 35 biomarkers and potentially impact the neurologic symptoms observed in PTLD. Applying WEAT analysis,[18] we observe that most of the 35 have some annotations, but they are mostly under-studied (Table S4).

Nonetheless, due to the strong alignment of the samples to the two principal components, we sought to identify the most singularly predictive genes to find an adequate proxy for the principal components. The top-ranked gene based on the permutation importance analysis is Kelch-like family member 11 (KLHL11), which is a known member of the cullin-RING-based BCR (BTB-CUL3-RBX1) E3 ubiquitin-protein ligase complex. KLHL11 expression levels are capable of distinguishing LD from healthy controls well but cannot distinguish acute LD from PTLD as well. Another gene, undifferentiated embryonic cell transcription factor 1 (UTF1), can distinguish acute LD from PTLD well but does not perform well in classifying LD from healthy controls. A classifier that combines KLHL11 and UTF1 together performs much better than a random classifier but not well enough to become a reliable diagnostic (Figure S3). Additionally, we assessed the performance of all singular marker genes on the four classification tasks and

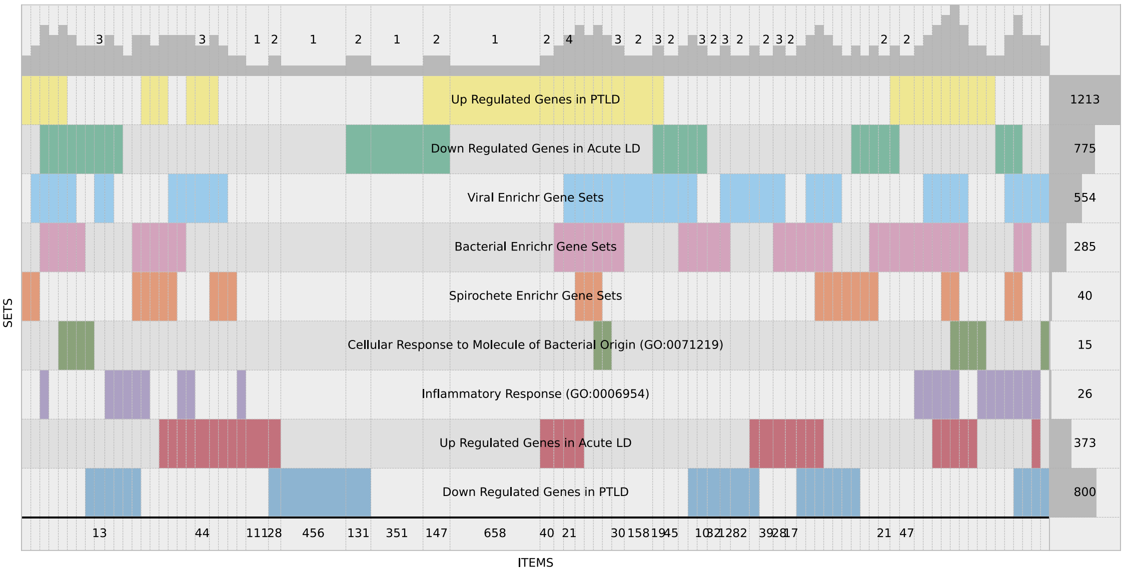


Figure 4. Super-Venn diagram visualization of overlapping gene sets between the cohorts and genes known to be related to inflammatory responses in other diseases

Counts on the top correspond to the number of sets that overlap, counts on the right correspond to the number of genes in the set, and counts at the bottom correspond to the number of genes in the intersecting set. Up- and downregulated genes in acute LD refer to significantly DEGs with respect to the healthy individuals at visit 1. Up- and downregulated genes in PTLD refer to significantly DEGs with respect to the healthy individuals. Viral, bacterial, and spirochete genes derived from Enrichr gene sets and filtered by genes that are also differentially expressed in LD or PTLD. Relevant gene sets from Enrichr are listed in Table S2.

computed AUROC curves in the same way on the test set (Figure 7).

DISCUSSION

LD and the infection-associated sequelae of PTLD continue to emerge as an important public health issue. Existing tests to confirm diagnosis have limited accuracy, especially in supporting the diagnosis of PTLD, which is made in persistently symptomatic individuals many months to years after appropriate diagnosis and treatment. Here, we have applied whole PBMC transcriptome analysis to a set of individuals in a well-defined cross-sectional cohort of individuals with PTLD. The PTLD signature was found to be distinct from the healthy controls and from individuals diagnosed with acute LD, which were part of a longitudinal cohort that we have reported on previously.[9] Importantly, the PTLD and acute LD RNA-seq signatures were sufficiently distinct, enabling us to design a set of mRNA markers that will be of value in distinguishing acute LD, PTLD, and healthy controls. Although the individuals diagnosed with PTLD have a similar immune activation to the individuals with acute LD, there is a component of this immune response that is diminished or altered. The reduction in immune activation is expected, but the observation that these affected individuals are markedly different from healthy controls is illuminating. The separation of participants with PTLD from the healthy controls and acute LD clusters may be explained by batch effects. However, removing the batch effects is difficult because the group labels correspond to the batch labels. The observation that the DEGs point to bacterial infection and inflammatory response suggest that the differences observed are not just due to batch effects. Some of the DEGs in PTLD point to common symptoms observed in these affected individuals. For example, neuropsychiatric symptoms that have been reported in individuals with PTLD and in individuals with Lyme encephalopathy[16,17] are consistent with the enrichment analysis that suggests potential genetic underpinning of such a phenotype. Expression data from RNA-seq applied to PBMCs collected from acute LD, PTLD, and healthy controls yields distinct separation of individuals along the first two principal-component axes of variance. This suggests that mRNA biomarkers may be feasibly identified to diagnose acute LD and PTLD. Gene set overlap analysis was used to identify consensus and divergent DEGs in acute LD and PTLD and compared with gene sets from other viral and bacterial infections. The resulting genes that are specific for LD were further reduced, and classifiers were constructed to assess the feasibility of developing a diagnostic. Overall, the gene classifiers we identified can categorize individuals with acute LD and PTLD using as few as two mRNA biomarkers.

Our results suggest the possibility of utilizing an mRNA-based diagnostic biomarker panel, in combination with precise clinical evaluations, to identify and/or categorize individuals in whom LD is suspected. This approach could be adapted to utilize whole blood, a readily accessible tissue, and would not rely on the detection of anti-*Borrelia* antibodies or bacterial DNA, approaches that have been shown to lack sensitivity. In addition, previous studies have shown that levels of chemokines (CCL19), serum metabolites, and a fecal microbiome signature have been associated with the development of PTLD.[19–21] Therefore, it is possible that an approach incorporating mRNA combined with other molecular

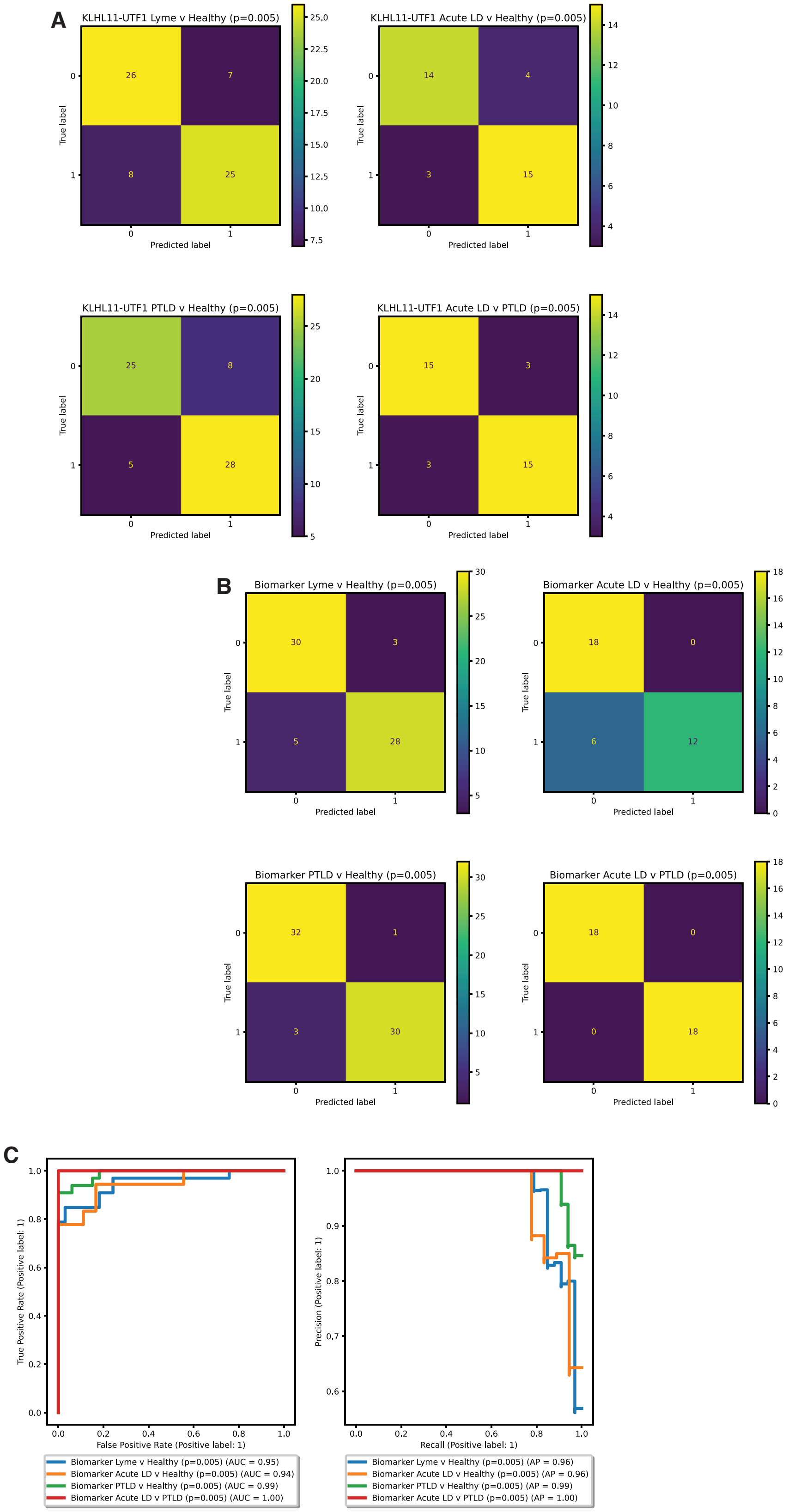


Figure 5. Performance of the classifier using 35 mRNA biomarker genes chosen with the training set

Performance is based on training four independent models and predicting the held-out samples not used during training or feature selection. The test set is randomly under-sampled such that classes are balanced. p values are based on a permutation test across different potential train-test splits. (A) Confusion matrices for a two-gene classifier. (B) Confusion matrices for the 35-gene biomarker set. (C) ROC and Precision Recall curves for the 35-gene-set biomarker.

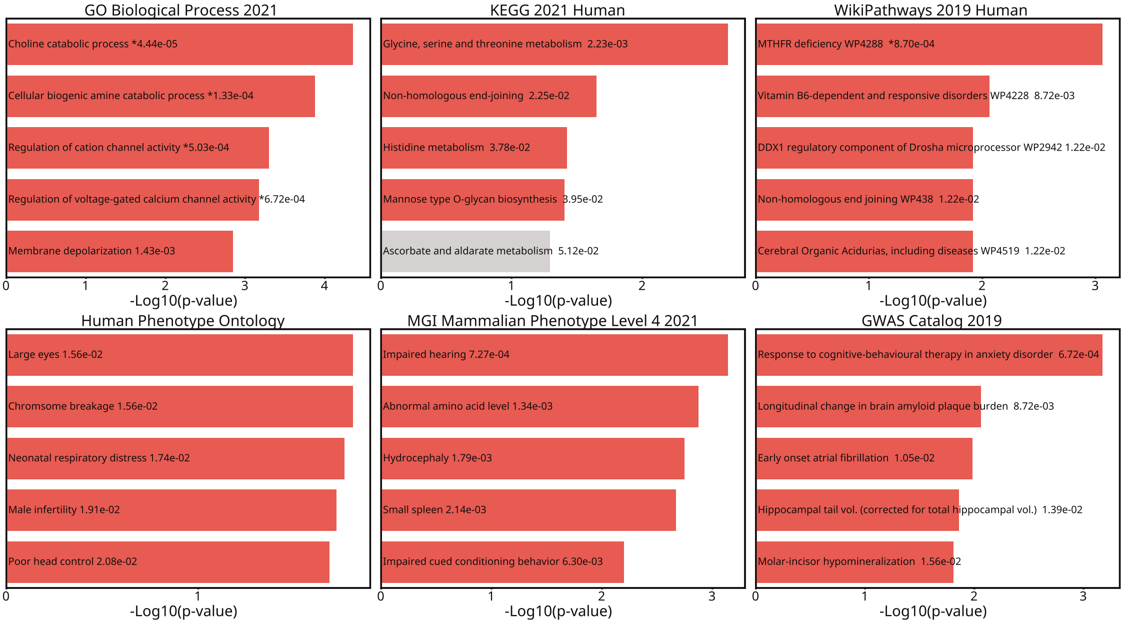


Figure 6. Enrichment results from Enrichr for the 35 biomarker genes

Full results across more libraries are available from https://maayanlab.cloud/Enrichr/enrich?dataset=41c885f2b79be29e03211733ca32d137.

signatures may lead to an accurate diagnostic with high sensitivity and specificity in diagnosing the multiple stages of human LD, including PTLD. This, however, will require a broader study design using samples from multiple PTLD cohorts as well as the study of cohorts from other disorders such as post-acute COVID-19 and chronic fatigue syndrome.

The finding that PBMCs from individuals with PTLD express an mRNA signature also indicates that PTLD has a specific underlying biology. Understanding this complex biology will be of great value in the development of novel treatment strategies. The gene expression data on this large group of individuals with PTLD expand on previous work, all of which lead to the identification of gene classifiers for acute LD.[8–10] In addition, the studies on acute LD identified an mRNA signature that was consistent with a strong immune response. In this study, which focuses on PTLD, analysis using a Super-Venn diagram showed that both up- and downregulated genes overlapped with host inflammatory response genes and genes linked to viral and bacterial infections. Comparing the genes that distinguish PTLD from acute LD identified several immune features. The complement pathway is identified using KEGG[22] and BioCarta pathways analysis, as well as GO Biological Processes[23] enrichment analysis. In addition, enrichment analysis using the Azimuth Cell Types library identified gene signatures from immune cell types including plasmablasts and proliferating CD4$^+$ and CD8$^+$ T cells. This would imply, as mentioned above, that several immune pathways are a part of the underlying biology of PTLD.

In addition, our analysis also identified non-immune features. For example, when comparing upregulated genes in PTLD versus healthy controls, WikiPathways[24] enrichment analysis identified a gene set related to primary cilium development and ciliopathies, and GO Cellular Components revealed cilium (GO: 0005929) and motile cilium (GO: 0031514). These are features that are associated with non-immune, non-bone-marrow-derived cell types such as epithelial cells. Of note, cell-type enrichment analysis using the Descartes Cell Types and Tissue 2021 library identified a signature associated with ciliated epithelial cells of the lung, and the Human Gene Atlas library identified enriched terms aligned with non-immune cell types. This signature could be from a non-bone-marrow-derived cell that bears primary cilia, or it could be from an immune cell that expresses cilia-like projection. Regarding this latter possibility, recent data have shown that effector T cells can form cilia-like projections, most notably in the process of immune synapse formation when effector T cells are interacting with targets cells.[25] These structures have been referred to as “frustrated cilia,” and the genes responsible overlap considerably with genes known to be involved in formation of primary cilia in epithelium and other non-immune cell types.[25] These genes include TTC26, TTC23, IFT 74, IFT81, IFT85, ARL13B, CEP83, CEP162, CEP76, and CEP44. CEP83 encodes a protein involved in centrosome docking on the plasma membrane and is critical for primary cilia and immune synapse formation.[25,26] *ARL13B* encodes a guanine exchange factor that regulates membrane composition and the recruitment of signaling molecules in both the cilium and immune synapses.[27] TTC genes encode tetratricopeptide repeat-containing proteins that can interact with IFT proteins and are critical for cilia formation and function.[28] Intraflagellar transport (IFT) proteins participate in the active sorting and transport of cytosolic and membrane proteins destined for the cilium, and this can include signaling molecules.[29] All these genes are upregulated in individuals with PTLD relative to healthy controls, suggesting the presence of an immune cell type in circulation that is in the process of assembling cilia-like structures.

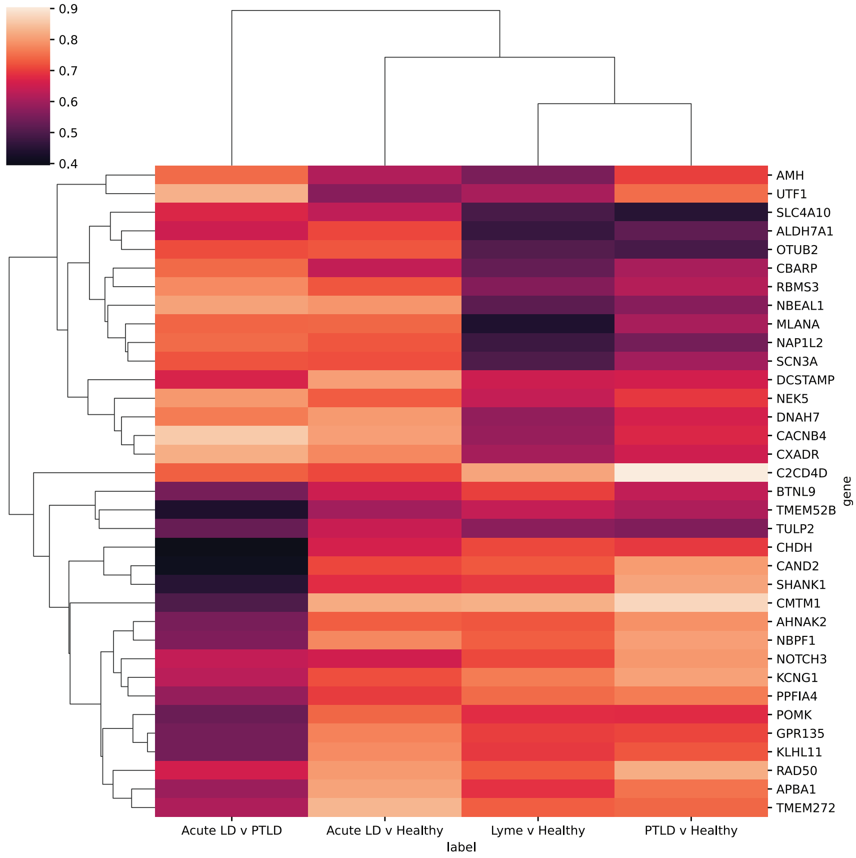


Figure 7. Classifier performance with single genes

Performance of singular mRNA biomarkers as the feature for the models. Performance is based on training four independent models with the training set and predicting the test samples not used during training. The test set is randomly under-sampled such that classes are balanced. AUROCs for each gene are reported as a heatmap.

This is consistent with the discussion above, suggesting that immune activation is a component of the underlying biology of PTLD. Further study utilizing single-cell transcriptomics would be of value in clearly identifying and validating this novel cell subset as well as help to understand what role such a cell type may play in disease pathogenesis.

The 35 biomarker genes identified as a classifier were not linked to obvious immune pathways. However, enrichment analysis with Enrichr did identify enrichment for several metabolomic pathways including glycine, serine, and threonine metabolism (BioPlanet[30] and KEGG), lysine catabolism (BioPlanet), and glycolysis (MSigDB[31]). This is consistent with previous work identifying a metabolomic signature in PTLD.[20] In addition, a recent study identifying a set of 31 genes for diagnosis of acute LD also noted that only a fraction of the genes are immune related. We believe that these observations underscore the complex biology that is part of PTLD. Understanding the roles that immune and non-immune pathways play in the various stages of LD including PTLD will likely lead to novel therapeutic targets and other therapeutic strategies. The 35 biomarker genes have two genes involved in calcium ion channel regulation (CBARP, CACNB4) and several genes that are highly co-expressed with CAM kinases 2 and 4 (SLC4A10, CACNB4, NAP1L2, APBA1, PPFIA4, and SHANK1, marked in Table S3), potentially forming a neuronal cell signaling pathway and further making a case for neurological symptoms.

In conclusion, this study produced a gene-expression profile for PTLD. This is just a first step that requires confirmation for diagnosis of PTLD. Gene expression can support the diagnosis of PTLD in individuals with a history of prior diagnosed and treated LD and persistent post-treatment symptoms. In addition, if a future diagnostic panel can suggest negative test results for PTLD, based on a reduced representation of gene-expression profile, this could be valuable in individuals with look-alike syndromes not associated with prior LD and would lead to further evaluation of these affected individuals to establish a definitive diagnosis.

Limitations of the study

The identified 35 biomarker genes may be useful as a diagnostic only if the same approach is applied to whole blood instead of PBMCs. PBMC isolation is expensive and currently requires academic laboratory expertise. To translate the test into primary health care for individuals with LD, a parallel test will have to be devised by experimentally comparing PBMCs with whole-blood results. This can be done computationally but more reliably by experimentally measuring gene expression from PBMCs and whole blood from the same large cohort of individuals diagnosed with LD and PTLD.

The study also has additional limitations. The healthy controls and individuals with acute LD samples were collected for a different study and as such were not processed altogether, although they were processed using the same methodology. Because of this, standard batch correction techniques cannot be adequately performed, and the possibility of a batch effect between PTLD and acute LD samples cannot be completely ruled out. Future studies should be designed with the controls processed at the same time as the cases. In addition, the individuals were profiled with RNA-seq in either their acute phase or when they had developed PTLD. However, we do not have sufficient samples from the same individuals that contributed samples in both phases. Hence, we cannot predict which acute individuals would eventually develop PTLD. The rigorous inclusion and exclusion criteria for these cohorts may also limit applicability to a wider set of potential individuals affected by PTLD. A follow-up study should include individuals with LD and PTLD with a broader range of clinical presentations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - The human cohort
  - Demographics of the human cohort
  - RNA-seq profiling of patients
- METHOD DETAILS
  - Isolation of PBMC
  - Preparation of the samples for RNA sequencing
  - RNA sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - RNA-seq processing
  - Enrichment analysis
  - Set overlap analysis
  - Classification model construction and candidate biomarker selection
  - Model evaluation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100816.

ACKNOWLEDGMENTS

The project was partially supported by funds from the Cohen Lyme & Tickborne Disease Initiative and NIH grant P30AR070254. We would like to thank Ting Yang from JHU for depositing the data into dbGAP.

AUTHOR CONTRIBUTIONS

D.J.B.C. performed the analysis, created the figures, and contributed to the writing of the manuscript; A.W.R. provided information about the clinical component of the study and contributed to the writing of the manuscript; J.F. processed the and annotated the samples; M.J.S. and J.N.A. designed the study, oversaw the analysis and interpretations of the results, and contributed to the writing of the manuscript; A.M. oversaw and contributed to the analysis and led the writing of the manuscript.

DECLARATION OF INTERESTS

The authors have applied for patents covering some aspects of the study.

INCLUSION AND DIVERSITY

The authors support inclusive, diverse, and equitable conduct of research.

Received: March 28, 2022
Revised: July 24, 2022
Accepted: October 14, 2022
Published: November 15, 2022

REFERENCES

1. Kugeler, K.J., Schwartz, A.M., Delorey, M.J., Mead, P.S., and Hinckley, A.F. (2021). Estimating the frequency of Lyme disease diagnoses, United States, 2010–2018. Emerg. Infect. Dis. *27*, 616–619.
2. Branda, J.A., and Steere, A.C. (2021). Laboratory diagnosis of Lyme borreliosis. Clin. Microbiol. Rev. *34*, e00018-19–e00019.
3. Porwancher, R., and Landsberg, L. (2021). Optimizing use of multi-antibody assays for Lyme disease diagnosis: a bioinformatic approach. PLoS One *16*, e0253514. https://doi.org/10.1371/journal.pone.0253514.
4. Wormser, G.P., Dattwyler, R.J., Shapiro, E.D., Halperin, J.J., Steere, A.C., Klempner, M.S., Krause, P.J., Bakken, J.S., Strle, F., Stanek, G., et al. (2006). The clinical assessment, treatment, and prevention of lyme disease, human granulocytic anaplasmosis, and babesiosis: clinical practice guidelines by the Infectious Diseases Society of America. Clin. Infect. Dis. *43*, 1089–1134. https://doi.org/10.1086/508667.
5. Aucott, J.N., Yang, T., Yoon, I., Powell, D., Geller, S.A., and Rebman, A.W. (2022). Risk of post-treatment Lyme disease in patients with ideally-treated early Lyme disease: a prospective cohort study. Int. J. Infect. Dis. *116*, 230–237. https://doi.org/10.1016/j.ijid.2022.01.033.
6. Lantos, P.M., Rumbaugh, J., Bockenstedt, L.K., Falck-Ytter, Y.T., Aguero-Rosenfeld, M.E., Auwaerter, P.G., Baldwin, K., Bannuru, R.R., Belani, K.K., Bowie, W.R., et al. (2021). Clinical practice guidelines by the infectious diseases society of America (IDSA), American academy of neurology (AAN), and American college of rheumatology (ACR): 2020 guidelines for the prevention, diagnosis and treatment of lyme disease. Clin. Infect. Dis. *72*, e1–e48. https://doi.org/10.1093/cid/ciaa1215.
7. Bobe, J.R., Jutras, B.L., Horn, E.J., Embers, M.E., Bailey, A., Moritz, R.L., Zhang, Y., Soloski, M.J., Ostfeld, R.S., Marconi, R.T., et al. (2021). Recent progress in lyme disease and remaining challenges. Front. Med. *8*, 666554. https://doi.org/10.3389/fmed.2021.666554.
8. Bouquet, J., Soloski, M.J., Swei, A., Cheadle, C., Federman, S., Billaud, J.-N., Rebman, A.W., Kabre, B., Halpert, R., Boorgula, M., et al. (2016). Longitudinal transcriptome analysis reveals a sustained differential gene expression signature in patients treated for acute Lyme disease. mBio *7*, e00100–e00116.
9. Clarke, D.J.B., Rebman, A.W., Bailey, A., Wojciechowicz, M.L., Jenkins, S.L., Evangelista, J.E., Danieletto, M., Fan, J., Eshoo, M.W., Mosel, M.R., et al. (2021). Predicting lyme disease from patients' peripheral blood mononuclear cells profiled with RNA-sequencing. Front. Immunol. *12*, 636289.
10. Petzke, M.M., Volyanskyy, K., Mao, Y., Arevalo, B., Zohn, R., Quituisaca, J., Wormser, G.P., Dimitrova, N., and Schwartz, I. (2020). Global transcriptome analysis identifies a diagnostic signature for early disseminated Lyme disease and its resolution. mBio *11*, e00047-20–e00020.
11. Marques, A., Schwartz, I., Wormser, G.P., Wang, Y., Hornung, R.L., Demirkale, C.Y., Munson, P.J., Turk, S.-P., Williams, C., Lee, C.-C.R., et al. (2017). Transcriptome assessment of erythema migrans skin lesions in patients with early Lyme disease reveals predominant interferon signaling. J. Infect. Dis. *217*, 158–167.

12. Jiang, R., Meng, H., Raddassi, K., Fleming, I., Hoehn, K.B., Dardick, K.R., Belperron, A.A., Montgomery, R.R., Shalek, A.K., Hafler, D.A., et al. (2021). Single-cell immunophenotyping of the skin lesion erythema migrans identifies IgM memory B cells. JCI insight *6*, 148035.

13. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/nar/gkw377.

14. Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L.I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015.

15. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517.

16. Matera, G., Labate, A., Quirino, A., Lamberti, A.G., BorzÃ, G., Barreca, G.S., Mumoli, L., Peronace, C., Giancotti, A., Gambardella, A., et al. (2014). Chronic neuroborreliosis by B. garinii: an unusual case presenting with epilepsy and multifocal brain MRI lesions. New Microbiol. *37*, 393–397.

17. Juric, S., Janculjak, D., Tomic, S., Butkovic Soldo, S., and Bilic, E. (2014). Epileptic seizure as initial and only manifestation of neuroborreliosis: case report. Neurol. Sci. *35*, 793–794. https://doi.org/10.1007/s10072-014-1648-1.

18. Fan, R., and Cui, Q. (2021). Toward comprehensive functional analysis of gene lists weighted by gene essentiality scores. Bioinformatics *37*, 4399–4404. https://doi.org/10.1093/bioinformatics/btab475.

19. Morrissette, M., Pitt, N., González, A., Strandwitz, P., Caboni, M., Rebman, A.W., Knight, R., D'onofrio, A., Aucott, J.N., Soloski, M.J., and Lewis, K. (2020). A distinct microbiome signature in posttreatment Lyme disease patients. mBio *11*, e02310-20–e02320.

20. Fitzgerald, B.L., Graham, B., Delorey, M.J., Pegalajar-Jurado, A., Islam, M.N., Wormser, G.P., Aucott, J.N., Rebman, A.W., Soloski, M.J., Belisle, J.T., and Molins, C.R. (2021). Metabolic response in patients with posttreatment Lyme disease symptoms/syndrome. Clin. Infect. Dis. *73*, e2342–e2349.

21. Aucott, J.N., Soloski, M.J., Rebman, A.W., Crowder, L.A., Lahey, L.J., Wagner, C.A., Robinson, W.H., and Bechtold, K.T. (2016). CCL19 as a chemokine risk factor for posttreatment Lyme disease syndrome: a prospective clinical cohort study. Clin. Vaccine Immunol. *23*, 757–766.

22. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

23. The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–D338.

24. Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. *46*, D661–D667.

25. Douanne, T., Stinchcombe, J.C., and Griffiths, G.M. (2021). Teasing out function from morphology: similarities between primary cilia and immune synapses. J. Cell Biol. *220*, e202102089. https://doi.org/10.1083/jcb.202102089.

26. Stinchcombe, J.C., Randzavola, L.O., Angus, K.L., Mantell, J.M., Verkade, P., and Griffiths, G.M. (2015). Mother centriole distal appendages mediate centrosome docking at the immunological synapse and reveal mechanistic parallels with ciliogenesis. Curr. Biol. *25*, 3239–3244. https://doi.org/10.1016/j.cub.2015.10.028.

27. Powell, L., Samarakoon, Y.H., Ismail, S., and Sayer, J.A. (2021). ARL3, a small GTPase with a functionally conserved role in primary cilia and immune synapses. Small GTPases *12*, 167–176. https://doi.org/10.1080/21541248.2019.1703466.

28. Xu, Y., Cao, J., Huang, S., Feng, D., Zhang, W., Zhu, X., and Yan, X. (2015). Characterization of tetratricopeptide repeat-containing proteins critical for cilia formation and function. PLoS One *10*, e0124378. https://doi.org/10.1371/journal.pone.0124378.

29. Taschner, M., and Lorentzen, E. (2016). The intraflagellar transport machinery. Cold Spring Harb. Perspect. Biol. *8*, a028092. https://doi.org/10.1101/cshperspect.a028092.

30. Huang, R., Grishagin, I., Wang, Y., Zhao, T., Greene, J., Obenauer, J.C., Ngan, D., Nguyen, D.-T., Guha, R., Jadhav, A., et al. (2019). The NCATS BioPlanet–an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. Front. Pharmacol. *10*, 445.

31. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

32. Rebman, A.W., Bechtold, K.T., Yang, T., Mihm, E.A., Soloski, M.J., Novak, C.B., and Aucott, J.N. (2017). The clinical, symptom, and quality-of-life characterization of a well-defined group of patients with posttreatment lyme disease syndrome. Front. Med. *4*, 224. https://doi.org/10.3389/fmed.2017.00224.

33. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

34. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

35. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*. R29–R17.

36. Smyth, G.K. (2005). Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, S. Dudoit., ed. (Springer), pp. 397–420.

37. The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–d338. https://doi.org/10.1093/nar/gky1055.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PTLD RNA-seq FASTQ files	dbGAP	phs002797.v1.p1
Acute LD RNA-seq FASTQ files	dbGAP	phs002795.v1.p1
Processed data	Zenodo	https://doi.org/10.5281/zenodo.7084176
Software and algorithms		
Deidentified processed RNA-seq data file https://github.com/LymeMIND/LM3-study-supporting-materials	GitHub	N/A
Jupyter Notebook with code to reproduce the analysis https://github.com/LymeMIND/LM3-study-supporting-materials	GitHub	N/A
Jupyter Notebook to reproduce figures	Zenodo	https://doi.org/10.5281/zenodo.7084176

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Avi Ma'ayan (avi.maayan@mssm.edu).

Materials availability

This study did not generate new reagents.

Data and code availability

- The processed deidentified data is available from: https://github.com/LymeMIND/LM3-study-supporting-materials/tree/main/data
- The raw RNA-seq data with more complete clinical metadata is available from dbGAP accession code: phs002795.v1.p1.
- All source code used for the analysis is provided at: https://github.com/LymeMIND/LM3-study-supporting-materials.
- DOI for source code: https://doi.org/10.5281/zenodo.7084176.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The human cohort

The cohorts of patients included in the current analysis are part of large, on-going studies to characterize patients with LD and PTLD. Samples from 152 patients with PTLD were drawn from a cross-sectional study for which detailed recruitment and eligibility information has been previously described.[32] Briefly, participants were largely recruited from a clinic-based population during the period of 2008-2018 and were required to have medical record-confirmed prior LD and appropriate antibiotic treatment. Eligibility criteria included evidence of documented erythema migrans rash, oligoarthritis with joint swelling, facial palsy, neuropathy, meningitis, encephalitis, carditis, or a viral-like illness, as well as concurrent laboratory evidence of infection performed by a laboratory following CDC recommendations for test interpretation. Additionally, participants were required to have continued fatigue, pain, or cognitive dysfunction that affected function, and were excluded for a range of specific medical conditions with significant symptom overlap with PTLD. For the current analysis, we did not require participants to have been ill for at least 6 months at the time of enrollment. The implications of this decision were tested, and the makeup of the biomarker set was largely preserved even without considering the convalescent cohort patients. This is because these patients are uniformly mixed with the other PTLD patients (Figures S1 and S2). Healthy controls were recruited from the same geographic region. They did not have a clinical history for LD and were CDC-negative on two-tier testing for antibodies to *B. burgdorferi*.

Demographics of the human cohort

The PTLD cohort consisted of 152 patients made of 66 females and 86 males. Their average age was 47.27 with a SD of 15.85. Three patients in this cohort were self-identified as Asian, five as Hispanic, and three as Black while the remaining self-identified as White. The acute LD cohort consisted of 72 patients made of 31 females and 41 males. Their average age was 47.19 with a SD of 15.68. One patient self-identified as Asian, and one as Black while the remaining self-identified as White. The healthy control cohort consisted of 45 patients made of 26 females and 19 males. Their average age was 50.29 with a SD of 15.28. Five patients in this cohort were self-identified as Black, five as Hispanic, and one as Native American while the remaining self-identified as White.

RNA-seq profiling of patients

The RNA-seq profiles of participants with PTLD were compared to data from 72 patients with acute Lyme (‘acute LD’) who were then followed longitudinally up to one year after completing appropriate antibiotic treatment (convalescent cohort which was not considered in this study). Participants with acute LD had a physician-diagnosed EM rash present and $\leq$72 h of appropriate antibiotic treatment at the time of enrollment. Finally, 44 healthy control participants without a clinical history of LD who were also two-tier sero-negative for LD were also included. Additional details of the acute LD and control participants, as well as prior analysis of their RNA-seq profiles, were previously published.[9]

The Institutional Review Board of the Johns Hopkins University School of Medicine approved this study, and all participants signed written consent prior to initiation of any study activities.

METHOD DETAILS

Isolation of PBMC

PBMCs were isolated from fresh whole blood using Ficoll (Ficoll-Paque Plus, GE Healthcare) and total RNA was extracted from 10^7 PBMCs using RLT Lysis Buffer (Qiagen) by following manufacturer’s instructions. The NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (Cat# E7765) was used to generate RNA-seq libraries.

Preparation of the samples for RNA sequencing

Poly A RNAs were isolated from total RNAs using NEBNext Poly(A) Magnetic Isolation Module (NEB #E7490) and then fragmented for cDNA synthesis. End repair is performed where 3′ to 5′ exonuclease activity of enzymes removes 3′ overhangs, and the polymerase activity fills in the 5′ overhangs. An ‘A’ base is then added to the 3′ end of the blunt phosphorylated DNA fragments which prepares the DNA fragments for ligation to the sequencing adapters, which have a single ‘T’ base overhang at their 3′ end. Ligated fragments are subsequently size selected through purification using the Sample Purification Beads included in the kit and undergo PCR amplification to prepare the ‘libraries. The BioAnalyzer is used for quality control of the libraries to ensure adequate concentration and appropriate fragment size free of adapter dimers. The resulting library insert size is 200bp-500bp with a median size around 300bp. Libraries were barcoded and pooled for HiSeq2500 sequencing.

RNA sequencing

The prepared samples were processed by an Illumina HiSeq2500 sequencing instrument at the Genomics Core Facility at the Icahn School of Medicine at Mount Sinai.

QUANTIFICATION AND STATISTICAL ANALYSIS

RNA-seq processing

All samples taken from both studies were processed with FastQC, aligned to the human genome (gh38) with the STAR RNA-seq aligner,[33] after which Picard tools were used for gene, exon, and transcript quantification. The RNA-seq gene counts were merged, genes filtered by edgeR[34] quasi-likelihood, log2 transformed, quantile normalized and z-scored. Top differentially expressed genes were calculated using a limma-voom[35,36] applied to the raw counts with a BH-corrected adjusted q-value cutoff of 0.01. Differentially expressed genes are computed by comparing pairwise between healthy controls, acute LD from patients at their first visit, and patients with PTLD.

Enrichment analysis

The differentially expressed genes were submitted to Enrichr,[13] an interactive web tool for performing Enrichment analysis. Each set was submitted independently, and a report of significant hits compiled.

Set overlap analysis

The differentially expressed gene sets were further investigated using the Super-Venn package to visualize multi-set comparisons, helping to contrast gene sets against those in Enrichr. Enrichment analysis of GO Biological Processes[37] using the consensus upregulated genes between acute LD and PTLD revealed significant enrichment for cellular response to molecules of bacterial origin but also to inflammatory response. Consensus and divergent genes between the two groups which did not appear as biomarkers for inflammatory response, or several other viral or bacterial infections were considered as candidate biomarkers.

Classification model construction and candidate biomarker selection

The candidate biomarkers were further filtered by a variance selection criterion which scores biomarkers by total variance divided by inter-group variance and by permutation importance using Logistic Regression classifiers on four classification problems: LD vs. healthy, acute LD vs. healthy, PTLD vs. healthy, and acute LD vs. PTLD. The biomarkers achieving high scores for each of these categories were selected as features for the classification task. Additionally, the top single-gene biomarkers capable of separating samples were highlighted.

Model evaluation

To benchmark the generalizability of the approach, we hold out a third of the patients stratified across: controls, acute LD, and PTLD. Then, we follow the same candidate biomarker selection approach to produce a biomarker set. We constructed four independent pipelines consisting of scaling to unit mean and variance followed by a Logistic Regression classifier trained on the test set and evaluated using the held-out patients. To mitigate class imbalance, we under sampled the test set to have equal number of samples in the positive and negative classes during validation. The performance on the held-out set is visualized with Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves, area under these curves (AUC) is computed and a confusion matrix produced from the true and false positives when considering a cutoff at 50% of the Logistic Regression Classifiers' assigned probability. Additionally, permutation testing is applied on different train test splits to ensure results are consistent across many runs.