scientific reports

OPEN

Recursive ensemble feature selection provides a robust mRNA expression signature for myalgic encephalomyelitis/chronic fatigue syndrome

Paula I. Metselaar[1], Lucero Mendoza-Maldonado[2], Andrew Yung Fong Li Yim[3], Ilias Abarkan[4], Peter Henneman[3], Anje A. te Velde[1], Alexander Schönhuth[5], Jos A. Bosch[6,7], Aletta D. Kraneveld[4] & Alejandro Lopez-Rincon[4,8]✉

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a chronic disorder characterized by disabling fatigue. Several studies have sought to identify diagnostic biomarkers, with varying results. Here, we innovate this process by combining both mRNA expression and DNA methylation data. We performed recursive ensemble feature selection (REFS) on publicly available mRNA expression data in peripheral blood mononuclear cells (PBMCs) of 93 ME/CFS patients and 25 healthy controls, and found a signature of 23 genes capable of distinguishing cases and controls. REFS highly outperformed other methods, with an AUC of 0.92. We validated the results on a different platform (AUC of 0.95) and in DNA methylation data obtained from four public studies on ME/CFS (99 patients and 50 controls), identifying 48 gene-associated CpGs that predicted disease status as well (AUC of 0.97). Finally, ten of the 23 genes could be interpreted in the context of the derailed immune system of ME/CFS.

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a complex disorder characterized by a persistent and debilitating fatigue that lasts for at least six months. According to the 1994 Fukuda case definition, ME/CFS includes four or more of the following symptoms: memory or concentration impairment, sore throat, tender glands, muscle pain, multi-joint pain, headaches, unrefreshing sleep and post-exertion malaise. The latter implies that fatigue symptoms worsen upon minimal mental or physical efforts[1]. Three types of ME/CFS can be distinguished, based on the medical history: 1) post-infection related, 2) (auto)immune disease related, and 3) of unknown origin. The worldwide prevalence of ME/CFS is approximately 0.76–3.28%[2]. Spontaneous recovery is less than 5% and current treatment options are considered suboptimal. ME/CFS is thus a chronic disorder that severely impacts patients' quality of life. The etiology remains elusive, although ME/CFS appears to have a heritable component[3] with genome-wide association studies identifying multiple risk loci[4,5]. Recent evidence suggests that non-genetic factors like infections[6] can induce epigenetic changes[7,8] that might be involved in etiology as well. Currently, there is no prognostic or diagnostic test available[9,10].

Several studies focused on mRNA expression to find biomarkers that support diagnosis and better understanding of etiology. Unfortunately, the results show little consistency. Fang et al.[11] found 164 mRNAs that were significantly differentially expressed in peripheral blood mononuclear cells (PBMCs) from fatigued versus non-fatigued participants ($n = 167$). Presson et al.[12] restructured this data and recoded fatigue severity for the 167 participants.They selected 118 participants for an Integrated Weighted Gene Co-expression Network Analysis

[1]Tytgat Institute for Liver and Intestinal Research, AGEM, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [2]Unidad de Citogenética, Hospital Civil de Guadalajara "Juan I. Menchaca", Guadalajara, Jalisco, Mexico. [3]Department of Clinical Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [4]Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands. [5]Genome Data Science, Faculty of Technology, Bielefeld University, Bielefeld, Germany. [6]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands. [7]Department of Medical Psychology, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [8]Department of Data Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. ✉email: a.lopezrincon@uu.nl

(IWGCNA) that identified 20 candidate genes related to ME/CFS severity. In other cohorts, Gow et al.[13] found 366 differentially expressed mRNAs in PBMCs of eight post-infectious ME/CFS patients versus seven healthy controls and Nguyen et al.[14] identified 176 mRNAs in whole blood of adolescents (18 healthy vs. 29 patients). Between those 366 and 176 mRNAs, measured in slightly different cell populations, only ten overlapped. Finally, Byrnes et al.[15] found no differential expression in peripheral blood leukocytes (PBLs) of 44 monozygotic twin pairs discordant for ME/CFS. In the results from four additional differential expression studies in PBLs, only one gene, *MSN* overlapped between two of them[15].

Consideration of epigenetic factors may help resolve some of the inconsistencies. Epigenetics performs transcriptional regulation in a mitotically heritable fashion, yet can be influenced by non-genetic factors[16]. It encompasses a wide variety of modifications that do not change the DNA sequence, but are thought to affect the accessibility and hence readability thereof. One modification is DNA methylation, the addition of a methyl group to the fifth carbon of cytosine. This process has been described predominantly for cytosines followed by guanine residues (CpG)[17]. An increased concentration of CpGs is often observed in the promoter region of genes, where such clusters are called CpG islands[18]. Promoter methylation is typically inversely correlated with the expression of the associated gene, where a high density of methylated CpGs correlates with lower mRNA expression and vice versa[19,20]. De Vega et al. have performed differential methylation analyses on PMBCs of ME/CFS patients[21–23] and reported between 1,192 and 12,608 differentially methylated CpG sites, related to 826 to 5,544 annotated genes. Using another microarray method with more probes, Trivedi et al.[24] found 17,296 differentially methylated CpG sites related to 6,368 genes. To the best of our knowledge, no study has yet integrated mRNA expression data and DNA methylation data of ME/CFS.

In light of the preceding discussion, we sought to integrate publicly available mRNA expression and DNA methylation data to innovate biomarker research in ME/CFS. To this end, we made use of recursive ensemble feature selection (REFS), a classification pipeline which, when applied in near-analogous settings, has proven to reliably shed light on the corresponding relationships[25]. REFS was used to identify a robust mRNA expression signature that could differentiate between cases and controls. Subsequently, we investigated whether genes found to be differentially expressed in ME/CFS were indicative of differential DNA methylation as well. To equate methods and results of previous studies, we also compared the performance of REFS to IWGCNA and univariate analyses. Finally, we sought to interpret the biological function and relevance of the genes identified by REFS as candidate biomarkers for ME/CFS.

Results

mRNA expression feature selection.

We ran the REFS algorithm ten times on the CAMDA mRNA expression dataset containing 118 samples and the algorithm found the optimal number of predictor genes to be 23 (Fig. 1a; Supplemental Table S1). Multivariate analysis of variance (MANOVA)[26] indicated a statistically significant difference between healthy controls and ME/CFS patients ($F(23, 95) = 5.15$, $p < .0001$; Wilk's $\lambda = 0.445$, R-squared (uncentered) = 0.555). mRNA expression of all 23 predictor genes was down-regulated in ME/CFS patients (Fig. 1b). Next, we compared our results with IWGCNA and univariate feature reduction using χ^2, which were used by Presson et al.[12] and Byrnes et al.[15], respectively, to identify genes that were associated with ME/CFS.

IWGCNA constructs a co-expression network and prioritizes modules of genes based on their association with a phenotype in combination with the presence of a disease-associated genetic variant. Essentially, IWGCNA selects the most important genes by $p - value$ score. Presson et al.[12] reported 20 candidate genes based on mRNA expression data from the CAMDA dataset, the same dataset in which we found 23 predictor genes using REFS. The results of Presson et al. and our REFS algorithm could thus be directly compared. Univariate feature selection reduces the number of features using univariate score metrics, in this case, χ^2. For the comparison, we applied the sci-kit learning implementation[27] to the CAMDA dataset, and selected features according to the highest χ^2 scores with a fixed $k = 23$. Next, we calculated the accuracy and the associated receiver-operating characteristic (ROC) curve for all three algorithms. The REFS algorithm had an average accuracy of 91.57% differentiating controls and patients with ME/CFS globally, with the Passive Aggressive classifier showing a 95.87% accuracy using only the 23 candidate genes (Table 1). Moreover, the area under the curve (AUC) of the ROC curve for REFS was 0.92 (Fig. 2a), which is considered outstanding[28,29]. The compared methods were only slightly better than chance (IWGCNA: AUC = 0.51 (Fig. 2b), univariate analysis: AUC = 0.56 (Fig. 2c). Comparing the genes of interest each of the three methods returned, we found that REFS outperformed both IWGCNA and univariate analysis based on the ROC curve and every performance classifier (Table 1).

Cross platform validation.

We applied the resulting mRNA expression signature to a separate mRNA dataset (GSE14577) as a cross platform validation. Due to the small sample size ($n = 15$), we did not perform feature reduction, but applied the 23 candidate genes directly to the dataset. As the data was measured with a different platform, not all genes overlapped with the CAMDA dataset. Using the DAVID gene functional tool[30], we retrieved eighteen genes. After training the eighteen-gene based model through 5-fold cross validation, we achieved a global classification accuracy of 91.11% differentiating controls and patients with ME/CFS (Table 1) with an ROC AUC of 0.95 (Fig. 2d).

DNA methylation feature selection.

We reduced the number of CpG probes to 278 by selecting for probes associated with the 23 predictor genes. Of these 278 CpGs, REFS identified 48 CpGs as predictive of ME/CFS after ten runs on the combined methylation data (Fig. 3a). MANOVA analysis showed a statistically significant difference between healthy controls and ME/CFS patients ($F(48, 101) = 50.6783$, $p < .0001$; Wilk's $\lambda = 0.034$, R-squared (uncentered) = 0.960). The 48 CpGs reached a global accuracy of 90.17% distinguishing

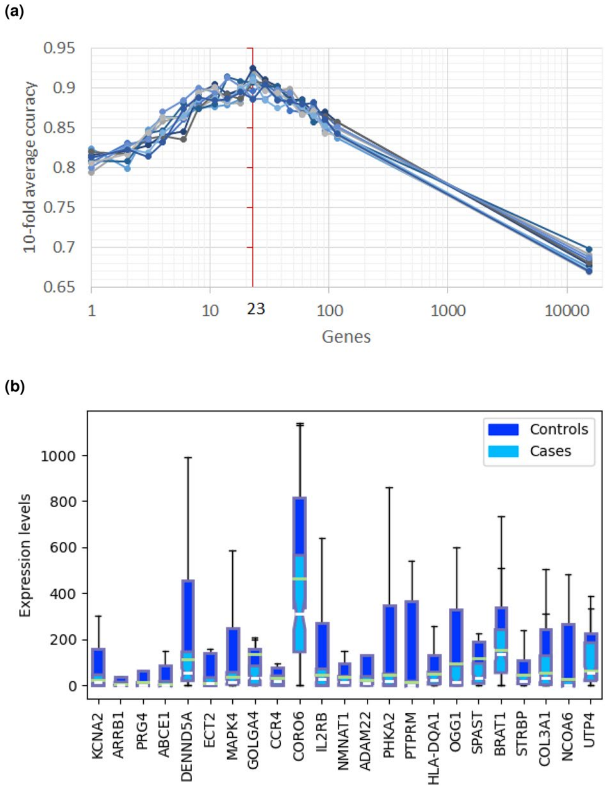

Figure 1. Results of the REFS algorithm run ten times on the mRNA expression data of 118 samples from the CAMDA dataset. (**a**) The optimal number of predictor genes to distinguish 93 cases from 25 controls was 23 (red vertical line). (**b**) mRNA expression levels of the 23 predictor genes for 93 cases (light blue) and 25 controls (dark blue) in box-and-whisker plots. Outliers were omitted for visualization purposes.

	REFS		IWGCNA		χ^2		mRNA validation		DNA methylation	
Classifier	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Gradient boosting (n_estimators=300)	0.898	0.0647	0.7417	0.0975	0.7801	0.0775	0.8111	0.1641	0.7852	0.1023
Random forest (n_estimators=300)	0.8575	0.0714	0.7298	0.0576	0.799	0.0459	0.9333	0.0943	0.8181	0.0867
Logistic regression	0.9595	0.0533	0.7138	0.0913	0.8052	0.0530	0.9333	0.0943	0.9457	0.042
Passive aggressive	0.9587	0.0675	0.6057	0.1744	0.8151	0.0693	0.9333	0.0943	0.98	0.0306
SGD	0.9595	0.0533	0.6264	0.0659	0.7564	0.0914	0.9333	0.0943	0.9324	0.0437
SVC (linear)	0.9421	0.0754	0.7472	0.0848	0.7962	0.0408	0.9333	0.0943	0.9667	0.0683
Ridge	0.9023	0.0691	0.7214	0.0858	0.8143	0.0616	0.9333	0.0943	0.9733	0.0327
Bagging (n_estimators=300)	0.8478	0.0495	0.7144	0.0841	0.7801	0.0940	0.8778	0.0875	0.8119	0.0586
Average	0.9157	0.0630	0.7001	0.0927	0.7933	0.0667	0.9111	0.1022	0.9017	0.0581

Table 1. Classification accuracy of the 23 predictor genes obtained with REFS, the 20 genes obtained with IWGCNA, and the 23 genes obtained with univariate analysis based on the same mRNA expression data of 118 samples from the CAMDA dataset Classification accuracy of eighteen of the 23 predictor genes applied to a validation mRNA dataset (GSE14577), and of 48 predictor CpGs obtained with REFS based on merged DNA methylation data. μ, mean; σ, standard deviation.

patients and healthy controls (Table 1). When comparing CpG methylation between cases and controls, most CpG sites showed enhanced methylation in ME/CFS patients (Fig. 3c). Finally, the calculated ROC curve had an AUC of 0.97 (Fig. 3b), pointing to a clear distinction between patients and controls when applying the 48 predictor CpG sites.

Biological interpretation of the gene signature. To put our findings in a biological context, we interrogated the existing literature on the proteins encoded by the 23 predictor genes produced by the REFS algorithm. These results were obtained in PBMCs, therefore we focused on ten proteins, MAPK4, ARRB1, GOLGA4,

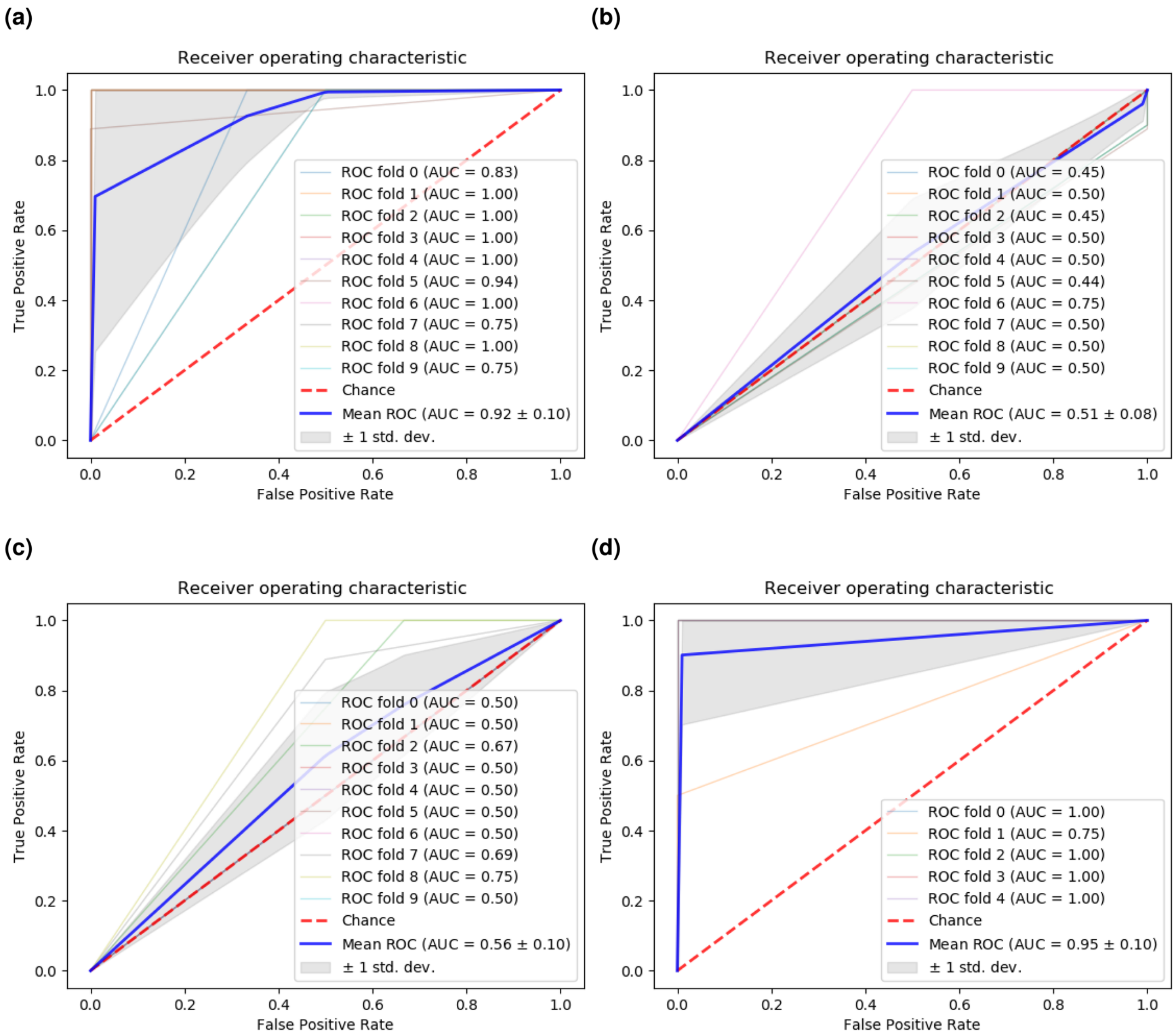

Figure 2. ROC curves for (**a**) REFS, (**b**) IWGCNA, and (**c**) univariate analysis applied to the same mRNA expression data of 118 samples from the CAMDA dataset. (**d**) ROC curve for the validation of the obtained 23-gene signature on a separate dataset (GSE14577). The 5-fold cross validation was performed with the eighteen genes available in GSE14577.

ABCE1, PHKA2, IL2RB, CCR4, HLA-DQA1, PRG4 and OGG1, that acted in immune pathways. The mRNA transcripts encoding these proteins were all downregulated in ME/CFS patients compared to healthy controls (Fig. 1a). Protein functions are visualized in Fig. 4 and briefly described below.

Intracellular signaling. In response to extracellular stimuli binding to immune cell receptors, mitogen-activated protein kinases (MAPKs) transduce intracellular signals through protein phosphorylation cascades to mediate the cellular response. MAPK4 (also known as ERK4) directly phosphorylates AKT, subsequently activating the mTOR signaling pathway and ultimately regulating cell survival, proliferation and metabolism. Decreased MAPK4 activity is associated with decreased AKT activity and proliferation[31]. Another intracellular mediator of the immune response is ARRB1 (β-arrestin 1, also known as arrestin 2). ARRB1 is found in high levels in PBLs where it inhibits G-protein coupled receptors that are phosphorylated by β-adrenergic receptor kinase. GOLGA4 (golgin A4) is a member of the golgin protein family. Upon macrophage LPS activation, formation of GOLGA4-demarked carriers is upregulated, which was found to increase TNF trafficking and subsequent secretion. Depletion of GOLGA4 consequently impairs TNF secretion[32]. ATP-binding cassette sub-family E member 1 (ABCE1), a cell membrane transporter, inhibits RNase L activity. RNase L is important in immune defense, as it degrades all RNA within the cell upon viral infection and releases antiviral IFNγ[33]. ABCE1 was downregulated in PBLs of ME/CFS patients, which correlated with upregulated RNase L[34]. Finally, PHKA2 is the hepatic isoform of the alpha subunit of phosphorylase b kinase. This enzyme phosphorylates glycogen phosphorylase b, converting it to the active glycogen phosphorylase a, which then breaks down stored glycogen to glucose. This glucose is converted through glycolysis to ATP, to meet metabolic demands. PHKA2 is thus important in provid-

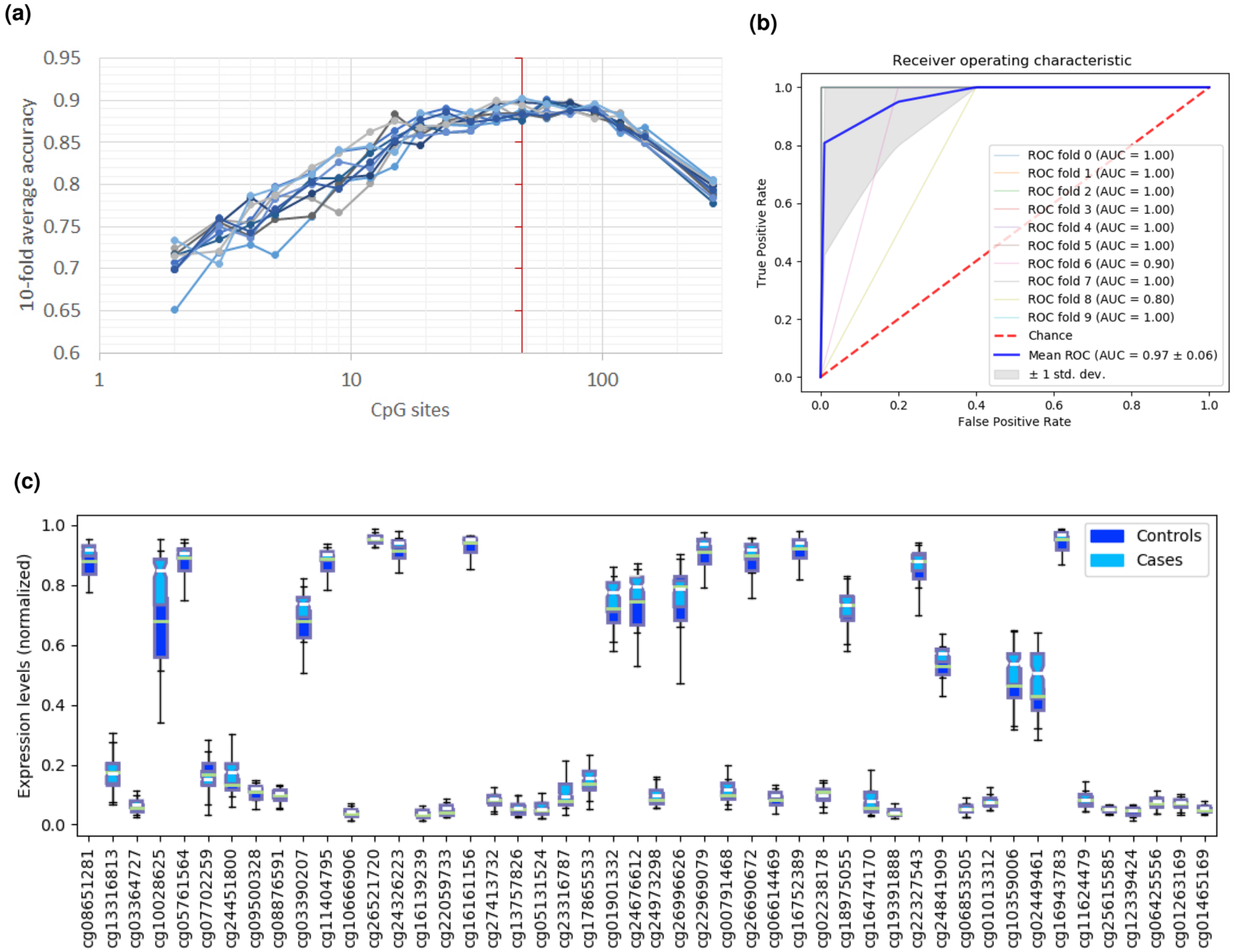

Figure 3. Results of the REFS algorithm run ten times on the merged DNA methylation datasets restricted to 278 probes associated with the 23 candidate genes. (**a**) The optimal number of predictor CpGs to distinguish 99 cases from 50 controls was 48 (red vertical line). (**b**) ROC curve of the 48 predictor CpGs. c) DNA methylation levels (normalized using *Standard* scaler) of the 48 predictor CpG sites for 99 cases (light blue) and 50 controls (dark blue).

ing energy to cells by maintaining glucose levels. Mutations in *PHKA2* gene cause glycogen storage disease type IXa, characterized by hypoglycemia, hepatomegaly, elevated liver enzymes, growth retardation and motor delay, hypercholesterolemia and hypertriglyceridemia. Some patients report fatigue and muscle weakness[35].

Immune cell receptors and ligands. IL2RB is the β-subunit of the IL2 receptor, a receptor for IL2 that is involved in the differentiation of regulatory-, effector-, memory-, type 1 and 2 helper T cells. IL2RB-deficient patients show dysregulated IL2 and IL15 signaling, enhanced natural killer cell levels, and subsequent immunodeficiency and impaired antiviral immunity[36]. A SNP in the 3-prime untranslated region of the gene was associated with ME/CFS, interfering with miRNA binding which could reduce IL2 receptor function[37]. CCR4 (CC chemokine type 4 receptor) is a G protein-coupled receptor expressed on regulatory and type 2 helper T cells that binds CCL17 and CCL22. These chemokines cause chemotaxis of the cell, traffic leukocytes and are involved in development, homeostasis and function of the immune system[38]. Blockage of CCR4 by mogamulizumab induced fatigue in human subjects[39]. HLA-DQA1 is the α-chain of the MHC class II receptor expressed on antigen-presenting cells such as macrophages, B lymphocytes and dendritic cells. The MHC class II receptor presents peptides to T cell receptors to activate the immune system upon viral infection. Proteoglycan 4 (PRG4) is an immune cell ligand. It is a lubricant, reducing friction between joints or boundaries, and altered expression is implicated in rheumatoid arthritis. In addition, PRG4 was found to 'coat' neutrophils and macrophages, perhaps by binding TLR2, -4, -5 or CD44, thereby preventing activation by low levels of pro-inflammatory cytokines. Upon inflammation, when an immune response is required, PRG4 was shed from the immune cells to allow activation[40].

DNA damage repair. DNA repair enzyme 8-oxoguanine DNA glycosylase (OGG1) specifically excises the most pre-mutagenic oxidative base lesion 8-oxoguanine. Such DNA damage caused by exposure to reactive oxygen

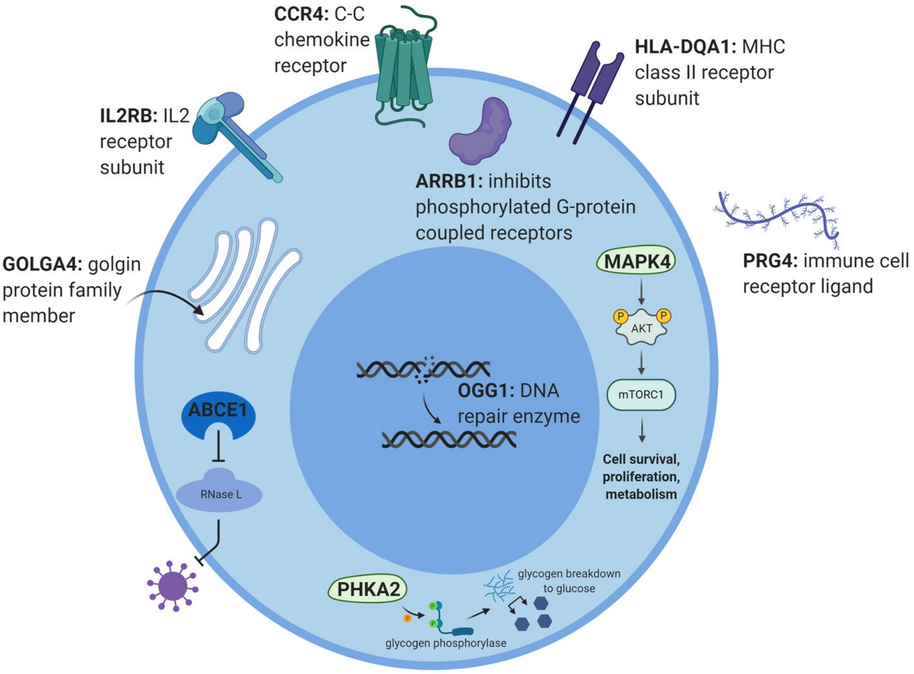
CCR4: C-C chemokine receptor
HLA-DQA1: MHC class II receptor subunit
IL2RB: IL2 receptor subunit
ARRB1: inhibits phosphorylated G-protein coupled receptors
PRG4: immune cell receptor ligand
GOLGA4: golgin protein family member
MAPK4
P
P
AKT
mTORC1
Cell survival, proliferation, metabolism
OGG1: DNA repair enzyme
ABCE1
RNase L
PHKA2
glycogen breakdown to glucose
glycogen phosphorylase

Figure 4. Visualization of the functions and locations of ten proteins in a hypothetical immune cell setting. All ten mRNA transcripts were downregulated in PBMCs of ME/CFS patients compared to healthy controls. Created with BioRender.com.

species (ROS) leads to mutagenesis or cell death. OGG1 activity prevents mutations and Alzheimer's disease patients with an OGG1 polymorphism show increased DNA damage[41].

Discussion

ME/CFS is a chronic disorder characterized by persistent, disabling fatigue for which no diagnostic or prognostic test nor complete treatment is available. Several studies have sought to define biomarkers for ME/CFS by performing differential mRNA expression or DNA methylation analysis. However, as Byrnes et al.[15] pointed out, these results were study-dependent and no definitive biomarkers were found. We used a state-of-the-art machine learning technique to distinguish ME/CFS patients from healthy controls across different platforms, several cohorts and on different levels of gene expression regulation. To our knowledge, this was the first time such a technique was used in mRNA expression data and validated in DNA methylation data.

In this study, we implemented the REFS algorithm on public mRNA expression data and found 23 genes whose changes in expression levels were able to distinguish ME/CFS patients from healthy controls. The 23 predictor genes differentiated between cases and controls with 91.57% global accuracy and returned a ROC AUC of 0.92. In addition, 48 CpG methylation sites associated with these genes were predictive of ME/CFS in four merged DNA methylation studies. Moreover, all 23 candidate genes were downregulated in ME/CFS patients while DNA methylation of almost all 48 CpG sites was enhanced. This inverse correlation between mRNA expression and DNA methylation, across different samples and studies, legitimizes the results of our study. As previously demonstrated[25], REFS identifies a more accurate, robust gene signature than previous methods. Comparing the gene signature returned by three different methods, based on the same data, REFS outperformed both IWGCNA and univariate analysis in separating ME/CFS patients and healthy controls with a ROC AUC of 0.92. The AUC of the gene signature applied to a different platform was 0.95, and the AUC even reached 0.97 when plotting the sensitivity and specificity of the 48 predictor CpG sites.

To show the relevance of the returned predictor genes, we investigated the biological functions of ten encoded proteins active in immune pathways. This decision was based on the mRNA expression being measured in PBMCs and the literature pointing towards an important role for the immune system in ME/CFS. Sotzny et al.[42] reviewed autoimmunity in ME/CFS, concluding that immunologic and metabolic alterations were often reported. The authors stress the potential importance of autoantibodies in the disorder and the proposed role of preceding infections. Downregulation of ABCE1, one of the encoded proteins identified in our study, concurs with the presence of previous viral infections, as the protein inhibits RNase L's viral RNA degrading activity. Similarly, ARRB1 protein was decreased after Epstein-Barr virus-infection in mice[43]. Its downregulation in our study concurs with this finding. Recently, Mandarano et al.[44] described evidence of immune involvement in their study of 53 ME/CFS patients. They specifically focused on T cells, showing that CD8+ T cells derived from patients had lower mitochondrial membrane potential, which points towards T cell exhaustion. PHKA2 is necessary for the first step in breaking down glycogen to glucose. Its downregulation could contribute to impaired glycolysis in immune cells. In ME/CFS, CD4+ and CD8+ T cells had impaired resting glycolysis, and plasma glucose was reduced. CD8+

T cells showed an impaired metabolic response to activation[44]. Another study found that glycogen metabolism regulates the immune functions of dendritic cells[45]. Inhibiting glycogen phosphorylase impaired their ability to produce inflammatory cytokines and stimulate T cells. These findings combined suggest that reduced PHKA2 in ME/CFS might inhibit glycogen phosphorylase activation and thus dendritic cell functioning.

Several (subunits of) immune cell receptors were also part of our gene signature and downregulated in ME/CFS. IL2RB, CCR4 and HLA-DQA1 are vital elements in proper immune response, dysregulation, whether up or down, could be evidence of a disturbed immune system or be the cause of it. The same holds true for decreased MAPK4 expression, an ubiquitous transducer of intracellular signals in response to immune cell receptor binding. Further downstream, GOLGA4 was upregulated in response to macrophage LPS activation to increase TNF secretion. TNF is the main pro-inflammatory cytokine secreted by inflammatory macrophages, and its release is important for enhancing the activation and recruitment of T cells, ensuring robust innate and adaptive immune responses. In our study however, GOLGA4 was downregulated in PBMCs of ME/CFS patients, potentially causing impaired TNF secretion and subsequently an impaired immune response to inflammation. Furthermore, decreased expression of PRG4 leads to reduced anti-inflammatory action of this protein. By binding immune cells receptors, PRG4 prevents activation by low levels of circulating pro-inflammatory cytokines. PRG4 could thus be important in low-grade inflammation causing ME/CFS[46].

Finally, evidence has emerged that oxidative stress levels are raised in ME/CFS, for example in response to exercise, perhaps causing some of the symptoms seen in ME/CFS[47,48]. DNA damage caused by exposure to ROS leads to mutagenesis or cell death, OGG1 specifically repairs this DNA damage. OGG1 depletion in human monocyte-derived dendritic cells inhibited enhanced cell surface molecule-expression and secretion of pro-inflammatory cytokines upon exposure to 8-oxoguanine base lesions. This suggests that OGG1 is important for dendritic cell activation in response to ROS[49]. Concurrently, 8-oxoguanine base lesions did not cause acute or systemic inflammation in Ogg1-deficient mice[50]. As we found that OGG1 is downregulated in ME/CFS patients, while oxidative stress levels are increased, DNA damage might be increased, which in turn causes the release of danger associated molecular patterns (DAMPs) and activates the innate immune system. We can conclude, from the various roles of these ten genes, that their downregulation may not only contribute to immune activation, but also towards a general dysregulation of the immune response. Whether all genes are causative of the ME/CFS phenotype, or some are mere consequences of immune mayhem in ME/CFS patients remains to be investigated.

Our results return a promising gene signature for ME/CFS that needs to be validated in a well characterized clinical cohort to study its use as a diagnostic tool. In this cohort, the number of cases and controls should be balanced, as the current study suffers from cases outnumbering controls in both the mRNA expression and DNA methylation datasets. We compensated for the imbalance with stratified folds in the cross-validation. Finally, the investigation of the predictor genes has thus far been limited to a literature review. In vitro experiments with PBMCs should provide additional information regarding gene function.

To conclude, we found a mRNA expression signature of 23 genes for ME/CFS capable of separating cases and controls. These candidate genes could potentially be used as biomarkers for diagnostic purposes. In addition, ten of these genes could be interpreted in the context of a derailed immune system in ME/CFS. Those genes could be investigated further for target finding and development of future treatments for ME/CFS.

Methods

mRNA expression data.

The mRNA expression data used in this study was retrieved as provided by Presson et al.[12], which was made available by the CDC for the 2006 Critical Assessment of Microarray Data Analysis (CAMDA) conference (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/CFS/). We refer to this dataset as the CAMDA dataset. The CAMDA- and subsequent datasets were chosen because they are the only large, publicly available studies performed on PBMCs. PBMCs are thought to be involved in ME/CFS pathophysiology[42,44]. Included in the CAMDA dataset are mRNA expression levels in PBMCs and fatigue severity status from 118 participants. ME/CFS was determined based on the Fukuda case definition criteria[1], and fatigue severity was estimated based on clustered scores from the SF-36 fatigue score, Multidimensional Fatigue Inventory, and Symptom Inventory Case Definition Score[12]. mRNA expression was measured on the MWG Biotech microarray platform, containing probes for approximately 20,000 transcripts.

To analyze the data, we first re-annotated the probes with HGNC gene symbols using DAVID[51], yielding 15,419 gene-associated probes. The data was then normalized using the *Standard* scaler from sci-kit learning toolbox[27]. Samples with low fatigue severity were encoded as 0 (controls) and samples with moderate or high fatigue severity were encoded as 1 (cases), based on the clustered scores described by Presson et al.[12]. Altogether, 93 cases and 25 controls were included for REFS. A 10-fold cross-validation with stratified folds was performed to accommodate the unbalanced classes.

As a test dataset for the algorithm, mRNA expression levels in PBMCs from post-infectious, male ME/CFS patients ($n = 8$) and healthy male controls ($n = 7$) were included as provided by Gow et al.[13] in the Gene Expression Omnibus (GEO)[52] GSE14577. ME/CFS was determined with the Fukuda case definition criteria[1] and mRNA expression was measured with Affymetrix Human Genome U133A Array. To analyze the GSE14577 data, the *series matrix* file was employed and scaled using the *Standard* scaler. Given the low number of samples, a 5-fold cross-validation was performed only for the analysis. A summary of the datasets is available in Table 2.

DNA methylation data.

DNA methylation data was retrieved from GEO (Table 2). The DNA methylation datasets pertained methylation profiles in PBMCs of ME/CFS patients and healthy controls measured with the Illumina HumanMethylation Infinium 450k BeadChip array (450k) and its successor, the Illumina HumanMethylation Infinium EPIC BeadChip array (EPIC). The 450k and the EPIC measured the methylation status of 485,577 and 865,859 CpGs, respectively, with approximately 90% of the 450k probes being present on the

GEO accession number/name	Reference	Samples (n)	Measuring platform	Probes (n)	Type
CAMDA	Presson et al.[12]	118	MWG Biotech microarray	15,419	mRNA expression
GSE14577	Gow et al.[13]	15	Affymetrix Human Genome U133A Array	22,284	mRNA expression
GSE59489	de Vega et al.[21]	24	450k (GPL13534)	485,577	DNA methylation
GSE93266	de Vega et al.[22]	75	450k (GPL13534)	485,512	DNA methylation
GSE102504	de Vega et al.[23]	25	450k (GPL13534)	467,971	DNA methylation
GSE111183	Trivedi et al.[24]	25	EPIC (GPL21145)	866,895	DNA methylation

Table 2. Characteristics of the mRNA expression and DNA methylation datasets.

EPIC[53]. In all four studies, ME/CFS was determined with the Fukuda and Canadian case definitions[1,54]. From the GEO repositories, *series matrix* files containing the normalized percentage methylation per CpG were obtained. The data was pre-processed by scaling the percentage methylation per sample using the *Standard* scaler, after which the individual datasets were merged. Subsequently, the data was split for 10-fold cross-validation with stratified folds. In total, the methylation data included 149 samples (99 cases and 50 controls).

REFS on mRNA expression data.

Several studies have employed univariate feature selection, with statistic metrics such as χ^2 or $F - value$, that reduce the features using the highest scores, yielding study-specific results. To overcome this issue, REFS was used to identify a mRNA expression signature that was able to classify disease status in a PMBC study of ME/CFS patients and healthy controls. The utility and accuracy of REFS compared to other methods was previously demonstrated in miRNA and mRNA datasets[25,55]. In short, the REFS algorithm reduces the number of features to the most significant ones by combining the results of separate classifiers of distinct topologies. The algorithm gives each feature a score, based on how each different algorithm used it. In case of a tree-based algorithm, it depends on how many times the feature appeared in the tree. If it is a coefficient based algorithm, it depends on the value of the coefficient, the highest being the most important. This scoring is consistent with the Borda method[56], the difference being that 10-fold cross validation is included for scoring. After each iteration, the most important features are selected (Fig. 5).

The REFS algorithm (Algorithm 1) was implemented on the 15,419 gene-associated probes through 10-fold cross-validation with the following classifiers: Bagging, Gradient Boosting, Random Forest, Logistic Regression, Passive Aggressive, SGD, SVC (linear), and Ridge. During the first iteration, the algorithm reduced the total features to the most significant ones, after which it reduced the number of variables by 20% per step with a stop parameter of 70% accuracy as indicated in Fig. 5.

Algorithm 1: Recursive Ensemble Feature Selection.

```
Normalize dataset on each of the $F$ features, Divide dataset in $N$ folds, Select $K$ classifiers, Choose the number of features in the signature $S$;
while Global Accuracy > 70% do
    for each fold $n$ of $N$ do
        for each classifier $k$ of $K$ do
            Train classifier $k_n$ on all folds minus $n$, using all features;
            Test classifier $k_n$ on fold $n$;
            Obtain sorted list $l_{kn}$ of features from $k_n$;
            Assign weight $w_{fnk}$ to each $f$ of the $F$ features;
            for each feature $f$ of $F$ do
                if $f$ is among the top $S$ features in $l_{kn}$ then
                    $w_{fnk} = 1$
                else
                    $w_{fnk} = 0$
    $N_c = N \cdot K$;
    for each feature $f$ do
        $N_t = \sum_n^N \sum_k^K w_{fnk}$;
        $s_f = N_t / N_c$;
    Select $S$-feature signature, from features with highest $s_f$;
    for each fold $n$ of $N$ do
        for each classifier $k$ of $K$ do
            Train classifier $k_n$ on all folds minus $n$, using signature;
            Test classifier $k_n$ on fold $n$;
    Compare performance of classifiers using all features and signature;
```

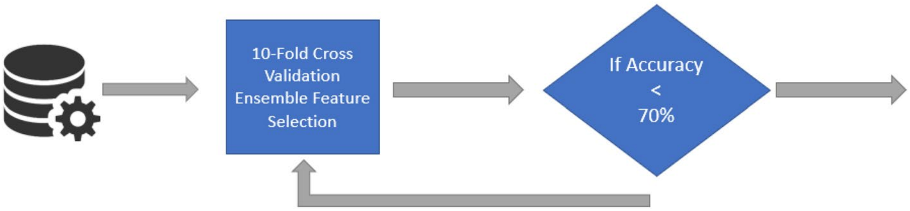

Figure 5. Overview of the REFS algorithm using the 10-fold global accuracy > 70% as a stop parameter.

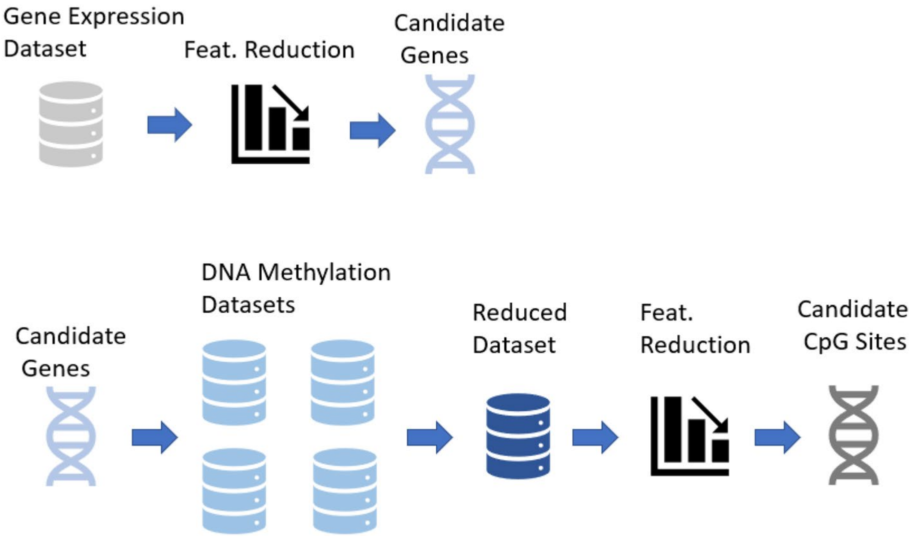

Figure 6. Overview of the pipeline to reduce the number of methylation data probes for REFS by selecting CpGs associated with the candidate genes.

REFS on DNA methylation data of candidate genes. The DNA methylation platforms were represented by a large amount of features, 485,577 and 865,859 features by the 450k and EPIC, respectively. Therefore, the search space was reduced by adopting a candidate gene approach, in which CpGs were only selected if they associated with the genes found using REFS on the CAMDA dataset. Practically, CpG probes were selected from the datasheet of the platform GPL13534 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc/=GPL13534) when the *UCSC_RefGene_Name* field matched any of the genes previously selected by the ensemble feature selection approach. This process is explained in Fig. 6.

Biological interpretation of the mRNA expression signature. To put the genes identified with the REFS algorithm in the mRNA expression dataset in a biological perspective, the protein encoded by each gene was investigated in the existing literature available on Pubmed and Google Scholar (March 2020). Because the analyses were performed with genetic material derived from PBMCs, protein function was investigated in the context of the immune system. After initial analysis, ten proteins that functioned in immune pathways were selected for further literature review.

Received: 22 August 2020; Accepted: 27 January 2021
Published online: 25 February 2021

References

1. Fukuda, K. *et al.* The chronic fatigue syndrome: A comprehensive approach to its definition and study. *Ann. Intern. Med.* **121**, 953–959. https://doi.org/10.7326/0003-4819-121-12-199412150-00009 (1994).
2. Johnston, S., Brenu, E. W., Staines, D. & Marshall-Gradisnik, S. The prevalence of chronic fatigue syndrome/myalgic encephalomyelitis: A meta-analysis. *Clin. Epidemiol.* **5**, 105. https://doi.org/10.2147/CLEP.S39876 (2013).
3. Walsh, C. M., Zainal, N. Z., Middleton, S. J. & Paykel, E. S. A family history study of chronic fatigue syndrome. *Psychiatr. Genet.* **11**, 123–128. https://doi.org/10.1097/00041444-200109000-00003 (2001).
4. Schlauch, K. A. *et al.* Genome-wide association analysis identifies genetic variations in subjects with myalgic encephalomyelitis/chronic fatigue syndrome. *Transl. Psychiatry* **6**, e730–e730. https://doi.org/10.1038/tp.2015.208 (2016).

5. Perez, M. *et al.* Genetic predisposition for immune system, hormone, and metabolic dysfunction in myalgic encephalomyelitis/chronic fatigue syndrome: A pilot study. *Front. Pediatr.* **7**, 206. https://doi.org/10.3389/fped.2019.00206 (2019).
6. Rasa, S. *et al.* Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *J. Transl. Med.* **16**, 268. https://doi.org/10.1186/s12967-018-1644-y (2018).
7. Almenar-Pérez, E. *et al.* Epigenetic components of myalgic encephalomyelitis/chronic fatigue syndrome uncover potential transposable element activation. *Clin. Therap.* **41**, 675–698. https://doi.org/10.1016/J.CLINTHERA.2019.02.012 (2019).
8. Herrera, S., de Vega, W. C., Ashbrook, D., Vernon, S. D. & McGowan, P. O. Genome-epigenome interactions associated with myalgic encephalomyelitis/chronic fatigue syndrome. *Epigenetics* **13**, 1174–1190. https://doi.org/10.1080/15592294.2018.1549769 (2018).
9. Smith, M. E. B. *et al.* Treatment of myalgic encephalomyelitis/chronic fatigue syndrome: A systematic review for a national institutes of health pathways to prevention workshop. *Ann. Intern. Med.* **162**, 841–50. https://doi.org/10.7326/M15-0114 (2015).
10. Cortes Rivera, M., Mastronardi, C., Silva-Aldana, T. C., Arcos-Burgos, M. & Lidbury, A. B. Myalgic encephalomyelitis/chronic fatigue syndrome: A comprehensive review. *Diagnostics*https://doi.org/10.3390/diagnostics9030091 *(2019).*
11. Fang, H. *et al.* Gene expression profile exploration of a large dataset on chronic fatigue syndrome. *Pharmacogenomics*https://doi.org/10.2217/14622416.7.3.429 *(2006).*
12. Presson, A. P. *et al.* Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst. Biol.* **2**, 95. https://doi.org/10.1186/1752-0509-2-95 (2008).
13. Gow, J. W. *et al.* A gene signature for post-infectious chronic fatigue syndrome. *BMC Med. Genom.* **2**, 38. https://doi.org/10.1186/1755-8794-2-38 (2009).
14. 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.* **15**, 102. https://doi.org/10.1186/s12967-017-1201-0 (2017).
15. Byrnes, A. *et al.* Gene expression in peripheral blood leukocytes in monozygotic twins discordant for chronic fatigue: No evidence of a biomarker. *PLoS ONE* **4**, e5805. https://doi.org/10.1371/journal.pone.0005805 (2009).
16. Nanney, D. L. Epigenetic control systems. *Proc. Natl. Acad. Sci.* **44**, 712–717. https://doi.org/10.1073/pnas.44.7.712 (1958).
17. Jang, H., Shin, W., Lee, J. & Do, J. Cpg and non-cpg methylation in epigenetic gene regulation and brain function. *Genes* **8**, 148. https://doi.org/10.3390/genes8060148 (2017).
18. Bird, A. P. Cpg-rich islands and the function of DNA methylation. *Nature* **321**, 209–213. https://doi.org/10.1038/321209a0 (1986).
19. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321. https://doi.org/10.1038/nature14192 (2015).
20. Stepper, P. *et al.* Efficient targeted DNA methylation with chimeric dcas9-dnmt3a-dnmt3l methyltransferase. *Nucleic Acids Res.* **45**, 1703–1713. https://doi.org/10.1093/nar/gkw1112 (2017).
21. de Vega, W. C., Vernon, S. D. & McGowan, P. O. DNA methylation modifications associated with chronic fatigue syndrome. *PLoS ONE* **9**, e104757. https://doi.org/10.1371/journal.pone.0104757 (2014).
22. de Vega, W. C., Herrera, S., Vernon, S. D. & McGowan, P. O. Epigenetic modifications and glucocorticoid sensitivity in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *BMC Med. Genom.* **10**, 11. https://doi.org/10.1186/s12920-017-0248-3 (2017).
23. de Vega, W. C., Erdman, L., Vernon, S. D., Goldenberg, A. & McGowan, P. O. Integration of DNA methylation & health scores identifies subtypes in myalgic encephalomyelitis/chronic fatigue syndrome. *Epigenomics* **10**, 539–557. https://doi.org/10.2217/epi-2017-0150 (2018).
24. Trivedi, M. S. *et al.* Identification of myalgic encephalomyelitis/chronic fatigue syndrome-associated DNA methylation patterns. *PLoS ONE* **13**, e0201066. https://doi.org/10.1371/journal.pone.0201066 (2018).
25. Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A. & Tonda, A. Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection. *BMC Bioinform.* **20**, 480. https://doi.org/10.1186/s12859-019-3050-8 (2019).
26. O'Brien, R. G. & Kaiser, M. K. Manova method for analyzing repeated measures designs: An extensive primer. *Psychol. Bull.* **97**, 316. https://doi.org/10.1037/0033-2909.97.2.316 (1985).
27. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
28. Šimundić, A.-M. Measures of diagnostic accuracy: Basic definitions. *EJIFCC* **19**, 203 (2009).
29. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d (2010).
30. Huang, D. W. *et al.* The David gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183. https://doi.org/10.1186/gb-2007-8-9-r183 (2007).
31. Wang, W. *et al.* Mapk4 overexpression promotes tumor progression via noncanonical activation of akt/mtor signaling. *J. Clin. Investig.* **129**, 66. https://doi.org/10.1172/JCI97712 (2019).
32. Lieu, Z. Z. *et al.* A trans-golgi network golgin is required for the regulated secretion of TNF in activated macrophages in vivo. *Proc. Natl. Acad. Sci. USA* **105**, 3351–3356. https://doi.org/10.1073/pnas.0800137105 (2008).
33. Tian, Y., Han, X. & Tian, D. L. The biological regulation of abce1. *IUBMB Life* **64**, 795–800. https://doi.org/10.1002/iub.1071 (2012).
34. Vojdani, A., Choppa, P. C. & Lapp, C. W. Downregulation of rnase l inhibitor correlates with upregulation of interferon-induced proteins (2–5a synthetase and rnase l) in patients with chronic fatigue immune dysfunction syndrome. *J. Clin. Lab. Immunol.* **50**, 1–16 (1998).
35. Beauchamp, N. J. *et al.* Glycogen storage disease type ix: High variability in clinical phenotype. *Mol. Genet. Metab.* **92**, 88–99. https://doi.org/10.1016/j.ymgme.2007.06.007 (2007).
36. Fernandez, I. Z. *et al.* A novel human il2rb mutation results in t and nk cell-driven immune dysregulation. *J. Exp. Med.* **216**, 1255–1267. https://doi.org/10.1084/jem.20182015 (2019).
37. Rajeevan, M. S., Dimulescu, I., Murray, J., Falkenberg, V. R. & Unger, E. R. Pathway-focused genetic evaluation of immune and inflammation related genes with chronic fatigue syndrome. *Hum. Immunol.* **76**, 553–560. https://doi.org/10.1016/j.humimm.2015.06.014 (2015).
38. Yoshie, O. & Matsushima, K. Ccr4 and its ligands: From bench to bedside. *Int. Immunol.* **27**, 11–20. https://doi.org/10.1093/intimm/dxu079 (2014).
39. Kasamon, Y. L. *et al.* Fda approval summary: Mogamulizumab-kpkc for mycosis fungoides and sezary syndrome. *Clin. Cancer Res.* **25**, 7275–7280. https://doi.org/10.1158/1078-0432.Ccr-19-2030 (2019).
40. Das, N., Schmidt, T. A., Krawetz, R. J. & Dufour, A. Proteoglycan 4: From mere lubricant to regulator of tissue homeostasis and inflammation: Does proteoglycan 4 have the ability to buffer the inflammatory response?. *BioEssays* **41**, 1800166. https://doi.org/10.1002/bies.201800166 (2019).
41. Dinçer, Y. *et al.* DNA repair gene ogg1 polymorphism and its relation with oxidative DNA damage in patients with Alzheimer's disease. *Neurosci. Lett.* **709**, 134362. https://doi.org/10.1016/j.neulet.2019.134362 (2019).
42. Sotzny, F. *et al.* Myalgic encephalomyelitis/chronic fatigue syndrome—Evidence for an autoimmune disease. *Autoimmun. Rev.* **17**, 601–609. https://doi.org/10.1016/j.autrev.2018.01.009 (2018).
43. Williams, M. V., Cox, B., Lafuse, W. P. & Ariza, M. E. Epstein-barr virus dutpase induces neuroinflammatory mediators: Implications for myalgic encephalomyelitis/chronic fatigue syndrome. *Clin. Ther.* **41**, 848–863. https://doi.org/10.1016/j.clinthera.2019.04.009 (2019).
44. Mandarano, A. H. *et al.* Myalgic encephalomyelitis/chronic fatigue syndrome patients exhibit altered t cell metabolism and cytokine associations. *J. Clin. Investig.* **130**, 1491–1505. https://doi.org/10.1172/JCI132185 (2020).

45. Thwe, P. M. *et al.* Cell-intrinsic glycogen metabolism supports early glycolytic reprogramming required for dendritic cell immune responses. *Cell Metab.* **26**, 558-567.e5. https://doi.org/10.1016/j.cmet.2017.08.012 (2017).
46. Montoya, J. G. *et al.* Cytokine signature associated with disease severity in chronic fatigue syndrome patients. *Proc. Natl. Acad. Sci.* **114**, E7150–E7158. https://doi.org/10.1073/pnas.1710519114 (2017).
47. Jammes, Y. & Retornaz, F. Understanding neuromuscular disorders in chronic fatigue syndrome. *F1000Res*https://doi.org/10.12688/f1000research.18660.1 *(2019).*
48. Kennedy, G. *et al.* Oxidative stress levels are raised in chronic fatigue syndrome and are associated with clinical symptoms. *Free Radic. Biol. Med.* **39**, 584–9. https://doi.org/10.1016/j.freeradbiomed.2005.04.020 (2005).
49. Pazmandi, K. *et al.* Oxidized base 8-oxoguanine, a product of DNA repair processes, contributes to dendritic cell activation. *Free Radic. Biol. Med.* **143**, 209–220. https://doi.org/10.1016/j.freeradbiomed.2019.08.010 (2019).
50. Visnes, T. *et al.* Small-molecule inhibitor of ogg1 suppresses proinflammatory gene expression and inflammation. *Science* **362**, 834–839. https://doi.org/10.1126/science.aar8048 (2018).
51. Dennis, G. *et al.* David: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, R60. https://doi.org/10.1186/gb-2003-4-9-r60 (2003).
52. Edgar, R. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210. https://doi.org/10.1093/nar/30.1.207 (2002).
53. Pidsley, R. *et al.* Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208. https://doi.org/10.1186/s13059-016-1066-1 (2016).
54. Carruthers, B. M. *et al.* Myalgic encephalomyelitis: International consensus criteria. *J. Intern. Med.* **270**, 327–338. https://doi.org/10.1111/j.1365-2796.2011.02428.x (2011).
55. Lopez-Rincon, A. *et al.* Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification. *Cancers* **12**, 1785 (2020).
56. Sarkar, C., Cooley, S. & Srivastava, J. Robust feature selection technique using rank aggregation. *Appl. Artif. Intell.* **28**, 243–257. https://doi.org/10.1080/08839514.2014.883903 (2014).

Acknowledgements

The funding for the study was provided by the division of Pharmacology, Department of Pharmaceutical Sciences, Faculty of Science, Utrecht University and Tytgat Institute for Liver and Intestinal Research, AGEM, Amsterdam UMC. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Author contributions

A.D.K., A.S., A.A.V., A.L.R. suggested the problem, A.L.R. prepared the datasets, performed the analyses and coded. P.I.M. wrote the manuscript and contributed the biological interpretation. L.M.M. contributed the biological interpretation, A.L.Y. and P.H. on epigenetics, I.A. on genetics, and A.A.V., A.S. and A.D.K. on all fields. A.D.K., A.S., A.A.V., and J.B. provided insights on ME/CFS. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83660-9.

Correspondence and requests for materials should be addressed to A.L.-R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.