Identifying Genetic Risk Factors for ME/CFS and Long COVID:
First Genetic Associations, Novel Targets, Actively Protective Biology, Diagnostics and Repurposing Opportunities

Steve Gardner, Sayoni Das, Krystyna Taylor, Jason Sardell, Matt Pearson, Karolina Chocian
PrecisionLife, Oxford

Introduction

PrecisionLife has developed a unique combinatorial approach to analyzing large scale genomic and other patient data. This captures the non-linear effects of interactions between multiple genes and exogenous (e.g., clinical, epidemiological, transcriptomic) factors. The inclusion of non-linear interactions enables far deeper insights than Genome-Wide Association Studies (GWAS).

We find more significant SNPs in complex disease patient data than GWAS, explain more disease variance, evaluate their causality, and our results translate between populations with different ancestries better than GWAS or polygenic risk scores.

Our mechanistic patient stratification approach correlates disease risk signatures with specific subgroups of patients who share similar disease drivers and treatment responses, to make precision medicine possible in multiple chronic diseases. We have applied our approach to 50 complex chronic diseases working with key opinion leaders, disease charities and patients to find better treatment options for patients with unmet medical needs, including ME/CFS and long COVID.

ME/CFS

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a massively debilitating chronic disease that presents with diverse symptoms including post-exertional malaise, chronic pain, and cognitive impairment. It affects 0.2-0.4% of the population, with an estimated 25% being housebound or bedbound. There are currently no diagnostic tests or approved disease modifying therapies for ME/CFS, and patients struggle, often for decades, to find ways of managing their symptoms.

There has never been any reproducible significant genetic associations identified for ME/CFS and only 1 gene identified for long COVID prior to these studies. This demonstrates the importance of non-linear signal in complex disease biology.

Methods

DATASET:

We analyzed genotype data from 2,382 patients reporting an ME/CFS diagnosis in the UK Biobank Pain Questionnaire matched against 4,764 controls in a case:control study design in the PrecisionLife platform.

Over 90% were of European genetic ancestry so this set were selected as the case cohort for this study. To ensure properly characterized control subjects, individuals were selected who had no evidence of diagnoses of chronic fatigue, post-exertional malaise, post-viral fatigue syndrome or myalgia. To avoid potential confounding, controls meeting these criteria were matched by genetic ancestry and gender against the cases in a 2:1 ratio (and this was repeated with a separate 4:1 ratio study).

COMBINATORIAL ANALYSIS:

The dataset was analyzed in the PrecisionLife platform to identify combinations of SNP genotypes that when observed together in a patient are strongly associated with ME/CFS.

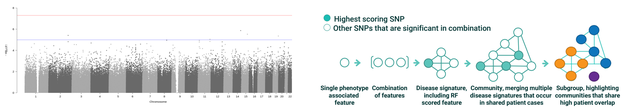
SNP combinations that have high odds ratios, low *p* values and high prevalence in cases are prioritized. This process undergoes 1,000 cycles of fully randomized permutations and combinations must meet a specified FDR threshold.

SNPs are scored using a Random Forest algorithm in a 5-fold cross validation framework and prioritized based on their ability to differentiate cases and controls.

The highest scoring SNPs are then mapped to genes and clustered by the patients they co occur in to generate a disease architecture.

Each ME/CFS dataset analysis took around 7 days (168) hours to complete, running on a server with 64 CPU cores and 4 × Nvidia GPUs.

Figure 1. Combinatorial Analysis of Genomic Data vs GWAS



Left. Manhattan plot of genome-wide *p*-values of association for the ME/CFS UK Biobank cohort, showing no significant SNPs (below $p<5 \times 10^{-8}$)
Right. Conceptual representation of features, combinations, disease signatures and communities used to build up the PrecisionLife disease architecture

GWAS	Combinatorial Analysis
Single SNP associations must be significant across large groups of patients, and they must be accurately diagnosed	Specific co-associated combinations of variants serve as a genetic stratification biomarker for each patient subgroup
Limited insights unless disease is caused by a small number of rare variants with large effect sizes (often in coding regions affecting protein 3D structure)	Patient subgroups with different causes of disease or even incorrect diagnoses can be distinguished (stratified) by different mechanistic etiology
Does not account for the effects of interactions between SNPs, genes and metabolic networks	Captures epistatic and non-linear additive effects of all interactions between SNPs, genes, environmental factors and metabolic networks

Figure 2. Combinatorial Analysis of Disjoint UK Biobank ME/CFS Cohorts

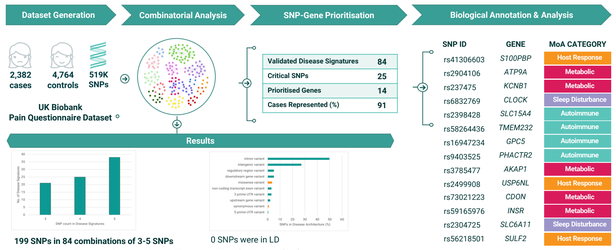
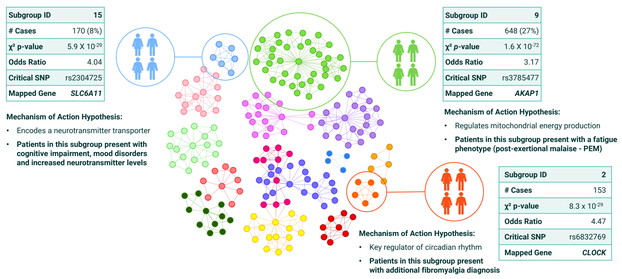


Figure 3. Mechanistic Patient Stratification and Validation using Patient Phenotypes



Retrospectively analyzing the patient stratification achieved by the hypothesis-free combinatorial analysis reveals that the patients' clinical presentation with different symptoms (fatigue/post-exertional malaise, cognitive impairment, mood disorders, sleep disturbance, autoimmunity etc) closely matches both the stratification and prevalences observed in the clinic.

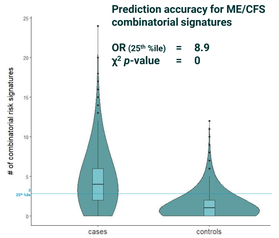
Figure 4. Predictive Accuracy of ME/CFS Disease Risk Signatures

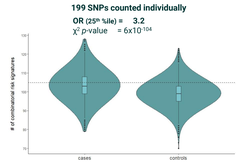
84 ME/CFS risk associated SNP combinations (199 unique SNPs) were found in the ME/CFS analysis (with good replication between the two disjoint ME/CFS cohorts studied).

By simply counting the number of risk signatures found in ME/CFS patients and controls, the distributions shown right were seen. Taking a 25th percentile cutoff, the odds ratio for this distribution is 8.9.

This is a massively unexpected predictive score, that exceeds established monogenic risk factors in other diseases, such as BRCA mutations predicting breast cancer, and is reproducible across different ME/CFS cohorts and in other complex diseases such as long COVID, endometriosis and motor neuron disease/ALS.

The additional predictive power of the combinatorial approach is shown below when evaluating the predictive power of the same 199 SNPs individually – an odds ratio of just 3.2 is achieved.





Replicates in other disjoint populations and gives similar results in other complex diseases

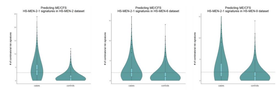
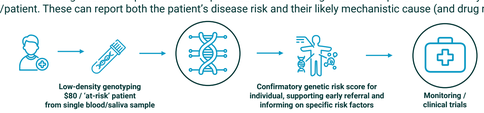


Figure 5. Novel Diagnostic Tools – Evaluating Risk and Differential Genotypic Triage

The results offer the opportunity to create a new kind of rapid, cheap diagnostic tool for complex chronic diseases. A test to count the number of risk signatures in a patient's makeup can be run from a single saliva sample on a low-density genotyping array @ $80/patient. These can report both the patient's disease risk and their likely mechanistic cause (and drug response).



This can also provide a differential triage tool to suggest the likely cause of easily confused common symptoms.

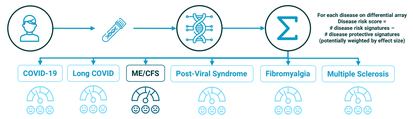
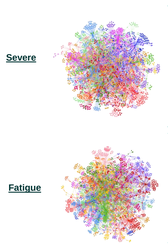


Figure 6. Overlap with Long Covid and other Fatigue Related Syndromes

73 unique genes identified in long COVID, 6 of which overlap with ME/CFS
(CLOCK, INSR, USP6NL, SLC15A4, GPC5, ATP9A)

13 SNPs in ME/CFS genes identified in Severe and/or Fatigue associated long COVID cohorts



43 genes identified: 30% found in other long COVID study results (e.g. transcriptomics)
Notable genes:
- ***TLR4***: long-term cognitive dysfunction after COVID-19
- ***SOX5***: T cell memory population associated with autoantibodies persisting post COVID-19 infection
- ***ADIPOQ***: adipokine stimulating AMPK phosphorylation, protective against SARS-CoV-2 induced lung injury
- ***ETS1***: downregulation may protect against SARS-CoV-2 cardiovascular insult

Overall themes: metabolic stress (insulin), COVID-19 susceptibility, inflammatory pathways

35 genes identified: 5 were significant from prior ME/CFS study, including *INSR* and *CLOCK*
Notable genes:
- ***NLGN1***: mediated by CLOCK, role in inflammatory pain, cognitive functioning and psychological stress vulnerability
- ***TNS1***: downstream of AMPK in maintaining energy balance of cell
- ***CNTN4***: deficient mice demonstrate increased stress responsiveness and maladaptive anxiety responses

Overall themes: metabolic stress (insulin, lipids), CNS-related pathways (stress, anxiety)

Figure 7. Repurposing Existing Drugs for Specific ME/CFS Cohorts based on PSBs

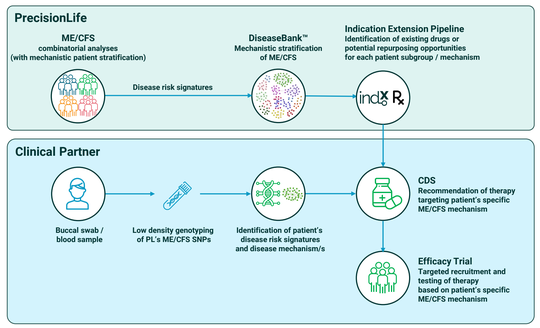
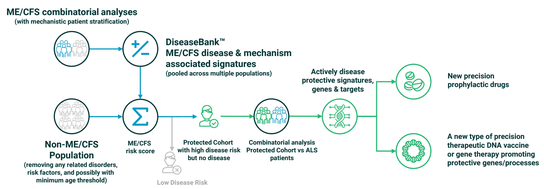
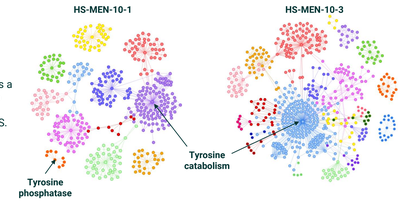


Figure 8. Actively Protective Biology working to Protect Against Disease



Results of analysis run to find actively protective genes in two disjoint ME/CFS populations identified the same gene involved in tyrosine catabolism. Tyrosine is a precursor for dopamine which is known to be associated with depression and/or mania, which can be symptoms of ME/CFS.



Discussion

The hypothesis-free combinatorial analytics approach implemented in the PrecisionLife platform identified 14 novel genetic associations with ME/CFS in a UK Biobank cohort[1], and 73 novel genes in the Sano GOLD long COVID dataset[2].

Several previous attempts at GWAS approaches have failed to validate a single SNP association or highlight significant risk genes in this ME/CFS cohort and the 50 group international consortium Covid-19 HGI has only reported one genetic association (*FOXP4*) from their meta-GWAS study on a population over 6x the size available to us.

This study has produced further evidence of the polygenic and heterogeneous nature of the disease and produced detailed patient stratification insights that describe the mechanistic etiology of the disease. This builds on our previous studies identifying genes underpinning all of the major symptoms in severe COVID-19 infection[3,4].

The study suggested a set of novel potential drug targets that may be relevant for the major ME/CFS (and long COVID) patient subgroups. These results have been augmented and validated by three additional studies.

The first uses PrecisionLife's INDx platform[5] to identify where a drug target newly associated with ME/CFS may have been investigated previously. If drug candidates modulate the target in the appropriate direction, they may well be useful for repurposing into patients. This is the fastest way to identify novel therapeutic options for patients, and for COVID-19 resulted in PrecisionLife identifying dutasteride, which has been shown to reduce infection severity and length by over 40% in double-blind RCTs.

Using the mechanistic patient stratification biomarkers (PSBs) identified by the study enables the selection of *n* of 50 patient cohorts with the specific mechanistic defects targeted by a given drug. These are ideal for small and fast efficacy trials of repurposing candidates.

The second and most unexpected result was demonstration of the predictive accuracy of the mechanistic biomarkers based on the disease risk (and protective) signatures. As well as demonstrating reproducible prediction accuracies better than established monogenic risk factors such as BRCA1/2, these tests also capture mechanistic information on the specific etiology of the disease within a patient, and the drug they are most likely to respond to.

The third approach is to use this high prediction accuracy of the risk signatures to enable the detection of a cohort of people who have all of the risk signatures that should lead them to have the disease, but who have not been affected by it. This cohort (selected for the longest/maximum exposure to disease risk factors) can be inferred to be protected from developing the disease by some other factors.

Comparing this cohort to ME/CFS patients allows us to remove the disease specific risk signatures and leave behind just those that appear to be protecting against the development of the disease even in patients otherwise at very high risk. These targets are likely to be actively protective biology that works to mitigate the effect of pathophysiological processes.

In the ME/CFS study we have identified a series of actively protective genes, all of which are associated with highly credible protective processes modulating known pathophysiology. As well as neurotransmitter metabolism, we have identified cell signaling pathways and genes known to be associated with specific endotypes.

References

1. Das, S., Taylor, K., Kozubek, J., Sardell J., Gardner, S. Genetic risk factors for ME/CFS identified using combinatorial analysis. *J Transl Med* **20**, 598 (Dec 2022) https://doi.org/10.1186/s12967-022-03815-8

2. Taylor, K., Pearson, M., Das, S., Sardell, J., Chocian, K., Gardner, S. Genetic risk factors for severe and fatigue dominant long COVID and commonalities with ME/CFS identified by combinatorial analysis. *J Transl Med* **21**, 775 (Nov 2023). https://doi.org/10.1186/s12967-023-04588-4

3. Das S, Pearson M, Taylor K, Bouchet V, Møller GL, Hall TO, Strivens M, Tzeng KTH, Gardner S. Combinatorial Analysis of Phenotypic and Clinical Risk Factors Associated With Hospitalized COVID-19 Patients. *Front. Digit. Health*, 2021 **3**:71-75 https://doi.org/10.3389/fdgth.2021.660809

4. Taylor, K., Das, S., Pearson, M., Kozubek, J., Pawlowski, M., Jensen, CE., Skowron, Z., Møller, GL, Strivens, M., Gardner, S. (Jun 2020). Analysis of genetic host response risk factors in severe COVID-19 patients Preprint at medRxiv. https://doi.org/10.1101/2020.06.17.20134015

5. Das S, Taylor K, Beaulah S, Gardner S. Systematic indication extension for drugs using patient stratification insights generated by combinatorial analytics. *Patterns* (N Y). 2022 Jun **10**;3(6):100496. https://10.1016/j.patter.2022.100496.

Acknowledgements

Research described in this study has been conducted using data from the UK Biobank Resource (application number 44288) and Sano Genetics. Input to the analyses was provided by Prof. Chris Ponting of U. Edinburgh and Sonya Chowdhury of Action for ME with whom we are collaborating on the Innovate UK funded LOCOME project using DecodeME data, by Solve ME, and by the Metrodora Institute. Thanks to the whole PrecisionLife team for their input and support.

precisionlife
improving health | for everyone