Supplementary Information

Title: A plasma protein biomarker signature that differentiates acute rheumatic fever from related clinical presentations

Authors: Emmy Okello[1,2,17], Timothy C. Barnett[3,4,17], Casey P. Shannon[5,6,7,17], Jenifer Atala[1], Ryan R. Brinkman[8], David I. Broadhurst[9], Guillaume Drouart[3], Christine Everest[3], Nina Kresoje[10], Amy H.Y. Lee[11], Wenna Lee[3], Peter Lwabi[1], Sebastiano Montante[8], David J. Martino[10], Emma Ndagire[1], Linda M. Oyella[1], Rym Ben-Othman[12], Jafesi Pulle[1], Craig A. Sable[13], Rachel Sarnacki[13], Michael Serralha[12], Scott J. Tebbutt[5,6,14], Andrea Beaton[15,16,18], Tobias R. Kollmann[12,18], Jonathan R. Carapetis[3,18]

Affiliations:

1. The Uganda Heart Institute, Mulago Hospital Complex, Kampala, Uganda
2. Department of Medicine, Makerere University, Kampala, Uganda
3. Wesfarmers Centre of Vaccines and Infectious Diseases, The Kids Research Institute Australia, University of Western Australia, Perth, WA, Australia
4. The Marshall Center for Infectious Diseases Research and Training, School of Biomedical Sciences, University of Western Australia, Perth, Australia
5. Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada.
6. Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada
7. Providence Research, Providence Health Care Research Institute, Vancouver, BC, Canada
8. BC Cancer Agency, Vancouver, BC, Canada
9. Centre for Integrative Metabolomics & Computational Biology, Edith Cowan University, Joondalup, WA, Australia
10. Wal-yan Respiratory Research Centre, The Kids Research Institute Australia, University of Western Australia, Perth, WA, Australia
11. Molecular Biology and Biochemistry, Simon Fraser University, British Columbia, Canada
12. The Kids Research Institute Australia, University of Western Australia, Perth, WA, Australia
13. Children's National Hospital, Washington, DC, USA
14. Division of Respiratory Medicine, Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada
15. Cincinnati Children's Hospital Medical Center, OH, USA
16. Department of Pediatrics, The University of Cincinnati School of Medicine, OH
17. These authors contributed equally: Emmy Okello, Timothy C. Barnett, Casey P. Shannon
18. These authors jointly supervised this work: Andrea Beaton, Tobias R. Kollmann, Jonathan R. Carapetis

This PDF file includes:

Supplementary Notes 1 and 2
Supplementary Methods
Figures S1 to S12
Supplementary references

Supplementary Note 1.

Multi-omic profiling reveals dysregulation of multiple peripheral white blood cell components in ARF.

Transcriptomics: We first compared gene expression from individuals with either known (n=9) or unknown (n=13) alternate diagnosis against those with definite ARF (n=20). Comparing known alternate diagnosis against those with definitive ARFs identified only 6 differentially expressed (DE) genes while unknown alternate diagnosis compared to definitive ARFs yielded 517 DE genes (adjusted p values <= 0.05; absolute fold changes >= 1.5; **Source Data**). Functional enrichment of these 517 DE genes using ReactomePA showed enrichment in eukaryotic transcriptional and translation processes **(Fig. S8a)**. Comparison of individuals with RHD to those with definite ARF at intake identified 863 DE genes; while comparison to healthy individuals yielded the biggest transcriptional differences (1074 DE genes). Of note, RHD individuals displayed additional pathways indicative of alteration in the purine pathways[44], as well as dysregulation of various cell senescence pathways including "Oxidative Stress Induced Senescence", "Diseases of programmed cell death", "Cellular response to starvation" and "Defective pyroptosis". Contrasting DE genes between Healthy Controls and Definite ARF indicates acute infectious response **(Fig. S8b)**. Pathway enrichment of transcriptional differences contrasting unknown alternative diagnosis (Unknown Alt), RHD or healthy individuals to Definite ARF individuals identified overlapping pathways related to transcription and translation.

Epigenetics: In our primary comparison, epigenetic changes in peripheral blood WBC between definite ARF cases (n=108) and participants with alternate diagnoses (n=39) could be detected but failed to provide sufficient discriminatory classification. In secondary analysis, there was evidence for changes in PBMC methylation profiles between healthy controls and participants with ARF and alternate diagnosis at intake. Both gain and loss of methylation was observed at different regions. Most regions were annotated to gene coding regions, and a small number of methylation changes were observed in non-coding regions. To determine the epigenetic contribution to ARF, logistic regression of autosomal loci was carried comparing Definite ARF v Healthy controls, and also Definite ARF v Alternate Diagnosis. Only one CpG locus reached genome-wide significance when comparing ARF to Healthy Controls (Adj P < 0.05) (**Fig. S9a**). Individual CpGs were binned into 1kb regions and we compared the smoothed average over regions in a differential region test analysis. A total of 14 differentially methylated regions (DMRs) comprising at least 4 CpGs were identified between RF cases and Healthy Controls (**Fig. S9a**). When comparing ARF cases against those labelled 'alternate diagnosis' only one individual CpG site passed the threshold for genome-wide significance (adj. P < 0.05,) (**Fig. S9b**). We identified 2 DMRs when comparing Definite ARF vs. Alternate Diagnosis groups that passed genome wide significance, and these were also identified in the previous analysis (**Fig. S9b**)

Flow Cytometry: There were no significant differences in cell composition using predefined canonical anchor markers between clinical diagnoses at intake or throughout the course of the disease **(Fig. S10)**. This indicated that any differences in peripheral white blood cell-based signals correlating with clinical diagnosis would unlikely be due to differences in cell composition, but more likely inherent to the cellular function of these WBC.

Supplementary Note 2

Metabolomic profiling reveals dysregulation of multiple peripheral white blood cell components in ARF.

Statistical analysis took the form of a covariate-adjusted generalized linear model (GLM - with identity link function & normal distribution), adjusting for age & sex, applied to each of the 768 metabolites in turn. Mulago data were used for biomarker discovery and, where possible, Mbarara data was used to validate potential biomarkers. Before modelling, the metabolite data was $\log_{10}$ transformed as is standard practice[45]. Two statistical comparisons were performed. Firstly, individuals with definite ARF diagnosis vs. Healthy Controls or RHD (without RF), and then definite ARF vs. known or unknown alternate diagnosis. **Fig. S11a** shows an upset plot describing the significant ($p < 0.05$) contribution of each model factor across the 768 models generated when comparing definite ARF diagnosis vs. Healthy Controls or RHD (without RF), for the Mulago samples. 244 metabolites were significantly associated with definite ARF. This number dropped to 105 when adjustment for False Discovery Rate was performed. **Fig. S11c** shows a volcano plot of the 768 adjusted p-values (q-values) vs. fold-change for the definite ARF diagnosis against Healthy Controls comparison. Metabolite q-values are provided in the Source Data file. As there were no Mbarara Healthy Control or RHD (without RF) samples, validation of these biomarkers was not possible. **Fig. S11b** shows an upset plot describing the significant ($p < 0.05$) contribution of each model factor across the 768 models generated when comparing definite ARF diagnosis vs. known or unknown alternate diagnosis. 53 metabolites were significantly associated with definite ARF. This number dropped to zero when adjustment for False Discovery Rate was performed. None of the 53 significant metabolites discovered in the Mulago data were found to be significant in the Mbarara samples, suggesting that these results should be disregarded.

Supplementary Methods

Flow Cytometry

Whole blood samples were processed for flow cytometry as described[46-48]. Flow cytometry data were acquired on an LSR Fortessa (BD Biosciences) and analysed manually using Flowjo software (version 9.9) following a pre-defined gating strategy (**Fig. S12**). In addition, immunophenotyping using unbiased automated gating was undertaken on the same samples using flowCut, flowDensity, and flowTypeFilter as well as biomarker visualisation (RchyOptimyx) as described[49].

Transcriptomics

Paxgene blood samples were used to manually extract total RNA using column based Qiagen whole bood extraction kits. Stranded libraries were prepared from high quality extracted RNA (RIN>7) using **Sureselect HS2** library preparation kit. The protocol includes poly A enrichment followed by fragmentation, reverse transcription, ligation with adaptors containing molecular identifiers followed by amplification for indexing. Library QC will be performed using Tapestation 4200 and Qubit, followed by QC sequencing on iSeq and deep sequencing on Illumina NovaSeq 6000. Sequencing was conducted to yield approximately 30 – 40 million raw paired reads (read1 + read2 = 1 paired read) per sample. PolyA RNAseq libraries from were then sequenced on three S2 flow cells at 2 X 50 cycles to yield approximately 40 million raw paired reads per sample. Demultiplexed FASTQ files were then analysed to understand baseline expression differences between definite ARF versus individuals with alternate diagnosis (alternate acute illnesses and inflammatory conditions), whole blood transcriptomics by RNAseq was conducted focused on intake samples. Whole blood was stored in Paxgene Blood RNA tubes (BD Biosciences), frozen and shipped to the Australian Genome Research Facility (AGRF; https://www.agrf.org.au/). Total RNA was extracted, followed by quantification and quality assessment of total RNA using an Agilent 2100 Bioanalyzer (Santa Clara, CA, USA). Library preparation of the polyA RNA was done using TruSeq mRNA Library Prep with polyA selection and unique dual indexing, followed by sequencing on the Illumina NovaSeq 6000 to generate paired-end sequences. Sequence quality was assessed using FastQC v0.12.1 and MultiQC v1.13[50]. The FASTQ sequence reads were aligned to the hg38 human genome (Ensembl GRCh38.98) using STAR v2.7[51] and mapped to Ensembl GRCh38 transcripts. Read-counts were generated using htseq-count (HTSeq) v2.0.2[52]. All data processing and subsequent differential gene expression analyses were performed using R version 4.3.1 and DESeq2 version 1.38.3[53]. Given our interest in differentiating Definite ARF from other acute febrile illnesses, we compared gene expression from individuals with either known or unknown alternate diagnosis against those with definite ARF.

Epigenomics

Genomic DNA was extracted from peripheral blood mononuclear cells using the Chemagic low volume blood extraction kit (Perkin Elmer, #CMG-1417). Samples were block randomised across plates so that each comparison group will be equally represented on each plate. Extracted DNA was assessed for quantity and quality on the Qubit fluorometer with gel analysis. DNA samples were submitted to the AGRF for sodium bisulphite treatment and hybridisation to Infinium HumanMethylation EPIC BeadChip. Raw intensity files were pre-processed using the Minfi package (v1.38)[54] from the bioconductor project (http://www.bioconductor.org) in the R statistical environment (http://cran.r-project.org/, version 4.1.2). Sample quality was assessed using control probes on the array. Between-array normalisation was performed using the stratified quantile method to correct for Type 1 and Type 2 probe bias. Probes exhibiting a *P*-detection call rate of >0.01 in 1 or more samples were removed prior to analysis. Probes containing SNPs at the single base extension site, or at the CpG assay site were removed, as were probes measuring non-CpG loci. Probes reported to have off-target effects[55,56] were also removed. After sample and probe filtering the final data set size was 275 samples and 757,904 probes. Methylation percentages were derived as *β* values with log 2 transformation to *M* values for statistical analysis[57]. Major blood cell proportions were deconvoluted from the methylation data set using the FlowSorted.Blood.450k package (v1.3). To perform differential methylation analysis, a bayesian logistic regression model (limma v3.48.3) was fit to autosomal M-values with ARF diagnosis as the main predictor, adjusted for sex, age, study site, blood cell counts and the first five principal components as a proxy for technical variation. Genome-wide significance was declared at a false discovery rate adjusted p-value of <= 0.05 to call differentially methylated regions. The regression model statistics were used as inputs to the DMRcate package[58] for *de novo* identification of differentially methylated regions using bandwidth smoothing window of 1kb, scaling factor of 2 and minimum 4 CpGs. Only DMRs with a minimum smoothed FDR p< 0.05 were reported.

Metabolomics

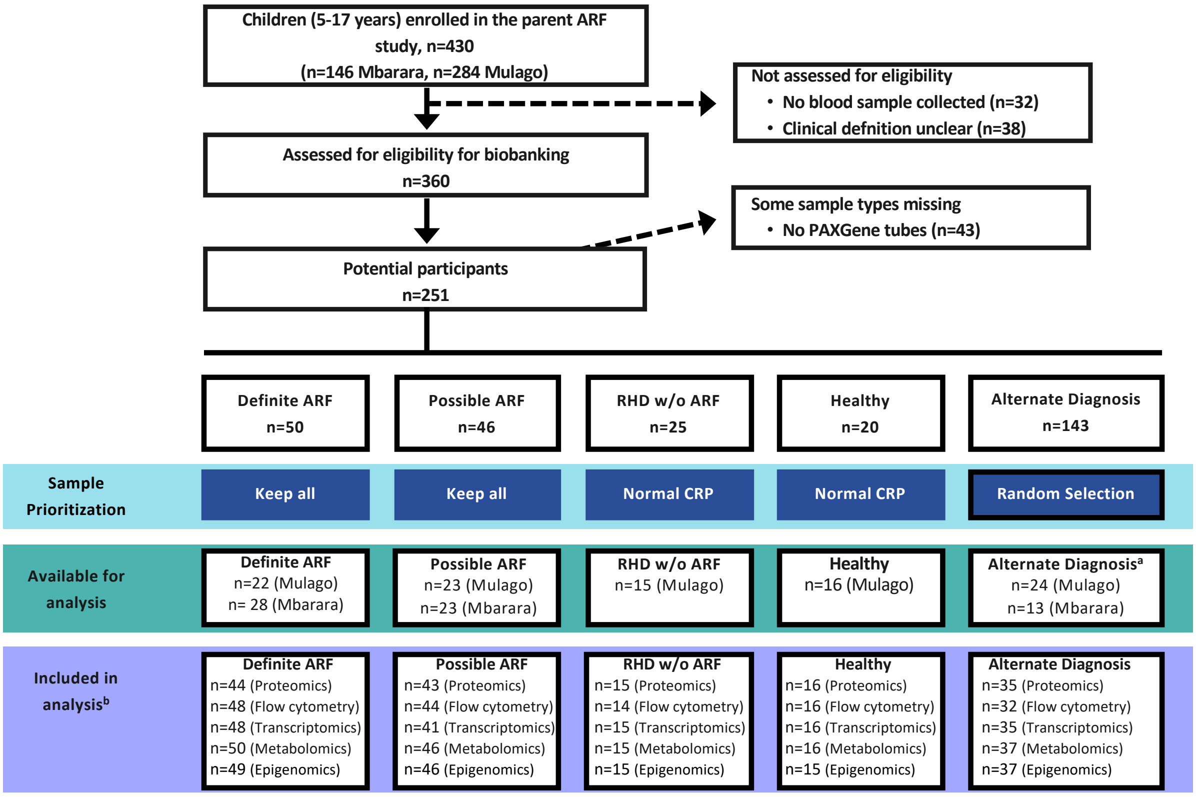
Untargeted metabolomic profiling was performed on plasma samples using liquid chromatography coupled to high-resolution mass-spectrometry (LC-MS). Data was acquired by Metabolon analytical services using three modes of operation: reverse-phase/ultraperformance liquid chromatography (UPLC)-MS/MS with positive ion mode electrospray ionisation (ESI), reverse-phase/UPLC-MS/MS with negative ion mode ESI, and Hydrophilic interaction (HILIC)/UPLC-MS/MS with negative ion mode ESI using their standard protocols[59]. All identified metabolites were annotated using appropriate orthogonal analytical techniques applied to the metabolite of interest against a chemical reference standard. 768 annotated endogenous metabolites were reproducibly detected.

Metabolomics is particularly sensitive to factors related to sample collection and biobanking[60]. Exploratory data analysis confirmed that there was a significant collection site confounder, and that age, sex were also potential confounders. As such, statistical analysis took the form of a covariate-adjusted generalized linear model (GLM - with identity link function & normal distribution, and adjusting for collection site, age & sex) applied to each of the 768 metabolites. Prior to statistical modelling, the metabolite data was $\log_{10}$ transformed as is standard practice[45]. Two statistical comparisons were performed. Firstly, individuals with definite ARF diagnosis vs. Healthy Controls or RHD (without RF), and then definitive ARF vs. known or unknown alternate diagnosis. Due to the site confounder, and uneven distribution of diagnostic classes between sites, samples from Mulago were used for discovery and where possible Mbarara for validation.

SUPPLEMENTARY FIGURES.

Supplementary Fig. S1.

STROBE DIAGRAM FOR PARTICIPANT SELECTION

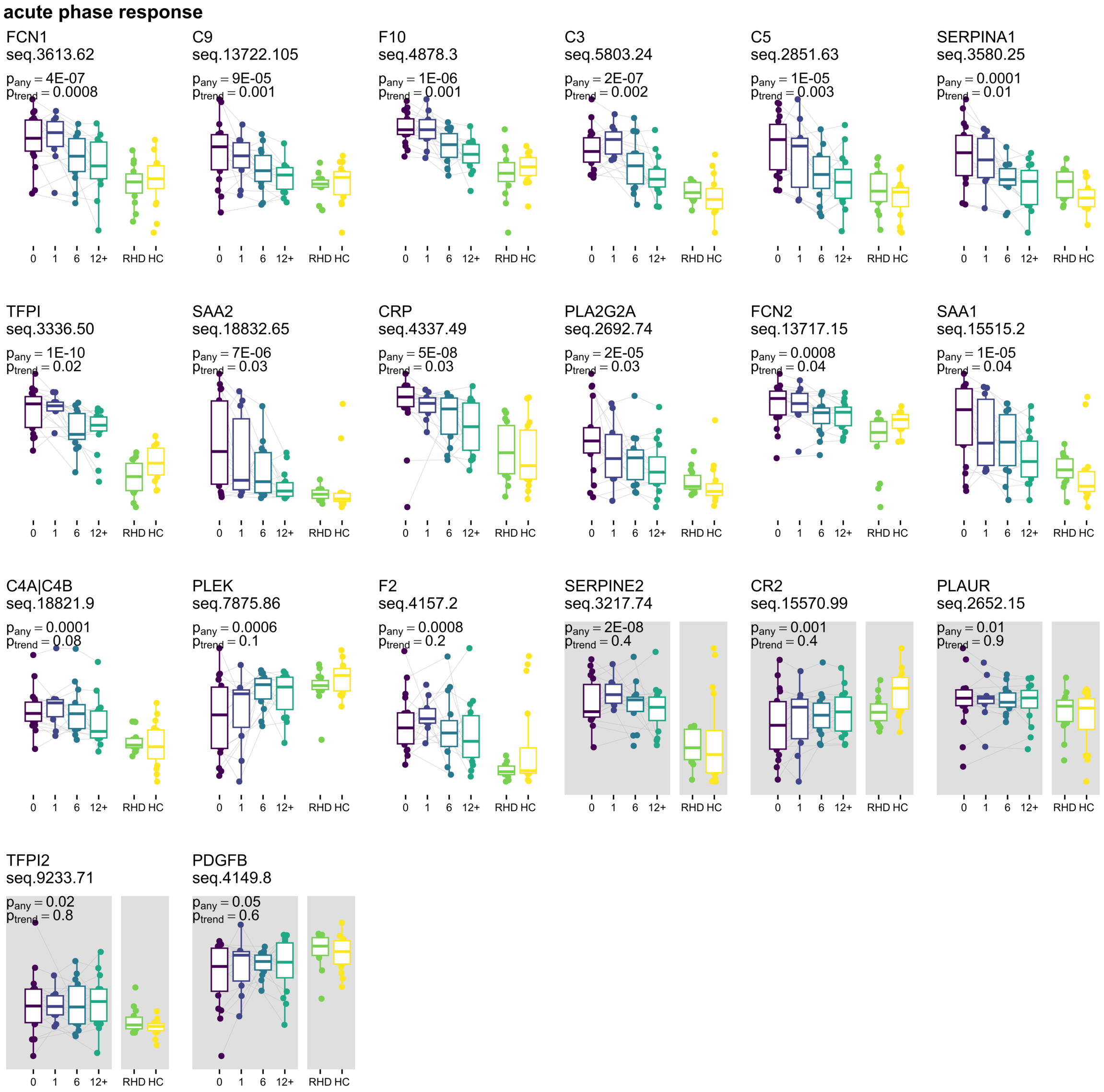


a. Includes: 28 unknown alternate diagnoses (n=15 Mulago, n=13 Mbarara) and 9 known alternate diagnoses (n=9 Mulago)

b. Individual participants included/excluded from individual analysis platforms is provided in the Supplementary Data file

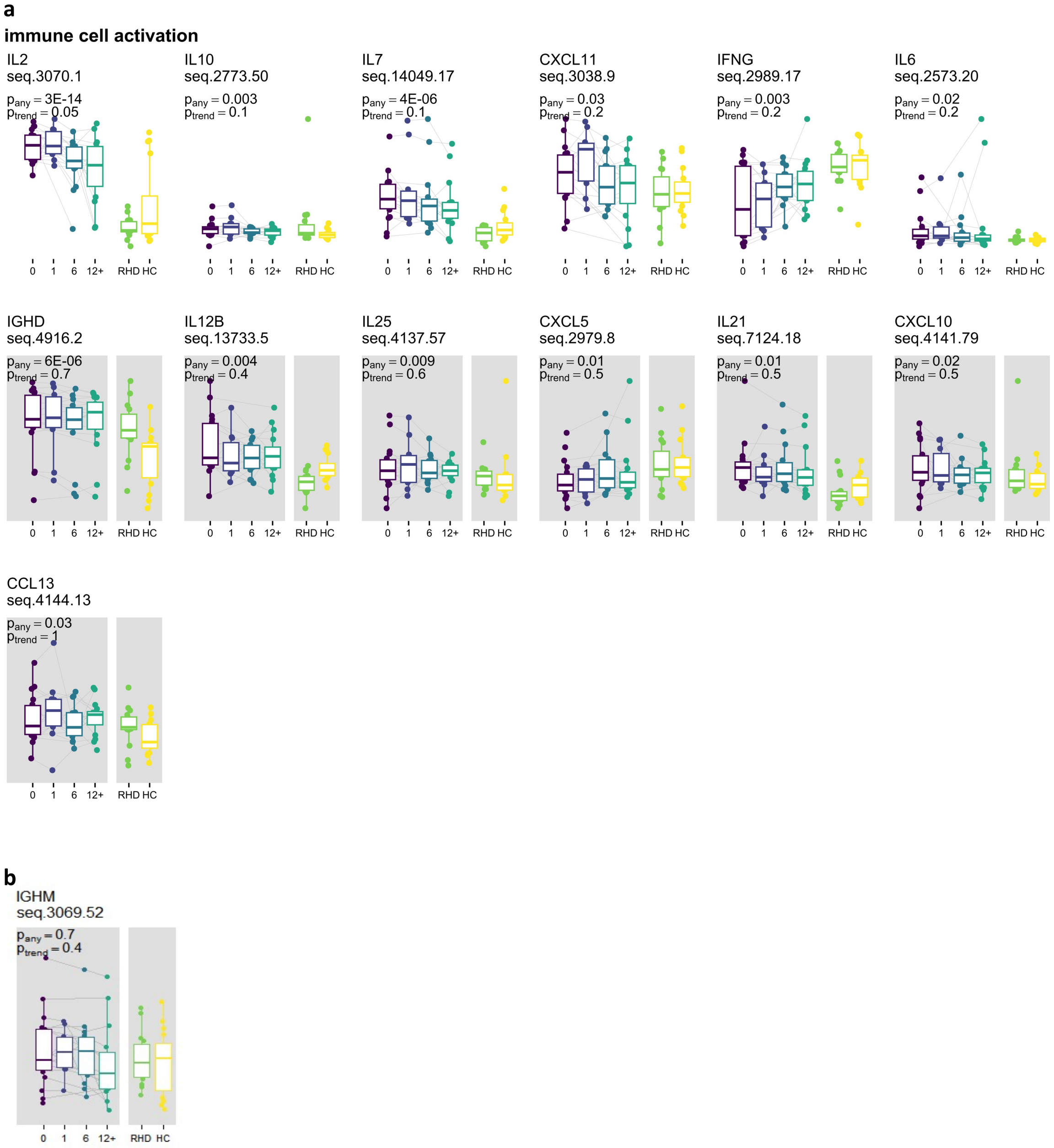
STROBE Diagram of Participant selection

Supplementary Fig. S2.



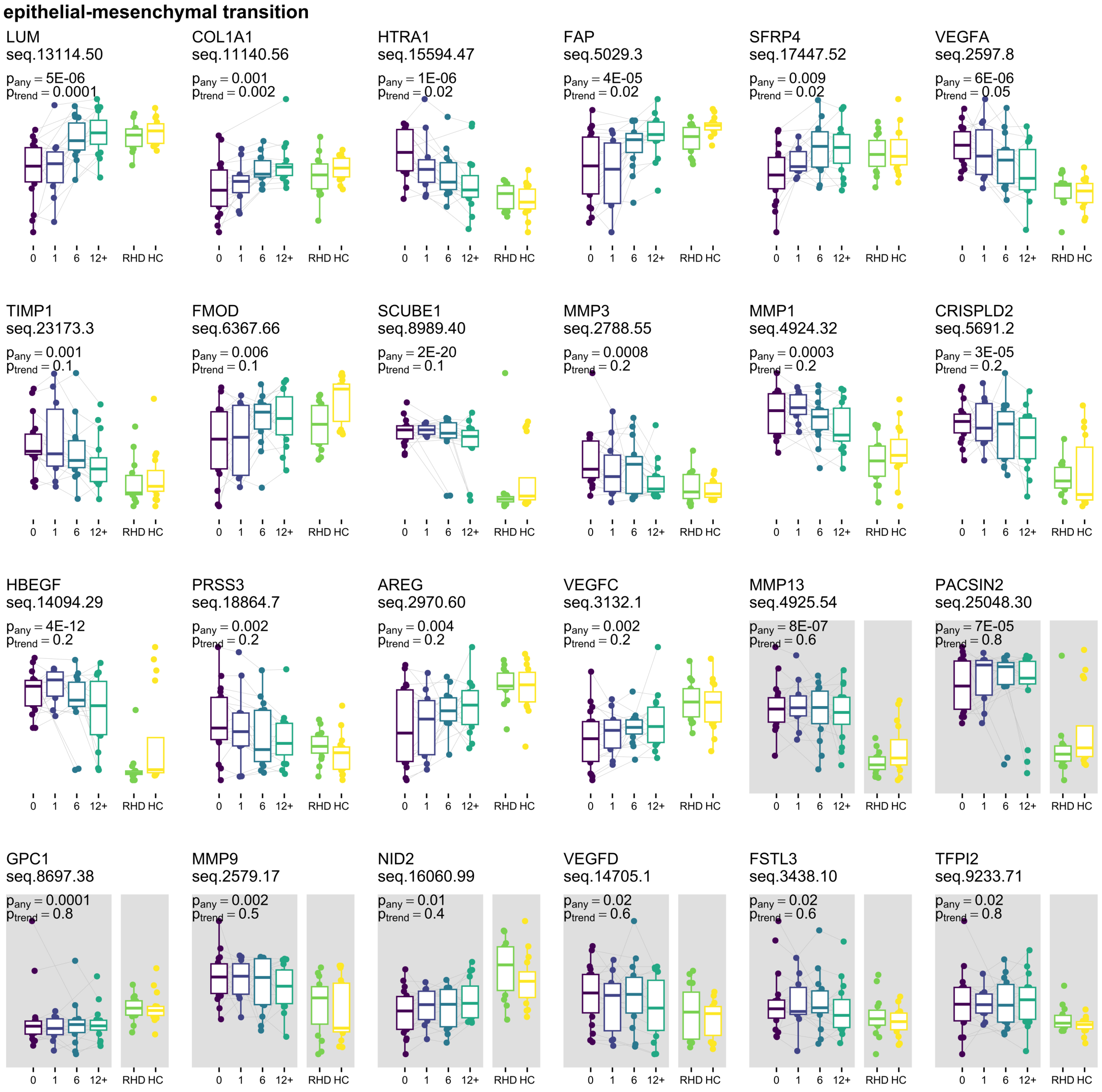
Temporal patterns of relative protein plasma concentration for major dysregulated proteins from the acute phase response and complement and coagulation pathways.

Supplementary Fig. S3.



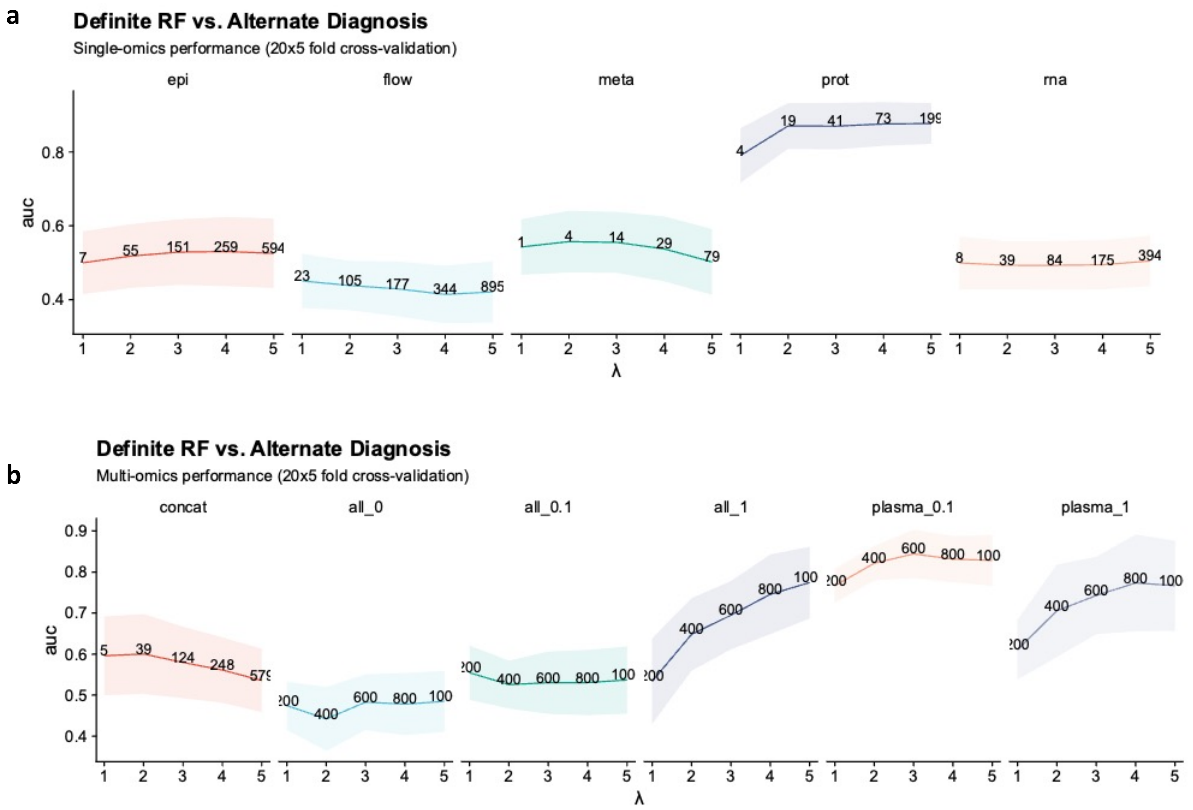
a, Temporal patterns of relative protein plasma concentration for major dysregulated proteins associated with immune cell activation. **b**, Temporal pattern of relative protein plasma concentration for IgM (IGHM).

Supplementary Fig. S4.



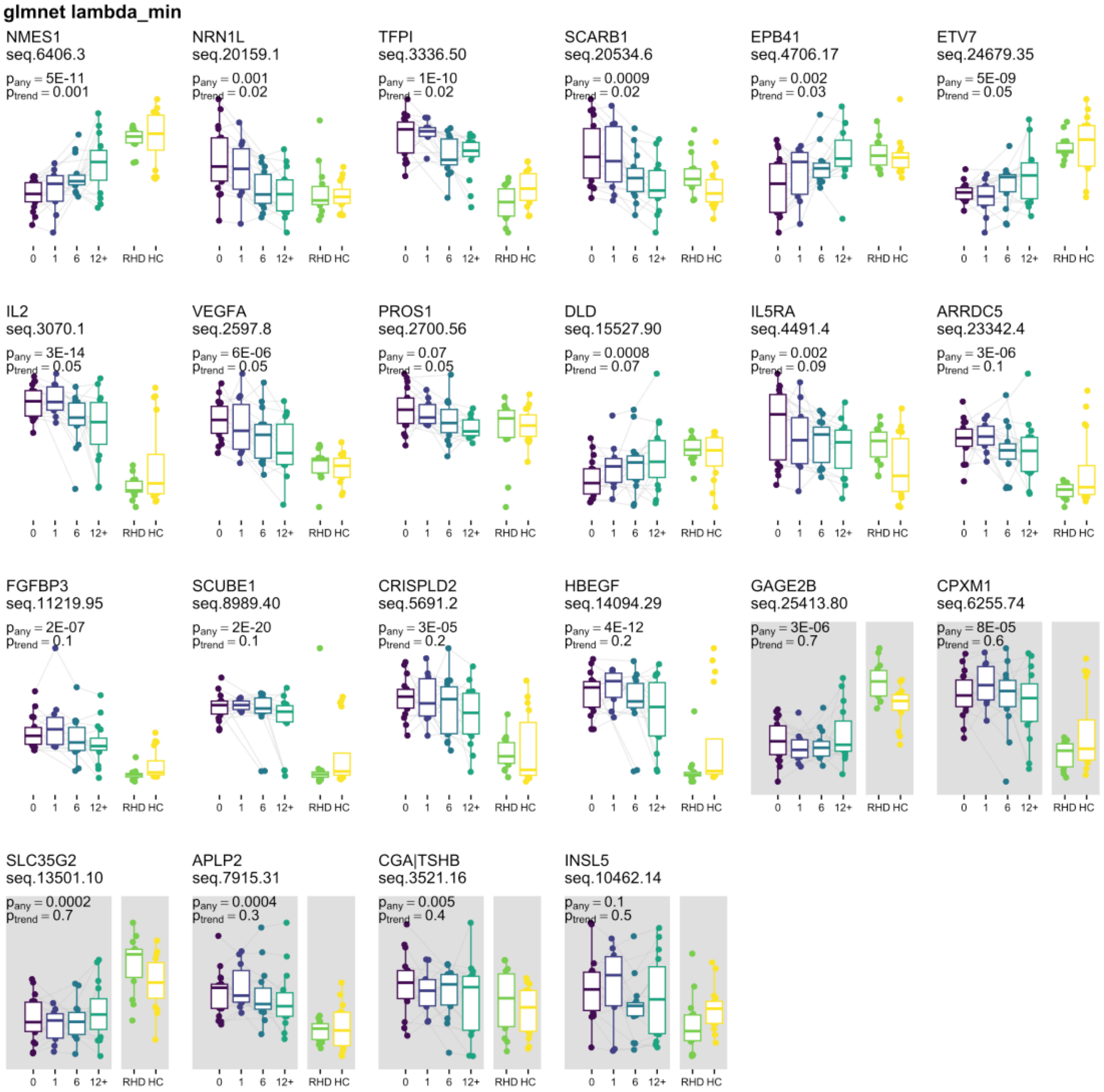
Temporal patterns of relative protein plasma concentration for major dysregulated proteins associated with epithelial-mesenchymal transition.

Supplementary Fig. S5.



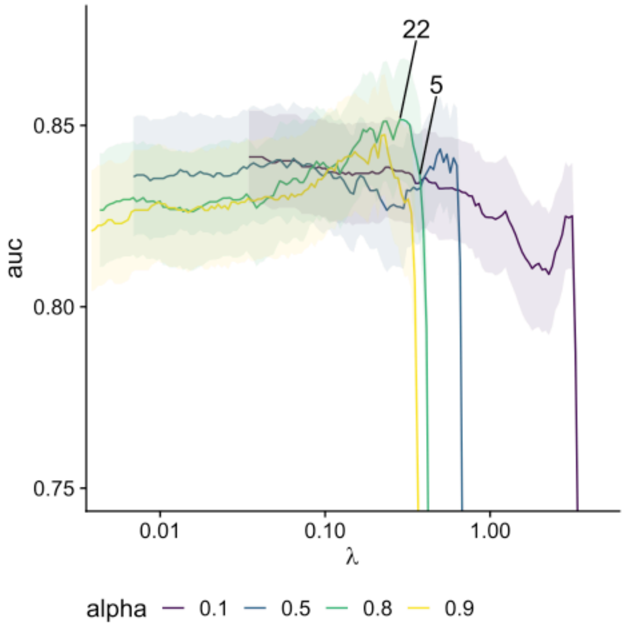
Comparison of diagnostic performance of definite ARF (Definite RF) and Alternate Diagnosis across epigenetics (epi), flow cytometry (flow), metabolomics (meta), proteomics (prot) and transcriptomics (rna). 20x5 fold cross validation. **a**, Single-omics performance. **b**, Multi-omics performance.

Supplementary Fig. S6.



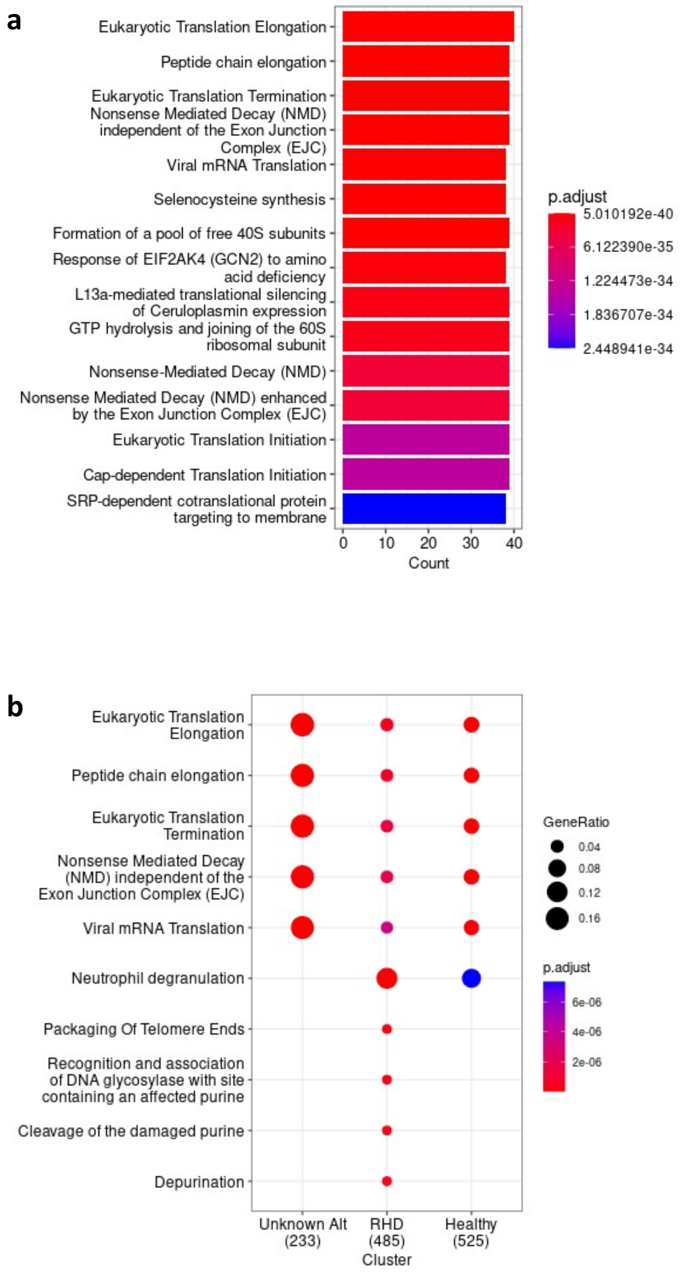
Temporal patterns of relative protein plasma concentration for each protein associated with 22-protein signature in Fig. 2a.

Supplementary Fig. S7.



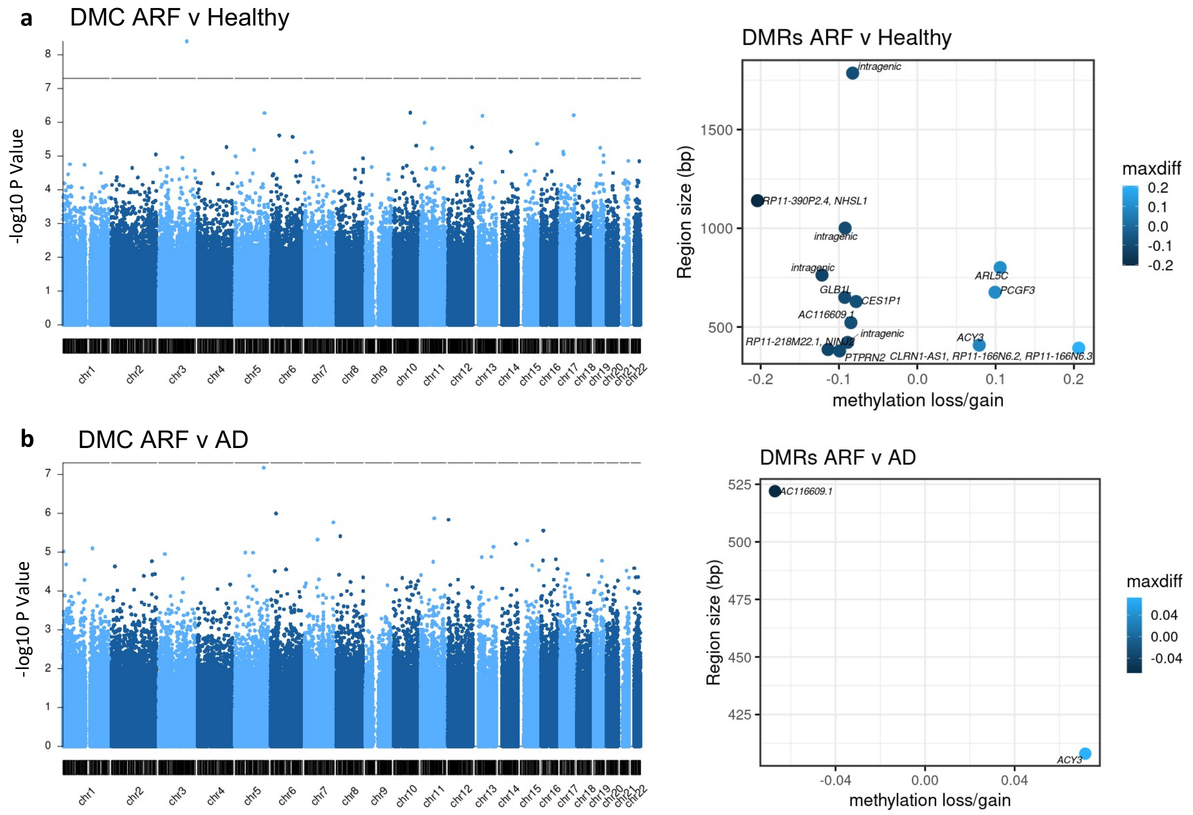
Exploring the elastic net parameter space. Cross-validation curve showing classification performance (area under the receiver-operator characteristic curve; AUC; y-axis), including upper and lower standard deviation (shaded area), as a function of the penalization parameter lambda (x-axis), and for different values of the elastic net mixing parameter alpha (colours; alpha=0: L2-penalization, ridge regression; alpha=1: L1-penalization, LASSO regression). Larger values of lambda result in smaller models (fewer features with non-zero coefficients; left-to-right). Numbered labels indicate number of proteins retained in the model at the corresponding point in the cross-validation curve). The smallest model that achieved within 1 SD of the highest performance achieved was selected for further evaluation in other available data.

Supplementary Fig. S8.



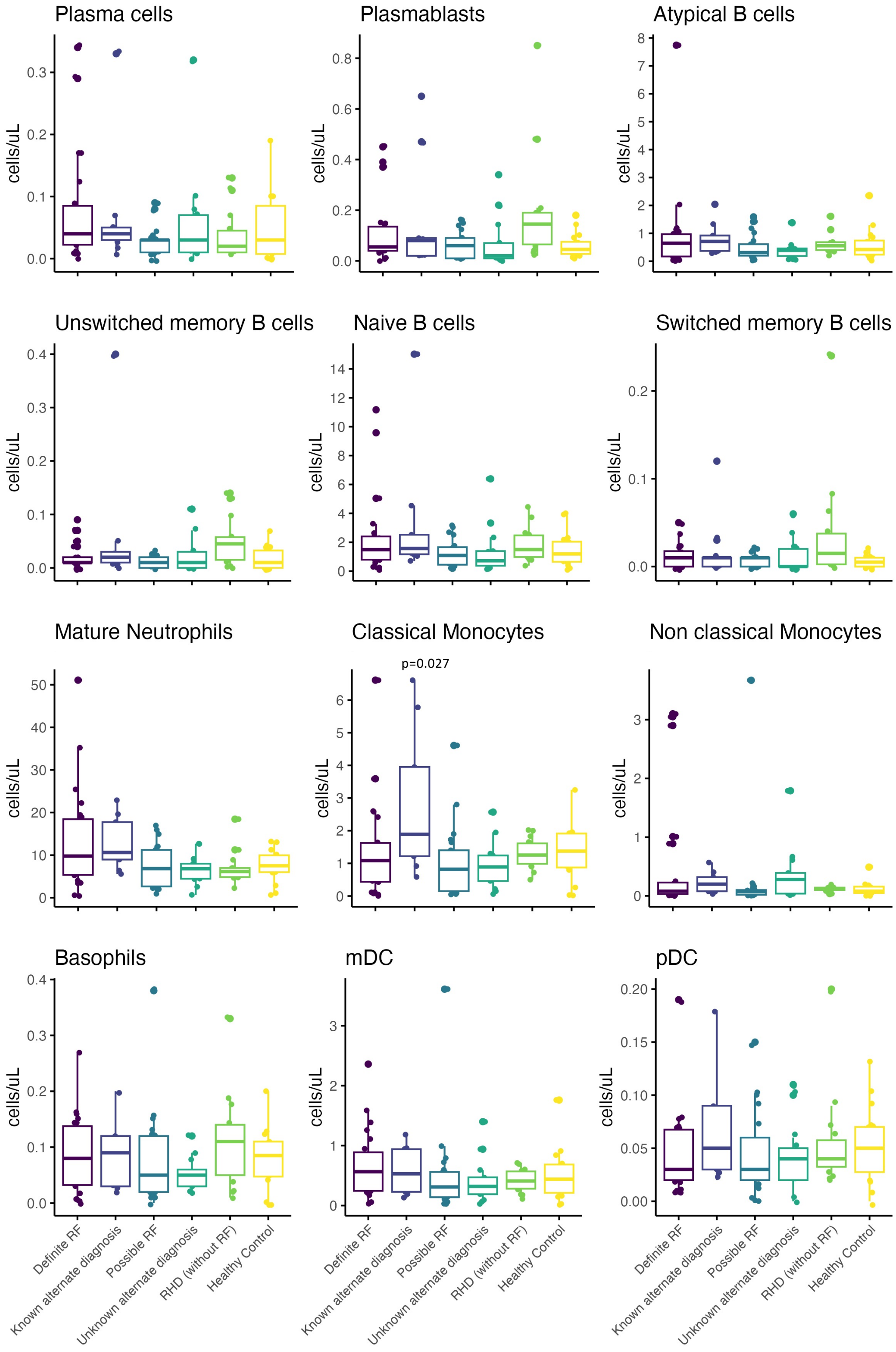
a, Pathway enrichment of differentially expressed genes comparing individuals with unknown alternative diagnosis to definite ARF. **b**, Comparison of enriched pathways from differentially expressed genes comparing individuals with unknown alternative diagnosis, RHD or healthy individuals to definite ARF.

Supplementary Fig. S9.



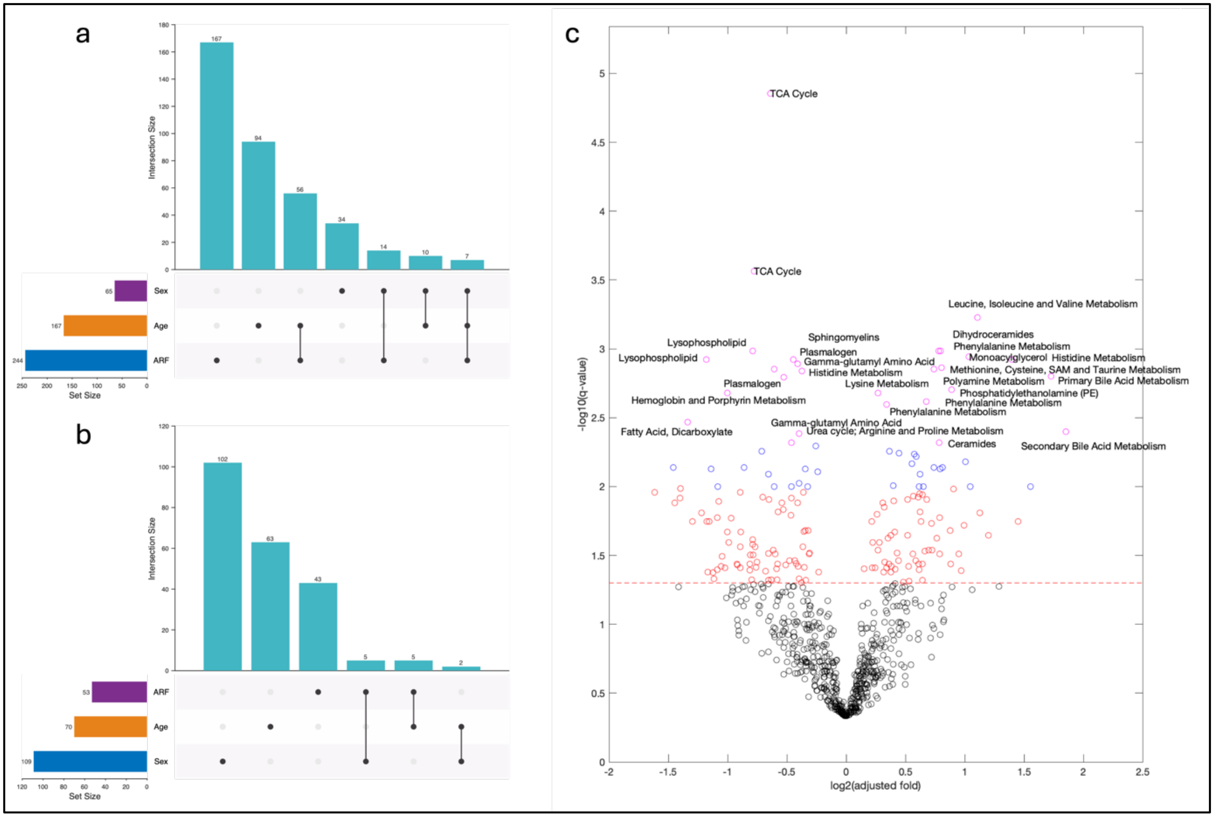
Summary statistics from epigenome-wide association analyses. **a**, Results from ARF v Alternate Diagnosis (AD) comparision. Manhattan plot (left) shows summary statistics for each CpG by chromosomal location. Genome-wide significance line is shown. Volcanoplot (right) shows DMRs. Points are coloured according to effect size. **b**, Results from ARF v AD comparison. *ARF= acute rheumatic fever. AD=Alternate diagnosis. Maxdiff is the maximum difference in methylation ratios (10-2) between cases and controls. DMR = differentially methylated region. DMC= differentially methylated CpG.*

Supplementary Fig. S10.



Distribution of immune cell populations across different groups determined by flow cytometry. With the exception of classical monocytes in the known alternate diagnosis group (p=0.027), there were no significant differences in populations between definite ARF and control groups.

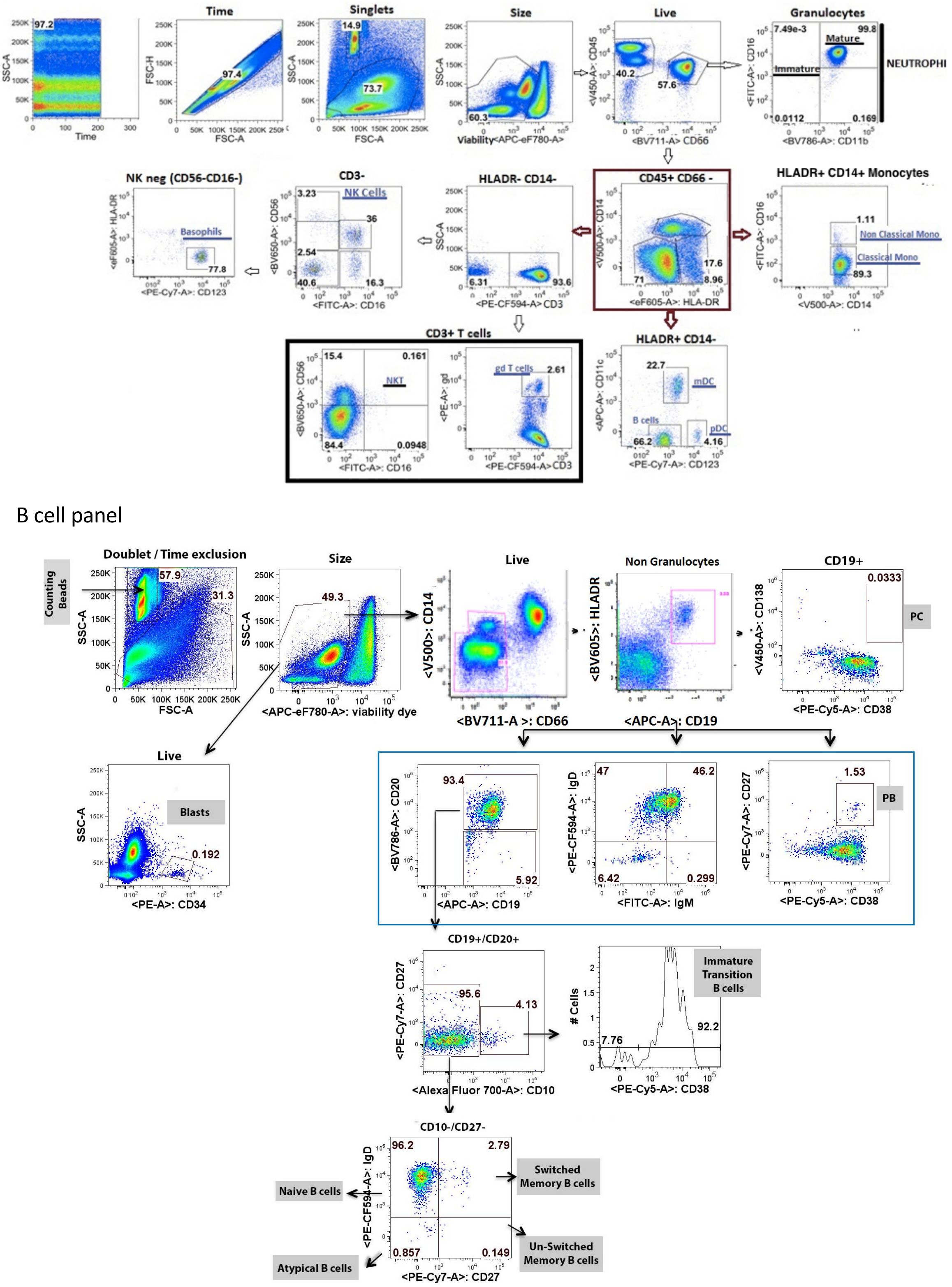
Supplementary Fig. S11.



Metabolomic profiling reveals dysregulation of multiple peripheral white blood cell components in ARF. **a**, UpSet plot describing the significant ($p < 0.05$) contribution of each model factor across the 768 covariate corrected regression models comparing definite ARF samples against Healthy Controls or RHD (without RF) in the Mulago data. **b**, UpSet plot describing the significant ($p < 0.05$) contribution of each model factor across the 768 covariate corrected regression models generated when comparing definite ARF diagnosis against either known or unknown alternate diagnosis in the Mulago data. **c**, Volcano plot of the 768 adjusted p-values (q-value) vs. fold-change for the definite ARF diagnosis against Healthy Controls or RHD (without RF) comparison, with individual metabolites labelled by metabolite pathway. A full list of metabolite names and associated statistics are provided in the Source Data file.

Supplementary Fig. S12.

Myeloid panel



B cell panel

Flow cytometry gating strategy for Myeloid (top) and B cell (bottom) panels.

Supplementary References

44 Das, S. *et al.* An Untargeted LC-MS based approach for identification of altered metabolites in blood plasma of rheumatic heart disease patients. *Sci Rep* **12**, 5238, doi:10.1038/s41598-022-09191-z (2022).

45 van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142, doi:10.1186/1471-2164-7-142 (2006).

46 Ben-Othman, R. *et al.* Systems Biology Methods Applied to Blood and Tissue for a Comprehensive Analysis of Immune Response to Hepatitis B Vaccine in Adults. *Front Immunol* **11**, 580373, doi:10.3389/fimmu.2020.580373 (2020).

47 Idoko, O. T. *et al.* Clinical Protocol for a Longitudinal Cohort Study Employing Systems Biology to Identify Markers of Vaccine Immunogenicity in Newborn Infants in The Gambia and Papua New Guinea. *Front Pediatr* **8**, 197, doi:10.3389/fped.2020.00197 (2020).

48 Iturriaga, C. *et al.* A cluster randomized trial of interferon ss-1a for the reduction of transmission of SARS-Cov-2: protocol for the Containing Coronavirus Disease 19 trial (ConCorD-19). *BMC Infect Dis* **21**, 814, doi:10.1186/s12879-021-06519-4 (2021).

49 Lee, A. H. *et al.* Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. *Nat Commun* **10**, 1092, doi:10.1038/s41467-019-08794-x (2019).

50 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

51 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

52 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

53 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

54 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369, doi:10.1093/bioinformatics/btu049 (2014).

55 McCartney, D. L. *et al.* Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genom Data* **9**, 22-24, doi:10.1016/j.gdata.2016.05.012 (2016).

56 Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208, doi:10.1186/s13059-016-1066-1 (2016).

57 Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587, doi:10.1186/1471-2105-11-587 (2010).

58 Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* **8**, 6, doi:10.1186/1756-8935-8-6 (2015).

59 Ford, L. *et al.* Precision of a Clinical Metabolomics Profiling Platform for Use in the Identification of Inborn Errors of Metabolism. *J Appl Lab Med* **5**, 342-356, doi:10.1093/jalm/jfz026 (2020).

60 Kirwan, J. A. *et al.* Preanalytical Processing and Biobanking Procedures of Biological Samples for Metabolomics Research: A White Paper, Community Perspective (for "Precision Medicine and Pharmacometabolomics Task Group"-The Metabolomics Society Initiative). *Clin Chem* **64**, 1158-1182, doi:10.1373/clinchem.2018.287045 (2018).