nature portfolio

Corresponding author(s): Jonathan Carapetis, Andrea Beaton, Tobias Kollmann

Last updated by author(s): Jan 13, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

All data collection software is either open source or commercially available and is listed below.
1. BD FACSDiva Software Version 8.0.1

Data analysis

Proteomics:
1. R v4.4.2 (2024-10-31 ucrt)
2. glmnet v4.1-8
3. SomaDataIO v6.0.0
4. Biobase v2.60.0
5. BiocGenerics v0.46.0
6. Tidyverse collection (v2.0.0), which encompassed dplyr (v1.1.4), tidyr (v1.3.1), ggplot2 (v3.4.4), readr (v2.1.5), forcats (v1.0.0), purrr (v1.0.2), stringr (v1.5.1), and tibble (v3.2.1).
7. lubridate v1.9.3
8. finalfit v1.0.7
9. mice v3.16.0
10. car v3.1-2

Flow Cytometry:
1. R v3.5.2
2. flowCut v1.16.0
3. flowDensity v1.40.0
4. flowTypeFilter v0.1.0

5. RchyOptimyx v1.2.0

Transcriptomics:
1. R v4.3.1
2. DESeq2 v1.38.3
3. FastQC v0.12.1
4. MultiQC v1.13
5. STAR v2.7
6. htseq-count (HTSeq) v2.0.2

Epigenomics:
1. R v4.1.2
2. EpiDISH v2.10.0
3. Limma v3.50.3
4. DMRcate v2.8.5
5. MissMethyl v1.28.0
6. Tidyverse collection (v2.0.0)
7. Minfi v1.40.0

Metabolomics:
1. MATLAB® (v24.2.0 - R2022b)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The proteomic, metabolomic and flow cytometry datasets and de-identified clinical data have been deposited in the Mendeley Data database under accession code 5ppxd6sgdb (DOI: 10.17632/5ppxd6sgdb.1). Due to ethical and legal restrictions, raw sequence and epigenomic files are not publicly available. Requests for access to these datasets or any other data supporting the findings of this study should be directed to the corresponding author [JRC]. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Subject sex was recorded in clinical notes and determined by genetic inference from whole blood samples.
Reporting on race, ethnicity, or other socially relevant groupings	No analyses were performed on race, ethnicity, or other socially relevant groupings. This information was not considered as part of the inclusion/exclusion criteria and was only reported in limited circumstances.
Population characteristics	Covariates were not controlled in the study design, and therefore where possible covariate-correction was performed as part of the statistical modeling process. To eliminate potential site-confounders, we leveraged samples from Mulago for initial discovery and performed validation using samples from Mbrara.
Recruitment	The detailed recruitment methods of the parent study have been previously published (Okello et al. doi:10.1161/JAHA.120.016053; Okello et al. doi:10.1016/S2214-109X(21)00288-6). ARF clinics were established at Mbarara Regional Referral Hospital (Western Region) and at Mulago National Referral Hospital (Central Region). Community and provider sensitization was conducted through direct education and mass media approaches. Children aged 3-17 years of age who presented with either history of fever in the past 48 hours and a joint complaint, suspicion of rheumatic carditis, or suspicion of chorea were invited to the clinic for additional evaluation. Additionally, for this sub-study, three control groups were recruited under an amendment of the original ethics from various hospital and community settings for each case of ARF, including one each from the categories of children with RHD without ARF, children with overlapping clinical presentation who had confirmed alternate infectious or inflammatory conditions (i.e., sickle cell crisis, malaria, influenza), and healthy controls with no signs or symptoms of acute infection or inflammatory condition. All parents signed a written informed consent prior to their child's participation, and children eight years of age and older also signed a written informed assent.
Ethics oversight	The study has approval from the Makerere University School of Medicine Research and Ethics committee (REC REF 2017-042), and from the Uganda National Council for Science and Technology (HS 2215).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample sizes are in the STROBE diagram (Supplementary Fig. S1).
Data exclusions	The sample sizes are in the STROBE diagram (Supplementary Fig. S1). 274 samples were available for analysis, including 163 samples from intake. Samples were excluded from analysis if there was insufficient sample amount or they failed QC. Excluded samples were: proteomics (n=13), flow cytometry (n=16), epigenomics (n=2), transcriptomics (n=9). All remaining samples were included in the analysis and the number of participants included in the various comparisons we carried out is specified in figure legends and the relevant results sections of the manuscript text.
Replication	The reproducibility of our results was assessed in two distinct ways. First, the model was trained using samples from the first study site (Mulago), with cross-validation performed to estimate the model's ability to generalize to unseen data (we report this cross-validated AUC). Subsequently, we applied the final trained model "off-the-shelf" (i.e., without further modification) to samples from a second study site (Mbarara), providing an independent estimate of the model's performance.
Randomization	Where samples were analysed in batches (epigenomics and flow cytometry), samples were were block randomized across plates so that each comparison group was equally represented on each plate. Samples for proteomics, metabolomics and transcriptomics were not randomised.
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Blood samples were processed to enable accurate cell population quantification, with all samples handled consistently within 10% of the total elapsed time across the study to ensure comparability. For sample preparation, 100 µl of a cells/RPMI mixture was aliquoted into four pre-labeled tubes and gently mixed by pipetting. 20mM EDTA (10 µl) and 1 µl of Fixable Viability Dye APC eFluor 780 (FVD) were added to each tube, followed by gentle mixing to ensure even distribution. The samples were incubated on ice (2–8°C) for 30 minutes, protected from light. Next, 175 µl of SmartTube Stable-Lyse V2 buffer was added to each tube, mixed gently, and incubated at room temperature for 15 minutes. A 1:1000 dilution of True Count beads in Stable-Store V2 buffer was prepared, and 500 µl of the bead-containing buffer was added to each sample. The samples were incubated at room temperature for an additional 15 minutes before being frozen at -80°C for storage for downstream analysis.
Instrument	BD LSRFortessa Cell Analyzer
Software	BD FACSDiva Software Version 8.0.1
Cell population abundance	Flow cytometry data were collected using a BD LSR II flow cytometer (BD Biosciences). Sample acquisition was performed with FACSDiva software (BD Biosciences), and data analysis was conducted using FlowJo 9.9. Cell populations were quantified based on the percentage of gated populations relative to total events following a pre-defined gating strategy.
Gating strategy	Compensation and gating strategies were standardized across samples, with fluorescence channel settings calibrated using single-stained controls and unstained samples. In addition, immunophenotyping using unbiased automated gating was undertaken on the same samples using flowCut, flowDensity, and flowTypeFilter as well as biomarker visualization (RchyOptimyx).

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.