nature research

Corresponding author(s): Rachel S Kelly

Last updated by author(s): Mar 26, 2021

# Reporting Summary

Nature Research 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 Research 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<br>*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<br>*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 | For the c8-pos, c18-neg and HILIC-pos metabolomic profiling platforms, raw metabolomic data was processed using TraceFinder software (Thermo Fisher Scientific) for targeted peak integration and manual review of a subset of identified lipids and using Progenesis QI (Nonlinear Dynamics) for peak detection and integration of both lipids of known identify and unknowns.<br>For the AMID-Neg platform raw data were processed using MassHunter Quantitative Analysis Software (Agilent). |
| Data analysis | All statistical analyses were conducted in R version 4.0.0. (with the exception of the genetic meta-analysis)<br>The metabo-endotypes were derived using Similarity Network Fusion (SNF) [R package: SNFtool version 2.2]; which included functionalities for fusing the data from the four metabolomic profiling platforms, spectral clustering to determine endotype membership and label propagation classification to assign the CAMP subjects to GACRS defined endotypes.<br>Differences between the endtyoes were assessed using the R fucntions 'aov()' and 'chisq.test()'<br>Power was computer using pwr.anova.test function from the R package 'pwr' [version 1.3-0].<br>For the genetic analyses SNPs were annotated to genes using biomaRt R package [version 3.12] and gene set enrichment analysis was conducted using gProfileR R package [version 0.7.0].<br>The metabolite meta-analysis was conducted using the meta R package [version 4.18-0]<br>The WGS meta-analysis was conducted using the '--meta-analysis' functionality of plink2 |

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The metabolomic data were generated as part of the NHLBI Trans-Omics for Precision Medicine Initiative (TOPMed). These data will be released to the scientific community in their entirety via NIH-designated repositories according to the TOPMed data release timeline. Full details can be found at https://www.nhlbiwgs.org/topmed-data-access-scientific-community.

# 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 | In GACRS: All children with suitable plasma samples collected at recruitment, with sufficient volume for metabolomic profiling were selected for these analyses<br>IN CAMP: all children with available plasma samples collected at the visit four years post baseline, with sufficient volume for metabolomic profiling were selected for these analyses<br>Power calcuations indicated the sample size in our dicscovery population (GACRS) provided good power to detect small effect differences (63.4% power based on the Cohen's F measure) and excellent power to detect medium or large differences (99% and 100% respectively) accross up to ten endotypes. (eTable2) |
| Data exclusions | Subjects were excluded if they had any missing data on age, sex, BMI (and Race in CAMP) only, as these variables were used to generate the metabolite residuals |
| Replication | We derived metabo-endotypes in GACRS, and then we recapitulated these endotypes in CAMP. This was achieved using a label propagation approach, whereby CAMP subjects were assigned to the GACRS derived endoypes based on their metabolomic profiles. In this way the GACRS and CAMP metab-endoypes were equivalent. We defined validation in this study as the observation of the same asthma-relevant clinical differences in the CAMP metabo-endotypes, as those seen in GACRS. The results of this validation can be seen in Figure 1 and in Tables 2 and 3. A number of the clinical differences, specifically those relating to lung function and lung obstruction validated between the two populations. We additionally did see clinical differences in GACRS that did not validate completely in CAMP. We hypothesise this was due to underlying differences between the cohorts relevant to these clinical metrics. In particular, with regard to the design of CAMP, the differing health systems of the nations from which the two populations originated and differences in the level of atopy between the two populations. We address this in the discussion. |
| Randomization | N/A. This was not a clinical trial, it was an observational study |
| Blinding | N/A. This was not a clinical trial, it was an observational study |

# 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 |
| ☐ | ☒ Human research participants |
| ☐ | ☒ Clinical data |
| ☐ | ☒ Dual use research of concern |

### Methods

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

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | A total of 1151 subjects with asthma from the GACRS discovery population had plasma samples available for metabolomic profiling (Table 1). These children were aged between 5.4 and 16.2 years (mean 9.22 years), and nearly 60% were male. The BMI of the participating children ranged from 8.2 to 41.4, and as GACRS represents a genetic population islet, all children were Hispanic.<br><br>In the CAMP validation population, 911 subjects with asthma had plasma samples extracted at the end of study and were eligible for inclusion. The age of this population at end of study ranged from 9.1-18.1 years (mean=12.9 years) and BMI ranged from 13.8 to 41.4. Sixty percent of children were male, 69.2% were White, 9.2% Black and 8.8% Hispanic<br><br>These metrics are described in Table 1 |
| Recruitment | GACRS recruited 1,165 children with asthma, aged 6-14 years, from 140 school across the Central Valley of Costa Rica between February 2001 and August 2008. Children were eligible for the study if they had asthma, as defined by physician-diagnosis and ≥2 respiratory symptoms or asthma attacks in the prior year, and a high probability of having ≥6 great-grandparents born in the Central Valley of Costa Rica.<br>This population was specifically selected to be a genetic population isolet, and as we note in the discussion this could limit the generalisabiltiy of our findings. However we do address this by means of a validation populaion- CAMP.<br><br>CAMP (Clinicaltrials.gov register: NCT00000575) was designed as a multi-center, randomized, double-masked, clinical trial designed to determine the long-term effects of inhaled treatments for mild to moderate asthma in children. From December 1993 to September 1995 CAMP recruited 1,041 children aged 5 to 12 at baseline with mild to moderate asthma from eight sites in North America (Albuquerque, Baltimore, Boston, Denver, San Diego, Seattle, St Louis and Toronto). Inclusion criteria included chronic asthma symptoms for at least 6 months in the year prior to interview and PC20 <12.5 mg/mL. Children were randomized to either nedocromil, budesonide or the placebo arm. We note that the findings of this trial were null, there was no difference in lung function or asthma relevant phenotypes between the study arms. We further note that the analysis conducted for this current manuscript is based on metabolomic profiling of blood samples collected after the completion of the trial and clinical indices measured after the completion of the trial. We do not consider the trial intervention or its primary outcomes in any of our analyses, and they are not relevant to our hypothesis or conclusions. Therefore we do not consider this to be a clinical trial, and we do not believe selection bias exists.<br><br>The only exclusion criteria for this study was missing data on a small number of covariates (age, bmi, sex or race) or insufficient volume or quality of plasma, and we determine that our study populations are representation of the wider GACRS and CAMP populations. |
| Ethics oversight | GACRS was approved by the Partners Human Research Committee at Brigham and Women's Hospital (Boston, USA); Protocol#: 2000-P-001130/55, and the Hospital Nacional de Niños (San José, Costa Rica).<br>CAMP study was approved by the institutional review board of Partners Healthcare (Partners Human Research Committee; Protocol#: 1999-P-001549/29), by all eight CAMP clinical centers and by the CAMP Data Coordinating Center. |

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

## Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

| | |
|---|---|
| Clinical trial registration | *Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.* |
| Study protocol | *Note where the full trial protocol can be accessed OR if not available, explain why.* |
| Data collection | *Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.* |
| Outcomes | *Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.* |

## Dual use research of concern

Policy information about dual use research of concern

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

| No | Yes | |
|---|---|---|
| ☒ | ☐ | Public health |
| ☒ | ☐ | National security |
| ☒ | ☐ | Crops and/or livestock |
| ☒ | ☐ | Ecosystems |
| ☒ | ☐ | Any other significant area |

## Experiments of concern

Does the work involve any of these experiments of concern:

| No | Yes | |
|---|---|---|
| ☒ | ☐ | Demonstrate how to render a vaccine ineffective |
| ☒ | ☐ | Confer resistance to therapeutically useful antibiotics or antiviral agents |
| ☒ | ☐ | Enhance the virulence of a pathogen or render a nonpathogen virulent |
| ☒ | ☐ | Increase transmissibility of a pathogen |
| ☒ | ☐ | Alter the host range of a pathogen |
| ☒ | ☐ | Enable evasion of diagnostic/detection modalities |
| ☒ | ☐ | Enable the weaponization of a biological agent or toxin |
| ☒ | ☐ | Any other potentially harmful combination of experiments and agents |