nature research

Corresponding author(s): DBPR NMED-A100113A

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 Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The microarray data used were publicly available and can be accessed via NCBI GEO (see below also.) Titer data were publicly available and can be retrieved for the following data sets through ImmuneSpace and ImmPort: NIH (SDY80 and through https://chi.niaid.nih.gov/DATA/chi/09-H-0239/), Yale (SDY400, SDY404), Stanford (SDY212). The Emory antibody titer data was received via personal communications and can be found in the figshare site accompanying this paper (see link in the Data Availability section below). Flow cytomery data were previously reported in Tsang et al, Cell 2014 and this data was used to gate the additional cell populations (see Methods). CITE-seq data were generated using the protocol in the Methods, including using a 10x Genomics (Pleasanton, CA) workflow. RNA sequencing analysis was performed on sorted B cell populations derived from peripheral blood mononuclear cells of healthy human donors.
Data analysis	Full details of software versions can be found in Supplementary Table 8. Flow cytometry data was gated using FlowJo v9.9.3 for OS X. Data analysis for the first part of the study (Figs. 1-3 and Ext. Data Figs. 1-7) was performed using R version 3.4.1. CITE-seq data analysis was performed using R version 3.6.1. R packages from CRAN and Bioconductor repositories were used and their versions are listed in Supplementary Table 8. R code to reproduce the analysis is available at: https://github.com/kotliary/baseline. Software used for low-level processing of the CITE-seq data are listed in the "Computational low-level processing of CITE-seq data" section in Methods. Statistical analysis for Extended Data Fig. 9 was performed using GraphPad Prism version 8.3.0.328 (GraphPad Software, San Diego, CA).

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/reviewers upon request. 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

All data used in the study including the flow cytometry and CITE-seq data in analysis-ready format are available at figshare: https://doi.org/10.35092/yhjc.c.4753772. All microarray data were publicly available prior the study. Titer data was publicly available for the following data sets through ImmuneSpace and ImmPort: NIH (SDY80), Yale (SDY400, SDY404), Stanford (SDY212). The Emory antibody titer data was received via personal communications and can be found via the figshare link above. See Supplementary Table 2a for details. The original/raw public gene expression data are available in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) under accession numbers: GSE47353, GSE41080, GSE59654, GSE59743, GSE29619, GSE74817, GSE13486, and GSE65391.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	Sample sizes for the data sets used are reported in Supplementary Table 2a. CITE-seq experiment was performed for 20 samples: 10 low and 10 high influenza vaccination responders (20 subjects in total). Gene expression of CD38 high B cell populations were assessed using 6 healthy donors not included in the influenza vaccination study. See Supplementary Table 9 for demographic information.
Data exclusions	All sample exclusions (due to QC or scientific reasons) were performed prior to data analysis (see Methods for details and reasons for the exclusion).
Replication	TGSig was assessed using independent data from multiple publicly available vaccination studies and in one SLE study (see manuscript for results), as well as technically replicated by the CITE-seq experiment.
Randomization	No randomization of subjects was performed in this study. The subjects were assigned to low, middle and high responder classes by analysis of the antibody response to vaccination (log fold-changes of post- vs. pre-vaccination titers, except for Yellow Fever where the initial titer readings were close to zero and we only assigned subjects based on the post-vaccination titers). SLE patients were assigned to patient groups by unsupervised analysis of their gene expression association with disease activity. SLE samples from multiple visits were categorized to low, middle and high disease activity according to the SLEDAI.
Blinding	The TGSig signature was developed using the NIH/CHI data only before we assessed it using other datasets. Similarly, signatures associated with disease activity were developed using only the SLE dataset first before they were assessed using vaccination data.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

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

Antibodies

Antibodies used	Antibodies from Biolegend were used in CITE-seq; please see Supplementary Table 10 for details. Antibodies used for flow sorting are described in the methods under heading "Sorting of B cell populations from healthy donors."
Validation	Quality control and reprodicibility statements are available on the following Biolegend TotalSeq-A website: https://www.biolegend.com/en-us/quality-control https://www.biolegend.com/en-us/reproducibility

Human research participants

Policy information about studies involving human research participants

Population characteristics	PBMC samples for CITE-seq were acquired previously (see Tsang et al, Cell 2014). Age, gender and race/ethnicity information for all subjects is publicly available (See Supplementary Table 9). Cohort characteristics for the other public data we used can be found in the original publications (and some may also be available in NCBI GEO and NIAID ImmPort.) See also Supplementary Table 2a.
Recruitment	PBMC used for CITE-seq were collected prior to the study; see Tsang et al, Cell 2014 for details on the original NIH cohort. See also Supplementary Table 2a. For sorting of B cell populations from healthy donors, 6 healthy donors were recruited under IRB approved protocol NCT00001281 (See Supplementary Table 9).

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	This study uses previously generated Flow Cytometry data. See the respective publications cited, including Tsang et al, Cell 2014.
Instrument	See the original publications cited, including Tsang et al. Cell, 2014.
Software	FlowJo ver.9.9.3 (Becton Dickinson Co., Ashland, OR) on Mac OS X; R version 3.4.1.
Cell population abundance	Cell frequencies are expressed as the percentage of the parent population.
Gating strategy	Using the flow cytometry data and predictive B cell populations defined in (Tsang et al., 2014), we gated four new cell subsets using the CD38 and CD10 markers) under the CD45+CD19+CD20+ live B-cell population (See Extended Data Fig. 1). Gate 1: CD38 + cells with CD38 (G610-A) fluorescent intensity over 1000; gate 2: CD38++ cells with fluorescent intensity over 10000; gate 3: CD38++CD10+ cells with CD10 (R780-A) fluorescent intensity over 1000; gate 4: CD38++CD10- cells with CD10 fluorescent intensity below 1000.

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