

Corresponding author(s): Yuanting Zheng, Leming ShiLast updated by author(s): Sep 3, 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
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | 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
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | 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

The details of methods for the data collection were described in the Methods section. The source codes are available at Github (<https://github.com/QiaochuChen/proteomics-batch-effect-correction-benchmarking>).

Data analysis

Software: R v4.4.1. R codes for data analysis and figure generation are deposited in Zenodo ([doi:10.5281/zenodo.17032531](https://doi.org/10.5281/zenodo.17032531)).

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 statistical data generated in this study have been deposited in the Figshare database at <https://doi.org/10.6084/m9.figshare.29567366.v2>. The processed Quartet and simulated proteomics data are available in the Figshare database at <https://doi.org/10.6084/m9.figshare.29567333.v2>. The processed ChiHOPE proteomics data are available at <https://doi.org/10.6084/m9.figshare.30028336>. All performance assessment results generated in this study are provided in the Source Data file. The raw Quartet and ChiHOPE proteomics data used in this study are available in a ProteomeXchange partner database under accession code

PXD045065 (<https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PX045065>) and accession code PXD068273 (<https://proteomecentral.proteomexchange.org/ui?search=PX068273>). 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

Quartet samples were derived from four healthy volunteers of a Chinese Quartet family from the Fudan Taizhou Cohort, including father (F7), mother (M8), and monozygotic twin daughters (D5 and D6).
In ChiHOPE data, a total of 1,431 plasma samples were collected at baseline or after a 24-week treatment from 750 subjects with T2D (274 females and 476 males, aged from 23 to 70 years); two plasma samples for use of quality control (QC) were collected from two healthy donors, of which one is male (P10) and the other is female (P11); one plasma sample (PM) was prepared by mixture of all plasma samples of patients with T2D. There were 32, 16 and 16 biological replicates of Pm, P10 and P11, respectively.

Reporting on race, ethnicity, or other socially relevant groupings

All participants are adults from Han Chinese.

Population characteristics

No additional covariates-relevant population characteristics were collected for each donor due to IRB approval restrictions.

Recruitment

Quartet family is randomly selected from families with monozygotic twin daughters recruited from Taizhou Longitudinal Study. Two healthy volunteers are recruited from Fudan University. In ChiHOPE proteomics data, 750 participants are a subset of 835 drug-naïve patients with T2D recruited from two multi-center, randomized, double-blind phase 3 trials of chiglitazar (CMAP, placebo-controlled, ClinTrials.gov registration no. NCT02121717; CMAS, sitagliptin-controlled, ClinTrials.gov registration no. NCT02173457).
No self-selection bias was expected to be introduced.

Ethics oversight

All Ethical approvals were obtained from Ethical Committees of the study centers. All procedures which performed in the study involving human participants were in accordance with the ethical standards of the institutional and (or) national research committee, and with the Declaration of Helsinki and its later amendments or comparable ethical standards. All participants provided written informed consent.

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

For Quartet data, there are a total of 72 MS runs from six labs; each lab injected and measured biological triplicates of four biological sample groups (D5/D6/F7/M8).
For simulated data, there are a total of 27 samples from three batches; each batch contains triplicates of four preset sample groups (Group1/2/3).
For ChiHOPE data, there are 1,495 MS runs, which are composed of 1,431 samples from 1,419 unique plasma specimens (12 technical duplicates) collected from 750 subjects at baseline and 24 weeks, and 64 accompanying QC samples.

Data exclusions

Due to unusually low protein quantities that was identified as an outlier, sample ExpB19 and ExpJ1 in ChiHOPE data was excluded in some supplementary figures, but all samples of the data were included for batch-effect correction and performance assessments. This has been noted in the supplementary figure legends.

Replication

For biological replicates, there are 18 replicates for each Quartet group, 9 replicates for each simulated group; 12 duplicated patient samples, 32 replicates of PMs, 16 replicates of P10s and 16 replicates of P11s in ChiHOPE data.
For replicates generated by data preprocessing workflow, details was illustrated explicitly in the Supplementary Tables.

Randomization

Aliquots of Quartet reference materials (digested protein) were randomly distributed to the lab.
Aliquots of plasma samples were randomized prior to proteomics sample preparation and data acquisition.

Blinding

Samples distributed to each lab was blinded to avoid specific experimental order affecting the objective assessment of lab proficiency.

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

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	The human immortalized B-lymphoblastoid cell lines of four healthy volunteers from a family Quartet, as part of the Taizhou Longitudinal Study in Taizhou, Jiangsu, China.
Authentication	The cell lines have been authenticated by STR profile, karyotype, PCR mycoplasma and sterility testing.
Mycoplasma contamination	No mycoplasma contamination found.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

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	CMAF, placebo-controlled, ClinTrials.gov registration no. NCT02121717; CMAS, sitagliptin-controlled, ClinTrials.gov registration no. NCT02173457
Study protocol	The primary publications describing the trial designs and the cohort (PMID: 36654286 & 36654287).
Data collection	Details are described at https://www.biosino.org/node/project/detail/OEP00002924 .
Outcomes	Our study aims to evaluate and compare the performance of different batch-effect correction strategies. Therefore, the primary and secondary outcome measures are defined around the performance assessment (described in the Methods section), not clinical results.

Plants

Seed stocks	<i>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.</i>
Novel plant genotypes	<i>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.</i>
Authentication	<i>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, mosaicism, off-target gene editing) were examined.</i>