nature research

Corresponding Author: David Reich

Date: 06/06/2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   We used simulations to determine the approximate number of individuals that would be needed to detect strong founder events (main text lines 457-490).

2. Data exclusions

   Describe any data exclusions.

   We used standard quality control metrics in the field to exclude SNPs and individuals based on genotyping completeness, batch effects, close relatives, or PCA outliers (main text lines 409-453).

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   We increased our sample size for groups detected to have strong founder events and repeated the analyses with the larger sample sizes.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Individuals were grouped based on anthroplogical information.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not relevant to the study, because it would not affect the results.

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Genetics: doi:10.1038/ng.3917

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We used PLINK version 1.7 (a genotype analysis tool for basic calculations of SNP data), smartpca (from EIGENSOFT to perform PCAs and Fst calculations), GERMLINE (version 1.5.1, to calculate IBD), ARGON (for simulations of population history), Beagle (version 3.3.2, to do phasing), and Haploscore (to find false positive IBD segments), qpgraph (from Eigensoft, to make a model of population history), ADMIXTOOLS (to do f3 statistics), and self-written code to do smaller data manipulations.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

The informed consents and permits associated with the newly reported data are not consistent with fully public release. Therefore, researchers who wish to analyze the data should send the corresponding authors a PDF of a signed letter containing the following language: "(a) We will not distribute the data outside my collaboration, (b) We will not post data publicly, (c) We will make no attempt to connect the genetic data to personal identifiers, (d) We will not use the data for commercial purposes."

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used have been authenticated OR state that no eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination OR state that no eukaryotic cell lines were used.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Individuals came from over 260 groups in India, Pakistan, Nepal, Sri Lanka, and Bangladesh. Age and gender were not provided. Patients with mucopolysaccharidosis type IVA and progressive pseudorheumatoid dysplasia were also studied. Their ethnic group, age and gender information were not provided or used for this study.

Nature Genetics: doi:10.1038/ng.3917