nature research

Corresponding author(s): Evan E. Eichler

Last updated by author(s): May 14, 2020

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.

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

BWA-MEM (version 0.7.15-r1140) - used for alignment of short reads to the reference genome.
SAMtools (version 1.7) - used for BAM file manipulation and sorting.
sambamba (version 0.6.6) - used for marking PCR duplicate reads after alignment.
minimap2 (version 2.14-r883) - used for alignment of long reads to the reference genome.
iVision-Mac (version 4.0.14; BioVision Technologies) - used to collect fluorescent FISH images.
Adobe Photoshop (version CS6) - used for postprocessing of collected FISH images.

Data analysis

R package primatR (https://github.com/daewoooo/primatR)(version 0.1.0) - used for in-depth analysis of detected inversions.
R package breakpointR (https://github.com/daewoooo/breakpointR,devel branch)(version 0.99.10) - used to detect inversion breakpoints
pbsv (version 2.0.2) - used to detect structural variants in long PacBio reads.
Sniffles (version 1.0.10) - used to detect structural variants in long PacBio reads.
DELLY (version 0.7.9) - used to detect structural variants in short reads.
Bayesian phylogenetics tool BEAST (version v2.5.0) - used to reconstruct evolutionary relationships among great apes.
RTG tool (RTG Core Non-Commercial version 3.9.1) - used to detect single-nucleotide variants in Strand-seq data.
cDNA_Cupcake tool (https://github.com/Magdoll/cDNA_Cupcake/) (version 7.2) - used to detect fusion genes.
R package regioneR (version 1.16.2) - used to perform enrichment analysis of various genomic features.
R package DEseq2 (version 1.24.0) - used to perform differential expression analysis.
R package UpsetR (version 1.4.0) - used to visualize overlaps between various datasets.
R package gggenes (version 0.4.0) - used to visualize inverted haplotypes in NHPs.
Long Ranger (version 2.2.2) - used for processing of 10x Genomics data.
Canu (version 1.9) - used to produce local haplotype assemblies for inversion validation.
DottedPython (https://github.com/ruiyangl/Dottedpython) - used to produce dotplots for local haplotype assemblies.
FreeBayes (10x provided version: v0.9.21-7-g7dd41db-dirty) - used calling for single-nucleotide variants.
Tracer (version 1.7.1) - used to determine the quality of phylogeny prediction.
Figtree (version 1.4.3) - used to visualize phylogenetic trees.

DensiTree (version 2.2.6) - used to visualize phylogenetic trees.
minimap2 (version 2.17) - used to align SDA contigs to GRCh38.
nucmer (version 3.1) - used for targeted alignment of SDA contigs against a reference sequence.

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. 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

Strand-seq data aligned to GRCh38 and ape-specific composite files are available at zenodo, DOI: 10.5281/zenodo.3818043
PacBio and Bionano datasets are reported in Supplementary Table 11 and 14.
Supplementary data: https://github.com/daewoooo/ApeInversion_paper
PacBio and Bionano inversion callset: https://github.com/daewoooo/ApeInversion_paper/Supplementary_datasets

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 this study we analyzed a given number of Strand-seq single-cell libraries: chimpanzee (n=62), bonobo (n=51), gorilla (n=81) and orangutan (n=60). Strand-seq libraries selected based on quality as described previously (Porubský et al. 2016; Sanders et al. 2017).
Data exclusions	We have excluded single-cell libraries with visible background reads (i.e., reads mapped to opposite direction on chromosomes that inherited template strands with the same directionality) and libraries with low (<50,000 reads) or uneven coverage were excluded, as detailed previously (Porubský et al. 2016; Sanders et al. 2017).
Replication	Replication of Strand-seq experiments was performed as required in order to achieve expected number of informative Strand-seq libraries (~50).
Randomization	As we analyzed only one individual for each studied great ape species, randomization is not applicable for this study.
Blinding	As we do not compare two distinct subject groups, blinding was not applicable in this 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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All cell lines used in the study were previously generated EBV-immortalized B-cell lymphoblastoid cell lines. The chimpanzee

Cell line source(s)	(Dorien) and bonobo (Ulindi) lines were generously provided by Svante Pääbo, while the gorilla (GG09) and orangutan (PPY10) lines were provided by Francesca Antonacci.
Authentication	The lineage of each cell line was authenticated based on the inversion predictions and phylogenetic analyses performed from the sequence data--all of which were consistent with previous studies.
Mycoplasma contamination	All lines were negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None used.

Human research participants

Policy information about studies involving human research participants

Population characteristics	We chose three parent–child trios (mother, father and child): a Han Chinese (CHS) trio (HG00513, HG00512, HG00514), a Puerto Rican (PUR) trio (HG00732, HG00731, HG00733) and a Yoruban (YRI) Nigerian trio (NA19238, NA19239, NA19240). The Han Chinese and Yoruban Nigerian families were representative of low and high genetic diversity genomes, respectively, while the Puerto Rican family was chosen to represent an example of population admixture. Data obtained from Chaisson et al. (2019)
Recruitment	Transformed lymphoblast cell lines from three parent-child trios belonging to the 1000 Genomes Project were obtained from the Coriell Cell Repository as part of the NHGRI catalog (Chaisson et al. 2019).
Ethics oversight	See Chaisson et al. (2019).

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