nature portfolio

Corresponding author(s): Eugenio Montini, Andrea Calabria

Last updated by author(s): Oct 18, 2024

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 *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	PCR amplification products were sequenced using the Illumina Myseq/HiSeq platform. BCL to FASTQ file converter Bfastq v2.20.0.422 sofware is used to create Fastq data. Bioanalyzer 2100, ddPCR, Nanodrop, Qubit, and ViiA7 qPCR thermocycler instrument were used for sample preparation and quantification.
Data analysis	Statistical analyses were performed with GraphPad Prism 10.0 and R (4.0.3, see below for details) with ISAnalytics software version 1.12, bioconductor BioC 3.12. Statistical significance for each CIS was established using the Grubbs test for outliers, as described in Biffi et al, 2011. GO has been realized using R packages for GO "clusterProfiler", the annotation DB "org.Hs.eg.db", "msigdbr". Semantic similarity has been done with the R package "GOSemSim"64. Feature annotations have been realized with RefGene table (UCSC database hg19). Circos plot generated by the R package "circlize". The following R libraries and software version have been used for IS analysis and statistics: R 4.0.3 with base packages (including 'stats' and 'splines') and the packages fpc (2.2-9), DescTools (0.99.47), cluster (2.1.4), vegan (2.6-4), permute (0.9-7), ISAnalytics (1.0.11), magrittr (2.0.3), factoextra (1.0.7) with the function 'pca', ggbreak (0.1.1), rstatix (0.7.0) with the function 'wilcox_test', ggpubr (0.4.0), circlize (0.4.15), openxlsx (4.2.5.1), Hmisc (4.7-1), Formula (1.2-4), survival (3.4-0), lattice (0.20-45), dplyr (1.0.8), reshape2 (1.4.4), psych (2.2.9), plyr (1.8.7), sqldf (0.4-11), RSQLite (2.2.18), gsubfn (0.7), proto (1.0.0), stringr (1.4.1), gridExtra (2.3), scales (1.3.0), gplots (3.1.3), RColorBrewer (1.1-3), pheatmap (1.0.12), ggrepel (0.9.1), ggplot2 (3.5.1). The following packages have been used for Good Turing and Bayesian regression: R version 4.2.2 (2022-10-31), plyr_1.8.9, tools_4.2.2, jsonlite_1.8.8, grid_4.2.2, tidyselect_1.2.0; Python 3.8.15, packaged by conda-forge, sklearn .0.2, joblib 1.2.0, numpy 1.24.1, scipy 1.10.1, threadpoolctl 3.1.0. See the repository for all code developed for this study: https://github.com/calabrialab/Code_HSPCdynamics.

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

Data availability
Sequencing data of mutations have been deposited into the Sequencing Read Archive (NCBI SRA) with the BioProject PRJNA1150995 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1150995). IS data have been deposited in Github at the following project archive: https://github.com/calabrialab/Code_HSPCdynamics. Source figure data are provided with this paper.

Code Availability
We released our source code in Github (https://github.com/calabrialab/Code_HSPCdynamics).

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	Gender analysis was performed and no differences in sex were observed in any analysis.
Reporting on race, ethnicity, or other socially relevant groupings	Race, ethnicity, or other socially relevant groupings were not considered in this study
Population characteristics	MLD patients were enrolled in a phase 1/2 clinical trial (NCT01560182) or treated with a hospital exemption program or CUP. A total of 29 patients had been treated with HSPC gene therapy clinical protocol for ARSA deficiency (Eudract no. 2009-017349-77). Sixteen of the treated patients (Pt16, Pt32, Pt47, Pt02, Pt20, Pt34, Pt31, Pt03, Pt37, Pt33, Pt08, Pt53, Pt23, Pt40, Pt25, Pt28) were affected by Late Infantile (LI) MLD in a pre-symptomatic stage and have been identified by molecular and biochemical tests in the presence of at least an affected older sibling, while 13 patients (Pt38, Pt01, Pt10, Pt43, Pt42, Pt04, Pt30, Pt51, Pt44, Pt18, Pt50, Pt14, Pt07) were affected by Early Juvenile (EJ) MLD in a pre- or early-symptomatic stage. Patients were treated with a myeloablative busulfan conditioning regimen administered before reinfusion of the engineered HSPCs . Fourteen male WAS patients for whom no human leukocyte antigen–identical sibling donor or suitable matched unrelated donor was available underwent lentiviral GT after a reduced conditioning regimen protocol. Patients Pt52, Pt21, Pt48, Pt13, Pt29, Pt11, Pt39 and Pt17 were enrolled in an open-label, non-randomized, phase 1/2 clinical study 32 registered with ClinicalTrial.gov (number NCT01515462) and EudraCT (number 2009-017346-32). The other WAS patients were treated under early access program, compassionate use program (CUP) or hospital exemption. Among β thalassemic patients, 3 adults and 6 children with β0 or severe β+ mutations were enrolled in a phase 1/2 trial (NCT02453477) for intrabone administration of GLOBE lentiviral vector–modified HSPCs after myeloablative conditioning with treosulfan–thiotepa. Population characteristics are reported in Supplementary Table 1.
Recruitment	Participants were referred by MLD/WAS/thalassemia centers, by patient societies or self-referred. All subjects interested in the trials received trial specific information and were screeed for inclusion and exclusion criteria. Subjects meeting all inclusion criteria and without all exclusion criteria were enrolled sequentially in the specific age cohort until completion of enrolment. Written informed consent was obtained from patients and/or parents.
Ethics oversight	Gene therapy patients included in the study were treated in the context of clinical trials or early access programs approved by ethical committee and competent regulatory authorities. The treatment was administered at the Bone Marrow Transplantation Unit at the San Raffaele Scientific Institute in Milan, Italy. We have complied with all the ethical regulations for retrieving biological materials from gene therapy patients. Parents signed informed consent for research protocols approved by the San Raffaele Scientific Institute's Ethics Committee (TIGET06 and TIGET09). All patients received autologous HSPC transduced with transgene encoding lentiviral vectors under the same transduction protocol, as previously described previously.

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	No statistical method was used to predetermine sample size of the cohort of patients because we used all available patients in the clinical studies. For IS analyses we use vector integration sites available from the patients enrolled.
Data exclusions	Technically validated results were always included to the analyses. We did not apply any exclusion criteria for outliers.
Replication	For IS retrieval fragmented DNA was split in technical replicates, based on the available material, prior to end repair and 3' adenylation process. At the end of the PCR procedure adopted for IS retrieval, each amplified sample was quantified in technical triplicates by qPCR (KAPA). Replicates were checked and repeated if the source material was enough. All attempts at replication were successful. The full list of samples is available in Supplementary Tables. IS analyses represent individual analyses of peripheral blood and bone marrow samples obtained at different time points post infusion.
Randomization	This is not a clinical trial. The experimental design did not include allocation of samples to randomised experimental group.
Blinding	The experimental design did not include allocation to groups nor to blinding given that this is not a new clinical 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
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

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	MLD phase 1/2 clinical trial (NCT01560182) and Eudract no. 2009-017349-77; WAS phase 1/2 clinical study (NCT01515462) and EudraCT No. 2009-017346-32). B-thalassemia phase 1/2 trial (NCT02453477)
Study protocol	The protocols for the clinical trials are available upon request by Telethon.
Data collection	Data were collected in paper Case Report Forms (CRF) and subsequently transferred to an electronic database by the Marketing Authorization Holder (MAH).
Outcomes	This is not a clinical trial

Plants

Seed stocks	*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.*
Novel plant genotypes	*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.*
Authentication	*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, mosiacism, off-target gene editing) were examined.*