nature research

Corresponding author(s): Jonathan Weissman, Alexander Meissner

Last updated by author(s): Mar 18, 2019

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	scRNA-seq data was processed and aligned using 10x Cell Ranger v2. The filtered gene-barcode matrices were then processed in Seurat v2.0 (https://satijalab.org/seurat/) for data normalization (global scaling method “LogNormalize”), dimensionality reduction (PCA), and generation of t-sne plots, which use the first 16 principal components. Amplicons were additionally processed using cutadapt v1.14 to remove sequence beyond the polyA (http://cutadapt.readthedocs.io/en/stable/) and BioPython v1.7 to build a consensus sequence from multiple, trimmed UMIs using parameters described in the methods. We use emboss water (v6.6.0) to align sequences to the target site reference sequence with the following parameters, which were determined empirically: [–asequence targetSiteRef.fa –sformat1 fasta –bsequence consensusUMI.fa –sformat2 fasta –gapopen 15.0 –gapextend 0.05 –outfile sam –aformat sam]. Additional processing of the resulting alignment files was done in perl.
Data analysis	Following processing using a custom software pipeline described above, data was analyzed using Python.

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

Lineage tracing data is available in GEO with accession number GSE117542. Wild type embryo data is available under GSE122187. All figures use raw data generated in this project.

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 sample size calculation was performed, the number of embryos reported is the number that we were able to generate in a reasonable cost and time frame. Each embryo generates a stochastic mutational path from the delivery of the target site to collection at E8.5, requiring the innovation of novel analytical tools and perspectives to confirm the reproducible generation of similar lineages, which was accomplished via strategies such as the shared progenitor score between different tissues. Each embryo collected demonstrated our reproducible ability to recover indels according to the rate of the guide series and to assign them to matched transcriptional profiles from the same cells. Moreover, the general composition of embryos was also consistent and fell between E8.0 and E8.5 when compared to a wild type compendium.

Data exclusions

We excluded one of the seven embryos for which we generated single cell data from detailed lineage analysis because it did not produce cells of the primitive heart tube, suggesting a developmental abnormality that may be due to the mutational nature of the randomly integrating, target site containing piggyBAC transposon.

Replication

We demonstrate the reproducible nature of our technology to be recovered from early embryos and to be assigned to a consistent make up of developmental cell types. The reproducibility of lineage relationships is more difficult to evaluate due to the stochastic nature of indel generation, though we confirm general trends between high complexity embryos by comparing shared progenitor scores as a proxy for the ancestral relationships between different tissues. Analysis related to the reproducibility of lineage relationship is presented for each of the six morphologically normal embryos in Figures 4, Extended Data Figures 8 and 9

Randomization

As our objective was to recover high complexity embryos with as large a number of integrated target sites as possible, embryos were selected for inclusion in this study based upon the uniform brightness and high intensity of the target-site linked reporter. Our study does not follow a hypothesis driven design, and as such, no groupings of embryos were made therefore randomization was not applicable.

Cells were assigned to states according to their Euclidean Distance to the 712 marker genes described in the methods and available as a supplementary file in GSE122187. The robust nature of these assignments were confirmed by comparing the distance for each assigned cell to its closest and next closest cluster center (see Extended Data Figure 5c).

To estimate shared progenitor scores, we downsampled the number of cells from each tissue before calculating: 150 cells were randomly sampled from each tissue and the tree was pruned to only include the sampled cells. For tissues with less than 150 cells, all cells were included. For embryo 2, we downsampled to 300 cells since it is a merger of two biological replicates and is therefore doubly sampled. The shared progenitor score was calculated from the pruned tree and the process was repeated 1000 times for each embryo. The median progenitor score is presented in the heatmap and was used instead of the mean to prevent potential outlier effects.

During tree reconstruction, we attempted one of two different approaches, "Biased Search Through Phylogenetic Space" and "Greedy." In Biased search, we generated trees by selecting indels either randomly or according to their frequency normalized weight (the fraction of alleles an indel is found divided by its independent frequency, see Figure 2c). In these cases, we generated >30,000 simulated trees and calculated the log likelihood of each by summing the likelihoods of all indels that appear in the tree and reported the one with the highest likelihood. We also employed a greedy algorithm that recursively splits cells into mutually exclusive groups based upon the presence or absence of a specific mutation, prioritizing mutations that appear frequently within the embryo but are improbably according to their independent likelihood (see Figure 2c). This approach yields only one tree, which was only selected if it performed better than the best tree recovered by our sampling approach. Finally, the cumulative tree for embryo 2 could only be generated with this approach is it includes too many cells to enable robust sampling over ~100,000 simulations.

Blinding

Tree building and cell state assignments operate with the same parameters independently of the embryo used. As such, there is no need to blind the investigator to the data being handled. Individual parameters were not altered according to the specific features of a given sample, with the following minor exceptions:

The number for assigning shared progenitor scores was set according to the overall complexity (number of cells within each tissue) to 300 for embryo 2 and 150 for all other embryos because embryo 2 was sampled ~2x more deeply.

The number of frequency normalized weighted tree simulations for each embryo depended on the number of alleles: higher allele numbers underwent fewer simulations due to increased processing times. In these cases, the greedy algorithm consistently yielded trees with appreciably higher probabilities.

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)	The K562 cell line originated from ATCC.
Authentication	Cytogenetic profiling by array comparative genomic hybridization closely matches previous characterizations of the K562 cell line (Naumann et al., 2001).
Mycoplasma contamination	Cell lines tested negatively for mycoplasma
Commonly misidentified lines (See ICLAC register)	None used

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Oocytes were isolated from B6D2F1 strain female mice (age 6 to 8 weeks, Jackson Labs) , sperm was isolated from 2 8 week old Gt(ROSA)26Sortm1.1(CAG=cas9*,EGFP)Fezh/J strain mouse (Jackson labs) or C57BL/6J strain mice. Blastocysts were transferred into CD-1 strain female mice (age 6-10 week old).
Wild animals	None used
Field-collected samples	None used
Ethics oversight	All procedures follow strict animal welfare guidelines as approved by Harvard University IACUC protocol (#28-21).

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

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	K562 cells were filtered to make a single cell suspension.
Instrument	LSR-II flow cytometer (BD Biosciences)
Software	FlowCytometryTools (http://eyurtsev.github.io/FlowCytometryTools/)
Cell population abundance	To isolate reporter cell lines, the population abundance was <10% of the unsorted population.
Gating strategy	Cells were sorted against a negative control, or gating thresholds were obvious from bimodality in the cell population for highly expressed fluorescent proteins.

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