

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                                                                                                                                                                                                                                                                                      |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement                                                                                                                               |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly                                                                                                                                    |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>                                                               |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested                                                                                                                                                                                                                     |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons                                                                                                                                        |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> 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) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted<br><i>Give P values as exact values whenever suitable.</i>                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings                                                                                                                                                                      |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes                                                                                                                                                |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), 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 | we used R MatchIt package for case-control matching in the RDW comparisons (Fig 2F)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Data analysis   | scRNA-seq data analysis: vireo (version 0.3.2, with cellSNP version 0.3.0) and souporcell (version 2.4) were used to identify doublets and assign cells to individuals. To produce UMI count matrices, cell-ranger version 3.1.0 was used in the analysis resulting in Figures 1-3 and EDFs 1-6, while cell-ranger version 7.0.1 was used in the analysis resulting in Figure 4 and EDFs 7-10. In both cases, cell-ranger count was used to align reads to hg38. metacell2 was used to construct all cohort and individual metacell models.<br>Targeted DNA-seq (MIP): Varscan was used to call SNPs from the genotyping MIP panel. BWA-MEM (v2) was used to align reads either to hg38, or to hg37 while LiftOver was used to convert hg37 genomic positions to hg38. For predicting MDS diagnosis, xgboost Python package (v2.0.3) was used.<br>Code to reproduce the figures is available at <a href="https://github.com/tanaylab/blood_aging">https://github.com/tanaylab/blood_aging</a> . The Metacell package is available at <a href="https://github.com/tanaylab/metacells">https://github.com/tanaylab/metacells</a> |

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](#)

scRNA-seq data generated in this study is available in Gene Expression Omnibus (GEO) under accession GSE285943, in CELLxGENE (<https://cellxgene.cziscience.com/collections/5542eeb0-96ef-4ab9-95ea-eb6abc178461>), and also as metacells at [https://apps.tanaylab.com/MCV/blood\\_aging/](https://apps.tanaylab.com/MCV/blood_aging/). Targeted DNA sequencing data of clonal hematopoiesis mutations by Molecular Inversion Probes is available in European Nucleotide Archive (ENA) under accession PRJEB85241. All of these data were uploaded in accordance with donor written informed consent.

## 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

No gender-related data was collected. Sex data was self-reported and verified based on XIST and RPS4Y1 expression. Sex data is reported in Supplementary Tables S1 and S9.

### Reporting on race, ethnicity, or other socially relevant groupings

No sociopolitical construct (such as race or ethnicity) were considered in this study. We acknowledge that due to practical considerations in recruitment, this study included mostly individuals of Jewish ancestry, and so our results should be validated in a larger and more ancestrally diverse cohort.

### Population characteristics

For the healthy cohort (analyzed in Figures 1-3) we collected blood samples from 148 consented healthy individuals, consisting of 79 males and 69 females aged 23-91 (median 61.5). The cytopenic cohort (analyzed in Figure 4) included a total of 83 individuals, 50 of which were labeled as MDS cases based on BM morphology and/or mutational and karyotypic abnormalities (as detected in the clinic, or by our CH panel and scRNA CNA analysis). The remaining 33 cytopenic patients not satisfying MDS criteria were labeled as cytopenia cases. 17 of the 83 cytopenic patients, presenting with asymptomatic mild cytopenia, were also included in the original 148-individual healthy cohort. Median age for the cytopenic cohort was 73 years (range 27-93), with males representing 53% of patients.

### Recruitment

Recruitment of the healthy cohort took place between Nov 2020 and Dec 2023. Blood donors were recruited from numerous sources (including the Weizmann Institute of Science and several primary care clinics) as detailed under "Methods". We profiled responsive volunteers with no known hematological malignancy / prior information on blood clonality, balancing sex and seeking a dispersed age distribution biased toward older individuals. We reassessed this strategy following initial sampling, and observed remarkable homogeneity in transcriptional states across individuals. We observed large variation in the composition of cell states across healthy individuals, which suggested that more sampling, without specific preferences, would be appropriate for further characterizing compositional variation. Recruitment of the cytopenic cohort (including MDS and non-MDS-related cytopenic patients) took place between Nov 2021 - Feb 2024. These patients were recruited from several outpatient hematological clinics by collaborating physicians to represent the wide clinical spectrum of MDS, from patients with moderate anemia and mild dysplasia as their sole BM abnormality to those with severe cytopenia and excess blasts on BM analysis. As different physicians use different criteria for BM analysis this can create a selection bias among the cytopenic patients. Furthermore, some of the cytopenic patients did not go through BM analysis as their physicians decided it was not appropriate, which might lead to selection bias in cytopenic cases. A significantly larger study across diverse clinical attributes and locations is needed to overcome these biases.

### Ethics oversight

All protocols were approved by the Weizmann Institute of Science ethics committee (under IRB protocol 283-1) or approved by the TSMC ethics committee (under IRB protocol 02-130).

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](https://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 calculations of sample size were made. Recruitment of the healthy cohort was intended to allow characterization of the normal variation in CHSPC states. As no such profiling had been previously performed, we could not assume much regarding the variance in the population or define sample size a-priori. We aimed to sample equally from different age groups and both genders.

|                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | In the analysis resulting in Figures 1-3 and EDFs 1-6, We filtered out cells with at least 20% mitochondrial expression, then removed mitochondrial genes (as well as few other batch-prone genes), and further filtered cells with $\leq 500$ UMIs (cell barcodes with less than 500 UMIs usually correspond to empty 10x droplets exposed to ambient RNA). In the analysis resulting in Figure 4 and EDFs 7-10, cells were also filtered if they were suspected to be contaminated by platelets, neutrophils or erythrocytes, as described in the Methods section. Few individuals were excluded from certain analyses due to lack of sufficient data. These instances are reported in all relevant figures, their accompanying legends and text. |
| Replication     | We ran technical replicates on 39 individuals and biological replicates on a follow-up cohort of 20 individuals, sampled approximately one year following their original sampling date. These are discussed and analyzed in the text. Specifically, the 20 follow-up replicates exhibited stable CLP and MEBEMP frequencies (Fig 2C).                                                                                                                                                                                                                                                                                                                                                                                                               |
| Randomization   | Simultaneously processed human samples were randomly assigned to different pools in order to minimize batch effects and single cells were randomly captured prior to sequencing.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Blinding        | Not relevant, as readouts were quantitative and not prone to subjective judgment.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |

## 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                                  |
|-------------------------------------|--------------------------------------------------------|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Antibodies         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants                        |

### Methods

| n/a                                 | Involved in the study                           |
|-------------------------------------|-------------------------------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

## Antibodies

|                 |                                                                                              |
|-----------------|----------------------------------------------------------------------------------------------|
| Antibodies used | The EasySep Human CD34 Positive Selection Cocktail uses a class II anti-CD34 antibody clone. |
| Validation      | Available from the manufacturer                                                              |

## 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, mosaicism, off-target gene editing) were examined.                                                                                                                                                                                                                                       |