nature portfolio

https://doi.org/10.1038/s41586-024-08250-x

## Supplementary information

# Long-term lineage commitment in haematopoietic stem cell gene therapy

**Long-term lineage commitment in hematopoietic stem cell gene therapy patients**

Andrea Calabria[1,@], Giulio Spinozzi[1], Daniela Cesana[1], Elena Buscaroli[2], Fabrizio Benedicenti[1], Giulia Pais[1], Francesco Gazzo[3,1], Serena Scala[1], Maria Rosa Lidonnici[1], Samantha Scaramuzza[1], Alessandra Albertini[1], Simona Esposito[1], Francesca Tucci[1,4], Daniele Canarutto[1,4,5], Maryam Omrani[1], Fabiola De Mattia[1], Francesca Dionisio[1], Stefania Giannelli[1], Sarah Marktel[1,4], Francesca Fumagalli[1,4], Valeria Calbi[1,4], Sabina Cenciarelli[1,4], Francesca Ferrua[1,4], Bernhard Gentner[1], Giulio Caravagna[2], Fabio Ciceri[4], Luigi Naldini[1,5], Giuliana Ferrari[1,5], Alessandro Aiuti[1,4,5], and Eugenio Montini[1,@]

[1]San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), IRCCS San Raffaele Scientific Institute, Milan, Italy

[2]Department of Mathematics, Informatics and Geosciences. University of Trieste.

[3]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[4]Pediatric Immunohematology and BMT, San Raffaele Hospital, Milan, Italy

[5]Vita Salute San Raffaele University, Milan, Italy

[@] Correspondence should be addressed to: Andrea Calabria (calabria.andrea@hsr.it) and Eugenio Montini (montini.eugenio@hsr.it)

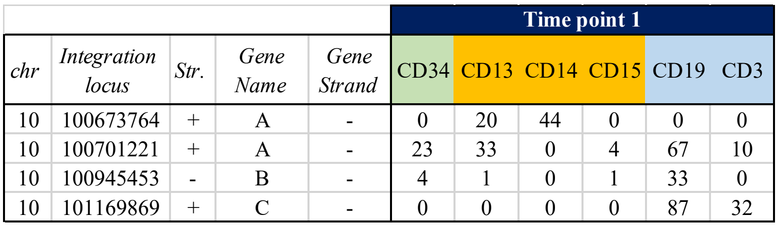
## Supplementary Information

### Table of Contents

# Supplementary Methods

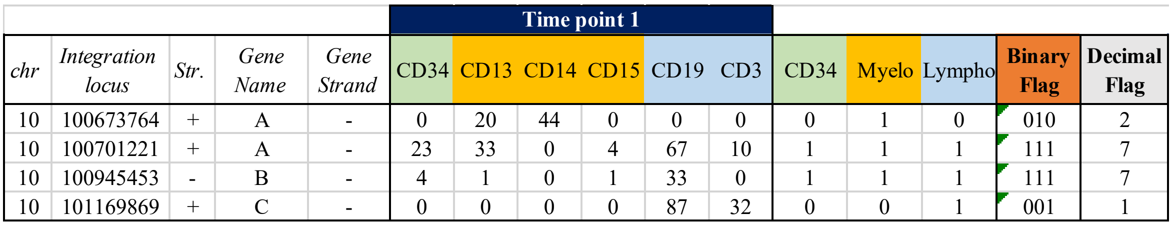
## HSC lineage commitment

Here we describe the computational approach to analyze HSC lineage commitment. On the computational side, we approached this problem by generating a binary table with all the possible combinations of each intersection for each time point. For instance, the table below shows a case of 4 ISs (in rows) observed in 6 cell markers corresponding to CD34+ cells, myeloid, and lymphoid cells, with the number of genomes as values. For simplicity, we will use only 3 labels here to represent intersection cases: “CD34,” “myeloid,” and “lymphoid.”



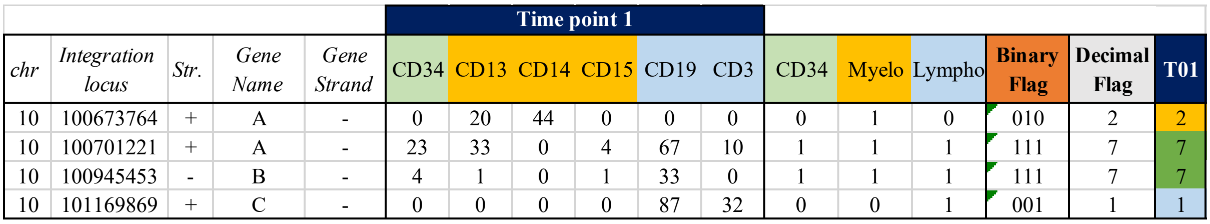
| | | | | | Time point 1 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *chr* | *Integration locus* | *Str.* | *Gene Name* | *Gene Strand* | CD34 | CD13 | CD14 | CD15 | CD19 | CD3 |
| 10 | 100673764 | + | A | - | 0 | 20 | 44 | 0 | 0 | 0 |
| 10 | 100701221 | + | A | - | 23 | 33 | 0 | 4 | 67 | 10 |
| 10 | 100945453 | - | B | - | 4 | 1 | 0 | 1 | 33 | 0 |
| 10 | 101169869 | + | C | - | 0 | 0 | 0 | 0 | 87 | 32 |

For this time point, we generate intersections by grouping the cell markers by label and then converting the number of genomes into a binary format.



| | | | | | Time point 1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *chr* | *Integration locus* | *Str.* | *Gene Name* | *Gene Strand* | CD34 | CD13 | CD14 | CD15 | CD19 | CD3 | CD34 | Myelo | Lympho | **Binary Flag** | **Decimal Flag** |
| 10 | 100673764 | + | A | - | 0 | 20 | 44 | 0 | 0 | 0 | 0 | 1 | 0 | 010 | 2 |
| 10 | 100701221 | + | A | - | 23 | 33 | 0 | 4 | 67 | 10 | 1 | 1 | 1 | 111 | 7 |
| 10 | 100945453 | - | B | - | 4 | 1 | 0 | 1 | 33 | 0 | 1 | 1 | 1 | 111 | 7 |
| 10 | 101169869 | + | C | - | 0 | 0 | 0 | 0 | 87 | 32 | 0 | 0 | 1 | 001 | 1 |

This binary format is then transformed into a binary string (“binary flag”), which corresponds to a decimal number (“decimal flag”). Each flag is uniquely associated with a specific intersection set. We classified each intersection set using the following rule: if an IS is observed in only one set, it is classified as uni-lineage and labeled with the name of that set (e.g., myeloid or lymphoid); if an IS is observed across multiple mature lineages, it is classified as multi-lineage. Any intersection involving CD34$^+$ cells does not affect the classification. In this example, the last column (“T01”) would be added as follows:



| | | | | | Time point 1 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *chr* | *Integration locus* | *Str.* | *Gene Name* | *Gene Strand* | CD34 | CD13 | CD14 | CD15 | CD19 | CD3 | CD34 | Myelo | Lympho | **Binary Flag** | **Decimal Flag** | **T01** |
| 10 | 100673764 | + | A | - | 0 | 20 | 44 | 0 | 0 | 0 | 0 | 1 | 0 | 010 | 2 | 2 |
| 10 | 100701221 | + | A | - | 23 | 33 | 0 | 4 | 67 | 10 | 1 | 1 | 1 | 111 | 7 | 7 |
| 10 | 100945453 | - | B | - | 4 | 1 | 0 | 1 | 33 | 0 | 1 | 1 | 1 | 111 | 7 | 7 |
| 10 | 101169869 | + | C | - | 0 | 0 | 0 | 0 | 87 | 32 | 0 | 0 | 1 | 001 | 1 | 1 |

We used this flag representation for each time point to assess lineage commitment for individual clones, reporting the percentage for each time point. A time course representation example is shown in the following table:

| *chr* | *Integration locus* | *str.* | *Gene Name* | *Gene Strand* | T01 | T03 | T06 | T09 | T12 | T18 | T24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 100673764 | + | A | - | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 10 | 100701221 | + | A | - | 7 | 3 | 1 | 4 | 2 | 0 | 2 |
| 10 | 100945453 | - | B | - | 7 | 0 | 0 | 3 | 0 | 1 | 0 |
| 10 | 101169869 | + | C | - | 1 | 1 | 0 | 3 | 1 | 3 | 7 |

This approach allowed us to classify each IS over time. To address potential sub-sampling issues at specific time points or in certain samples, which could affect the multi/uni-lineage labels, we refined the labels by analyzing lineage tracking over time as follows:

- Rule 1: Since we aim to observe progressive lineage commitment over time, any fluctuations in labels from uni-lineage to multi-lineage can be smoothed out. Specifically, if an IS was labeled as uni-lineage at a time point between two multi-lineage time points, we reclassified the uni-lineage label as multi-lineage. For example, in the previous case, T06 initially shows a uni-lineage lymphoid label, followed by a multi-lineage label, and then by two uni-lineage myeloid labels. Here, we converted the uni-lineage lymphoid label to multi-lineage.

| *chr* | *Integration locus* | *str.* | *Gene Name* | *Gene Strand* | T01 | T03 | T06 | T09 | T12 | T18 | T24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 100673764 | + | A | - | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 10 | 100701221 | + | A | - | 7 | 3 | 7 | 4 | 2 | 0 | 2 |
| 10 | 100945453 | - | B | - | 7 | 0 | 0 | 3 | 0 | 1 | 0 |
| 10 | 101169869 | + | C | - | 1 | 1 | 0 | 3 | 1 | 3 | 7 |

- Rule 2: To avoid misassigning a uni-lineage label to multiple mature lineages over time, we set the label as multi-lineage if an IS was labeled as uni-lineage but in different lineages at different times. In the earlier example, this rule would adjust the labels for the last IS at time points T01 and T03 as follows:

| *chr* | *Integration locus* | *str.* | *Gene Name* | *Gene Strand* | T01 | T03 | T06 | T09 | T12 | T18 | T24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 100673764 | + | A | - | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 10 | 100701221 | + | A | - | 7 | 3 | 7 | 4 | 2 | 0 | 2 |
| 10 | 100945453 | - | B | - | 7 | 0 | 0 | 3 | 0 | 1 | 0 |
| 10 | 101169869 | + | C | - | 3 | 3 | 0 | 3 | 1 | 3 | 7 |

## Multivariate Bayesian regression

In a *linear model*, the response variable $y$, is defined as a linear combination of input features **X**

(1) $y_i = \omega_0 + \sum_{d=1}^{D} \omega_d x_{i,d} + \sigma_i, i = 1, \ldots, N.$

Here, $y$ is a $N$-dimensional vector of target values, $X$ is a $N \times D$ matrix of input predictors, $\omega = (\omega_1, \ldots, \omega_D)^T$ and $\sigma$ represents the source of random error in the model (here Gaussian with mean 0 and variance 1). In *Bayesian linear regression*, instead, the target variable $y$ is assumed to follow a Gaussian distribution with mean parameter defined by equation (1), the linear combination of the input data **X** and the set of latent coefficients $\omega$ and variance $\sigma$, representing the noise of the model. The likelihood function of the Bayesian model is then defined as

(2) $p(y|X, \omega, \sigma) = \mathcal{N}(X\omega, \sigma)$,

where $\mathcal{N}$ is a Gaussian distribution with mean $X\omega$ and covariance $\sigma$.

The Bayesian model formulation introduces an implicit measure of uncertainty in the estimated parameters through the specification of prior distributions over the parameter space[1]. In this model the prior over the random errors $\sigma = (\sigma_1, \ldots, \sigma_N)$ is a Gaussian with zero mean and unit variance. We specify the prior over the latent coefficients $\omega$ as a zero-mean Gaussian distribution

(3) $p(\omega|\lambda) = \mathcal{N}(0, \Sigma^{-1})$,

with diagonal covariance matrix $\Sigma^{-1}$, i.e., with diagonal entries $[\lambda_1, \ldots, \lambda_D]$, and the prior distribution over $\lambda_i$ defined as a Gamma distribution. Each covariate coefficient $\omega_d, d = 1, \ldots, D,$ has therefore its own standard deviation $\frac{1}{\lambda_d}$. This method is called Automatic Relevance Determination (ARD), and is relevant in the Machine Learning field[1].

## Good-Turing estimator

The Good-Turing formulation[2] has been extensively used in ecology to estimate the bias between the observed and true number of species in a specific area. If we treat each IS as a species and cell markers as distinct assemblages, we can apply the same formulation to refine the richness of ISs in distinct cell markers and the count of shared ISs between CD34 and each marker, accounting for undetected species.

**One-assemblage formulation.** In the single assemblage formulation, we assume the presence of $S$ species with their true relative abundances. In a sample of $N$ individuals selected with replacement, $X_i$ represents the species frequency for the $i-th$ species in the samples, with $i = 1, \ldots, S$ and $\sum_{X_i>0} X_i = N$. Species with frequency $X = 0$ exist in the assemblages but go undetected, hence are not included in the sample data.

We define $S_{obs}$ the number of species observed in the sample and $f_0$ as the count of undetected species. Thus, we can express the true number of species as the sum of the observed and unobserved ones, such that

(4) $S = S_{obs} + f_0$.

The value of $f_0$ is unknown and in the Good-Turing formulation a lower bound is estimated solely based on the rarest observed species, as

(5) $\hat{f_0} \geq \frac{N-1}{N} \frac{{f_1}^2}{2f_2}$,

where $f_1$ and $f_2$ represent the number of singletons and doubletons, respectively.

**Two-assemblages formulation.** The one-assemblage formulation can be extended to estimate the number of shared species between two assemblages. Assuming there are $S$ species in a combined assemblage, we define $S_{shared} = S_{12}$ as the true count of species observed in both assemblages. By collecting random samples of $N_1$ and $N_2$ individuals from each assemblage, we can determine the observed shared species, $S_{shared,obs}$, and the observed abundance for each species in each assemblage. Let $f_{rv}$ represent the number of shared species observed $r$ times in the first assemblage and $v$ times in the second one. Similarly, $f_{r+}$ denotes the number of shared species observed $r$ times in the first sample and at least once in the second one, with similar symmetry for $f_{+v}$. Also, let $f_{++}$ be the count of shared species observed at least once in both samples.

Similar to the one-assemblage formulation, we can define the true number of shared species, accounting for species undetected in one or both assemblages, as

(6) $S_{shared} = S_{shared,obs} + f_{0+} + f_{+0} + f_{00}.$

Using the Good-Turing formulation, we can derive an approximation to the unknown elements of the formula ($f_{0+}$, $f_{+0}$ and $f_{00}$) as

(7) $\widehat{f_{0+}} \geq \frac{N_1-1}{N_1}\frac{{f_{1+}}^2}{2f_{2+}},$

(8) $\widehat{f_{+0}} \geq \frac{N_2-1}{N_2}\frac{{f_{+1}}^2}{2f_{+2}},$

(9) $\widehat{f_{00}} \geq \frac{N_1-1}{N_1}\frac{N_2-1}{N_2}\frac{{f_{11}}^2}{4f_{22}}$

**Application.** Using the Good Turing estimators, our goal is to approximate the count of undetected species based on the occurrences of the rarest detected species and population sizes. In the formulation for a single assemblage, the count of undetected species, $f_0$, is solely determined by the population size ($N$) and the number of singletons ($f_1$) and doubletons ($f_2$), representing species occurring once and twice, respectively. In the two-assemblages formulation, our objective is to adjust the number of species shared between two populations. The observed count of shared species is corrected by considering species unobserved in both populations ($f_{00}$), those missing in the first assemblage but observed in the second ($f_{0+}$), and likewise those detected in the first assemblage but missing in the second ($f_{+0}$). To estimate these unknowns, we rely on the occurrences of shared species observed once and twice in the first sample, and at least once in the second one ($f_{1+}$ and $f_{2+}$), with similar symmetry for $f_{+1}$ and $f_{+2}$, the count of shared species observed exactly once and twice in both assemblages ($f_{11}$ and $f_{22}$), along with the population sizes ($N_1$ and $N_2$).

In the analyses of *linage output*, the evaluation of CD34$^+$ output is carried out in terms of the sharing ratio of ISs identified in CD34$^+$ and recaptured across mature cell markers. In this formula, the numerator reflects the count of species observed in both CD34$^+$ cells and each mature cell marker, while the denominator corresponds to the overall size of the CD34$^+$ population. An additional bias is introduced in this analysis due to the challenge of detecting rare species. To address this, we incorporated the Good-Turing estimator for shared species between two assemblages as the numerator, and the GT formula to estimate the richness of the CD34$^+$ population as the denominator of the sharing ratio. To compute the adjusted richness of the CD34$^+$ population, we gather, for each patient, the set of ISs observed in CD34$^+$ BM cells and their respective abundances over time. This vector indicates the number of times each IS is observed, representing the species frequency detected in the collected sample. This allows us to quantify the population size ($N$) as the sum of species abundances, and the number of singletons ($f_1$) and doubletons ($f_2$) as the ISs occurring once and twice. Another variable of interest is the number of shared species between each mature cell marker and CD34$^+$ BM cells, approximated via the two-assemblages GT formulation. The vector of ISs abundances observed in CD34$^+$ BM population is retrieved for each patient and summed over time, and similarly collected for each mature cell marker, time point, patient and tissue. These vectors are used to compute the unknowns of equation (6) ($f_{00}$, $f_{0+}$ and $f_{+0}$), providing an estimation of the number of shared species between each mature cell marker and CD34$^+$ cells. The adjusted sharing ratio is then employed as a response variable in the multivariate Bayesian linear regression model. Patients' heterogeneity with respect to age is analyzed by stratifying patients into three groups and fitting the regression independently, incorporating the age of the subjects as additional covariate in the model.

In the analysis of *lineage commitment*, each IS at every time point is uniquely categorized as multi- or uni-lineage. This is evaluated as a sharing ratio calculated for each time point, representing the ratio between the number of IS observed in a specific group and the total number of clones captured at that time. To address the challenge of unseen species, we applied the one-assemblage formulation of the GT estimator. For each patient, timepoint, and class (multi-lineage or one of the uni-lineage), we extract the class-specific vector of ISs and compute the overall abundance across markers. We quantify the population size ($N$), the number of singletons ($f_1$) and doubletons ($f_2$) from the vector of ISs frequency and apply the GT formulation to yield an adjusted value for the numerator of the sharing ratio. Similarly, for each time point we retrieve and sum over all cell markers the observed ISs abundances to estimate the number of captured ISs via GT. Subsequently, the computed sharing ratio undergoes correction for potential confounding factors through a multivariate Bayesian linear regression.

## Somatic mutations with Myeloid panel

Obtained sequences were aligned with BWA-MEM to the human reference genome (hg19/GRChg37), resulting in more than 100,000,000 reads correctly aligned on the targeted exon panel, with an average of 4400 and 4300 reads/base in β-Thal and MLD patients respectively, covering 278 amplicons among 38 genes (**Supplementary Table 5-6**). List of genes: ABL1, ASXL1, BCOR, BRAF, CALR, CBL, CEBPA, CSF3R, DNMT3A, ETV6, EZH2, FLT3, GATA2, HRAS, IDH2, IKZF1, JAK2, KIT, KRAS, MPL, MYD88, NF1, NRAS, PHF6, PRPF8, PTPN11, RB1, RUNX1, SETBP1, SF3B1, SH2B3, SRSF2, STAG2, TET2, TP53, U2AF1, WT1, ZRSR2.

We then generated the pileup files using Samtools (samtools mpileup) (options: -B -q 1) and we performed a variant calling on these files using VarScan2[3] (mpileup2snp: --min-coverage 100 --min-var-freq 0.01 --min-reads2 1 --p-value 0.05 --output-vcf 1, mpileup2indel: --min-coverage 100 --min-var-freq 0.01 --min-reads2 1 --p-value 9,1 --output-vcf 1). To remove the false positives, we applied the following filters: (1) removal of mutations present in more than one independent sample (patients); (2) removal of mutations present in low-covered amplicons (less than 200 reads); (3) removal of mutations clearly germline (heterozygous or homozygous, 49 < VAF < 51 or VAF > 99); (4) removal of mutations present in the last 3 bp of the reads; (5) removal of mutations in regions enriched in poly-T or poly-A (manually curated).

The average sequencing depth in β-Thal and MLD patients was 4,400 ± 283 and 4,300 ± 1789 reads/base respectively (**Supplementary Table 6**). Moreover, we removed mutations with a Varian Allele Frequency (VAF) suggestive of heterozygous or homozygous germline variants (49 < VAF < 51 or VAF > 99). Most somatic mutations (85 out of 96) exhibited a Variant Allele Frequency (VAF) of less than 2%. Overall, in β-Thal patients we found on average 7.1 ± 6 mutations (range 2 to 21) in 21 different genes. The detected mutations underwent annotation utilizing the Genome Aggregation Database (gnomAD)[4], the Database of Single Nucleotide Polymorphisms (dbSNP)[5] and the ClinVar database[6]. In MLD patients we found on average 1.5 ± 0.7 mutations (range 1 to 3) in 14 different genes. Out of the identified mutations, only 4 were annotated as known variants with no role as drivers in clonal hematopoiesis or cancer.

# Supplementary Tables

**Supplementary Table 1. Patients' summary.**

Summary table of treated patients under analysis, showing details of the disease, the ID of the patient ("Patient ID"), the body weight at treatment, the cell dose expressed as millions of CD34$^+$ cells 10/kg, the transduction efficiency, the GT source (expressed as "BM", "MPB" for mobilized PB, or "BM and MPB"), the VCN at treatment, the age of the patient at treatment, and the maximum follow up available with molecular data, and the gender.

**Supplementary Table 2. Patients' Samples and details with ISs.**

The table presents all details of each patient sample used in this work, using the following columns: Disease, Patient ID, Tissue of sample origin (PB and BM), Lineage, Cell Marker, Timepoint (expressed in months after GT), vector copy number (VCN), amount of genomic material (reported as ng of DNA), PCR Method for IS retrieval (between LAM-PCR and SLiM-PCR), sequencing technology and platform, N. reads supporting the ISs, N. ISs, population diversity index (using Shannon H index).

**Supplementary Table 3. Normalized integration frequency by gene.**

Pair-wise comparison of gene normalized integration frequency for MLD, WAS, and β-Thal patients, after homogeneous subsampling and randomization. For each gene (column "GeneName") and pair-wise comparison (column "Comparison") we reported the average number of ISs for study, and the final corrected p-value (column "corrected p-value avg").

**Supplementary Table 4. Estimated number of active HSPC.**

For each available patient, the table reports the number of estimated HSPCs observed at early time points (<24 months) and from 24 months (column "Pop. Size >24 m") calculated with the Chao1 model and corrected with the VCN (if VCN>1). We also reported the number of IS retrieved <24 months and >24 months, and the number of shared IS between the two sets. The last column reports the fraction of the estimated HSPCs >12 months on the total number of observed IS >24 months.

**Supplementary Table 5. Mean Depth Coverage of somatic mutations data.**

The table shows the mean depth coverage reached for each MLD and β-Thal patient coming from the Illumina's AmpliSeq™ Myeloid Panel Targeted exome sequencing used for the somatic mutations analysis. Columns report: Disease, Patient ID, time point in days of the sample, mean depth coverage.

**Supplementary Table 6. Somatic Mutations.**

The table presents all the somatic mutations for each MLD and β-Thal patient coming from the Illumina's AmpliSeq™ Myeloid Panel Targeted exome sequencing. Columns report: Disease, Patient ID, time point in days of the sample, cell type, total number of reads in the position of the mutation, number of reads supporting the reference variant, number of reads supporting the mutated variance, Variant Allele Frequency (VAF), chromosome, starting position, ending position, nucleotide reported by the reference, nucleotide reported by the mutation, functional annotation of the position (intron, exon), gene, exon modification, aminoacidic change, effect of the mutation reported by ClinVar, frequency of the mutation in the population reported by gnomAD, ID of the mutation reported by avsnp147.

**Supplementary Table 7. Primers used for the 1st (Exponential) PCR.**

List of primers used for the first exponential PCR with details about Name, Type, and Sequence

**Supplementary Table 8. Primers used for the 2nd (Fusion) PCR.**

List of primers used for the second fusion PCR with details about Name, Type, Sequence and Barcode (8 nucleotides) of each primer.

**Supplementary Table 9. Linker cassette sequences.**

The linker cassette sequences used in the SLiM-PCR procedure. Each linker cassette was generated by the annealing of the Short oligo with a Long oligo. The sequence of each oligo and its Barcode (8 nucelotides) is reported.

# Supplementary Discussion

## IS sensitivity

In our cohort of patients, IS were collected from samples with different characteristics in terms of the amount of DNA, VCN, PCR technologies, sequencing platforms, and other variables. Differential proliferation in specific cell subsets after transplantation can lead to varying clonal kinetics, which may affect the retrieval of ISs. To address potential subsampling and account for sample variability, we implemented mathematical models that accounted for confounding factors (Bayesian model) and recapturing probabilities in assemblages (Good Turing). Subsampling issues, which might affect the classification of an integration site (IS) as either multilineage or unilineage committed, can be addressed through accurate filtering procedures (e.g., based on clonal abundance) and evaluated using bootstrapping methods that provide confidence intervals for each observation. However, analyzing more data reduces the impact of subsampling biases. In fact, avoiding data sparsity enhances the accuracy of the results. Conducting multiple longitudinal samplings for each patient and incorporating multiple technical replicates in IS analysis can increase the confidence in observations, making them suitable for mathematical corrections and preventing extreme data rarefaction. The ability to track the clonal repertoire for several years after treatment has allowed us to study hematopoiesis in homeostatic conditions, as indicated by the paucity of newly retrieved ISs from most patients after reaching the plateau at 12-24 months post gene therapy.

# Supplementary References

1 Bishop, C. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. (2007).

2 Chao, A. *et al.* Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. *Ecology* **98**, 2914-2929, doi:10.1002/ecy.2000 (2017).

3 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

4 Gudmundsson, S. *et al.* Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **597**, E3-E4, doi:10.1038/s41586-021-03758-y (2021).

5 Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**, 352-355, doi:10.1093/nar/28.1.352 (2000).

6 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).