Peer Review File

Manuscript Title: Long-term lineage commitment in hematopoietic stem cell gene therapy patients

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Calabria and colleagues report a clonal tracking study after lentiviral hematopoietic stem and progenitor cell gene therapy treatment on 53 patients (MLD, WAS, and β-Thal) through the use of vector integration sites. With this method, the authors estimated the HSPC sizes after transplantation and investigated the clonal diversity, as well as lineage biases. The authors showed differences of clonal activity across different disease conditions and conclude that HSPCs acquire and retain a memory that influence different behaviors due to patient's underlying disease.

The authors provide a very precious resources to study HSPC clonal behaviors in humans in a transplantation context. The number of patients (53) and the time (up to 8 years) of follow up is impressive. However, I have some concerns regarding the major conclusion that HSPCs have preexisting memories dictated by a patient's clinical condition or genetic background, which could be confounded by several covariates. Addressing the potential confounders is recommended in order to be able to make these provocative conclusions. In addition, since the study is performed in a transplantation context, the interpretations of some biological insights, such as HSC numbers should be more cautious as this does not necessarily reflect the native physiologic state.

Here are specific comments:

- The HSPCs transduced by different gene expression vectors are used for different diseases. It would be helpful to provide more information of the different vectors used. Could the differences in HSPC features observed in this paper may be caused by the difference in the vectors used for different diseases?

- The author show that the insertion sites tend to be enriched in gene-dense regions. The author claimed high correlation of IS gene GO terms across different diseases. However, Extended Data Fig 1D does not seem to clearly indicate that this is the case. In addition, can the authors perform a direct comparison at the gene level or genomic region level across different diseases to investigate whether there is any insertion preference specific across disease conditions. The possibility that the HSPC clonal behaviors being biased by insertion preference needs to ruled out.

- There are a number of confounders in interrogating HSPC clonal behaviors in different disease conditions that need to be assessed and accounted for. This includes the use of different conditioning regimens, gender of patients, infused cell number, PCR methods for amplification, etc. It is recommended to have an overall model (such as a multivariate regression) to model and control potential confounders together.

- How do the authors control the labeling efficiency and the ability to detect ISs? For example, would the difference of PCR efficiency across different insertion context result in biases being observed potentially? This might particularly become an issue, if insertion context varies considerably across different disease contexts. By comparing across time points, how can the authors distinguish whether one barcode is technically not detected vs. it exisiting in quiescent clones that did not contribute to hematopoiesis until a later time point? What is the grey bar in Extended data fig1F attempting to show? Are those non-recaptured clones?

- The authors discuss HSC lineage biases. Can the author provide evidence to justify the robustness in defining the uni-lineage vs multi lineage clones? Would it be possible that the “multilineage” clones are the one with overall better detection, while the “uni-lineage” have more dropout?

Referee #2 (Remarks to the Author):

Reviewer Comments: Long-term lineage commitment is modulated by the underlying disease in hematopoietic stem cell gene therapy patients

Summary of Manuscript: This manuscript compares the clonal outputs of lentivirally gene corrected hematopoietic stem cells (LGC-HSCs) in gene therapies for three congenital diseases: Metachromic Leukodystrophy (MLD), Wiskott-Aldrich Syndrome (WAS) and beta-Thalassemia (BTHAL). Impressively, it analyzes more than 6,700 peripheral blood (PB) and bone marrow (BM) samples from 53 patients up to 8 years after treatment.

The authors note that the number of engrafted, long-term LGC-HSCs is positively correlated with the number of infused CD34+ cells. Importantly, they assert that they do not see any evident plateau for the total number of LT-LGC-HSCs [at least not over the range of dosages used in these trials].

From their analyses, the authors conclude that in all disease conditions 50% of clones demonstrate multilineage potential. They assert that the remainder show preferential lineage commitment that is specific to the disease condition. The authors hypothesize that this is due to LT-LGC-HSC retaining “memory” of pre-gene therapy cell states.

Major Points:
1. There are several technical confounders that could potentially mimic lineage skewing and therefore

deserve more careful evaluation and discussion.

Confounders include:
(A) Gene-therapy did not in all cases fully correct initial disease conditions.

In the BTHAL trial (Marktel et al, 2019) all three adult and one of the pediatric patients continued to be transfusion dependent. The other evaluable children remained anemic. This would suggest that all patients continued to have a strong erythropoietic drive and that factors extrinsic to the LT-LGC-HSCs may have influenced lineage skewing.

Likewise in the WAS trial (Ferrua et al, 2019) patients generally remained thrombocytopenic after therapy which also may have extrinsically influenced lineage skewing.

Mitogen (EPO/TPO) levels might add useful information. At the very least a fuller disclosure in the text and caveats to the conclusion would be reasonable.

(B) Differences in input DNA amounts impact the sensitivity to detecting clones and has the potential to mimic lineage skewing.

Offering specifics about the amount of input DNA for all samples is important. Concomitantly, it is essential to detail any corrections in the inference of the number of HSCs (e.g. via the sample-size-based rarefaction/extrapolation formulae for estimating diversity from a sample of a single assemblage) or that were applied to estimate overlaps in clonal outputs (e.g. via the Good-Turing estimators for the number of species shared between two assemblages).

This point may be particularly pertinent to the analyses of the sharing ratio, where corrections due to sample size can be important.

(C) Comparisons of IS from cells where clones are geographically segregated (i.e. from the BM) can result in misinterpretions of lineage potential.

Erythroid, Myeloid and B cells in BM are locally produced and the clones are geographically separated for a time after therapy. In primates it can take up to 2 years for clonal geographic segregation to disappear (Verovskaya et al, JEM 2014; Wu et al, JEM, 2018; Chung et al, Blood 2018). Comparisons to cells from contaminating blood or T cells which develop outside the BM may lead to erroneous interpretations with regard to lineage skewing.

(D) Misidentification of multi-insertion clones as uni-insertion clones can result in both (a) overcounting of inferred number of HSC clones and (b) to the extent that multi-insertion clones have low-prevalence and concomitant less complete recovery of all ISs, overcounting of lineage-restricted clones.

Essential are a more complete description of the number of vector insertions per HSC along with an

explanation of any corrections applied to the computation of the number of HSCs and their lineage restriction.

2. Beyond technical issues affecting whether clonal output is truly skewed, there are several other mechanisms besides LT-LGC-HSC intrinsically retaining “memory” of pre-gene therapy cell states that could result in putative lineage skewing.

Plausible mechanisms include:
(A) Persistence of the pre-gene therapy environment for hematopoiesis

See point 1(A) above.

(B) Differences in the conditioning regimens and their resultant effects on the hematopoietic environment

The authors note in that different conditioning regimens were used in the therapies for different diseases. A lymphodepleting regimen was given to WAS patients; consequently there was more rapid “filling” of the lymphoid compartment with LGC cells. Filling may have originated from ST-HSPCs that did not persist, thus producing ‘uni-lineage’ T cells and separate from LT-HSPCs. Once filled homeostatic proliferation maintains T cell clones independent of on-going production from LT-HSPCs. By contrast, much slower refilling of the T cell compartment occurred in BTHAL (where thiotepa and treosulfan were used for conditioning) and MLD (where busulfan was used for conditioning).

This point deserves fuller disclosure in both the abstract and the conclusions.

(C) Persistence of heritable epigenetic changes intrinsic to HSPCs, which arose prior to treatment

Specific mechanisms might include (a) intrinsic disparities in the rates by which HSCs differentiate to specific lineages versus (b) intrinsic differences in the proliferation rates of committed progenitor states. If this is a characteristic of HSPCs independent of their environment, this might be apparent in in vitro differentiation and proliferation studies.

A more definitive investigation would include either bulk ATAC-seq on HSPCS and selected subsets or scATAC-seq.

(D) [Probably less likely] persistent changes in the cells comprising the hematopoietic niche; either because these have remained uncorrected or due to their exposure to the original pre-gene therapy environment.

Although a less likely explanation, it could be investigated after other possibilities have been ruled out.

3. It would be of interest to show abundance by various classes of clone.

Beyond the computation of Shannon indices, remaining analyses focus on binary (absence of presence within a lineage) classifications of clones. By lineage abundance fractions offer useful insights and are typically used in murine primate studies. For instance, are the uni-lineage clones small and therefore might sampling be an issue?

Minor Points:
1. Please note explicitly in the text (rather than figure legends and supplemental text) that HSPC number inference was done with VCN corrections.
2. For the one WAS and two BTAHL patients with dramatically greater numbers of cumulative ISs, it would be useful to also see a plot of the proportion that were detected over the long-term.
3. Why is 24 months regarded as long-term/stable in many analyses (e.g. Figure 2), but for numbers of HSCs and ISs and their comparisons in Table S2, 12 months is used as the cut-off between short- and long-term. 24 months appears more relevant from the data in the rest of the paper.
4. The word “cell” is confusingly used synonymously to individual IS (e.g. figure 2A and in the methods supplement)

Summary Reviewer Opinion: This is an important paper, based on a huge, long-term and rich dataset that will be of interest to both gene therapy and hematopoiesis investigators. Important conclusions include the observation that, for all disease conditions, the number of active HSPCs is positively correlated to the dosage of CD34+ cells without evident plateau. The hypothesis that the prior disease condition imprints LT-HSCs is interesting but requires more careful analysis.

Referee #3 (Remarks to the Author):

This manuscript by A. Calabria et al analyzes a large and detailed dataset for assessing the safety and post-transduction kinetics of engraftment and stable hematopoiesis after lentiviral gene therapy for hereditary disorders. Integration site analysis was used to characterize these dynamics as they relate to the diversity and lineage-specificity of engrafting clones, analyzed in samples collected over nearly a decade of follow-up. The authors report an intriguing finding, which is that the underlying disease appears to influence expansion of the transduced rescued lineage, which is influenced more broadly by patient age at treatment, VCN and transduction efficiency, and the tempo of hematopoietic reconstitution. This is an important report that will be very useful to understanding the dynamics of hematopoiesis and safety after gene therapy in hematopoietic stem cells. While there are limitations in the report regarding the mechanistic underpinnings of these observations and extrapolating the data to predict clinical outcomes from baseline characteristics in the patient and/or in drug product, this is probably only the beginning of a very important story.

Major comments:

With regard to the late appearing IS's >24 month post-infusion, particularly in the older thalassemia patients, was there any evidence that these emerged/were recruited in the setting of hematopoietic stress in which a proliferative stem cell expansion might be triggered in lieu of quiescence? In other studies of gene therapy for thalassemia, it has been observed that features of stress erythropoiesis persist even after establishing RBC transfusion independence (skewed M:E ratio in the marrow favoring erythroid progenitors, persistently elevated markers of ineffective erythropoiesis, etc). The question of exhaustion of true HSCs following cell proliferation signals in this setting is also unclear. The authors argue that early appearance of ISs post-infusion and their drop-out indicates these were HSPCs and not true HSCs but it is possible HSC exhaustion and drop out has not been excluded, particularly in the period of recovery and rapid expansion that follows pre-infusion myeloablation/conditioning. This would tend to select a smaller subset of clones with better proliferative activity. It is also sobering to observe that a very small number of true HSCs ultimately establish steady-state hematopoiesis, under conditions that would appear to select clones with robust proliferative capacity. This also raises the question if stochastic events might skew abundance of some clones, simply because they are more proliferative and then enriched further by way of natural selection of a particular lineage, as occurs in thalassemia for example, where in allogeneic HCT donor-host chimerism also favors enrichment of corrected donor cells in erythroid progenitors? A similar analysis would find an erythroid skewing of CD34+ cells of donor origin, even when there is a minority of donor cells.

One strength of the analysis with its careful assessment of the timing and persistence of clones post-infusion is that the emergence of stable clones established from true HSCs might take up to 24 months post-infusion to emerge. This indicates the importance of a longer period of follow up in these patients is needed before stable hematopoiesis from transduced HSCs can be ascertained. The new information will be enormously valuable to gene therapy teams and in directing long-term followup assessments. Because there relative few true HSCs contributing to steady-state hematopoiesis, the potential for clonal hematopoiesis must be monitored over the long-term, probably decades.

While the kinetics of clonal hematopoiesis and the size of this cohort across three disparate hereditary disorders is very impressive, the mechanistic basis for the phenomenon observed – a clonal bias favoring a particular lineage over another – has not yet been defined, although it will be critical to do so. While perhaps beyond the scope of this study, an obvious question is whether there are epigenetic marks in lineage specific loci that might establish and favor the expansion of a single lineage from these HSCS? Does the lentiviral vector tropism for integration near chromatin and histone-modification loci favor the re-capitulation of chromatin configurations in the HSCs that direct lineage differentiation? Would it be useful to conduct a study of snRNA-seq to better delineate the progenitor populations as these expand after engraftment? The manuscript would have been improved by including some of these studies (if feasible since snRNA-seq would require fresh marrow samples) to better understand mechanistic underpinnings of these very interesting observations.

In Fig 4C, the sharing ratio of B and T-lineages in WAS was not as significant as the sharing ratio of the erythroid lineage observed in thal. In fact, the sharing ratio significance was not prominent in B & T lineages between WAS and thal. Does this indicate that marking and enrichment for T and B cells in WAS

was not as strong as erythroid selection pressure in thal? It would be interesting to evaluate the sharing ratio in GT recipients with X-SCIDs, in whom the sharing ratio for T-lineage might be especially pronounced. If observed, this would suggest the strength of the natural selection for corrected clones might be predicted to follow the impact of the mutation. Or is this finding simply reflective of the transduction efficiency in WAS (70 – 90% LVV+) compared with thal (30 – 77%) as shown in Table 1. This would tend to exert a stronger selection in the minority of erythroid progenitors with the transgene compared with residual cells and cells from drug product lacking vector, as both the latter populations will be prone to ineffective erythropoiesis and apoptosis. This was also reflected in the older thal patients having lower VCN/%LVV+ HSPCs with higher active HSPCs. This is supported by the association with transduction efficiency depicted in the PCA in Fig 3B.

Minor comments:

Line 147 – this appears to be missing a statement that the erythroid lineage had higher clonal complexity in thalassemia compared with the other 2 disorders.

Line 180 – was the 'depth' of myeloablation more complete in thalassemia and MLD than in WAS recipients, accounting for the larger drop-off IS's compared with estimated HSPCs in the steady-state phase. Might this also be related to lower numbers of long-term HSCs in MLD and thal, where selection of the most-fit proliferative clones under the stress hematopoiesis with engraftment might have occurred?

Line 240 – it would be very interesting to determine if the older patients with thal in whom the lineage sharing of CD34+ cells with erythroid cells was most striking also had driver mutation SNPs characteristic of clonal hematopoiesis. It is acknowledged that IS clonal expansion was not observed in this analysis, but age-driven accumulation of driver mutations is a recognized phenomenon, and might occur in thal as appears to be the case in sickle cell disease.

Author Rebuttals to Initial Comments:

Revisions – Point by point reply to referees:

Calabria A et a., *Long-term lineage commitment is modulated by the underlying disease in hematopoietic stem cell gene therapy patients.*

General answer to Reviewers and Editor

We are grateful for considering our work of potential interest and the positive Reviewers' comments on the importance and quality of our work:

As quoted by Reviewer 1: "*The authors provide a very precious resource to study HSPC clonal behaviors in humans in a transplantation context. The number of patients (53) and the time (up to 8 years) of follow up is impressive".*

By reviewer 2: *"This is an important paper, based on a huge, long-term and rich dataset that will be of interest to both gene therapy and hematopoiesis investigators. Important conclusions include the observation that, for all disease conditions, the number of active HSPCs is positively correlated to the dosage of CD34+ cells without evident plateau".*

And by Reviewer 3: "*This is an important report that will be very useful to understanding the dynamics of hematopoiesis and safety after gene therapy in hematopoietic stem cells*".

Moreover, we are grateful for the constructive suggestions on how to improve our analyses and thus eventually reinforce or disprove our claims. These suggestions included the application of models to correct for confounding factors, and additional comparisons to evaluate and correct for specific technical biases. As you will appreciate in the point-by-point reply to Reviewers we implemented several novel analyses and included additional experimental data that allowed us to further strengthen our claims.

General reply to all Reviewers:

1) To better compare the datasets with different numerosity and heterogeneous clonal abundances we applied the Good-Turing frequency estimator. To reduce the impact of technical confounders we applied a Bayesian multivariate linear regression model that considers multiple technical confounding factors and the possible interactions between variables simultaneously. The variables (confounding factors) included: the PCR method, amount of DNA used, dose of $CD34^+$ infused per Kg, vector copy number, sequencing depth, patient's gender, and age. We added entire new methodological sections and reviewed all the analyses (applied to new **Figure 1C, Figures 2 and 3 (and the related Extended Data Figures 2, 5, 6 and 9).**

   These additional corrections and modeling of confounding variables resulted in relatively minor changes in HSPC output and commitment analyses. The only noteworthy difference between our previous and novel results was that the $CD34^+$ output towards B-cells in WAS patients was reduced compared to the results before correction. This observation is in line with the notion that the selective advantage provided by WASP expression in T cells is stronger than B cells as reported in previous studies, which further reinforced the confidence in our observations.

2) We devised several novel analyses to understand if differences in the genomic integration profile among clinical programs and if differences in clonal abundance could bias the lineage output or commitment. None of these factors appear to significantly impact our analyses. We added a new extended table to include these results per gene (**Extended Data Table 2).**

3) We devised novel analyses to further investigate the dynamics of lineage commitment over time at single clone resolution, in which long-lived clones identified at early and late phases of hematopoietic reconstitution (<24 months and >24 months respectively) were selected and classified into clones that transitioned from multilineage to uni-lineage, or that were found to be persistently uni-lineage committed or persistently multi-lineage (Multi-Multi) since the early to late phases of hematopoietic reconstitution. This single clone level analysis allowed us to compare if and how the different disease conditions impact the rate of lineage commitment over time as well as the relative contribution of long-lived clones already committed since the early phases of hematopoietic reconstitution. The results indicate about 1/3 of uni-lineage committed clones originates from multilineage clones, while the remaining 2/3 were identified since the early phases of hematopoietic reconstitution. These data suggest that uni-lineage committed clones are in part produced under a disease-specific "pressure" during

reconstitution and likely by HSPCs already committed before transplantation. All these analyses are reported in the Results section and supported by the **new Figure 3E-I (and related Extended Data Figures 7 and 8)** and the new methodological sections.

4) We expanded our analyses of lineage output and commitment to **10 additional SCID-X1 (XSCID) HSPC GT patients** treated in another institution (*De Ravin S., et al Nature Communications 2022*). These datasets comprised ISs retrieved overtime (max follow-up of 84 months) from $CD34^+$ cells (16,650 IS), $CD14^+$ myeloid cells (16,640 IS), $CD3^+$ T-cells (95,362 IS) and from $CD19^+$ B-cells (58,317 IS), for a total of 186,969 IS. Our analyses showed that, like WAS, also XSCID patients the HSPC output was skewed towards the lymphoid lineage together with a pronounced uni-lineage T-cell commitment and for B-cells although to a lesser extent. We added the results of these analyses in the Results section supported by the new **Extended Data Figures 5C and 9.**

5) The finding that the marked HSCP output and commitment towards lymphoid and erythroid lineages observed in WAS and β-Thal patients respectively, remained stable over time, even after years after GT, suggested that an increased "pressure" on HSPCs to specifically "replenish" these cell compartments is still maintained.

6) A possible explanation is that WAS patients, despite the overall undeniable beneficial effects of the therapy, remained thrombocytopenic, indicating the lack of a full normalization of the hematopoietic defects in this disease. Similarly, among β-Thal patients, despite the improved production of hemoglobin, the three adults and one of the pediatric patients remained transfusion dependent and in the three remaining pediatric patients, although being transfusion independent, some signs of ineffective erythropoiesis were still present. Thus, the pathophysiological condition before and after therapy may have altered the microenvironment and the concentration of specific cytokines to instruct the HSPCs to produce and to commit long-term towards the most needed lineages. To address these hypotheses, **we performed a longitudinal analysis of thrombopoietin (TPO) and erythropoietin (EPO) levels in WAS and β-Thal GT patients** respectively. TPO levels in WAS patients before GT were near the normal levels, which was expected because WAS patients before therapy were under TPO receptor agonist treatment. However, **TPO levels increased at 1 year after GT, regardless of the age at treatment, and then decreased to almost normal levels at 4 years from GT, suggesting that the beneficial effects of the therapy in these patients may have alleviated the need for enhanced production of this important cytokine albeit not entirely.** In 4 out of 6 β-Thal patients transplanted at ages ranging from 4 to 13 years, EPO levels were higher compared to normal levels at all time points

analyzed (up to 3 years from GT). On the other hand, 2 out of the 3 β-Thal patients treated at age >30 years who remained transfusion dependent and showed the highest output toward the erythroid lineage, showed on average a 3-5 fold higher EPO level compared to the younger cohort, especially at 1 year from GT. Moreover, 2 β-Thal transfusion independent pediatric patients showed EPO normal levels and a marked erythroid commitment. **Thus, EPO levels negatively correlate with the therapeutic outcome of the therapy, albeit partially, suggesting that other factors could modulate the behavior of HSPCs in β-Thal.** Overall, our data agree with the notion that HSPCs activity is modulated to better respond to the demands imposed by the specific pathological condition. The results of this analysis are illustrated in the new **Figure 4**.

7) Intrigued by the recent finding that sickle cell disease (SCD) patients have an increased frequency of potential driver mutations associated with myeloid cancer or clonal hematopoiesis (*Spencer Chapman M. et al., Nature Medicine 2023*), we wanted to address whether if this was also the case in β-Thal patients. Thus, **we performed an exhaustive analysis of somatic mutations in exons of 40 genes involved in clonal hematopoiesis and myeloid cancer in the 9 β-Thal patients and 23 MLD patients**. The search for somatic mutations was performed on genomic DNA from $CD34^+$ cells prior infusion and PBMCs harvested at 2 years and >5 years after transplantation from our cohort of adult and pediatric β-Thal and MLD patients. We found that somatic mutations do not accumulate over time and no somatic mutations, known to drive clonal hematopoiesis or myeloid cancer, were found in any patient. However, **β-Thal patients showed a significantly higher somatic mutation rate (>7 fold) than MLD patients**. This finding suggests that **the increased frequency and accumulation of somatic mutations β-Thal patients are not directly associated with the HSC gene therapy treatment itself but rather with the intense and progressive hematopoietic stress inherent in this disease**. Overall, these findings are in line with the notion that in hemoglobinopathies such as SCD and β-Thal, the prolonged proliferation stress caused by the need to produce erythroid cells and the increased oxidative stress caused by iron overload result in increased levels of genomic and mitochondrial DNA damage, telomere erosion, cellular senescence, and progressive damage of the bone marrow niche. We extended the Results section and added **Figure 5** and **Extended Data Table 4 and 5** to support our findings.

Besides the abovementioned general points, we answered all remaining questions raised by the Reviewers, as detailed in the point-by-point reply to the Reviewers.

Please note that the changes in the main test and figure legends are in blue.

Point by point reply to Reviewers.

Referee #1 (Remarks to the Author):

Calabria and colleagues report a clonal tracking study after lentiviral hematopoietic stem and progenitor cell gene therapy treatment on 53 patients (MLD, WAS, and β-Thal) through the use of vector integration sites. With this method, the authors estimated the HSPC sizes after transplantation and investigated the clonal diversity, as well as lineage biases. The authors showed differences of clonal activity across different disease conditions and conclude that HSPCs acquire and retain a memory that influence different behaviors due to patient's underlying disease.

The authors provide a very precious resources to study HSPC clonal behaviors in humans in a transplantation context. The number of patients (53) and the time (up to 8 years) of follow-up is impressive. However, I have some concerns regarding the major conclusion that HSPCs have preexisting memories dictated by a patient's clinical condition or genetic background, which could be confounded by several covariates. Addressing the potential confounders is recommended in order to be able to make these provocative conclusions. In addition, since the study is performed in a transplantation context, the interpretations of some biological insights, such as HSC numbers should be more cautious as this does not necessarily reflect the native physiologic state.

Here are specific comments:

- The HSPCs transduced by different gene expression vectors are used for different diseases. It would be helpful to provide more information of the different vectors used.

We added the description of the vectors and the source references in the in the manuscript's Results section.

-Could the differences in HSPC features observed in this paper may be caused by the difference in the vectors used for different diseases?

The vectors have the same backbone but different promoters and transgenes. The transgene has a role in the repopulation process, specifically where the transgene expression provides a selective advantage, such as in WAS patients. We consider the effects of the transgene expression to be linked to the disease background rather than a specific vector feature.

On the other hand, we do not think that the different promoters could explain the bias in lineage output and commitment among the trials, at least not in a such a massive scale. While it is possible that vector integrations might activate genes that impact the cellular phenotype, fate and/or cellular fitness, as we previously demonstrated that in HIV1 infected individuals under antiretroviral therapy had a significant enrichment in insertions activating the expression of STAT5B or BACH2 specifically on T-regulatory cells (*Cesana D. et al., Nature Communications 2017*), but it these events occur at low frequency, well below the frequency of HSPC lineage output and commitment observed in this study.

Moreover, the promoters used in these trials are of cellular origin which never have been implicated in insertional mutagenesis. Indeed, we do not have any evidence of the presence of common insertion sites or aberrant clonal expansions, hallmarks of insertional mutagenesis, nor enrichment of targeted gene classes that could explain the role of differentiation and commitment towards specific lineages.

- The author show that the insertion sites tend to be enriched in gene-dense regions. The author claimed high correlation of IS gene GO terms across different diseases. However, Extended Data Fig 1D does not seem to clearly indicate that this is the case. In addition, can the authors perform a direct comparison at the gene level or genomic region level across different diseases to investigate whether there is any insertion preference specific across disease conditions. The possibility that the HSPC clonal behaviors being biased by insertion preference needs to ruled out.

We realize now that the graph in **Extended Data Fig 1D** did not convey the message clearly. For this reason, we changed the panel with dot plot graph showing the level of semantic similarity in a pairwise fashion. In addition, as suggested by the Reviewer, we performed a genome-wide comparison of the distribution of IS by using Fisher's exact test to compare the number of IS assigned to the nearest gene for

each trial. No differences in the gene targeting frequency were observed in any of the different trials. These results are reported in the new **Extended Data Table 2**. Given that the integration profiles across the different clinical programs were essentially the same, it is unlikely that vector integrations may be influencing the behavior of clones.

- There are a number of confounders in interrogating HSPC clonal behaviors in different disease conditions that need to be assessed and accounted for. This includes the use of different conditioning regimens, gender of patients, infused cell number, PCR methods for amplification, etc. It is recommended to have an overall model (such as a multivariate regression) to model and control potential confounders together.

We thank the reviewer for this excellent suggestion. In the new version of the manuscript, we applied the Good-Turing model to correct the impact of the different numerosity of the datasets to be compared and avoid biases due to the different dataset sizes. This correction was applied to the analysis of the linage output and commitment in the different trials. Moreover, we used a Bayesian multivariate linear regression model to remove technical confounding factors such as the PCR method, amount of DNA used, dose of $CD34^+$ infused per Kg, vector copy number, sequencing depth, patient's gender, and age.

The only noteworthy difference between our previous and novel results was that the $CD34^+$ output towards B-cells in WAS patients was reduced compared to the results before correction. This observation is in line with the notion that the selective advantage provided by WASP expression in T cells is stronger than B cells (**Konno A., et al., 2004 Blood and Ferrua F., et al., Lancet Hematology 2019**).

- How do the authors control the labeling efficiency and the ability to detect ISs? For example, would the difference of PCR efficiency across different insertion context result in biases being observed potentially? This might particularly become an issue, if insertion context varies considerably across different disease contexts.

Regarding the "labelling efficiency" (we interpret this as vector marking efficiency), we consider the average vector copy number (VCN) per cell as a measure of vector marking levels and adopt specific

corrections depending on the analysis. For example, to correctly calculate the active HSPCs we correct the estimated number by dividing it by the VCN value if >1 as described previously (*Six E., et al., Blood 2020*).

Regarding the ability to detect ISs, SLiM-PCR performed on 10 ng of DNA with a vector copy number of 1, can retrieve and accurately quantify IS at 0.16% of relative abundance (corresponding to a total of 7 molecules/genomes in absolute terms (*as reported in Cesana et al, Nat Medicine 2021 and Wagner et al, Nat Comm 2022*). LAM-PCR has a reduced efficiency in the retrieval of IS up to 10-fold when compared to SLiM-PCR, as reported in a recent communication (*Benedicenti et al, Abstract #180, ASGCT 2023, Supplement1 Mol Ther, Issue 4, Vol 31 and manuscript in preparation*).

We do not think that different efficiency in IS retrieval between LAM-PCR and SLiM-PCR may bias the commitment or any observed dynamics in a such disease-specific fashion. Moreover, we want to point out that all WAS-derived IS and part of the MLD-derived IS were retrieved by LAM-PCR, yet the differences in lineage output and commitment between these two diseases remain specific. These results indicate that the difference in IS retrieval techniques is not responsible for the observed differences. Moreover, as explained above, we used a Bayesian multivariate linear regression model to remove confounding factors in all subsequent analyses such as the HSPC lineage output and commitment. The factors considered in this model also included the PCR method.

It is possible that different genomic regions flanking the integration site may be amplified with different efficiencies. Moreover, about 30% of IS landing in repeated regions, that cannot be univocally mapped, are discarded (*Spinozzi G., et al., BMC Bioinformatics 2017*). However, we consider that the different retrieval efficiency, or discarded IS, unlikely would result in such specific biases found in the different clinical programs. Indeed, we are always using LVs, which although having a different design, have essentially the same genomic distribution, as shown in the new **Extended Data Table 2**. Therefore, such bias would be the same across trials and thus it cannot explain the observed differences.

By comparing across time points, how can the authors distinguish whether one barcode is technically not detected vs. it exisiting in quiescent clones that did not contribute to hematopoiesis until a later time point?

In our data, for each sample, we cannot distinguish if an IS was not captured because the clone harboring it is quiescent or because there was a subsampling. Indeed, given the complexity of the clonal populations in the transplanted patients, despite the sensitivity of our PCR technology and even at extremely high

deepness of sequencing, it is almost impossible to retrieve all IS present in a single DNA sample (100 ng of vector marked DNA at VCN 1 will contain ~15,000 IS). Therefore, there are for sure many clones/IS that will not be detected because of the subsampling issue. For this reason, IS that are missing in one time point/sample are not considered as 0 abundance but rather undetermined, or not captured. This is important for statistical analyses.

Moreover, we performed multiple sequential samplings for IS retrieval totalizing hundreds of thousands of ISs and found that the cumulative frequency of new IS over time tends to plateau meaning that we essentially covered nearly the entire clonal repertoire. The high coverage in terms of number of samples/time points, amounts of DNA used, and sequencing depth is fundamental to avoid biases in lineage output and commitment caused by subsampling, as recently shown by the Hans-Peter Kiem's lab in a recent clonal tracking study in non-human primates subjected to HSPC-GT (*Radtke S et al., Blood 2023*).

- What is the grey bar in Extended data fig1F attempting to show? Are those non-recaptured clones?

We apologize for the lack of clarity. The gray bar represents the pool of IS with an abundance <1%. We amended the Figure legend with the description of the gray color in the bars.

- The authors discuss HSC lineage biases. Can the author provide evidence to justify the robustness in defining the uni-lineage vs multi lineage clones?

Would it be possible that the "multilineage" clones are the one with overall better detection, while the "uni-lineage" have more dropout?

We consider IS/clones only when recaptured in at least two time points and we consider only those IS represented by a genome count ≥3. These filters have been shown to be fundamental to reduce biases caused by subsampling as shown in previous studies (*Biasco L., et al., Cell stem cell 2016 and Radtke S., et al., Blood 2023*). In the recent publication from Radkle S., et al Blood 2023, the clonal abundance was positively correlated to the multilineage potential. As explained in more detail below, we did not see a significant bias in IS retrieval caused by clonal abundance, probably because we selected only IS with a genome count ≥3 and if captured in at least two time points, (essentially focusing only on robust IS, which

does not mean are the most abundant or dominant) and because we analyzed many samples obtained during several years of follow-up.

We extended the main text with the following paragraph to explain our analyses.

*A recent study in nonhuman primates[52] demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we conducted comparisons of abundances at early and late time points for clones transitioning from multilineage to uni-lineage, revealing no statistically significant differences (**Extended Data Fig. 8A**). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (**Extended Data Fig. 8B**). Collectively, these findings suggest that, at least in our dataset derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment.*

Referee #2 (Remarks to the Author):

Reviewer Comments: Long-term lineage commitment is modulated by the underlying disease in hematopoietic stem cell gene therapy patients

Summary of Manuscript: This manuscript compares the clonal outputs of lentivirally gene corrected hematopoietic stem cells (LGC-HSCs) in gene therapies for three congenital diseases: Metachromic Leukodystrophy (MLD), Wiskott-Aldrich Syndrome (WAS) and beta-Thalassemia (BTHAL). Impressively, it analyzes more than 6,700 peripheral blood (PB) and bone marrow (BM) samples from 53 patients up to 8 years after treatment.

The authors note that the number of engrafted, long-term LGC-HSCs is positively correlated with the

number of infused CD34+ cells. Importantly, they assert that they do not see any evident plateau for the total number of LT-LGC-HSCs [at least not over the range of dosages used in these trials].

From their analyses, the authors conclude that in all disease conditions 50% of clones demonstrate multilineage potential. They assert that the remainder show preferential lineage commitment that is specific to the disease condition. The authors hypothesize that this is due to LT-LGC-HSC retaining "memory" of pre-gene therapy cell states.

Major Points:

1. There are several technical confounders that could potentially mimic lineage skewing and therefore deserve more careful evaluation and discussion.

Confounders include:

(A) Gene-therapy did not in all cases fully correct initial disease conditions.

In the BTHAL trial (Marktel et al, 2019) all three adult and one of the pediatric patients continued to be transfusion dependent. The other evaluable children remained anemic. This would suggest that all patients continued to have a strong erythropoietic drive and that factors extrinsic to the LT-LGC-HSCs may have influenced lineage skewing.

Likewise in the WAS trial (Ferrua et al, 2019) patients generally remained thrombocytopenic after therapy which also may have extrinsically influenced lineage skewing.

Mitogen (EPO/TPO) levels might add useful information. At the very least a fuller disclosure in the text and caveats to the conclusion would be reasonable.

The point is well taken. We realized that we did not comment appropriately the aspect that gene therapy did not in all cases fully correct initial disease conditions and did not explore if levels of very important factors such as thrombopoietin (TPO) or erythropoietin (EPO) could be correlated to lineage skewing observed in the different clinical programs.

In this revised version of the manuscript, we performed a longitudinal analysis of TPO levels in blood-plasma samples in 14 WAS patients (harvested before GT, at a follow-up time 1 year from transplant and a FU time >2-4 years from transplant and EPO levels in 9 β-Thal patients (harvested, at a follow-up time of 1 year, 2 years and 3 years from transplant).

we found that the TPO levels in WAS patients before GT were near the normal levels, which was expected because WAS patients before therapy were under TPO receptor agonist treatment. However, TPO levels increased at 1 year after GT, regardless of the age at treatment, and then decreased to almost normal levels at 4 years from GT, suggesting that the beneficial effects of the therapy in these patients may have alleviated the need for enhanced production of this important cytokine albeit not entirely. In 4 out of 6 β-Thal patients transplanted at age ranging from 4 to 13 years EPO levels were higher compared to normal levels at all time points analyzed (up to 3 years from GT). On the other hand, 2 out of the 3 β-Thal patients treated at age >30 years who remained transfusion dependent and showed the highest output toward the erythroid lineage, showed on average a 3-5 fold higher EPO level compared to the younger cohort, especially at 1 year from GT. Moreover, 2 β-Thal transfusion independent pediatric patients showed EPO normal levels and a marked erythroid commitment. Thus, EPO levels negatively correlate with the therapeutic outcome of the therapy, albeit partially, suggesting that other factors could modulate the behavior of HSPCs in β-Thal. Overall, our data agree with the notion that HSPCs activity is modulated to better respond to the demands imposed by the specific pathological condition. This analysis is illustrated in new **Figure 4**.

(B) Differences in input DNA amounts impact the sensitivity to detecting clones and has the potential to mimic lineage skewing.

Offering specifics about the amount of input DNA for all samples is important. Concomitantly, it is essential to detail any corrections in the inference of the number of HSCs (e.g. via the sample-size-based rarefaction/extrapolation formulae for estimating diversity from a sample of a single assemblage) or that were applied to estimate overlaps in clonal outputs (e.g. via the Good-Turing estimators for the number of species shared between two assemblages).

This point may be particularly pertinent to the analyses of the sharing ratio, where corrections due to sample size can be important.

We provided details on the amounts of input DNA for all samples in the new **Extended Data Table 1**.

We corrected the HSC number estimations by dividing by the vector copy number when VCN>1 of the specific cell population under study (which is a surrogate readout of the overall marking level) as previously described (*Six E., et al., Blood 2023*). Moreover, as suggested by the Reviewer, for calculations involving the sharing levels between $CD34^+$ and cell lineages for the analysis of the output and commitment, we applied the Good Turing model to the analyses to remove the biases caused by consecutive subsampling between assemblages. Finally, we used the Bayesian multivariate linear regression algorithm to model and correct biases induced by different technical confounding factors simultaneously (which also include the amount of DNA used). We used the Bayesian multivariate linear regression algorithm to model and correct biases induced by different technical confounding factors simultaneously. These factors included PCR method, amount of DNA used, dose of $CD34^+$ infused per Kg, VCN, sequencing depth, and patient's gender. To be consistent across the different analyses, we used the same approach to correct the clonal diversity analyses (**Figure 1C**). We edited **Figure 2** and **Figure 3** with the new results, and **Extended Data Figures 2, 5-6, 9**.

Overall, the corrections resulted in slight changes in the sharing levels of all lineages, so the relative proportions among the different clone classes remained essentially the same, with only one exception: before correction, the B cell output in WAS patients was similar T-cell output, but after correction was significantly decreased (although still significantly higher than in the other two clinical programs). This is in line with the notion that the selective advantage provided by WASP expression in T cells is stronger than B cells (**Konno A., et al., 2004 Blood and Ferrua F., et al., Lancet Hematology 2019**).

As you can appreciate, after Good-Turing and the Bayesian corrections, the differences between lineages (in terms of output and commitment) are enhanced.

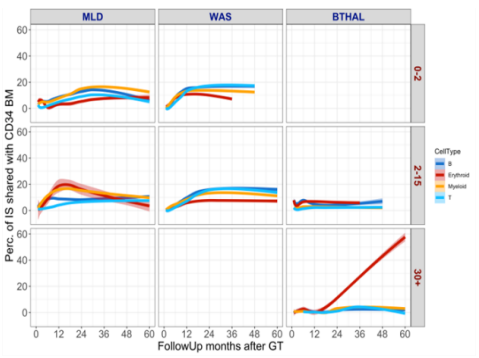


Figure 2D: CD34+ output ***before correction***

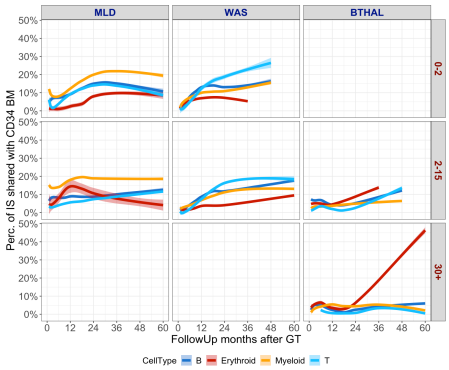


Figure 2D: CD34+ output ***after correction***

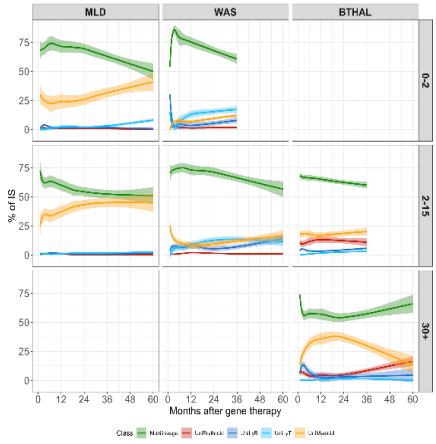


Fig. 3D: lineage commitment ***before correction***

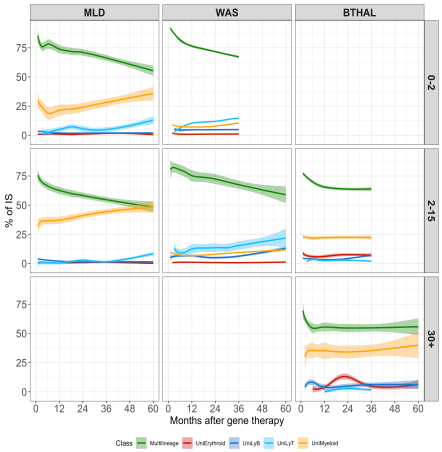


Fig. 3D: lineage commitment ***after correction***

(C) Comparisons of IS from cells where clones are geographically segregated (i.e. from the BM) can result in misinterpretions of lineage potential.

Erythroid, Myeloid and B cells in BM are locally produced and the clones are geographically separated for a time after therapy. In primates it can take up to 2 years for clonal geographic segregation to disappear (Verovskaya et al, JEM 2014; Wu et al, JEM, 2018; Chung et al, Blood 2018). Comparisons to cells from contaminating blood or T cells which develop outside the BM may lead to erroneous interpretations with regard to lineage skewing.

We are not sure we understood this point as it is unclear (to us) how the geographical segregation of clones in bone marrow could skew the lineage output and commitment in such a disease-specific fashion. Our IS

datasets were obtained before and well beyond a follow-up time of two years, and we do not see relevant differences between the time point from one year to two years after transplant and later time points (as possible to observe in all figures).

Moreover, comparison of the IS retrieved in BM and PB, and their skewing in terms of lineage output as well as commitment showed some differences in WAS patients where T-cells in PB appear to have a higher frequency of lineage commitment compared to the T cells found in BM. Since T cells are produced in the thymus and released in the bloodstream, the analyses on PB could better reflect the lineage commitment. Thus, while the initial geographical segregation of vector marked clones in BM could somehow bias the interpretations on lineage skewing it did not appear to be relevant in our settings in which a large number of samples and periods of follow up well beyond 2 years after therapy have been analyzed with great sequencing depths.

(D) Misidentification of multi-insertion clones as uni-insertion clones can result in both (a) overcounting of inferred number of HSC clones and (b) to the extent that multi-insertion clones have low-prevalence and concomitant less complete recovery of all ISs, overcounting of lineage-restricted clones. Essential are a more complete description of the number of vector insertions per HSC along with an explanation of any corrections applied to the computation of the number of HSCs and their lineage restriction.

We now corrected the HSC number estimations by dividing the HSPC number by the VCN when > 1 as described in *Six E., et al Blood 2020*. This type of correction allows to avoid the inflation of the HSC numbers caused by clones with multiple integrations (updated **Extended Data Table 3**, and **Figures 1E-F** and **Extended Data Figure 3A** with cut-off time point at 24 months).

Moreover, to eliminate any impact of vector marking levels on lineage skewing we implemented a correction based on a Bayesian multivariate linear regression model which corrects for technical variables including the VCN (together with the PCR method for IS retrieval and the amount of DNA used, cell dose). The results, after these additional corrections, further reinforced our claims.

2. Beyond technical issues affecting whether clonal output is truly skewed, there are several other mechanisms besides LT-LGC-HSC intrinsically retaining "memory" of pre-gene therapy cell states that could result in putative lineage skewing.

Plausible mechanisms include:

(A) Persistence of the pre-gene therapy environment for hematopoiesis

See point 1(A) above.

The point is well taken, and we provided an answer to point 1A above.

(B) Differences in the conditioning regimens and their resultant effects on the hematopoietic environment

The authors note in that different conditioning regimens were used in the therapies for different diseases. A lymphodepleting regimen was given to WAS patients; consequently there was more rapid "filling" of the lymphoid compartment with LGC cells. Filling may have originated from ST-HSPCs that did not persist, thus producing 'uni-lineage' T cells and separate from LT-HSPCs. Once filled homeostatic proliferation maintains T cell clones independent of on-going production from LT-HSPCs.

By contrast, much slower refilling of the T cell compartment occurred in BTHAL (where thiotepa and treosulfan were used for conditioning) and MLD (where busulfan was used for conditioning).

This point deserves fuller disclosure in both the abstract and the conclusions.

We did not comment on the different repopulation kinetics of the different lineages in each clinical program as they have already been described in previous publications, albeit separately. We now disclose the differences in conditioning and repopulation kinetics in the introduction section as follows:

"In each of the 3 clinical programs, different conditioning regimens have been adopted. MLD patients, received full myeloablative conditioning[5,27,28] with busulfan at doses ranging from 10 to 14 mg/Kg. WAS

patients received a reduced-intensity conditioning regimen designed to achieve depletion of HSPC and lymphoid cells consisting of the combined administration of a monoclonal antibody against CD20, busulfan (7.6 to 10.1 mg/Kg) and fludarabine (60 mg/m^2)[4,19,21,39]. In β-Thal patients, the conditioning consisted in thiotepa (6-8 mg/Kg) and treosulfan (14 g/m^2) administration[10].”

Regarding the possibility that in WAS patients the T-cell reconstitution may have originated from ST-HSPCs that did not persist, thus producing ‘uni-lineage’ T cells and separate from LT-HSPCs:

We agree that uni-lineage T-cells may arise from already committed HSPCs. This concept applies not only on long-lived uni-lineage committed T-cell clones but also to uni -myeloid, -B and -erythroid restricted clones which are also long-lived probably, although the HSPCs output and commitment vary specifically depending on the disease, age of treatment and disease burden. However, because uni-lineage committed clones persist for years, it is difficult to explain that these are originating from ST-HSPCs.

Moreover, we devised a new analysis to further investigate the dynamics of lineage commitment over time and found that 1/3 of the long-term uni-lineage committed clones originated from multilineage clones that during the early phase of hematopoietic reconstitution (<2 years from transplant) turned into uni-lineage committed and persisted long-term (> 2 years from transplant), while the remaining 2/3 were found to be committed since the early to late phases of hematopoietic reconstitution. The pressure to transit from multilineage to uni-lineage committed was specific for each clinical program.

(C) Persistence of heritable epigenetic changes intrinsic to HSPCs, which arose prior to treatment Specific mechanisms might include (a) intrinsic disparities in the rates by which HSCs differentiate to specific lineages versus (b) intrinsic differences in the proliferation rates of committed progenitor states. If this is a characteristic of HSPCs independent of their environment, this might be apparent in in vitro differentiation and proliferation studies.

(D) [Probably less likely] persistent changes in the cells comprising the hematopoietic niche; either because these have remained uncorrected or due to their exposure to the original pre-gene therapy environment. Although a less likely explanation, it could be investigated after other possibilities have been ruled out.

A more definitive investigation would include either bulk ATAC-seq on HSPCS and selected subsets or scATAC-seq.

Answer to points C and D:

Our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level studies to unravel biological differences in each disease condition. These studies are still in progress and will be reported in the future as an independent study.

It would be of interest to show abundance by various classes of clone.

Beyond the computation of Shannon indices, remaining analyses focus on binary (absence of presence within a lineage) classifications of clones. By lineage abundance fractions offer useful insights and are typically used in murine primate studies. For instance, are the uni-lineage clones small and therefore might sampling be an issue?

To address this point, we designed a new analysis aimed at unraveling if differences in clonal abundance could bias the linage output or commitment. The description of the analysis reported in the results section is the following:

*"A recent study in nonhuman primates[46] demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we conducted comparisons of abundances at early and late time points for clones transitioning from multilineage to uni-lineage, revealing no statistically significant differences (**Extended Data Figure 8A**). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (**Extended Data Figure 8B**). Collectively, these findings suggest that, at least in our dataset*

derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment."

Therefore, we did not see a significant bias in IS retrieval caused by clonal abundance, probably because we selected only IS with a genome count ≥3 and if captured in at least two time points, (essentially focusing only on robust IS, which does not mean are the most abundant or dominant) and because we analyzed many samples obtained during several years of follow-up.

Minor Points:

1. Please note explicitly in the text (rather than figure legends and supplemental text) that HSPC number inference was done with VCN corrections.

We mentioned in the main text that the HSPC number inferences were done with VCN corrections when the VCN is >1.

2. For the one WAS and two BTAHL patients with dramatically greater numbers of cumulative ISs, it would be useful to also see a plot of the proportion that were detected over the long-term.

We calculated the proportion of IS found in these 3 patients. We did not plot these results but mentioned specifically in the result section (from line 218 to 222) as follows:

"We calculated the percentage of IS detected over the long-term (>24 months after therapy) for the 3 patients that showed a great number of IS. In the WAS patient with ~450,000 IS (Pt17) the proportion of IS detected long term was 4.22%, while in the two β-Thal patients with ~450,000 and ~280,000 IS (respectively Pt36 and Pt41) the proportions were ~3.5% and 1.74% respectively."

3. Why is 24 months regarded as long-term/stable in many analyses (e.g. Figure 2), but for numbers of HSCs and ISs and their comparisons in Table S2, 12 months is used as the cut-off between short- and long-term. 24 months appears more relevant from the data in the rest of the paper.

In some analyses, we used the 12 months as a cutoff because some patients had a relatively short follow up (<36 months) and would be excluded from the analysis (12 MLD and 7 WAS and 3 β-Thal). Therefore, to include these patients in the analyses, we reduced the cutoff to 12 months. However, for consistency, we now harmonized the analyses to 24 months and updated figures and data accordingly (Figure 2, Extended data 3, Supplementary Figure S3A). The new results for MLD and WAS clinical programs are essentially the same, showing that there is a significant decrease in HSPC size over time. However, with this cutoff the significant decrease HSPC size was lost in β-Thal patients. Therefore, the drop in HSPC size in β-Thal patients occurs before the other two clinical applications, possibly because were transplanted with mobilized $CD34^+$ cells which lead to a faster recovery and stabilization when compared to BM-derived $CD34^+$ cells. For this reason, for β-Thal patients we show two panels with the cutoff to 12 months and 24 months as well.

4. The word "cell" is confusingly used synonymously to individual IS (e.g. figure 2A and in the methods supplement)

We harmonized the terminology.

Referee #3 (Remarks to the Author):

This manuscript by A. Calabria et al analyzes a large and detailed dataset for assessing the safety and post-transduction kinetics of engraftment and stable hematopoiesis after lentiviral gene therapy for hereditary disorders. Integration site analysis was used to characterize these dynamics as they relate to the diversity and lineage-specificity of engrafting clones, analyzed in samples collected over nearly a decade of follow-up. The authors report an intriguing finding, which is that the underlying disease appears to influence expansion of the transduced rescued lineage, which is influenced more broadly by patient age at treatment, VCN and transduction efficiency, and the tempo of hematopoietic reconstitution. This is an important report that will be very useful to understanding the dynamics of hematopoiesis and safety after gene therapy in hematopoietic stem cells. While there are limitations in the report regarding the mechanistic underpinnings of these observations and

extrapolating the data to predict clinical outcomes from baseline characteristics in the patient and/or in drug product, this is probably only the beginning of a very important story.

Major comments:

With regard to the late appearing IS's >24 month post-infusion, particularly in the older thalassemia patients, was there any evidence that these emerged/were recruited in the setting of hematopoietic stress in which a proliferative stem cell expansion might be triggered in lieu of quiescence?

We believe that a good part of these clones arises from committed progenitors as well as LT-HSCs prompted by hematopoietic stress.

In adult β-Thal patients the exacerbation of the erythroid output over time is suggestive of hematopoietic stress. Indeed, in adult and transfusion dependent β-Thal patients after 24 months the long-lasting clones were about 2% of the entire clonal population while in MLD and WAS patients ranged between 8 to 10 %. This difference suggests that in β-Thal patients there is more HSC activity on the long term as new IS (clones) appear also in late stages compared to MLD and WAS patients. Moreover, it should be noted that in the pediatric β-Thal patients the erythroid output is significantly higher than in MLD and WAS patients, but it was lower than adult β-Thal patients. Interestingly the significant increase in erythroid output in adults appears only 24 months after transplantation and increased progressively over time, suggesting that stress resulting from inefficient erythropoiesis accumulated over time is responsible of the increased observed output.

In other studies of gene therapy for thalassemia, it has been observed that features of stress erythropoiesis persist even after establishing RBC transfusion independence (skewed M:E ratio in the marrow favoring erythroid progenitors, persistently elevated markers of ineffective erythropoiesis, etc).

The point is well taken. We realized that we did not comment appropriately the aspect that gene therapy did not in all cases fully correct initial disease conditions and did not explore if levels of very important factors such as thrombopoietin or erythropoietin could be correlated to lineage skewing observed in the different clinical programs.

In this revised version of the manuscript, we performed a longitudinal analysis of thrombopoietin levels in blood-plasma samples in 14 WAS patients (harvested before GT, at a follow-up time 1 year from transplant and a FU time >2-4 years from transplant and erythropoietin levels in 9 β-Thal patients (harvested, at a follow-up time of 1 year, 2 years and 3 years from transplant).

From this analysis we found that the TPO levels in WAS patients before GT were near the normal levels, which is expected as WAS patients before therapy were under TPO receptor agonist treatment. However, TPO levels increased at 1 year after GT, regardless of the age at treatment, and importantly decreased to almost normal levels at 4 years from GT, suggesting that the beneficial effects of the therapy in these patients may have alleviated the need for enhanced production of this important cytokine albeit not entirely.

In 4 out of 6 β-Thal patients transplanted at age ranging from 4 to 13 years EPO levels were higher compared to normal levels and at all time points analyzed (up to 3 years from GT). On the other hand, the 2 out 3 β-Thal patients treated at age >30 years which remained transfusion dependent and showed the highest output toward the erythroid lineage, showed on average a slight 3-5-fold increase of EPO level compared to the younger cohort, especially at 1 year from GT. Moreover, 2 β-Thal transfusion independent pediatric patients showed EPO normal levels and a marked erythroid commitment. Thus, EPO levels correlate positively with the therapeutic outcome of the therapy, although partially, suggesting that other factors could modulate the behavior of HSPCs in β-Thal. Overall, our data agree with the notion that HSPCs activity is modulated to better respond to the demands imposed by the specific pathological condition.

Moreover, in the result section we introduced these analyses with this paragraph which discuss the possible factors influencing the hematopoietic output and commitment in WAS and β-Thal patients:

"The preferential lineage output and uni-lineage commitment, specific for WAS and β-Thal patients, appears to persist even after years from GT, suggesting that despite the (overall) positive therapeutic effect

of the treatment in WAS patients there is still the need for a heightened output for T and B -cells and in β-Thal patients, especially those which remained transfusion dependent, for erythroid cells.

Because hematopoietic cytokines can instruct the lineage fate of HSPCs and provide survival and proliferation signals to both multipotent progenitors as well as committed progenitors and are regulated by disease states we wanted to assess if in WAS patients, plasmatic levels of thrombopoietin (TPO) could be indicative for the hematopoietic recovery by favoring the output lymphoid lineages and megakaryocytic/platelet maturation and if in β-Thal patients, elevated plasmatic levels of erythropoietin (EPO), a classic hallmark of anemic β-thalassemia patients and gene therapy patients[47-49], could be associated with the increased erythroid output".

The question of exhaustion of true HSCs following cell proliferation signals in this setting is also unclear. The authors argue that early appearance of ISs post-infusion and their drop-out indicates these were HSPCs and not true HSCs but it is possible HSC exhaustion and drop out has not been excluded, particularly in the period of recovery and rapid expansion that follows pre-infusion myeloablation/conditioning. This would tend to select a smaller subset of clones with better proliferative activity.

We apologize for the lack of clarity on this point. We do not exclude that in the early phases of hematopoietic reconstitution the drop-out can be caused by exhaustion of HSCs. However, we believe that the dropout of early IS (early clones) is mostly due to the disappearance of short-lived progenitors and in part by exhausted HSCs. In support of this hypothesis, our result of the lineage commitment over time on singleton IS (**Figure 3J**), which are those ISs found only in a single timepoint but can be found in multiple lineages, show that at the earliest time point which stars 1 month after transplant, about 50% of the clones were found to be myeloid in all the disease conditions. However, the frequency of these myeloid-restricted ISs decreased dramatically as early as 6 to 12 months after transplant and stabilized thereafter to about 5-10%. On the other hand, the multilineage singleton ISs (IS found only in a single time point in multiple lineages) constituted about 10-20% from 1 to 3 months after transplant and decreased to 1-5% thereafter. We showed that the percentage of myeloid-restricted singleton ISs is about 50 % in early phases (1 to 3 months after transplant) while the multilineage singleton ISs (which could be originated by HSPCs) were much less, suggesting that the early dropout in mostly caused by myeloid committed progenitors.

Moreover, in β-Thal patients, erythroid committed singleton IS during the early phases after transplant reached 25% in pediatric patients and remained stable thereafter while in adult patients reached >60% from 1 to 3 months after transplant and decreased to 25% after 6 months from transplant. These data suggest that in β-Thal patients, especially in adults, large numbers of erythroid committed cells were transplanted.

Finally, we devised a new analysis to further investigate the dynamics of lineage commitment over time and found that 1/3 of the long-term uni-lineage committed clones originated from multilineage clones that during the early phase of hematopoietic reconstitution (<2 years from transplant) turned into uni-lineage committed and persisted long-term (> 2 years from transplant), while the remaining 2/3 were found to be committed since the early to late phases of hematopoietic reconstitution. The pressure to transit from multilineage to uni-lineage committed was specific for each clinical program.

It is also sobering to observe that a very small number of true HSCs ultimately establish steady-state hematopoiesis, under conditions that would appear to select clones with robust proliferative capacity.

This also raises the question if stochastic events might skew abundance of some clones, simply because they are more proliferative and then enriched further by way of natural selection of a particular lineage, as occurs in thalassemia for example, where in allogeneic HCT donor-host chimerism also favors enrichment of corrected donor cells in erythroid progenitors? A similar analysis would find an erythroid skewing of CD34+ cells of donor origin, even when there is a minority of donor cells.

To address this point, we designed a new analysis aimed at unraveling if differences in clonal abundance could bias the linage output or commitment.

The results of this analysis were reported in the Results section as follows:

"A recent study in nonhuman primates[46] demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we conducted comparisons of abundances at early and late time points

*for clones transitioning from multilineage to uni-lineage, revealing no statistically significant differences (**Extended Data Fig. 8A**). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (**Extended Data Fig. 8B**). Collectively, these findings suggest that, at least in our dataset derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment."*

Therefore, we did not see a significant bias in IS retrieval caused by clonal abundance, probably because we selected only IS with a genome count ≥3 and if captured in at least two timepoints, (essentially focusing only on robust IS, which does not mean are the most abundant or dominant) and because we analyzed many samples obtained during several years of follow-up.

While the kinetics of clonal hematopoiesis and the size of this cohort across three disparate hereditary disorders is very impressive, the mechanistic basis for the phenomenon observed – a clonal bias favoring a particular lineage over another – has not yet been defined, although it will be critical to do so. While perhaps beyond the scope of this study, an obvious question is whether there are epigenetic marks in lineage specific loci that might establish and favor the expansion of a single lineage from these HSCS?

The reviewer is right, our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level and will be reported in the future as an independent study.

Does the lentiviral vector tropism for integration near chromatin and histone-modification loci favor the re-capitulation of chromatin configurations in the HSCs that direct lineage differentiation?

We performed a genome-wide comparison of the distribution of IS by using Fisher's exact test to compare the number of IS assigned to the nearest gene for each trial. No differences in the gene targeting frequency

were observed in any of the different trials. These data indicate that differences in integration profiles cannot explain differences in $CD34^+$ lineage outputs nor commitment.

These results are reported in the new **Extended Data Table 2**.

Would it be useful to conduct a study of snRNA-seq to better delineate the progenitor populations as these expand after engraftment? The manuscript would have been improved by including some of these studies (if feasible since snRNA-seq would require fresh marrow samples) to better understand mechanistic underpinnings of these very interesting observations.

As explained above our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level and will be reported in the future as an independent study.

In Fig 4C, the sharing ratio of B and T-lineages in WAS was not as significant as the sharing ratio of the erythroid lineage observed in thal. In fact, the sharing ratio significance was not prominent in B & T lineages between WAS and thal. Does this indicate that marking and enrichment for T and B cells in WAS was not as strong as erythroid selection pressure in thal? It would be interesting to evaluate the sharing ratio in GT recipients with X-SCIDs, in whom the sharing ratio for T-lineage might be especially pronounced. If observed, this would suggest the strength of the natural selection for corrected clones might be predicted to follow the impact of the mutation. Or is this finding simply reflective of the transduction efficiency in WAS (70 – 90% LVV+) compared with thal (30 – 77%) as shown in Table 1. This would tend to exert a stronger selection in the minority of erythroid progenitors with the transgene compared with residual cells and cells from drug product lacking vector, as both the latter populations will be prone to ineffective erythropoiesis and apoptosis. This was also reflected in the older thal patients having lower VCN/%LVV+ HSPCs with higher active HSPCs. This is supported by the association with transduction efficiency depicted in the PCA in Fig 3B.

In adult β-Thal patients and who did not reach transfusion independence, we observed an extreme progressive bias to produce erythroid cells with any sign of stabilization, while in the pediatric patients (where 3 out of 4 reached the transfusion independence) was significantly lower (although still significantly

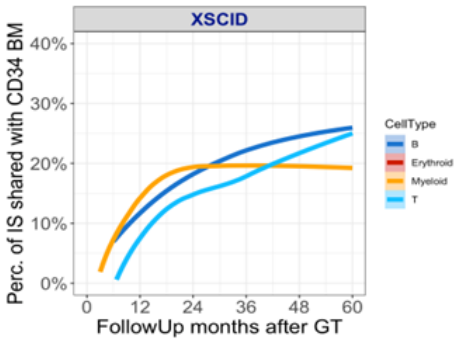
higher than MLD or WAS patients) and stabilized since the early phases of hematopoietic reconstitution. Thus, we believe that the lower skewing observed in WAS compared to β-Thal is related to the therapeutic outcome of the different treatments and the different disease burdens accumulated over many years in adult β-Thal patients.

As anticipated in the general response to the Reviewers and Editor we did study the linage output and commitment in X-SCID patients from another clinical trial (*De Ravin., et al Nature Communications 2022*). The analyses of $CD34^+$ output and commitment were included in the Results section as follows:

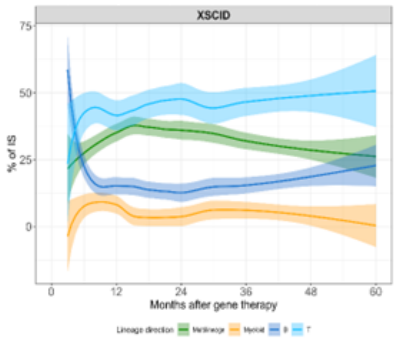
*"To confirm that the increased lineage output for the lymphoid B and T cell lineages in WAS patients was not specific to the disease but a general characteristic of the lymphoid impairment observed also in other immunodeficiencies, we analyzed previously published [23] IS datasets of 10 XSCID patients that received HSC GT with a lentiviral vector expressing the common gamma chain cDNA after a mild non-myeloablative conditioning with busulfan (6 mg/kg). These datasets comprised ISs retrieved over time (max follow-up of 84 months) from $CD34^+$ cells (16,650 IS), $CD14^+$ myeloid cells (16,640 IS), $CD3^+$ T-cells (95,362 IS) and from $CD19^+$ B-cells (58,317 IS), for a total of 186,969 IS. As described above, we applied the Good-Turing correction to eliminate biases related to the comparison of IS datasets with unbalanced numerosity. The output of $CD34^+$ cells towards the T and B cell lineages was initially low (<2%) but increased progressively up to 25% at 60 months, the last time point of the analysis (**Extended Data Fig. 5C**). Therefore, our data shows that the preferential output towards T and B cells is common between these two lymphoid immunodeficiencies. Surprisingly, the B-cell output was similar if not superior and faster than the output of T-cells in XSCID patients when compared to WAS, which is unexpected since, as in XSCID patients the selective advantage of T-cells is considered to be superior compared to B cells[45]. The reasons for this discrepancy are unclear".*

And

….. Lineage commitment analysis performed for the 10 XSCID patients described above showed that the T cell committed clones were already 25% at the early time points and increased up to 50% at 60 months ***(Extended Data Fig. 9****). Conversely, multilineage clones exhibited an initial rise from 25% to 37% at 12 months but progressively decreased to 25% at the latest available time point. Committed B-cell clones, on the contrary, appeared to be >50% at 3 months but rapidly decreased to ~13% at 8 months and slowly increased up to <25% at the last time point. Myeloid cells showed a commitment of ~6%, resembling the pattern observed in WAS patients. These findings indicate that as in WAS patients also XSCID patients have a pronounced uni-lineage T cell commitment and to a lesser extent for B cells (****Extended Data Fig. 9****).*



Extended Data Figure 5C



Extended Data Figure 9

Minor comments:

Line 147 – this appears to be missing a statement that the erythroid lineage had higher clonal complexity in thalassemia compared with the other 2 disorders.

Thank you for noticing the mistake. We fixed it now.

Line 180 – was the ‘depth’ of myeloablation more complete in thalassemia and MLD than in WAS recipients, accounting for the larger drop-off IS’s compared with estimated HSPCs in the steady-state phase. Might this also be related to lower numbers of long-term HSCs in MLD and thal, where selection of the most-fit proliferative clones under the stress hematopoiesis with engraftment might have occurred?

We think the condition protocol is absolutely a factor that impacts the kinetic of hematopoietic reconstitution in terms of the speed in which some lineages repopulate.

As mentioned in the introduction section:

"In each of the 3 clinical programs, different conditioning regimens have been adopted. MLD patients received full myeloablative conditioning[5,26,27] with busulfan at doses ranging from 10 to 14 mg/Kg. WAS patients received a reduced-intensity conditioning regimen consisting of the combined administration of a monoclonal antibody against CD20, busulfan (7.6 to 10.1 mg/Kg) and fludarabine (60 mg/m^2)[4,19,21,32]. In β-Thal patients, the conditioning consisted in thiotepa (6-8 mg/Kg) and dose- adjusted treosulfan (14 g/m^2) administration[10]".

Line 240 – it would be very interesting to determine if the older patients with thal in whom the lineage sharing of CD34+ cells with erythroid cells was most striking also had driver mutation SNPs characteristic of clonal hematopoiesis. It is acknowledged that IS clonal expansion was not observed in this analysis, but age-driven accumulation of driver mutations is a recognized phenomenon and might occur in thal as appears to be the case in sickle cell disease.

This is a very intriguing point. For this reason, we did search for somatic mutations in exons of 40 genes involved in clonal hematopoiesis and myeloid cancer in all nine β-Thal patients and 23 MLD patients. As explained in the in the results

"We then performed an exhaustive analysis of somatic mutations in exons of 40 genes involved in clonal hematopoiesis and myeloid cancer by using the Illumina's AmpliSeq™ Myeloid Panel Targeted exome sequencing kit. We searched for somatic mutations in transduced CD34$^+$ cells before infusion and PBMCs obtained at ~2 years after treatment and at the last available timepoint (range 2.5 to 7.5 years after transplant) in all nine β-Thal patients and 23 MLD patients. For each time point we analyzed 20 ng of genomic DNA (2,700 equivalent genomes).

*Overall, we obtained >300,000,000 sequence reads, which after filtering for sequence quality, genomic mappability, and removing amplicons with sequencing depth <200 reads per exon, yield >100,000,000 sequencing reads correctly aligned on the targeted exon panel (see **Methods**). The average sequencing depth in β-Thal and MLD patients was 4,400 ± 283 and 4,300 ± 1789 reads/base respectively (**Extended***

Data Table 4**). Variant calling analysis was performed using VarScan2[50] with custom filtering parameters (see **Methods**) to eliminate false positives, we removed mutations present in more than one patient, or present in the last 3 bp of the reads and mutations in regions enriched in poly-T or poly-A (manually curated). Moreover, we removed mutations with a Varian Allele Frequency (VAF) suggestive of heterozygous or homozygous germline variants (49 < VAF < 51 or VAF > 99). The detected mutations underwent annotation utilizing the Genome Aggregation Database (gnomAD)[51], the Database of Single Nucleotide Polymorphisms (dbSNP)[52] and the ClinVar database[53]. Out of the identified mutations, only 4 were annotated as known variants with no role as drivers in clonal hematopoiesis or cancer. Most somatic mutations (85 out of 96) exhibited a Variant Allele Frequency (VAF) of less than 2%. The most abundant mutation discovered in an MLD patient, involving the p53 gene with an average VAF of 15%, was annotated as benign **(Extended Data Table 5)**. In all β-Thal patients, we found a total of 68 somatic mutations, 67 of which were single nucleotide variants (SNVs) and one single nucleotide deletion. In each β-Thal patient were found from 2 to 24 somatic mutations (average 7.5 ± 6.7 mutations per patient). The average number of mutations in β-Thal patients remained consistent across all time points, showing no statistically significant variations (p-value>0.9 by Friedman test) (**Fig. 5A**). Considering that the sequenced genomic interval corresponds to 76,715 bps and that we analyzed a total of 8,100 equivalent genomes per patient, the resulting mutation rate in β-Thal patients was $1.21x10^{-8}$ mutations/bp. Conversely, we collectively observed 26 somatic mutations in 16 out of 23 MLD patients, 25 of which were SNVs and one single nucleotide deletion. These patients acquired from 1 to 3 somatic mutations (average 1.6 ± 0.7 mutations per patient), resulting in a mutation rate of $2.6x10^{-9}$ mutations/bp. The average number of mutations in MLD patients was similar at all time points without any statistically significant variations (p>0.9 by Friedman test) (**Fig. 5B**). The average mutation rate in the 3 adult β-Thal patients was significantly higher than the rate measured in the 6 pediatric patients (11.3 ± 11 Vs. 5.6 ± 3.5, p<0.05 by one way ANOVA). Thus, the average somatic mutations rates in adult and pediatric β-Thal patients were significantly higher than in MLD patients treated at 0-2 years or 2-15 of age (by one-way ANOVA test) (**Fig. 5C).** Five out of the 96 mutations (4 in β-Thal and 1 in MLD) were found in more than one time point and none showed a progressive increase in abundance, suggesting that these mutations did not confer a selective/proliferative advantage to the mutated cell clones (**Fig. 5D).

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have appropriately addressed all of my concerns. They have now applied Good-Turing estimation and Bayesian corrections to account for the undetected IS as well as various confounders, which significantly enhance confidence in the conclusions. The authors also added TPO/EPO levels, as well as somatic mutation analysis.

Overall, this study presents a valuable dataset based on long term tracking of insertion sites and provides important insights on the HSC dynamics after transplantation.

The estimation of HSC clonal lineage commitment over time is interesting and I appreciate the author's explanation for how they categorize cells as multi- or unilineage. However, subsampling is an important limiting factor as the authors state. Given this limitation, it would be valuable to add to the discussion about this issue and ways this could be addressed in the future. A relevant question: Can the author comment on whether there exist HSC clones that change their commitment over time (for example, from multi-lineage to uni-lineage, or from one lineage to another)? If the current dataset is not sufficient to answer this question, is it possible to address this in the future?

The author state that 50% of transplanted clones exhibit multilineage behavior. I wonder whether in the HSC clones labeled as "uni-lineage", there are some that could actually be multi-lineage, but appear as "uni-lineage" due to data sparsity or a lack of reads. Is it possible to provide a confidence level for each assignment (label)?

In addition, my only minor point is that the layout of some plots could be improved to more clearly present the data.

Referee #1 (Remarks on code availability):

The code is clearly presented and is a valuable resource.

Referee #2 (Remarks to the Author):

Major Points

1(A) Thank you for the additional important data on mitogen levels in HSC-GT patients and the additional figure 4. The figures are plotted in log scale, which is helpful for increasing dynamic range for but tends to minimize differences. The data indicate a transient rise in TPO and a sustained elevation of

EPO after HSC-GT.

Mitogen-driven proliferation of progenitors can result in higher detection rates of IS in the corresponding mature lineage; even while these IS remain below the limit of detectability in other lineages. This can be seen from the Chao1 species richness estimator:

$S_{Chao1}=S_{obs}+(n-1)/2n\ (f_1^2)/f_2$

where Sobs is the number of observed IS, n is the size of the sample, fr is the number of IS with precisely r reads (usually termed the abundance frequency count). The second summand estimates the number of unobserved IS. Higher clonal outputs reduce $(f_1^2)/f_2$ and thereby also reduce the number of unobserved IS in that lineage.

1(B) Thank you for providing data on the amount of DNA used for all samples. The table suggests a very wide distribution of input DNA:

Figure 1 Full distribution of input DNA (See attached Reviewer Comments pdf)

Figure 2 Distribution of input DNA, truncated at the 70% quantile (See attached Reviewer Comments pdf)

With the range of input DNA varying over 3 orders of magnitude, sensitivities to IS will vary greatly. The may explain why the number of detected IS fluctuates so much even for a single patient. For instance for MLD Pt43, the number of IS detected from 1 month to 60 months varies from 118 to 1750 with a standard deviation of 426.3, which is 45% of the mean number of IS detected. (This suggests that about half of the IS are not detected at a given interval.). The distribution does not stabilize during 'HSPC homeostasis'. This does not seem biologically plausible and suggests variation in sampling.

Thank you for applying the Good-Turing correction. We do not believe it was carried out properly. The Good-Turing correction to the calculation of multipotent clones is potentially very significant; please see the detailed notes we attach on this point. (Please see attached Comparison of Assemblages pdf)

1(C) Comparing locally produced erythroid cells in marrow for the erythroid lineage versus myeloid and B cells that are both locally produced and in contaminating blood versus T cells that are all produced outside the marrow is potentially problematic, at least early before mixing has occurred; primate data suggests it can take up to two years before geographic segregation disappears.

1(D) Thank you for providing the data on VCN and for introducing a Bayesian multivariate linear regression model to correct for this.

2(A) Please see our remarks above.

2(B) Thank you for adding details about the conditioning regimen. Differential proliferation in specific subsets of cells in order to restore homeostasis can result in high sensitivity to detecting IS and deserves mention.

2(C) We believe that, given the lack of sensitivity of the IS analysis, single cell scRNA-seq and scATAC-seq studies will be more illuminating.

2(D) Thank you. We look forward to reading this study.

3 Thank you. Since (i) multilineage potential was defined by IS detected in two distinct lineages at any time point and (ii) number of detected IS varied widely by time point, it would be most useful to see the distribution of clonal read counts at the time point it appeared in two lineages versus time points when it did not. Mean [longitudinal] abundances per patient obscure this.

Minor Points:

(1) Thank you for providing a table of VCNs and the development of a Bayesian multivariate linear regression model to correct for this.

(2) Thank you for highlighting this information for us. [Would be great to know if the more abundant clones were more likely to be detected long term, because this would raise concerns about sensitivity. In the Kiem paper the top 100 clones were all multilineage and sensitivity of ISA was limitation.]

(3) Thank you for this change. [The reason for showing two cutoffs for BTHAL patients deserves mention in the discussion if not already there.]

(4) Thank you for the changes which help to make the presentation clearer.

Summary Reviewer Opinion: This is an important paper, based on a huge, long-term and rich dataset that will be of interest to both gene therapy and hematopoiesis investigators. Important conclusions include the observation that, for all disease conditions, the number of active HSPCs is positively correlated to the dosage of CD34+ cells without evident plateau. The hypothesis that the prior disease condition imprints LT-HSCs is interesting but, given the shortcomings of the IS analysis, is not convincingly supported and should be stated in a more qualified manner.

Referee #2 (Remarks on code availability):

The code has been published in a separate manuscript.

Attachment to Referee #2's review

Reviewer Comments: Long-term lineage commitment is modulated by the underlying disease in hematopoietic stem cell gene therapy patients

Summary of Manuscript: This manuscript compares the clonal outputs of lentivirally gene corrected hematopoietic stem cells (LGC-HSCs) in gene therapies for three congenital diseases: Metachromic Leukodystrophy (MLD), Wiskott-Aldrich Syndrome (WAS) and beta-Thalassemia (BTHAL). Impressively, it analyzes more than 6,700 peripheral blood (PB) and bone marrow (BM) samples from 53 patients up to 8 years after treatment.

The authors note that the number of engrafted, long-term LGC-HSCs is positively correlated with the number of infused CD34+ cells. Importantly, they assert that they do not see any evident plateau for the total number of LT-LGC-HSCs [at least not over the range of dosages used in these trials].

From their analyses, the authors conclude that in all disease conditions 50% of clones demonstrate multilineage potential. They assert that the remainder show preferential lineage commitment that is *specific* to the disease condition. The authors hypothesize that this is due to LT-LGC-HSC retaining "memory" of pre-gene therapy cell states.

Major Points:

1. There are several technical confounders that could potentially mimic lineage skewing and therefore deserve more careful evaluation and discussion.

   Confounders include:

   (A) Gene-therapy did not in all cases fully correct initial disease conditions.

   In the BTHAL trial (Marktel et al, 2019) all three adult and one of the pediatric patients continued to be transfusion dependent. The other evaluable children remained anemic. This would suggest that all patients continued to have a strong erythropoietic drive and that factors *extrinsic* to the LT-LGC-HSCs may have influenced lineage skewing.

   Likewise in the WAS trial (Ferrua et al, 2019) patients generally remained thrombocytopenic after therapy which also may have extrinsically influenced lineage skewing.

   Mitogen (EPO/TPO) levels might add useful information. At the very least a fuller disclosure in the text and caveats to the conclusion would be reasonable.

   The point is well taken. We realized that we did not comment appropriately the aspect that gene therapy did not in all cases fully correct initial disease conditions and did not explore if levels of very important factors such as thrombopoietin (TPO) or erythropoietin (EPO) could be correlated to lineage skewing observed in the different clinical programs.

In this revised version of the manuscript, we performed a longitudinal analysis of TPO levels in blood-plasma samples in 14 WAS patients (harvested before GT, at a follow-up time 1 year from transplant and a FU time >2-4 years from transplant and EPO levels in 9 β-Thal patients (harvested, at a follow-up time of 1 year, 2 years and 3 years from transplant). we found that the TPO levels in WAS patients before GT were near the normal levels, which was expected because WAS patients before therapy were under TPO receptor agonist treatment. However, TPO levels increased at 1 year after GT, regardless of the age at treatment, and then decreased to almost normal levels at 4 years from GT, suggesting that the beneficial effects of the therapy in these patients may have alleviated the need for enhanced production of this important cytokine albeit not entirely. In 4 out of 6 β-Thal patients transplanted at age ranging from 4 to 13 years EPO levels were higher compared to normal levels at all time points analyzed (up to 3 years from GT). On the other hand, 2 out of 26 of the 3 β-Thal patients treated at age >30 years who remained transfusion dependent and showed the highest output toward the erythroid lineage, showed on average a 3-5 fold higher EPO level compared to the younger cohort, especially at 1 year from GT. Moreover, 2 β-Thal transfusion independent pediatric patients showed EPO normal levels and a marked erythroid commitment. Thus, EPO levels negatively correlate with the therapeutic outcome of the therapy, albeit partially, suggesting that other factors could modulate the behavior of HSPCs in β-Thal. Overall, our data agree with the notion that HSPCs activity is modulated to better respond to the demands imposed by the specific pathological condition. This analysis is illustrated in new ***Figure 4****.*

Thank you for the additional important data on mitogen levels in HSC-GT patients and the additional figure 4. The figures are plotted in log scale, which is helpful for increasing dynamic range for but tends to minimize differences. The data indicate a transient rise in TPO and a sustained elevation of EPO after HSC-GT.

Mitogen-driven proliferation of progenitors can result in higher detection rates of IS in the corresponding mature lineage; even while these IS remain below the limit of detectability in other lineages. This can be seen from the Chao1 species richness estimator:

$$S_{Chao1} = S_{obs} + \frac{n-1}{2n}\frac{f_1^2}{f_2}$$

Where S_{obs} is the number of observed IS, n is the size of the sample, f_r is the number of IS with precisely r reads (usually termed the abundance frequency count). The second summand estimates the number of *unobserved* IS. Higher clonal outputs reduce f_1^2/f_2 and thereby also reduce the number of unobserved IS in that lineage.

(B) Differences in input DNA amounts impact the sensitivity to detecting clones and has the potential to mimic lineage skewing.

Offering specifics about the amount of input DNA for all samples is important. Concomitantly, it is essential to detail any corrections in the inference of the number of HSCs (e.g. via the sample-size-based rarefaction/extrapolation formulae for estimating diversity from a sample of a single assemblage) or that were applied to estimate overlaps in clonal outputs (e.g. via the Good-Turing estimators for the number of species shared between two assemblages).

Attachment to Referee #2's review

This point may be particularly pertinent to the analyses of the sharing ratio, where corrections due to sample size can be important.

We provided details on the amounts of input DNA for all samples in the new Extended Data Table 1.

We corrected the HSC number estimations by dividing by the vector copy number when VCN>1 of the specific cell population under study (which is a surrogate readout of the overall marking level) as previously described (Six E., et al., Blood 2023). Moreover, as suggested by the Reviewer, for calculations involving the sharing levels between CD34+ and cell lineages for the analysis of the output and commitment, we applied the Good Turing model to the analyses to remove the biases caused by consecutive subsampling between assemblages. Finally, we used the Bayesian multivariate linear regression algorithm to model and correct biases induced by different technical confounding factors simultaneously (which also include the amount of DNA used). We used the Bayesian multivariate linear regression algorithm to model and correct biases induced by different technical confounding factors simultaneously. These factors included PCR method, amount of DNA used, dose of CD34+ infused per Kg, VCN, sequencing depth, and patient's gender. To be consistent across the different analyses, we used the same approach to correct the clonal diversity analyses (Figure 1C). We edited Figure 2 and Figure 3 with the new results, and Extended Data Figures 2, 5-6, 9.

Overall, the corrections resulted in slight changes in the sharing levels of all lineages, so the relative proportions among the different clone classes remained essentially the same, with only one exception: before correction, the B cell output in WAS patients was similar T-cell output, but after correction was significantly decreased (although still significantly higher than in the other two clinical programs). This is in line with the notion that the selective advantage provided by WASP expression in T cells is stronger than B cells (Konno A., et al., 2004 Blood and Ferrua F., et al., Lancet Hematology 2019).

As you can appreciate, after Good-Turing and the Bayesian corrections, the differences between lineages (in terms of output and commitment) are enhanced.

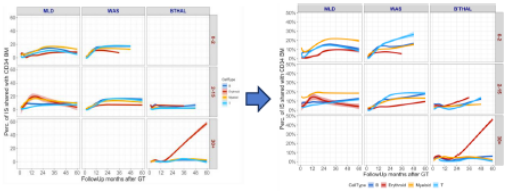


Figure 2D: CD34+ output **before correction**

Figure 2D: CD34+ output **after correction**

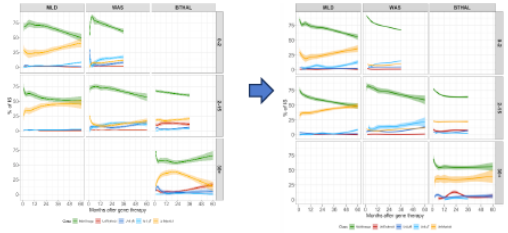


Fig. 3D: lineage commitment **before correction**

Fig. 3D: lineage commitment **after correction**

Attachment to Referee #2's review

Thank you for providing data on the amount of DNA used for all samples. The table suggests a very wide distribution of input DNA:

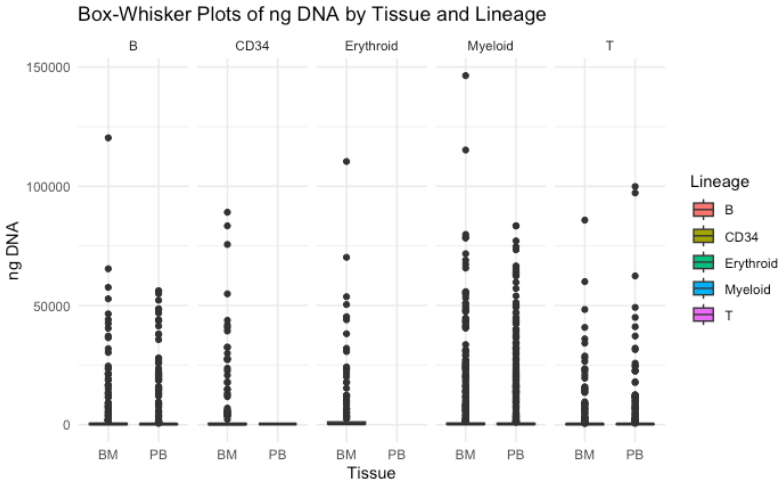


Figure 1 Full distribution of input DNA

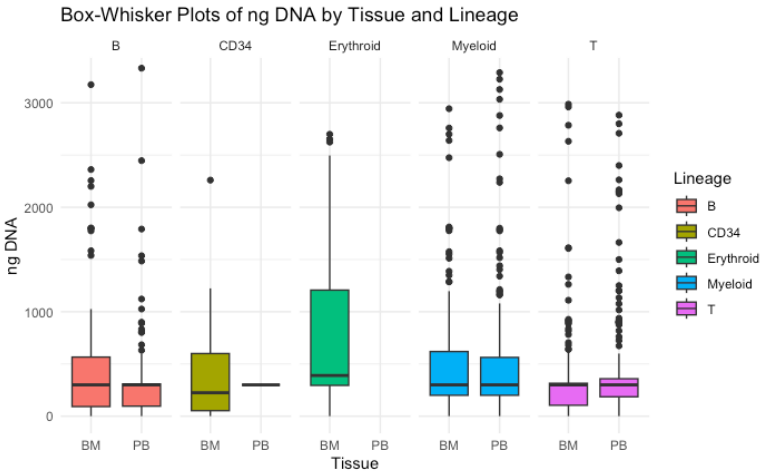


Figure 2Distribution of input DNA, truncated at the 70% quantile.

With the range of input DNA varying over 3 orders of magnitude, sensitivities to IS will vary greatly. The may explain why the number of detected IS fluctuates so much even for a single patient. For instance for MLD Pt43, the number of IS detected from 1 month to 60 months varies from 118 to 1750 with a standard deviation of 426.3, which is 45% of the mean number of IS detected. (This suggests that about half of the IS are not detected at a given interval.). The distribution does not stabilize during 'HSPC homeostasis'. This does not seem biologically plausible and suggests variation in sampling.

Thank you for applying the Good-Turing correction. We do not believe it was carried out properly. The Good-Turing correction to the calculation of multipotent clones is potentially very significant; please see the detailed notes we attach on this point.

(C) Comparisons of IS from cells where clones are geographically segregated (i.e. from the BM) can result in misinterpretions of lineage potential.

Erythroid, Myeloid and B cells in BM are locally produced and the clones are geographically separated for a time after therapy. In primates it can take up to 2 years for clonal geographic segregation to disappear (Verovskaya et al, JEM 2014; Wu et al, JEM, 2018; Chung et al, Blood 2018). Comparisons to cells from contaminating blood or T cells which develop outside the BM may lead to erroneous interpretations with regard to lineage skewing.

We are not sure we understood this point as it is unclear (to us) how the geographical segregation of clones in bone marrow could skew the lineage output and commitment in such a disease-specific fashion. Our IS datasets were obtained before and well beyond a follow-up time of two years, and we do not see relevant differences between the time point from one year to two years after transplant and later time points (as possible to observe in all figures).

Moreover, comparison of the IS retrieved in BM and PB, and their skewing in terms of lineage output as well as commitment showed some differences in WAS patients where T-cells in PB appear to have a higher frequency of lineage commitment compared to the T cells found in BM. Since T cells are produced in the thymus and released in the bloodstream, the analyses on PB could better reflect the lineage commitment. Thus, while the initial geographical segregation of vector marked clones in BM could somehow bias the interpretations on lineage skewing it did not appear to be relevant in our settings in which a large number of samples and periods of follow up well beyond 2 years after therapy have been analyzed with great sequencing depths.

Comparing locally produced erythroid cells in marrow for the erythroid lineage versus myeloid and B cells that are both locally produced and in contaminating blood versus T cells that are all produced outside the marrow is potentially problematic, at least early before mixing has occurred; primate data suggests it can take up to two years before geographic segregation disappears.

(D) Misidentification of multi-insertion clones as uni-insertion clones can result in both (a) overcounting of inferred number of HSC clones and (b) to the extent that multi-insertion clones have low-prevalence and concomitant less complete recovery of all ISs, overcounting of lineage-restricted clones.

Essential are a more complete description of the number of vector insertions per HSC along with an explanation of any corrections applied to the computation of the number of HSCs and their lineage restriction.

We now corrected the HSC number estimations by dividing the HSPC number by the VCN when > 1 as described in Six E., et al Blood 2020. This type of correction allows to avoid the inflation of the HSC numbers caused by clones with multiple integrations (updated Extended Data Table 3, and Figures 1E-F and Extended Data Figure 3A with cut-off time point at 24 months).

Moreover, to eliminate any impact of vector marking levels on lineage skewing we implemented a correction based on a Bayesian multivariate linear regression model which corrects for technical variables including the VCN (together with the PCR method for IS

retrieval and the amount of DNA used, cell dose). The results, after these additional corrections, further reinforced our claims.

Thank you for providing the data on VCN and for introducing a Bayesian multivariate linear regression model to correct for this.

2. Beyond technical issues affecting whether clonal output is truly skewed, there are several other mechanisms besides LT-LGC-HSC *intrinsically* retaining "memory" of pre-gene therapy cell states that could result in putative lineage skewing.

   Plausible mechanisms include:
   (A) Persistence of the pre-gene therapy environment for hematopoiesis

   See point 1(A) above.

   The point is well taken, and we provided an answer to point 1A above.

   Please see our remarks above.

   (B) Differences in the conditioning regimens and their resultant effects on the hematopoietic environment

   The authors note in that different conditioning regimens were used in the therapies for different diseases. A lymphodepleting regimen was given to WAS patients; consequently there was more rapid "filling" of the lymphoid compartment with LGC cells. Filling may have originated from ST-HSPCs that did not persist, thus producing 'uni-lineage' T cells and separate from LT-HSPCs. Once filled homeostatic proliferation maintains T cell clones independent of on-going production from LT-HSPCs. By contrast, much slower refilling of the T cell compartment occurred in BTHAL (where thiotepa and treosulfan were used for conditioning) and MLD (where busulfan was used for conditioning).

   This point deserves fuller disclosure in both the abstract and the conclusions.

   We did not comment on the different repopulation kinetics of the different lineages in each clinical program as they have already been described in previous publications, albeit separately. We now disclose the differences in conditioning and repopulation kinetics in the introduction section as follows:

   "In each of the 3 clinical programs, different conditioning regimens have been adopted. MLD patients, received full myeloablative conditioning5,27,28 with busulfan at doses ranging from 10 to 14 mg/Kg. WAS patients received a reduced-intensity conditioning regimen designed to achieve depletion of HSPC and lymphoid cells consisting of the combined administration of a monoclonal antibody against CD20, busulfan (7.6 to 10.1 mg/Kg) and fludarabine (60 mg/m2)4,19,21,39. In β-Thal patients, the conditioning consisted in thiotepa (6-8 mg/Kg) and treosulfan (14 g/m2) administration10."

   Regarding the possibility that in WAS patients the T-cell reconstitution may have originated from ST-HSPCs that did not persist, thus producing 'uni-lineage' T cells and separate from LT-HSPCs:

We agree that uni-lineage T-cells may arise from already committed HSPCs. This concept applies not only on long-lived uni-lineage committed T-cell clones but also to uni -myeloid, -B and -erythroid restricted clones which are also long-lived probably, although the HSPCs output and commitment vary specifically depending on the disease, age of treatment and disease burden. However, because uni-lineage committed clones persist for years, it is difficult to explain that these are originating from ST-HSPCs.

Moreover, we devised a new analysis to further investigate the dynamics of lineage commitment over time and found that 1/3 of the long-term uni-lineage committed clones originated from multilineage clones that during the early phase of hematopoietic reconstitution (<2 years from transplant) turned into uni-lineage committed and persisted long-term (> 2 years from transplant), while the remaining 2/3 were found to be committed since the early to late phases of hematopoietic reconstitution. The pressure to transit from multilineage to uni-lineage committed was specific for each clinical program.

Thank you for adding details about the conditioning regimen. Differential proliferation in specific subsets of cells in order to restore homeostasis can result in high sensitivity to detecting IS and deserves mention.

(C) Persistence of heritable epigenetic changes intrinsic to HSPCs, which arose prior to treatment

Specific mechanisms might include (a) *intrinsic* disparities in the rates by which HSCs differentiate to specific lineages versus (b) *intrinsic* differences in the proliferation rates of committed progenitor states. If this is a characteristic of HSPCs independent of their environment, this might be apparent in *in vitro* differentiation and proliferation studies.

A more definitive investigation would include either bulk ATAC-seq on HSPCS and selected subsets or scATAC-seq.

We believe that, given the lack of sensitivity of the IS analysis, single cell scRNA-seq and scATAC-seq studies will be more illuminating.

(D) [Probably less likely] persistent changes in the cells comprising the hematopoietic niche; either because these have remained uncorrected or due to their exposure to the original pre-gene therapy environment.

Although a less likely explanation, it could be investigated after other possibilities have been ruled out.

Answer to points C and D:

Our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level studies to unravel biological

differences in each disease condition. These studies are still in progress and will be reported in the future as an independent study.

Thank you. We look forward to reading this study.

3. It would be of interest to show abundance by various classes of clone.

   Beyond the computation of Shannon indices, remaining analyses focus on binary (absence of presence within a lineage) classifications of clones. But lineage abundance fractions offer useful insights and are typically used in murine primate studies. For instance, are the uni-lineage clones small and therefore might sampling be an issue?

   To address this point, we designed a new analysis aimed at unraveling if differences in clonal abundance could bias the linage output or commitment. The description of the analysis reported in the results section is the following:

   "A recent study in nonhuman primates46 demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we conducted comparisons of abundances at early and late time points for clones transitioning from multilineage to uni-lineage, revealing no statistically significant differences (Extended Data Figure 8A). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (Extended Data Figure 8B). Collectively, these findings suggest that, at least in our dataset derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment."

   Therefore, we did not see a significant bias in IS retrieval caused by clonal abundance, probably because we selected only IS with a genome count ≥3 and if captured in at least two time points, (essentially focusing only on robust IS, which does not mean are the most abundant or dominant) and because we analyzed many samples obtained during several years of follow-up.

   Thank you. Since (i) multilineage potential was defined by IS detected in two distinct lineages at any time point and (ii) number of detected IS varied widely by time point, it would be most useful to see the distribution of clonal read counts at the time point it appeared in two lineages versus time points when it did not. Mean [longitudinal] abundances per patient obscure this.

Minor Points:

1. Please note explicitly in the text (rather than figure legends and supplemental text) that HSPC number inference was done with VCN corrections.

   We mentioned in the main text that the HSPC number inferences were done with VCN corrections when the VCN is >1.

Thank you for providing a table of VCNs and the development of a Bayesian multivariate linear regression model to correct for this.

2. For the one WAS and two BTAHL patients with dramatically greater numbers of cumulative ISs, it would be useful to also see a plot of the proportion that were detected over the long-term.

   We calculated the proportion of IS found in these 3 patients. We did not plot these results but mentioned specifically in the result section (from line 218 to 222) as follows:

   "We calculated the percentage of IS detected over the long-term (>24 months after therapy) for the 3 patients that showed a great number of IS. In the WAS patient with ~450,000 IS (Pt17) the proportion of IS detected long term was 4.22%, while in the two β-Thal patients with ~450,000 and ~280,000 IS (respectively Pt36 and Pt41) the proportions were ~3.5% and 1.74% respectively."

   Thank you for highlighting this information for us. [Would be great to know if the more abundant clones were more likely to be detected long term, because this would raise concerns about sensitivity. In the Kiem paper the top 100 clones were all multilineage and sensitivity of ISA was limitation.]

3. Why is 24 months regarded as long-term/stable in many analyses (e.g. Figure 2), but for numbers of HSCs and ISs and their comparisons in Table S2, 12 months is used as the cut-off between short- and long-term. 24 months appears more relevant from the data in the rest of the paper.

   In some analyses, we used the 12 months as a cutoff because some patients had a relatively short follow up (<36 months) and would be excluded from the analysis (12 MLD and 7 WAS and 3 β-Thal). Therefore, to include these patients in the analyses, we reduced the cutoff to 12 months. However, for consistency, we now harmonized the analyses to 24 months and updated figures and data accordingly (Figure 2, Extended data 3, Supplementary Figure S3A). The new results for MLD and WAS clinical programs are essentially the same, showing that there is a significant decrease in HSPC size over time. However, with this cutoff the significant decrease HSPC size was lost in β-Thal patients. Therefore, the drop in HSPC size in β-Thal patients occurs before the other two clinical applications, possibly because were transplanted with mobilized CD34+ cells which lead to a faster recovery and stabilization when compared to BM-derived CD34+ cells. For this reason, for β-Thal patients we show two panels with the cutoff to 12 months and 24 months as well.

   Thank you for this change. [The reason for showing two cutoffs for BTHAL patients deserves mention in the discussion if not already there.]

4. The word "cell" is confusingly used synonymously to individual IS (e.g. figure 2A and in the methods supplement)

   We harmonized the terminology.

Attachment to Referee #2's review

Thank you for the changes which help to make the presentation clearer.

Summary Reviewer Opinion: This is an important paper, based on a huge, long-term and rich dataset that will be of interest to both gene therapy and hematopoiesis investigators. Important conclusions include the observation that, for all disease conditions, the number of active HSPCs is positively correlated to the dosage of CD34+ cells without evident plateau. The hypothesis that the prior disease condition imprints LT-HSCs is interesting but, given the shortcomings of the IS analysis, is not convincingly supported and should be stated in a more qualified manner.

Comparison of Assemblages

March 22, 2024

1 Introduction

The following is a summary and mild generalization of methods described in A. Chao et al, Ecology 98(2017):2914-2929. The methods were originally developed by Turing and Good for cryptography but have found wide application in ecology and more recently in clonal lineage tracing. The basic idea is to infer the number of *unobserved* labels (species in ecology; insertion sites or barcodes in clonal lineage tracing) from the tail of the distribution of *observed* label abundances.

Assessment of biodiversity is an active focus in ecology, which is where many of the methods discussed here were applied. Due to practical limitations it is virtually impossible to detect all species, especially in hyperdiverse settings with an abundance of rare species. Almost certainly some proportion of the species will remain undetected. This has necessitated the development of good statistical tools.

2 Estimating the true number of different labels from a sample

A taxonomically related group of species populations that occur together in space is called an *assemblage* in ecology. Let p_i be the relative abundance of the i-th species and S be the true number of species in the assemblage. Assume that from this assemblage we choose a sample of n individuals with replacement[1]. Let X_i be the abundance of the i-th species in the sample ($\sum_{i=1}^{S} X_i = n$); only species with $X_i > 0$ are detected. One defines the abundance frequency count, f_r, as the number of species that are represented by precisely r individuals in the sample. For instance, f_1 is the number of 'singletons', i.e. the number of species represented in the sample by a single individual.; f_2 is the number of 'doubletons', i.e. the number of species represented by exactly two individuals; etc. The total number of species *observed* in the sample is

$$S_{obs} = \sum_{r>0} f_r \tag{1}$$

[1]To apply these ideas to clonal lineage tracing, we implicitly assume that the sample is small compared to the compartments from which the cells are drawn.

while the number of unobserved species is just f_0; therefore $S = S_{obs} + f_0$.

For a given sample, one can compute the expected ratio of true relative abundances, p_i of all species that appear exactly r times in the sample over their abundance frequency count

$$\alpha_r = \frac{1}{f_r} \sum_{i=1}^{S} p_i I(X_i = r) \tag{2}$$

where $I(C)$ is the indicator function that condition, C is true, i.e $I(C) = 1$, if C is true and 0 otherwise. Here $\sum_{i=1}^{S} p_i I(X_i = r)$ is the total true frequency of all species appearing exactly r times in the sample; dividing by f_r gives the mean per species of the relative frequencies. Note that the coverage deficit, the true proportion of the total individuals that belong to species that were not detected in the sample is given by

$$\alpha_0 f_0 \sum_{i=1}^{S} p_i I(X_i = 0) \tag{3}$$

Turing found an estimator for α_r (never published by Turing; but, with Turing's approval, Good published a proof using a Bayesian approach in 1953).

$$\widehat{\alpha_r} = \frac{r+1}{n} \frac{f_{r+1}}{r_r} \tag{4}$$

Improvements to this estimator were eventually published by Chiu et al (2014)

$$\widehat{\alpha_r} = \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_r + 1} \approx \frac{(r+1)f_{r+1}}{(n-r)f_r} \tag{5}$$

For $r = 0$, the mean population relative frequency for undetected species is approximately

$$\widehat{\alpha_0} \approx \frac{f_1}{nf_0} \tag{6}$$

which, in this form, is not computable from the observed data (f_0 is not observable). Note that the coverage deficit is however computable

$$\widehat{\alpha_0 f_0} \approx \frac{f_1}{n} \tag{7}$$

This estimate was used to establish a lower bound on a the number of unobserved species by Chao et al (1984). From the above approximation, it is apparent that

$$\widehat{\alpha_1} \approx \frac{2f_2}{(n-1)f_1} \tag{8}$$

If the mean relative frequency of all undetected species is less than the mean relative frequency of all singletons (which is intuitively reasonable), then $\alpha_0 \leq$

α_1, and therefore

$$\begin{aligned}\widehat{f_0} =& \frac{\widehat{\alpha_0 f_0}}{\widehat{\alpha_0}} \geq \frac{\widehat{\alpha_0 f_0}}{\widehat{\alpha_1}} \\ =& \frac{\frac{f_1}{n}}{\frac{2f_2}{(n-1)f_1}} = \frac{(n-1)}{n}\frac{f_1^2}{2f_2}\end{aligned} \tag{9}$$

This ultimately leads to the species richness estimator by Chao (often referred to as *Chao1*).

$$S_{Chao1} = \begin{cases} S_{obs} + \frac{n-1}{2n}\frac{f_1^2}{f_2} & \text{if} f_2 > 0, \\ S_{obs} + \frac{n-1}{2n} f_1(f_1 - 1) & \text{otherwise} \end{cases} \tag{10}$$

Chao also estimated the variance

$$\begin{aligned}\text{var}\left(S_{Chao1}\right) =& f_2\left(\frac{n-1}{2n}\right)^2\left(\frac{f_1}{f_2}\right)^4 + f_2\left(\frac{n-1}{n}\right)^2\left(\frac{f_1}{f_2}\right)^3 \\ &+ f_2\left(\frac{n-1}{2n}\right)\left(\frac{f_1}{f_2}\right)^2\end{aligned} \tag{11}$$

which can be used to establish confidence intervals for species richness.

3 Estimating the true number of shared labels from two samples

The above methods can be used in the comparison of two different assemblages (in the case of ecology, this can be a comparison of the biodiversity in two different locations or at the same location at two different time points; in the case of clonal lineage tracing, it can be a comparison in the clonal output in two different lineages or in the same lineage at different points in time). Random samples of sizes n_1 and n_2 are taken from two assemblages. Let S be the total number of species across both assemblages (i.e. the union of species in both assemblages, but not necessarily present in both); and let $X_i^{(\mu)}$ be the observed abundance of the i-th species in the μ-th assemblage. Analogously to the single assemblage case, one defines the number of shared species that are observed r times in assemblage 1 and s times in assemblage 2 as

$$f_{rs} = \sum_{i=1}^{S_{shared}} I\left(X_i^{(1)} = r \wedge X_i^{(2)} = s\right) \tag{12}$$

where S_{shared} is the total number of species shared between the two assembles. Typically, this is not equal to $S_{shared,obs}$, the number of species *observed* to be

shared between both samples. The following are of interest

$$\begin{aligned} f_{r+} &= \sum_{i=1}^{S_{shared}} I\left(X_i^{(1)} = r \wedge X_i^{(2)} > 0\right) = \sum_{s>0} f_{rs} \\ f_{+s} &= \sum_{i=1}^{S_{shared}} I\left(X_i^{(1)} > 0 \wedge X_i^{(2)} = s\right) = \sum_{r>0} f_{rs} \\ f_{++} &= \sum_{i=1}^{S_{shared}} I\left(X_i^{(1)} > 0 \wedge X_i^{(2)} > 0\right) = \sum_{r,s>0} f_{rs} = S_{shared,obs} \end{aligned} \tag{13}$$

The Good-Turing formulas generalize to the two assemblage case

$$\begin{aligned} \widehat{\alpha_{r+}} &= \frac{(r+1)f_{r+1,+}}{(n_1 - r)f_{r+}} \\ \widehat{\alpha_{+s}} &= \frac{(s+1)f_{+,s+1}}{(n_2 - s)f_{+s}} \end{aligned} \tag{14}$$

which may be used to estimate the true number of species shared between the two assemblages as follows:

$$\begin{aligned} S_{shared} &= S_{shared,obs} + f_{0+} + f_{+0} + f_{00} \\ \widehat{S}_{Chao1,shared} &= S_{shared,obs} + k_1 \frac{f_{1+}^2}{f_{2+}} + k_2 \frac{f_{+1}^2}{f_{+2}} + k_1 k_2 \frac{f_{11}^2}{f_{22}} \end{aligned} \tag{15}$$

where $k_\mu = \frac{(n_\mu - 1)}{2n_\mu}$ for $\mu = 1, 2$.

4 Estimating the true number of multipotent clones from four cross-sectional samples

To find the number of multipotent clones, we need to determine the number of IS shared between all four lineages. The generalization to more more that two assemblages is straightforward though it becomes combinatorically complex. This time we define the number of shared species that are observed r_1, r_2, r_3, and r_4 times in each of the four assemblages:

$$f_{r_1 r_2 r_3 r_4} = \sum_{i=1}^{S_{shared}} I\left(I(X_i^{(1)} = r_1 \wedge X_i^{(2)} = r_2 \wedge X_i^{(3)} = r_3 \wedge X_i^{(4)} = r_4\right) \tag{16}$$

where S_{shared} is the [true] total number of species shared between the four assemblages. Again, typically this is not equal to $S_{shared,obs}$, the number of species

observed to be shared between both samples. The following are of interest

$$\begin{aligned}
f_{r_1r_2r_3+} &= \sum_{i=1}^{S_{shared}} I\left(I(X_i^{(1)} = r_1 \wedge X_i^{(2)} > 0 \wedge X_i^{(3)} > 0 \wedge X_i^{(4)} > 0\right) \\
&= \sum_{r_4>0} f_{r_1r_2r_3r_4} \\
f_{r_1r_2++} &= \sum_{i=1}^{S_{shared}} I\left(I(X_i^{(1)} = r_1 \wedge X_i^{(2)} = r_2 \wedge X_i^{(3)} > 0 \wedge X_i^{(4)} > 0\right) \\
&= \sum_{r_3,r_4>0} f_{r_1r_2r_3r_4} \\
f_{r_1+++} &= \sum_{i=1}^{S_{shared}} I\left(I(X_i^{(1)} = r_1 \wedge X_i^{(2)} = r_2 \wedge X_i^{(3)} = r_3 \wedge X_i^{(4)} > 0\right) \\
&= \sum_{r_2,r_3,r_4>0} f_{r_1r_2r_3r_4} \\
f_{++++} &= \sum_{i=1}^{S_{shared}} I\left(I(X_i^{(1)} = r_1 \wedge X_i^{(2)} = r_2 \wedge X_i^{(3)} > 0 \wedge X_i^{(4)} > 0\right) \\
&= \sum_{r_1,r_2,r_3,r_4>0} f_{r_1r_2r_3r_4} = S_{shared,obs}
\end{aligned} \tag{17}$$

We note that similar formulae apply to $f_{r_1r_2+r_4}$, $f_{r_1+r_3r_4}$, $f_{+r_2r_3r_4}$, $f_{r_1+r_3+}$, $f_{+r_2r_3+}$, $f_{r_1++r_4}$, $f_{+r_2+r_4}$, $f_{++r_3r_4}$, f_{+r_2++}, f_{++r_3+}, and f_{+++r_4}.

The Good-Turing formulae generalize to the four assembly case as follows

$$\begin{aligned}
\widehat{\alpha_{r+++}} &= \frac{(r+1)f_{r+1,+++}}{(n_1-r)f_{r+++}} \\
\widehat{\alpha_{rr++}} &= \frac{(r+1)^2 f_{r+1,r+1,++}}{(n_1-r)(n_2-r)f_{rr++}} \\
\widehat{\alpha_{rrr+}} &= \frac{(r+1)^3 f_{r+1,r+1,r+1,+}}{(n_1-r)(n_2-r)(n_3-r)f_{rrr+}} \\
\widehat{\alpha_{rrrr}} &= \frac{(r+1)^4 f_{r+1,r+1,r+1,r+1}}{(n_1-r)(n_2-r)(n_3-r)(n_4-r)f_{rrrr}}
\end{aligned} \tag{18}$$

and corresponding formulae for the other permutations of the indices. Then

$$\begin{aligned}
\widehat{f_{0+++}} &\approx k_1 \frac{f_{1+++}^2}{f_{2+++}} \\
\widehat{f_{00++}} &\approx k_1k_2 \frac{f_{11++}^2}{f_{22++}} \\
\widehat{f_{000+}} &\approx k_1k_2k_3 \frac{f_{111+}^2}{f_{222+}} \\
\widehat{f_{0000}} &\approx k_1k_2k_3k_4 \frac{f_{1111}^2}{f_{2222}}
\end{aligned} \tag{19}$$

where $k_\mu = \frac{n_\mu - 1}{2n_\mu}$ and similarly for the other permutations of the indices. Then the *Chao1* estimator of the multipotenti progenitors is

$$\begin{aligned}
\widehat{S}_{Chao1,shared} =& S_{shared,obs} + f_{0+++} + f_{+0++} + f_{++0+} + f_{+++0} \\
&+ f_{00++} + f_{0+0+} + f_{0++0} + f_{+00+} + f_{+0+0} + f_{++00} \\
&+ f_{000+} + f_{00+0} + f_{0+00} + f_{+000} + f_{0000} \\
\approx& S_{shared,obs} + k_1 \frac{f_{1+++}^2}{f_{2+++}} + k_2 \frac{f_{+1++}^2}{f_{+2++}} + k_3 \frac{f_{++1+}^2}{f_{++2+}} + k_4 \frac{f_{+++1}^2}{f_{+++2}} \\
&+ k_1 k_2 \frac{f_{11++}^2}{f_{22++}} + k_1 k_3 \frac{f_{1+1+}^2}{f_{2+2+}} + k_1 k_4 \frac{f_{1++1}^2}{f_{2++2}} \\
&+ k_2 k_3 \frac{f_{+11+}^2}{f_{+22+}} + k_2 k_4 \frac{f_{+1+1}^2}{f_{+2+2}} + k_3 k_4 \frac{f_{++11}^2}{f_{++22}} \\
&+ k_1 k_2 k_3 \frac{f_{111+}^2}{f_{222+}} + k_1 k_3 k_4 \frac{f_{1+11}^2}{f_{2+22}} + k_1 k_2 k_4 \frac{f_{11+1}^2}{f_{22+2}} + k_2 k_3 k_4 \frac{f_{+111}^2}{f_{+222}} \\
&+ k_1 k_2 k_3 k_4 \frac{f_{1111}^2}{f_{2222}}
\end{aligned} \tag{20}$$

The formula is complex in appearance, but straightfoward to implement via dynamic programming. The terms correcting the naive $S_{shared,obs}$ are likely to be sizable since for IS clonal tracing, f_{1+++}, f_{11++}, f_{111+}, f_{1111}, etc. tend to be sizable.

Referee #3 (Remarks to the Author):

I have reviewed the rebuttal statement and the expanded, revised manuscript. I am satisfied that all the critiques have been adequately addressed and I have no further comments.

Author Rebuttals to First Revision:

Point by Point reply to the Editor and Reviewers to manuscript "*Long-term lineage commitment is modulated by the underlying disease in hematopoietic stem cell gene therapy patients*" (2022-09-14800B)

Thank you for giving us the opportunity to revise our manuscript in response to the positive and constructive comments from the Referees.
As quoted by Reviewer # 1: *Overall, this study presents a valuable dataset based on long term tracking of insertion sites and provides important insights on the HSC dynamics after transplantation.*
Reviewer #2: *This is an important paper, based on a huge, long-term and rich dataset that will be of interest to both gene therapy and hematopoiesis investigators.*
And
Reviewer #3: *I have reviewed the rebuttal statement and the expanded, revised manuscript. I am satisfied that all the critiques have been adequately addressed and I have no further comments.*

As pointed out, Reviewer 1 and especially Reviewer 2 have still some concerns regarding potential biases caused by data heterogeneity (different amounts of DNA used, impact of clonal abundance on the probability of recapturing the same IS in different lineages). In this revised version of the manuscript, we implemented additional analyses to address the above-mentioned remarks and expanded specific sections of the results and discussion as requested by the reviewers.
Changes in the text and figures are highlighted in yellow.

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have appropriately addressed all of my concerns. They have now applied Good-Turing estimation and Bayesian corrections to account for the undetected IS as well as various confounders, which significantly enhance confidence in the conclusions. The authors also added TPO/EPO levels, as well as somatic mutation analysis.

Overall, this study presents a valuable dataset based on long term tracking of insertion sites and provides important insights on the HSC dynamics after transplantation.

The estimation of HSC clonal lineage commitment over time is interesting and I appreciate the author's explanation for how they categorize cells as multi- or unilineage. However, subsampling is an important limiting factor as the authors state. Given this limitation, it would be valuable to add to the discussion about this issue and ways this could be addressed in the future.

We thank the reviewer for the suggestion. We added a short paragraph in the discussion section describing the caveat related to subsampling and the use of optimized statistical analyses to avoid biases caused by confounding variables and sparse data.

> *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. 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*

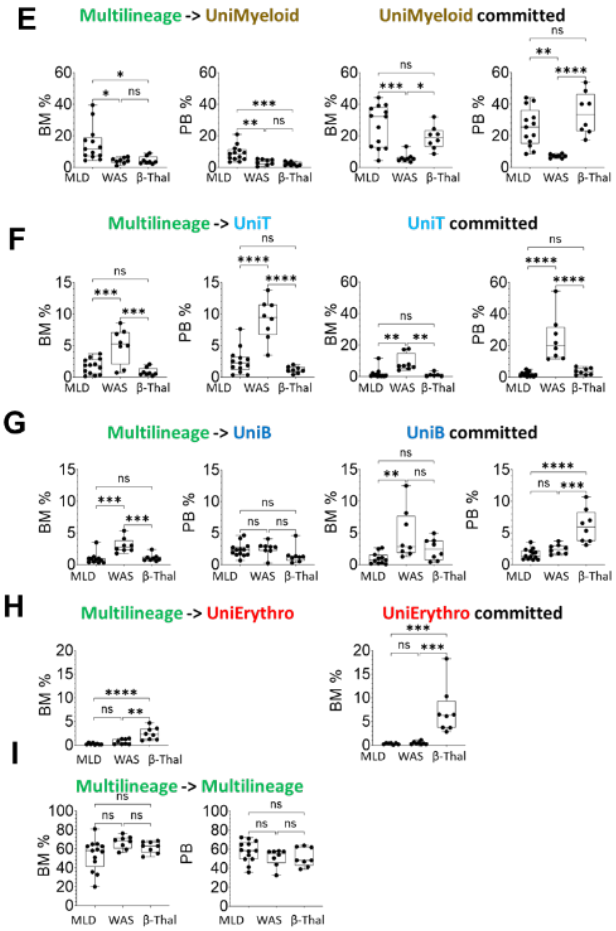
integration site (IS) as either multilineage or unilineage committed, can be addressed through accurate filtering procedures (e.g., based on clonal abundance) and evaluated using bootstrapping methods that provide confidence intervals for each observation. However, analyzing more data reduces the impact 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.

A relevant question: Can the author comment on whether there exist HSC clones that change their commitment over time (for example, from multi-lineage to uni-lineage, or from one lineage to another)? If the current dataset is not sufficient to answer this question, is it possible to address this in the future?

We did not classify as unilineage clones that turned from one lineage into another one because it would be classified as a clone with multilineage potential (please see the Methods section). On the other hand, we did analyze how clones from multilineage turn into unilineage committed clones over time and in the different disease conditions, as shown in new **Figure 3E-I. and Extended Data Fig. 7**. In the result section we describe this analysis and the results as follows:

*"To further investigate the dynamics of lineage commitment over time, we devised a new analysis in which long-lived clones identified at early and late phases of hematopoietic reconstitution (<24 months and >24 months respectively) were selected and classified into classes: clones that transitioned from multilineage to uni-lineage (referred to as multi-uniMyelo, multi-uniT, multi-uniB or multi-uniErythro), or persistently uni-lineage committed (uniMyelo, uniT, uniB or uniErythro) or persistently multi-lineage (Multi-Multi) clones. This single clone level analysis allowed us to compare if and how the different disease conditions impacted the rate of lineage commitment over time as well as the relative contribution of long-lived clones already committed since the early phases of hematopoietic reconstitution. A significantly higher myeloid commitment (multi-uniMyelo transition) in BM and PB of MLD patients (10-12%) was observed when compared to WAS and β-Thal patients (<5%). Myeloid committed clones that remained committed in the late phases of hematopoietic reconstitution (uniMyelo), reached from 20 to 30% in both MLD and β-Thal patients, while in WAS patients were significantly lower (<5%) (**Fig. 3E**). These data indicate that in MLD patients there is a specific bias of multilineage clones to transition into myeloid committed clones, and that permanently committed uniMyeloid clones were present in MLD and β-Thal patients at a significantly higher proportion when compared to WAS patients. The lineage commitment of T cell clones (multi-uniT transition multi-uniB), in BM and/or PB was significantly higher in WAS when compared to MLD and β-Thal patients (**Fig. 3F, G**). Intriguingly, the uniB permanently committed clones in PB were significantly higher in β-Thal when compared to MLD and WAS patients (**Fig. 3G**). On the other hand, the erythroid commitment in β-Thal patients was always higher than in MLD and WAS patients both for transitioning multi-uniErythro and the permanently committed uniErythro clones (**Fig. 3H**). In agreement with our previous findings, the clones preserving multilineage potential over time remained abundant in all clinical programs reaching >50% (**Fig. 3I**). We then addressed the impact of patients' age at treatment on multilineage potential and commitment over time by comparing the percentages of the different linage commitment classes as described above in patients of the same clinical program stratified by age. MLD and WAS patients from the age range 0 to 2 years were compared to the patients from the age range of 2 to 15 years, while β-Thal patients from the age range of 2 to 15 years were compared to patients with > 30 years of age at treatment (**Extended Data Fig. 7A-E**). From this analysis, we found that the 2 to 15 years MLD patients have a significantly higher amount of uniMyeloid committed clones when compared to the younger 0 to 2 years cohort. Similarly, uniMyeloid committed clones in adult (< 30 years) β-Thal patients were significantly higher than the younger 2 to 15 years cohort (**Extended Data Fig. 7A**). Conversely, 2 to 15 years MLD patients showed a significantly decreased multi-uni T cell commitment as well as uniT committed clones when compared to the younger 0 to 2 years cohort (**Extended Data Fig. 7B**). On the other hand, > 30 years β-Thal patients showed a significant decrease in persisting multilineage clones when compared to the younger 2 to 15 years cohort (**Extended Data Fig. 7E**). Age did not*

appear to impact significantly on the commitment of other lineages in any clinical program, at least in the age ranges analyzed in this study".



(E) Box plot representation of myeloid lineage commitment during early and late phases of hematopoietic reconstitution. The first and second panels from left represent the % of multilineage clones transitioning into uni-myeloid committed clones in BM and PB respectively. The third and fourth panels represent the % of unilineage myeloid committed clones that remain committed from early to lathe phases of hematopoietic reconstitution in BM and PB respectively. Comparisons were performed by one-way ANOVA, , p-values are labeled "ns" if $p > 0.05$; $p < 0.05$ = "*"; $p < 0.01$ = "**"; $p < 0.001$ = "***"; $p \leq 0.0001$ = "***"; number or elements reported below each boxplot).

(F) Box plot representation of T-cell lineage commitment during early and late phases of hematopoietic reconstitution. Panel order from left to right and statistical analysis as in E

(G) Box plot representation of B-cell lineage commitment during early and late phases of hematopoietic reconstitution. Panel order from left to right and statistical analysis as in E

(H) Box plot representation of T-cell lineage commitment during early and late phases of hematopoietic reconstitution in BM. The panel on the left represents the % of multilineage clones transitioning into uni-Erythroid committed clones, the panel on the right represents the % of unilineage erythroid clones committed clones that remain committed from early to lathe phases of hematopoietic reconstitution. Statistical analysis as in E.

(I) Box plot representation of multilineage clones that remain multilineage in BM (left panel) and PB (right panel). Statistical analysis as in E.

Figure S7

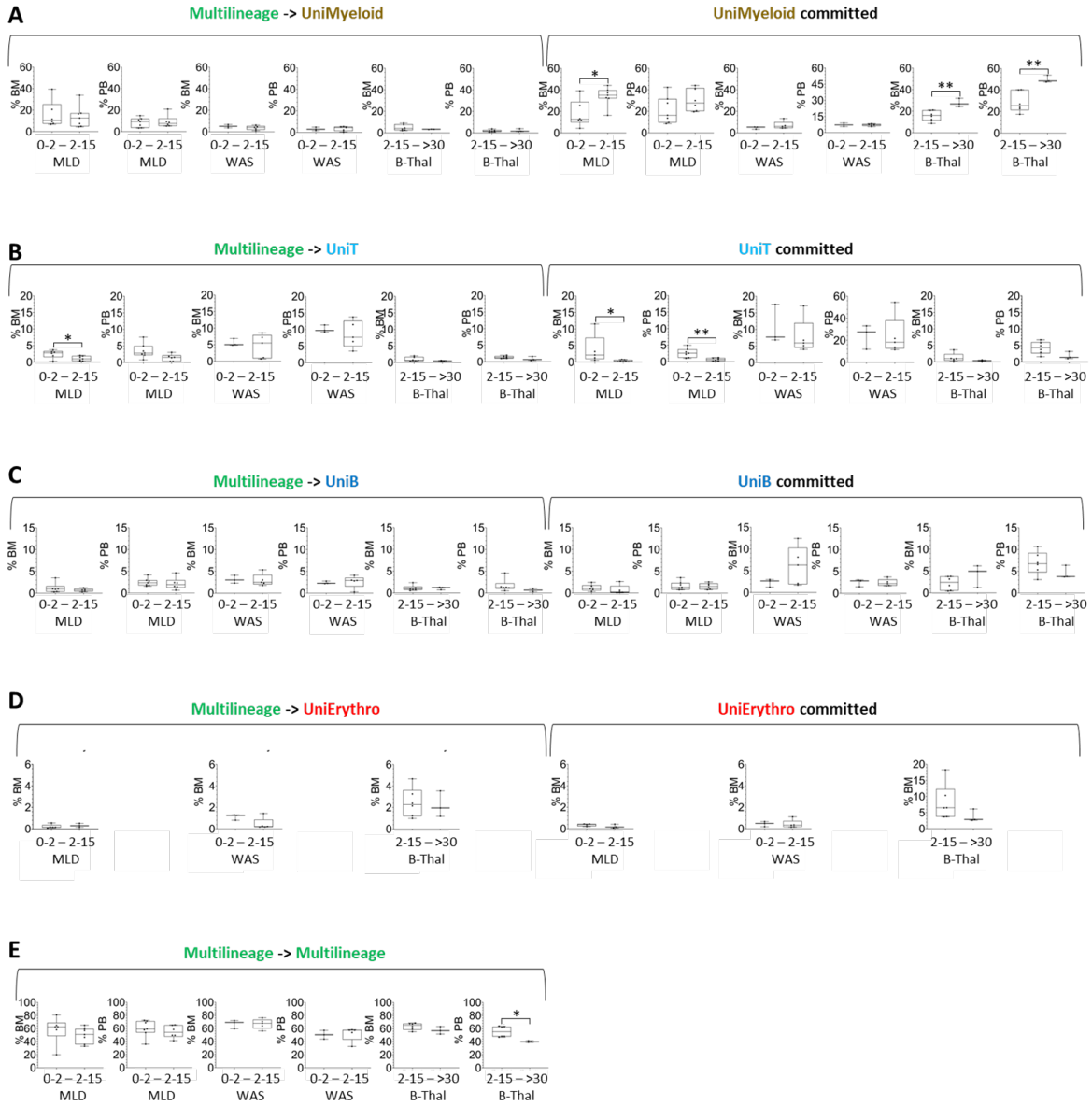


Figure S7. Comparison of lineage commitment during early and late phases of hematopoietic reconstitution in patients stratified by age.

(A) The first 6 panels (form left) represent the percent (y-axis) of multilineage clones identified <24 months from transplant in BM and PB (indicated as % BM and %PB respectively) that transition into uniMyeloid committed clones at the late phases of hematopoietic reconstitution (>24 months after transplant). The remaining 6 panels represent the percent of uniMyeloid committed clones that remain committed in the late phase of hematopoietic reconstitution. Each panel represent the pairwise comparison of linage commitment for each clinical program stratified by age of treatment. For MLD and WAS clinical programs the patients were stratified in by age at treatment 0- 2 years and 2-15 years and for β-Thal 2-15 and >30 years. Comparisons were performed by student's T-test, p-values are labeled "ns" if $p > 0.05$; $p < 0.05$ = "*"; $p < 0.01$ = "**"; $p < 0.001$ = "***"; $p \leq 0.0001$ = "****"; number or elements reported below each boxplot).

(B) Box plot representation of T-cell lineage commitment during early and late phases of hematopoietic reconstitution. Panel order from left to right and statistical analysis as in (A).

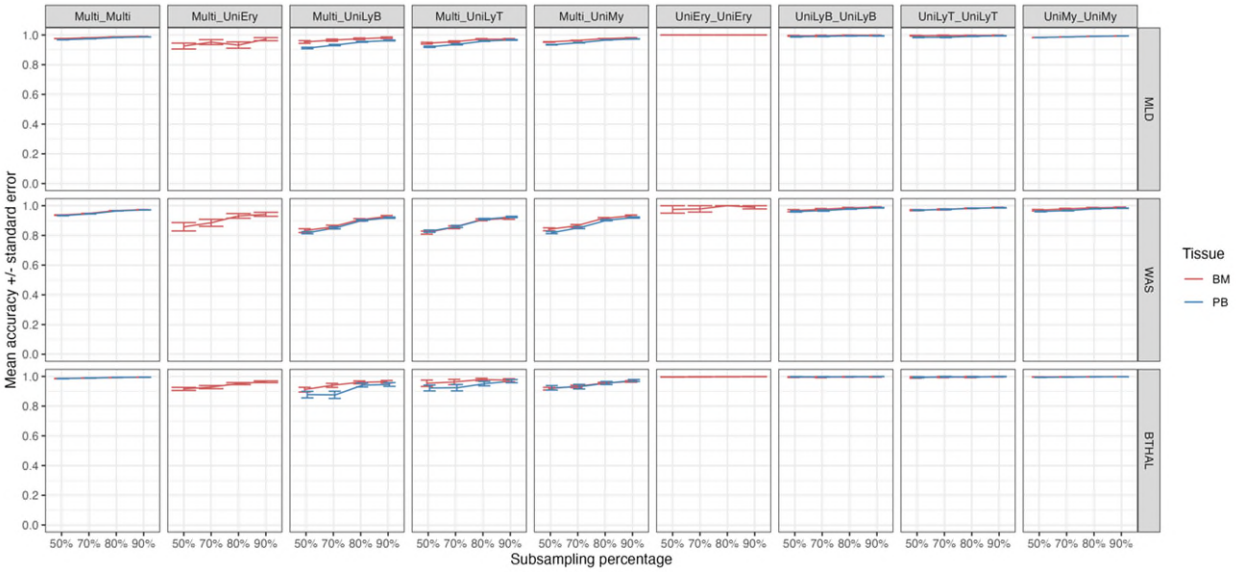
(C) Box plot representation of B-cell lineage commitment during early and late phases of hematopoietic reconstitution. Panel order from left to right and statistical analysis as in (A).

(D) Box plot representation of erythroid lineage commitment during early and late phases of hematopoietic reconstitution in BM. The panel on the left represents the % of multilineage clones transitioning into uni-Erythroid committed clones, the panel on the right represents the % of unilineage erythroid clones committed clones that remain committed from early to lathe phases of hematopoietic reconstitution. Statistical analysis as in (A).

(E) Box plot representation of multilineage clones that remain multilineage in BM and PB. Statistical analysis as in (A).

The author state that 50% of transplanted clones exhibit multilineage behavior. I wonder whether in the HSC clones labeled as "uni-lineage", there are some that could actually be multi-lineage, but appear as "uni-lineage" due to data sparsity or a lack of reads. Is it possible to provide a confidence level for each assignment (label)?

We thank the Reviewer for this observation. Clones with mixed labels (meaning clones that exhibit a mixture of classifications over time) have been processed to avoid misleading conclusions due to potential subsampling. In particular, we first applied specific rules to correct labels (see Methods) in clones that were labeled with a mixture of uni-lineage classes and in clones that exhibited sporadic multilineage classification. For the first ones, we converted in multi-lineage the labels of the mixed timepoints since we imposed a well-established knowledge in HSC biology for which if a stem cell clone can produce different mature lineages is considered multi-lineage. For the second correction, the rationale is that if a clone has been observed multilineage at a certain timepoint, this potential was present up to that timepoint and backward.
Moreover, to account for subsampling, we introduced 2 mathematical models, reducing the impact of confounding factors (the Bayesian model) and to account for the sample size in different assemblages (the Good Turing approach). We are now including extensive documentation of all the models used and relative corrections in the repository (https://github.com/calabrialab/Code_HSPCdynamics/tree/main/code/3.Notebook).
To add a confidence interval for each IS, we performed a bootstrap approach. We sub-sampled reads from our source data at different incremental percentages (50%, 70%, 80%, 90%) with 10 randomizations, to account for sampling variability, and re-processed the analysis of clonal commitment for each patient for each sub-sampled data. We then computed the confidence interval at single IS as the probability of maintaining its state of commitment between early (<24 months) and late (≥24 months). Our results displayed that the commitment per IS stably and consistently maintained a minimum average accuracy > 0.9 from 80% (>0.82 at 50%).



We extended the main text to include this new analysis and added a new figure **Extended Data Fig. 9** with the average accuracy of all CIs for all ISs per commitment state, and a dedicated section in the **Methods**. Moreover, we released the new matrix with per IS accuracy in our GitHub repository (https://github.com/calabrialab/Code_HSPCdynamics/tree/main/data/Accuracy).

In addition, my only minor point is that the layout of some plots could be improved to more clearly present the data.
We have now improved the plots and fonts.

Referee #1 (Remarks on code availability):

The code is clearly presented and is a valuable resource.

We thank the Reviewer for the positive feedback.

Referee #2 (Remarks to the Author):

Major Points

1(A) Thank you for the additional important data on mitogen levels in HSC-GT patients and the additional figure 4. The figures are plotted in log scale, which is helpful for increasing dynamic range for but tends to minimize differences. The data indicate a transient rise in TPO and a sustained elevation of EPO after HSC-GT.

Mitogen-driven proliferation of progenitors can result in higher detection rates of IS in the corresponding mature lineage; even while these IS remain below the limit of detectability in other lineages. This can be seen from the Chao1 species richness estimator:

$S_{Chao1}=S_{obs}+\frac{(n-1)}{2n}\frac{(f_1^2)}{f_2}$

where Sobs is the number of observed IS, n is the size of the sample, fr is the number of IS with precisely r reads (usually termed the abundance frequency count). The second summand estimates the number of unobserved IS. Higher clonal outputs reduce $(f_1^2)/f_2$ and thereby also reduce the number of unobserved IS in that lineage.

We are in full agreement with the Reviewer. Indeed, EPO and TPO have a well-established role in promoting eritro and trombo -poiesis respectively. Our data highlights how these cytokines are driving, at least in part, the hematopoiesis in these patients, likely in response to incomplete correction of the disease.

1(B) Thank you for providing data on the amount of DNA used for all samples. The table suggests a very wide distribution of input DNA:

Figure 1 Full distribution of input DNA (See attached Reviewer Comments pdf)
Figure 2 Distribution of input DNA, truncated at the 70% quantile (See attached Reviewer Comments pdf)

With the range of input DNA varying over 3 orders of magnitude, sensitivities to IS will vary greatly. The may explain why the number of detected IS fluctuates so much even for a single patient. For instance for MLD Pt43, the number of IS detected from 1 month to 60 months varies from 118 to 1750 with a standard deviation of 426.3, which is 45% of the mean number of IS detected. (This suggests that about half of the IS are not detected at a given interval.). The distribution does not stabilize during 'HSPC homeostasis'. This does not seem biologically plausible and suggests variation in sampling.

We appreciate the Reviewer's analysis, and, indeed, we applied the Bayesian regression model to explicitly correct for confounding factors (including the amounts of DNA used and other variables as well). We respectfully disagree with the Reviewer that there is no stabilization over time as we show that the number of novel integrations over time reached a plateau in most patients analyzed. Of course, the variation in the sampling remains also in the late phases of hematopoietic reconstitution which is not related to the lack of stabilization at the biological level but to technical variability, corrected by the Bayesian regression model as described above.

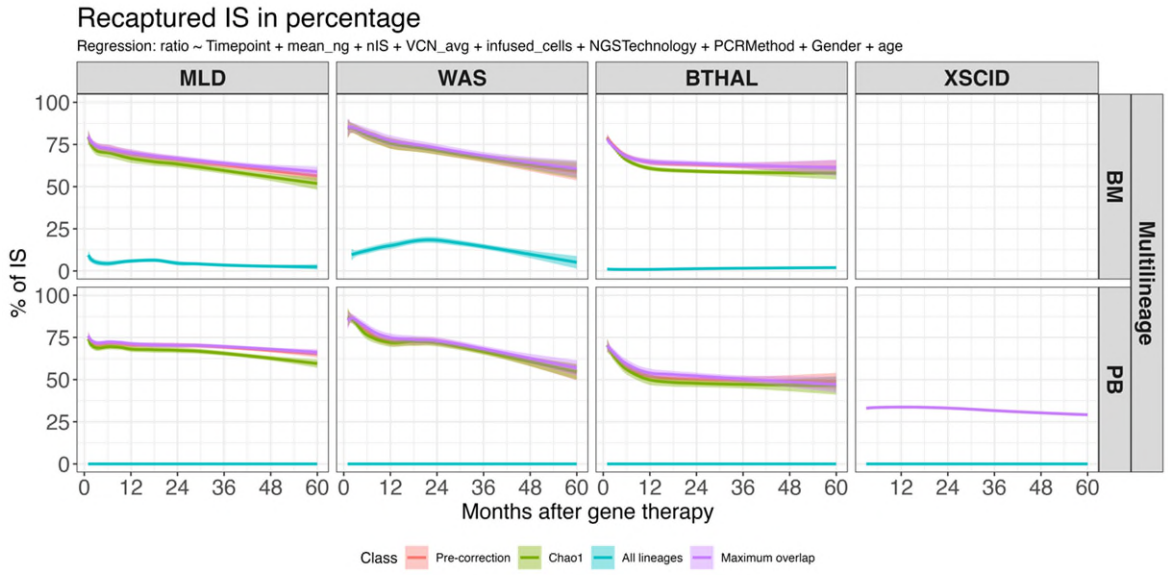
Thank you for applying the Good-Turing correction. We do not believe it was carried out properly. The Good-Turing correction to the calculation of multipotent clones is potentially very significant; please see the detailed notes we attach on this point. (Please see attached Comparison of Assemblages pdf)

We respectfully disagree with the Reviewer. First, at the mathematical level, our analyses are identical to the ones proposed by the Reviewer and were carried out correctly. The only formal difference between our analysis and the one proposed by the Reviewer is the number of lineages to be considered for each assemblage to classify clones as multilineage. We imposed the rule that to be considered multilineage an IS must be retrieved in at least two different mature lineages, while the Reviewer is asking to revise this rule by considering multilineage clones only if shared across all mature lineages, myeloid, erythroid, B- and T cells (quoting the Reviewer: "*To find the number of multipotent clones, we need to determine the number of IS shared between all four lineages*". We think that the new rule proposed by the Reviewer is unnecessarily stringent and will trigger a strong bias in the analyses for several reasons:

(1) There is no reason to define HSPC as a cell where its integration site is found in all 4 lineages. Indeed, if we find an IS shared between mature T cells and erythroid datasets it must be derived from a HSPC with multilineage potential. This is a well-established paradigm in the biology of hematopoiesis.

(2) The need to define multilineage clones as only those that are shared in all four lineages will lead to an artificial underestimation of clones with multilineage potential. Indeed, the probability of retrieving the same integration site in all four lineages will be the product of the probabilities of retrieval in each lineage. In the case of a polyclonal reconstitution (like the one observed in our patients) the chance to capture these events will be exceptionally low, while in the case of patients with oligoclonal reconstitution the probability to recapture these clones will be higher, thus paradoxically we would be more likely able to find multilineage clones in patients with oligoclonal reconstitution.

(3) Even without considering the concerns expressed above, when applying the requirement to define multilineage clones as only those that are shared in all four lineages (as suggested by the Reviewer), the results do not fit with the biology. In the graph below we compared different methods to estimate the relative percent of multilineage clones over time: In all analyses, the variable of interest is defined for each time-point as the ratio between the number of ISs observed in a specific group (i.e., multi-lineage or one of the uni-lineages) and the total number of ISs captured.



a) pre-correction: computed as the raw ratio of the numerator and denominator without carrying out any correction).
b) Chao1: The ratio is computed after performing the Chao1 richness correction.
c) Multi-lineage: to compute the correct value of shared multi-potent species, we apply the Good Turing estimator generalized to four assemblages. In this way, we are considering only the ISs shared by all mature lineages and correcting this value accounting for the number of singletons and doubletons in combinations of assemblages.
d) Maximum Overlap: we apply the Good Turing estimator generalized to multiple assemblages. Differently from the multi-lineage analysis, in this case for each IS we consider the subset of lineages in which we observe the IS. In this way, the correction accounts for the IS abundance in distinct lineages.

By this comparative analysis, it is possible to appreciate that the "*precorrection*", "*Chao1*" and "*maximum overlap*" show similar results (~50% of clones are multilineage over time) while the "*all lineages*" correction showed multilineage clones were essentially absent in all patients and clinical programs. In other words, the patients' hematopoiesis would be driven and sustained only by unilineage clones. This hypothesis and result conflict with biology, with previous reports, and with the clinical outcomes of the patients.
Given these considerations, we believe that the Good-Turing correction as applied is correct and does not require the suggested revision that would (as observed) impact negatively on the interpretation of the biological phenomenon.
We added an exhausting notebook representing our analysis in the code repository of this paper (https://github.com/calabrialab/Code_HSPCdynamics/tree/main/code/3.Notebook; this file must be downloaded and opened in a web browser).

1(C) Comparing locally produced erythroid cells in marrow for the erythroid lineage versus myeloid and B cells that are both locally produced and in contaminating blood versus T cells that are all produced outside the marrow is potentially problematic, at least early before mixing has occurred; primate data suggests it can take up to two years before geographic segregation disappears.

As explained in our previous response the geographical separation does not appear to be an issue for our calculations regarding the lineage output and commitment. Indeed, our analysis goes well beyond the 2 years from transplant, and we did not observe significant variations in our results at early vs late time points.

An important point we need to clarify is that the T cells found in bone marrow are not "contaminating" by any mean. These are patrolling T-cells which exert an important function in immunological defense in different tissues, including bone marrow. Therefore, not analyzing these patrolling clones will be an unjustified omission given their important role in immunity.

1(D) Thank you for providing the data on VCN and for introducing a Bayesian multivariate linear regression model to correct for this.

We appreciated the Reviewer's comment.

2(A) Please see our remarks above.

Regarding this point, we reported our comment in the previous reply "1(B)" including our explanation about the Good Turing model. Moreover, we introduced accuracy measurements to quantify the reliability of our readouts.

2(B) Thank you for adding details about the conditioning regimen. Differential proliferation in specific subsets of cells in order to restore homeostasis can result in high sensitivity to detecting IS and deserves mention.

We thank the Reviewer for this comment.

We added a paragraph in the discussion related to confounding factors and corrections:

> "*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. 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.*"

2(C) We believe that, given the lack of sensitivity of the IS analysis, single cell scRNA-seq and scATAC-seq studies will be more illuminating.

We appreciate the Reviewer's comment for proposing these experiments. As mentioned in our previous revision, our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level studies to unravel biological differences in each disease condition. These studies are still in progress and will be reported in the future as an independent study.

2(D) Thank you. We look forward to reading this study.

Thank you.

3 Thank you. Since (i) multilineage potential was defined by IS detected in two distinct lineages at any time point and (ii) number of detected IS varied widely by time point, it would be most useful to see the distribution of clonal read counts at the time point it appeared in two lineages versus time points when it did not. Mean [longitudinal] abundances per patient obscure this.

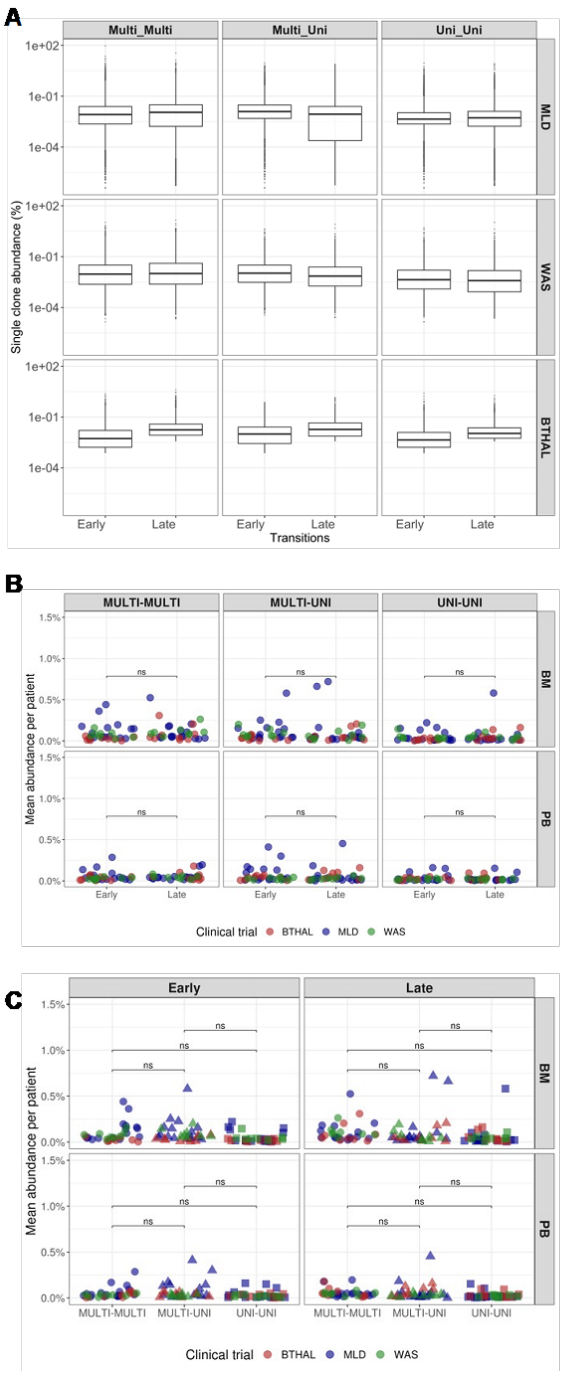
We apologize for the lack of clarity that has likely led to a misinterpretation on how we classify the unilineage and multilineage clones. We classify a clone as multi or uni-lineage at each time point, meaning that if a clone has been captured in different (>=2) mature lineages is classified as multi, otherwise (=1) uni-lineage. The correction by time occurs only after the first classification and removes mixtures of uni-lineage labels that are plausible only in the presence of a multi-lineage clone. Indeed, we have extensively analyzed the issue of clonal abundance and the probability of being shared across multiple lineages. The Good Turing model is in place to correct for mis-classifications.

Moreover, we analyzed at a single clonal level the difference of abundance in early versus late time points, reporting in the final manuscript the statistical analysis of the means using thousands of clones per patient (**Extended Data Fig. 8**). We have now also added the distributions of the abundance of the clones as **Extended Figure 8 A**. These plots showed that the distribution of the clonal abundances (relative percent) between early and late phases in the different assemblages (multi- or uni- lineage classes and transitioning multi-uni) is similar, meaning that multilineage, unilineage, and transitioning multi-uni clones displayed similar distributions of abundances (***Extended Data Fig. 8***).

We have extended the main text with the following paragraph to explain our analyses.

> *A recent study in nonhuman primates[46] demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we analyzed the distributions of the abundance of the clones between early and late phases in the different assemblages (multi- or uni- lineage classes and transitioning multi-uni) (**Extended Figure 8 A**) which resulted similar, meaning that multilineage, unilineage, and transitioning multi-uni clones displayed similar distributions of abundances. Further statistical comparisons revealed no significant differences (**Extended Data Fig. 8B**). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (**Extended Data Fig. 8C**). To further confirm the robustness of our findings, we added a confidence interval to each IS generated through a bootstrap approach using incremental percentage of reads' sampling (50%, 70%, 80%, 90%) and 10 randomizations (see **Methods**), obtaining in each class (multi or uni-lineage) CI > 0.9 on average (**Extended Data Fig.9**). Collectively, these findings suggest that, at least in our dataset derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment.*

Figure S8



Minor Points:

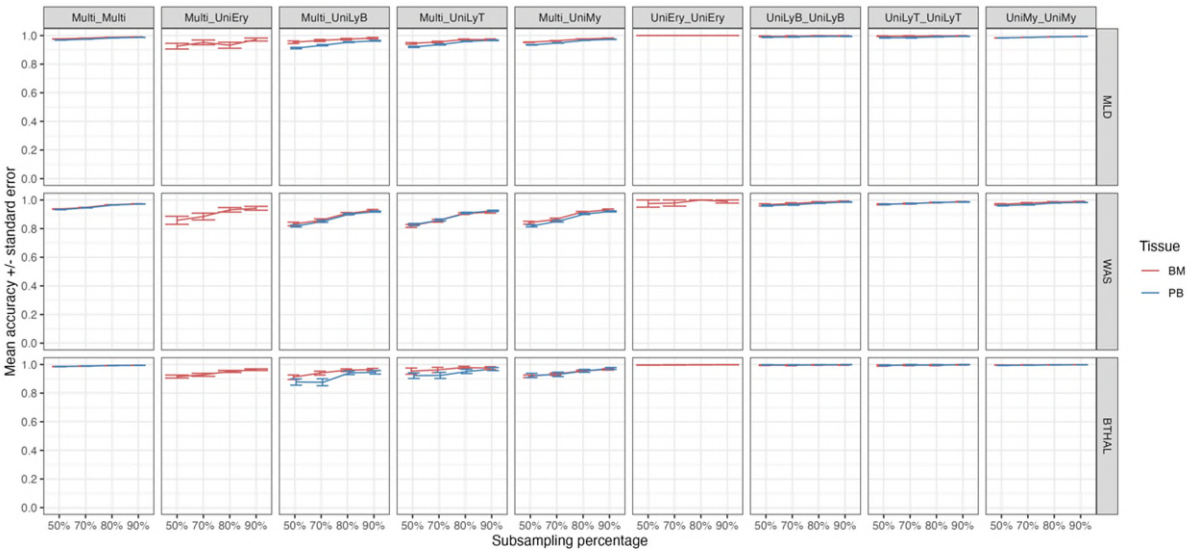
(1) Thank you for providing a table of VCNs and the development of a Bayesian multivariate linear regression model to correct for this.

We appreciated the Reviewer's comment.

(2) Thank you for highlighting this information for us. [Would be great to know if the more abundant clones were more likely to be detected long term, because this would raise concerns about sensitivity. In the Kiem paper the top 100 clones were all multilineage and sensitivity of ISA was limitation.]

Please see our answer to the previous point of the clonal abundances observed in early (<24 months) vs late (≥24 months) timepoints.

To account for clonal abundances and their impact on our core analyses, we tested the robustness of our commitment results by computing a confidence interval per IS. We developed a bootstrap approach to subsample incremental percentage of reads from the source observations (50%, 70%, 80%, 90%) with 10 randomizations, and we re-run the whole procedure of the commitment. We then evaluated the robustness and stability of the assigned label class (multi or uni- lineage) in the two phases of the reconstitution (early, <24 months, versus late, ≥24 months). With this analysis, we confirmed that the average accuracy per IS was at least > 0.82 (in 50%) and >0.9 from 80% subsampling, granting accuracy and the lack of biases caused by subsampling. We have added this analysis in the manuscript (**Results** and **Methods**) with a new figure **Extended Data Fig. 9**.



(3) Thank you for this change. [The reason for showing two cutoffs for BTHAL patients deserves mention in the discussion if not already there.]

We appreciated the Reviewer's comment. We have extended the main text with the following paragraph in Results:

> *"With the cutoff of 24 months after GT to estimate the number of active HSPCs β-Thal patients did not result in a significant decrease in HSPC size compared to the number estimated before 24 months after GT. On the other hand, decreasing the cutoff to 12 months we were able to evidence a significant decrease. Therefore, the drop in HSPC size in β-Thal patients occurs before the other two clinical applications, possibly because were transplanted with mobilized CD34+ cells which lead to a faster recovery and stabilization when compared to BM-derived CD34+ cells. For this reason, for β-Thal patients we show two panels with the cutoff to 12 months and 24 months as well* ***(Extended Data Fig. 3A).***"

(4) Thank you for the changes which help to make the presentation clearer.

We appreciated the Reviewer's comment.

Summary Reviewer Opinion: This is an important paper, based on a huge, long-term and rich dataset that will be of interest to both gene therapy and hematopoiesis investigators. Important conclusions include the observation that, for all disease conditions, the number of active HSPCs is positively correlated to the dosage of CD34+ cells without evident plateau. The hypothesis that the prior disease condition imprints LT-HSCs is interesting but, given the shortcomings of the IS analysis, is not convincingly supported and should be stated in a more qualified manner.

Referee #2 (Remarks on code availability):
The code has been published in a separate manuscript.

Referee #3 (Remarks to the Author):

I have reviewed the rebuttal statement and the expanded, revised manuscript. I am satisfied that all the critiques have been adequately addressed and I have no further comments.

We appreciated the Reviewer's note.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I have carefully revised the revised manuscript. The authors have done a fantastic job of addressing all of my remaining concerns.

Referee #1 (Remarks on code availability):

Clearly documented code.

Referee #2 (Remarks to the Author):

Major Points:

1(A) Author response: We are in full agreement with the Reviewer. Indeed, EPO and TPO have a well-established role in promoting erithro- and thrombopoiesis respectively. Our data highlights how these cytokines are driving, at least in part, the hematopoiesis in these patients, likely in response to incomplete correction of the disease.

Reviewer comments: Environmental signals (including mitogens and cytokines) rather that intrinsic factors (epigenetics) may drive skewed – but not exclusive - production towards deficient mature lineages. Greater abundance of these lineages implies higher detection rates and contributions to other lineages may be below the threshold of detection. We feel this caveat needs to be clearly stated in the discussion.

1(B) Author response: We appreciate the Reviewer's analysis, and, indeed, we applied the Bayesian regression model to explicitly correct for confounding factors (including the amounts of DNA used and other variables as well). We respectfully disagree with the Reviewer that there is no stabilization over time as we show that the number of novel integrations over time reached a plateau in most patients analyzed. Of course, the variation in the sampling remains also in the late phases of hematopoietic reconstitution which is not related to the lack of stabilization at the biological level but to technical variability, corrected by the Bayesian regression model as described above.

[…]

Reviewer response: The Bayesian linear regression modeling is unconvincing. The authors provide no meaningful way to assess whether the nonlinearity between response and predictors was adequately accounted for through a second-order spline transformation. Lacking also are the standard demonstrations of normality of errors, absence of correlations between residuals, homoscedasticity or absence of collinearity.

<Refer to figure supplied by the authors in rebuttal>

The low detection rate of multilineage clones across 4 lineages highlights our main point about limitations in the sensitivity of this data. Data from a variety of animal models shows that on-going multipotent production is readily captured in barcoding experiments and is reflected in a significantly higher rate of clones found across 4 lineages.

2(B) Author response: We thank the Reviewer for this comment.

We added a paragraph in the discussion related to confounding factors and corrections:
“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. 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.”

Reviewer comments: Please refer to our comments about Bayesian linear regression in 1(B). Filtering based on clonal abundance can also introduce selection bias and can artificially inflate effect size and result in loss of information. We recommend against including this paragraph in the paper.

2(C) Author response: We appreciate the Reviewer's comment for proposing these experiments. As mentioned in our previous revision, our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level studies to unravel biological differences in each disease condition. These studies are still in progress and will be reported in the future as an independent study.

Reviewer comments: The data and its statistical analysis leave us unconvinced. This is a huge, long-term and rich dataset that is very valuable. The sensitivity of the analyses here limits them to hypothesis generation. We feel this limitation needs to be better reflected in the discussion.

(3) Author response: We have extended the main text with the following paragraph to explain our analyses.

A recent study in nonhuman primates46 demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we analyzed the distributions of the abundance of the clones between early and late phases in the different assemblages (multi- or uni- lineage classes and transitioning multi-uni) (Extended Figure 8 A) which resulted similar, meaning that multilineage, unilineage, and transitioning multi-uni clones displayed similar distributions of abundances. Further statistical comparisons revealed no significant differences (Extended Data Fig. 8B). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (Extended Data Fig. 8C). To further confirm the robustness of our findings, we added a confidence interval to each IS generated through a bootstrap approach using incremental percentage of reads' sampling (50%, 70%, 80%, 90%) and 10 randomizations (see Methods), obtaining in each class (multi or uni-lineage) CI > 0.9 on average (Extended Data Fig.9). Collectively, these findings suggest that, at least in our dataset derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment.

Reviewer response: Thank you.

Minor Points:

(2) Author response: To account for clonal abundances and their impact on our core analyses, we tested the robustness of our commitment results by computing a confidence interval per IS. We developed a bootstrap approach to subsample incremental percentage of reads from the source observations (50%, 70%, 80%, 90%) with 10 randomizations, and we re-run the whole procedure of the commitment. We then evaluated the robustness and stability of the assigned label class (multi or uni- lineage) in the two phases of the reconstitution (early, <24 months, versus late, ≥24 months). With this analysis, we confirmed that the average accuracy per IS was at least > 0.82 (in 50%) and >0.9 from 80% subsampling, granting accuracy and the lack of biases caused by subsampling. We have added this analysis in the manuscript (Results and Methods) with a new figure Extended Data Fig. 9.

Reviewer response: Thank you.

(3) Author response: We appreciated the Reviewer's comment. We have extended the main text with the following paragraph in Results:

"With the cutoff of 24 months after GT to estimate the number of active HSPCs β-Thal patients did not result in a significant decrease in HSPC size compared to the number estimated before 24 months after GT. On the other hand, decreasing the cutoff to 12 months we were able to evidence a significant decrease. Therefore, the drop in HSPC size in β-Thal patients occurs before the other two clinical applications, possibly because were transplanted with mobilized CD34+ cells which lead to a faster recovery and stabilization when compared to BM-derived CD34+ cells. For this reason, for β-Thal patients we show two panels with the cutoff to 12 months and 24 months as well (Extended Data Fig. 3A)."

Reviewer response: Thank you.

Attachment to Referee #2's review

Reviewer Comments: Long-term lineage commitment is modulated by the underlying disease in hematopoietic stem cell gene therapy patients

Major Points:

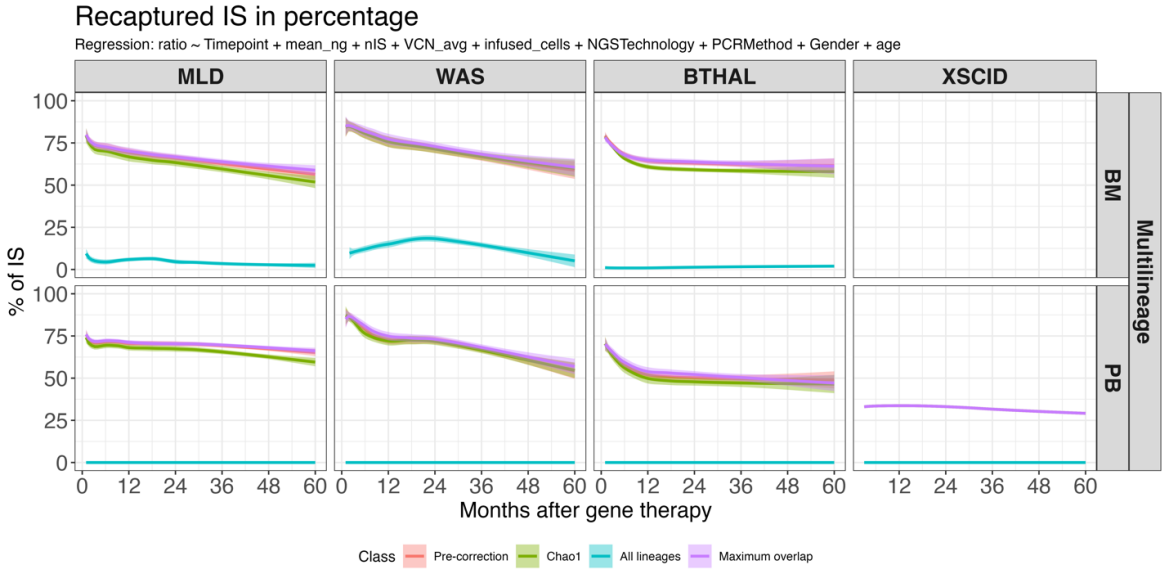
1(A) Author response: We are in full agreement with the Reviewer. Indeed, EPO and TPO have a well-established role in promoting erithro- and thrombopoiesis respectively. Our data highlights how these cytokines are driving, at least in part, the hematopoiesis in these patients, likely in response to incomplete correction of the disease.

Environmental signals (mitogen and cytokine) rather that intrinsic factors (epigenetics) may drive skewed – but not exclusive - production towards deficient mature lineages. Greater abundance of these lineages implies higher detection rates and contributions to other lineages may be below the threshold of detection. We feel this needs to be clearly stated in the discussion.

1(B) Author response: We appreciate the Reviewer's analysis, and, indeed, we applied the Bayesian regression model to explicitly correct for confounding factors (including the amounts of DNA used and other variables as well). We respectfully disagree with the Reviewer that there is no stabilization over time as we show that the number of novel integrations over time reached a plateau in most patients analyzed. Of course, the variation in the sampling remains also in the late phases of hematopoietic reconstitution which is not related to the lack of stabilization at the biological level but to technical variability, corrected by the Bayesian regression model as described above.

[…]

The Bayesian linear regression modeling is unconvincing. The authors provide no meaningful way to assess whether the nonlinearity between response and predictors was adequately accounted for through a second-order spline transformation. Lacking also are the standard demonstrations of normality of errors, absence of correlations between residuals, homoscedasticity or absence of collinearity.



Attachment to Referee #2's review

The low detection rate of multilineage clones across 4 lineages highlights our main point about limitations in the sensitivity of this data. Data from a variety of animal models shows that on-going multipotent production is readily captured in barcoding experiments and is reflected in a significantly higher rate of clones across 4 lineages.

2(B) Author response: We thank the Reviewer for this comment.

We added a paragraph in the discussion related to confounding factors and corrections:

"*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. 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.*"

Please refer to our comments about Bayesian linear regression in 1(B). Filtering based on clonal abundance can also introduce selection bias and can artificially inflate effect size and result in loss of information. We recommend against including this paragraph in the paper.

2(C) Author response: We appreciate the Reviewer's comment for proposing these experiments. As mentioned in our previous revision, our collaborators in this manuscript are analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level studies to unravel biological differences in each disease condition. These studies are still in progress and will be reported in the future as an independent study.

The data and its statistical analysis leave us unconvinced. This is a huge, long-term and rich dataset that is very valuable. The sensitivity of the analyses here limits them to hypothesis generation. We feel this limitation needs to be better reflected in the discussion.

(3) Author response: We have extended the main text with the following paragraph to explain our analyses.

*A recent study in nonhuman primates[46] demonstrated that highly abundant clones are more readily detected in vector integration studies, thus implying that clonal abundance can potentially introduce bias into analyses based on the sharing levels of ISs between assemblies, encompassing lineage output and commitment. To assess the impact of relative clonal abundance on the likelihood of being detected as multilineage or uni-lineage clones, we compared clonal abundances in uni-lineage committed (erythroid, B, T, and myeloid) and clones with multilineage potential during two timeframes (early, <24 months, and late, >24 months). Furthermore, we analyzed the distributions of the abundance of the clones between early and late phases in the different assemblages (multi- or uni- lineage classes and transitioning multi-uni) (**Extended Figure 8 A**) which resulted similar, meaning that multilineage, unilineage, and transitioning multi-uni clones displayed similar distributions of abundances. Further statistical comparisons revealed no significant differences (**Extended Data Fig. 8B**). Subsequently, we compared clonal abundances between early and late datasets in the multilineage, uni-lineage, and multi-uni categories, finding no statistically significant differences (**Extended Data Fig. 8C**). To further confirm the*

*robustness of our findings, we added a confidence interval to each IS generated through a bootstrap approach using incremental percentage of reads' sampling (50%, 70%, 80%, 90%) and 10 randomizations (see **Methods**), obtaining in each class (multi or uni-lineage) CI > 0.9 on average (**Extended Data Fig.9**). Collectively, these findings suggest that, at least in our dataset derived from a comprehensive set of samples and time points along with ultradeep sequencing, clonal abundances did not significantly influence lineage output or commitment.*

Thank you.

Minor Points:

(2) Author response: To account for clonal abundances and their impact on our core analyses, we tested the robustness of our commitment results by computing a confidence interval per IS. We developed a bootstrap approach to subsample incremental percentage of reads from the source observations (50%, 70%, 80%, 90%) with 10 randomizations, and we re-run the whole procedure of the commitment. We then evaluated the robustness and stability of the assigned label class (multi or uni- lineage) in the two phases of the reconstitution (early, <24 months, versus late, ≥24 months). With this analysis, we confirmed that the average accuracy per IS was at least > 0.82 (in 50%) and >0.9 from 80% subsampling, granting accuracy and the lack of biases caused by subsampling. We have added this analysis in the manuscript (**Results** and **Methods**) with a new figure **Extended Data Fig. 9**.

Thank you.

(3) Author response: We appreciated the Reviewer's comment. We have extended the main text with the following paragraph in Results:

*"With the cutoff of 24 months after GT to estimate the number of active HSPCs β-Thal patients did not result in a significant decrease in HSPC size compared to the number estimated before 24 months after GT. On the other hand, decreasing the cutoff to 12 months we were able to evidence a significant decrease. Therefore, the drop in HSPC size in β-Thal patients occurs before the other two clinical applications, possibly because were transplanted with mobilized CD34+ cells which lead to a faster recovery and stabilization when compared to BM-derived CD34+ cells. For this reason, for β-Thal patients we show two panels with the cutoff to 12 months and 24 months as well **(Extended Data Fig. 3A).**"*

Thank you.

Author Rebuttals to Second Revision:

Revision 3, Point by point reply to Reviewers' comments (Manuscript ID 2022-09-14800).

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)

Point by point reply

Referees' comments:
Referee #2 (Remarks to the Author):

We would like to renew our thanks to the Reviewer and here we provide the changes we have made.

Major Points:
1(A) Reviewer comments: Environmental signals (including mitogens and cytokines) rather that intrinsic factors (epigenetics) may drive skewed – but not exclusive - production towards deficient mature lineages. Greater abundance of these lineages implies higher detection rates and contributions to other lineages may be below the threshold of detection. We feel this caveat needs to be clearly stated in the discussion.

To address this point, we introduced the following sentence in the discussion:
"*Environmental signals, such as mitogens and cytokines, together with intrinsic or epigenetic factors, may lead to a biased but not exclusive production of deficient mature*

lineages. The increased abundance of these lineages suggests higher detection rates, while contributions to other lineages might fall below the detection threshold."

1(B) Reviewer response: The Bayesian linear regression modeling is unconvincing. The authors provide no meaningful way to assess whether the nonlinearity between response and predictors was adequately accounted for through a second-order spline transformation. Lacking also are the standard demonstrations of normality of errors, absence of correlations between residuals, homoscedasticity or absence of collinearity.

In our results and methods section, we aimed to avoid overwhelming readers with excessive information. However, in response to the Reviewer's request, we have now added the analytical checks for the regression as a Python notebook, which is now available in our code repository:
https://github.com/calabrialab/Code_HSPCdynamics/blob/bdd7e4ec6dd9046f69e884ad106c3fc83ae88c09/code/Notebook/check_regression.html

The low detection rate of multilineage clones across 4 lineages highlights our main point about limitations in the sensitivity of this data. Data from a variety of animal models shows that on-going multipotent production is readily captured in barcoding experiments and is reflected in a significantly higher rate of clones found across 4 lineages.

We agree with the reviewer that several published animal models have shown a greater sharing of vector-marked cells (or barcodes). The barcoding technology has an intrinsic higher level of sensitivity compared with IS analysis (Adair et al, Molecular Therapy Methods and Clinical Dev 2020, DOI: https://doi.org/10.1016/j.omtm.2020.03.021). We acknowledge that this reduced sensitivity of IS analysis may have an impact to our data as follows:

"*It is important to note that we classify HSPC clones as multilineage if they are detected in at least two mature lineages, rather than in all four lineages analyzed. Our choice is driven by the high polyclonality of our patients which reduce of recapture in all four linages, whereas barcoding tracking studies in animal models displayed higher sensitivity.*"

2(B) Please refer to our comments about Bayesian linear regression in 1(B). Filtering based on clonal abundance can also introduce selection bias and can artificially inflate effect size and result in loss of information. We recommend against including this paragraph in the paper.

We thank the Reviewer for this suggestion. Since Reviewer 1 requested the addition of this paragraph in revision 2022-09-14800B, and considering this Reviewer's feedback, we have determined that the best compromise is to relocate the paragraph to the Supplementary Discussion. We also recognize that filtering clonal abundance could introduce biases. However, the potential noise from not filtering IS based on the number of reads (<3) could lead to even greater biases and has been utilized in several previous clonal tracking studies (i.e. Biasco et al, Cell Stem Cell 2016, doi: 10.1016/j.stem.2016.04.016).

2(C) The data and its statistical analysis leave us unconvinced. This is a huge, long-term and rich dataset that is very valuable. The sensitivity of the analyses here limits them to hypothesis generation. We feel this limitation needs to be better reflected in the discussion.

We added this sentence in the Discussion:

"*Future experiments aimed at analyzing the expression signatures and epigenetic states of HSPCs subsets in bulk and at the single cell level will allow to better elucidate the biological mechanisms underlying the biological differences in each disease condition.*"