blood

American Society of Hematology
2021 L Street NW, Suite 900,
Washington, DC 20036
Phone: 202-776-0544 | Fax 202-776-0545
editorial@hematology.org

Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs

Tracking no: BLD-2019-002350R1

Emmanuelle Six (INSERM U1163, France) Agathe Guilloux (Université Paris-Saclay, Univ Evry, France) Adeline Denis (INSERM UMR 1163, France) Arnaud Lecoules (INSERM UMR 1163, France) Alessandra Magnani (Necker Enfants malades university Hospital, France) Romain Vilette (Evry University, France) Frances Male (University of Pennsylvania, United States) Nicolas Cagnard (Bioinformatics Plateforme University Paris-Descartes, France) Marianne DELVILLE (4. Department of Biotherapy, Necker Children's Hospital, Assistance Publique-Hôpitaux de Paris, France) Elisa Magrin (Hôpital Necker Enfants Malades, France) Laure CACCAVELLI (Necker hospital, France) Cécile Roudaut (Hopital Necker - Enfants Malades, France) Clemence Plantier (Necker Enfants malades university Hospital, France) Steicy Sobrino (Instutut IMAGINE, Hôpital Necker, France) John Gregg (University of Pennsylvania, United States) Christopher Nobles (University of Pennsylvania, United States) John Everett (University of Pennsylvania, United States) Salima Hacein-Bey-Abina (Université Paris Descartes, France) Anne Galy (Genethon, France) Alain FISCHER (INSTITUT IMAGINE, France) Adrian Thrasher (UCL Institute of Child Health, United Kingdom) Isabelle André (Institut Imagine, France) Marina Cavazzana (APHP, France) Frederic Bushman (University of Pennsylvania, United States)

Abstract:
In gene therapy with human hematopoietic stem and progenitor cells (HSPCs), each gene-corrected cell and its progeny are marked in a unique way by the integrating vector. This feature enables lineages to be tracked by sampling blood cells and using DNA sequencing to identify the vector integration sites. Here, we studied five cell lineages (granulocytes, monocytes, T cells, B cells, and natural killer cells) in patients having undergone HSPC gene therapy for Wiskott-Aldrich syndrome or beta hemoglobinopathies. We found that the estimated minimum number of active, repopulating HSPCs (which ranged from 2,000 to 50,000) was correlated with the number of HSPCs per kg infused. We sought to quantify the lineage output and dynamics of gene-modified clones; this is usually challenging because of (i) sparse sampling of the various cell types during the analytical procedure, (ii) contamination during cell isolation, and (iii) different levels of vector marking in the various lineages. We therefore measured the residual contamination and corrected our statistical models accordingly, in order to provide a rigorous analysis of the HSPC lineage output. A cluster analysis of the HSPC lineage output highlighted the existence of several stable, distinct differentiation programs, including myeloid-dominant, lymphoid-dominant and balanced cell subsets. Our study evidenced the heterogeneous nature of the cell lineage output from HSPCs, and provided methods for analyzing these complex data.

Conflict of interest: No COI declared

COI notes:

Preprint server: No;

Author contributions and disclosures: A.M., M.D., E.M., A.F., A.J.T., S.H.-B.-A., A.Ga. and M.C. designed and conducted the clinical trials. E.S., I.A.S, M.C. and F.D.B. designed the experiments. L.C., C.R. ,C.P. and S.S. performed and analyzed the experiments. F.M., J.G., C.N. and F.D.B. sequenced and analyzed the samples. A.Gu., A.D., A.L., R.V., N.C. designed and performed the statistical analysis. The paper was written by E.S., A.Gu., M.C. and F.D.B. All authors discussed the results and commented on the manuscript. Conflict-of-interest disclosure: The authors declare no competing financial interests.

Non-author contributions and disclosures: No;

Agreement to Share Publication-Related Data and Data Sharing Statement: public deposit : - SRA deposit for sequencing data : SRP139090 - GitHub repository for analysis code : https://github.com/BushmanLab/HSC_diversity

Clinical trial registration information (if any):

Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs

Emmanuelle Six*[1,2], Agathe Guilloux[3], Adeline Denis[1,2], Arnaud Lecoules[1,2], Alessandra Magnani[4], Romain Vilette[3], Frances Male[5], Nicolas Cagnard[2,6], Marianne Delville[1,2], Elisa Magrin[4], Laure Caccavelli[4], Cecile Roudaut[4], Clemence Plantier[4], Steicy Sobrino[1,2], John Gregg[5], Christopher L Nobles[5], John K. Everett[5], Salima Hacein-Bey-Abina[7,8], Anne Galy[9,10], Alain Fischer[2,11,12,13], Adrian J. Thrasher[14], Isabelle André[1,2], Marina Cavazzana[1,2,6,15], Frederic D. Bushman*[5,15].

1. INSERM UMR 1163, Laboratory of Human Lymphohematopoiesis, Paris, France
2. Paris Descartes–Sorbonne Paris Cité University, Imagine Institute, Paris, France
3. LaMME, CNRS, Evry University, Paris-Saclay University, Evry, France
4. Biotherapy Clinical Investigation Center, Groupe Hospitalier Universitaire Ouest, AP-HP, INSERM, Paris, France
5. Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, USA
6. SFR Necker, Bioinformatic platform, Paris, France
7. Clinical Immunology Laboratory, Groupe Hospitalier Universitaire Paris-Sud, Kremlin-Bicêtre Hospital, AP-HP, Le Kremlin Bicetre, France
8. UTCBS, CNRS UMR 8258, INSERM U1022, Faculté de Pharmacie de Paris, Université Paris Descartes, Chimie ParisTech, Paris, France
9. Genethon, Evry, France
10. Inserm UMR S951, Evry University, Paris-Saclay University, Evry, France
11. Pediatric Hematology-Immunology and Rheumatology Unit, Necker-Enfants Malades Hospital, Assistance Publique-Hôpitaux de Paris (APHP), Paris, France
12. INSERM UMR 1163, Paris, France
13. College de France, Paris, France
14. Great Ormond Street, Institute of Child Health, Molecular and Cellular Immunology, University College London, UK.
15. These authors contributed equally

* Corresponding authors: emmanuelle.six@inserm.fr (E.S.), bushman@pennmedicine.upenn.edu (F.D.B.)

Running title: HSPC clonal tracking in gene therapy patients

Key Points

In the context of gene therapy, the estimated number of active, repopulating HSPCs was correlated with the number of HSPCs per kg infused.

An analysis of human HSPC clonal lineage outputs highlighted the presence of myeloid-dominant, lymphoid-dominant and balanced cell subsets.

Abstract

In gene therapy with human hematopoietic stem and progenitor cells (HSPCs), each gene-corrected cell and its progeny are marked in a unique way by the integrating vector. This feature enables lineages to be tracked by sampling blood cells and using DNA sequencing to identify the vector integration sites. Here, we studied five cell lineages (granulocytes, monocytes, T cells, B cells, and natural killer cells) in patients having undergone HSPC gene therapy for Wiskott-Aldrich syndrome or beta hemoglobinopathies. We found that the estimated minimum number of active, repopulating HSPCs (which ranged from 2,000 to 50,000) was correlated with the number of HSPCs per kg infused. We sought to quantify the lineage output and dynamics of gene-modified clones; this is usually challenging because of (i) sparse sampling of the various cell types during the analytical procedure, (ii) contamination during cell isolation, and (iii) different levels of vector marking in the various lineages. We therefore measured the residual contamination and corrected our statistical models accordingly, in order to provide a rigorous analysis of the HSPC lineage output. A cluster analysis of the HSPC lineage output highlighted the existence of several stable, distinct differentiation programs, including myeloid-dominant, lymphoid-dominant and balanced cell subsets. Our study evidenced the heterogeneous nature of the cell lineage output from HSPCs, and provided methods for analyzing these complex data.

Introduction

Hematopoietic stem cells (HSCs) are defined by their ability to self-renew while producing daughter cells capable of differentiation and thus enabling the sustained production of all blood cell lineages. Literature data from *in vitro* differentiation and transplantation assays in murine models have suggested that HSCs differentiate into multipotent progenitors, which in turn give rise to early committed progenitors that progressively lose their self-renewal ability. The early committed progenitors segregate into common myeloid progenitors and common lymphoid progenitors [1,2]. However, this classical model has been challenged by the identification of other self-renewing progenitors, including lymphomyeloid-restricted progenitors (i.e. cells having lost their megakaryocyte and erythroid potential) and myeloid-restricted progenitors (i.e. cells having retained their long-term myeloid and megakaryocyte potential) [3–7]. Cells may thus lose their multipotency while retaining the ability to self-renew and produce a restricted number of lineages [8]. The classical model has been further challenged by the documented heterogeneity of murine HSC self-renewal and reconstitution [9], and the identification of stem cells that can give rise to cell populations with different myeloid:lymphoid ratios [5,10,11]. Most recently, the combination of genetic barcoding and labeling methods with murine transplantation studies has increased the accuracy of clonal tracking and confirmed the existence of discrete HSC subsets [12–16] and multilineage/oligolineage HSC clones [17].

A clonal tracking study of lentiviral integration sites (ISs) in macaques documented the existence of three groups of HSCs with different myeloid and lymphoid potentials [18]. In the same non-human primate model, Dunbar et al. recently used a quantitative barcoding approach to observe relatively stable, multipotent, long-term, clonal HSC outputs, together with clones whose output was biased toward myeloid or lymphoid lineages [19,20]. Taken as a whole, the results of animal studies suggest that long-lived clones can be subdivided into several functional groups.

In humans, decades of therapeutic stem cell transplantation have shown that long-term repopulating HSCs are part of the CD34+ subset or (according to some studies) the CD133+ cell subset [21] , that comprise a mixture of hematopoietic stem and progenitor cells (HSPCs). Xenotransplantation in immunodeficient NOD-SCID gammaC-/- (NSG) mice can be used as a surrogate to distinguish between committed progenitors on one hand and HSCs capable of long-term engraftment on the other [22]. Barcoding analyses of human CD34+ HSPCs engrafted in NSG mice also suggest that the HSPC potential is heterogeneous in humans [23,24]. However, the long-term repopulation capacity is limited by the animal’s life span, and the interpretation of these data in mice is complicated by a skewing of human cell differentiation towards lymphoid lineages.

Human gene therapy based on the *ex vivo* transduction of CD34+ cells with an integrating vector provides an opportunity to directly track stem cell activity in humans [25]. Integration of the therapeutic vector marks the genome at unique positions in each cell, and this mark is transmitted to the cell’s progeny. Thus, tracking ISs in fractionated blood cell lineages enables the clonal tracking of stem cell progeny. Initial reports on gene therapy trials for Wiskott-Aldrich syndrome (WAS) and metachromatic leukodystrophy (MLD) [26,27] showed that lymphoid lineages, myeloid lineages and bone marrow (BM) CD34+ cells shared ISs. Recently, Biasco *et al.* performed a more detailed analysis of clonal dynamics in four patients from a WAS gene therapy trial by taking advantage of marking with integrated vectors. Their data suggested that reconstitution had occurred in two waves, with a 12-month time interval between cell transplantation and the establishment of steady-state hematopoiesis [28,29]. Hence, the myeloid and lymphoid lineages appeared to have segregated relatively late in development. This pioneering study provided the most in-depth look to date of human hematopoiesis as revealed by IS tracking in gene therapy patients.

In order to investigate human HSPC function in more detail, we applied IS mapping to track HSPC dynamics in six patients from two gene therapy trials in which lentiviral vectors had been used to introduce copies of “healthy” genes. We extended Biasco *et al.*’s study by (i) analyzing more than one disease, (ii) quantifying residual cell contamination and differences in sampling between cell lineages, and (iii) generating a statistical model of the quantitative clonal lineage output. It may be difficult to draw firm conclusions about the biology of human hematopoietic cells when the corrected cells may or may not have a selective advantage over their non-corrected counterparts; hence, studies of patients with different disorders are preferable. Here, we analyzed four patients treated for WAS [30], one patient treated for sickle cell disease (βS/βS) [31] and one patient treated for beta thalassemia (β0/βE)

[32]. Peripheral blood samples were fractionated into five blood cell lineages, and HSPC dynamics and lineage outputs were tracked by using vector ISs as markers.

Methods

Patients

Four patients with WAS were included in the present study, and have been described previously as part of a gene therapy trial [30]. The patient designations used here are the same as in the earlier publication. The two patients with beta hemoglobinopathy (one with βS/βS sickle cell disease [31] and one with β0/βE beta thalassemia [32]) were participating in the HGB205 clinical trial, and were chosen because their long follow-up period. Autologous HSPCs were derived from BM or from mobilized peripheral blood (MPB, Figure 1A) collected by apheresis following the administration of granulocyte colony-stimulating factor (G-CSF) and the CXCR4 antagonist plerixafor [33]. All patients received a myeloablative conditioning regimen that promoted the engraftment of gene-corrected HSPCs. Samples were obtained through gene therapy protocols set up in the Biotherapy Clinical Investigation Center at Necker Children's Hospital (Paris, France). All gene therapy and follow-up protocols were approved by the local institutional review board (CPP Ile-de-France II, Paris, France; reference for WAS: 2014-04-02-MS1; reference for beta thalassemia and sickle cell disease: 2013/35). The number of corrected HSPC infused was determined by multiplying the number of CD34+ cells infused in the patient by the VCN measured in CD34+ cells (for VCN<1).

Identification and quantification of ISs

Integration sites were amplified for sequencing and analyzed using the INSPIIRED pipeline, as described previously [34]. The sites were isolated from each patient's genomic DNA using nested ligation-mediated PCR after unbiased fragmentation (Covaris System). The samples were tracked by dual indexing, cycling through variations of 96 linkers, and error-correcting the barcodes. Sample and library concentrations were quantified using a KAPA SYBR FAST Universal qPCR Kit, and libraries were pooled on the basis of the sample measurements. Libraries were diluted or concentrated to 1–4 nM prior to the Miseq loading protocol. AMPure XP beads were used to purify and concentrate the DNA. Sequenced reads were aligned with the hg38 reference genome (>95% identity) using BLAT (BLAST-like alignment tool v35). Integration sites marked single clones, whose clonal abundances were determined with the SonicAbundance method; the length of the DNA fragments flanking ISs is used to document independent isolations of integration events and thus provides an estimate of the numbers of cells associated with each unique IS [34–36]. Multiple replicates (4 – 20) of each sample were analyzed independently, in order to reduce founder effects during PCR and the stochastic sampling. PCR contamination during library preparation can be a source of error. To suppress PCR crossover, each sheared DNA sample was ligated to a unique linker. A dual barcoding strategy was then used to filter out PCR crossovers [34]. To monitor possible PCR contamination, a total of 41 control samples of human DNA lacking ISs were analyzed in parallel with the patient samples. After sequence acquisition and analysis, 40 control samples did not show any no ISs, and one sample showed two ISs. We concluded that PCR contamination was not a significant confounder in our analysis.

Data sharing statement

The full code listing and the set of post-processed data used for the analysis and the figures are available online at https://github.com/BushmanLab/HSC_diversity.

All the sequence data used in the present study are available in the NCBI SRA (reference: SRP139090).

Full details of methods for other procedures are provided in the Supplemental Methods.

Results

Clonal tracking in patients following gene therapy with integrating vectors

In order to determine the activity of vector-modified progenitor cells, we analyzed four patients treated for WAS [30] and two patients treated for beta hemoglobinopathies [31,32] (Figure 1A; the patients' reference numbers are the same as in the primary publications). All had successfully undergone gene therapy with integrating lentiviral vectors. Autologous HSPCs were derived from either BM (WAS4, WAS5, and the βS/βS patient) or MPB (WAS2, WAS7, and the β0/ βE patient) (Figure 1A).

For each patient, five cell types were sorted from peripheral blood samples: granulocytes, monocytes, T cells, B cells, and natural killer (NK) cells (referred to as G, M, T, B and K, respectively; Supplemental Figure 1). DNA from these samples was analyzed to determine the IS distribution and their clonal abundance, using the INSPIIRED pipeline and the sonicAbundance method [34–36]. Each patient provided samples for at least two time points, with the first around one year after gene correction and the last at least two years after gene correction (Figure 2). The patients showed gene marking in all the blood cell lineages analyzed, as evidenced by the vector copy number (VCN, Supplemental Figure 2A).

At the latest time point, the number of unique ISs detected per patient ranged from 2941 (in the WAS5 patient) to 98185 (in the patient with β0/ βE) (Supplemental Table 1). The number of ISs per cell type varied (Figure 2) with the level of gene marking (i.e. the VCN) and the amount of DNA used to build the library (Supplemental Figure 2).

We observed high clonal diversity for ISs within the sampled cell types, as measured by the Shannon diversity index [37] (Supplemental Figure 3). The diversity values increase with an increase in the numbers of ISs or an increase in the evenness of distribution; hence, a high degree of diversity reflects successful polyclonal gene correction.

To gain an overview of the IS distributions, we first mapped the sites relative to genomic features (Supplemental Figure 4A) and epigenetic marks (Supplemental Figure 4B). In all five cell lineages, the vector integrated preferentially (i) within transcription units and gene-rich regions, and (ii) near marks associated with active transcription - as expected for lentiviral vectors [38–40]. There was no obvious selection during hematopoietic reconstitution, i.e. there were no increases in integration frequency near particular annotated chromosomal features.

To check for clonal skewing caused by insertional mutagenesis, we next looked for increased frequencies of clones harboring ISs near specific genes. To this end, we compared pooled data from the earliest time point with pooled data from the last time point (Supplemental Figure 4C). No obvious selective clonal outgrowth was detected. Thus, if clonal skewing had taken place, it had not strongly affected proliferation between the first and last time points. We did not observe an increase over time in the frequency of integration near cancer-associated genes (Supplemental Table 2).

Estimation of the minimum number of active HSPCs capable of long-term engraftment

In order to estimate the minimum true population size of gene-corrected and active HSPCs, we used the Chao1 estimator [41]. The latter provides an estimate of the minimum total number of unique ISs and thus the minimum total number of gene-corrected cells (see the Supplemental Methods for more details). We concentrated on the latest time point, in order to estimate the populations of long-term repopulating cells and granulocytes; due to their short lifespans (1-5 days), these populations best reflect HSPC dynamics [42]. The estimated minimum population of active, granulocyte-generating HSPCs ranged from 1943 to 52444 (Figure 3A).

The estimated minimum frequency of active, long-term engrafting clones producing granulocytes ranged from 0.0007% to 0.01% of the initial corrected CD34+ cell population; this corresponded to 7 to 98 repopulating HSPCs per 10^6 infused CD34+ cells (Supplemental Figure 5). We observed a positive correlation between the estimated number of active HSPCs in each patient and the total number of corrected CD34+ cells infused (data not shown) but the highest correlation was detected with the number of corrected CD34+ cells per kg infused (Figure 3B).

Analysis of highly active gene-modified clones

In order to model HSPC activity using IS data, one would ideally reconstruct each clone’s cell lineage output from the sequencing data specified by the IS positions. However, the data generation process involves several steps that add uncertainty to the estimated population compositions. We therefore used statistical techniques to model the total populations (Figure 1B and Supplemental Methods).

Quantifying the ISs present in combinations of cell lineages (Figure 4A and Supplemental Figure 6A) revealed clones from all five lineages (493 for WAS4 at m48). However, a large proportion of the clones were present at a low abundance in a single lineage (e.g. the 5267 clones detected in T cells only for WAS4 m48). To assess the efficiency of our method for IS identification, we compared the results from library technical replicates and cell sorting biological replicates [19]. We found that many IS were present in a single replicate only, which highlights the challenge posed by sparse sampling for this type of analysis (Supplemental Figures 7-9). Hence, in subsequent analyses, we focused on highly active clones detected consistently at high levels in the replicates. An assessment of the number of cells sampled showed that clones comprising at least six cells were shared at 97% between replicates.

Analysis of well-sampled clones revealed a diversity of compositions: although most of the clones (n=424 for WAS4 m48) were detected in all five lineages, a significant number of clones were detected in 2 or 1 lineages (n=83 in G and M, and 62 in T for WAS4 m48; Figure 4B and Supplemental Figure 6A). Importantly, we found that the detection of ISs in 4 or 5 cell lineages is not sufficient to define a multipotent HSPC lineage output, since there were sometimes large differences in abundance between lineages (Supplemental Figure 6B). Thus a quantitative investigation of cell lineages is required to discriminate between a multipotent lineage output and a biased lineage output.

A quantitative investigation should also take account of errors introduced during cell sorting and inter-lineage differences in sampling (due to varying levels of gene marking in each lineage, and varying amounts of cellular DNA available for IS analysis; Supplemental Figure 2). In our pipeline, we corrected for both sources of error (Supplemental Methods and Supplemental Table 4); this enabled us to model the cell lineage proportions produced by the most active progenitors within a rigorous statistical framework.

Lineage bias among active HSPCs

In the following analysis, we focused on the corrected data for four patient/time-point combinations with the most active clones (WAS4, WAS5, βSβS and β0/βE at m48, m55, m24 and m48, respectively) (Figure 5A). This corresponded to 331 to 1094 highly active clones (corresponding to 8 to 12% of all the ISs for the patients WAS4, WAS5 and βSβS and 0.3% for the β0/βE patient). The clones contributed between 9% and 69% of the total hematopoietic output (Figure 5B). We first evaluated lineage bias by analyzing Pearson’s correlation coefficient for corrected clonal abundances in pairs of cell types (Supplemental Figure 10A). We observed the strongest correlations for granulocytes vs. monocytes (0.77 to 0.97), intermediate correlations for granulocytes vs. B cells (0.31 to 0.62) and for T cells vs.

B cells (-0.01 to 0.67), and the lowest correlations for T cells vs. the other cell lineages (-0.011 to 0.47).

To investigate these differences further, we evaluated abundance ratios and potential bias toward certain cell types (defined as at least a 10-fold difference between a pair of cell types [19]). For each pair of cell types, ISs can variously be assigned to balanced clones (i.e. bias = 1) or clones with more than a 10-fold bias towards one lineage or another (Figure 5C-D and Supplemental Figure 10B). We mostly detected balanced contributions for granulocytes vs. monocytes. However, our analysis of granulocytes vs. T cells revealed that (i) the majority of clones were biased toward one cell type, and (ii) few clones had a balanced contribution. The long half-life of lymphoid cells (relative to granulocytes) might complicate the interpretation of this result by favoring the accumulation of a lymphoid-biased population. However, an analysis of pooled data from the four available time points for WAS4 gave similar results (Supplemental Figure 11) and thus emphasized the heterogeneity of the HSPC lineage output.

Identification of distinct HSPC subsets

We next analyzed the human HSPC lineage output in the five cell lineages for each highly active clone, and estimated the lineage potential. In an initial analysis, each IS was quantitatively mapped to one of the 31 reference cell type compositions containing all combinations of one, two, three, four or five cell types (Supplemental Figure 12). A Kullback-Leibler divergence analysis was used to define the closest reference composition (see Supplemental Methods); it revealed a broad variety of cell compositions, ranging from unipotent clones to clones contributing to all five lineages.

We next looked for possible HSPC subtypes via unsupervised K-means clustering of the data (see Supplemental Methods). Each patient displayed four to six predominant clusters, each of which was characterized by a specific lineage composition (potentially corresponding to a class of HSPC). Ternary plots showed the various IS clones' respective lineage outputs (Figure 6 and Supplemental Figure 13). The lineage composition of each class is described by the group's centroid (Supplemental Figure 14).

In all four patients, at least half of the active clones (52-80%) clustered within a group whose centroid composition was close to GMBT, GBKT or GMBK (in purple, Figure 6) - as would be expected for multipotent HSPCs. The analysis also revealed another significant lymphoid-dominant group of clones in three patients (i.e. leading almost exclusively to the production of T cells), which appeared at the ternary plot's T apex (in red in Figure 6). This group accounted for 9% to 13% of the IS clones, and was detected in all patients other than the βS/βS individual (in whom a BT group was detected instead). The next most frequent group (accounting for 6% to 19% of IS clones, in three patients) corresponded to myeloid-dominant clones, which appeared at the ternary plot's GM apex (in blue in Figure 6). In the remaining groups of clones, we noted an NK group (in turquoise) in two patients that accounted respectively for 2% and 20% of the clones. Thus, the results of the K-means cluster analysis suggested the coexistence of myeloid-dominant and lymphoid-dominant (mostly T-dominant) HSPC subsets and more balanced, multipotent HSPC subsets in each patient. Further longitudinal sampling will be required to determine whether the absence of a myeloid-dominant subset (in WAS5) or a T-dominant subset (in βS/βS) corresponds to interpatient heterogeneity or was due to the smaller number of clones analyzed (relative to WAS4).

In order to further characterize these distinct HSPC clones, we quantified and compared their ability to produce granulocytes and T cells (Supplemental Figure 15). The granulocyte abundance was similar in the myeloid-dominant group and multipotent groups (WAS4) or was slightly lower in the myeloid-dominant group than in the multipotent groups (βS/βS and β0/βE) (Supplemental Figure 15A). The T cell abundance was similar in the T-cell-dominant group (WAS5) or was slightly higher in the T cell dominant group than in the multipotent groups (Supplemental Figure 15B). These results suggest that multipotent and lineage bias clones produced their progeny with a very similar level of efficiency.

Lastly, we developed a modelling approach and used it to assess our clonal tracking pipeline's limit of detection (Supplemental Figure 16A, and Supplemental Methods). We simulated the whole blood population containing gene-marked cells (Supplemental Figure 16B) as (i) a homogeneous multipotent population or (ii) a heterogeneous population containing three HSPC subsets (60% multipotent HSPC, 20% myeloid-dominant HSPC and 20% T cell dominant HSPC), to which the errors having occurred during sample processing were applied. The results of the simulation indicated that consideration of all detected clones (Supplemental Figure 16C) can give rise to sparse sampling and thus the false detection of lineage-biased clones, as shown by our lineage bias analysis of granulocytes vs. T cells and our cluster analysis. In this sampled population, it was not possible to distinguish between the two HSPC configurations. In contrast, highly abundant clones (Supplemental Figure 16D) closely resemble the whole blood population (Supplemental Figure 16B) in terms of both lineage bias and clustering. Concentrating on highly abundant clones enables one to discriminate between homogeneous and heterogeneous HSPC populations.

By combining an HSPC lineage output analysis with a simulation approach, we were able to highlight the heterogeneity of human HSPCs (i.e. the coexistence of multipotent, myeloid-dominant and lymphoid dominant HSPC subsets).

Changes over time in HSPC subsets

In order to investigate changes over time in HSPC clones, we focused on patient WAS4; four time points were available, and m12 and m48 were sufficiently well sampled (with good correlation between replicates; Supplemental Figure 9) to allow reliable cluster analysis. At m12, 2652 highly abundant clones accounted for 56% of the total cell count (Supplemental Figure 17A-B). A cluster analysis revealed the presence of multipotent, T-cell-dominant, and myeloid-dominant clones (Supplemental Figure 17C). An analysis of clones present at m12 and m48 showed that the proportion of multipotent clones did not change markedly over time (52% at m12 and 72% at m48, Figure 7A). The same was true of the myeloid-dominant clones. In contrast, a large proportion of the T-lymphoid-dominant clones present at m12 (96%, corresponding to 1035 IS) were not present at m48 (Figure 7A), and might therefore correspond to progenitors lacking the ability to self-renew.

We next sought to determine whether the lineage potential of the various clonal subsets were stable over time by concentrating on the fraction of clones shared between m12 and m48. We found that the majority of multipotent clones at m12 (81% GMBK) were multipotent at m48 (GBKT), and that the majority of myeloid-dominant and T-cell-dominant clones at m12 were stable with regard to their lineage output up to m48 (Figure 7B and Supplemental Figure 17D). The stability of these three main HSPC subsets was further confirmed by exploring the changes in the most prevalent clones over the four available time points (Figure 7C). This analysis further supported the existence of various HSPC subsets, as defined by distinct lineage outputs with stable properties.

Discussion

In the present study, we exploited cell marking by vector integration during human gene therapy to analyze human HSPC function. CD34+ HSPCs were gene-corrected *ex vivo* and then transplanted into patients; this enabled us to recover, sort and characterize vector-marked cells from peripheral blood cell lineages. Our results highlighted the difficulties associated with using this type of information to track stem cell activity: differential sampling of cell lineages, sparse sampling of cell populations, and the imperfect separation of cell fractions. We therefore modeled each of these effects and reconstructed the initial cell populations. One novel aspect of our study was the analysis of hematopoiesis in two distinct pathophysiologic contexts, which helps to clarify possible sources of bias related to the selective advantage of the gene-corrected cell populations. We were thus able to draw several inferences about human HSPC function. Our results suggested that in at least two distinct genetic diseases, human hematopoiesis after gene therapy is maintained by several distinct HSPC subsets, rather than a single subset of multipotent HSPCs.

In order to infer human HSPC function under physiological conditions, the various cell lineages should be uniformly marked during gene therapy. This was the case for the beta hemoglobinopathy patients, whose gene-corrected cells did not have a selective advantage. In WAS patients, however, the VCN was higher in lymphoid lineages than in myeloid lineages as a result of the selective advantage of gene-corrected lymphoid cells; this finding is in line with data from a murine model [43]. Hence, this selective advantage could lead to the false detection of lymphoid-biased clones in WAS patients. Our HSPC clonal tracking approach is novel because by normalizing against the vector input, we corrected for the imbalance in marking between cell types and thus increase the reliability of the analysis.

A focus on highly active clones enables the output lineages to be quantified with more confidence, since it circumvents the issue of sparse sampling. This advantage is further emphasized by the results of our simulation, showing that data recovered from highly active clones should closely parallel the true blood HSPC output using our sampling scheme. However, our approach does not address the question of lineage relationships in less abundant clones. Our results also emphasizes the value of analyzing later time points, since ISs detected more than two years after cell infusion are thought to represent HSPCs capable of long-term repopulation [28].

Our analysis of abundant long-term HSPCs suggested the existence of several human HSPC subsets with distinct lineage outputs. We detected multipotent HSPCs (GMBKT and GMBT clusters) accounting for about two-thirds of the total number of clones in the four patients analyzed. We also observed myeloid-dominant clones (accounting for an average of 14% of the clones) in 3 of 4 patients studied here, and thus in the context of two different types of disease. Lymphoid-dominant clones (T-cell-dominant clones, more specifically) accounted for more than 10% of the clones - even in the absence of a selective advantage in patients with a beta hemoglobinopathy. The restricted clones' stability over time suggests that they correspond to intrinsically defined HSPC subsets. However, the true dynamics of the T-cell-dominant subset are more difficult to assess, given the long half-life of T cells. Thus, our finding of biased clonal output in humans mirrors the HSC heterogeneity reported in mice [5,11]. This conclusion is also supported by the long-term results of two clinical trials showing that cells with a limited lineage potential (T-cell-dominant and myeloid-dominant cells in both trials) have long-term self-renewal and proliferation capacities [44].

The existence of lymphoid-biased progenitors with self-renewal capacity in humans has been previously suggested by a comprehensive IS study of peripheral clones and BM HSPC subsets in another WAS gene therapy trial; independent ISs were detected in multilymphoid progenitors and HSCs [29]. By applying a sonication-based IS analysis and error correction, our novel pipeline enabled us to characterize the clonal lineage output through a quantitative measure of HSPC clone size. This approach confirmed the existence of myeloid-biased HSPCs in humans, and identified T-cell-dominant HSPC subsets for the first time. The existence of the latter subsets agrees with the results of a study of mixed chimerism following HSCT transplantation in sickle cell disease patients[45]. The researchers demonstrated that T cell chimerism was not correlated with myeloid or B-cell chimerism, and that donor engraftment was usually lower for T cells than for other lineages. These results suggest that T cell reconstitution might be driven by an independent HSPC subset [45].

However, one can question whether the heterogeneous lineage output in fact reflects abnormal emergency hematopoiesis driven by the transplantation conditioning regimen. A recent study showed that in the absence of irradiation, HSC clones contributed homogeneously to myeloid and B-cell lineages, whereas conditioning resulted in lineage bias after transplantation [46]. In contrast, other studies of steady-state hematopoiesis have also suggested the existence of distinct HSPC lineage outputs [14,15,47,48]. Thus, on the basis of the present work and reports from other researchers, it appears that HSPC lineage output heterogeneity is a part of normal hematopoiesis and is modulated by various intrinsic and environmental factors.

Tracking ISs in gene therapy patients also allowed us to estimate the total number of active HSPCs. Our estimate (from 2,000 to 50,000 clones) was similar to the values of 2,000 to 12,000 derived from WAS and MLD trials [26,27]. This estimate suggests that fewer than 0.01% of the total corrected CD34+ cells infused were active, repopulating clones; this is more than 10 times lower than the frequency of potentially engrafting clones estimated from phenotyping with known human HSC markers (CD34+CD38−CD45RA−CD90+CD49f+) [22,49]. The ex vivo culture probably contribute to reduce the number of long-term repopulating clones and it is very likely that an additional fraction of HSPC clones (perhaps 30% of the total) is quiescent and thus cannot be detected [50,51]. However, we cannot exclude some remaining heterogeneity in the CD34+CD38−CD45RA−CD90+CD49f+ HSC population and the results of other clonal tracking approaches have also suggested that only a small number of active HSPCs contribute to the homeostasis of human hematopoiesis. In an analysis of the age-related change in the X-chromosome inactivation ratio, the steady-state number of HSPC clones was estimated to be around 1,200[52]. More recently, HSPC clonal dynamics were reconstructed with high resolution by tracking somatic mutations identified in BM HSPCs and in their mature peripheral blood progeny; the number of active HSPCs in a healthy donor was estimated to range from 50,000 to 200,000 [53]. This result is in line with the minimum estimate of 50,000 clones in our β0/βE patient engrafted with the highest cell dose (13.6x10^{6} cells/kg). The comparison of gene therapy/transplantation settings with human steady state hematopoiesis (using our present approach and analysis of somatic mutations [53,54]) would be a valuable way of further optimizing gene therapy protocols.

Our HSPC clonal tracking pipeline constitutes a robust platform for comparing different trials and assessing the impact of various factors on long-term hematopoietic reconstitution. In the present study, the number of highly active clone was lower in the β0/βE patient than in the other three patients; this might have been due to the high total number of clones detected or the distinct HSPC source. In both autologous and allogeneic transplantation protocols, cells sourced from MPB have largely replaced those sourced from BM [55,56]. Mobilization using G-CSF (either alone or combined with Plerixafor) increases the CD34+ stem cell content and accelerates the restoration of blood cell counts[57], relative to BM-sourced HSPCs – perhaps as a result of the larger number of progenitors. Many studies have shown that MPB-sourced HSPCs (mobilized with G-CSF alone, plerixafor alone, or a combination of the two) and BM-sourced HSPCs display intrinsic differences in their long-term multipotency and cell cycle characteristics [33,58,59]. Further longitudinal follow-up and the analysis of additional patients will be needed to assess putative intrinsic differences in stem cell sources and their impact on the long-term lineage output. *Ex vivo* culture is another key parameter that might modify HSPC function and lineage output [60–62]. Recent research has suggested that HSPC engineering might benefit from shorter culture times or the use of compounds that stimulate HSPC self-renewal [63,64].

In summary, the present study introduced a rigorous statistical approach for tackling the technical challenges associated with the use of IS data for HSPC tracking. Our results emphasized the heterogeneity of human HSPCs. Looking ahead, the long-term follow-up of gene therapy patients will facilitate the characterization of HSPC subset dynamics and the investigation of hematopoietic hierarchies in humans. Ultimately, this type of assessment might help to optimize the isolation and handling of HSPCs for gene therapy and transplantation.

Acknowledgments

We thank Charles Berry for pivotal guidance during the early phases of this work. We also thank Bluebird bio for support with the beta hemoglobinopathy trial (NCT 02151526), and Généthon for sponsoring the Wiskott-Aldrich syndrome trial (NCT02333760). We are grateful to Olivier Pellé at the Necker Cytometry Facility for assistance with cell sorting. We acknowledge Rawya Zreik for help with statistical analysis. We are grateful to members of the Bushman laboratory for help and advice. This work was funded by NIH grants AI 052845-13, AI 082020-05A1, AI 045008-15, U19AI117950-01, and UMIAI126620 to F.D.B., the Penn Center for AIDS Research, and the European Research Council (ERC Regenerative Therapy 269037 and Gene for Cure 693762) to M.C. This work was also funded by The Wellcome Trust (090233/Z/09/Z) and the NIHR Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust, and University College London (to A.J.T.).

Authorship

A.M., M.D., E.M., A.F., A.J.T., S.H.-B.-A., A.Ga. and M.C. designed and conducted the clinical trials. E.S., I.A.S, M.C. and F.D.B. designed the experiments. L.C., C.R. ,C.P. and S.S. performed and analyzed the experiments. F.M., J.G., C.N. and F.D.B. sequenced and analyzed the samples. A.Gu., A.D., A.L., R.V., N.C. designed and performed the statistical analysis. The paper was written by E.S., A.Gu., M.C. and F.D.B. All authors discussed the results and commented on the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

References

1. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000).
2. Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661–672 (1997).
3. Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295–306 (2005).
4. Arinobu, Y. *et al.* Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell* **1**, 416–427 (2007).
5. Yamamoto, R. *et al.* Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154**, 1112–1126 (2013).
6. Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13**, 102–116 (2013).
7. Miyawaki, K. *et al.* CD41 marks the initial myelo-erythroid lineage specification in adult mouse hematopoiesis: redefinition of murine common myeloid progenitor. *Stem Cells* **33**, 976–987 (2015).
8. Nimmo, R. A., May, G. E. & Enver, T. Primed and ready: understanding lineage commitment through single cell analysis. *Trends Cell Biol.* **25**, 459–467 (2015).

9. Benveniste, P. *et al.* Intermediate-Term Hematopoietic Stem Cells with Extended but Time-Limited Reconstitution Potential. *Cell Stem Cell* **6**, 48–58 (2010).

10. Müller-Sieburg, C. E., Cho, R. H., Thoman, M., Adkins, B. & Sieburg, H. B. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* **100**, 1302–9 (2002).

11. Dykstra, B. *et al.* Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell* **1**, 218–229 (2007).

12. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29**, 928–933 (2011).

13. Grosselin, J. *et al.* Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. *Stem Cells* (2013). doi:10.1002/stem.1383

14. Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542–546 (2015).

15. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **advance on**, (2014).

16. Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229–232 (2013).

17. Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).

18. Kim, S. *et al.* Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. *Cell Stem Cell* **14**, 473–485 (2014).

19. Wu, C. *et al.* Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* **14**, 486–499 (2014).

20. Koelle, S. J. *et al.* Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants. *Blood* **129**, 1448–1457 (2017).

21. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120–136 (2012).

22. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218–221 (2011).

23. Cheung, A. M. S. *et al.* Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* **122**, 3129–3137 (2013).

24. Knapp, D. J. H. F. *et al.* Single-cell analysis identifies a CD33+ subset of human cord blood cells with high regenerative potential. *Nat. Cell Biol.* **20**, 710–720 (2018).

25. Naldini, L. Gene therapy returns to centre stage. *Nature* **526**, 351–360 (2015).

26. Biffi, A. *et al.* Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science* **341**, 1233158 (2013).

27. Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **341**, 1233151 (2013).

28. Biasco, L. *et al.* In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. *Cell Stem Cell* **19**, 107–119 (2016).

29. Scala, S. *et al.* Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.* (2018). doi:10.1038/s41591-018-0195-3

30. Hacein-Bey Abina, S. *et al.* Outcomes following gene therapy in patients with severe Wiskott-Aldrich syndrome. *JAMA* **313**, 1550–1563 (2015).

31. Ribeil, J.-A. *et al.* Gene Therapy in a Patient with Sickle Cell Disease. *N. Engl. J. Med.* **376**, 848–855 (2017).
32. Thompson, A. A. *et al.* Gene Therapy in Patients with Transfusion-Dependent β-Thalassemia. *N. Engl. J. Med.* **378**, 1479–1493 (2018).
33. Broxmeyer, H. E. *et al.* Rapid mobilization of murine and human hematopoietic stem and progenitor cells with AMD3100, a CXCR4 antagonist. *J. Exp. Med.* **201**, 1307–1318 (2005).
34. Sherman, E. *et al.* INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes. *Mol. Ther. - Methods Clin. Dev.* **4**, 39–49 (2017).
35. Berry, C. C. *et al.* INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions. *Mol. Ther. - Methods Clin. Dev.* **4**, 17–26 (2017).
36. Berry, C. C. *et al.* Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* **28**, 755–762 (2012).
37. Shannon, C. E. A Mathematical Theory of Communication," Bell Systems Techm. (1948).
38. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17**, 1186–94 (2007).
39. Schröder, A. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–9 (2002).
40. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, E234 (2004).
41. Chao, A. Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* 265–270 (1984).
42. Pillay, J. *et al.* In vivo labeling with 2H2O reveals a human neutrophil lifespan of 5.4 days. *Blood* **116**, 625–627 (2010).
43. Westerberg, L. S. *et al.* WASP confers selective advantage for specific hematopoietic cell populations and serves a unique role in marginal zone B-cell homeostasis and function. *Blood* **112**, 4139–4147 (2008).
44. Cavazzana-Calvo, M. *et al.* Is normal hematopoiesis maintained solely by long-term multipotent stem cells? *Blood* **117**, 4420–4424 (2011).
45. Magnani, A. *et al.* Extensive multilineage analysis in patients with mixed chimerism after allogeneic transplantation for sickle cell disease: insight into hematopoiesis and engraftment thresholds for gene therapy. *Haematologica* (2019). doi:10.3324/haematol.2019.227561
46. Lu, R., Czechowicz, A., Seita, J., Jiang, D. & Weissman, I. L. Clonal-level lineage commitment pathways of hematopoietic stem cells in vivo. *Proc. Natl. Acad. Sci.* **116**, 1447–1456 (2019).
47. Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
48. Yu, V. W. C. *et al.* Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell* **167**, 1310-1322.e17 (2016).
49. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
50. Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118–29 (2008).
51. van der Wath, R. C., Wilson, A., Laurenti, E., Trumpp, A. & Liò, P. Estimating dormant and active hematopoietic stem cell kinetics through extensive modeling

of bromodeoxyuridine label-retaining cell dynamics. *PLoS One* **4**, e6972 (2009).

52. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460–4466 (2011).
53. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
54. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e22 (2019).
55. Körbling, M. & Anderlini, P. Peripheral blood stem cell versus bone marrow allotransplantation: does the source of hematopoietic stem cells matter? *Blood* **98**, 2900–8 (2001).
56. Saraceni, F., Shem-Tov, N., Olivieri, A. & Nagler, A. Mobilized peripheral blood grafts include more than hematopoietic stem cells: the immunological perspective. *Bone Marrow Transplant.* **50**, 886–891 (2015).
57. Bensinger, W. I. *et al.* Transplantation of bone marrow as compared with peripheral-blood cells from HLA-identical relatives in patients with hematologic cancers. *N. Engl. J. Med.* **344**, 175–81 (2001).
58. Larochelle, A. *et al.* AMD3100 mobilizes hematopoietic stem cells with long-term repopulating capacity in nonhuman primates. *Blood* **107**, 3772–3778 (2006).
59. Lagresle-Peyrou, C. *et al.* Plerixafor enables safe, rapid, efficient mobilization of hematopoietic stem cells in sickle cell disease patients after exchange transfusion. *Haematologica* **103**, 778–786 (2018).
60. Miller, P. H. *et al.* Early production of human neutrophils and platelets posttransplant is severely compromised by growth factor exposure. *Exp. Hematol.* **44**, 635–640 (2016).
61. Glimm, H., Oh, I. H. & Eaves, C. J. Human hematopoietic stem cells stimulated to proliferate in vitro lose engraftment potential during their S/G(2)/M transit and do not reenter G(0). *Blood* **96**, 4185–93 (2000).
62. Larochelle, A. *et al.* Bone marrow homing and engraftment of human hematopoietic stem and progenitor cells is mediated by a polarized membrane domain. *Blood* **119**, 1848–1855 (2012).
63. Zonari, E. *et al.* Efficient Ex Vivo Engineering and Expansion of Highly Purified Human Hematopoietic Stem and Progenitor Cell Populations for Gene Therapy. *Stem Cell Reports* **8**, 977–990 (2017).
64. Fares, I. *et al.* Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science (80-.).* **345**, 1509–1512 (2014).

Figure legends

Figure 1. HSPC lineage tracking in human gene therapy trials.

(A) Characteristics of the gene therapy patients studied here.

WAS: Wiskott-Aldrich syndrome; MPB: mobilized peripheral blood; BM: bone marrow; Bu: busulfan; Flu: fludarabine; VCN: vector copy number; y: year; m: month; AUC: area under the curve; corr.: corrected.

(B) Workflow for the IS clonal tracking analysis, from patient blood sample processing, IS library construction and IS data preparation to functional studies of the HSPCs.

Figure 2. Numbers of unique ISs in gene-corrected patients.

The number of unique ISs for granulocytes, monocytes, B cells, NK cells, and T cells at various time points (in months (m)) after gene therapy. Patients received HSPCs sourced from BM (circles) or MPB (squares).

Figure 3. Estimation of the number of active HSPCs.

(A) The estimated number of active, long-term-repopulating HSPCs, corresponding to the estimated granulocyte population size at last follow-up with a normalized Chao1 estimate (the Chao1 estimate divided by VCN for VCN>1: see the supplemental Methods).

(B) The correlation between the number of active, long-term-repopulating HSPCs (Chao1 estimate of the size of the granulocyte population) and the number of corrected CD34+ cells infused per kg, for each patient.

Figure 4. Clonal detection of ISs in the various cell lineages.

(A) An UpSet plot showing the number of ISs detected in a single lineage and any combination of 2, 3, 4 or 5 cell lineages for patient WAS4 at m48 of follow-up, considering the IS dataset as a whole. The mean abundance in each category is shown by the heat map at the bottom of the graph.

(B) An UpSet plot showing the number of ISs detected in a single lineage and any combination of 2, 3, 4 or 5 cell lineages for patient WAS4 at m48 of follow-up, considering only the highly abundant ISs. The mean abundance in each category is shown by the heat map at the bottom of the graph.

Figure 5. Clonal contributions to various hematopoietic cell types.

(A) The number of highly active HSPC clones for each patient at last follow-up (unique ISs with at least 6 cells of one of the 5 cell types).

(B) Contribution of highly active clones to total hematopoiesis output. The proportion of the total cell count (from all cell types) in the whole IS dataset accounted for by highly active clones is shown.

(C-D) Lineage bias analysis for different combinations of cell lineage abundance and in each patient, represented as a bias histogram. For each IS, the ratio of abundance is calculated for the two cell types and then classified as balanced (1) through a >10-fold difference.

(C) Lineage bias for granulocytes vs. monocytes (G vs M).

(D) Lineage bias for granulocytes vs. T cells (G vs T).

Figure 6. HSPC subsets and HSPC heterogeneity.

Ternary plots for each patient, showing the K-means cluster analysis of highly active ISs and highlighting the presence of 4 to 6 predominant cell type compositions (HSPC subsets). The lineage output inferred for each IS clone is projected onto a triangle that combines the five cell lineages and three axis: GM/T/BK (with the proportion of granulocytes and monocytes on the left axis, the proportion of T cells on the right axis, and the proportion of B+NK cells on the bottom axis). The corrected relative abundance of each clone at a specific position is indicated by the dot density. The total number of IS clones at a specific position is indicated by the dot size. The clusters are named according to their centroid (shown in Supplemental Figure 14), using the closest reference composition (according to the minimum Kullback-Leibler distance to the 31 reference compositions; see the Supplemental Methods). The proportions of IS in each cluster related to the total number of IS clones are indicated.

Figure 7. Dynamic of HSPC subsets.

(A) Proportion of ISs persisting over time in each IS subset.

(B) An alluvial plot showing the changes in the different IS subsets comprising the 768 clones that were present at both m12 and m48.

(C) A heatmap showing the TOP15 IS clones from multipotent, myeloid-dominant and T-cell-dominant subsets at m48, for the four available time points (m12, m36, m48 and m60).

Figure 1

A

	Pathology	Age at treatment	Source	Mobilizing agent	Conditioning regimen	Busulfan AUC	corr. CD34 10^{e}6/kg	Total corr. CD34 infused	VCN in CD34
WAS4	WAS	10m	BM	-	Bu +Flu	18,103	7.3	65.7	2.8
WAS5	WAS	3y	BM	-	Bu +Flu	17,601	4.08	70.584	0.6
βS/βS	Sickle Cell Disease	13y	BM	-	Bu	19,363	5.6	226.8	1 / 1.2
WAS2	WAS	15y	MPB	G-CSF + Plerixafor	Bu +Flu	NA	11	660	1.3
WAS7	WAS	3.5y	MPB	G-CSF + Plerixafor	Bu +Flu	17,204	9	177.3	0.6
β0/βE	Beta-thalassemia	16y	MPB	G-CSF + Plerixafor	Bu	20,848	13.6	557.6	2.1

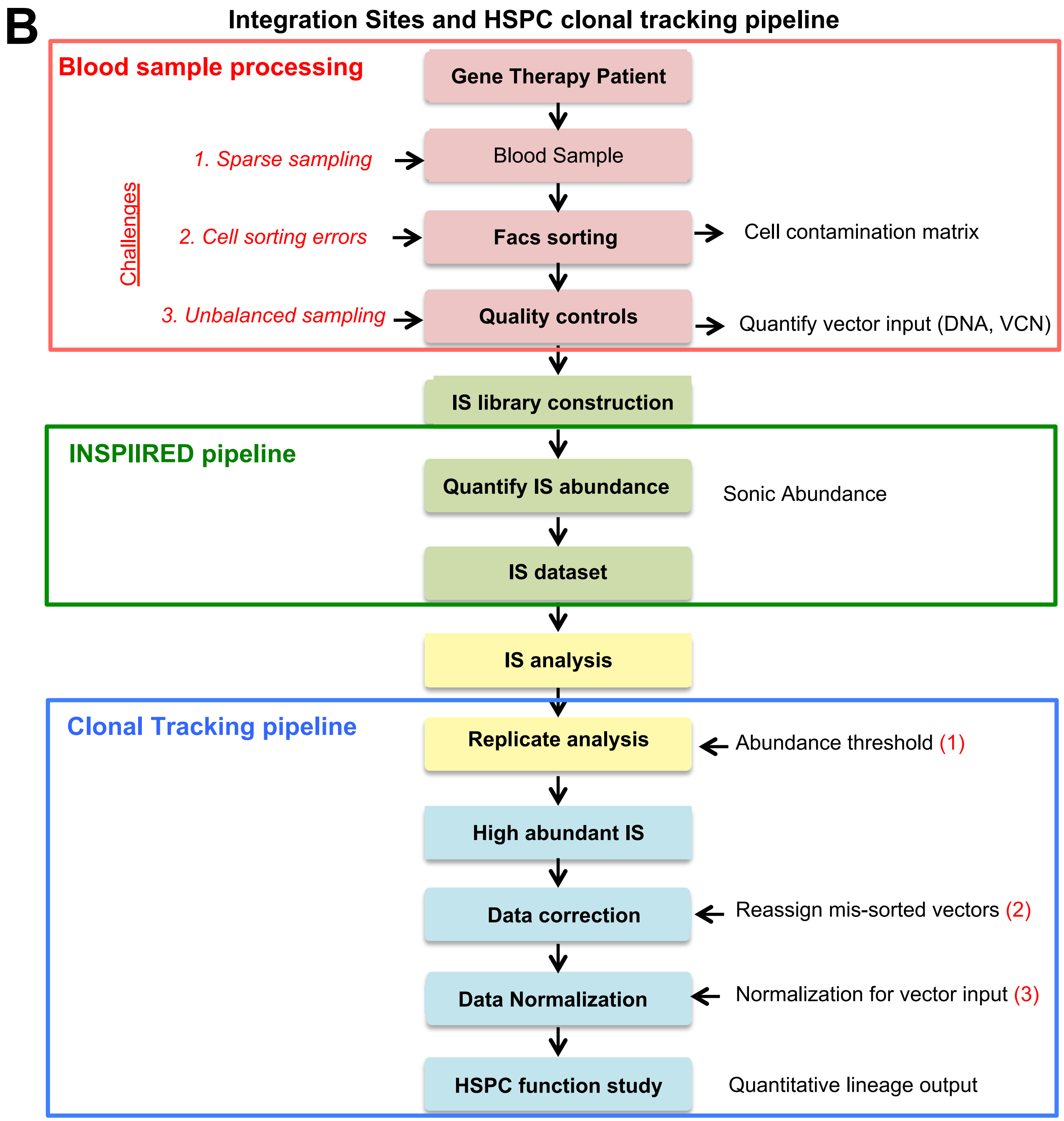


Figure 2

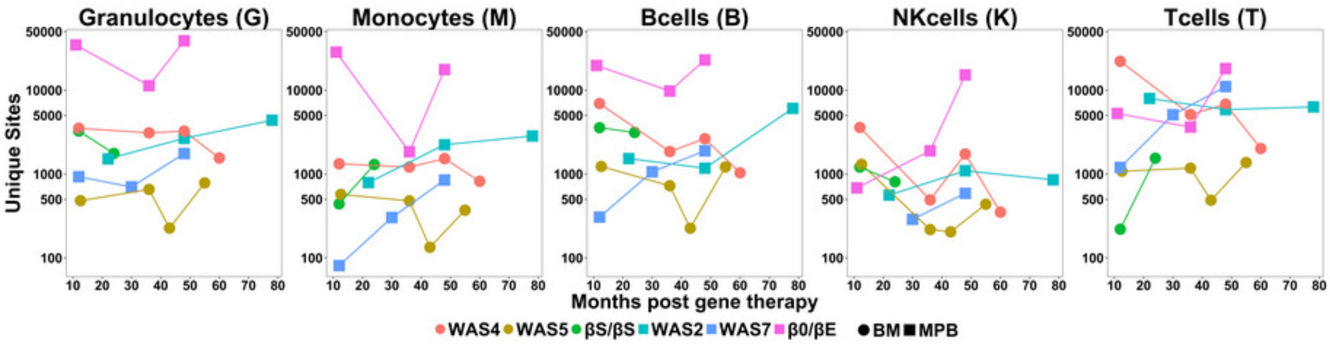

Granulocytes (G)
Monocytes (M)
Bcells (B)
NKcells (K)
Tcells (T)
Unique Sites
50000
10000
5000
1000
500
100
10 20 30 40 50 60 70 80
Months post gene therapy
WAS4
WAS5
βS/βS
WAS2
WAS7
β0/βE
BM
MPB

Figure 3

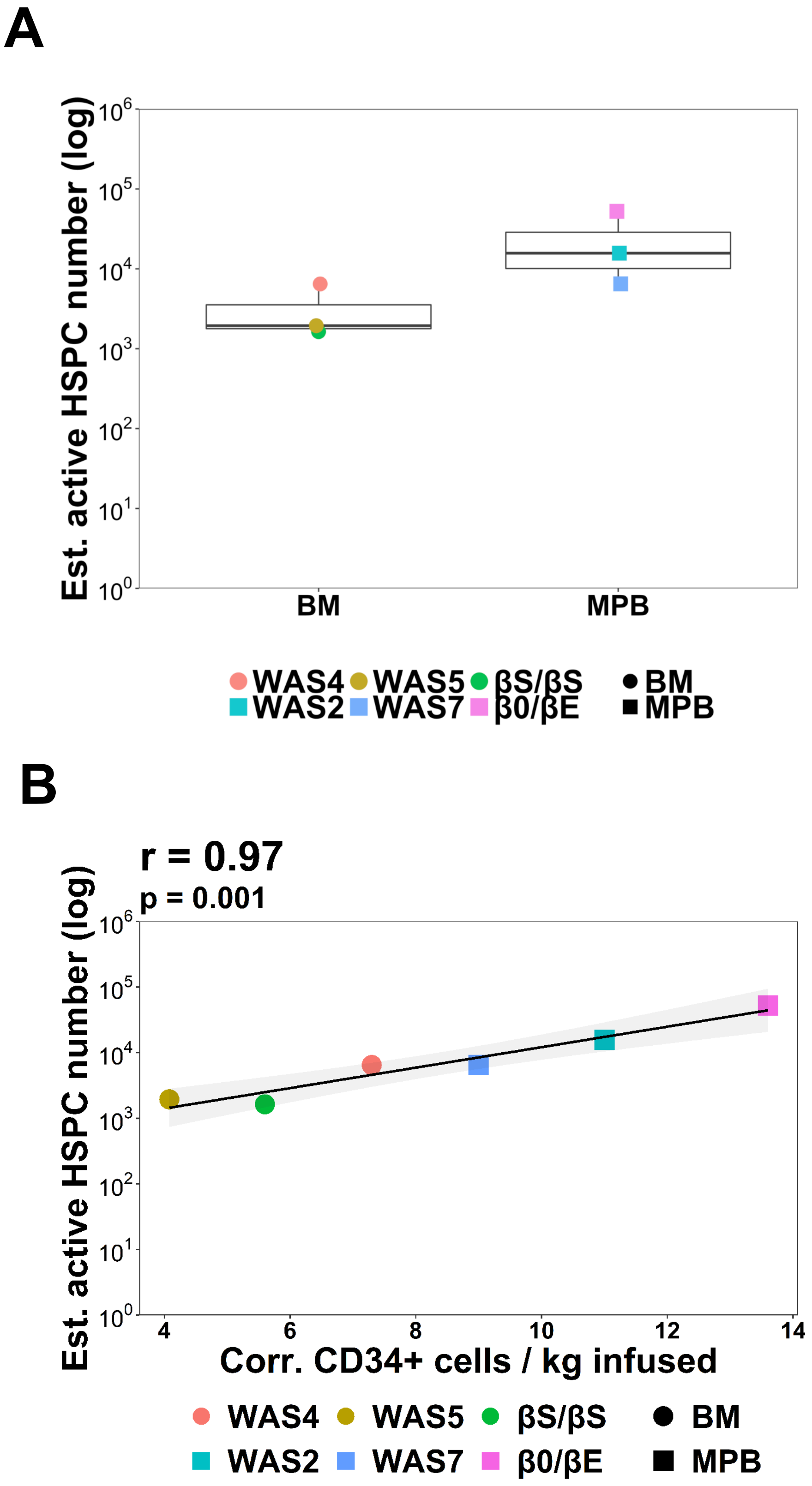

A
Est. active HSPC number (log)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
BM
MPB
● WAS4 ● WAS5 ● βS/βS ● BM
■ WAS2 ■ WAS7 ■ β0/βE ■ MPB
B
r = 0.97
p = 0.001
Est. active HSPC number (log)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
4
6
8
10
12
14
Corr. CD34+ cells / kg infused
● WAS4 ● WAS5 ● βS/βS ● BM
■ WAS2 ■ WAS7 ■ β0/βE ■ MPB

Figure 4

WAS4 m48

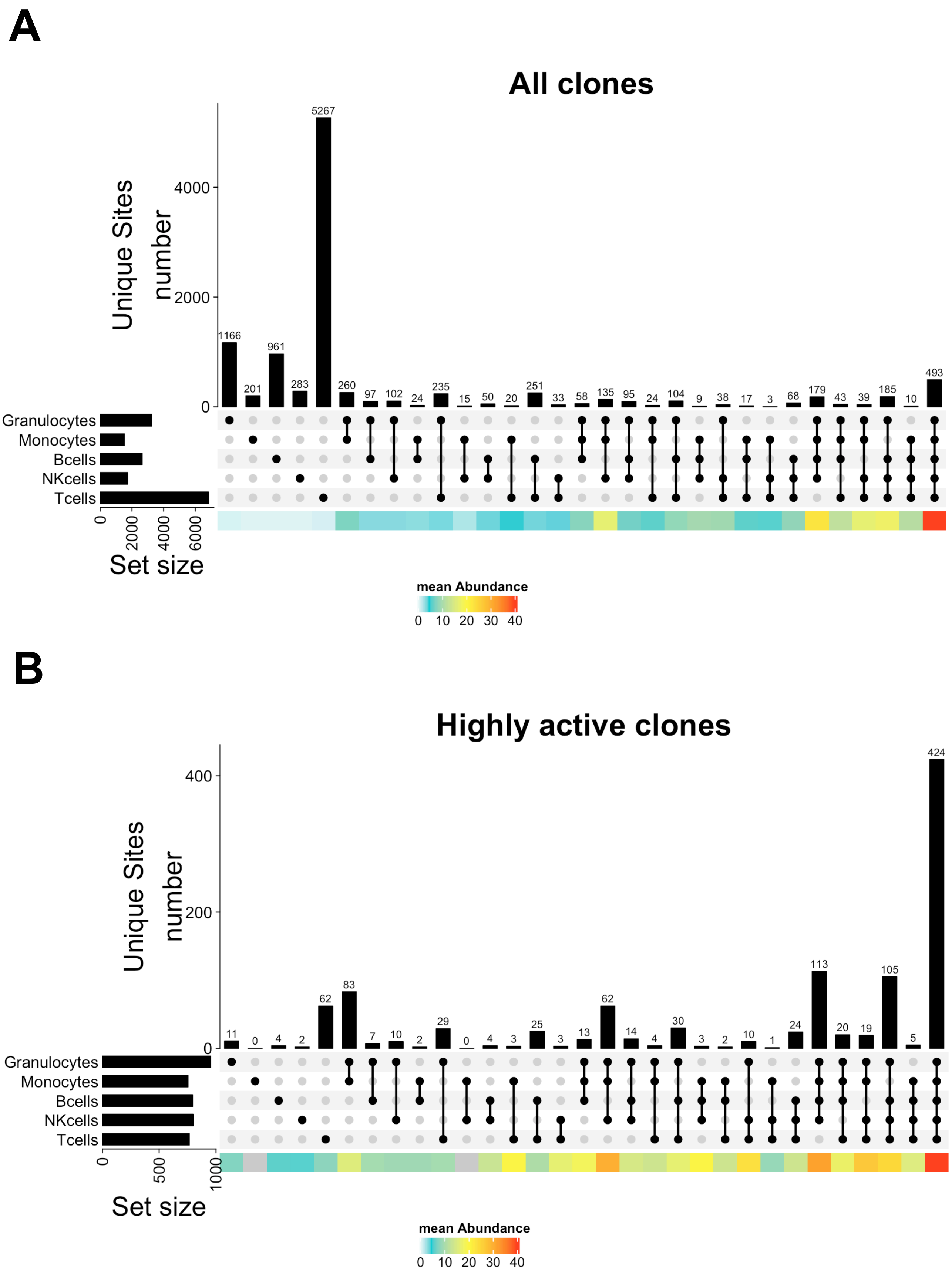


Figure 5

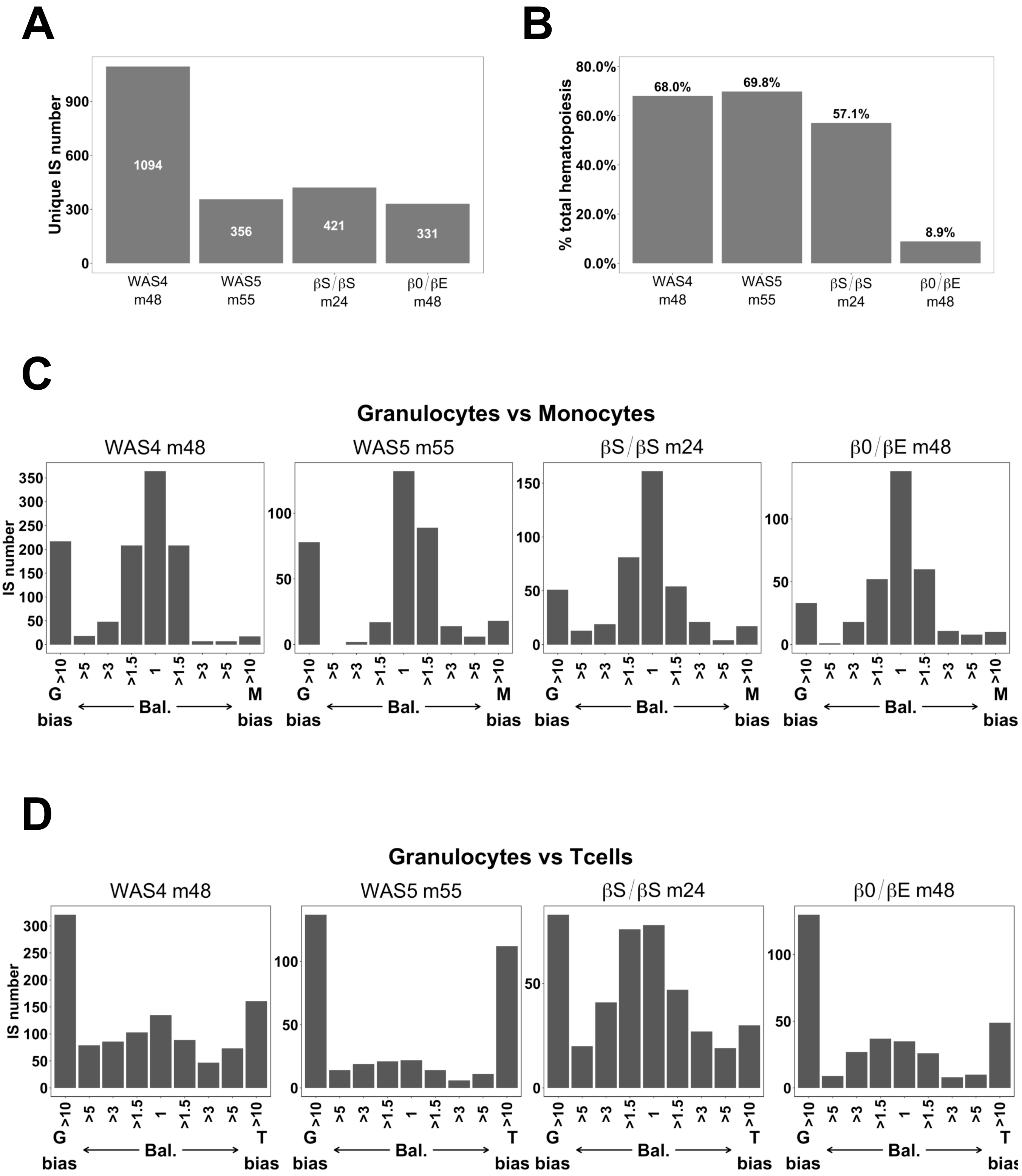
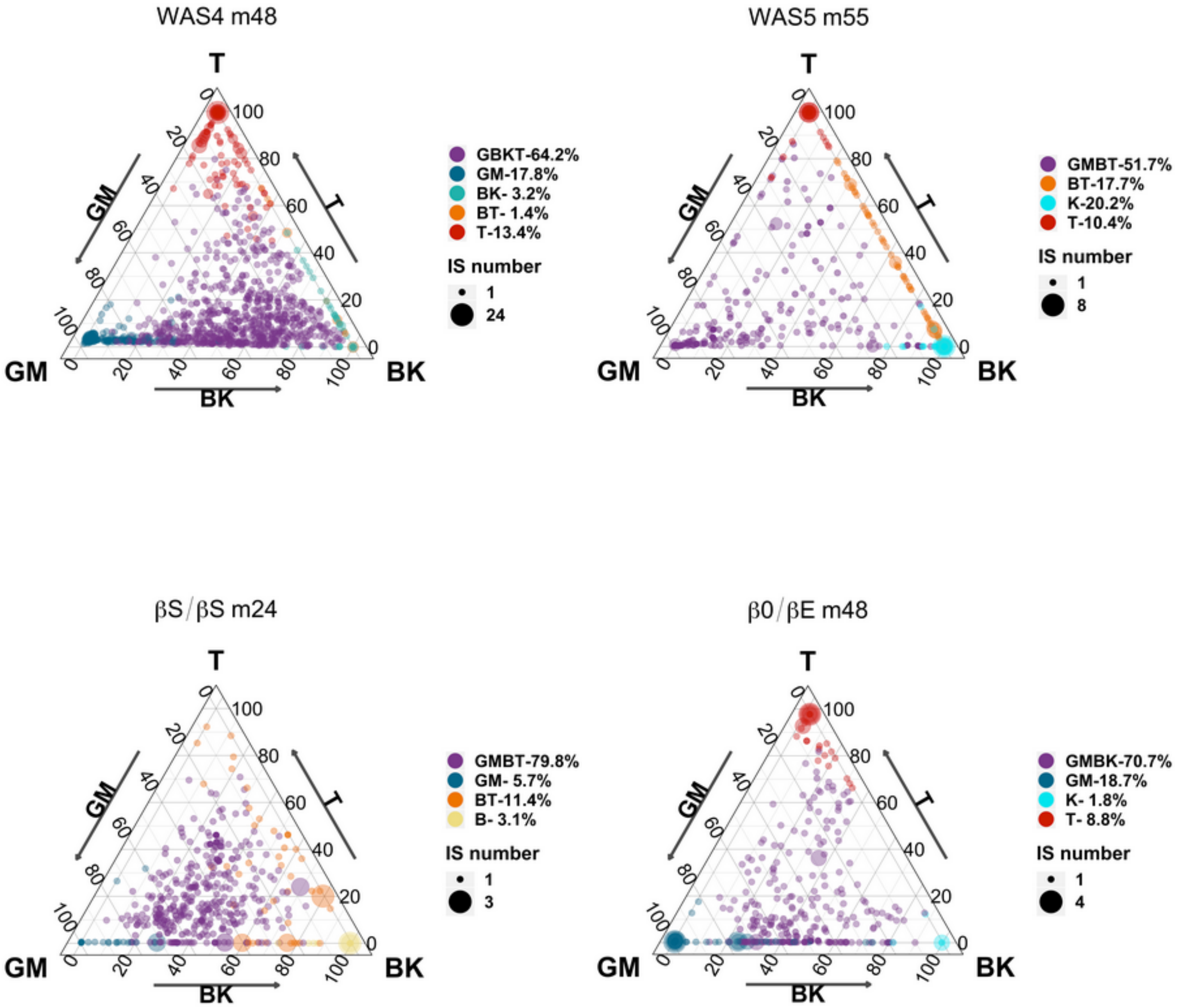

A
Unique IS number
0
300
600
900
1094
356
421
331
WAS4 m48
WAS5 m55
βS/βS m24
β0/βE m48
B
% total hematopoiesis
0.0%
20.0%
40.0%
60.0%
80.0%
68.0%
69.8%
57.1%
8.9%
WAS4 m48
WAS5 m55
βS/βS m24
β0/βE m48
C
Granulocytes vs Monocytes
WAS4 m48
WAS5 m55
βS/βS m24
β0/βE m48
IS number
0
50
100
150
200
250
300
350
>10
>5
>3
>1.5
1
>1.5
>3
>5
>10
G bias
Bal.
M bias
D
Granulocytes vs Tcells
WAS4 m48
WAS5 m55
βS/βS m24
β0/βE m48
IS number
0
50
100
150
200
250
300
>10
>5
>3
>1.5
1
>1.5
>3
>5
>10
G bias
Bal.
T bias

Figure 6



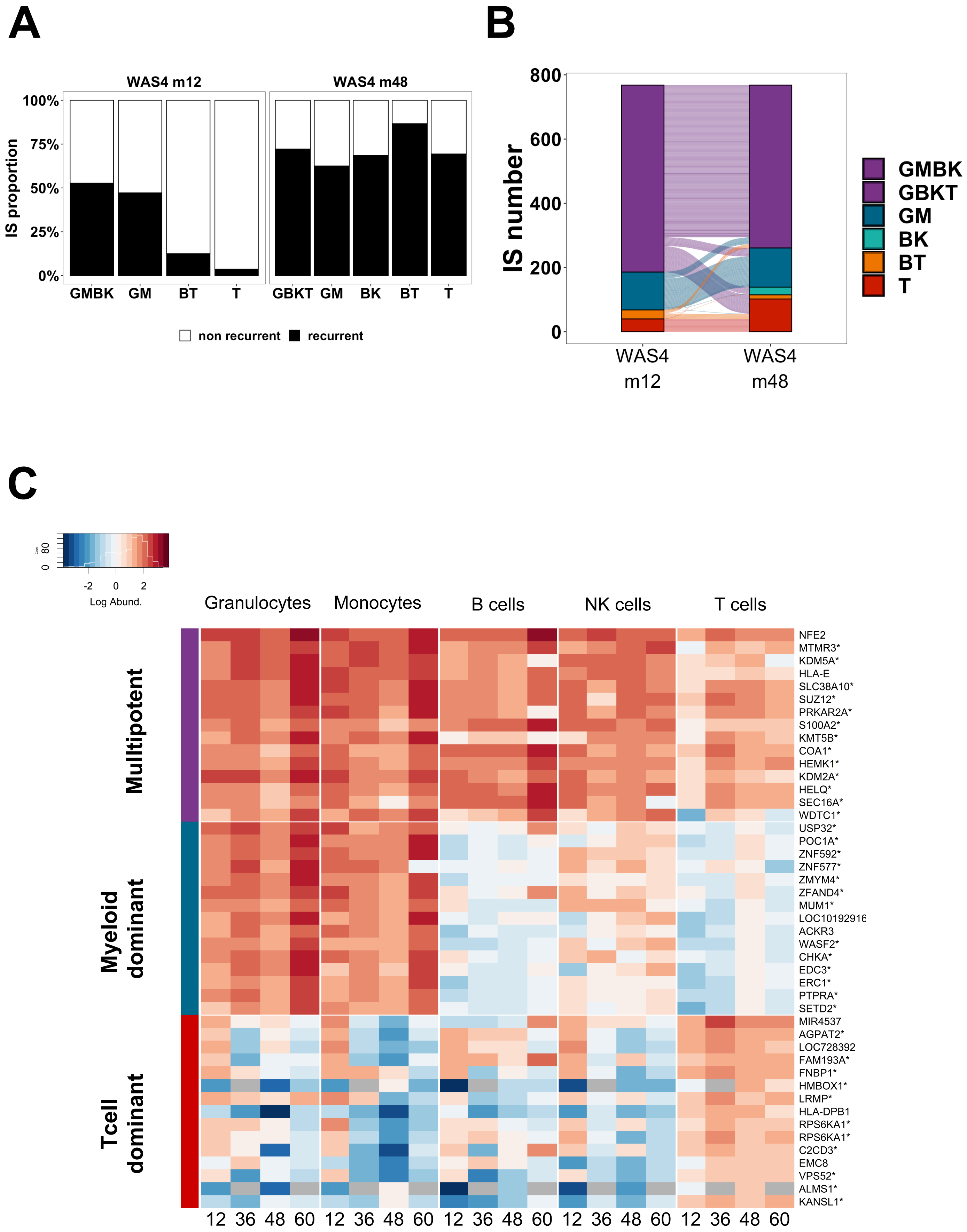
WAS4 m48

WAS5 m55

βS/βS m24

β0/βE m48

Figure 7



A
WAS4 m12
WAS4 m48
100%
75%
50%
25%
0%
IS proportion
GMBK
GM
BT
T
GBKT
GM
BK
BT
T
non recurrent
recurrent
B
800
600
400
200
0
IS number
WAS4 m12
WAS4 m48
GMBK
GBKT
GM
BK
BT
T
C
Count
0
80
-2
0
2
Log Abund.
Granulocytes
Monocytes
B cells
NK cells
T cells
Multipotent
Myeloid dominant
Tcell dominant
NFE2
MTMR3*
KDM5A*
HLA-E
SLC38A10*
SUZ12*
PRKAR2A*
S100A2*
KMT5B*
COA1*
HEMK1*
KDM2A*
HELQ*
SEC16A*
WDTC1*
USP32*
POC1A*
ZNF592*
ZNF577*
ZMYM4*
ZFAND4*
MUM1*
LOC10192916
ACKR3
WASF2*
CHKA*
EDC3*
ERC1*
PTPRA*
SETD2*
MIR4537
AGPAT2*
LOC728392
FAM193A*
FNBP1*
HMBOX1*
LRMP*
HLA-DPB1
RPS6KA1*
RPS6KA1*
C2CD3*
EMC8
VPS52*
ALMS1*
KANSL1*
12 36 48 60 12 36 48 60 12 36 48 60 12 36 48 60 12 36 48 60