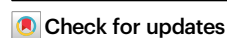


Modeling integration site data for safety assessment with MELISSA

Received: 2 October 2024

Accepted: 1 August 2025

Published online: 23 August 2025



Tsai-Yu Lin ^{1,6}, Giacomo Ceoldo ^{2,6}, Kimberley House¹, Matthew Welty¹,
Thao Thi Dang¹, Denise Klatt², Christian Brendel ^{2,3}, Michael P. Murphy⁴,
Kenneth Cornetta ¹ & Danilo Pellin ^{2,3,5} 

Gene and cell therapies pose safety concerns due to potential insertional mutagenesis by viral vectors. We introduce MELISSA, a regression-based statistical framework for analyzing Integration Site (IS) data to assess insertional mutagenesis risk, by estimating and comparing gene-specific integration rates and their impact on clone fitness. We characterized the IS profile of a lentiviral vector on Mesenchymal Stem Cells (MSCs) and compared it with that of Hematopoietic Stem and Progenitor Cells (HSPCs). We applied MELISSA to published IS data from patients enrolled in gene therapy clinical trials, successfully identifying both known and novel genes that drive changes in clone growth through vector integration. MELISSA offers a quantitative tool to bridge the gap between IS data and safety and efficacy evaluation, facilitating the generation of comprehensive data packages supporting Investigational New Drug (IND) and Biologics License (BLA) applications and the development of safe and effective gene and cell therapies.

Gene therapy using Hematopoietic Stem and Progenitor Cells (HSPC) modified with viral vectors has emerged as a promising approach for treating rare monogenetic diseases^{1,2}. This strategy involves introducing functional copies of therapeutic genes into patients' HSPCs, enabling them to produce the missing or malfunctioning protein and restore normal cellular function. Due to their efficient gene delivery capabilities, viral vectors have also found extensive application in novel cell therapies such as Chimeric Antigen Receptor (CAR)-T immunotherapies, where Lenti/Retroviral Vectors are used to deliver CAR genes in a patient's T cell.

The capability for viral vectors to integrate their cargo DNA into unpredictable locations within the host genome carries a certain risk of Insertional Mutagenesis (IM), defined as the disruption or dysregulation of a gene caused by a vector insertion. IM can lead to enhanced cell growth due to the activation of oncogenes or the disruption of tumor suppressor genes by the integrated DNA^{3–6}. Clonal dominance underscores the risk of a particular clone or a small subset of clones of transduced cells gaining a growth

advantage, potentially skewing the clonal composition of the repopulating cells towards monoclonal or oligoclonal configurations that pose additional long-term safety and efficacy concerns, including oncogenic transformation^{7–9}.

Regulatory agencies like the US FDA require rigorous preclinical safety and efficacy evaluations and 15 years of IM monitoring for patients treated with genetically modified HSPCs¹⁰. Recently, the US FDA extended monitoring to individuals receiving BCMA- or CD19-directed autologous CAR T cell immunotherapies. A comprehensive risk-benefit analysis is crucial for evaluating the potential for IM, yet the scientific community still lacks standardized indices, metrics, and methods to clearly distinguish between safe and unsafe integration profiles.

The association of the integrome, namely the distribution of IS across the host genome, with genomic annotation data has elucidated the role of chromatin conformation, transcriptional activity, and the cellular genome 3D nuclear organization on IS frequency^{11–15}. These insights, combined with the analysis of oncogenic transformation

¹National Gene Vector Biorepository, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA. ²Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ³Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ⁴Department of Surgery, Indiana University School of Medicine, Indianapolis, IN, USA. ⁵Harvard Data Science Initiative, Cambridge, MA, USA. ⁶These authors contributed equally: Tsai-Yu Lin, Giacomo Ceoldo. ✉ e-mail: danilo.pellin@dfci.harvard.edu

mechanisms observed in clinical trials, enabled the generation of novel and safer viral vectors for HSPC gene therapies.

Using IS analysis to monitor the clonal composition over time, scientists can gain insights into the diversity and dynamics of the engrafting cell population, assess the long-term efficacy, and the risk of clonal dominance^{7,16–18}. From a safety perspective, IS analyses serve as the primary screening tool for identifying potential risks associated with uncontrolled clonal expansion by detecting abundant clones based on their relative contribution and mapping IS location within the host cell genome. Conventionally, the identification of abundant clones has been based on predefined percentage thresholds calculated on an individual sample basis. Leveraging IS information from multiple datasets to estimate the integrome and clonal contributions can improve the characterization of gene targeting preferences and associated risks and enhance the detection of clonal expansion dynamics.

This paper introduces a set of statistical tools, MELISSA (Modeling IS for Safety Analysis), to translate IS data into actionable safety assessment and evaluation insights. MELISSA provides statistical models for measuring and comparing gene targeting rates and their effects on clone growth within a gene-based approach. MELISSA modeling consists of a regression approach that analyzes and combines data from complex experimental designs, including datasets with multiple patients, donors, replicates, and additional covariates of interest. MELISSA facilitates the quantitative comparisons of different conditions and includes rigorous statistical testing, visualization, and annotation to aid the biological interpretation of results.

From an experimental standpoint, we constructed a unique dataset to compare the integration and safety profiles of Mesenchymal Stem Cells (MSCs) isolated from two different sources and mobilized CD34⁺ HSPCs. This enabled us to evaluate the effectiveness of our proposed framework as a tool for analyzing preclinical experiments and guiding the development of novel gene therapy strategies.

Additionally, we analyzed published IS data from patients enrolled in gene therapy clinical trials for Beta-thalassemia (β -thal), Sickle Cell Disease (SCD), Wiskott-Aldrich Syndrome (WAS)¹⁸, and for X-linked Severe Combined Immunodeficiency (SCID-X1)¹⁹. By leveraging longitudinal modeling of clone sizes, our analysis identified both known and novel genes with the potential to influence clonal fitness when targeted by an IS. These findings were consistent across a range of clinical scenarios, from cases with high relative abundance of dominant clones (e.g., patient 3 in SCID-X1, 57%) to those without clearly concerning expansions (e.g., β -thal, 1%; SCD, 2.4%; WAS, 9.5%). This highlights MELISSA's sensitivity and ability to detect early signs of

clonal expansion, demonstrating its effectiveness in identifying risk factors before they become prominent.

MELISSA is a quantitative tool for supporting researchers, clinicians, and regulatory agencies in advancing the development of safe and effective gene and cell therapies. Its goal is to bridge the gap between the IS datasets generated from preclinical experiments to support IND applications and clinical samples for safety and efficacy evaluation.

Results

MELISSA is a robust statistical framework available as an R package designed to analyze and interpret IS data for safety evaluation in viral vector-based gene and cell therapies. We evaluated MELISSA using a comprehensive Monte Carlo simulation study to assess performance metrics, specifically focusing on the Positive Predictive Value (PPV) and the detection rate. This allowed us to understand the proportion of true positive results among all positive test results, reflecting the reliability of a positive test outcome and the likelihood of detecting a real effect or difference when it exists.

As shown in the schematic in Fig. 1, it requires three primary inputs. The IS tables are provided in bed file format and contain clone size estimates (as read counts, UMIs, or shear site data)^{20–22}. The design matrix includes sample-specific covariates, such as conditions, replicates, cell type, time after therapy, and others. MELISSA can analyze the full set of genes using genome annotation files, providing a comprehensive overview of potential IM risks. Alternatively, it allows users to provide a list of genomic regions of interest for targeted analysis, such as the cytogenetic bands and overlapping genomic intervals of varying sizes used in this manuscript.

MELISSA provides a comprehensive suite of functionalities. It generates descriptive statistics summarizing IS distribution across various genomic features and calculates samples' clonality indexes commonly referenced in the literature. It tackles critical safety-related questions: (1) estimating gene-specific IS rates within a single group or comparing them across groups to identify potential differences across conditions; (2) estimating whether IS within a gene influences clone growth, again for single or multiple groups.

Furthermore, MELISSA facilitates downstream analysis through various methods. Researchers can generate informative graphical representations and utilize gene-scoring tables to identify potential associations between IS data and underlying biological mechanisms with tools such as Gene Ontology (GO) enrichment analysis, Gene Set Enrichment Analysis (GSEA) and KEGG and REACTOME pathway analysis^{23–25}. Additionally, the distribution of genomic annotations,

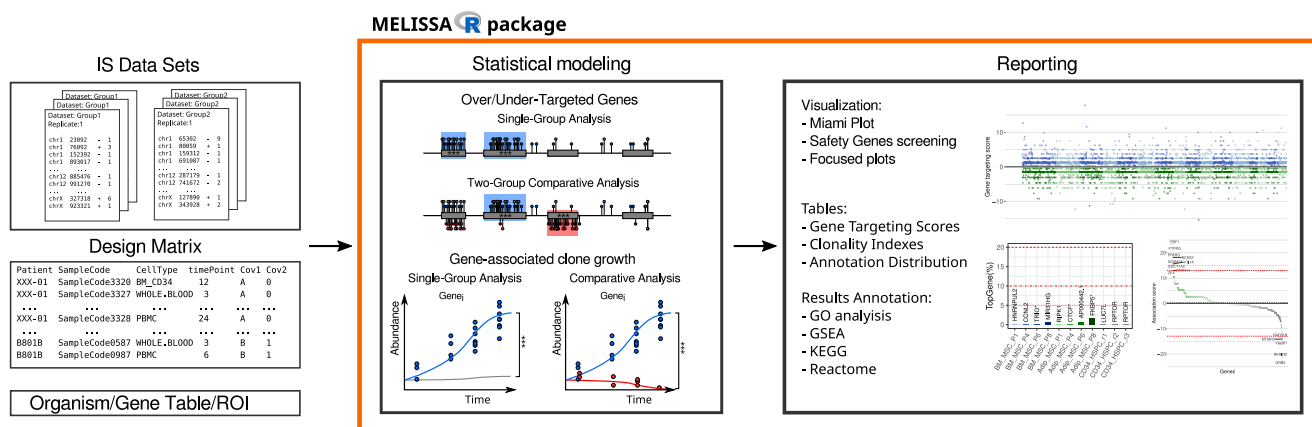


Fig. 1 | Schematic of MELISSA workflow. MELISSA receives as input 1. IS data in a bed file format, 2. a design matrix with all the relevant samples' annotation, 3. a reference genome or a bed file containing genomic region to be tested for IS enrichment. MELISSA combines input data and performs two main types of

analysis: gene targeting rates and gene/clone fitness alterations. MELISSA also has a dedicated function for calculating clonality indexes, graphical representation of the results, and downstream pathway enrichment analyses.

provided in BED file format (e.g., epigenetic profiles) within IS flanking regions, offers deeper insights into the mechanisms driving IS selection process.

Statistical modeling

MELISSA implements two complementary statistical models that address distinct biological questions related to gene therapy safety: (1) IS targeting rate analysis, which assesses whether specific genomic regions (e.g., genes) are preferentially targeted by IS events, and (2) clone fitness analysis, which evaluates whether IS within a region affects the expansion dynamics of clones over time. Each model can be applied in either a one-condition or a two-condition setting. The one-condition analysis supports datasets with multiple samples or replicates from a single group, while the two-condition analysis compares groups, such as different cell types or experimental conditions. In all cases, MELISSA returns two scoring metrics for each gene tested: a score equivalent to the Likelihood Ratio Test (LRT) statistic, which quantifies statistical significance, and a regression coefficient that captures the magnitude and direction of the effect.

The gene targeting rate analysis is based on logistic regression and estimates the likelihood that an IS falls within a specific gene compared to the rest of the genome. The model treats IS presence or absence at each genomic coordinate as a binary outcome and accounts for features such as gene length and overall IS frequency. In the one-condition setting, MELISSA compares a baseline model that assumes uniform targeting across all genes to an alternative where the gene of interest has its own targeting rate. The gene targeting score (LRT statistic) measures the significance of deviation from the genome-wide average IS rate, while the IS-enrichment (regression coefficient) quantifies the strength of the enrichment. In the two-condition setting, MELISSA tests whether targeting rates for a given gene differ between groups by comparing models with shared versus condition-specific targeting parameters. The differential gene targeting score is derived by assigning the sign of the regression parameter to the LRT statistic to reflect the direction of the change.

The clone fitness model evaluates whether IS in a given gene is associated with altered clone expansion over time. It assumes that IS may perturb gene function through mechanisms such as promoter interference or enhancer disruption^{4,26}. Clone size trajectories are modeled using logistic regression for binomial count data. In the one-condition setting, the model tests whether clones with IS in a given gene grow at a rate different from the global baseline, comparing a model with a shared growth rate to one that allows for a gene-specific effect. The regression coefficient associated with time represents the growth rate increment and indicates how clone contribution changes per time unit beyond the genome-wide trend. In the two-condition setting, the model tests whether gene-associated growth rates differ across groups. As with targeting, both the LRT statistic and the signed growth coefficient are used to interpret the results.

Multiple testing correction is applied using the False Discovery Rate (FDR) method²⁷ or the Holm–Bonferroni method²⁸, as specified by the user.

Simulation study

To evaluate MELISSA's performance in detecting variations in IS targeting rates and gene-specific effects on clone fitness, we conducted a simulation study on chromosome 15 (hg38) in which we included five artificial genes (TestGenes) with a length of 5, 10, 20, 40, and 80 kb pairs not overlapping with the existing 1027 endogenous genes. The study assessed how the PPV and TestGenes' detection rate varied with sample size and effect strength (Fig. 2, Suppl. Table 1). For each setting, 1000 simulations were performed to assess the model's performance. For simplicity, IS locations were simulated by sampling genomic coordinates under a piecewise uniform distribution, without accounting for mapping biases associated with low-complexity or

repetitive genomic sequences. Further details of the simulation design are provided in the “Methods” section, and the simulated datasets are shown in Suppl. Fig. 1.

Two strategies were used to simulate IS datasets for gene targeting rate analysis. In the single-group analysis (Fig. 2A), $N = 100, 200, 400, 800$ IS elements were generated, corresponding to genome-wide IS datasets ranging from 3000 to 25,000 IS, in line with preclinical and clinical datasets. IS were simulated using a two-part model: a genome-wide baseline and elevated integration rates for over-targeted genes, with fold increases of $h = 1$ (no effect), 2, 4, 8, 16, 32, and 64. For differential targeting analysis (Fig. 2B), each gene was assigned a unique rate, consistent across datasets, with group 1 showing elevated integration probabilities compared to group 2.

To analyze clone fitness and detect variations in clone growth rates, we simulated time-series data of IS clone sizes across six time points ($t = 0$ to 5). We first generated a database of 10,000 potential IS coordinates using a constant baseline and a variable gene-specific targeting rate. Clone sizes were determined by sampling IS from this database, with replacement, at each time point. Sampling probabilities for IS in TestGenes linearly increase over time (Fig. 2C), mimicking selective clone expansion. In the single-group scenario, datasets with cumulative clone sizes of $M = 1000, 2000, 4000, 8000$ were generated. For differential analysis (Fig. 2D), 12 datasets were generated, representing two groups with identical sampling probabilities, except for TestGenes where group 1 had higher growth rates controlled by the effect size h .

We evaluated the impact of P value adjustment methods on the performance of MELISSA, with results shown in Suppl. Fig. 2. Using adjusted P values, MELISSA's PPV consistently outperformed results obtained using raw p -values, emphasizing the importance of correction methods in improving predictive accuracy. The Holm and FDR adjustment methods yielded comparable results. Computational times under different scenarios are detailed in Suppl. Table 2.

Our simulation study demonstrated that MELISSA's PPV and detection rates consistently improved with increasing sample sizes and effect strengths. In single-group analyses, detecting higher gene targeting rates in short genes proved challenging, while clone fitness analysis showed strong sensitivity, even for low effect sizes. For the differential setting, the results highlighted that detecting differences in gene targeting rates is more efficient than detecting variations in clone fitness. Importantly, the model produced no false positives, confirming its high specificity.

Using MELISSA to evaluate the clonal dynamics of transduced MSC over time

MSC is a heterogeneous cell population found in bone marrow, adipose tissue, placenta, umbilical cord and cord blood, dental pulp, and other tissues that show potential for cell therapy applications due to their regenerative potential. There are over 1500 clinical trials exploring their application across diverse diseases (www.clinicaltrials.gov), and lentiviral vector (LV) are being evaluated to increase the therapeutic potential of MSCs. While extensive research has been conducted on the integrone in other cell types, our understanding of LV integration and its potential risks in MSCs remains limited.

Low passage healthy donor Bone Marrow-derived (BM), Adipose-derived (Ad) MSCs, and CD34+ healthy donor mobilized HSPCs were transduced with an LV encoding EGFP. IS analysis was performed on all cell types early after transduction and at various passages of in-vitro culture for MSCs (P1, P4, P6, P8). Cells transduced at the initiation of the culture will give rise to a progeny of cells (IS clone) that share the exact IS coordinates (i.e., the LV IS). More details on the MSC transduction, IS retrieval protocol and bioinformatic analysis are available in the “Methods” section.

Accurate clone size estimation is crucial for IS analysis since it allows for understanding clones' relative contribution and

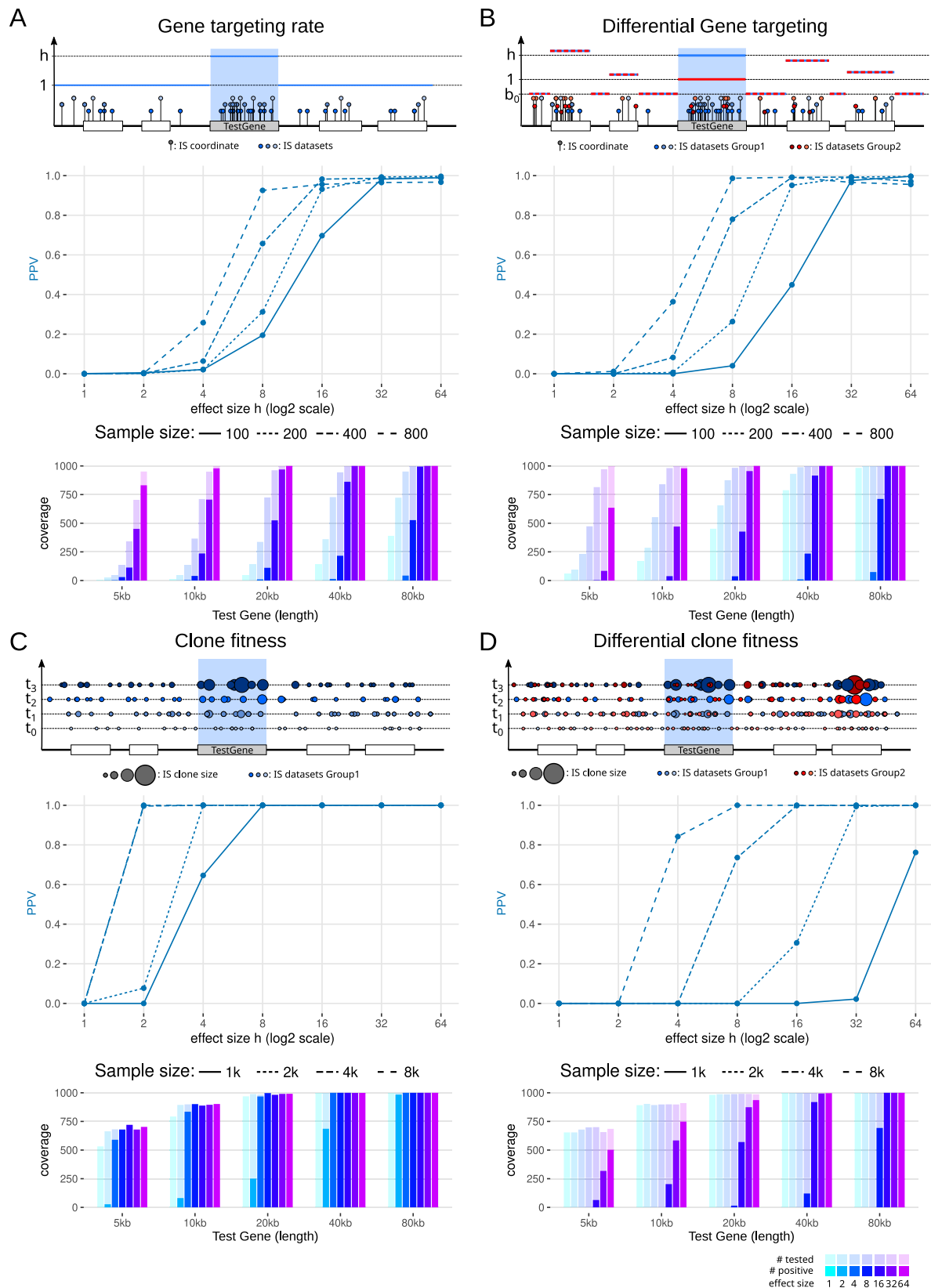


Fig. 2 | Simulation study results. Average Positive Predicted Value (PPV) across 1000 simulations for model (A) gene targeting rate; (B) differential gene targeting; (C) clone fitness; (D) differential clone fitness. Solid, dashed, dotted, and dash-dot lines correspond to different sample sizes. Seven increasing effect sizes have been tested, ranging from $h = 1$ (no effect) to 64. Detection rate bar graphs (bottom of each

panel) report the number of times each *TestGene* has been tested (semi-transparent bars) and the number of times in which the *TestGene* is detected as significantly over-targeted/gene growth associated, for the simulations with sample size $N = 400$ for (A, B), and $M = 4000$ for (C, D).

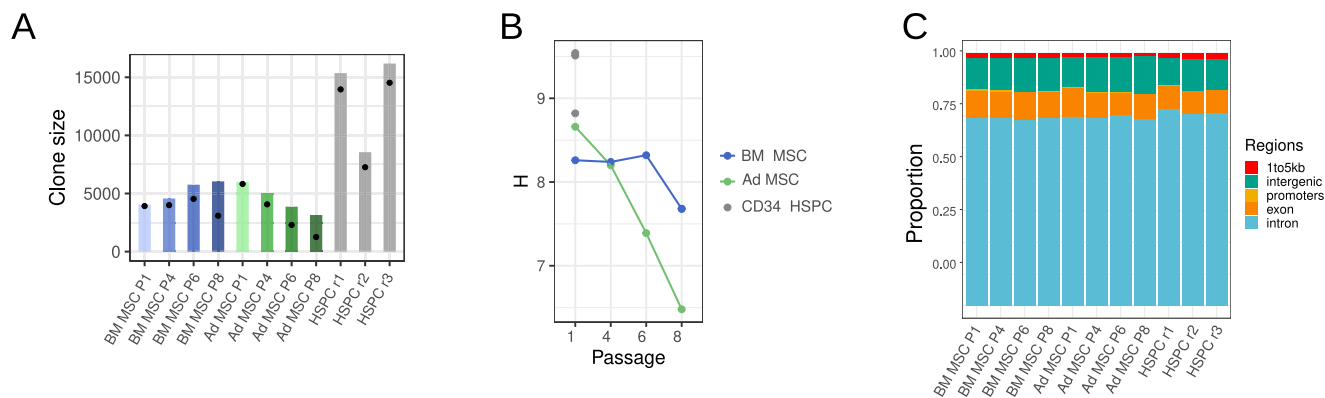


Fig. 3 | IS datasets descriptive statistics. A Integration site counts and cumulative clone size: Bar chart depicting the cumulative clone size for each dataset. The number of distinct IS retrieved in each sample is reported with dots. **B** Shannon entropy (H) index. Line plot representing the longitudinal dynamics of the diversity during in-vitro culture with subsequent passages of MSC. CD34 HSPC were

analyzed at a single timepoint. **C** Integration site annotation distribution: stacked bar chart displaying the distribution of integration sites according to genomic annotation, delineating the genomic regions where integration events are most prevalent.

characterizing the composition of the population. Here we estimated clone sizes using the sonicLength method²⁹, albeit alternative quantification strategies (read counts, UMIs) are compatible with the MELISSA framework. As shown in Fig. 3A and Suppl. Table 3, the cumulative clone size in BM MSCs shows a consistent positive trend throughout the in-vitro culture period. In contrast, the number of distinct clones steadily increases in P1 through P6, with a drop observed at P8. Ad MSCs displayed a declining IS count and a similar trend in total clone size. These observations are further supported by the Shannon diversity index analysis (Fig. 3B), indicating stable diversity in BM MSCs and a progressive loss of diversity in Ad MSC.

LV integration preferences have been primarily reported in hematopoietic cells. The distribution observed in our HSPC datasets confirmed the known tendency of LV vectors to integrate into gene bodies. Interestingly, we found that this pattern also showed remarkable consistency in MSCs^{30–32}. In the gene-based analysis reported in the following sections, we extended the gene bodies (region from the transcription start site to the end of the transcript) by including the 1-kb gene promoter region upstream of a gene's transcriptional start site, thereby encompassing -85.2% of the IS database (74.9% intronic, 10% exonic, and 0.3% resided in gene promoter regions).

MELISSA identifies both shared and cell type-specific over-targeted genes in MSCs and HSPC

We analyzed IS datasets using MELISSA to identify the most targeted genes in BM MSC, Ad MSC and HSPC to investigate the presence of differentially targeted genes.

Figure 4A shows the gene scores from the HSPC control dataset analysis. Among 1959 significant genes, known Recurrent Integration Genes (RIGs) like *PACSI*, *KDM2A*, and *GRB2* have higher scores³³, confirming our approach effectively captures established integration patterns.

In BM MSC, we found 5313 genes had at least 1 IS within their extended gene body, of which 1134 were significantly over-targeted. In Ad MSC, 718 genes out of 5401 were significantly over-targeted.

As shown in Fig. 4B, 157 genes were over-targeted in all cell types, 109 genes appear to be MSC specific, and the overlap between BM MSC and HSPC (295 genes) is higher than the shared genes between Ad MSC and HSPC (134 genes).

We defined a set of High-Risk genes that includes cancer-associated genes and genes related to IM events and verified how their targeting rates varied among cell types³⁴. The heatmap in Fig. 4C highlights how the IS patterns are very similar, especially on the highly targeted High-Risk genes. The one exception was *SPPL2B* which was

specific for BM MSC. Interestingly, genes associated with documented clonal expansion in humans, such as *LMO2*, have no detected IS, and genes such as *MECOM* and *HMG2* had scores higher in HPSC than MSCs.

We provide a more detailed representation of the IS distribution around two genes in Fig. 4D. *NPLOC4* appeared among the top ten targeted genes across all datasets, while *PLEC* is highly enriched in IS in both BM MSC and Ad MSC datasets but there was no IS noted in the HSPC dataset. The complete results of the analysis are available in the Source Data file, the genome-wide gene targeting score visualization is provided in Suppl. Fig. 4 and the pathway analysis for each dataset in Suppl. Fig. 5–7.

Gene targeting rates have been linked to transcriptional activity, primarily through the interaction of LEDGF/p75 with H3K36me3-marked chromatin^{35,36}. To assess whether this association can be explained by gene expression levels, we analyzed published bulk RNA-seq data from healthy donor-derived HSPC³⁷ (8 replicates) and BM MSCs^{38,39} (8 replicates). We then performed a regression analysis to assess the correlation between gene targeting scores and expression levels. As shown in Fig. 4E, a positive correlation was observed in both HSPCs ($\beta = 0.9$, $p = 2e-11$, $R^2 = 1.33e-2$) and BM MSCs ($\beta = 0.27$, $p = 0.15$, $R^2 = 1.64e-3$), although statistically significant only for HSPC. However, the explained variance was low, indicating that gene expression levels are a poor quantitative predictor of targeting rates. The gene expression analysis is described in the “Methods” section.

Differential gene targeting analysis highlights differences between MSC and HSPC and minor differences within MSC integromes

In addition to the gene sets intersection analysis, MELISSA allows us to compare the safety profiles in two cell types with a dedicated model, avoiding loss of relevant information and focusing exclusively on detecting the differences between, for example, a novel and a reference integrome.

First, we compared BM MSC to HSPC (Fig. 5A) and identified 119 genes over-targeted in BM MSC and 47 in HSPC. The most over-targeted gene in BM MSC was *PLEC* (Fig. 4D), while *DACH1* was the most over-targeted gene in HSPC and showed no IS in BM MSC.

We investigated whether differences in gene expression could account for the variation in targeting rates between HSPCs and BM MSCs. To this end, we computed gene expression fold changes between the two cell types and assessed their association with differential targeting scores. As shown in Fig. 5B, this analysis revealed a weak but statistically significant positive correlation ($\beta = 0.13$, $p < 2e-$

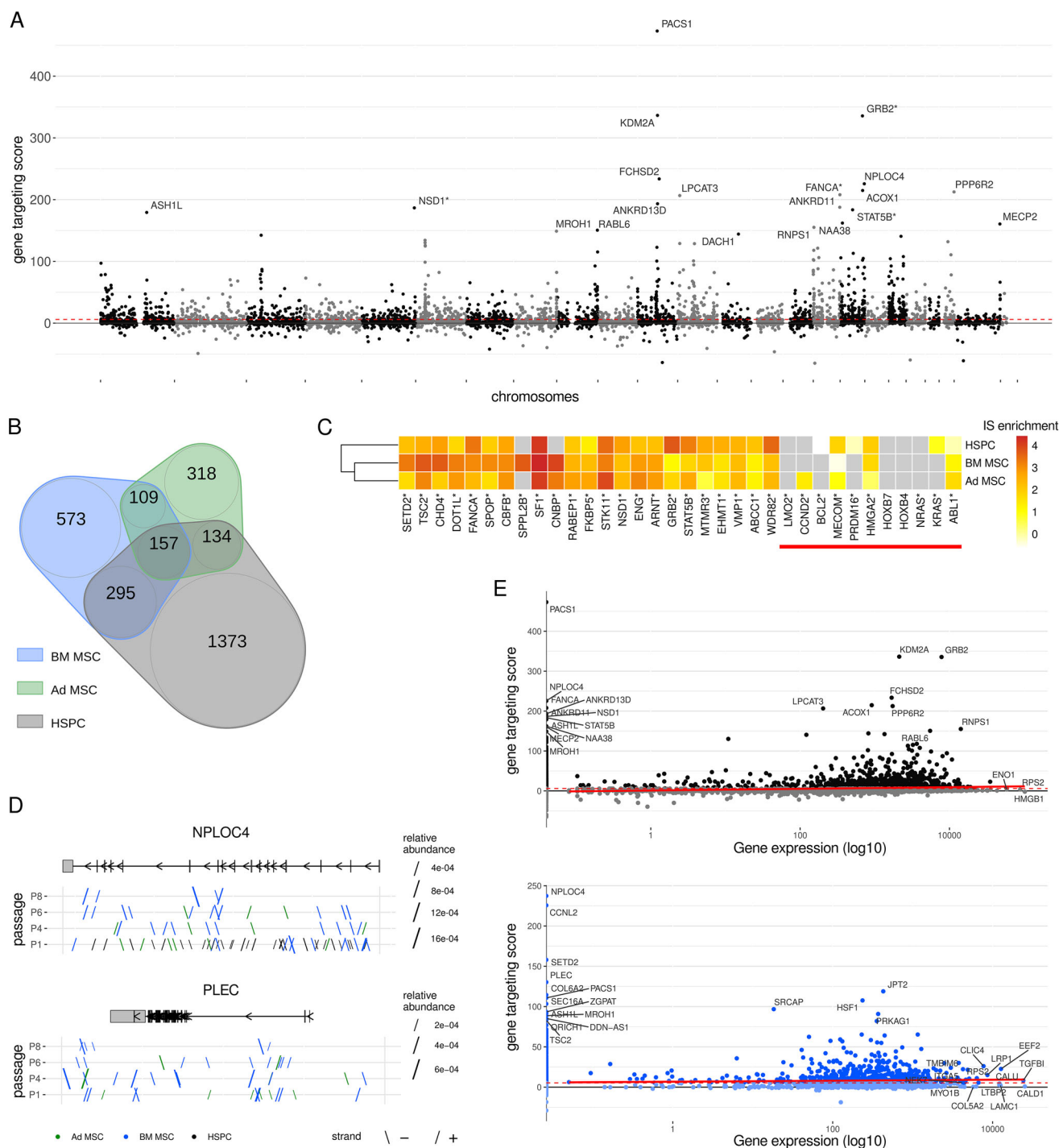
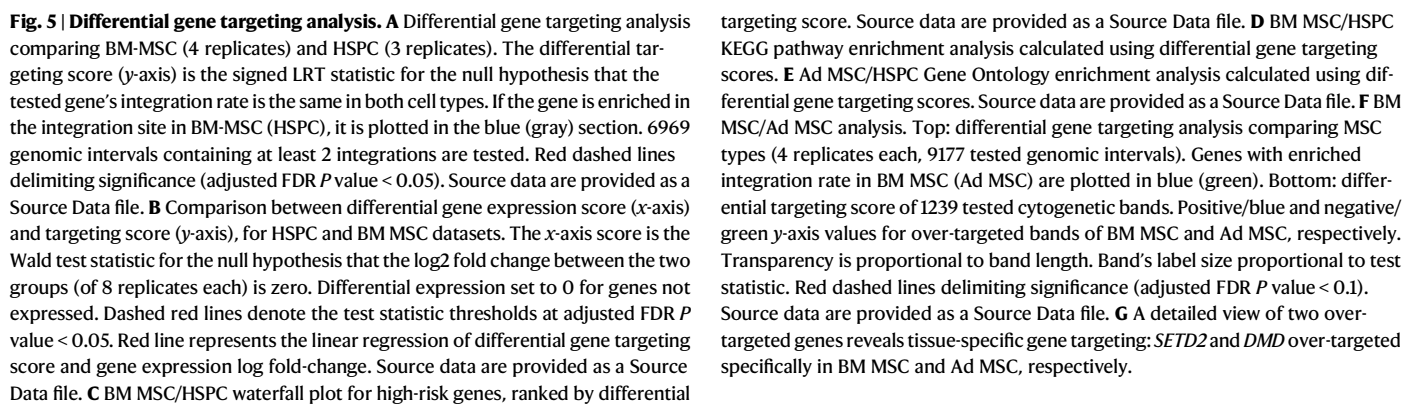


Fig. 4 | Results for the single group gene targeting rates analysis. A Gene targeting score for the HSPC dataset. Dots correspond to genes and are ordered according to chromosome location, with genes in even-numbered chromosomes having a lighter color. Y-axis value is the signed likelihood ratio test statistics (score) for the null hypothesis that the tested gene's integration rate is equal to genome-wide baseline. Genes marked with a * are included in the high-risk gene list. The dashed red line represents the threshold for statistically significant enrichment (586 genes, with adjusted FDR P value < 0.05). 3 replicates, 8204 tested genomic intervals (containing at least one integration), corresponding to 1959 genes. Source data are provided as a Source Data file. **B** Overlap between significantly over-targeted (adjusted FDR P value < 0.05) gene sets. 4 replicates for both MSC types, 3019 and 6288 tested genomic intervals for BM MSC and Ad MSC respectively. Source data are provided as a Source Data file. **C** Most targeted genes in MSCs and HSPC. IS enrichment scores for the top 10 most frequently targeted genes in each

dataset. Genes associated with documented clonal expansion in gene therapy clinical trials are included in the heatmap, marked by a red line under the gene name. Colors in the heatmap represent the IS enrichment, quantified as the log odds (regression coefficient estimate) associated with the gene effect. Source data are provided as a Source Data file. **D** Detailed view of two over-targeted genes reveals cell type-specific and shared integration patterns: *NPLOC4* is over-targeted among all cell types, *PLEC* is over-targeted exclusively in both BM MSC and Ad MSC. **E** Comparison between the expression level (x axis, log10 scale) and the targeting rates (y axis) defined in Fig. 4A. The expression level is the average across 8 samples of raw counts, normalized for differences in sequencing depth across samples. A default zero expression level is assigned to genes not expressed. Red line represents the linear regression of gene targeting score and gene log10 expression score. Source data are provided as a Source Data file.



high-risk genes being over-targeted in BM MSC (top genes: *SPPL2B*, *PPARD*, *STAG2*, *SETD2*) and 13 in HSPC (top genes: *GRB2*, *PTPRD*, *IKZF2*), the risk profiles of the two integromes are well-balanced overall, with an average risk score equal to 0.4 (0 = equal risk). Pathway analysis calculated using gene targeting scores is reported in Suppl. Fig. 8.

When evaluating genes with high IS rates in the Ad MSC and HSPC analysis (Suppl. Figs. 9–10), 9 out of 51 genes in the Ad MSC group and 5 out of 21 in the HSPC group were present in the High-Risk gene list.

The KEGG and the GO analysis in Fig. 5D, E were calculated using the gene's IS rates scores from the comparisons between MSCs and HSPC. The analysis shows that IS rates can discriminate cell type-specific pathways, such as in immune responses and signaling pathways in HSPC and extracellular matrix regulation in Ad MSC.

When comparing BM MSC to Ad MSC (Fig. 5E, and Suppl. Fig. 11), we found only 2 differentially targeted genes. *SETD2* (Fig. 5G), a critical chromatin regulator with potential tumor suppressor activity⁴⁰ is over-targeted in BM MSC. The *DMD* gene was over-targeted in Ad MSC. Despite the *DMD* gene not being part of the High-Risk gene sets, it is noteworthy that a growing body of evidence suggests a potential role of the *DMD* gene in tumor development and progression^{41,42}.

Given the limited number of significant genes detected, we applied a gene-agnostic approach using cytogenetic bands (1283 regions) to also capture potential intergenic effects that may be missed when focusing solely on gene boundaries, allowing for a genome-wide IS distribution comparison between MSC cell types. As shown in Fig. 5F (bottom), this strategy identified both broad chromosome-level and more localized Mb-scale regions with significant differential targeting scores, in some cases exceeding those observed for individual genes.

The most enriched cytogenetic bands in BM MSCs was 17q25.3 (6 Mbp; score = 33.0, $p = 3 \times 10^{-6}$), a gene-dense region containing *RIG*, *NPLOC4*, and 13 High-Risk genes. In comparison, sub-band 3p21.31, which includes *SETD2*, had a lower score (score = 10.0, adjusted $p = 0.03$). In Ad MSCs, the entire chromosome X, which includes *DMD*, was significantly over-targeted (score = -40.4, $p = 2 \times 10^{-7}$), as was sub-band 6q25 (score = -21.6, $p = 4 \times 10^{-4}$).

All targeting score tables corresponding to the comparisons described in this section are available in the Source Data file.

Different gene sets influence MSC clone growth dynamics in-vitro

While analyzing gene-targeting rates reveals potential safety risks, it does not directly assess whether IS within specific genes alter cell behavior. To address this crucial safety concern, we can quantify how targeting a particular gene impacts the fitness of individual cell clones. MELISSA's model tracks the growth of clones with IS within specific genes, comparing them to average growth rates and potentially across different conditions. This approach helps discern whether clone selection is driven by stochastic factors or a direct consequence of the IS.

Using a clone growth rate (regression parameter) calculated from the single group statistical analysis, we measured the impact for genes targeted in BM MSC and Ad MSC. Interestingly, the gene scores shown in Fig. 6A and available in the Source Data file reveal distinct sets of top genes associated with growth for each cell type. In contrast, genes with documented potential for clonal expansion have no significant impact. In the Ad MSC dataset, we identified more genes significantly affecting clone growth rates (Ad MSC: 48, BM MSC: 32) and a higher number of High-Risk genes (Ad MSC: 17, BM MSC: 7).

Figure 6B highlights a key difference: IS in 19 genes show a significant positive impact on Ad MSC clone growth, with *FKBP5* showing the most substantial effect. Conversely, only one gene, *RORA*, positively influenced BM MSC growth. A genome-wide analysis using a sliding window approach produced consistent results, identifying only two regions with significantly higher growth rates in BM MSCs. These included a region intersecting the *RORA* gene (chr15:60760219–61542469; score = 12.66, $p = 0.03$) and chr21:26869616–27651866 (score = 17.43, $p = 0.003$).

Figure 6C further details these findings, visualizing the individual and cumulative contributions of clones harboring IS events within

FKBP5, reaching a 3% total contribution with 9 IS, while the *RORA* effect is less pronounced and limited to 0.3%.

Integration preferences in MSC and HSPC show different correlations with epigenetic marks

MELISSA offers several tools to help in the biological interpretation of the findings derived from the IS rates and clone fitness analysis

We investigated the relationship between the distribution of IS and chromatin status in a cell-type-specific manner by taking advantage of the Roadmap Epigenomics project, which published Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq) data on the histone modifications H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3. We utilized data reports for BM MSC, Ad MSC and mobilized CD34 + HSPC (Suppl. Table 4). Our analysis assessed the enrichment of these epigenetic marks within a 100 kb window centered on the IS, similar to the analysis proposed in ref. 33. A heatmap representation of distance histograms is shown in Fig. 6D.

The results for the HSPC dataset are consistent with the profiles previously reported in the literature for all the six marks analyzed. For the open chromatin markers H3K4me3 and H3K36me3 and the repressive mark H3K27me3, we found a consistent pattern of distance distribution in both MSC cell types and HSPC, with the typical ± 2 kb sharp short-distances depletion for H3K4me3 and the progressive and smooth short-distances enrichment and depletion for H3K36me3 and H3K27me3 respectively. However, for the marker of active enhancers H3K4me1 and the repressive mark H3K9me3, MSC cell types and HSPC IS showed different patterns. In HSPC, short distances (± 2 kb) were underrepresented for both histone modifications, yet this pattern was not found in BM MSC and Ad MSC.

Gene targeting and clone fitness analysis in gene therapy clinical trials

In the manuscript by ref. 18, IS tables were generated from sorted myeloid (monocytes and granulocytes) and lymphoid (B, T, and NK cells) subpopulations isolated from peripheral blood samples collected at a minimum of two time points post-therapy. The study included patients enrolled in gene therapy trials for β -thal (1 patient), SCD (1 patient), and WAS (3 patients). Suppl. Fig. 12A summarizes the dataset, showing the number of IS and cumulative clone size for each cell type, time point, and patient.

The datasets display considerable heterogeneity in size across trials and patients. Notably, the β -thal patient exhibited a substantially higher number of IS compared to the SCD patient, while within the WAS cohort, patient WAS2 had a higher IS count than the other WAS patients. MELISSA models are designed to handle such imbalances and effectively account for longitudinal changes due to progressive clonal selection over time. Figure 7A displays the IS-enrichment score for significant genes. Miami plots for individual clinical trials are provided in Suppl. Fig. 13.

Compared to the 1944 genes significantly over-targeted in HSPC, fewer over-targeted genes were identified in the clinical trial datasets. No major differences were observed between myeloid (β -thal: 835, SCD: 687, WAS: 1215) and lymphoid lineages (β -thal: 875, SCD: 702, WAS: 1285). However, the clinical data exhibited a broader range of gene targeting scores than HSPC. Notably, there was a strong increase in highly targeted genes in specific genomic regions, such as chromosome 6, and a consistent decrease in targeting scores in regions like chromosomes 4 and 18 across lineages and trials.

To assess whether gene targeting rates differ significantly between clinical samples and control HSPC data, we employed MELISSA differential models. Figure 7B highlights the comparison between WAS, with data from four patients and the longest follow-up period (Suppl. Fig. 14 for β -thal and SCD), to determine whether IS within specific genes are more likely to pass the bottleneck of engraftment, characterizing the integrome of long-term repopulating

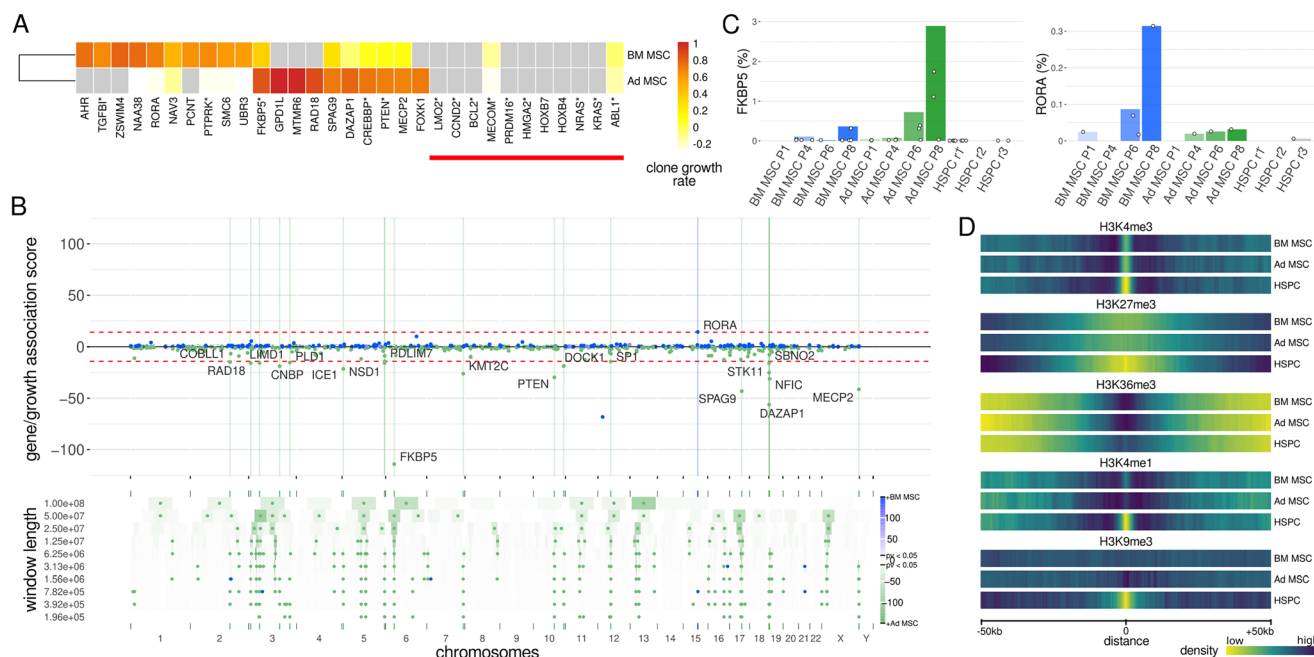


Fig. 6 | Integration site gene targeting effects on clone growth dynamics.

A Clone growth rate scores for the top 10 genes identified using the clone growth dynamics modeling in a single group setting (1 replicate, 4 time points). Genes associated with documented clonal expansion in gene therapy clinical trials are included in the heatmap, marked by a red line under the gene name. The color in the heatmap represents the estimated clone growth rates, measuring the average increase in clone size at each passage for IS mapping to a specific gene. Source data are provided as a Source Data file. **B** Enrichment score for differential clone growth dynamics in BM MSC and Ad MSC. The plot illustrates the growth association score, quantifying the differential impact of integration site targeting genes on the growth dynamics of clones within BM MSC and Ad MSC samples longitudinally over passages. Top plot: dots represent genes with high clone growth in BM MSC (blue) or Ad MSC (green). The score is the signed LRT statistic for null hypothesis of same growth rate in tested genomic interval (8637 tests). The dashed red line represents

the threshold for statistically significant differential clone dynamics (adjusted p value 0.05). Colored vertical lines highlight statistically significant genes. Bottom plot: sliding windows of different lengths (y-axis), color scale denotes the magnitude of the LRT statistic (blue BM MSC, green Ad MSC, 18002 tested windows). Colored dots inside the windows denote the statistically significant ones with adjusted FDR p value < 0.05 . Source data are provided as a Source Data file. **C** Detailed view of the relative contribution during the in-vitro culture of IS within *FKBP5* (Ad MSC) and *RORA* (BM MSC). The bar graph represents the IS's cumulative contribution (%) within the gene, while dots correspond to individual clones. **D** Integration site association with histone modifications in the different cell types. Comparison of the distance between IS datasets and the (top to bottom) H3K4me3, H3K27me3, H3K36me3, H3K4me1, and H3K9me3 histone modification profiles in the corresponding cell types. The histogram (bin size 300 bps) of distances is converted into a heatmap.

HSCs. In alignment with the single-group analyses, several broad genomic regions were significantly differentially targeted. For instance, regions such as chr6 p22.1-p21.31, chr1 p36.13-p36.11, and chr11 q12.2-q13.4, while enriched in targeted genes in HSPC, exhibited such an increase in their relative contribution in the clinical datasets that clones harboring IS in these regions have a significantly higher probability of long-term engraftment. Conversely, genes in chr13 and chr18, significantly over-targeted in HSPC, are likely lost upon in vivo selection.

We conducted a clone fitness analysis across all available clinical trial data, incorporating cell type as a covariate to identify evidence of IM. Given the limited number of patients in the clinical data, to minimize potential false-positive gene associations caused by stochastic clonal selection (without mechanistic interaction), we implemented an additional robust analysis step that filters out gene/clone growth associations that are not significant after excluding measurements from the most abundant clone in the tested gene. Further details on this procedure are provided in the “Methods” section.

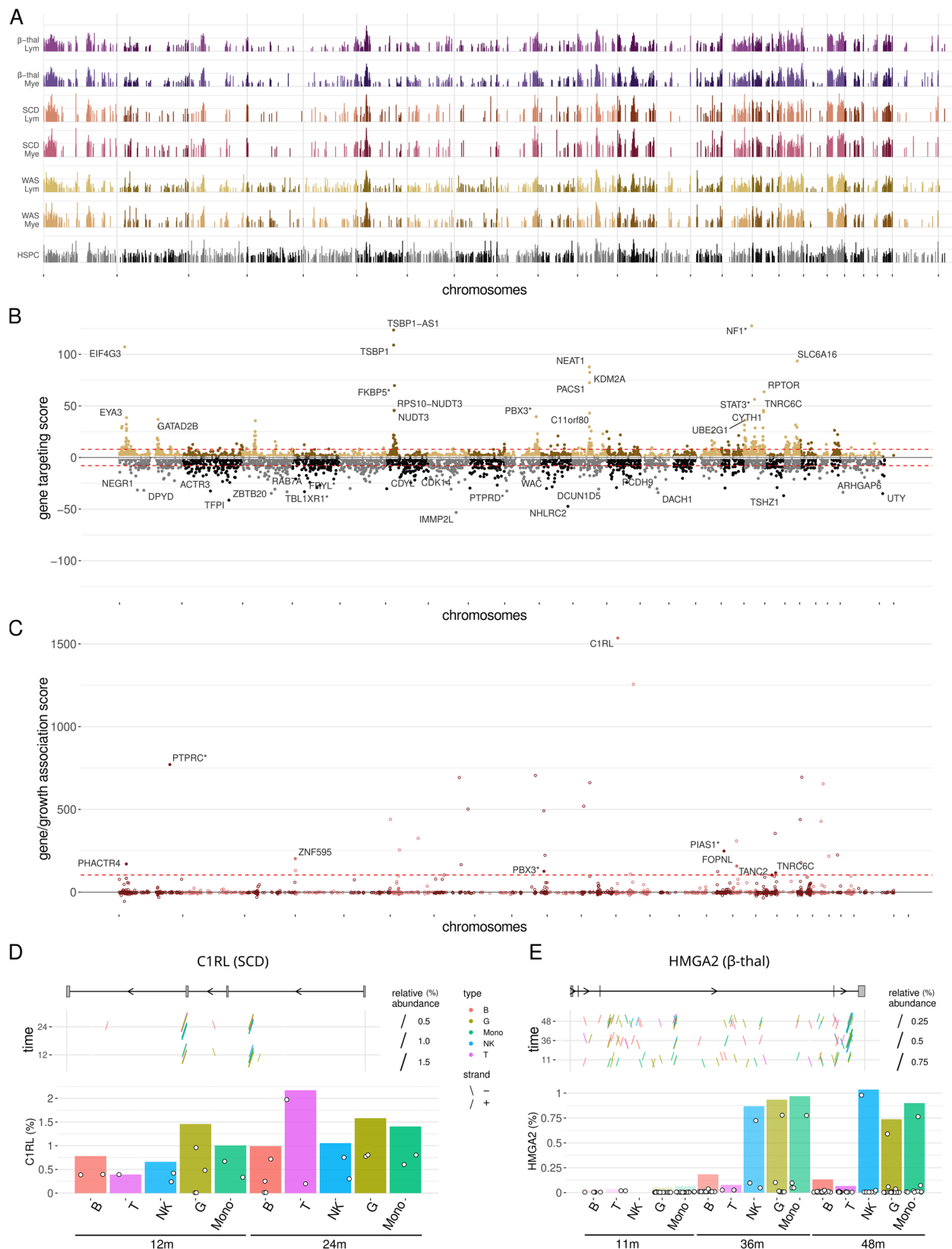
In the SCD dataset (Fig. 7C), six genes were identified with significant increases in growth rate. The most significant, *CIRL* (score = 1558, $p < 1e-11$), encodes Complement Component 1 Receptor-Like protein, involved in immune response modulation. A detailed representation of IS location and abundance in *CIRL* is shown in Fig. 7D. Despite a maximum cumulative contribution of 2%, the consistent increase across multiple independent clones and lineages suggests a potential mechanism linking *CIRL* to IM events in gene therapy. *CIRL*

has been significantly associated with survival in a cohort of acute myeloid leukemia patients and is widely expressed in myeloid leukemia cell lines, including K-562 (chronic myelogenous leukemia), THP-1 (myelomonocytic leukemia), HL-60, and NB4 (acute promyelocytic leukemia), as well as glioma^{43,44}.

In the β -thal dataset, IS in three genes were identified as potentially altering clone fitness and conferring a proliferative advantage (Suppl. Fig. 15). The most significant, *HMGA2* (Fig. 7E, score = 2732, $p < 1e-16$), encodes the AT hook 2 transcription factor, known for binding AT-rich DNA regions. This IM event, extensively described in previous studies^{45–47}, suggests that IS in the last intron generates mRNA lacking let-7 miRNA binding sites, leading to reduced *HMGA2* degradation and driving proliferative advantage. Despite the expansion of *HMGA2*-integrated clones, no progression to a leukemic phenotype was observed, and the event was classified as benign due to its stabilization over time. MELISSA's statistical model detected this IM event using IS tables from just three time points (11, 36, and 48 months), with a cumulative IS contribution in *HMGA2* of only 1%. As shown in Suppl. Fig. 16, this effect is specific to the myeloid lineage, consistent with published data. The second most significant gene, *SFPQ* (score = 458, $p < 1e-5$), is an epigenetic regulator with histone deacetylase binding activity.

Detecting clonal expansion in SCID-X1 gene therapy trial

Enhanced proliferation of clones harboring IS in *HMGA2* has been reported in all eight patients enrolled in the SCID-X1 gene therapy



clinical trial¹⁹, with cumulative clonal contributions up to 15%. This phenomenon has been attributed to the presence of a cryptic splice acceptor site within the cHS4 chromatin insulator element included in the LV vector used for this trial, which can lead to premature transcription termination or the generation of aberrant fusion transcripts. To evaluate whether MELISSA can consistently detect this expansion

across patients, we analyzed IS datasets generated from bead-selected myeloid (CD14⁺, PMN) and lymphoid (CD3⁺ T, CD19⁺ B, and NK) cells isolated from peripheral blood, comprising ~2 million IS. A graphical summary of the dataset is shown in Suppl. Fig. 12B. One patient (p2) was excluded from the analysis due to death at 2.5 years of follow-up from causes unrelated to the treatment.

Fig. 7 | Gene targeting and clone fitness analysis on Six et al. (2020) gene therapy clinical trials IS data. **A** IS enrichment score for genes found significantly over-targeted in at least one group of datasets, using the single-group gene targeting analysis (LRT for null hypothesis that targeting rate in tested genomic interval is equal to baseline). Only statistically significant genes (adjusted FDR p value < 0.05) are plotted. Each bar corresponds to a over-targeted gene with x-axis and y-axis that contain the chromosome location and the estimated regression parameter, respectively. Number of datasets / numbers of tested genomic intervals is 9/7522, 6/8604, 6/1671, 4/1275, 41/9022, 28/4170, 3/5250 for β -thal lymphoid, β -thal myeloid, SCD lymphoid, SCD myeloid, WAS lymphoid, WAS myeloid, HSPC, respectively. **B** Differential gene targeting analysis comparing WAS (69 datasets) and HSPC (3 replicates), 11436 tested genomic intervals. Gene with enriched integration rate in WAS (HSPC) are plotted in brown (gray). Genes marked with a * are

included in the high-risk gene list. The dashed red line represents the threshold for statistically significant enrichment (adjusted FDR p value < 0.05). **C** Gene/growth association in the SCD trial. The gene/growth association score provides a metric for assessing the significance of the fitness advantage or disadvantage conferred by IS within a gene. The score is the signed LRT statistic for null hypothesis that the size of clones with integration inside the tested gene does not increase with time. Genes marked with a * are included in the high-risk gene list. The dashed red line represents the threshold for statistically significant differential clone dynamics (adjusted FDR p value < 0.05). Empty dots show genes that have an adjusted p value for the robust analysis (largest clone in tested gene removed when computing the score) higher than 0.05. **D** Focus plots for clones harboring IS in *CIRL* gene in the SCD trial. **E** Focus plots for clones harboring IS in *HMG2* gene in the β -thal trial. Top: IS position. Bottom: individual and cumulative abundance.

We performed patient-specific clone fitness analyses, progressively incorporating additional IS data over time to determine when MELISSA could first detect the *HMG2* effect. As shown in Fig. 8A, *HMG2* consistently ranked as the top gene influencing clonal fitness in six out of seven patients when the full follow-up period was considered. In four patients (p3, p4, p5, and p6), *HMG2*-associated clonal expansion was already detectable as early as 6 months post-gene therapy.

Figure 8B illustrates the contribution of individual IS to the cumulative *HMG2* clonal abundance in β -thal and SCID-X1 patients. In SCID-X1 patients, sense-oriented IS consistently contributed over 90% of the total *HMG2* signal at all evaluated time points (12, 36, and 48 months), while antisense-oriented IS contributed less than 10%. Spatially, IS in sense between exons 3 and 5 showed a uniform distribution and gradual increase in contribution over time. In contrast, the β -thal patient displayed a less pronounced orientation bias, particularly at early time point (sense/antisense ratio ~70%/30%) but a strong proliferative effect associated with IS located in intron 4, suggesting that positional effects may outweigh strand effects in this context.

While most MELISSA analyses focus on testing groups of IS mapped to predefined genomic regions (e.g., genes or cytogenetic bands) to limit the number of false positive due to stochastic clonal selection and PCR amplification biases, it can also be applied to the analysis of individual clones. For this application, MELISSA requires an individual IS to be detected in at least three distinct samples (time points or cell types) to be evaluated. Figure 8C reports the fitness scores for the IS with a significant growth advantage from the SCID-X1 patients (adjusted p value < 0.1). The highest-scoring clone was identified in Patient 6, with an IS located at chr19:S812269 in *NRTN* (57% contribution), a gene encoding a growth factor known to enhance cell proliferation and survival, particularly in the enteric nervous system and cancer microenvironments.

Interestingly, although *HMG2* consistently ranked highest in gene-level growth association analysis, no single clone with an IS in *HMG2* was identified as the top proliferating clone in any patient, underscoring the value of complementary gene-level and clone-level assessments.

Discussion

Ensuring the safety of gene and cell therapies is a top concern for all stakeholders in the field, from researchers and regulators to patients and their advocates. While IS analysis is a cornerstone of evaluating safety outcomes in preclinical and clinical studies using cell products transduced with viral vectors, the existing analytical frameworks primarily focus on sample report generation⁴⁸, data handling⁴⁹, and modeling of clonal tracking data^{50–52}. This leaves a critical gap in evaluating safety concerns such as gene targeting rates and the presence of clones with altered fitness and proliferative advantage. To address this gap, we developed an analytical framework named MELISSA.

MELISSA is designed as an R package and facilitates the analysis, combination, and comparison of IS datasets generated using diverse conditions and complex experimental design. This powerful tool allows researchers to leverage the flexibility of regression models, generating gene-referenced quantitative scores that measure both targeting rates and their effect on clone fitness, and include dedicated functionality to explore the biological relevance of the results.

The simulation study presented in Fig. 2 highlights the remarkable performance of our methodology. Across all scenarios, the PPV increased as both effect size and sample size grew, showing the consistency of the testing procedure. This is true for target genes of all simulated lengths, although, for long TestGenes (40–80 Kb), the PPV shows a stronger increase with effect size.

The PPV curves for gene targeting (Fig. 2A, B) are similar, despite the differential analysis using twice the data of the single-group analysis. In our simulations, genes with fewer than 2 IS were excluded (users defined parameters). Shorter genes (5–10 Kb) were less likely to be included in the single-group analysis but, when tested, were more easily detected as significant. For genes of typical length (20–40 Kb), effect sizes of 4 (single-group) and 2 (differential) ensured sufficient coverage, though detection rates exceeded 50% only at effect sizes of 16.

In terms of clone fitness analysis, which is crucial for detecting risk of IM, our modeling demonstrated the best performance across all metrics in the single-group scenario. It excelled in terms of PPV, coverage, and detection rate, even for shorter TestGenes and low effect sizes, highlighting its sensitivity. In differential settings, the detection rate was lower because the model primarily leveraged the growth rates of a restricted number of clones located in TestGenes in the two groups.

Recent research highlights the potential of MSCs for gene therapy applications. Their immunomodulatory and regenerative properties, ease of isolation, rapid ex vivo expansion, and multilineage differentiation capabilities make them promising vehicles for targeted therapies in various contexts, including cancer gene therapy⁵³. However, studies have revealed heterogeneity among MSCs derived from different tissues, suggesting this variation might contribute to the diverse levels of efficacy and outcomes observed in animal studies and MSC-based clinical trials.

To address this limitation, we characterized and made available to the scientific community a detailed LV integrome in BM MSCs and Ad MSCs. We then compared their safety profiles, using HSPC transduced with the same LV vector as a reference. We chose HSPCs due to the extensive data available on their LV integration distribution and associated effects, providing a well-established benchmark for comparison.

Figure 3 highlights critical differences in clonal dynamics between the cell types despite both BM MSCs and Ad MSCs demonstrating high transduction efficiency. BM MSCs exhibited a more stable clonal composition, suggesting a more homogeneous cell population that better adapts to the culture conditions. Ad MSCs displayed a marked

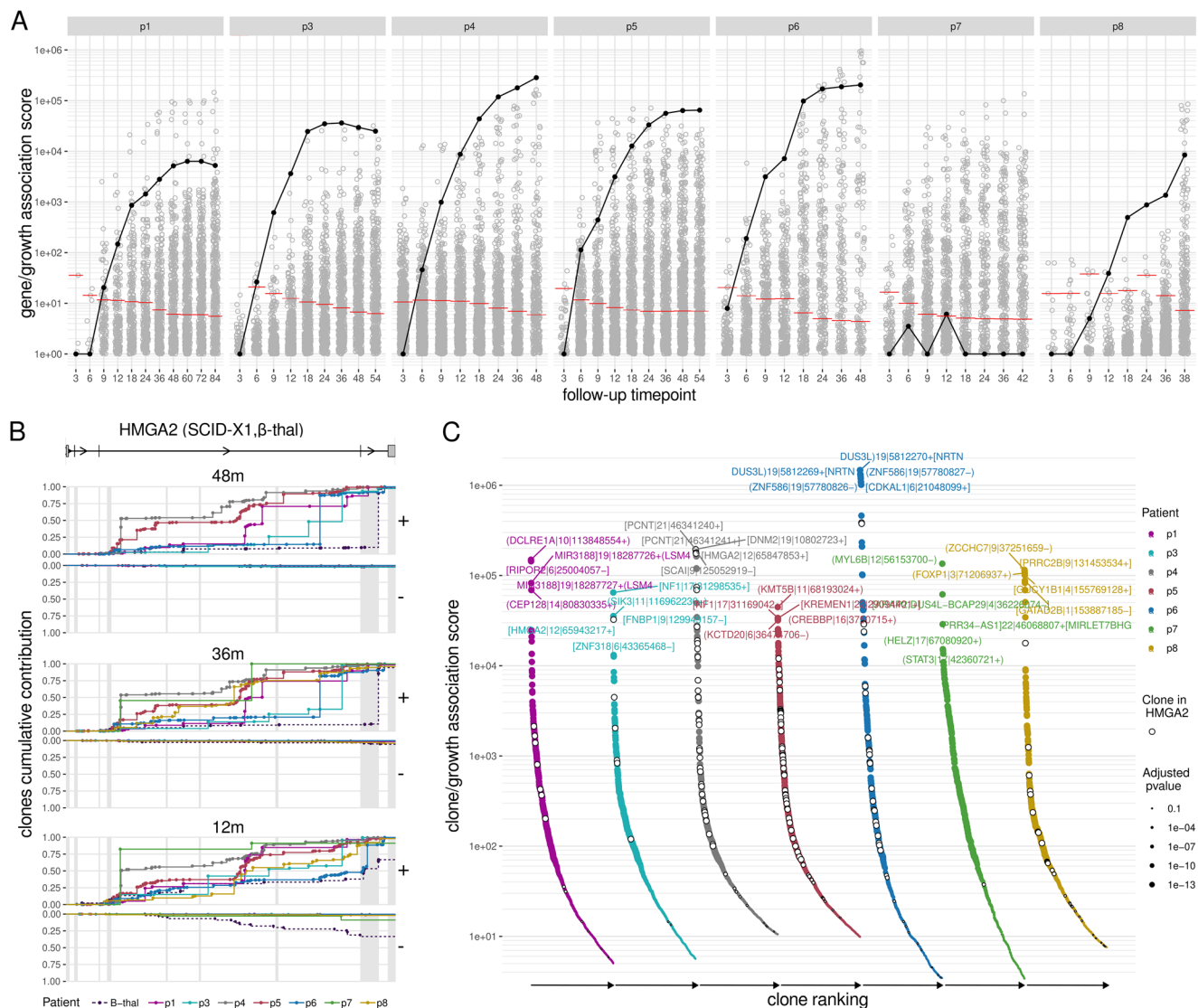


Fig. 8 | Clone fitness analysis on De Ravin et al. (2022) gene therapy clinical trial for SCID-X1. A Single-patient clone fitness analysis including all IS data (of six cell types) up to a given time point (x-axis, in months). The gray dots denote genes with fitness score (y-axis, log10 scale) greater than 1. The fitness score is the signed LRT statistic for the null hypothesis that the size of clones with integration inside the tested gene does not increase with time. Black connected dots show the fitness score for *HMGA2* gene (set to 1 if lower). Red bars delineate significance threshold (adjusted FDR p value < 0.05). At the lowest time point (3 months) the number of tested windows ranges from 233 (patient 1) to 8655 (patient 4). At the largest time point, the range is from 5572 (patient 7, time 42) to 13008 (patient 5, time 54). **B** Focus plot on *HMGA2* showing clone size evolution at three time points. The y-

axis represents the empirical cumulative distribution of clone sizes, stratified by strand of integration; integration site location is on the x-axis. Patients from De Ravin et al. (2022) and one β -thal patient from Six et al. (2022, dashed lines) are shown. Gray areas indicate exons in at least one *HMGA2* transcript. **C** Waterfall plot of single-patient, single-clone fitness analysis. Each dot represents a clone (high-lighted if in *HMGA2* with an adjusted FDR p value < 0.1 ; dot size reflects significance, and the y-axis shows the LRT statistic (log10 scale) for the null hypothesis of no clone growth. The number of tested clones in the seven patients is 30074, 24292, 99426, 116176, 58139, 10388, 22720, respectively. For each patient, the five clones with the highest growth are annotated.

decline in diversity and IS count over the culture period, suggesting a progressive loss of transduced cells and a limited proliferative activity compared to BM MSCs. Therefore, we did not identify an IS profile that could discriminate more primitive culture-initiating cells from the whole MSC cell population. This is in contrast to the study by ref. 54 where primitive culture-initiating cell expansion was observed. It must be noted that MSC in their study were obtained from whole umbilical cord pieces and may well have a different phenotype from both BM and Ad MSCs.

Encouragingly, the distribution of IS with respect to gene annotation was consistent across cell types. We found that the high propensity of LV to integrate within gene bodies extends to MSC and this feature strengthens the relevance of our gene-based approach, with

only 15% of the IS data points not considered in our analysis due to their intergenic position.

Our study generated a valuable resource for researchers working with LV-HSPC gene therapies consisting of a dataset composed of over 35,000 IS across three replicates. This dataset closely aligns with existing knowledge on HSPC integromes regarding RIG and overall association with the epigenetic profile (Figs. 4, 6D).

LV vectors preferentially integrate into transcriptionally active, open chromatin regions, which might suggest that higher gene expression levels would be associated with increased integration frequency. However, as shown in Fig. 4E, our analysis reveals that gene expression levels are a poor predictor of targeting rates in both HSPCs and BM MSCs, with comparable effect sizes across the two cell types.

This suggests that factors beyond transcriptional activity, such as higher-order genome organization and the spatial positioning of genomic regions near nuclear pores, may play a more decisive role in shaping the integrome.

Analysis of gene-specific IS rates revealed both shared and MSC-specific integration preferences. Among the shared targeted genes, *NPLOC4* serves as an example, showing high targeting not only in MSCs but also in ex vivo transduced HSPCs from healthy donors and clinical trial subjects^{55–58}, as well as in LV transduced T cells⁵⁹. In contrast, *PLEC* exemplifies an MSC-specific target, as it was the most frequently targeted gene in MSCs and absent in the IS profiles of HSPCs (Fig. 5A). The frequent targeting of *PLEC* in MSCs is particularly noteworthy given previous findings of high-frequency mutations in the *PLEC* gene in bone marrow-derived MSCs from patients with Acute Myeloid Leukemia, suggesting a potential link between *PLEC* dysfunction and MSC involvement in AML development⁶⁰. This raises potential safety concerns for LV-modified MSCs, especially regarding *PLEC*-targeted integrations.

Furthermore, our study revealed similar targeting of high-risk cancer genes across MSC and HSPC cell types with some distinctions. For example, *SPPL2B*, a protease that cleaves the transmembrane domain of tumor necrosis factor- α to release the intracellular domain and is known for its role in immune regulation and inflammation, was targeted exclusively in BM MSC. More deep studies involving in vivo experiments will be crucial to identify MSC-specific IM risk genes, as our results found low targeting rates and no clonal expansion associated with genes previously known to cause clonal expansion in animal models and human subjects involved with LV/RV HSPC gene therapy.

Gene therapy safety assessment often relies on comparative IS data analysis from different sources. This can involve, as demonstrated in this study, comparing gene targeting rates between alternative cell sources (e.g., BM MSC vs. Ad MSC), vector designs, or transduction protocols.

Figure 5 highlights MELISSA's capability of capturing differentially targeted genes. This allows for in-depth investigation of individual genes and calculation of genome-wide high-risk scores. The weak correlation between gene expression levels and gene targeting scores is further supported by the differential analysis comparing BM MSCs and HSPCs (Fig. 5B). However, enrichment analysis based on differential targeting scores identifies cell type-specific pathways, suggesting that while individual gene expression is a poor predictor of integration, broader transcriptional programs may still shape integration preferences (Fig. 5D, E).

In the comparative analysis focusing on MSC cells (Fig. 5F, G), only *SETD2*, an epigenetic regulator with potential tumor suppressor activity⁴⁰, was identified as differentially targeted in BM MSC, and *DMD* in Ad MSC. However, as demonstrated in the simulation study, the ability of MELISSA to detect differentially targeted genes is contingent upon the strength of the effect and the size of the dataset. It is plausible that by augmenting the number of IS retrieved from the MSC experiment, more differentially targeted genes with weaker associations could be identified, suggesting the potential for further exploration for a more conclusive safety assessment comparison. The analysis based on cytogenetic band annotation covers both intra- and intergenic regions, allowing for a more comprehensive genome-wide comparison of integration profiles. Notably, several genomic regions showed a region targeting scores higher than the top-ranked genes. This highlights a key limitation of gene-based analyses. While these approaches are useful for evaluating potential genotoxic risks associated with known genes, they may overlook the impact of perturbing important regulatory elements, such as distant enhancers, that fall outside gene boundaries but still play critical roles in genome function and safety.

Using MELISSA's clone fitness analysis, we explored the relationship between LV integration and clone fitness in MSCs. Figure 6

demonstrates that distinct sets of genes significantly impacted in-vitro clone growth dynamics within each MSC population. Notably, genes previously linked to clonal dominance or IM in HSPC-based gene therapy exhibited minimal influence on MSC expansion. This reinforces the crucial need for cell type-specific considerations when evaluating gene sets for their potential impact on clone fitness and associated safety risks.

The differential analysis identified a single gene, *RORA*, as significantly impacting clone growth in BM MSCs. In contrast, Ad MSCs displayed a more complex profile, with a greater number of significant genes and stronger growth effects. Notably, clones with IS in *FKBP5*, a known oncogene, are estimated to double their contribution with each passage of in vitro culture. An additional analysis using a sliding window approach further confirmed that Ad MSCs harbor more genomic regions with a significant positive impact on clone growth compared to BM MSCs. It is important to note that these associations might also be influenced by the observed consistent and significant loss of transduced cells in Ad MSCs. This reduces overall cumulative clone size and exacerbates the contribution of surviving clones.

This is the first report of IS analysis of MSC. While the differences noted in BM MSC and Ad MSC are interesting, further studies evaluating different donors are warranted. Gene modification of MSC is likely to transfer chemokine, cytokine, or other genes aimed at improving their potency but could also impact fitness and therefore worthy of future study. It is also worth noting that LV vectors are currently the most prevalent delivery vehicle for patients treated with FDA approved gene therapies and in many ongoing clinical trials.

The association observed between histone modifications and IS distribution in HSPCs aligns with findings reported in SCID-X1³³ and wild-type HIV¹⁵. However, the epigenomic profiles of MSCs only exhibit partial correlation with their respective integromes. Notably, H3K4me1 and H3K9me3 display distinct patterns of association between MSCs and HSPCs with their respective integromes. This inconsistency could be explained by the co-localization of H3K9me3 and H3K4me1 specific to HSPCs or the availability of tethering factors differing between cell types. Further studies are needed to fully understand the biological reasons behind this lack of consistency. However, it highlights the importance of in-depth characterization of how host cell status drives the IS selection process in a cell type-specific manner.

Our analysis across the multiple gene therapy clinical trials reported in ref. 18, as shown in Fig. 7, reveals significant heterogeneity between the integrome observed in vitro within HSPCs and that in myeloid and lymphoid subpopulations isolated from the peripheral blood of GT patients. These differences can largely be attributed to the nature of the cell populations and the timing of data collection. HSPCs are a diverse group, comprising both short-term progenitors and long-term repopulating HSCs. The HSPC data tables were generated shortly after transduction, without undergoing the selective pressures of engraftment seen in clinical data. Conversely, the clinical IS data, generated from peripheral blood samples at least 11 months post-gene therapy, are more reflective of the integrome of long-term HSCs. The consistent patterns of clonal selection observed across different trials, particularly in spatial distribution, suggest that these changes are driven by in vivo clonal selection processes rather than by the specific disease context.

The identification of known gene/growth associations, such as *HMG2* in the β -thal trial, provides evidence that the MELISSA model can capture mechanistic effects despite the limited data available. In the SCD trial, we found clones with IS in *CIRL* to exhibit a similar expansion rate to that of *HMG2* in β -thal. This emphasizes the importance of considering significant associations, especially in situations where data is limited to a single or few patients, as potential indicators of IM events that warrant close monitoring to differentiate between benign and malignant proliferative events and avoid false-positive associations. However, we anticipate that as the

number of patients in a dataset increases, MELISSA will be able to detect more subtle effects and further reduce the incidence of false positives, enhancing these findings' robustness.

In the SCID-X1 trial, MELISSA identified significant expansion of *HMG2A*-integrated clones using IS data from the first 12 months post-gene therapy for six out of seven patients analyzed, demonstrating early detection capability. Timely detection of expanding clones is critical for gene and cell therapy safety, as it enables clinical investigators to monitor clonal dynamics closely and intervene when necessary. As data from subsequent time points were included, the significance of the growth rate enhancement became more pronounced.

The integration of IS clone size dynamics and spatial distribution, as shown in Fig. 8B, provides insight into the potential mechanisms underlying fitness perturbations. Future work will build on this framework to develop methods for inferring the most likely mechanisms of IM and guiding the design of safer therapeutic vectors.

While MELISSA provides valuable insights, it is important to acknowledge its limitations. Relying solely on a gene-centric approach may overlook potential safety concerns. Since different viral vectors exhibit distinct preferences for specific genomic features, a substantial proportion of IS may occur in regions not annotated as genes and thus remain unexamined.

Gene-independent strategies, such as those based on cytogenetic bands or genomic windows of variable length (Figs. 5F, 6B), allow for broader screening across the genome. While establishing a causal relationship between IS location and its impact on clone fitness is more challenging in these contexts, these approaches can uncover IM events involving the disruption of distal regulatory elements. The statistical power and accuracy of such analyses improve with the number of regions tested, although this also increases computational demands.

To address these challenges, future work will focus on developing a data-driven approach to identify genomic intervals that maximize targeting or growth effects, aiming at reducing reliance on user-defined regions, improving result robustness, and enhance computational efficiency.

In conclusion, we present MELISSA, a framework for analyzing IS data to assess the safety of viral vector-mediated gene and cell therapies. MELISSA offers diverse functionalities for gene targeting analysis, differential analysis, clone growth analysis, and visualization. Applying MELISSA MSCs revealed distinct clonal dynamics, cell type-specific gene targeting, and the importance of cell type considerations for evaluating safety. We envision MELISSA to become a valuable tool for empowering researchers, clinicians, and regulators to ensure the safety of novel gene and cell therapies.

Methods

Lentiviral vectors

A third-generation⁶¹ LV encoding EGFP (pcDNA-CSCGW⁶² kindly provided by P. Zoltick, Children's Hospital of Philadelphia, Philadelphia, PA) was produced in a 5 L shaking flask (Thompson, Oceanside, CA, USA) using the packaging plasmids pMDL⁶¹, pRSV-Rev⁶¹, and pMD.G⁶³. Serum-free media adapted suspension HEK293T(IUS) cells, derived by Indiana University School of Medicine (IUSM) from HEK293T cells (Stanford University), were cultured in Expi293 media (Gibco, Cat# A1435102, Grand Island, NY, USA) at 2.5×10^5 cells/mL using shake flasks in a 37 °C 5% CO₂ incubator for at least 2 passages before transfection. Cells were counted on the day of transfection and resuspended in TransFX-H media (Cytiva, Cat# SH3093902, Marlborough, MA, USA) supplemented with 4 mM of GlutaMAX (Gibco, Cat# 35050079, Grand Island, NY, USA). The transfectant, containing OptiMEM (Gibco, Cat# 31985070, Grand Island, NY, USA), 4 plasmid DNA, and PEIpro (Polyplus, Cat# 101000026, Illkirch-Graffenstaden, France), was prepared at 5% of the total cell culture volume. Per 10⁶ cells, 2 µg of total plasmid DNA were added. Plasmids pcDNA-CSCGW,

pMDL, pMD.G, and pRSV-Rev were mixed into OptiMEM at the weight ratio of 4:2:1.3:1, respectively. PEIpro was added to the plasmid mix at a ratio of 1 µL of PEIpro per 1 µg of total plasmid DNA and allowed to polyplex for 15 minutes at room temperature. The transfectants were added to the HEK293T(IUS) cells for a final concentration of 3×10^6 cells/mL; transfected HEK293T(IUS) cells were cultured in a 37 °C incubator supplied with 5% CO₂ for 16 hours. After 16 hours of incubation, TransFX media containing 10 mM of sodium butyrate (Cayman Chemical Co., Cat# 13121, Ann Arbor, MI, USA) was added to the cell culture at a ratio of 1:1. Forty-eight hours post-transfection, culture media were harvested and filtered through 0.4 µm filters (Fisherbrand, Cat# FB12566505, Waltham, MA, USA) to remove cell debris. Filtered LV was allotted into cryovials (Thermo Scientific, Cat# 5000-1020, Waltham, MA, USA), 1 mL/vial, and stored at -80 °C freezer until use.

Transduction of MSC

Bone marrow-derived MSCs (BM-MSCs) were acquired from ATCC (Cat# PCS-500-012, Manassas, VA, USA) and grown in Rooster Media (Rooster Bio, Cat# K82003, Frederick, MD, USA). Adipose-derived MSCs (AT-MSCs) were acquired from ATCC (Cat# PCS-500-011, Manassas, VA, USA) and grown in DMEM/F12(Gibco, Cat# 11320082, Grand Island, NY, USA) supplemented with 5% FBS(Cytiva, Cat# SH30910.03, Marlborough, MA, USA), 10 ng/mL FGF-2 (R&D Systems, Cat# BT-FGFB-AFL-025, Minneapolis, MN, USA), 5 ng/mL EGF (R&D Systems, Cat# 236-EG-200, Minneapolis, MN, USA), and 250 µM ascorbic acid (Lanza, Cat# CC-4398, Walkersville, MD, USA). The cells were maintained in a 37 °C 5% CO₂ incubator. For both BM- and AT-MSC experiments, cells at passage 1 (P1) were seeded at 1×10^4 cells/cm² in 6-well dishes (Corning, Cat# 3516, Corning, NY, USA) for both the control and transduced groups. The following day, the media was aspirated from the transduced group, and 1 mL of LV diluted 1:100 (with fresh culture media) was added to the wells along with Polybrene (final concentration 8 µg/mL) (Millipore Sigma, Cat# TR1003, Burlington MA, USA). Four hours later, the LV solution was aspirated, and fresh culture media was added. Once the wells reached 70-80% confluence, the cells were washed with DPBS (Gibco, Cat# 14040133, Grand Island, NY, USA), harvested using TrypLE (Gibco, Cat# 12605028, Grand Island, NY, USA), and seeded 3×10^3 cells per cm². At passage 4 (P4), the vector copy number (VCN) per cell for BM-MSC and Ad-MSC was 3.91 and 2.69, respectively. At passage 8 (P8), the VCN per cell was 3.45 and 2.82 for BM MSC and Ad MSC, respectively. This procedure was repeated for both the control and transduced samples until passage 5 (P5). At P5, some of the control wells were transduced in the same manner as previously described. The passaging procedure was continued for all three groups until passage 11 (P11). The cells were collected at various time points for insertion site analysis. The cultures were also observed using a fluorescent microscope to observe GFP expression.

Transduction of CD34⁺HSPCs

GCSF-mobilized peripheral blood-derived (mPB) CD34⁺ HSPCs were pre-stimulated for 24 hours and two consecutive overnight transduction were performed at a concentration of 5×10^5 cells/mL at a multiplicity of infection (MOI) of 75 per transduction in the presence of the transduction enhancers synperonic F108 (200 µg/mL final concentration, Sigma-Aldrich, Cat# 07579, St. Louis, MO, USA) and protamine sulfate (4 µg/mL final concentration, Sigma-Aldrich, Cat# P3369, St. Louis, MO, USA) in serum-free stem cell growth medium (CellGenix® GMP SCGM, Cat# 20806-0500, Sartorius CellGenix, Freiburg, Germany,) supplemented with 1% Penicillin/Streptomycin (Gibco, Cat# 10-378-016, Grand Island, NY, USA), 1% L-glutamine (Gibco, Cat# 25030081, Grand Island, NY, USA), 100 ng/mL rhSCF (Gibco, Pepro-Tech®, Cat# 300-07-10UG, Grand Island, NY, USA), 100 ng/mL rhTPO (Gibco, PeproTech®, Cat# 300-18-10UG, Grand Island, NY, USA) and 100 ng/mL rhFLT3L (Gibco, PeproTech®, Cat# 300-19-10UG, Grand

Island, NY, USA). The cells were cultured in supplemented SCGM for ten days before genomic DNA isolation.

Droplet digital PCR for vector copy numbers in transduced cells

Sample DNA was extracted from cells using QIAamp DNA Mini Kit (Qiagen, Cat# 51304, Germantown, MD, USA). The primers and probe nucleotide sequences for detecting LV are forward primer LV-F (5'-ACTTGAAAGCGAAAGGGAAC-3'), reverse primer LV-R (5'-CACCCA TCTCTCTCCTTAGCC-3'), and probe TP-LV (5'-6FAM-AGCTCTCTC GACGCAAGACTCGGC-TAMRA-3'). Apolipoprotein B (APB) gene was used as an endogenous control; the primers and probe sequences for APB are forward primer APB-F (5'-TGAAGGTGGAGGACATTCCTCTA-3'), reverse primer APB-R (5'-CTGGAATTGCGATTCTGGTAA-3'), and probe TP-APB (5'-VIC-CGAGAATCACCTGCCAGACTTCCGT-TAMRA-3'). The primers and probe nucleotides were synthesized by a commercial vendor (Integrated DNA Technologies, Coralville, IA, USA).

Droplet digital PCR (ddPCR) was performed on the Bio-Rad QX200 system to quantify vector copy numbers in transduced MSCs by following the manufacturer's procedures. The controls are no-template controls (water), negative controls (DNA from untransduced CD34(-) cells), and positive controls (1.5 × 10⁵ copies of target sequences). Each sample was tested in triplicates. An excess master mix of reagents without DNA was prepared according to the total number of testing articles (including controls) and the following conditions: for each reaction, a final concentration of 1× no dUTP ddPCR super mix for probe (Bio-Rad, Cat# 1863023, Hercules, CA, USA), 900 nM of LV-F, 900 nM of LV-R, 250 nM of TP-LV, 900 nM of APB-F, 900 nM of APB-R, 250 nM of TP-APB, and 5 units (in 1 µL) of HindIII-HF restriction enzyme (New England BioLabs, Cat# R3104S, Ipswich, MA, USA) were assembled. The master mix was distributed to a microtube before adding 1 µg of testing article DNA; DNase/RNase-free water was added to bring the final volume to 20 µL in each reaction. The 20 µL samples were loaded into a cartridge (Bio-Rad, Cat# 1864008, Hercules, CA, USA) with 70 µL of generation oil (Bio-Rad, Cat# 1864005, Hercules, CA, USA) loaded in all the bottom wells for droplet generation in a droplet generator (Bio-Rad, Cat# 1864002, Hercules, CA, USA). The formed droplets were transferred to a 96-well ddPCR plate (40 µL droplets/well) (Bio-Rad, Cat# 12001925, Hercules, CA, USA) before being sealed with a pierceable foil heat seal (Bio-Rad, Cat# 1814040, Hercules, CA, USA) at 180 °C using a heat sealer (Bio-Rad, Cat# 1814000, Hercules, CA, USA). The sealed plate was placed on a C1000 Touch Thermal Cycler (Bio-Rad, Cat# 1851197, Hercules, CA, USA), and the PCR was run using the following program – Step 1 (95 °C for 10 minutes), Step 2 (40 cycles of 94 °C for 30 seconds followed by 60 °C for 1 minute), and Step 3 (98 °C for 10 minutes). After Step 3, the thermocycler was held at 12 °C until the droplets were read using a QX200 droplet reader (Bio-Rad, Cat# 1864003, Hercules, CA, USA). The data was analyzed using Bio-Rad QX200 Manager Software.

DNA Library preparation for integration site analysis

DNA library for integration site analysis was prepared according to the INSPIRED^{48,64} protocol with slight modifications described below. The genomic DNA from LV-transduced cells was extracted and eluted in DNase-/RNase-free water with QIAamp DNA Mini Kit (Qiagen, Cat# 51304, Germantown, MD, USA) by following the manufacturer's procedure. The extracted genomic DNA was fragmented at the Indiana University School of Medicine Center for Medical Genomics Service Core (IUSM-CMG). DNA fragmentation was performed by sonication using the Covaris ME220 sonication system with 40 Watts peak power, 10% duty factor, 1000 cycles/burst at 25 °C for 30 seconds, Oligonucleotides (linkers, primers, and blocking oligos) for DNA library preparation were synthesized, according to the INSPIRED protocol⁴⁸, by Integrated DNA Technologies, Inc. The DNA library was evaluated using the Agilent High Sensitivity DNA kit (Agilent Technologies, Cat# 5067-4626, Santa Clara, CA, USA) on an Agilent Bioanalyzer 2100

Electrophoresis System at IUSM-CMG to ensure library quality. A maximum of 5 prepared DNA libraries were pooled and sequenced using MiSeq Reagent Kit v2 300-cycle (Illumina, Cat# MS-102-2002, San Diego, CA, USA) on the MiSeq system at IUSM-CMG.

Statistical modeling

The four statistical analyses considered in this manuscript are the following.

A) Gene targeting. The tested gene has a higher IS rate than the rest of the genome. B) Differential gene targeting. The rate of IS in the tested gene is different in the two groups. C) Clone fitness. In the tested gene, clone sizes grow faster (on average, over time) than in the rest of the genome. D) Differential clone fitness. In the tested gene, there is a difference between group 1 and 2 in clones' growth rate. For analyses A) and B), the presence/absence of an IS at each genomic coordinate is modeled as a binary variable (1 if an IS maps to the specific location and 0 otherwise). This binary variable is then considered as the response variable in a logistic regression, where the number of IS inside and outside the tested gene ($y=1$) is subtracted from the length of the tested gene and the total length of the investigated genome ($y=0$). For analyses C) and D), we model clone size growth using logistic regression for proportion data as the response variable and include time in the set of covariates. In all analyses, we performed a likelihood ratio test for each gene. The test statistic is used to assign a p-value to each gene. The p-values are then corrected to account for multiplicity. In the simulation and experimental data analysis, the false discovery rate (FDR) correction method is used. All analyses can be done with an iterative approach, where significant genes are iteratively removed from the dataset until no significant genes are detected. For the clone fitness analyses (C) and (D), we implemented an optional robust statistical calculation to analyze the gene therapy clinical data. This approach was designed to reduce the likelihood of false positives—instances where significant gene-growth associations arise, not from mechanistic modifications of clone fitness due to the IS, but from clonal population dynamics. The clonal selection process, driven by the engraftment bottleneck and the need to maintain homeostasis, is expected to expand certain clones. These expansions are not caused by altered clone fitness due to IS location and are, therefore, likely to occur randomly across the genome. However, with a limited number of independent datasets (e.g., data from only one patient for β -thal and SCD), it becomes challenging to differentiate between mechanistic and clonal dynamics effects. To address this issue, we developed a procedure to reduce our false positive rate. For each gene, the robust statistic is calculated by excluding the IS with the highest count (summed across all cell types and time points) within that gene, mitigating the bias introduced by clonal population dynamics. The robust statistic serves as a filter, with the standard statistic reported only for genes that remain statistically significant when the robust statistic is applied.

Simulation study

The simulation study focuses on chromosome 15, which is ~108 Mb in length. Five new genes, referred to as TestGenes (with length of 5, 10, 20, 40, 80 Kb), were added to the gene list at non-overlapping locations. The final list of 1032 genes, including the TestGenes, was used for simulating integrations across all simulations. The study explores various effect sizes ($h=1, 2, 4, 8, 16, 32, 64$), where $h=1$ represents the null hypothesis, and sample sizes $N=100, 200, 400, 800$ IS for gene targeting analysis (A, B); and cumulative clone sizes $M=1000, 2000, 4000, 8000$ for clone fitness analyses (C, D). For each combination of h and sample size, 1000 simulations were performed.

Gene targeting (A): IS were sampled without replacement, with probability proportional to 1 for IS outside the TestGenes and proportional to h for IS within TestGenes. Four datasets, each with N integrations, were generated for each simulation. Differential gene targeting (B): IS were sampled without replacement, with probability in

TestGenes proportional to: h for IS within TestGenes in group 1, 1 for IS within TestGenes in group 2. For the 1027 endogenous genes, the IS probability (same in both groups) was proportional to a value sampled from an exponential distribution with parameter 1. The baseline IS rate outside of genes was set at 0.001. Each simulation produced four datasets of size N for each group. Clone fitness (C): A total of 10,000 IS were generated using the method described for group 2 in analysis B, representing the total population of IS. At each time point, M elements from the total population are sampled (with replacement). IS clone sizes are defined by the times each IS is sampled. Selective growth for IS within TestGenes is simulated by increasing the sampling probability over time for IS within TestGenes, while for IS outside TestGenes remained constant. The unnormalized probability for an IS within TestGenes evolved as ($q_{it} = 1 + (h-1)t$), while it remained $q_{it} = 1$ for those outside. Differential clone fitness (D): Two sets of 10,000 IS were generated for group 1 and group 2, similar to (C). Clones were sampled for both groups at each time point, with IS outside TestGenes evolving with a random growth rate sampled from a uniform distribution [$\min = 0$, $\max = 2$] for both groups. For TestGenes, group 1 had sampling probabilities evolving as ($q_{it} = 1 + ht$), while group 2 followed ($q_{it} = 1 + t$).

Gene expression analysis

To compare gene targeting results with gene expression profiles, we performed transcript quantification and differential expression analysis using the DESeq2 R package⁶⁵. For the analysis shown in Fig. 4E, raw transcript counts were normalized across datasets and subsequently averaged. Gene symbols were then mapped to the transcript-level scores to enable direct comparison with gene targeting results. For the analysis presented in Fig. 5B, raw transcript counts from HSPC and BM MSC datasets were used to compute differential expression. Test statistics corresponding to the null hypothesis of no differential expression between cell types were extracted, and the resulting signed values were used as differential expression scores.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw data generated in this study on MSC and HSPC lentiviral integrations are available in the NCBI BioProject database under accession code PRJNA1090149 at (<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1090149>) and in the Genome Sequence Archive under accession code SRA1832300. Integration site data from clinical trials for Beta-thalassemia, Sickle Cell Disease and Wiskott-Aldrich Syndrome have been published in ref. 18 and are available in the NCBI BioProject database under accession code PRJNA449379 at (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA449379>). Integration site data from clinical trials for SCID-X1¹⁹ are deposited in Zenodo (<https://zenodo.org/record/8147763>). Source data are provided with this paper.

Code availability

The code used to develop the model, perform the analyses, and generate results in this study is publicly available and has been deposited in MELISSA paper at (<https://github.com/pellinlab/MELISSApaper>) under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. The specific version of the code associated with this publication is archived in Code Ocean capsule available at (<https://codeocean.com/capsule/9281121/tree>).

References

- Ferrari, G., Thrasher, A. J. & Aiuti, A. Gene therapy using haematopoietic stem and progenitor cells. *Nat. Rev. Genet.* **22**, 216–234 (2021).
- Naldini, L. Gene therapy returns to centre stage. *Nature* **2015** **526**, 351–360 (2015).
- Baum, C. et al. Chance or necessity? insertional mutagenesis in gene therapy and its consequences. *Mol. Ther.* **9**, 5–13 (2004).
- Bushman, F. D. Retroviral insertional mutagenesis in humans: evidence for four genetic mechanisms promoting expansion of cell clones. *Mol. Ther.* **28**, 352–356 (2020).
- Cornetta, K., Lin, T. Y., Pellin, D. & Kohn, D. B. Meeting FDA guidance recommendations for replication-competent virus and insertional oncogenesis testing. *Mol. Ther. Methods Clin. Dev.* **28**, 28 (2023).
- Woods, N. B., Bottero, V., Schmidt, M., Von Kalle, C. & Verma, I. M. Therapeutic gene causing lymphoma. *Nature* **2006** **440**, 1123–1123 (2006).
- Corre, G. & Galy, A. Evaluation of diversity indices to estimate clonal dominance in gene therapy studies. *Mol. Ther. Methods Clin. Dev.* **29**, 418–425 (2023).
- Fehse, B. & Roeder, I. Insertional mutagenesis and clonal dominance: biological and statistical considerations. *Gene Ther.* **15**, 143–153 (2008).
- Nienhuis, A. W., Dunbar, C. E. & Sorrentino, B. P. Genotoxicity of retroviral integration in hematopoietic cells. *Mol. Ther.* **13**, 1031–1049 (2006).
- Fda & Cber. Long term follow-up after administration of human gene therapy products; guidance for industry. **1**, 1–35 (2020). https://www.fda.gov/files/vaccines%2C%20blood%20%26%20biologics/published/Long-Term-Follow-Up-After-Admin-Human-GT-Products_Jan_2020.pdf.
- Biasco, L. et al. Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol. Med.* **2**, 1757–4684 (2011).
- Cattoglio, C. et al. High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood, J. Am. Soc. Hematol.* **116**, 5507–5517 (2010).
- Goff, S. P. Host factors exploited by retroviruses. *Nat. Rev. Microbiol.* **2007** **5**, 253–263 (2007).
- Hofmann, A. et al. Epigenetic regulation of lentiviral transgene vectors in a large animal model. *Mol. Ther.* **13**, 59–66 (2006).
- Lelek, M. et al. Chromatin organization at the nuclear pore favours HIV replication. *Nat. Commun.* **6**, 6483 (2015).
- Pellin, D., Biasco, L., Scala, S., Di Serio, C. & Wit, E. C. Tracking hematopoietic stem cell evolution in a Wiskott-Aldrich clinical trial. **17**, 1841–1860 (2023).
- Scala, S. et al. Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.* **24**, 1683 (2018).
- Six, E. et al. Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs. *Blood* **135**, 1219–1231 (2020).
- De Ravin, S. S. et al. Lentivector cryptic splicing mediates increase in CD34+ clones expressing truncated HMGA2 in human X-linked severe combined immunodeficiency. *Nat. Commun.* **13**, 1–15 (2022).
- Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
- Berry, C. C. et al. Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* **28**, 755–762 (2012).
- Brady, T. et al. A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.* **39**, e72 (2011).
- Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).
- Fabregat, A. et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142 (2017).

25. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353 (2017).
26. Ranzani, M., Annunziato, S., Adams, D. J. & Montini, E. Cancer gene discovery: exploiting insertional mutagenesis. *Mol. Cancer Res* **11**, 1141–1158 (2013).
27. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
28. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
29. Berry, C. C. et al. Estimating abundance using the soniclelength package. *Bioinformatics* **28**, 755–762 (2014).
30. Ambrosi, A. et al. Estimated comparative integration hotspots identify different behaviors of retroviral gene transfer vectors. *PLoS Comput. Biol.* **7**, e1002292 (2011).
31. Cattoglio, C. et al. Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* **110**, 1770–1778 (2007).
32. Schröder, A. R. W. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
33. Yan, K. K. et al. Integrome signatures of lentiviral gene therapy for SCID-X1 patients. *Sci. Adv.* **9**, eadg9959 (2023).
34. Sadelain, M., Papapetrou, E. P. & Bushman, F. D. Safe harbours for the integration of new DNA in the human genome. *Nat. Rev. Cancer* **2011** **12**, 51–58 (2011).
35. Engelman, A. & Cherepanov, P. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. *PLoS Pathog.* **4**, e1000046 (2008).
36. Gijssbers, R. et al. LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Mol. Ther.* **18**, 552 (2010).
37. Brunmeir, R. et al. EZH2 modulates mRNA splicing and exerts part of its oncogenic function through repression of splicing factors in CML. *Leukemia* **2025** **39**, 650–662 (2025).
38. Billing, A. M. et al. Comprehensive transcriptomic and proteomic characterization of human mesenchymal stem cells reveals source specific cellular markers. *Sci. Rep.* **2016** **6**, 1–15 (2016).
39. Heydt, Q. et al. Adipocytes disrupt the translational programme of acute lymphoblastic leukaemia to favour tumour survival and persistence. *Nat. Commun.* **12**, 5507 (2021).
40. Chen, R., Zhao, W. Q., Fang, C., Yang, X. & Ji, M. Histone methyltransferase SETD2: a potential tumor suppressor in solid cancers. *J. Cancer* **11**, 3349–3356 (2020).
41. Alnassar, N. et al. Downregulation of dystrophin expression occurs across diverse tumors, correlates with the age of onset, staging and reduced survival of patients. *Cancers (Basel)* **15**, 1378 (2023).
42. Naidoo, M. et al. Duchenne muscular dystrophy gene expression is an independent prognostic marker for IDH mutant low-grade glioma. *Sci. Rep.* **2022** **12**, 1–14 (2022).
43. Laverdière, I. et al. Complement cascade gene expression defines novel prognostic subgroups of acute myeloid leukemia. *Exp. Hematol.* **44**, 1039–1043.e10 (2016).
44. Wang, J. et al. Immunological and clinicopathological characteristics of C1RL in 2120 glioma patients. *BMC Cancer* **20**, 931 (2020).
45. Cavazzana-Calvo, M. et al. Transfusion independence and HMGA2 activation after gene therapy of human $\beta\beta$ -thalassaemia. *Nature* **467**, 318–322 (2010).
46. Kumar, P. et al. HMGA2 promotes long-term engraftment and myeloerythroid differentiation of human hematopoietic stem and progenitor cells. *Blood Adv.* **3**, 681–691 (2019).
47. Mansoori, B. et al. HMGA2 as a critical regulator in cancer development. *Genes (Basel)* **12**, 1–23 (2021).
48. Sherman, E. et al. INSPIRED: a pipeline for quantitative analysis of sites of new DNA integration in cellular genomes. *Mol. Ther. -Methods \ Clin. Dev.* **4**, 39–49 (2017).
49. Merelli, I. et al. InCliniGene enables high-throughput and comprehensive in vivo clonal tracking toward clinical genomics data integration. *Database (Oxford)* **2023**, baad069 (2023).
50. Espinoza, D. A., Mortlock, R. D., Koelle, S. J., Wu, C. & Dunbar, C. E. Interrogation of clonal tracking data using barcodetrackR. *Nat. Comput. Sci.* **1**, 280–289 (2021).
51. Pais, G. et al. ISAnalytics enables longitudinal and high-throughput clonal tracking studies in hematopoietic stem cell gene therapy applications. *Brief Bioinform* **24**, bbac551 (2023).
52. Wagner, S. et al. Clonal reconstruction from co-occurrence of vector integration sites accurately quantifies expanding clones in vivo. *Nat. Commun.* **13**, 1–12 (2022).
53. Mohammadi, M., Jaafari, M. R., Mirzaei, H. R. & Mirzaei, H. Mesenchymal stem cell: a new horizon in cancer gene therapy. *Cancer Gene Ther.* **23**, 285–286 (2016).
54. Selich, A. et al. Massive clonal selection and transiently contributing clones during expansion of mesenchymal stem cell cultures revealed by lentiviral RGB-barcode technology. *Stem Cells Transl. Med* **5**, 591–601 (2016).
55. Aiuti, A. et al. Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science (1979)* **341**, 6148 (2013).
56. Biffi, A. et al. Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science (1979)* **341**, 1233158 (2013).
57. Cartier, N. et al. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**, 818–823 (2009).
58. Abina, S. H.-B. et al. Outcomes following gene therapy in patients with severe Wiskott-Aldrich syndrome. *JAMA* **313**, 1550–1563 (2015).
59. Malach, P. et al. Identification of a small molecule for enhancing lentiviral transduction of T cells. *Mol. Ther. Methods Clin. Dev.* **31**, 101113 (2023).
60. Fracchiolla, N. S., Fattizzo, B. & Cortelezzi, A. Mesenchymal stem cells in myeloid malignancies: a focus on immune escaping and therapeutic implications. *Stem Cells Int* **2017**, 6720594 (2017).
61. Dull, T. et al. A third-generation lentivirus vector with a conditional packaging system. *J. Virol.* **72**, 8463–8471 (1998).
62. Endo, M. et al. Gene transfer to ocular stem cells by early gestational intraamniotic injection of lentiviral vector. *Mol. Ther.* **15**, 579–587 (2007).
63. Naldini, L., Blomer, U., Gage, F. H., Trono, D. & Verma, I. M. Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. *Proc. Natl. Acad. Sci. USA* **93**, 11382 (1996).
64. Berry, C. C. et al. INSPIRED: quantification and visualization tools for analyzing integration site distributions. *Mol. Ther. -Methods \ Clin. Dev.* **4**, 17–26 (2017).
65. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

Acknowledgements

This project has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N92019D00018 K.C. Additional resources utilized include the Indiana University Melvin and Bren Simon Comprehensive Cancer Center Flow Cytometry Core (P30 CA082709 K.C.) and the Sequencing analysis was carried out in the Center for Medical Genomics at Indiana University School of Medicine, which is partially supported by the Indiana University Grand Challenges Precision Health Initiative.

Author contributions

T.Y.L., K.C. and D.P. conceived and designed the experiments. D.P. wrote the manuscript with input from all the authors. G.C., T.Y.L. and K.C. contributed to manuscript editing and revisions. D.P. designed the statistical method. G.C. ran analysis, software implementation, and a simulation study. D.K. and C.B. performed HSPC transduction. T.Y.L., K.C., K.H., M.W., T.T.D., M.P.M. and D.P. performed integration site analysis on MSC and HSPC. T.Y.L., K.C., G.C., and D.P. provided analysis and interpretation. D.P. supervised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s41467-025-63017-w>.

Correspondence and requests for materials should be addressed to Danilo Pellin.

Peer review information *Nature Communications* thanks Eugenio Montini and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at

<http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025