Molecular Therapy
Original Article

Predicting genotoxicity of viral vectors for stem cell gene therapy using gene expression-based machine learning

Adrian Schwarzer,[1,2,10] Steven R. Talbot,[3,10] Anton Selich,[1] Michael Morgan,[1] Juliane W. Schott,[1] Oliver Dittrich-Breiholz,[4] Antonella L. Bastone,[1] Bettina Weigel,[1] Teng Cheong Ha,[1] Violetta Dziadek,[1] Rik Gijsbers,[5] Adrian J. Thrasher,[6] Frank J.T. Staal,[7] Hubert B. Gaspar,[6] Ute Modlich,[8] Axel Schambach,[1,9,10] and Michael Rothe[1,10]

[1]Institute of Experimental Hematology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [2]Department of Hematology, Hemostasis, Oncology and Stem Cell Transplantation, Hannover Medical School, Hannover, Germany; [3]Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany; [4]Research Core Unit Genomics, Hannover Medical School, Hannover, Germany; [5]Molecular Virology and Gene Therapy, KU Leuven, Leuven, Belgium; [6]Molecular and Cellular Immunology Section, UCL Great Ormond Street Institute of Child Health, London, UK; [7]Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [8]Research Group for Gene Modification in Stem Cells, Division of Veterinary Medicine, Paul Ehrlich Institute, Langen, Germany; [9]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Hematopoietic stem cell gene therapy is emerging as a promising therapeutic strategy for many diseases of the blood and immune system. However, several individuals who underwent gene therapy in different trials developed hematological malignancies caused by insertional mutagenesis. Preclinical assessment of vector safety remains challenging because there are few reliable assays to screen for potential insertional mutagenesis effects *in vitro*. Here we demonstrate that genotoxic vectors induce a unique gene expression signature linked to stemness and oncogenesis in transduced murine hematopoietic stem and progenitor cells. Based on this finding, we developed the surrogate assay for genotoxicity assessment (SAGA). SAGA classifies integrating retroviral vectors using machine learning to detect this gene expression signature during the course of *in vitro* immortalization. On a set of benchmark vectors with known genotoxic potential, SAGA achieved an accuracy of 90.9%. SAGA is more robust and sensitive and faster than previous assays and reliably predicts a mutagenic risk for vectors that led to leukemic severe adverse events in clinical trials. Our work provides a fast and robust tool for preclinical risk assessment of gene therapy vectors, potentially paving the way for safer gene therapy trials.

INTRODUCTION

Hematopoietic stem cell gene therapy with retroviral vectors has demonstrated effectiveness in clinical trials for treatment of monogenetic diseases.[1] However, transplantation of genetically modified hematopoietic stem cells led to myelodysplastic syndromes and leukemias in some gene therapy trials.[2–4] These severe adverse events (SAEs) were caused by integration of the provirus in the vicinity of proto-oncogenes, such as *MECOM* and *LMO2*, which were subsequently upregulated by the strong viral promoter and enhancer sequences.[5] Research efforts toward safer gene therapy led to removal of the long terminal repeat (LTR) enhancer elements in first-generation vectors. Instead, the field now mostly uses internal promoters in self-inactivating (SIN) retroviral vector designs.[6–8] However, safety tests of integrating retro- and lentiviral vectors remain a bottleneck for transition from basic research to clinical application. Tumor-prone mice can be used to assess the mutagenic potential of integrating vectors, but these models are laborious, require large numbers of animals, and suffer from long readout times.[9] Another commonly required safety analysis is the integration site pattern of a vector and a screen for clonal dominance in mouse models. However, judging the results of integration site studies regarding clonal dominance versus normal clonal fluctuation in mouse models can be difficult[10] and suffers from poor predictability of the clinical occurrence of SAEs. Hence, efficient and reliable *in vitro* assays to screen for insertional mutagenesis are instrumental for clinical vector development. We previously developed the *in vitro* immortalization (IVIM) assay to quantify the risk of vector-induced cellular transformation.[11] In this assay, murine hematopoietic progenitor cells are expanded after transduction with retroviral vectors. Following limiting dilution, non-immortalized cells stop proliferating, whereas insertional mutants give rise to clonal outgrowth. The incidence of vector-induced immortalization can be used to quantify and compare the mutagenicity of different vector types. Although the IVIM assay mainly detects mutants with insertions near the *Mecom* (also known as *Evi1*) locus, it reliably uncovers the ability of a given vector to activate neighboring proto-oncogenes. Therefore, IVIM results have been accepted by regulatory agencies in Europe, the United States, Canada, and Australia as part of the preclinical safety assessment for gene therapy vectors.[12–15]

Received 26 February 2021; accepted 7 June 2021;
https://doi.org/10.1016/j.ymthe.2021.06.017.

[10]These authors contributed equally

Correspondence: Michael Rothe, Institute of Experimental Hematology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany.
E-mail: rothe.michael@mh-hannover.de

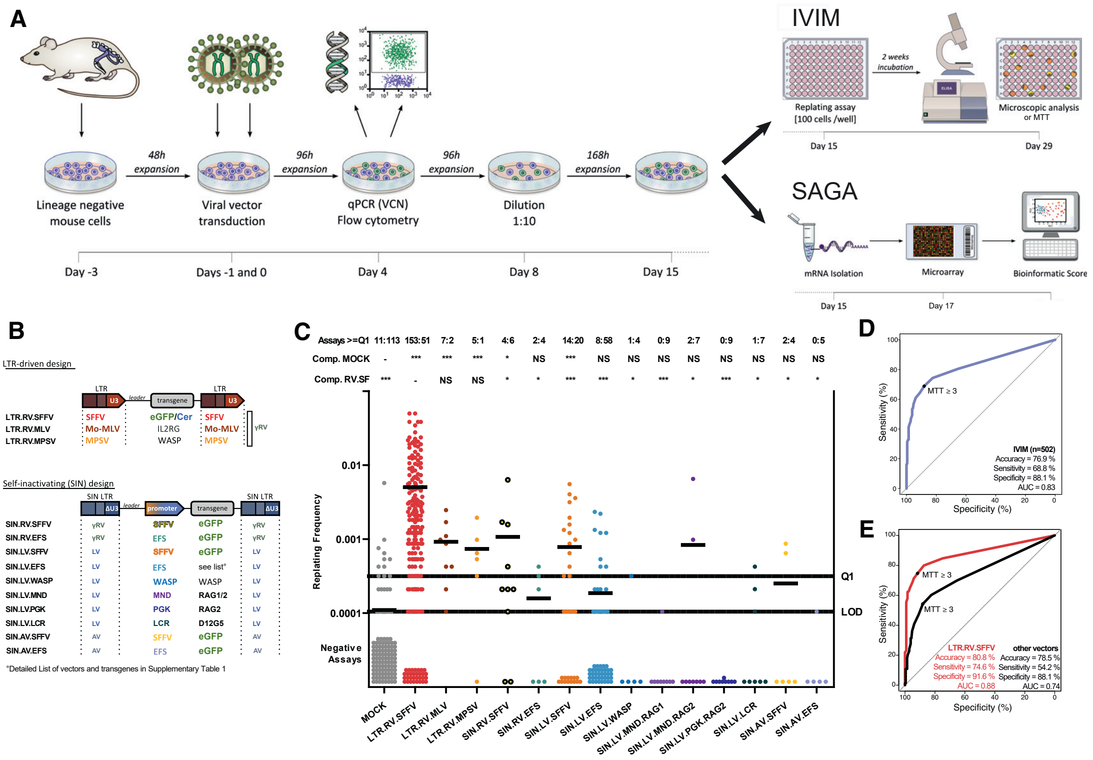

Figure 1. IVIM and SAGA assays to detect vector genotoxicity *in vitro*

(A) Workflow of the *in vitro* genotoxicity assays. (B) Vector designs used in this study. Indicated are the various promoters and transgenes tested in our study (for details, see Table S1). (C) Replating frequencies (RFs) of different IVIM samples (n = 502) measured in 68 IVIM assays. Each dot represents one individual sample. RFs above Q1 (Q1 = 0.75 quantile of the RF for LTR.RV.SFFV) are counted as positive assays. LOD, limit of detection. Above the graph, the ratios of assays with RFs above and below Q1 are shown. Differences in the incidence of positive and negative assays relative to mock- or LTR.RV.SFFV-transduced cells were analyzed by Fisher's exact test with Benjamini-Hochberg correction (*p < 0.05, ***p < 0.001; NS, not significant). Bars indicate mean RF. (D) Receiver operating characteristic (ROC) of the IVIM assay for samples (n = 502) with known activity in the IVIM assay. (E) Same as in (D) with separate curves for strongly transforming vectors (LTR.RV.SFFV) and mock controls (red curve) and weakly transforming vectors, safe vectors, and mock controls (black curve) for which the classification based on repeated testing in the IVIM assay was known.

However, the IVIM assay is limited in terms of sensitivity and suffers from considerable inter-assay variability.

Transformation of healthy hematopoietic progenitors to preleukemic cells and leukemia is linked to specific gene expression programs that cause dysregulation of stemness pathways, growth, and perturbed differentiation.[16–18] By creating a resource of transcriptional responses to vector integration, we show that integrating genotoxic vectors activate a gene expression program linked to transformation, stemness, and cancer de-differentiation. We hypothesized that this transcriptional signature can be exploited to create better predictors of vector-induced genotoxicity. To this end, we develop the surrogate assay for genotoxicity assessment (SAGA) classifier, which uses machine learning to detect dysregulation of this gene expression signature in transduced murine hematopoietic stem and progenitor cells (HSPCs). We compare results from the IVIM assay and SAGA for a variety of integrating benchmark vectors, including the three gammaretroviral vectors that triggered leukemias in clinical trials. The molecular readout of SAGA enhances sensitivity and reproducibility and eliminates the need to rely on the variable replating phenotype of the IVIM assay, reducing assay duration. In addition, we provide the SAGA analysis pipeline as a freely available R package.

RESULTS

Cell culture-based assays for *in vitro* genotoxicity prediction

For IVIM and SAGA, murine lineage-negative (Lin^-) HSPCs are transduced with high multiplicities of infection (MOIs) to reach at least 3 vector copies per cell (Figure 1A). After transduction, bulk cultures are expanded for 15 days in myeloid differentiation-promoting medium. On day 8 after transduction, cells are diluted to increase proliferative selection pressure until day 15. For the IVIM assay, cells are replated at low density and cultured for another 14 days before microscopic and enzymatic detection of growing insertional mutants. The results of the IVIM assay showed that non-transduced mock control cells rarely proliferated under limiting dilution conditions (11 of 124 assays positive), whereas cells transduced with a gammaretroviral vector with strong spleen focus-forming virus (SFFV) promoter/enhancer elements (LTR.RV.SFFV) showed a high incidence of insertional mutants (153 of 204 assays positive). These results are congruent with clinical data because this vector design led to myeloid malignancy in clinical trials for chronic granulomatous disease because of insertional activation of *EVI1* and *PRDM16*. We tested a variety of gammaretroviral, alpharetroviral, and lentiviral vectors with different designs and mutagenic potential in both assays, including several vectors that have been or are currently used in clinical trials (Figures 1B and 1C; Table S1). The IVIM assay identified the potential hazard of two other vectors with clinically demonstrated genotoxicity: LTR.RV.MLV.IL2RG (also known as MFGγC), which caused leukemia in X-SCID studies,[2] and LTR.RV.MPSV.WASP (also known as CMMP-WASP), which led to several primary and secondary leukemias in individuals with Wiskott-Aldrich syndrome (WAS).[4,19] However, the replating frequencies and, hence, the power of the IVIM assay to uncover the mutagenic potential was substantially lower for vectors that contain weaker LTRs or SIN lentiviral vectors with SFFV as an internal promoter. We summarized the outcome of 502 IVIM assays in a receiver operating characteristic (ROC) curve (Figure 1D) to determine the predictive power of the replating phenotype as a proxy for vector mutagenicity. Overall, the IVIM assay showed a low false negative rate (specificity, 88.1%) and a sensitivity of 68.8%, reaching an overall area under the receiver operating curve (AUROC) of 0.827 (Figure 1D). Next we evaluated the results separately for mock control cells against the strongly transforming vector LTR.RV.SFFV (Figure 1E, red curve; sensitivity, 74.6%; $\text{AUROC}_{\text{LTR.RV.SFFV}}$, 0.88) or other vectors with known mutagenicity (Figure 1E, black curve; sensitivity, 54.2%; $\text{AUROC}_{\text{other}}$, = 0.74). Hence, the IVIM assay has significant predictive power to detect vector-induced immortalization in mouse hematopoietic cells, but for vectors other than the LTR.RV.SFFV, the IVIM assay suffers from low sensitivity and has to be repeated many times to obtain a reliable prediction of vector safety.

Transforming vectors impose an oncogenic gene expression signature

We hypothesized that gene expression changes induced by transforming vectors might be a more accurate and sensitive predictor of vector-induced genotoxicity than the occurrence of a poorly defined clonal outgrowth after long periods of *in vitro* culture. Therefore, we analyzed the transcriptome from HSPC bulk cultures on day 15 after transduction. t-distributed stochastic neighbor embedding (t-SNE)[20] and heatmaps of single assays (Figures 2A–2D) revealed that transduction with transforming vectors (based on the IVIM assay) imposed a distinct gene expression signature in HSPCs, clearly distinguishing them from the mock samples. In contrast, SIN vectors with weaker internal promoters, such as the EF1α-short (EFS) promoter, clustered with the non-transduced mock controls (Figures 2C and 2D). Importantly, gene expression changes were linked to but independent of the full immortalization phenotype in the IVIM assay. Cultures that were transduced with transforming vectors but did not immortalize in the IVIM assay (indicated as closed circles in Figure 2A) still showed similar gene expression changes compared with LTR.RV.SFFV immortalized samples (Figures 2A and 2B). The transformation-associated gene expression changes were observed across different vector genera as transforming gammaretroviral, lentiviral, and alpharetroviral vectors clustered together (Figures 2C and 2D; Figure S1A). The most consistently dysregulated genes (absolute $\log_2\text{FC} > 1.0$, $P_{\text{adj.}} < 10^{-5}$; Table S2, tabs 1 and 2) in samples transduced with transforming vectors included stem cell-associated genes (*Aldh1a1*, *Smim5*, and *Ifitm6*), proto-oncogenes (*Zbtb16*, *Sox4*, *Pdgfrb*, and *Fgf3*), stem cell transcription factors or their target genes (*Spns2*, *Ces2g*, and *Myct1*), and myeloid markers (*Mpo* and *Cebpe).* This oncogenic signature was detected as early as day 4 after transduction with transforming vectors (Figures S1B and S1C) and across three gene expression platforms (microarrays, qPCR, and RNA sequencing [RNA-seq]; Figures S1D and S1E). Gene set enrichment analysis (GSEA) showed upregulation of gene sets linked to positive cell cycle regulation, hematopoietic stem cells, erythroid/megakaryocytic differentiation, and *Evi1* target genes[21] in samples transduced with transforming vectors compared with mock controls and safe vector designs (Figures 2E, 2F, 2H, and 2I; Table S3, tab 2). Interestingly, samples transduced with safe vector designs displayed a similar enrichment of cell cycle gene sets and genes linked to erythroid/megakaryocytic differentiation compared with mock controls (Figure 2G). However, in contrast to transforming vectors, they showed downregulation of stemness and *Evi1* target genes and a reduction of myeloid gene sets (Figures 2G and 2J; Table S3, tab 8). We hypothesized that the accelerated cell cycle and upregulation of genes associated with non-myeloid lineages might be a sign that assay progression and myeloid differentiation were generally delayed in transduced samples, independent of the vector type. Therefore, we compared early mock samples from day 8 with mock samples from day 15. Indeed, non-transduced HSPCs from this early time point displayed a similar enrichment of cell cycle, erythroid, and megakaryocytic gene sets but no upregulation of *Evi1* target genes (Figure 2K; Figure S1F; Table S3, tab 14), reflecting incomplete myeloid differentiation and faster proliferation at that time. Most importantly, only genotoxic vectors upregulated hematopoietic stem cell transcriptional programs and *Evi1* target genes compared with mock and non-transforming vectors. The upregulation of myeloid differentiation genes and stemness programs by transforming vectors (Figures 2E and 2F) underscores that differentiation and transformation are not mutually exclusive, as described for *Evi1*-driven leukemogenesis.[22] By probing more than 8,000 gene sets from the MSigDB collection,[22] we sought to obtain a more global view of the biological processes and pathways altered by transforming vectors. We found that transforming vectors triggered an early transcriptional signature that already included several "hallmarks of cancer,"[23] including upregulation of gene sets linked to DNA replication, stemness, cancer de-differentiation, and therapeutic resistance and an enrichment of interferon signaling genes (Figure 2L; Table S3, tab 11). These data demonstrate that integrating vectors with the propensity to transform hematopoietic cells induce a unique oncogenic gene expression signature that distinguishes them from non-transforming vectors.

Dataset preparation for classifier development

Having shown that genotoxic vectors impose a specific stemness-related gene expression profile, we sought to develop a machine learning algorithm distinguishing transforming vectors from safe

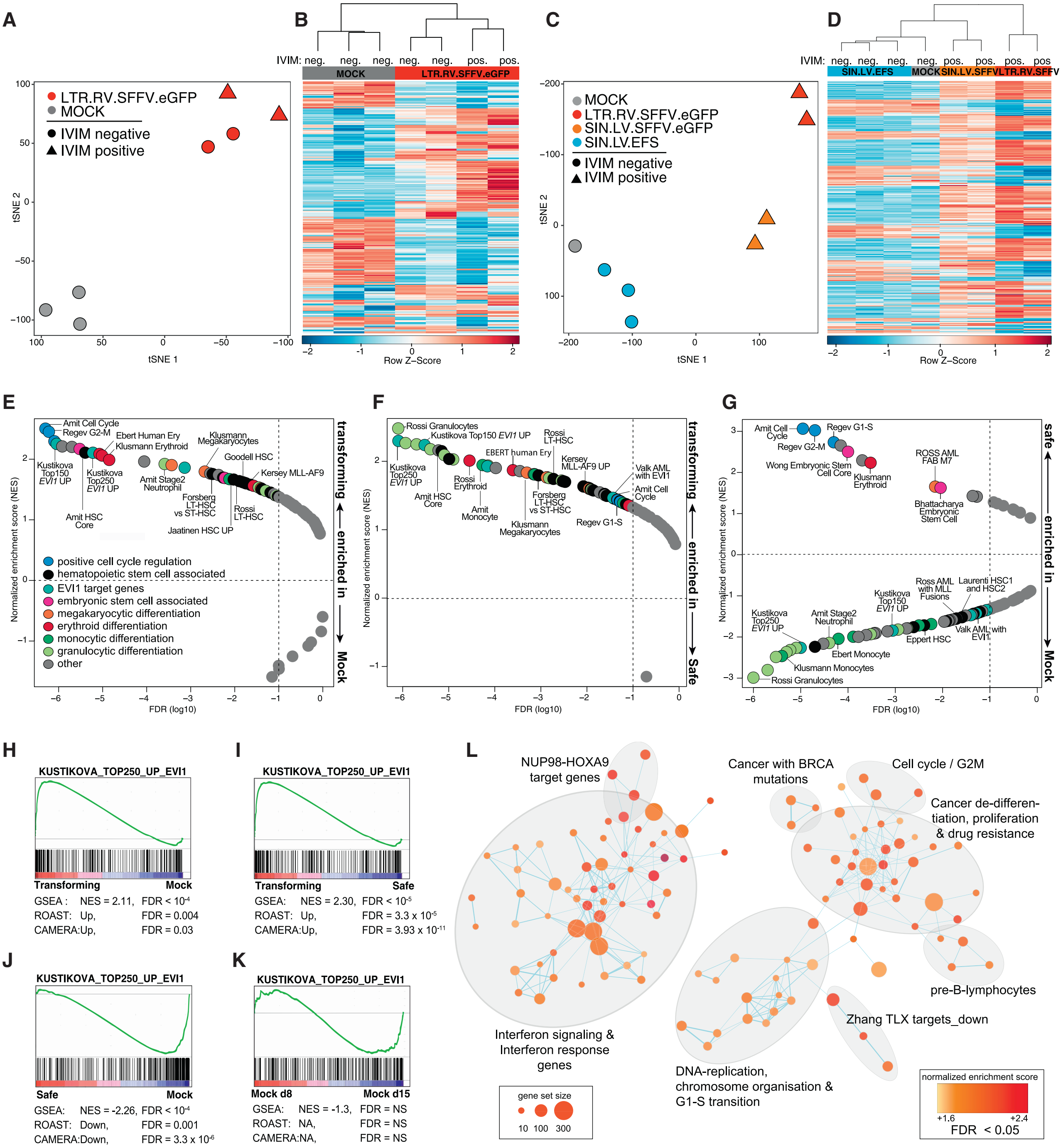

Figure 2. Transforming vectors impose an oncogenic gene expression signature in murine HSPCs

(A) t-distributed stochastic neighbor embedding (t-SNE) of three mock samples (gray) and 4 samples transduced with LTR.RV.SFFV (red) from one SAGA assay (ID 120411) using all 36,226 annotated probes. (B) Hierarchical clustering of the samples shown in (A) based on the most variable probes (top 1%). (C) t-SNE of a second SAGA assay (ID 150128, 36,226 annotated probes). (D) Hierarchical clustering of the samples shown in (C) based on the most variable probes (top 1%). (E) Gene set enrichment analysis (GSEA) of hematopoiesis-associated gene sets (Table S3, tab 1) of samples transduced with IVIM-transforming vectors versus mock controls. Plotted are normalized enrichment scores (NESs) against the false discovery rate (FDR). The enrichment cutoff (FDR < 0.1) is indicated by the dashed line. (F) GSEA of IVIM-transforming vectors against IVIM-safe vectors. (G) GSEA of samples transduced with IVIM-safe vectors against mock controls. (H–K) GSEA plots for EVI1 target genes[20] for the contrasts (H) transforming versus mock, (I) transforming versus safe, (J) safe versus mock, (K) mock day 8 versus mock day 15. (L) Enrichment map of highly upregulated (FDR < 0.005) gene sets from MSigDB in samples transduced with transforming vectors compared with mock control and safe samples.

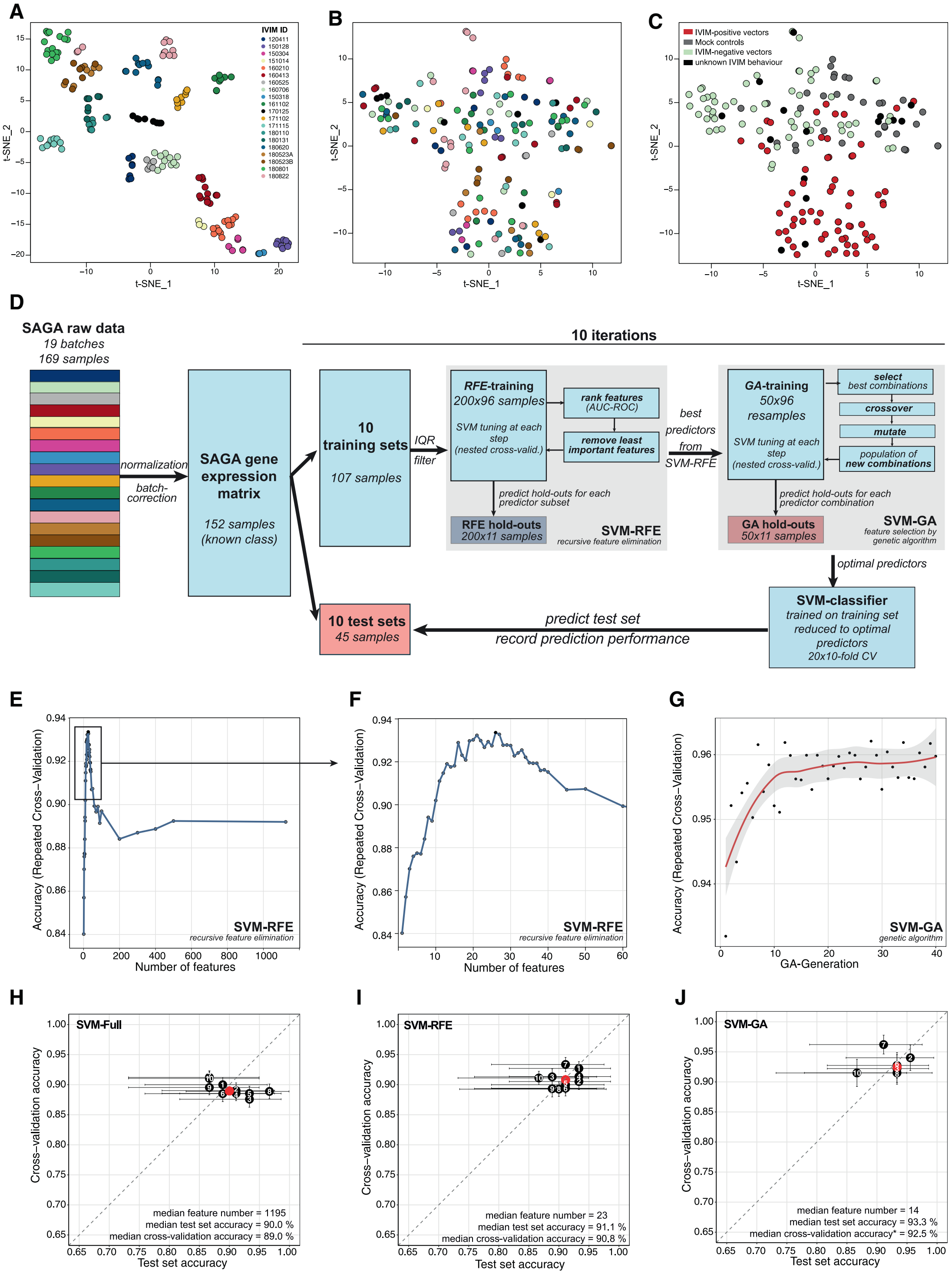

Figure 3. Development phase of an SVM classifier to predict genotoxicity

(A–C) Data preprocessing. (A) t-SNE representation of all 169 SAGA assays after quantile normalization using all 39,428 probes. The coloring scheme encodes individual SAGA assays. (B) t-SNE of the 169 SAGA-samples after quantile normalization and ComBat correction using the same color key as in (A). (C) t-SNE plot as in (B) with the samples color coded according to vector properties in the IVIM assay. IVIM positive, transforming vectors; IVIM negative, nontransforming vectors; mock, untransduced controls; unknown, IVIM data inconclusive. (D) Scheme of classifier development during the development phase. The complete raw dataset was quantile normalized and

(legend continued on next page)

designs. We started with the full dataset consisting of 169 SAGA microarrays with 39,428 probes each, resulting in more than 6 million data points. Similar to other cell culture assays analyzed with high-throughput methods,[24] we observed systematic differences between assays because of different primary cell material, reagent lots, instruments, time, and personnel ("batch effects"). These batch effects caused clustering of samples by assay and processing date after normalization (Figure 3A). Thus, we first implemented a robust normalization and batch effect correction method to reduce the unwanted variation between individual SAGAs. We found that a combination of quantile normalization[25] (Figures S2A and S2B) followed by ComBat[26] effectively normalized and removed batch effects, allowing us to analyze all samples within a common gene expression space where transforming and safe/mock groups formed two different but overlapping groups (Figures 3B and 3C). Importantly, both methods are capable of "add-on adjustment" of new test batches, allowing cross-batch predictive modeling by leaving training data and classification rules fixed when new test data are adjusted.[27] For the classifier development phase, the jointly normalized and batch-corrected gene expression matrix was reduced to samples with known properties in the IVIM assay (transforming/non-transforming/mock, n = 152; Figure 3D). We split the dataset into 10 different training and test sets, with the training sets comprised of 70% and corresponding test sets comprised of 30% of the samples. Development of models was performed on the training sets using repeated cross-validation to assess model performance during feature selection and hyperparameter tuning (Figure 3D; Figure S3). The test sets were then used to assess the performance of the final model fit and to control for overfitting of the classifier; for instance, because of feature selection bias or hyperparameter tuning.[28,29]

Genetic algorithm-enhanced feature selection for prediction of genotoxicity

Initial testing of several machine learning approaches revealed that support vector machines (SVMs) offered a good classification performance on our dataset. Because the performance and computational cost of SVMs are negatively influenced by non-informative predictors, we performed feature selection on the training sets to find smaller predictor subsets with higher predictive power. A second aim for feature selection was to reduce the number of predictors as far as possible for potential later transfer of SAGA to other technical platforms that offer higher sample throughput but can interrogate fewer predictors. Starting with all 36,226 annotated probes, an unsupervised filtering step was applied to remove all probes with little or no variation across the training set, leaving a median of 1,195 probes (Figure 3D; Figure S3; Table S4, tab 1). Next we performed recursive feature elimination (SVM-RFE),[30] which ranks all predictors according to their individual predictive power and then iteratively removes the least important predictors. The performance profiles across the different predictor subset sizes showed performance maxima between 3 and 45 predictors for the different training sets (median, 23 predictors; Figures 3E and 3F; Table S4, tab 1). On average, SVM-RFE removed more than 98% of the probes from the dataset, which resulted in a slight but significant boost in prediction performance, as measured by cross-validation (median accuracy full models, 89.0%; median accuracy RFE models, 90.8%; median P_{paired}, 0.038; Table S4, tab 1). On the separate test sets, the median accuracy for the full models was 90.0%, whereas the SVM-RFE models achieved a median accuracy of 91.1% (Table S4, tabs 1 and 2). Notably, the SVM-RFE models required less than a tenth of the computation times of the full models because of the smaller number of predictors.

SVM-RFE is a greedy algorithm that is effective in eliminating large numbers of less important probes, but it does not perform an exhaustive search to find the best combination of retained predictors. We hypothesized that an optimal combination of probes with high predictive power would allow us to further reduce the number of required predictors while maintaining or even increasing prediction performance. To this end, we next employed a genetic algorithm (GA) to find the best combination of probes retained by SVM-RFE. GAs search for the best solution in a given feature space, guided by evolutionary principles.[31] GAs have been shown to efficiently find optimal or near-optimal solutions for complex optimization problems, including feature selection.[29] We implemented the GA together with support vector machine-based modeling (SVM-GA) and used cross-validation to asses predictive performance during the feature selection process. For SVM-GA, a population of 40 candidate solutions (individuals) was initially created from random subsets of the most informative probes found by the preceding SVM-RFE step. The predictor subsets with the highest fitness (prediction performance) of each generation had the best chances to survive and produce the next generation of predictor subsets by random crossover and mutation, producing more and more optimized probe combinations over time (Figure 3D; Figure S3). We performed feature selection using SVM-GA for all training/test set splits where SVM-RFE had retained more than 10 predictors. This ensured that the GA could choose from a sufficient number of predictors to create the initial population of predictor subsets. In three of four cases where SVM-RFE alone had arrived at less than 10 predictors, cross-validation accuracy (median accuracy full models, 89.0%; median accuracy RFE models, 90.7%; median P_{paired}, 0.024) and test set accuracies (median accuracy full models, 90.0%; median accuracy RFE models, 92.2%; Table S4, tab 1) were already improved,

batch corrected. The dataset was split 10 times into training (70% of samples) and test sets (30% of samples). Feature selection by SVM-RFE and SVM-GA was performed by further splitting the training sets using repeated cross-validation and monitoring prediction performance using the hold-out samples. Tuning of the SVM was performed at each step of the feature selection routines using nested cross-validation. An SVM with radial kernel was trained on the training set reduced to the optimal predictors found by SVM-RFE and SVM-GA and used to predict the test set. (E and F) Performance profile of SVM-RFE: accuracy on the hold-out samples plotted against the number of remaining probes during SMV-RFE for a representative training set (split 7). (G) Performance profile of SVM-GA: accuracy on the hold-out samples plotted against generation of the GA for training set 7. (H and I) Estimates of the prediction accuracy for the full models (H), RFE models (I), and GA models (J) using the test set (x axis) or repeated cross-validation (y axis). The horizontal and vertical bars represent the 95% confidence intervals using the test set and resampling approach, respectively.

making a second round of feature selection unnecessary. The cross-validation performance over the 40 generations of SVM-GA was aggregated into a performance profile demonstrating the progress of the algorithm and to choose the optimal generation for the entire dataset (Figure 3G). The algorithm further reduced the median number of predictors from 36 to 14 for the six training/test set splits subjected to SVM-GA. Importantly, cross-validation accuracy and test set accuracy showed improved predictive power compared with the full and SVM-RFE models (median cross-validation accuracy SVM-GA, 92.5%; median cross-validation accuracy full model, 89.0%; median test set accuracy SVM-GA, 93.3%; median test set accuracy full model, 90.0%; Figures 3H–3J; Table S4, tabs 1 and 2).

Because the performance of the final SAGA classifier, which was built on all 152 samples available, could only be estimated via cross-validation, we assessed whether our cross-validation strategy was trustworthy or whether it produced overly optimistic results because of feature selection bias or overfitting. Therefore, we plotted the accuracy estimates from cross-validation within the training sets against the test set accuracies (Figures 3H–3J). The performance estimates obtained by cross-validation and the test set accuracies showed good agreement, with the median of both estimates for the different splits (red dots in Figures 3H–3J) being located near the identity line. The wider confidence intervals of the accuracies calculated from single test sets demonstrated the higher uncertainty of performance estimates obtained from a test set of limited size compared with properly implemented resampling.[28] Next we defined SAGA as the compound classifier obtained on a training set when SVM-RFE retained less than 10 predictors for this training set and used SVM-RFE followed by SVM-GA otherwise.

Assessing the predictive performance of SAGA

In the classifier development phase, the test samples were selected by random sampling from a jointly preprocessed gene expression matrix to obtain test sets of sufficient size and with the same class distributions as the training set. However, this approach did not fully reflect the later test scenario, where a new test set is to be predicted by SAGA using a preprocessed and fixed training set and classifier. To realistically assess the predictive performance of SAGA with unseen data in the absence of external validation data, we employed a jack-knife, leave-one-batch-out approach. The SAGA dataset was comprised of 19 individual SAGAs (batches). For each iteration, one complete batch was set aside as an independent test set (Figure 4A). The remaining 18 batches were used as the training set to which the preprocessing and feature selection pipeline developed above was applied (Figure 4A; Figure S3). On median, 10 optimal predictors were derived from the training sets (Table S5, tab 1), and an SVM was trained on the training set reduced to the optimal predictors (Figure 4B). Next, the batch that served as the independent test set was add-on normalized and add-on batch corrected. This adjusted the test set to the training set without altering the latter (Figure 4C), ruling out data leakage from the test samples into the training set or the classifier.[32,33] Finally, the add-on adjusted test set was reduced to the optimal predictors, and the class labels were predicted. We repeated this procedure for all 19 batches and aggregated the prediction results over the 19 iterations (Figures 4D–4I; Figure S4; Table S5, tabs 1 and 2). Compared with the IVIM assay, SAGA outperformed the IVIM assay in terms of AUROC ($\text{AUROC}_{\text{SAGA}}$, 0.940; $\text{AUROC}_{\text{IVIM}}$, 0.827; $P_{\text{DELONG}} < 10^{-4}$; Figure 4D), overall accuracy ($\text{accuracy}_{\text{SAGA}}$, 90.9%; $\text{accuracy}_{\text{IVIM}}$, 76.9%; Figure 4D), and the area under the precision recall curve ($\text{AUPRC}_{\text{SAGA}}$, 0.944; $\text{AUPRC}_{\text{IVIM}}$, 0.89; Figure 4G). Specifically, SAGA had a markedly higher sensitivity (87.7%) compared with the IVIM assay (68.8%). Most importantly, SAGA detected the genotoxicity of strongly transforming vectors ($\text{accuracy}_{\text{SAGA}}$, 97.1%; $\text{accuracy}_{\text{IVIM}}$, 80.8%; Figures 4E and 4H) and vectors with weaker transforming potential with higher predictive power than the IVIM assay ($\text{accuracy}_{\text{SAGA}}$, 88.9%; $\text{accuracy}_{\text{IVIM}}$, 78.5%; Figures 4F and 4I). The negative predictive value of SAGA was much higher than that of the IVIM assay (negative predictive value $[\text{NPV}]_{\text{SAGA}}$, 0.91; NPV_{IVIM}, 0.67; Table S5, tab 2); therefore, there is much higher confidence to classify a vector as "safe" for clinical use when using SAGA than IVIM. The number of predictors at each step and the classification performance metrics over the 19 iterations of the leave-one-batch-out approach are summarized in Figures S4A–S4D. To assess the stability of the feature selection process, we quantified how often individual predictors were included in the set of optimal predictors in each of the 19 iterations (Table 1; Table S5, tab 3). We found a high degree of overlap between the sets of optimal predictors found during the different iterations with a core set of highly potent predictors, such as Naip1 and Itih5, which were included in most of the sets (Table 1). These predictors also showed a high degree of overlap with the features found for the random test sets during the development phase (Table 1; Table S4, tab 3), as well as with the final list of 11 optimal features (Table 1; Table S4, tab 5) obtained on the complete SAGA dataset (described below).

Construction of the final SAGA model

Next, using the pipeline developed above, we built the final SAGA classifier on the entire set of available samples (n = 152) for use as the training set in the SAGA R package. After variance-based filtering, 1,243 features (Table S6, tabs 1 and 2) were supplied to SVM-RFE (Figure 5A), which retained 20 predictors (Table S6, tab 3). The following SVM-GA step found an optimal combination of 11 predictors after 14 iterations for the complete dataset (Figure 5B; Table S6, tab 4). Principal-component analysis of the SAGA dataset reduced to these 11 probes showed a clear separation of IVIM-transforming vectors against mock controls and IVIM-neutral vectors (Figure 5C), whereas a separation of classes was not discernible when the dataset was reduced to 11 randomly selected probes (Figure 5D). Finally, we queried transcriptome data from the Immunological Genome Consortium[34] to determine which cell types of the hematopoietic system expressed the 20 most important predictors found by SVM-RFE. A majority of these predictors were highly expressed in the most immature hematopoietic stem cells (Figure 5E), whereas the predictors that were retained after unsupervised filtering and used as input into the feature selection process showed no such association with HSCs (Figure S5). The remaining genes, such as *Frat2* and *Traf4*, were mainly associated with the lymphoid lineage, whereas none of these genes was expressed in mature granulocytes, a route of

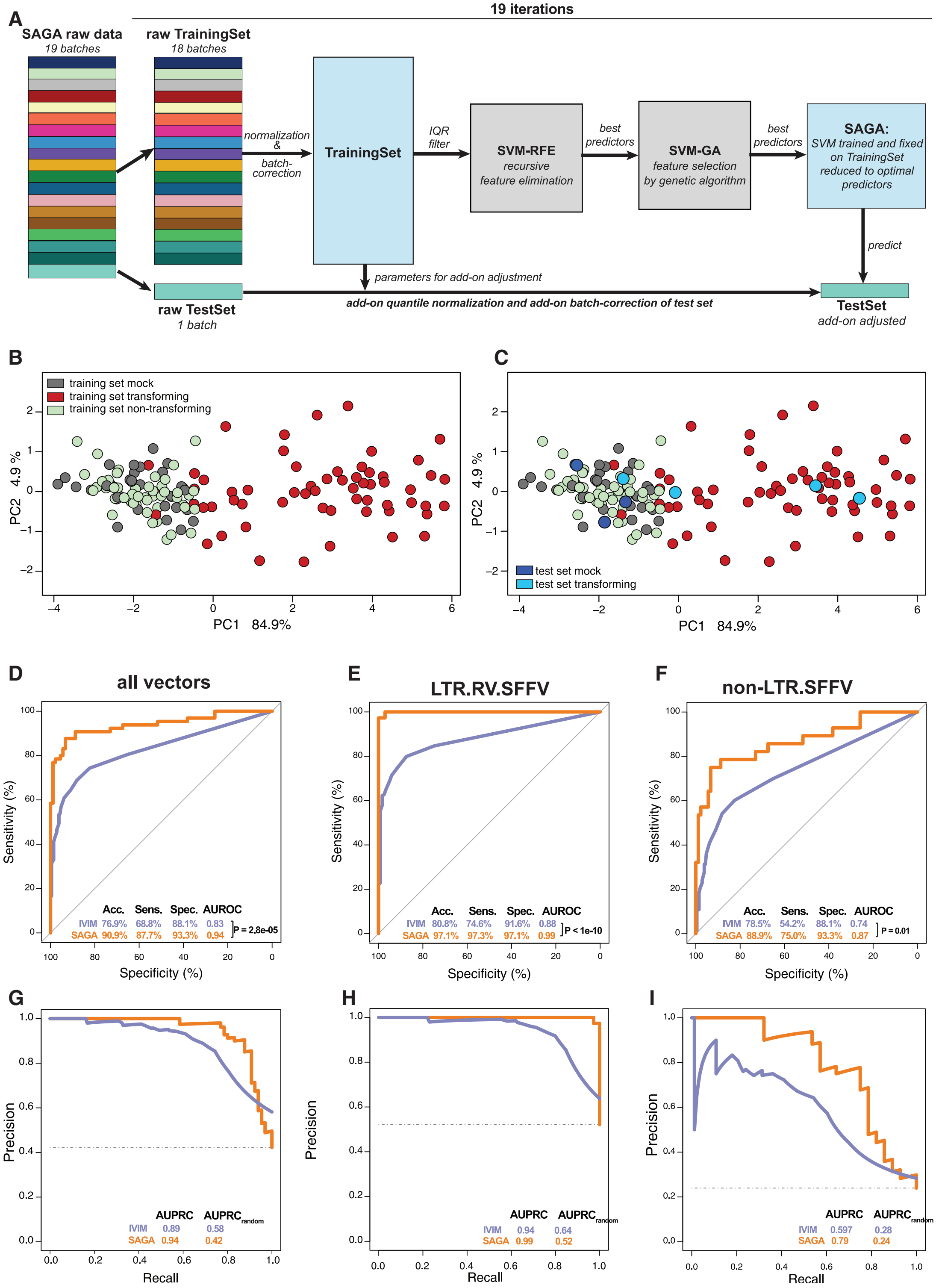

Figure 4. Estimation of model performance via the leave-one-batch-out approach

(A) Scheme of the leave-one-batch-out approach used to estimate SAGA performance. Details are given in the main text. (B) PCA representation of training set 01 reduced to the 8 optimal predictors derived from the training set and used to train the SVM. (C) Projection of add-on adjusted test set 01 samples into the PCA plot spanned by training set 01. (D–I) Aggregated prediction results over 19 iterations for the leave-one-batch-out approach versus a conventional IVIM assay. (D) AUC-ROC for all vector genera. (E) AUC-ROC for strongly transforming LTR.RV.SFFV vectors. (F) AUC-ROC for non-LTR.RV.SFFV vectors. (G) AUC-PRC for all vector genera. (H) AUC-PRC for strongly transforming LTR.RV.SFFV vectors. (I) AUC-PRC for non-LTR.RV.SFFV vectors.

Table 1. Top 20 predictors most often selected in the leave-one-batch-out approach and random test set approach used during the development phase

Leave-one-batch-out approach				Random test set approach		
Probe ID	Gene symbol	Times selected (of 19)	Variable importance (AUC)	Probe ID	Gene symbol	Times selected (of 10)
A_51_P289392	Naip1*	16	94.49	A_55_P2077048	Itih5*	10
A_55_P2077048	Itih5*	15	98.06	A_55_P2024155	Zbtb16*	8
A_51_P106059	Traf4*	14	94.65	A_51_P289392	Naip1*	7
A_66_P135106	Slco3a1*	12	94.18	A_66_P135106	Slco3a1*	7
A_55_P1987984	Zfpm1	12	93.64	A_55_P2018929	Spns2*	7
A_55_P2018929	Spns2*	11	96.98	A_66_P122559	Myct1	5
A_55_P2024155	Zbtb16*	9	100.00	A_51_P334942	Aldh1a1	4
A_51_P486121	Aff3*	8	96.59	A_55_P2108248	Art4*	4
A_55_P2057587	Arx	8	95.15	A_55_P1987984	Zfpm1	4
A_55_P1976882	4930519L02Rik	7	95.00	A_55_P2136426	Prss57	3
A_55_P2108248	Art4*	7	95.62	A_52_P6828	Xk	3
A_52_P56682	Sla2*	6	95.77	A_55_P1976882	4930519L02Rik	2
A_51_P177171	Tie1*	6	93.49	A_55_P2472735	A530032D15Rik	2
A_51_P334942	Aldh1a1	5	90.31	A_51_P486121	Aff3*	2
A_52_P73475	Fam78a	5	93.76	A_55_P2057587	Arx	2
A_52_P162957	Frat2*	5	93.52	A_52_P73475	Fam78a	2
A_52_P68221	Gria3	5	92.75	A_52_P162957	Frat2*	2
A_51_P115626	Shank3	5	94.49	A_52_P68221	Gria3	2
A_55_P2146034	Abca4	4	91.82	A_52_P663904	Lhfpl1	2

Tabulated are the number of times a predictor was included in the list of optimal predictors for a given training set and the global variable importance (AUC-ROC) of each individual predictor computed on the complete dataset of 152 SAGA samples. Indicated with asterisk (*) are 11 optimal predictors of the final SAGA classifier. Complete lists can be found in Table S4, tab 3; Table S5, tab 3; and Table S6, tab 4.

differentiation normally supported by the cytokine conditions used for the IVIM assay and SAGA. The most important predictors to detect genotoxic vectors were linked to stemness, differentiation arrest, and non-myeloid cell fates reflecting early steps of leukemogenesis that precede full cellular transformation and leukemia.[35]

SAGA-GSEA

A critical step in the SAGA-SVM procedure is correct estimation and correction of batch effects to project the new samples into a common gene expression space together with the training samples. This can be error prone for assays with few samples, a profoundly skewed class distribution, or particularly severe batch effects. For these cases, we sought to implement a more robust classifier that can be used within each individual SAGA independently (Figure 6A). We first examined whether the predictors found by our feature selection approach could be used in GSEA to discriminate genotoxic from safe vector designs. Indeed, we observed strong enrichment of the 11 optimal predictors from the final SAGA classifier in transforming vectors compared with mock controls (Figure 6B), whereas this signature was coordinately downregulated in safe vector designs compared with mock samples (Figure 6C). To estimate the predictive performance of this approach, we performed SAGA-GSEA within each of the leave-one-batch-out iterations by using the optimal predictors found for each of the training sets by our feature selection routine as a gene set for GSEA. We then examined the enrichment of the optimal predictor gene sets by performing GSEA for each sample against the mock controls in the left-out batches, yielding an AUC-ROC of 0.91 over all iterations (Figure 6D). To determine the optimal normalized enrichment score (NES) cutoff, we performed a ROC analysis, yielding an NES of greater than 1.7 as the ideal cutoff point for the complete dataset (Figure 6D). However, we found that this was confounded by inclusion of many samples of the strongly genotoxic LTR.RV.SFFV vector, which served as positive control and displayed very strong enrichment of the predictor gene sets. Therefore, we determined the cutoff again on the dataset without LTR.RV.SFFV samples, for which a NES of greater than 1.3 was the optimal threshold (Figure 6D). Using this NES cutoff, SAGA-GSEA outperformed the IVIM assay, albeit with a lower specificity than with the SVM-based SAGA classifier (AUROC$_{\mathrm{GSEA}}$, 0.91; AUROC$_{\mathrm{IVIM}}$, 0.827; P_{DELONG}, 0.005; accuracy$_{\mathrm{GSEA}}$, 84.8%; accuracy$_{\mathrm{IVIM}}$, 76.9%; AUPRC$_{\mathrm{GSEA}}$, 0.91; AUPRC$_{\mathrm{IVIM}}$, 0.89; Table S7, tab 1; Figures 6E and 6F). Similar to SAGA-SVM, we used the 11 final predictors (Table S6, tab 4) that were derived from the complete dataset as a GSEA gene set for the final SAGA-GSEA classifier. The optimal NES cutoff for this 11-predictor gene set was determined by ROC analysis on the complete dataset after exclusion of the LTR.RV.SFFV

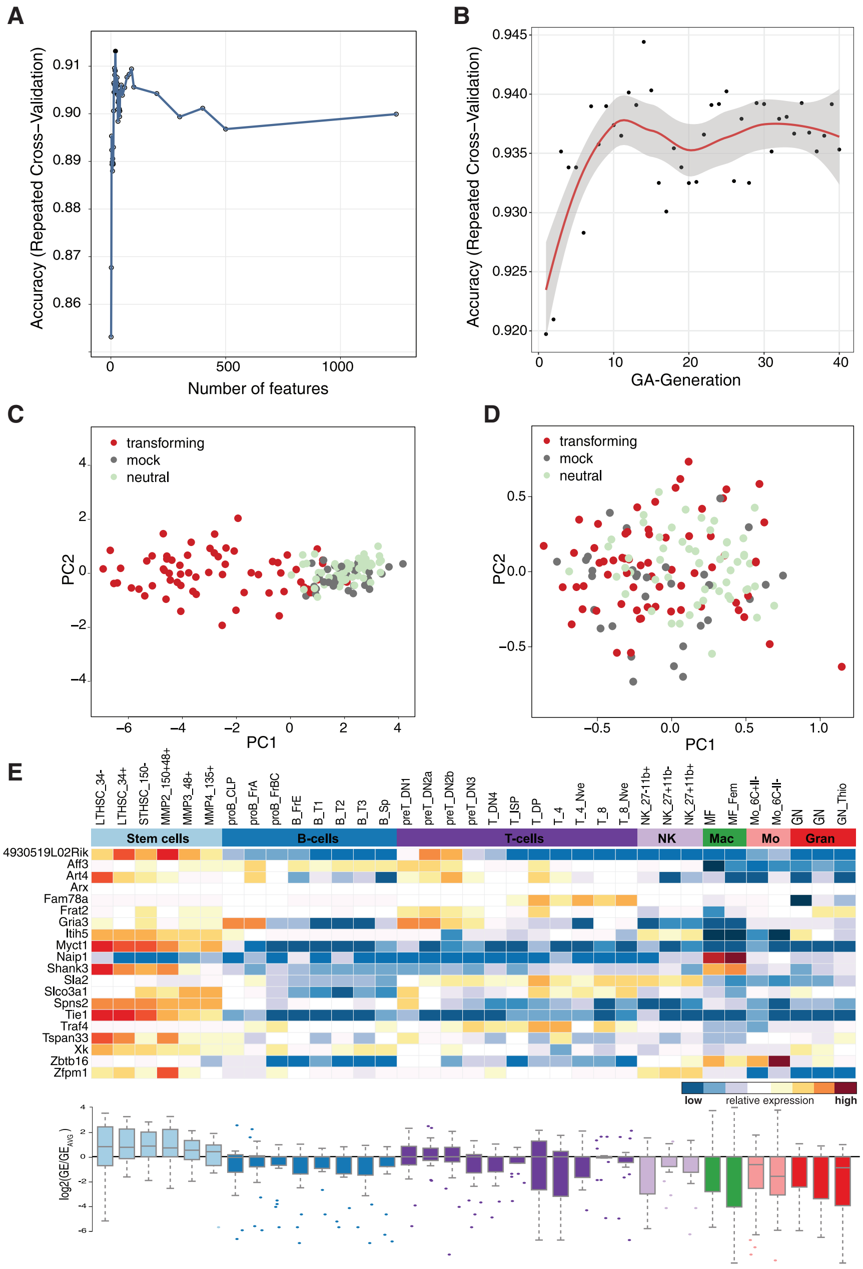

Figure 5. Construction of the final SAGA classifier

(A) Performance profile of the SVM-RFE procedure for the complete set of 152 samples. The filled circle represents the predictor subset with the highest performance comprised of the 20 most important predictors. (B) Performance profile of the SVM-GA procedure for the complete set of 152 samples over 40 generations of the GA. (C) Principal-component analysis (PCA) of 152 samples with known IVIM activity on the 11 optimal probes found by SVM-GA. (D) PCA of 152 samples with known IVIM activity on 11 randomly selected probes of 36,226 annotated probes. (E) Heatmap representing expression of the 20 genes with the highest predictive power from SVM-RFE across murine hematopoiesis.[34] The boxplot below the heatmap represents the expression of genes in each column relative to the expression of all genes. LT-HSC, long-term HSC; ST-HSC, short-term HSC; MPP, multipotent progenitor; Mac/MF, macrophage; Mo, monocyte; Gran/GN, granulocyte.

samples (NES > 1.0; Figure 6G). LTR-driven gammaretroviral, SIN lentiviral, or alpharetroviral vectors with strong promoters and transforming properties in the IVIM assay showed a mean NES between 1.22 and 2.15 (Figure 6H), whereas potentially safer vector architectures with weaker internal promoters did not or only rarely showed this enrichment (Figure 6H; Table S7, tab 2). SAGA-GSEA presents an alternative classifier that circumvents the caveats of cross-batch prediction when correct add-on adjustment is difficult to achieve but critically depends on the integrity of the mock samples.

DISCUSSION

One important bottleneck for gene therapy is the necessity to assess potential safety risks. Since its inception,[11], the IVIM assay has become the *de facto* gold standard *in vitro* assay for risk assessment of gene therapy vectors, and multiple groups have used the IVIM assay to test their vector constructs.[12–15] Here we show that the IVIM assay uncovers the genotoxic potential of vectors that caused SAEs in clinical trials for CGD[3] (LTR.RV.SFFV), X-SCID[2] (LTR.RV.MLV), and WASP[4] (LTR.RV.MPSV). However, transformation potential is affected by vector design, integration sites, vector copy number, the transgene itself, and the disease background. In some contexts, such as ADA-SCID, even LTR-driven vector designs seemed to have an acceptable safety profile,[36] although one individual treated with Strimvelis developed T cell leukemia linked to an insertional event.[37] Consequently, most gene therapy trials now use SIN vectors. IVIM assays for the mutagenic vector design SIN.LV.SFFV revealed the genotoxic risk in only 40% of assays. Hence, even though the IVIM has an excellent specificity because of its low sensitivity, it has to be repeated multiple times to produce an informative and reliable result. We developed SAGA as a robust, standardized pipeline that efficiently identifies genotoxic vectors with higher accuracy by coupling a shortened IVIM assay with a molecular readout. By performing gene expression profiling on murine hematopoietic progenitors transduced with vectors with known IVIM properties, we show that only genotoxic vectors upregulate a specific gene expression signature that is reminiscent of immature HSC transcriptional programs, myeloid differentiation, and early

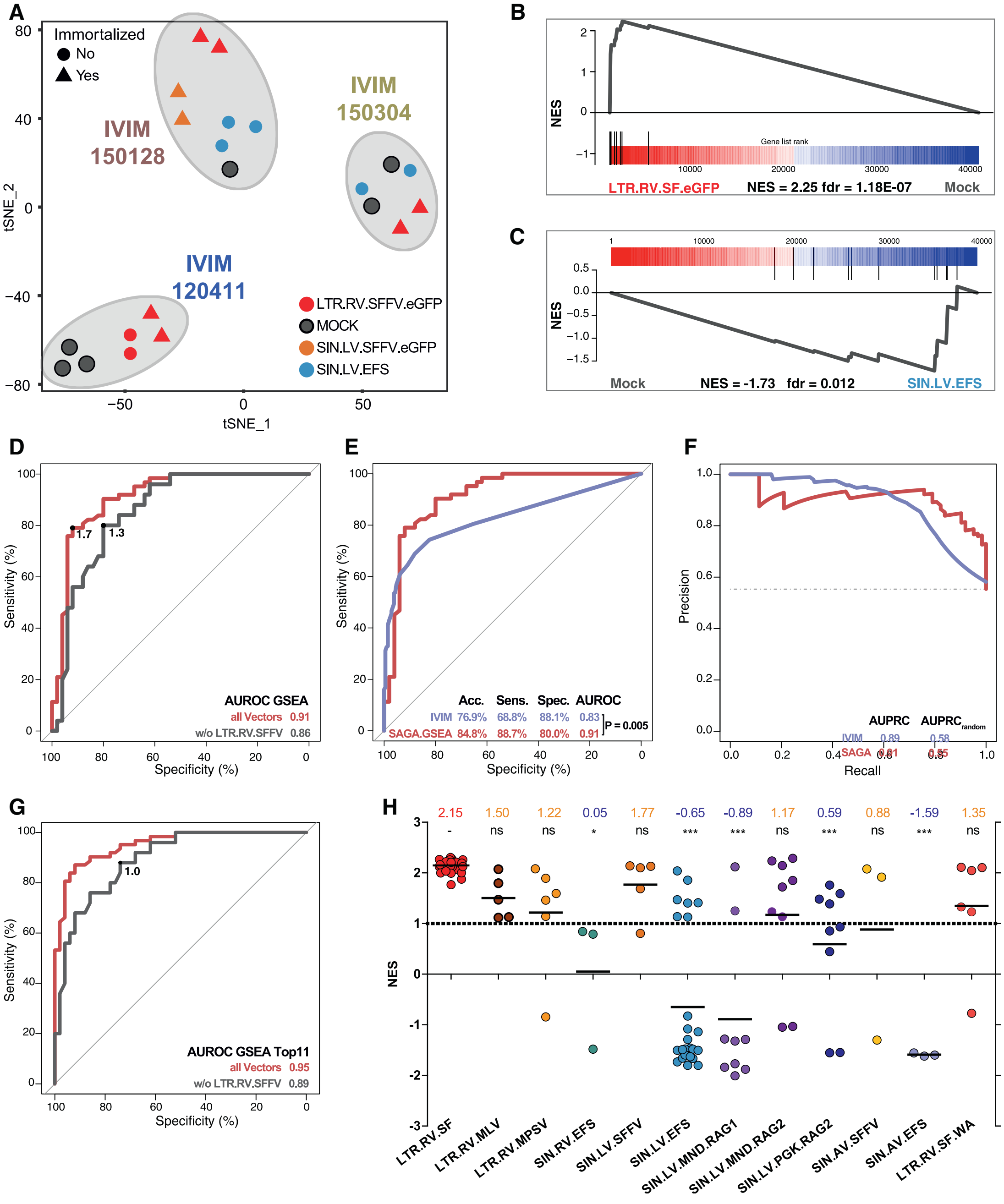

Figure 6. SAGA-GSEA

(A) t-SNE representation of gene expression data from three independent SAGAs without batch correction. (B) GSEA plot for the 11 optimal predictors from the final classifier for LTR.SFFV.EGFP (sample X4991) versus mock from IVIM 3 (shown in A). (C) GSEA plot for the 11 optimal predictors for SIN.LV.EFS (sample X4997) versus mock from IVIM 3. (D) AUC-ROC aggregated from the leave-one-batch-out approach for all vector genera (red) and without strongly transforming LTR.RV.SFFV vectors (gray). The points on the curve indicate the best NES cutoff. (E) AUC-ROC for all vector genera (same curve as in D) versus AUC-ROC of the IVIM assay. (F) AUC-PRC aggregated from the leave-one-batch-out approach for all vector genera versus IVIM. (G) AUC-ROC using the 11 optimal predictors from the final classifier on all IVIM batches for all vector genera (red) and without strongly transforming LTR.RV.SFFV vectors (gray). The point on the curve indicates the best NES cutoff. (H) SAGA-GSEA results for all tested vectors. Plotted are the NESs of the 11-probe gene set from the final classifier over the different vector genera. The dashed line denotes NES ≥ 1.0, indicating evidence of genotoxicity as determined from the ROC analysis (Figure 6G) for genotoxic vectors when the strongly transforming LTR.SFFV samples were disregarded. Above the graph, mean NES values are shown for each vector type. The level of evidence whether the NES is significantly different from the positive control is indicated (ns = not significant, *$p < 0.05$, ***$p < 0.001$; p values were calculated using a Kruskal-Wallis test with Dunn's post hoc test).

transformation. One challenge was to identify a set of optimal predictors from the highly dimensional predictor space that allow precise classification to keep computational costs low and make the model more interpretable and as a potential starting point for possible later transfer of SAGA to simpler technical platforms. However, even with efficient feature reduction using RFE, the predictor space to be explored to find the best combination of features retained after SVM-RFE is vast. For instance, finding the best combination of 10–15 of 30 retained predictors would require building and testing over 500 million models. Instead of a complete search of the predictor space, we show here that Darwinian natural selection embedded in a GA can be used efficiently for a guided search of the predictor space. Thus, harnessing principles of population biology for complex optimization tasks is a powerful approach, as shown before for optimization of a gene expression-based classifier using particle swarm optimization.[38] However, because of the nature of the GA, the solution represents a local optimum for each training set, and it cannot be excluded that better solutions may exist. Initializing the GA with different random seeds yielded slightly different lists of optimal predictors. However, the solutions mostly differed by only one or two predictors, indicating that the feature selection procedure was stable and the GA found a near-optimal solution. Importantly, as more samples are added to the dataset, improved solutions will be found, and SAGA will continue to evolve.

We tested a SIN lentiviral vector for RAG1 and RAG2 deficiency, in which transgene expression is controlled by a strong MND or a weaker PGK promoter.[39] The PGK promoter did not display genotoxic potential in IVIM or SAGA. Conversely, the MND promoter constructs differed in their risk profile; vector SIN.LV.MND.RAG1 was determined to be safe by IVIM (0 of 9 positive assays) and SAGA (2 of 9 samples with an NES $\geq$ 1.0). In contrast, SIN.LV.MND.RAG2 showed a replating phenotype in 2 of 9 IVIM assays and a core set enrichment in 7 of 9 SAGA tests. This underscores the higher sensitivity and predictive potential of SAGA compared with IVIM. Based on these data, the preferred vector for potential treatment of RAG2-SCID is PGK-RAG2 rather than MND-RAG2. Similarly, when we tested an integrase mutant (LTR.RV.SFFV.W390A) of the strongly transforming LTR.RV.SFFV vector with an altered and potentially safer integration profile,[40],[41] SAGA detected a decrease in the mean NES to 1.35 (compared with 2.15 of the wild-type integrase vector). Thus, by providing a continuous score rather than a digital outcome, SAGA provides a higher resolution of genotoxic risk.

Recently, Zhou et al.[42] observed that murine thymocytes transduced with mutagenic vectors show developmental arrest during T-lymphocyte development. The arrested progenitors overexpressed *Lmo2*, *Mef2c*, and *Prdm16*. The transcription factor LMO2 was the most clinically relevant dysregulated proto-oncogene in vector-associated transformation in clinical trials for X-SCID and WASP. Importantly, *Lmo2* and *Mef2c* upregulation was detected in SAGA samples transduced with mutagenic vectors ($\log_2$FC *Lmo2*, 0.67; $p = 5 \times 10^{-24}$; $\log_2$FC *Mef2c*, 0.79; $p = 9 \times 10^{-15}$; moderated t test with BH adjustment; Table S2, tab 1). Hence, SAGA detects perturbation of proto-oncogenes of the lymphoid lineage, something that is beyond the capacity of the conventional IVIM assay. In addition, more work is needed to determine whether SAGA can detect genotoxic potential caused by transformation mechanisms other than *cis* activation of proto-oncogenes, such as aberrant splicing,[43],[44] and whether the SAGA principle can be transferred to non-hematopoietic target tissues. In the future, single-cell RNA-seq experiments might help to further fine-tune the SAGA signature. Our work provides insights into the early molecular events of genotoxicity following transduction of hematopoietic cells with integrating vectors and presents a powerful machine-learning approach to prospectively estimate the mutagenic potential of integrating vector systems for gene therapy.

MATERIALS AND METHODS

Study design

The study aimed to develop a gene expression-based diagnostic classifier to distinguish potentially genotoxic gene therapy vectors from safe vector designs. For SAGA, murine Lin$^-$ HSPCs were transduced with vectors of interest (Table S1) at high MOIs (target, >3 vector copies per cell) and expanded in myeloid growth-promoting medium (Figure 1A). On day 15, RNA was extracted for microarray analysis on Agilent Whole Mouse Genome 4x44K v.2 microarrays. The data were analyzed using R 3.5.1 and Bioconductor 3.7. All available microarrays (n = 169) were read in, quantile normalized, and batch corrected. 152 SAGA samples with known behavior in the IVIM assay (65 transforming, 55 safe, and 32 mock samples; Supplemental materials and methods; Table S8, tabs 1 and 2) were analyzed for differential expression (Table S2), GSEA (Table S3), and development of the SAGA classifier (Table S4). During classifier development, the jointly preprocessed gene expression matrix of all 152 SAGA samples was split into 10 different training (70% of samples) and test sets (30%) using stratified resampling to ensure comparable class distributions in test and training sets (Figure S3). Development of models was performed on the training sets only using cross-validation to assess model performance during feature selection and nested cross-validation for hyperparameter tuning. The test sets were not used at any point for feature selection or model tuning. Three different feature selection routines were applied to the training data to reduce the number of predictors as far as possible. First, an unsupervised filter was applied to exclude probes showing little variation in the dataset, followed by RFE, which iteratively removes the least important predictors before applying a GA to find a near-optimal combination of predictors retained by the preceding steps. After feature selection, an SVM was trained on the training data reduced to the optimal predictors and used to predict the test sets. For estimation of classifier performance, a jack-knife, leave-one-batch-out procedure was employed by leaving one batch of SAGA assays completely out of the model building process. The complete feature selection and model training pipeline was applied to the remaining batches. After the optimal predictors had been derived from the training set and the classifier had been trained and fixed, the batch that served as the independent test set was add-on normalized, add-on batch corrected, and predicted. The procedure was repeated for each of the 19 experimental SAGA batches, and the predictions results on the left-out batches were aggregated to determine performance. The final SAGA classifier was built on the entire set of available samples (n = 152) for use as the training set in the SAGA R package. For SAGA-GSEA, we used the optimal predictors found by the feature selection routines on each training set for each iteration as a gene set for GSEA. We then examined the enrichment of these optimal predictor gene sets by performing GSEA for each test set sample against the mock controls within the left-out test sets.

Cell culture for IVIM and SAGA

Lin^- cells were isolated from tibiae, femora, and iliac crests of 8- to 12-week-old female C57BL/6J animals (Janvier) using the mouse lineage cell depletion kit (Miltenyi Biotec). Cells were frozen in aliquots of 5×10^5 cells in 90% fetal bovine serum (FBS) (PAA Laboratories) and 10% DMSO (Merck). After thawing, one aliquot per assay was cultured for 48 h in StemSpan (STEMCELL Technologies) supplemented with 50 ng/mL rm-SCF, 100 ng/mL rh-Flt-3L, 100 ng/mL rh-interleukin-11 (IL-11), and 20 ng/mL rm-IL-3 (PeproTech). For transduction, 250 μL of viral supernatant (or medium for the mock controls) was preloaded on 24-well suspension plates coated with RetroNectin (TaKaRa) to reach a defined MOI. Following the preloading, 1×10^5 cells were added to the wells in a total volume of 250 μL and incubated overnight. The preloading procedure was repeated for the second round of transduction. For this, suspension cells from the first transduction round were harvested, and cells still bound to RetroNectin were incubated with cell dissociation buffer (Gibco), pelleted, and resuspended in 250 μL of fresh culture medium before being added to the suspension harvest. Subsequently, 750 μL of the cell suspension was added to the wells preloaded for the second transduction. Cells were incubated for 24 h; harvested as described; mixed with 1.6 mL IMDM (Biochrom) containing 10% FBS, 1% penicillin/streptomycin (PAN Biotech), 2 mM glutamine (Biochrom). and cytokines as described above; and seeded onto 12-well suspension plates. On day 4 after (the second round of) transduction, we isolated DNA and/or RNA from 10% and used 2.5% of the cell material for flow cytometry analysis of transgene expression. After feeding the cells with 1.9 mL IMDM containing supplements ($IMDM^+$), cells were incubated for 48 h on 6-well suspension plates before adding another 2.2 mL of medium. On day 8 post transduction (p.t.), samples were diluted (~1:10) by seeding 1×10^6 cells in 4 mL of $IMDM^+$ in 6-well suspension plates. On days 11 and 13 p.t., cells were given 1.2 mL of $IMDM^+$. For the IVIM replating step, cells were re-seeded on day 15 p.t. at 100 cells/well in 96-well flat-bottom suspension plates. Following 14 days of incubation, plates were screened microscopically for growth of insertional mutants. Afterward, 20 μL of 0.25% thiazolyl blue tetrazolium bromide (Sigma) in DPBS (Pan Biotech) was added to the wells and incubated for 2–3 h at 37°C. Cells were lysed by addition of 100 μL of 20% SDS (Sigma). Plates were set on a shaker overnight at room temperature before absorption was measured at 540 nm with a SpectraMax 340PC (Molecular Devices). After background subtraction, the highest absorption value from the mock plate was used as a threshold to determine positive wells unless the value was higher than the mean absorption value of immortalized wells from a meta-analysis of 22 assays (5.61 times the expression value of a microscopically negative well). In this case, the second-highest mock value was used as a cutoff. Differences in the incidence of positive and negative assays relative to mock or LTR.RV.SFFV. EGFP-transduced cells were analyzed by Fisher's exact test with the Benjamini-Hochberg multiple testing correction procedure.

SAGA sample processing and bioinformatics

A complete list of SAGA samples with a detailed description of RNA isolation, microarray acquisition, and all bioinformatic procedures, including *de novo* annotation of microarrays, raw data preprocessing, t-SNE, and principal-component analysis (PCA) visualizations, differential expression analysis, GSEA, description of the classifier development and performance measurements, the SAGA R package, qRT-PCR, and RNA-seq are outlined in the Supplemental materials and methods.

Statistical analysis

The incidence of positive and negative IVIM assays in Figure 1C was analyzed by a Fisher's exact test with Benjamini-Hochberg correction. The SAGA-GSEA results in Figure 6H were compared using a Kruskal-Wallis test with Dunn's post hoc test. For a detailed description of the different R package versions and statistical tests used in each bioinformatic step of SAGA, refer to the respective Supplemental material and methods section.

Data and materials availability

All data associated with this paper can be found in the main text or the Supplemental materials. Raw and processed expression data from all experiments have been deposited in the Gene Expression Omnibus under GEO: GSE109391. The R code for the SAGA genotoxicity prediction package and all other computations is available as source code and compiled R package via https://github.com/rothemi/SAGA. For convenience, an Amazon Machine Image running R3.6 and SAGA is available, including test samples from two different SAGA assays.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ymthe.2021.06.017.

ACKNOWLEDGMENTS

We are grateful to Prof. Christopher Baum (Charité, Berlin) for scientific support and valuable discussions. We thank Anne Galy and Fulvio Mavilio (Genethon, France) for providing vector LV.WASP.WASP. We are grateful to Prof. Christoph Klein (LMU, Munich) for providing the CMMP.WASP plasmid. We thank David A. Williams and Christian Brendel (Boston Children's Hospital) for vector supernatant SIN.LV.LCR.D12G5. We thank Chris Mason (AVROBIO) for providing the lentiviral supernatants LV2AGA, LV2GBA and pCCL.CTNS. This work was supported by the German Research Foundation (RO 5102/1-1, cluster of excellence REBIRTH [Exc 62/2], and SFB738), the Federal State of Lower Saxony (research project R2N), and the CRACK-IT initiative from the National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs; NC/C015102/1). Results incorporated in this study received funding from the European Union Horizon 2020 Research and Innovation Program (755170) and the FP7 Project CELL-PID (261387). A. Schwarzer is a fellow of the Clinician-Scientist Program "Young Academy" of Hannover Medical School and received funding from the Else Kröner-Fresenius-Stiftung (2017_A74) and Joachim Herz Stiftung. A.J.T. received funding from The Wellcome Trust (090233/Z/09/Z).

AUTHOR CONTRIBUTIONS

M.R., V.D., B.W., A.L.B., and O.D.-B. performed the experiments. A. Schwarzer, S.R.T., and M.R. designed and analyzed experiments, performed bioinformatic analyses, and wrote the manuscript. S.R.T. programmed the SAGA R package. A.J.T., R.G., and H.B.G. provided materials and revised the manuscript. A. Selich, J.W.S., F.J.T.S., U.M., M.M., and T.C.H. discussed data, helped to adjust experimental designs, and revised the manuscript. A. Schwarzer supervised the study and revised the manuscript.

DECLARATION OF INTERESTS

A patent application has been filed under registration number EP3394286A1 (Analytical process for genotoxicity assessment).

REFERENCES

1. Morgan, R.A., Gray, D., Lomova, A., and Kohn, D.B. (2017). Hematopoietic Stem Cell Gene Therapy: Progress and Lessons Learned. Cell Stem Cell *21*, 574–590.

2. Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., Radford, I., Villeval, J.-L., Fraser, C.C., Cavazzana-Calvo, M., and Fischer, A. (2003). A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. N. Engl. J. Med. *348*, 255–256.

3. Stein, S., Ott, M.G., Schultze-Strasser, S., Jauch, A., Burwinkel, B., Kinner, A., Schmidt, M., Krämer, A., Schwäble, J., Glimm, H., et al. (2010). Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. Nat. Med. *16*, 198–204.

4. Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., Modlich, U., Beier, R., Gohring, G., Steinemann, D., et al. (2014). Gene Therapy for Wiskott-Aldrich Syndrome–Long-Term Efficacy and Genotoxicity. Sci. Transl. Med. *6*, 227ra33.

5. Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150.

6. Schambach, A., Zychlinski, D., Ehrnstroem, B., and Baum, C. (2013). Biosafety features of lentiviral vectors. Hum. Gene Ther. *24*, 132–142.

7. Baum, C., Schambach, A., Bohne, J., and Galla, M. (2006). Retrovirus vectors: toward the plentivirus? Mol. Ther. *13*, 1050–1063.

8. Rothe, M., Schambach, A., and Biasco, L. (2014). Safety of gene therapy: new insights to a puzzling case. Curr. Gene Ther. *14*, 429–436.

9. Cesana, D., Ranzani, M., Volpin, M., Bartholomae, C., Duros, C., Artus, A., Merella, S., Benedicenti, F., Sergi Sergi, L., Sanvito, F., et al. (2014). Uncovering and dissecting the genotoxicity of self-inactivating lentiviral vectors in vivo. Mol. Ther. *22*, 774–785.

10. Gonin, P., Buchholz, C.J., Pallardy, M., and Mezzina, M. (2005). Gene therapy biosafety: scientific and regulatory issues. Gene Ther. *12* (*Suppl 1*), S146–S152.

11. Modlich, U., Bohne, J., Schmidt, M., von Kalle, C., Knöss, S., Schambach, A., and Baum, C. (2006). Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. Blood *108*, 2545–2553.

12. Punwani, D., Kawahara, M., Yu, J., Sanford, U., Roy, S., Patel, K., Carbonaro, D.A., Karlen, A.D., Khan, S., Cornetta, K., et al. (2017). Lentivirus Mediated Correction of Artemis-Deficient Severe Combined Immunodeficiency. Hum. Gene Ther. *28*, 112–124.

13. Huang, J., Khan, A., Au, B.C., Barber, D.L., López-Vásquez, L., Prokopishyn, N.L., Boutin, M., Rothe, M., Rip, J.W., Abaoui, M., et al. (2017). Lentivector Iterations and Pre-Clinical Scale-Up/Toxicity Testing: Targeting Mobilized $CD34^+$ Cells for Correction of Fabry Disease. Mol. Ther. Methods Clin. Dev. *5*, 241–258.

14. Wolstein, O., Boyd, M., Millington, M., Impey, H., Boyer, J., Howe, A., Delebecque, F., Cornetta, K., Rothe, M., Baum, C., et al. (2014). Preclinical safety and efficacy of an anti-HIV-1 lentiviral vector containing a short hairpin RNA to CCR5 and the C46 fusion inhibitor. Mol. Ther. Methods Clin. Dev. *1*, 11–14.

15. Negre, O., Bartholomae, C., Beuzard, Y., Cavazzana, M., Christiansen, L., Courne, C., Deichmann, A., Denaro, M., de Dreuzy, E., Finer, M., et al. (2015). Preclinical evaluation of efficacy and safety of an improved lentiviral vector for the treatment of β-thalassemia and sickle cell disease. Curr. Gene Ther. *15*, 64–81.

16. Krivtsov, A.V., Twomey, D., Feng, Z., Stubbs, M.C., Wang, Y., Faber, J., Levine, J.E., Wang, J., Hahn, W.C., Gilliland, D.G., et al. (2006). Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. Nature *442*, 818–822.

17. Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J., et al. (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. Nat. Med. *17*, 1086–1093.

18. Ng, S.W.K., Mitchell, A., Kennedy, J.A., Chen, W.C., McLeod, J., Ibrahimova, N., Arruda, A., Popescu, A., Gupta, V., Schimmer, A.D., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature *540*, 433–437.

19. Boztug, K., Schmidt, M., Schwarzer, A., Banerjee, P.P., Díez, I.A., Dewey, R.A., Böhm, M., Nowrouzi, A., Ball, C.R., Glimm, H., et al. (2010). Stem-cell gene therapy for the Wiskott-Aldrich syndrome. N. Engl. J. Med. *363*, 1918–1927.

20. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

21. Kustikova, O.S., Schwarzer, A., Stahlhut, M., Brugman, M.H., Neumann, T., Yang, M., Li, Z., Schambach, A., Heinz, N., Gerdes, S., et al. (2013). Activation of Evi1 inhibits cell cycle progression and differentiation of hematopoietic progenitor cells. Leukemia *27*, 1127–1138.

22. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

23. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

24. Goh, W.W.B., Wang, W., and Wong, L. (2017). Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol. *35*, 498–507.

25. Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315.

26. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

27. Hornung, R., Causeur, D., Bernau, C., and Boulesteix, A.-L. (2017). Improving cross-study prediction through addon batch effect adjustment or addon normalization. Bioinformatics *33*, 397–404.

28. Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling (Springer), p. 67.

29. Kuhn, M., and Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models, First Edition (Chapman and Hall/CRC).

30. Guyon, I., Weston, J., and Barnhill, S. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn. *46*, 389–422.

31. Chatterjee, S., Laudato, M., and Lynch, L.A. (1996). Genetic algorithms and their statistical applications: an introduction. Comput. Stat. Data Anal. *22*, 633–651.

32. Castaldi, P.J., Dahabreh, I.J., and Ioannidis, J.P.A. (2011). An empirical assessment of validation practices for molecular classifiers. Brief. Bioinform. *12*, 189–202.

33. Hornung, R., Boulesteix, A.L., and Causeur, D. (2016). Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. BMC Bioinformatics *17*, 27.

34. Yoshida, H., Lareau, C.A., Ramirez, R.N., Rose, S.A., Maier, B., Wroblewska, A., Desland, F., Chudnovskiy, A., Mortha, A., Dominguez, C., et al.; Immunological Genome Project (2019). The cis-Regulatory Atlas of the Mouse Immune System. Cell *176*, 897–912.e20.

35. Stavropoulou, V., Kaspar, S., Brault, L., Sanders, M.A., Juge, S., Morettini, S., Tzankov, A., Iacovino, M., Lau, I.-J., Milne, T.A., et al. (2016). MLL-AF9 Expression in Hematopoietic Stem Cells Drives a Highly Invasive AML Expressing EMT-Related Genes Linked to Poor Outcome. Cancer Cell *30*, 43–58.

36. Cicalese, M.P., Ferrua, F., Castagnaro, L., Pajno, R., Barzaghi, F., Giannelli, S., Dionisio, F., Brigida, I., Bonopane, M., Casiraghi, M., et al. (2016). Update on the safety and efficacy of retroviral gene therapy for immunodeficiency due to adenosine deaminase deficiency. Blood *128*, 45–54.

37. (2020). Strimvelis: risk of lymphoid T-cell leukaemia? React. Wkly. *1830*, 6.

38. Best, M.G., Sol, N., In 't Veld, S.G.J.G., Vancura, A., Muller, M., Niemeijer, A.N., Fejes, A.V., Tjon Kon Fat, L.A., Huis In 't Veld, A.E., Leurs, C., et al. (2017). Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets. Cancer Cell *32*, 238–252.e9.

39. Garcia-Perez, L., van Eggermond, M., van Roon, L., Vloemans, S.A., Cordes, M., Schambach, A., Rothe, M., Berghuis, D., Lagresle-Peyrou, C., Cavazzana, M., et al. (2020). Successful Preclinical Development of Gene Therapy for Recombinase-Activating Gene-1-Deficient SCID. Mol. Ther. Methods Clin. Dev. *17*, 666–682.

40. El Ashkar, S., De Rijck, J., Demeulemeester, J., Vets, S., Madlala, P., Cermakova, K., Debyser, Z., and Gijsbers, R. (2014). BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. Mol. Ther. Nucleic Acids *3*, e179.

41. El Ashkar, S., Van Looveren, D., Schenk, F., Vranckx, L.S., Demeulemeester, J., De Rijck, J., Debyser, Z., Modlich, U., and Gijsbers, R. (2017). Engineering Next-Generation BET-Independent MLV Vectors for Safer Gene Therapy. Mol. Ther. Nucleic Acids *7*, 231–245.

42. Zhou, S., Fatima, S., Ma, Z., Wang, Y.-D., Lu, T., Janke, L.J., Du, Y., and Sorrentino, B.P. (2016). Evaluating the Safety of Retroviral Vectors Based on Insertional Oncogene Activation and Blocked Differentiation in Cultured Thymocytes. Mol. Ther. *24*, 1090–1099.

43. Scholz, S., Fronza, R., Bartholomae, C., Cesana, D., Montini, E., Kalle, C.V., Gil-Farina, I., and Schmidt, M. (2017). Lentiviral Vector Promoter is Decisive for Aberrant Transcript Formation. Hum. Gene Ther. *28*, 875–885.

44. Knight, S., Bokhoven, M., Collins, M., and Takeuchi, Y. (2010). Effect of the internal promoter on insertional gene activation by lentiviral vectors with an intact HIV long terminal repeat. J. Virol. *84*, 4856–4859.

YMTHE, Volume *29*

Supplemental Information

Predicting genotoxicity of viral vectors for stem cell gene therapy using gene expression-based machine learning

Adrian Schwarzer, Steven R. Talbot, Anton Selich, Michael Morgan, Juliane W. Schott, Oliver Dittrich-Breiholz, Antonella L. Bastone, Bettina Weigel, Teng Cheong Ha, Violetta Dziadek, Rik Gijsbers, Adrian J. Thrasher, Frank J.T. Staal, Hubert B. Gaspar, Ute Modlich, Axel Schambach, and Michael Rothe

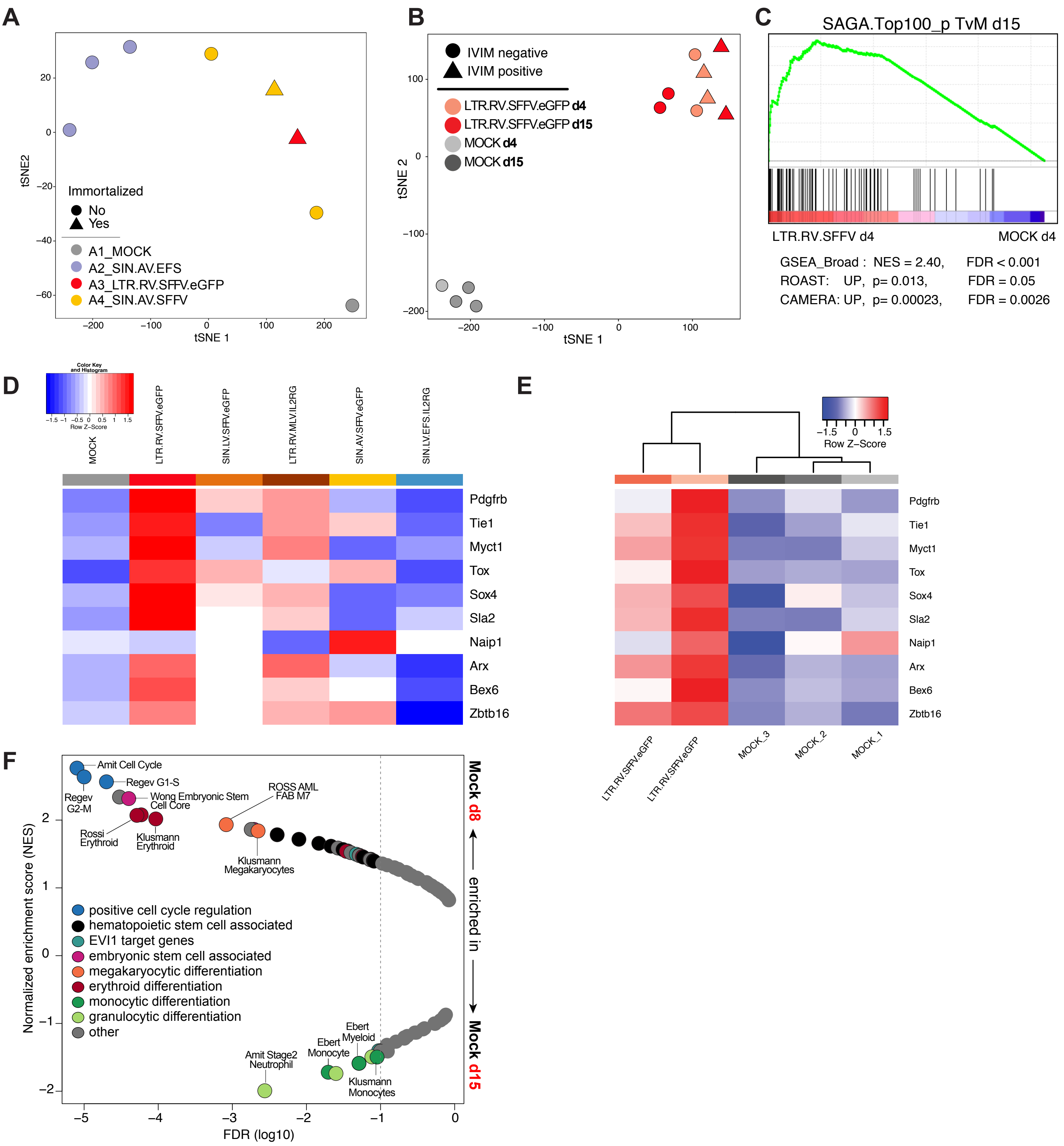

Figure S1. Transforming vectors impose specific gene expression changes in murine hematopoietic progenitors A) t-SNE representation of the gene expression profiles in samples transduced with SIN.AV.EFS, LTR.RV.SFFV and SIN.AV.SFFV. B) t-SNE representation of SAGA assays measured on day 4 (pale colors) and day 15 after batch correction using the assay date as a batch variable. C) GSEA-plot showing the enrichment of the Top 100 genes (upregulated on day 15 in transforming samples vs mock samples, rank based on p-values) in the day 4 samples transduced with LTR.RV.SFFV.eGFP compared to the MOCK control. Below the plot, the statistics for three different GSEA tests (Broad gene_set permutation; ROAST self-contained GSEA and CAMERA competitive GSEA test) are given as discussed in Materials and Methods. D) validation of gene expression changes by qPCR: row-scaled heatmap of qPCR based gene expression genes showing the highest $\log_2$FC by transforming vectors in the first three IVIM assays. e) validation of gene expression changes by RNASeq: row-scaled heatmap of RNA-Seq based gene expression (rlog transformed normalized counts) of the top-upregulated genes by transforming vectors. F) Gene set enrichment analysis of expression changes in 106 hematopoiesis-associated gene sets (Supplementary Table 3) in mock-samples from d8 versus mock-samples from day 15. Plotted are normalized enrichment scores (NES) against the false discovery rate (FDR) obtained by gene set permutation. Significant enrichment (FDR < 0.1) is indicated by the dashed line.

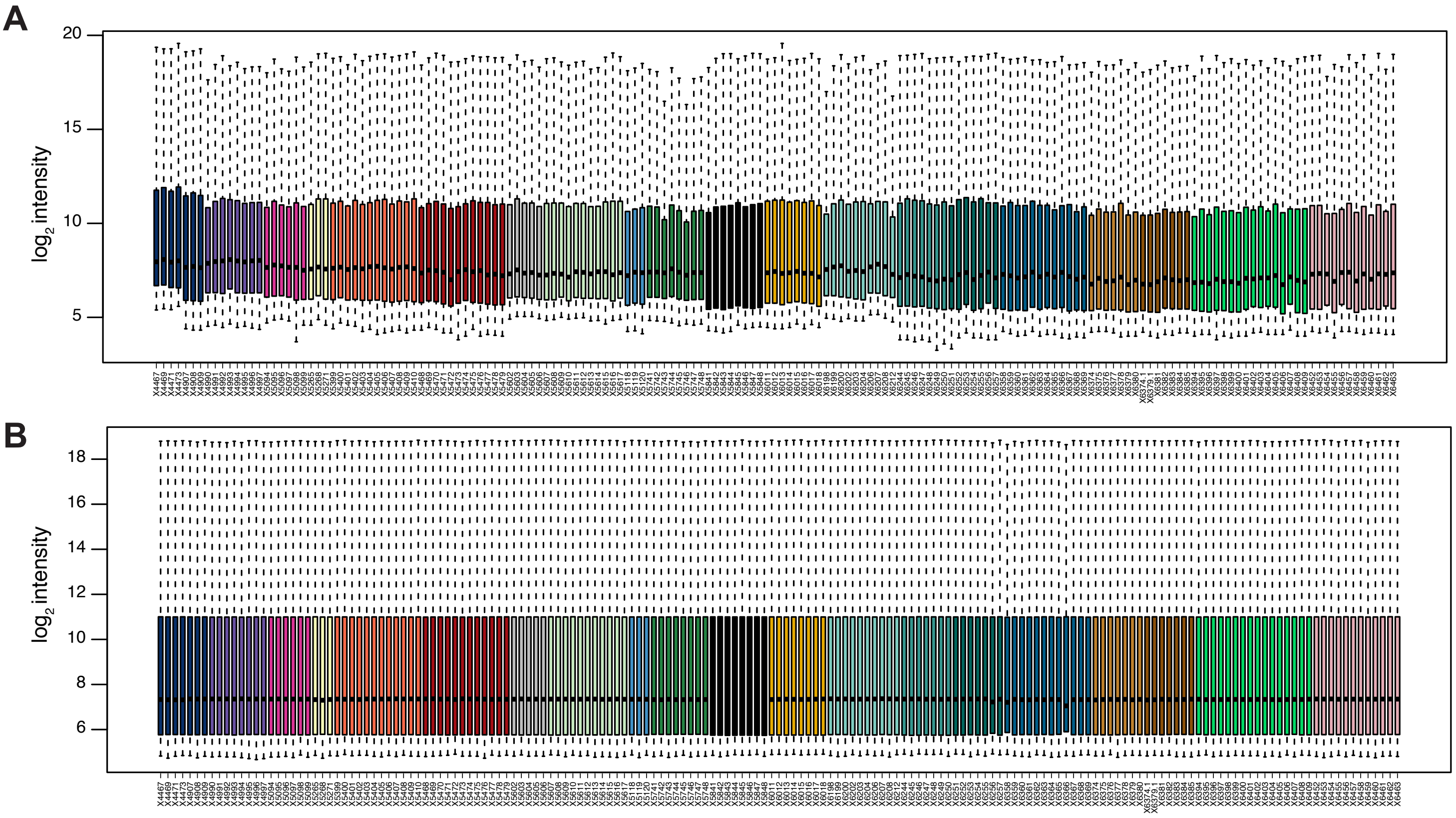

Figure S2 Quantile normalization of individual SAGA assays. A) Boxplot of $\log_2$ raw intensities of 169 SAGA samples (167 individual SAGA samples plus 2 mock duplicates from IVIM ID 180523; see Materials and Methods) hybridized to Agilent Microarrays. The coloring scheme denotes individual assays (batches). **B)** Boxplot of $\log_2$ intensities of 169 SAGA samples after quantile-normalization and averaging of quadruplicate probes

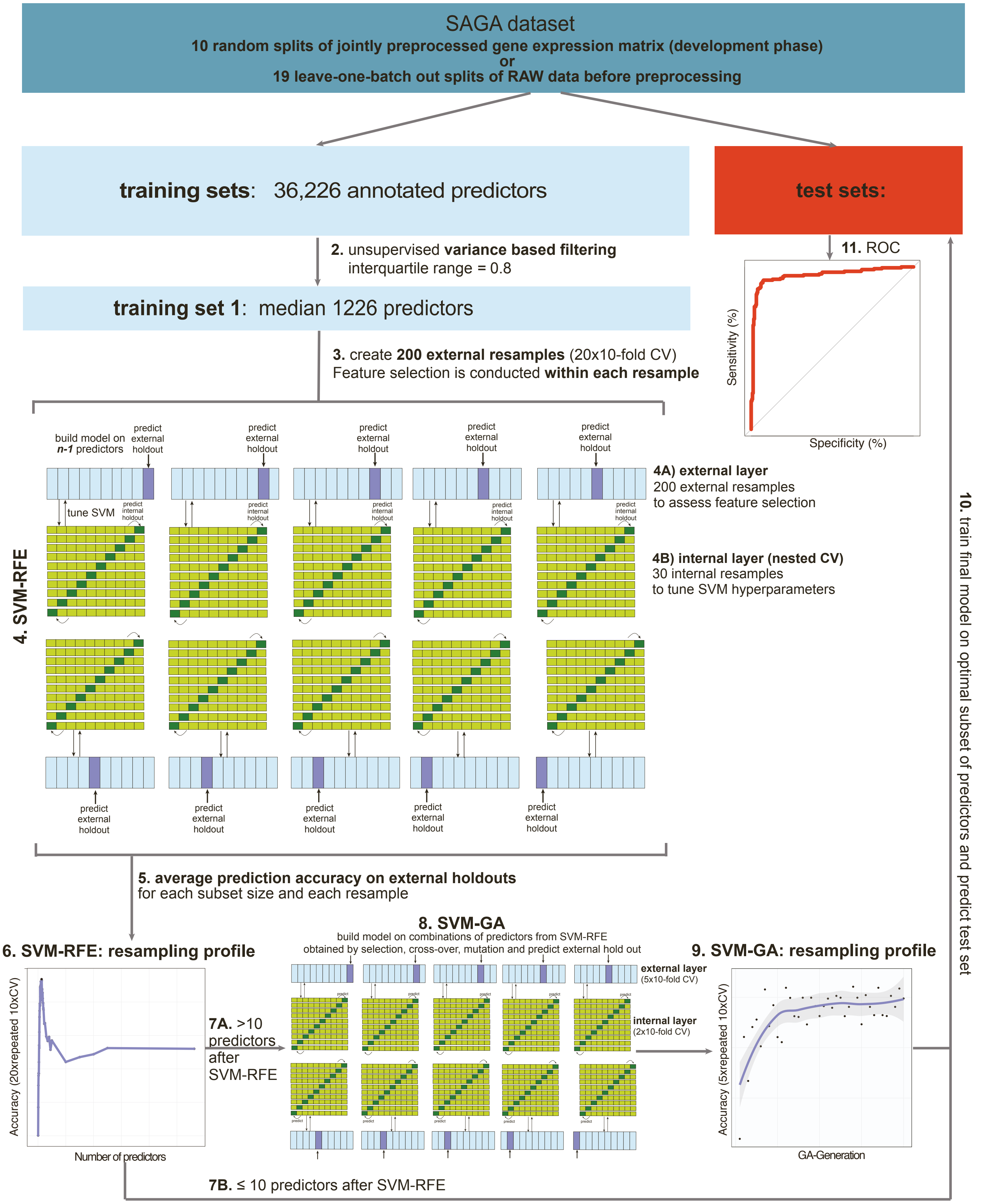
SAGA dataset
10 random splits of jointly preprocessed gene expression matrix (development phase)
or
19 leave-one-batch out splits of RAW data before preprocessing
training sets: 36,226 annotated predictors
test sets:
2. unsupervised variance based filtering
interquartile range = 0.8
training set 1: median 1226 predictors
3. create 200 external resamples (20x10-fold CV)
Feature selection is conducted within each resample
11. ROC
Sensitivity (%)
Specificity (%)
build model on n-1 predictors
predict external holdout
tune SVM
predict internal holdout
4A) external layer
200 external resamples
to assess feature selection
4B) internal layer (nested CV)
30 internal resamples
to tune SVM hyperparameters
4. SVM-RFE
10. train final model on optimal subset of predictors and predict test set
5. average prediction accuracy on external holdouts
for each subset size and each resample
6. SVM-RFE: resampling profile
Accuracy (20xrepeated 10xCV)
Number of predictors
7A. >10 predictors after SVM-RFE
8. SVM-GA
build model on combinations of predictors from SVM-RFE
obtained by selection, cross-over, mutation and predict external hold out
external layer
(5x10-fold CV)
internal layer
(2x10-fold CV)
predict
9. SVM-GA: resampling profile
Accuracy (5xrepeated 10xCV)
GA-Generation
7B. ≤ 10 predictors after SVM-RFE

Figure S3. Training / Test set splitting and resampling during feature selection. 1) The dataset is split into training and test sets 2) An unsupervised filter is applied to the training set to filter out probes with low variance 3) 200 resamples of the training set are generated using 20 times repeated 10-fold CV. Each resample is comprised of 90 % of the training set (97 samples, light blue) for feature selection and a 10% hold-out sample (15 samples, purple) to assess prediction performance ("external layer"). 4A) Feature selection is performed within each resample by training/tuning the model using all n predictors and prediction of the external hold-out sample. Variable importance is calculated via AUC for each predictor and the process is repeated using the n-1 most important predictors 4B) Tuning of hyperparameters is performed at each iteration of feature selection using an "internal layer" that further splits each training sample from the outer loop using 3x10-fold CV. The optimal hyperparameter is passed to the outer loop to build the model. 5) results from the outer loop are aggregated into 6) a performance profile over the tested predictor subsets. 7) If SVM-RFE retains more than 10 predictors 8) a genetic algorithm is employed to find the best combination of retained features using a similar resampling scheme. 9) the resampling accuracy of each generation of the genetic algorithm is recorded and the optimal iteration is selected. 10) The final model is build using the optimal predictors on the complete training set and the test set is predicted. 11) The whole process is repeated for the remaining 9 training / test set splits and the prediction accuracies on the test sets are aggregated into a ROC statistic.

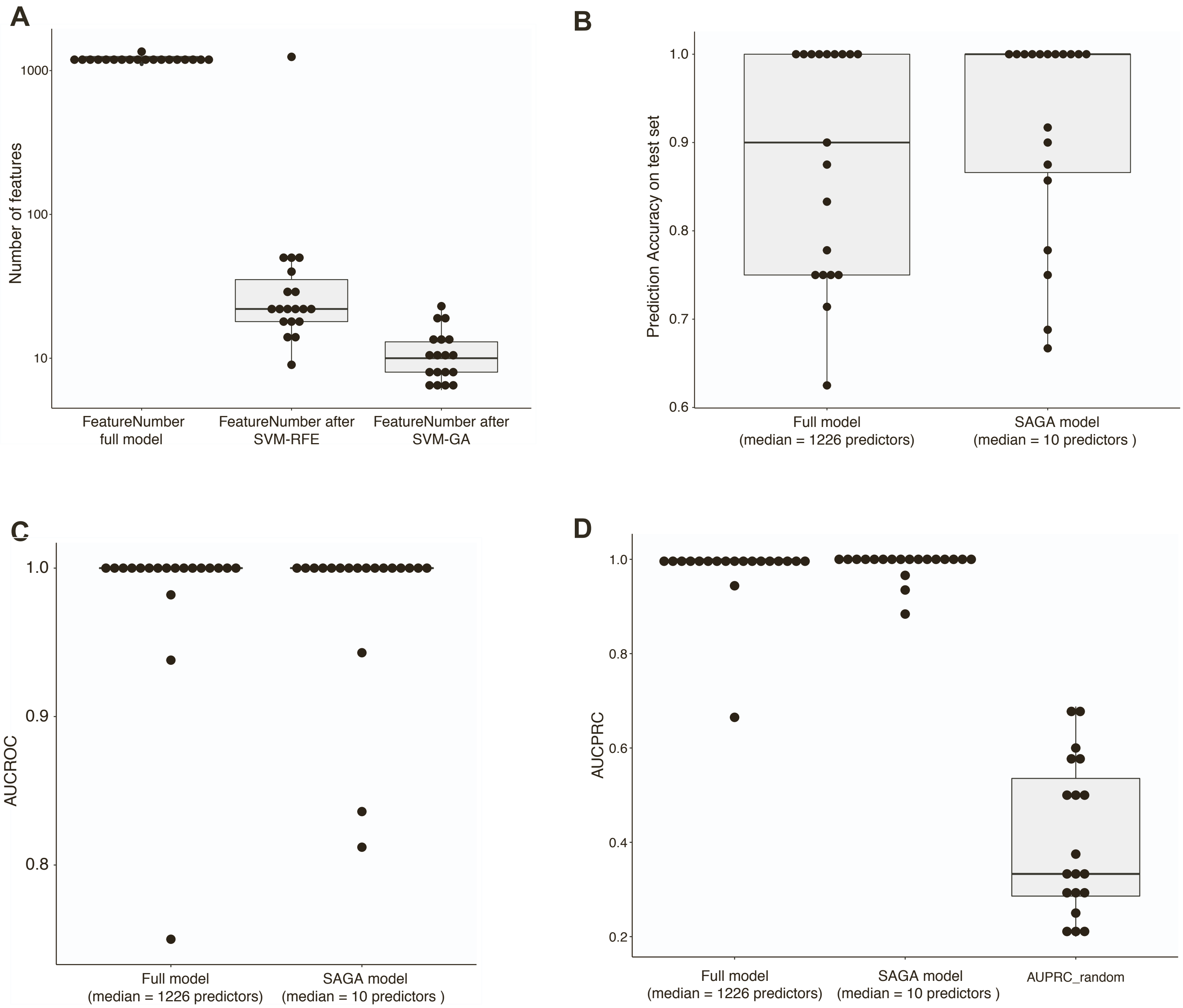

Figure S4 Classification performance metrics over the 19 iterations of the leave-one-batch-out approach. A) number of predictors used for the full model and as input for the feature selection routines (median = 1226), after SVM-RFE (median=22) and after SVM-GA (median = 10) **B)** boxplot of prediction accuracy on the 19 hold-out batches for the full model (mean prediction accuracy = 88.0 %) and SAGA (mean prediction accuracy = 91.7%, $P_{\text{Paired t-test}}$= 0.242) **C)** boxplot of AUC-ROC on the 19 hold-out batches for the full model (mean AUCROC = 0.98) and SAGA (mean AUCROC = 0.98) **D)** boxplot of AUC-PRC on the 19 hold-out batches for the full model (mean AUCPRC = 0.98) and SAGA (mean AUCPRC = 0.99)

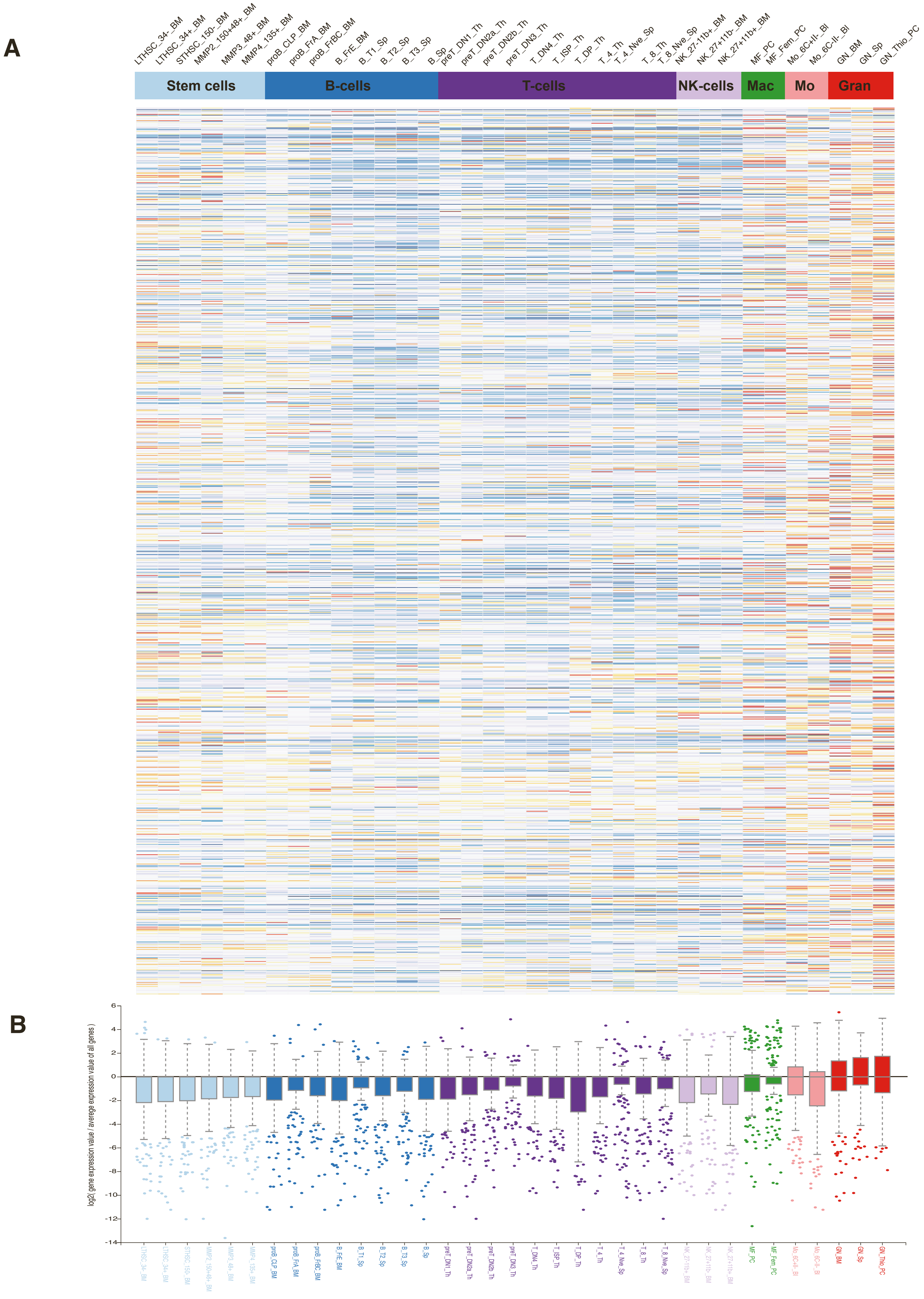

Figure S5 Expression of non-specifically filtered genes across the murine hematopoietic system. A) row- scaled heatmap of 1243 probes/genes retained after unsupervised filtering (IQR = 0.8) of the quantile normalized and batch-corrected expression matrix of n=152 SAGA samples. **B)** Boxplots of expression of genes in each column relative to the expression of all genes demonstrates relative enrichment of myeloid genes in the selection. Abbreviations: LT-HSC: long term-HSC, ST-HSC: short term-HSC, MPP: multipotent progenitors, Mac/MF: Macrophages, Mo: Monocytes, Gran/GN: Granulocytes.

Supplemental Methods and Materials

SAGA samples Table S8 tab 1 and tab 2 give an overview of all SAGA assays, experimental batches and the corresponding microarray samples used at each step. One SAGA assay consists of all SAGA samples generated in the same cell culture experiment. Each SAGA assay that was run independently is one experimental batch, with the following exemptions: SAGA assays with the IDs #160525 and #160706 were run in parallel and constitute one batch (batch 7). Due to severe class imbalance, one SAGA assay (ID #180523) had to be split into two separate batches (#180523A: batch 16, #180523B: batch 17) for normalization and batch correction. Each batch contained the two mock samples from assay #180523 and four or five LTR.RV.SFFV samples, respectively. **Table S8** lists all 179 SAGA samples used in this work, including 169 SAGA samples from day 15 (including the two mock duplicates X6374.1 and X6379.1 from assay #180523B), 5 SAGA samples from day 4, and 5 SAGA samples from day 8. For the computation of differentially expressed genes and pathways between mock, safe and transforming vectors, and for development of the SAGA classifier only samples from day 15 were used. These 169 samples were used as input into the microarray preprocessing pipeline, whose individual steps are visualized in the t-SNE plots of **Figures 3A-C** and **Figures S2A** and **2B** and described in detail in the paragraph "Microarray data processing". After preprocessing, the two mock duplicates (X6374.1 and X6379.1) and 15 samples for which the class label was unknown due to an insufficient number of IVIM assays or inconclusive IVIM results were removed from the analysis, resulting in a final dataset of 152 unique SAGA samples (65 transforming, 55 safe and 32 mock samples), which was used for differential expression, gene set enrichment analysis and development of the SAGA classifier. For the subsequent leave-one-batch-out approach, batch 17 was treated as independent test set with the two mock duplicates X6374.1 and X6379.1 included, resulting in a dataset of 19 test sets and 154 samples in total.

RNA isolation and microarray acquisition On day 15 p.t., cells from bulk cultures were pelleted (5 x10^5 to 2.5 x10^6 cells) and resuspended in 700 µl of RNAzol B reagent (WAK-Chemie Medical) and frozen at -80°C. Total RNA was isolated employing the Direct-Zol RNA MiniPrep Kit (Zymo Research) with on-column DNAse treatment. Four different microarray designs were used in this study, all representing a refined version of the Whole Mouse Genome Microarray 4x44K v2 (Design ID 026655, Agilent Technologies) comprised of all probes of this array in quadruplicates: (1) '026655AsQuadruplicatesOn4x180k' (Design ID 048306) was developed by the Research Core Unit Genomics (RCUG) of Hannover Medical School. Microarray design was created at Agilent's eArray portal using a 4x180K design format for mRNA expression as template. All non-control probes of design ID 026655 were printed four times within one 180K region. (2) '048306On1M' (Design ID 066423), (3) '048306On1M_V3' (Design ID 084107) and (4) '026655QM_RCUG_MusMusculus' (Design ID 084956) were also developed by RCUG, using a 1x1M design format for mRNA expression as template. All non-control probes of design ID 026655 were printed four times within a region comprising a total of 181560 features (probes) (170 columns x 1068 rows). Four of such regions were placed within one 1M region giving rise to four microarray fields per slide to be hybridized individually (Customer Specified Feature Layout). Control probes required for proper Feature Extraction software operation were determined and placed automatically by eArray using recommended default settings. 100 ng of total RNA was used to prepare Aminoallyl-UTP-modified (aaUTP) cRNA (Amino Allyl MessageAmp™ II Kit; #AM1753; Thermo Fisher Scientific) applying one round of amplification as directed by the company, except for a two-fold downscaling of all reaction volumes. Prior to the reverse transcription reaction, 1 µl of a 1:5000 dilution of Agilent's One-Color spike-in Kit stock solution (#5188-5282, Agilent Technologies) was added to 100 ng of total RNA of each analyzed sample. Labeling of aaUTP-cRNA was performed with Alexa Fluor 555 Reactive Dye (#A32756; Thermo Fisher Scientific) as recommended in the manual of the Amino Allyl MessageAmp™ II Kit (two-fold downscaled reaction volumes). cRNA fragmentation, hybridization and washing steps were carried out as recommended in the 'One-Color Microarray-Based Gene Expression Analysis Protocol V5.7', except that 500 ng of each fluorescently labeled cRNA population were used for hybridization. Slides were scanned using the Agilent Micro Array Scanner G2565CA (pixel resolution 3 µm, bit depth 20). Data extraction was performed with the 'Feature Extraction Software V10.7.3.1' with the extraction protocol file 'GE1_107_Sep09.xml'.

Microarray annotation Since microarray probe annotation may change as the genome annotation advances, we re-annotated the 39,428 probes on the Agilent Whole Mouse Genome Oligo Microarray 4x44K v2 (Design ID 026655) by mapping the 60mer sequences to a recent release of the murine transcriptome (Gencode version M18[1], GRCm38.p6, release 07/2018). The transcript databases were downloaded as FASTA files for the 64,732 protein coding transcripts (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M18/gencode.vM18.pc_transcripts.fa.gz) and for all 136,535 coding and noncoding transcripts of the reference transcriptome (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M18/gencode.vM18.transcripts.fa.gz). The annotation was performed using R 3.5.1, Bioconductor 3.7[2] and the R package "Biostrings"[3]. Due to their higher expression compared to non-coding transcripts[4] we prioritized protein coding sequences by aligning the 60mers first to all protein coding transcripts of Gencode M18, and second to all transcripts of Gencode M18 allowing a maximum of 3 mismatches per 60mer. Using these parameters 33,361 out of 39,428 probes were successfully mapped to the Gencode M18 transcriptome. The mapping process retrieved an Ensembl-GeneID (e.g. "ENSMUSG00000020743") for each probe with a hit in the Gencode transcriptome. The Ensembl-GeneID was further annotated using the "BiomaRt" [5] R package to retrieve gene symbols, description and gene type from the Ensembl 94 database. For probes that could not be annotated by Gencode, annotation was taken from the latest annotation file for the Whole Mouse Genome Oligo Microarray 4x44K v2 downloaded from Agilent eArray web service (https://earray.chem.agilent.com/earray/, ID 026655, released October 2017) resulting in annotation of 2872 additional probes and 36,226 annotated probes in total. The R script for the annotation and all files used are available in the GitHub repository accompanying this manuscript.

Microarray data processing The data was analyzed using R 3.5.1 and Bioconductor 3.7[2]. Raw files were read in separately for each array design and a merged dataset was created by extracting all probes derived from the original Agilent Mouse Genome Oligo Microarray 4x44K v2 array from the four array platforms and combining them using the function "cbind.EList" from "limma". The probe with the ID "A_55_P2337033" interrogating the gene "2310065F04Rik" was excluded from the dataset since it strongly cross-reacted with the sequence of EGFP. Array quality was assessed by interrogation of probe intensity distributions and by principal component analysis of $\log_2$-transformed unprocessed data. The Raw data was $\log_2$-transformed and quantile-normalized using the "limma" package. The success of preprocessing was verified by inspection of probe intensity distributions before and after preprocessing (**Figure S2A** and **2B)**. The four within-array replicates of each probe were collapsed using the "avereps" function from the R package "limma" resulting in a dataset with 39,428 unique probes. Probes interrogating the same gene were not collapsed further since most genes were only interrogated by one probe on this platform. In the quantile-normalized data, a substantial batch effect between different SAGA assays was observed (**Figure 3A)**. Batch correction between different SAGA assays was performed on quantile-normalized $\log_2$-values using the parametric ComBat algorithm as implemented in the R package "sva" [6] (**Figures 3B** and **3C)** with the SAGA number as batch variable and all other parameters set to default.

t-SNE and PCA visualizations Two-dimensional representation of gene expression profiles was visualized by t-distributed stochastic neighbor embedding (*t*-SNE)[7]. The Barnes-Hut implementation of *t*-SNE from the "Rtsne"-package[8] without prior dimension reduction was used for all t-SNE representations. For each t-SNE plot, Barnes-Hut t-SNE was run 1000 times with different random seeds and the iteration with the lowest Kullback-Leibler divergence was selected for visualization as a 2D plot. For t-SNE visualizations of the whole dataset (**Figures 3A-C**), all 39,428 probes were used and the perplexity was set to 16, since this exceeded the average number of samples within each cluster/SAGA assay and is within the range of 5-50 proposed by the authors of t-SNE[7]. For the t-SNE plots in **Figures 2A**, **2C** and **Figure S1A**, 36,226 annotated probes were used and the perplexity was set to 2, which was the maximum value allowed for this sample size. The "prcomp" function from the R package "stats" was used to perform principal component analysis. The function "heatmap.2" from the R package "gplots" was used to generate heatmaps on the number of probes indicated in the figure legend. Heatmaps were row-scaled with the color key indicated below the heatmap. Variance-based filtering of probes for unsupervised analysis was performed using the interquartile range (IQR) function in the package "genefilter" resulting in the number of probes indicated in the figure legend.

Differential expression analysis Differentially expressed probes between the subgroups were computed on the quantile normalized and batch corrected expression matrix of 36,226 annotated probes using the moderated t-test of the ”limma” package[9] with Benjamini-Hochberg multiple testing correction. We computed the Toplists (differentially expressed genes) for the following contrasts: ”transforming – mock” (**Table S2 tab 3**) , “safe – mock” (**Table S2 tab 4**), “transforming – safe” (**Table S2 tab 5**) and “transforming – (mock+safe)/2” (**Table S2 tab 1**) for 152 SAGA samples with known IVIM properties (65 transforming, 32 mock and 55 safe).

Gene set enrichment analysis The quantile normalized and batch corrected SAGA expression matrix (36,226 annotated probes, 152 samples with known IVIM properties (transforming, mock and non-transforming (”safe”)) was first filtered for gene symbols that appear at least once in the interrogated MSigDB.v6.2 (C2, C3, C5, C6, hallmark) gene set collections[10]. In cases with multiple probes per gene, the probe with the highest standard deviation across the samples was selected, resulting in a gene expression matrix consisting of 15,376 probes/rows interrogating 15,376 unique genes. From this matrix .gct files were generated containing all 65 transforming and 32 mock samples (contrast “transforming vs mock”, **Table S3 tab 2 – tab 4**), 65 transforming and 55 safe samples (contrast “transforming vs safe”, **Table S3 tab 7**), 55 safe and 32 mock samples (contrast “safe vs mock”, **Table S3 tab 8 – tab 10**), 65 transforming, 32 mock and 55 safe samples (contrast “transforming vs mock and safe”, **Table S3 tab 11 – tab 13**). For samples from day 4, cultures LTR.SF.EGFP (n=4) and one mock sample were used (contrast “transforming vs mock day 4”, **Figure S1C**). For the comparison of day 8 and day 15, samples (**Table S3 tab 14 – tab 16**) were preprocessed together with all 169 SAGA samples and treated as a separated batch in COMBAT. For the GSEA contrast “d8 mock vs d15 mock”, the two mock samples from day 8 were compared to 32 mock samples from day 15. The .gct files were used as input for the Broad GSEA software[10] together with a .chip file containing the annotation for the 15,376 probes. GSEA was performed with ranking the probes according to signal to noise ratio and the permutation type set to “gene_set” (10,000 permutations). First, we used 106 custom gene sets related to hematopoiesis and leukemia (**Table S3 tab 1**)[4]. In addition, 8286 gene sets were tested for enrichment from MSigDB.v6.2 (C2, C5, hallmark gene sets). The enrichment results were visualized by plotting the normalized enrichment score (NES) against the FDR (**Figures 2F-2H**). For visualization purposes, gene sets with a nominal FDR of zero were assigned a $\log_{10}$ FDR between -5 and -6 in **Figure 2F-2H** and **Figure S1F**. **Table S3 tab 2 – tab 16** contain all exact results of GSEA computations. Competitive gene set tests using permutation of genes assume statistical independence of genes in the gene sets, which is unrealistic in most cases. It has been shown that inter-gene correlation can lead to falsely significant P-values in these tests [11]. In contrast, permutation of the sample labels preserves inter-gene correlation, but requires a substantial number of samples in each group, suffers from low statistical power and inevitably alters the hypothesis being tested. Therefore, we additionally performed GSEA with ROAST (rotation gene set tests for complex microarray experiments[12]) and CAMERA (competitive gene set test accounting for inter-gene correlation[13]) from the limma package by applying both functions to the matrix of 15,376 probes and computing the same contrasts as with the Broad GSEA tool. The parameters for ROAST were set to 50,000 rotations and set.statistic="mean" (default value). For CAMERA the inter.gene.cor parameter was set to 0.01, as proposed by the authors[13]. CAMERA and ROAST allow for non-independence of genes by estimating the inter-gene correlation (CAMERA) or using rotation of residuals to generate a valid null distribution (ROAST)[12]. Importantly, both methods test different null hypotheses: whereas CAMERA is a competitive test that interrogates whether genes within the gene set of interest are significantly more often differentially expressed compared to genes outside of the gene set, ROAST is a self-contained test that tests whether a defined proportion of genes within the gene set is differentially expressed at all. However, while both methods have been shown to control the FDR correctly compared to methods based on gene permutation[12,13], they do not report a normalized enrichment score or a similar measure, making it difficult to assess how strong the gene set is enriched at the top or bottom of the ranked gene list. This also makes comparisons between different gene sets difficult. Therefore, we report both the results of GSEA with the intuitive and widely used NES (normalized enrichment score) and the results of CAMERA/ROAST based on a rigorous test statistic. All gene sets labeled in **Figures 2F-2H** and **Figure S1F** were found to be significantly enriched (FDR < 0.1) by at least one additional method (ROAST or CAMERA), whereas most of the gene sets were found by both additional methods (**Table S3 tab 2 – tab 16)**. For the enrichment map network shown in **Figure 2M**), the output from the GSEA analysis querying 8,286 gene sets from MSigDB.v6.2 was used as input for the Enrichment Map Tool[14] for Cytoscape 3.7.1. Gene sets with a nominal FDR < 0.05 were selected for visualization in the network graphs. The color of the nodes encodes normalized enrichment score as shown in the color key. A similarity cutoff of 0.375 (combined Jaccard and overlap) was used.

Classifier development phase The development of the predictive model was implemented using the R package “caret”[15] based on a support vector machine with a radial basis function kernel (method = "svmRadial"). Unless otherwise specified, all calls to functions mentioned in this paragraph belong to the “caret” package with key parameters specified in parentheses after the name of the function or directly discussed in the text. Computations allowing multiple cores, e.g. the feature selection routines, were run on a c5.18xlarge Amazon Web Service EC2 instance with 72 cores and 144 Gb RAM running RStudio 1.1.456 and R 3.5.1. The data splitting and resampling scheme to assess the performance of the models and control for overfitting is outlined in **Figure S3**. First, the quantile normalized and batch corrected expression matrix (36,226 annotated probes, 152 samples with known IVIM properties) was partitioned into a training set comprised of 70% of the samples (107 samples) and an independent test set of 30% of the samples (45 samples). The test set was not used at any point for feature selection or model tuning. To allocate samples to the test or training set the caret function "createDataPartition" (p=0.7) was used, which performs stratified sampling based on the class labels to keep the distribution of transforming and nontransforming samples equal between the training and test sets. Since a single training / test set split can lead to a biased assessment of model building and feature selection[16] ten stratified random training / test set splits of the dataset were created and the complete model building pipeline was run for ten times for a more unbiased and reliable assessment of the predictive modeling process. Predictive performance of many models, especially support vector machines, can be significantly affected by large numbers of irrelevant predictors[17]. Furthermore, models using fewer predictors are quicker to compute, less prone to overfitting and generally better interpretable than models based on thousands of predictors[16]. Therefore, a combination of feature selection steps was performed to reduce the number of predictors as far as possible while maintaining or increasing predictive power. First, we applied an unsupervised filter to each training set to exclude probes interrogating genes that were not expressed at all or show only little variation in the dataset. This step helped to reduce computation time and avoided the selection of features by the subsequent *SVM-RFE* step that have a good discriminatory power between the classes based on their AUROC, but display only a small absolute fold-change between the different classes. The R-package "genefilter" was used to discard probes with an interquartile range (IQR) of $\log_2$-expression values less than 0.8 in the quantile-normalized and batch corrected training cohort, which retained a median of 1,195 out of 36,226 annotated probes (**Table S4 tab 1**). IQR = 0.8 was chosen empirically, since it consistently selected around 1,000 features in all test/training set splits. Setting the IQR lower (e.g. IQR= 0.5) retained too many features (median around 4,500), leading to a substantial increase in overall computation time as well as a failure to reduce the number of features in the subsequent *SVM-RFE* step in 3 out of 10 training/test splits. In contrast, setting IQR=1.2 selected on average around 250 features, which could be efficiently handled by *SVM-RFE*. However, at IQR=1.2 important predictors, such as A_55_P2077048/Itih5 (AUROC= 0.98) were already discarded before the actual feature selection step. The implementations using IQR 0.5 and IQR = 1.2 are available at GITHUB. Next, we performed recursive feature elimination (*SVM-RFE*) on the training set using the function "rfe". Since feature selection is part of the model building process, it needs to be conducted inside of a resampling layer ("external resampling layer", **Figure S3**) to assess the impact of the selection process on the model performance and to prevent overfitting of the model to the predictors. To establish the external resampling layer, 200 resamples of the training set were created by twenty times repeated 10-fold cross-validation using the function "createMultiFolds" (Parameters: k=10, times = 20). The function divides the entire training set (107 samples) into 10 subsets (folds) of equal size and the first fold (11 samples, "external holdouts") is predicted by a model fit to the remaining 9 folds (96 samples, "external training") of the data. This is repeated with the second fold after the first one has been returned to the training set and so on, resulting in 10 resamples for each of the twenty repeats of 10-fold CV. Importantly, the 200 identical resamples were used to fit the full models using all predictors, to allow a direct comparison of the *SVM-RFE* model and the full model using the resampling accuracies. The 200 resamples were submitted to the helper function "rfeControl", which controls the details of the external resampling process of the function "rfe". The feature selection process itself was carried out for each of the 200 resamples separately and computed in parallel by setting the "rfeControl" parameter: "allowParallel = TRUE". To ensure reproducibility of the analysis, a fixed set of random seeds that "rfe" uses at each resampling iteration was created and submitted to "rfeControl" via the "seeds" parameter. Within each resample, *SVM-RFE* ranks all predictors according to their individual receiver operating characteristic (ROC) on the 96 training samples. In each iteration, less important predictors are removed, the model is fitted to the 96 training samples and the 11 holdout samples are predicted. The metric to be maximized by "rfe" was set to "Accuracy". After initial inspection of the resampling profiles, we noted that accuracy peaked most often between 5-30 predictors. For maximum resolution within these ranges, all subset sizes from 1-40 predictors were tested. Outside of this range, wider intervals were used (45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 predictors), resulting in 52 subset sizes in total. For each tested subset within each resample of the external layer an additional "inner layer" of resampling had to be established to determine the tuning parameters of the SVM-model. The details of the inner resampling layer were specified by the helper function "trainControl" and set to three times repeated 10-fold cross-validation (30 resamples). To be precise, each training set from the external layer (96 samples) was partitioned further into 30 internal resamples comprised of 86 "internal training" and 10 "internal holdout" samples, respectively (**Figure S3)**. For each value of tuning parameters and each internal resample, the SVMrad model was fit to the 86 internal training samples and the remaining 10 internal holdout samples were predicted. The prediction accuracy from the 30 internal resamples over the different tested hyperparameter values was used to determine the optimal value for the tuning parameters and these parameters were passed to the external layer to fit the model and predict the external hold-outs. *SVM-RFE* with a radial basis function kernel has two tuning parameters: cost (penalty parameter) and sigma (inverse width of the gaussian kernel). For the cost parameter, the parameter "tuneLength" of the "rfe"- function was set to 20, resulting in cost values ranging from 2^{-2} - 2^{17}. For the sigma parameter an analytical estimate was used which is calculated by "rfe" internally by calling the function "sigest" from the R-package kernlab[18]. "sigest" uses the methodology proposed by Caputo et al[19] to estimate a value for sigma which results in a good prediction performance when used with a radial kernel SVM. We validated this approach initially by a manual search for sigma over a wide range of values (2^{-15} - 2^{0}), but could not find substantially better solutions for our dataset than suggested by "sigest" (data not shown). Hence, using a fixed value for sigma estimated with "sigest" and tuning the SVM over the cost parameter only resulted in a substantially smaller hyperparameter space and reduced computation time for *SVM-RFE*. To find the best subset size for the entire training set, the prediction accuracy of the external holdout samples for each subset size and each resample was averaged into a resampling profile (**Figures 3E and 3F, Figure 5A**), which allowed to determine the best average subset size across all resamples. To generate the final set of predictors, "rfe" repeated the process on the complete training set with the optimal subset size determined from the resampling profile. The performance of the *SVM-RFE* model was compared to the full model using all predictors using the caret functions "resamples" and "diff", which compare resampling results of different models on a common data set comprised of identical resamples using a paired t-test[20]. The resampling-based results for the ten training / test set splits and the final model are tabulated in **Table S4 tab 1** (P-value_Resampling_full_vs_rfe). The GA procedure was implemented using the function "gafs" and its helper function "gafsControl" from the R package "caret". The gene expression matrix reduced to the probes found by the preceding *SVM-RFE* step was used as input into GA. Similarly to the *SVM-RFE* implementation, SVM-GA was conducted inside an external resampling layer to assess the performance of the GA-model over the generations (external resampling accuracy). 50 external resamples of the training sets or the final dataset were created with the function "createMultiFolds" (k=10, times = 5) and passed to the function "gafsControl", which controls the outer resampling process of the GA. The computational burden of *SVM-GA* is higher than for SVM-RFE, so only 50 external resamples were used to complete the analysis in a reasonable amount of time. The prediction performances on the external hold-out samples at each generation across all external resamples were averaged into the external resampling profile (**Figures 3G** and **5B),** which was used to determine the optimal number of iterations the algorithm should proceed **(Figure S3)**. To determine the final feature set, "gafs" applied the GA to the entire training set for the optimal number of generations from the resampling process. Further parameters of "gafsControl" were set to enable parallel computing for the external layer, to maximize the test statistic (accuracy) and to use fixed random seeds for reproducibility. In initial runs, using the default settings of "gaf" feature reduction was quite inefficient, leading to the removal of only 3-5 predictors on average. For a more effective reduction of feature numbers, the size of the initial predictor subsets (chromosomes) in the starting population was reduced. Therefore, the helper function of GA (caretGA$initial) that creates the initial population was modified to produce chromosomes comprised of a random 40% of predictors, instead of creating initial subsets ranging from 10% to 90% of predictors. The GA procedure itself was run for 40 generations, with a population size of 40, a crossover probability of 0.7, a mutation probability of 0.1. Elitism was set to 3, meaning that the best three solutions survive to the next generation. The metric to optimize was set to "accuracy", the classification method to "svmRadial". Similarly to the *SVM-RFE* process, the GA had an additional inner layer of resampling conducted at each generation within each resample and for each chromosome to tune the SVM. The inner resampling layer of GA was set to two times repeated 10-fold cross-validation (20 resamples) by the helper function "trainControl". For the cost parameter of the SVM, the parameter "tuneLength" was set to 12~~,~~ for cost values between 2^{-2} – 2^{9}. The reduced tune length was chosen to save computation time after it had been determined from the preceding steps that the optimal cost parameter for the SVM was in the range of 2^{-2} - 2^{7}. For the sigma parameter, the estimate computed by "sigest" function from "kernlab" was used as described above. For the analysis of gene expression of the selected predictors across murine haematopoiesis (20 probes from *SVM-RFE* and 1243 probes after unsupervised filtering, **Table S6 tab 2,3 and Figures 5E** and **S5**), the online resource of the Immunological Genome Consortium[21] (http://rstats.immgen.org/MyGeneSet_New/index.html) was queried using the corresponding gene symbols of the probes as input.

Classifier performance metrics Samples in the test sets were predicted after training a support vector machine with radial kernel on the training set using all predictors (full model) or reduced to the optimal predictors found by *SVM-RFE* and *SVM-GA* (reduced models) by using the caret functions "train" and "predict", respectively. For training the full and the reduced SVM-models, identical parameters and resamples were specified in the "train" function (method = "svmRadial", metric = "Accuracy", tuneLength = 20, twenty repeats of 10-fold cross-validation). The function "predict" was used with the parameter "type" set to "prob", which computes the probability that a sample belongs to a given class. An unknown sample was considered belonging to the class "transforming" when the probability for class "transforming" was greater than 0.5. Performance estimates (sensitivity, specificity, accuracy, kappa) for the predicted test sets were computed using the function "confusionMatrix" on the predicted and the true class labels, respectively. For **Figures 3H-3J**, the resampling accuracies and their confidence intervals were determined using the function "resamples" for the full models, *SVM-RFE* and *SVM-GA* and plotted on the y-axis. The values on the x-axis represent the test set accuracies and the corresponding confidence intervals as output by the function "confusionMatrix". The "pROC" R-package[22] (v1.15.3) was used to compute and visualize the ROC curves for the test sets using the function "roc" on the probability for class "transforming" as output by the "predict" function. *P* values to compare the difference between the AUROC of two unpaired ROC curves were performed with the "roc.test" function using the "delong" method and the alternative hypothesis set to “greater”. Precision recall curves were generated using the R-package “PRROC” (v.1.3.1). As delineated in the main text, we defined SAGA as the compound model based on the predictions from *SVM-RFE* when this process yielded equal or less than 10 optimal predictors and from *SVM-RFE* followed by *SVM-GA* otherwise. For **Figures 4D-4I**, the prediction results for SAGA for all 19 independent test batches were aggregated and compared to the performance of the IVIM assay via AUROC, AUPRC and calculation of the confusion matrices and associated performance estimates (**Table S5)**.

Performance estimation via leave-one-batch-out approach Raw intensities of 169 arrays from 19 experimental batches were read in and combined into an “EListRaw” object without further modification. 15 samples with unknown ground truth were subsequently removed from the dataset, resulting in 154 assays including two mock duplicates (X6374.1, X6379.1 from batch 17, **Table S8**). For iteration 1, the raw data of batch 1 (IVIM #120411) was set aside as an independent test set, all other batches (2-19) were used as training set and were quantile normalized, averaged and batch corrected as described above. The preprocessed training set was subjected to *SVM-RFE* and *SVM-GA* using the same parameters as above, except for the numbers of subset sizes to assess during SVM-RFE, which were reduced to 1,2,3…,40,45,50, all predictors = 43 predictor subsets in total to limit computational costs. After having determined the optimal predictors in the training set, the raw training set was again quantile normalized and batch-corrected by the R package “bapred”[23], in order to estimate and store the parameters necessary for the later add-on correction of the test set. An SVM with radial kernel was trained on the bapred-adjusted training set reduced to the optimal predictors found by the feature selection routines. The hyperparameters of the SVM (sigma and cost) were determined by 20 times repeated 10-fold cross-validation as described above. At this point, the optimal features had been determined and the classifier had been trained and fixed using the training set only, whereas the test set had not been used. This was followed by add-on quantile normalization and add-on batch correction of the raw-test set using the bapred functions “qunormaddon” and “combatbaaddon”, respectively. Add-on adjustment prevents the alteration of the training set by the addition of test set samples (information leakage) by applying the necessary adjustments to the test data using parameters estimated on the training data only[23]. The add-on adjusted test set was reduced to the optimal predictors determined on the training set (e.g. for the first iteration: 8 predictors) and predicted using the SVM trained before and the caret function “predict”. The complete procedure was repeated 18 additional times with every available batch to be used one time as independent test set. The results from the 19 iterations of building SAGA and predicting the independent test batches are summarized in **Figure 4, Table S5** and **Figure S4.**

SAGA R package The R implementation of the SAGA classifier is available online (https://github.com/mytalbot/saga_package) and its functionality is described in detail in the package vignette. The SAGA package depends on R ≧ 3.6. The SAGA package expects data from microarrays based on Agilents Whole Mouse Genome 4x44K v2 platform as input. The Agilent Design IDs of compatible arrays are given in the section “*RNA isolation and microarray acquisition”* above. SAGA is a support vector machine with radial kernel that is trained on the complete SAGA dataset of 152 arrays reduced to the 11 optimal predictors derived from this dataset by applying the pipeline developed above (quantile normalization, batch correction and feature selection) to all 152 SAGA samples with known IVIM behavior (Table S6). The SVM is trained by using the “caret” function “train” with the following parameters: method = "svmRadial", metric = "Accuracy", tuneLength = 20 and five repeats of tenfold cross-validation for tuning the cost parameter, the sigma parameter is estimated internally by “train” as outlined before. The unknown samples are read in using the “limma” function “read.maimages” followed by add-on quantile normalization and add-on batch correction using the functions “qunormaddon” and “combatbaaddon” from the R package “bapred”[23]. Add-on adjusted test sets are then reduced to the 11 optimal SAGA predictors. Prediction of the unknown samples is performed by the function “predict” with the parameter “type” set to “prob” as described above. An unknown sample is considered belonging to the class “transforming” when the probability for class “transforming” is greater than 0.5. Prediction of unknown samples by *SAGA-GSEA* follows the procedure described in the paragraph *SAGA-GSEA*.

SAGA-GSEA For the implementation of *SAGA-GSEA*, complete assays were read in batch-wise, quantile-normalized, averaged and $\log_2$-transformed within each assay using the R package “limma”. The preprocessed and unfiltered expression matrix with the Agilent ProbeIDs as row names was directly converted into an “epheno” object using the function “ExpressionPhenoTest” from the package “phenoTest”[24] with the phenotype variable (“Group”) set to 1 for all mock samples in each assay and a unique value {2,3,…,n} for each of the samples to be tested against the mock samples. The normalized enrichment score, p-values and fdr were calculated for every sample against the mock samples using the function “gsea” from “phenoTest”. During the leave-one-batch-out procedure, the optimal predictors found by *SVM-RFE* and *SVM-GA* for the training set of each iteration were used as geneset for GSEA. The raw data of the left-out test set was read in and preprocessed as described above followed by GSEA. IVIM #171102 was excluded from *SAGA-GSEA* since it had no mock samples available. The GSEA results were aggregated over the remaining 18 test sets. The ROC curve for *SAGA-GSEA* and the best NES cutoff were computed using the function “roc” on the normalized enrichment scores and the true class labels with the parameter “threshold” set to “best”, which determines the NES associated with the point farthest to the diagonal line[22]. A vector was assigned to the class "transforming" when its NES was greater than the optimal ROC-cutoff computed on the dataset after exclusion of the strongly transforming LTR.SFFV.eGFP samples (leave one-batch-out: NES>1.3, **Figure 6D**). For the final implementation of *SAGA-GSEA* to be used in the R-package, the 11 optimal predictors determined on the complete dataset for the final SAGA classifier (see above) are used as geneset. The optimal NES threshold for this geneset was determined by ROC-analysis after performing *SAGA-GSEA* on the 18 SAGA batches with mock controls available (NES > 1.0, **Figure 6G**).

Quantitative real-time PCR For quantitative real-time PCR (q-RT-PCR), 200 ng total RNA from day 15 samples were reverse transcribed with the QuantiTect Reverse Transcription kit (QIAGEN, Hilden, Germany). cDNA samples (20 µl reaction volume) were diluted with 20 µl water and 2 µl were used for each q-RT-PCR replicate. For quantification of gene expression in duplicate measurements, we used a mastermix of 7.5 µl 2x QuantiTect SYBR Green (QIAGEN), 0.75 µl 20x PrimeTime qPCR Assays (Mm.PT.39a.22214843.g, Mm.PT.56a.9170255, Mm.PT.58.10065691, Mm.PT.58.11560570, Mm.PT.58.5431010, Mm.PT.58.32478304.g, Mm.PT.58.41635140, Mm.PT.58.41288607, Mm.PT.58.12595646, Mm.PT.58.41494395, Mm.PT.58.5925960, all from Integrated DNA Technologies, Coralville, USA), 4.75 µl water and 2 µl diluted cDNA. The program in the StepOnePlus thermocycler (Thermo Fisher Scientific, Inc.) was 15 min 95°C, 50 cycles of 30 sec 94°C, 30 sec 60°C, 30 sec 72°C and a melt curve analysis with 15 min 95°C, 1 min 60°C and a gradual increase to 95°C for 15 min (2.3°C/min). Target gene expression was analyzed by the delta-delta Ct-method relative to *Actb*[25]. All transduced samples were compared to the mock control of the respective assay.

RNA-Seq RNA from three SAGA assays on day 15 was isolated as described above. RNA samples were sent for sequencing to Novogene Bioinformatics Technology Co., Hong Kong. The sample quality was verified with Agilent 2100. After mRNA enrichment with the NEBNext Poly(A) mRNA Magnetic Isolation Module, sequencing libraries were generated using the NEB Next® Ultra™ RNA Library Prep Kit from NEB. Samples were sequenced on an Illumina HiSeq2500. RNA-seq reads (on average 55 x 10^6 read counts per sample) were aligned to the Gencode mouse reference genome (GRCm38.p5) using Tophat2[26], which generated 44.5 x10^6 uniquely mapped reads on average. Count matrices were computed for Gencode defined transcripts and all reads that were unambiguously assigned to annotated exons were submitted to further expression analysis with DESeq2[27]. Heatmaps were generated by using rlog-transformed (Regularized logarithm transformation) values of normalized counts.

Supplemental References

1. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2018). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research *47*, D766–D773.

2. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biology *5*, R80-16.

3. Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2018). Biostrings: Efficient manipulation of biological strings. R package.

4. Schwarzer, A., Emmrich, S., Schmidt, F., Beck, D., Ng, M., Reimer, C., Adams, F.F., Grasedieck, S., Witte, D., Käbler, S., et al. (2017). The non-coding RNA landscape of human hematopoiesis and leukemia. Nature Communications *8*, 1–16.

5. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols *4*, 1184–1191.

6. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

7. Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 2579–2605.

8. Maaten, L. van der (2014). Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research *15*, 3221–3245.

9. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research *43*, e47–e47.

10. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545–15550.

11. Gatti, D.M., Barry, W.T., Nobel, A.B., Rusyn, I., and Wright, F.A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. BMC Genomics *11*, 574.

12. Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.L., Visvader, J.E., and Smyth, G.K. (2010). ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics *26*, 2176–2182.

13. Wu, D., and Smyth, G.K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research *40*, e133–e133.

14. Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. PLoS ONE *5*, e13984-12.

15. Kuhn, M. (2018). caret: Classification and Regression Training.

16. Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling 1st ed.

17. Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics *23*, 2507–2517.

18. Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab- An S4Package for Kernel Methods in R. journal of statistical software *11*, 1–20.

19. Caputo, B., Sim, K., Furesjo, F., and Smola, A. (2002). Appearance–Based Object Recognition Using SVMs: Which Kernel Should I Use? Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision.

20. Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005). The Design and Analysis of Benchmark Experiments. Journal of Computational and Graphical Statistics *14*, 675–699.

21. Yoshida, H., Lareau, C.A., Ramirez, R.N., Rose, S.A., Maier, B., Wroblewska, A., Desland, F., Chudnovskiy, A., Mortha, A., Dominguez, C., et al. (2019). The cis-Regulatory Atlas of the Mouse Immune System. Cell, 1–37.

22. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77–8.

23. Hornung, R. (2016). Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. BMC Bioinformatics, 1–19.

24. Planet, E. (2017). phenoTest Package. R Vignette, 1–15.

25. Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Research *29*, e45.

26. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology *14*, R36.

27. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.