Science Advances
AAAS

Supplementary Materials for

Integrome signatures of lentiviral gene therapy for SCID-X1 patients

Koon-Kiu Yan *et al.*

Corresponding author: Jiyang Yu, jiyang.yu@stjude.org; Stephen Gottschalk, stephen.gottschalk@stjude.org

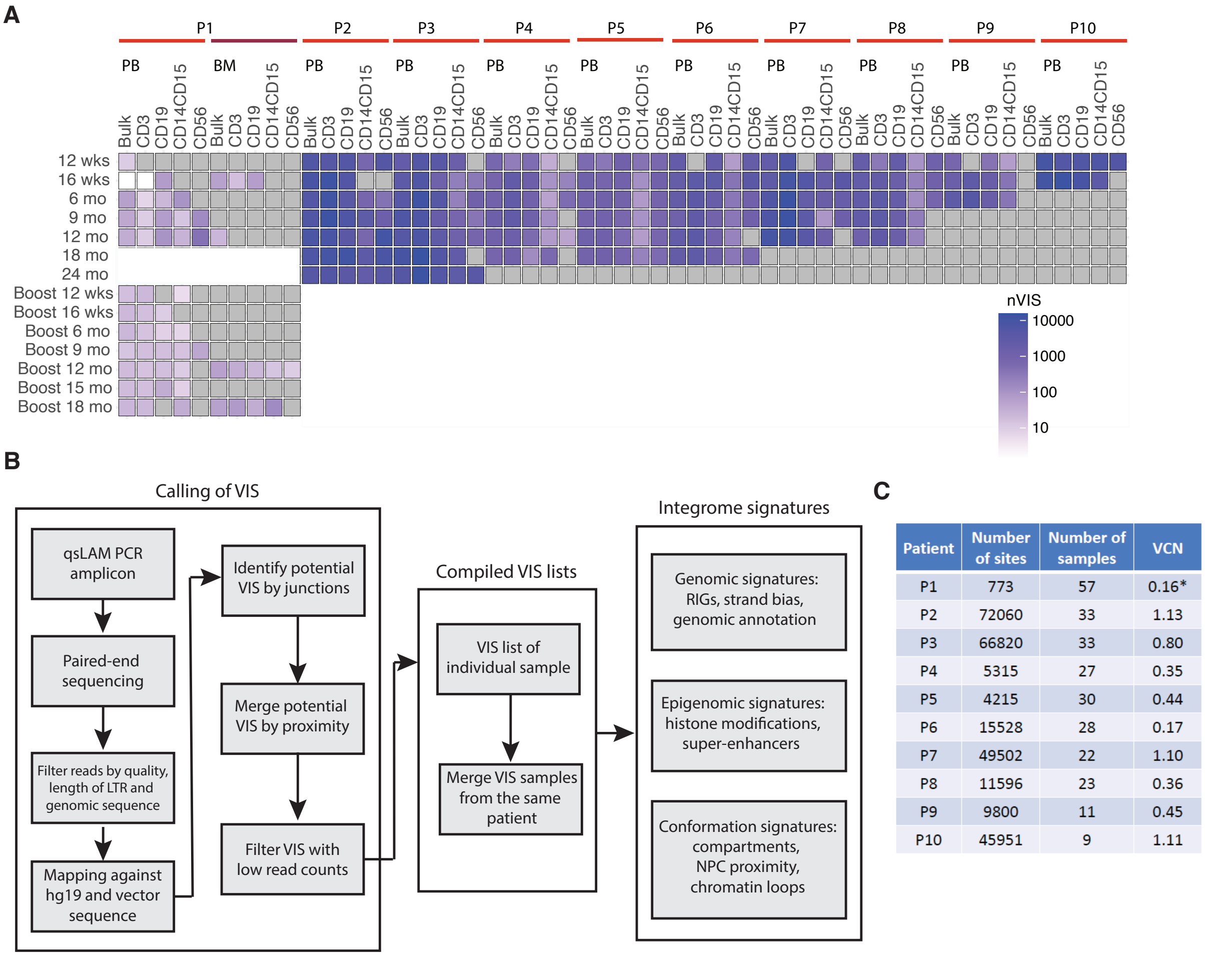
Sci. Adv. **9**, eadg9959 (2023)
DOI: 10.1126/sciadv.adg9959

The PDF file includes:

Figs. S1 to S17
Legends for tables S1 and S2
Legend for data S1

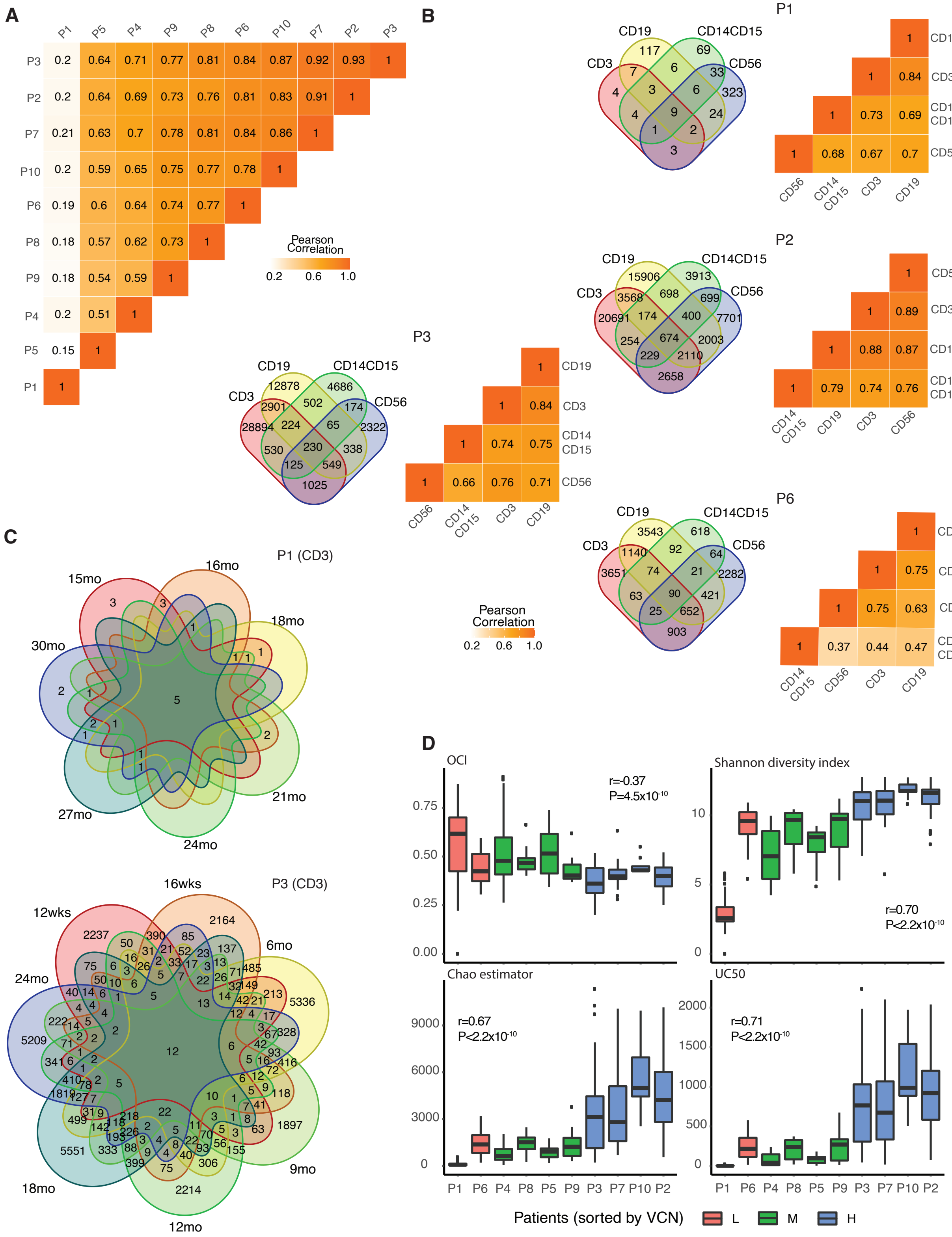
Other Supplementary Material for this manuscript includes the following:

Tables S1 and S2
Data S1



Patient	Number of sites	Number of samples	VCN
P1	773	57	0.16*
P2	72060	33	1.13
P3	66820	33	0.80
P4	5315	27	0.35
P5	4215	30	0.44
P6	15528	28	0.17
P7	49502	22	1.10
P8	11596	23	0.36
P9	9800	11	0.45
P10	45951	9	1.11

Fig. S1: VIS analysis pipeline. (**A**) Patient samples used in the study. The color of a square represents the number of VIS in the corresponding sample. Grey squares mean the samples are not available. **(B)** From VIS calling to integrome signatures. (**C**) The number of VISs and the number of samples analyzed for each patient (* patient 1 received a boost 12 months after infusion, with a VCN of 0.22).



	P1	P5	P4	P9	P8	P6	P10	P7	P2	P3
P3	0.2	0.64	0.71	0.77	0.81	0.84	0.87	0.92	0.93	1
P2	0.2	0.64	0.69	0.73	0.76	0.81	0.83	0.91	1	
P7	0.21	0.63	0.7	0.78	0.81	0.84	0.86	1		
P10	0.2	0.59	0.65	0.75	0.77	0.78	1			
P6	0.19	0.6	0.64	0.74	0.77	1				
P8	0.18	0.57	0.62	0.73	1					
P9	0.18	0.54	0.59	1						
P4	0.2	0.51	1							
P5	0.15	1								
P1	1									

P1

CD56	CD14 CD15	CD3	CD19	
			1	CD19
		1	0.84	CD3
	1	0.73	0.69	CD14 CD15
1	0.68	0.67	0.7	CD56

P3

CD56	CD14 CD15	CD3	CD19	
			1	CD19
		1	0.84	CD3
	1	0.74	0.75	CD14 CD15
1	0.66	0.76	0.71	CD56

P2

CD14 CD15	CD19	CD3	CD56	
			1	CD56
		1	0.89	CD3
	1	0.88	0.87	CD19
1	0.79	0.74	0.76	CD14 CD15

P6

CD14 CD15	CD56	CD3	CD19	
			1	CD19
		1	0.75	CD3
	1	0.75	0.63	CD56
1	0.37	0.44	0.47	CD14 CD15

Fig. S2: Lentiviral VIS profiles are consistent across patients with SCID-X1. (A) Correlation coefficients for VIS density profiles across patients. VIS density for each patient was calculated

by binning the genome into 100-kb bins. (**B**) Correlation coefficients for VIS profiles across different cell types and VISs shared between lineages for patients 1, 2, 3, and 6. (**C**) Sharing of VISs across time points in CD3+ samples from patients 1 and 3. (**D**) Clonal diversity of patients depends on the VCN of the graft. Metrics, including the UC50, OCI, entropy, and Chao estimator, show the same result. Patients are divided into 3 bins based on the VCN of the infused graft: Low (<0.25): patients 1 and 6; Middle (0.25 to 0.5): patients 4, 5, 8, and 9; High (>0.5): patients 2, 3, 7, and 10. Pearson correlation coefficients were calculated by correlating VCN (fig. S1C) with sample diversity.

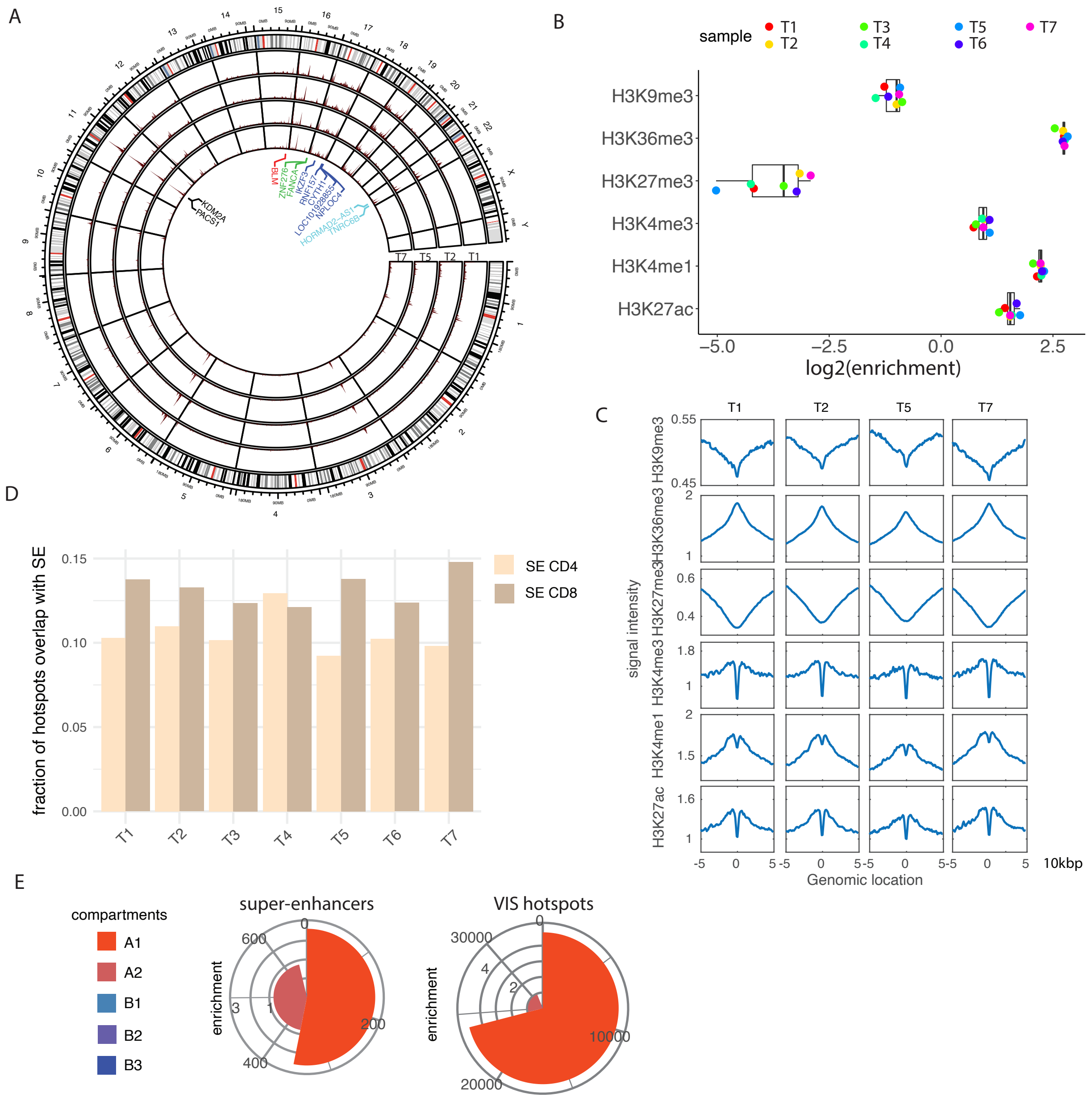


Fig. S3: Lentiviral-transduced CAR T cells share the same lentiviral integrome signatures. (**A**) Seven CAR T–cell samples were analyzed (patients: T1–T4; healthy donors: T5–T7). A circular projection of the human genome with the density of integration sites from four selected CAR T–cell samples is shown (T1, T2: patients; T5, T7: healthy donors). (**B**) Enrichment of VISs in six histone marks. (**C**) Aggregation plots showing the signal intensity of the six histone marks in CD3+ cells at the VIS and its flanking regions. (**D**) Hotspots in CAR T–cell samples overlap with SEs. Hotspots identified in samples sorted for CD4 and CD8 were matched with SEs in CD4+ and CD8+ cells. (**E**) Mapping of SE and VIS hotspots in CD3+ T cells from CAR T–cell samples to the five genome sub-compartments.

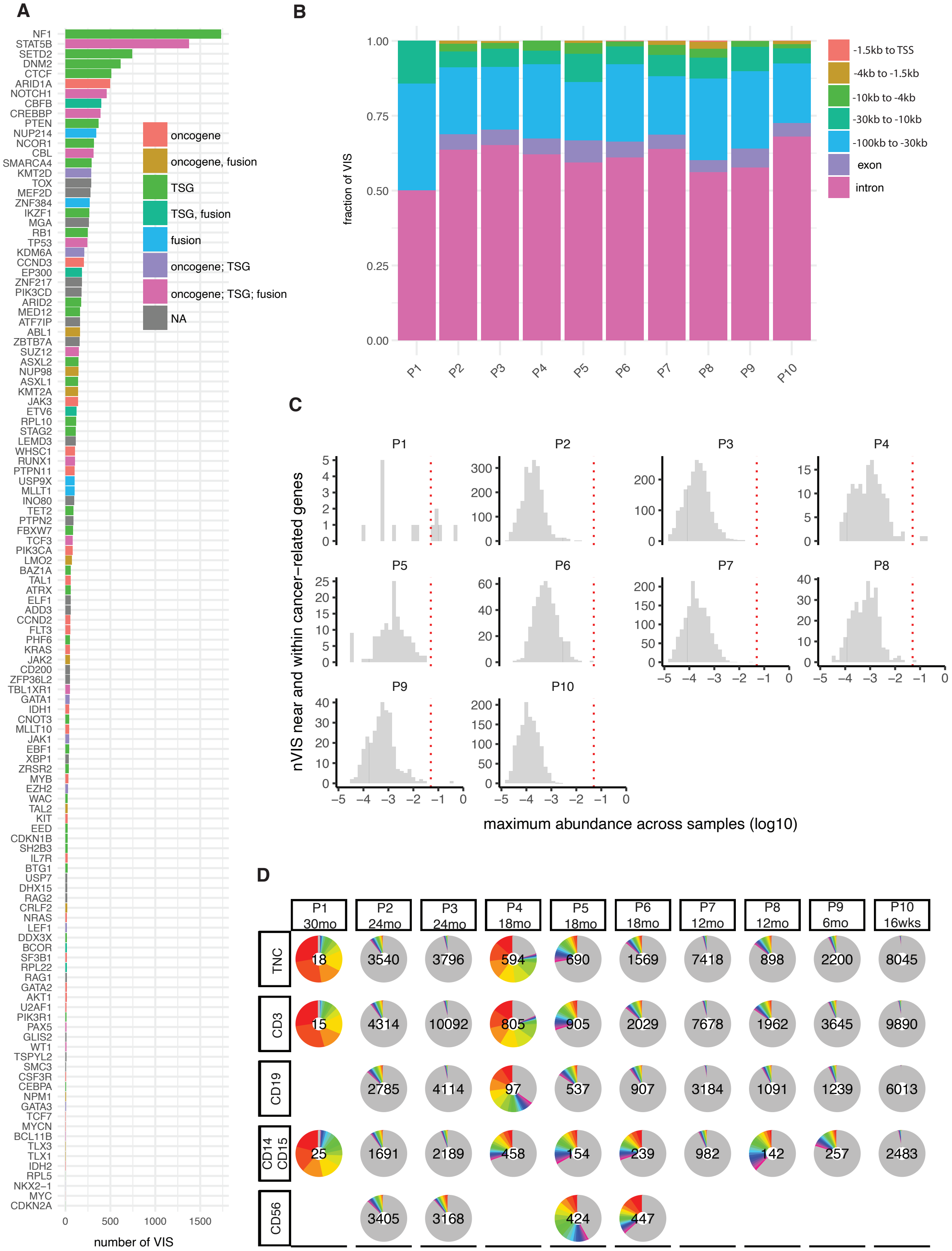


Fig. S4: VISs near cancer genes. (A) Number of vector integration sites within or near a set of

cancer-related genes that are associated with TALL, AML, BALL, or MDS. Vector integration sites profiled in all patient samples within the genes or located up to 300kb of the TSS were counted. (**B**) Location of vector integration sites within or near the cancer-related genes relative to TSS. VISs are rarely found in the promoters. (**C**) Relative abundance of vector integration sites within or near the cancer-related genes. For each VIS, only the maximum relative abundance across all available samples is shown. The red line indicates the cutoff 0.05. No genes related VIS (except one in patient 1) has abundance higher than the cutoff. (**D**) VIS frequency in sorted peripheral blood. Each pie chart shows the relative proportion of the top 20 most frequent VIS within the indicated sample as various colored sections. The grey regions show the remaining proportion of all other less frequent VIS. The total number of VIS detected in each sample is shown at the center of each pie. Columns are labeled by the patients. For each patient, we show the samples obtained at the latest time point. The rows are labeled by the sorting markers used for cell isolation; TNC: total nucleated cells.

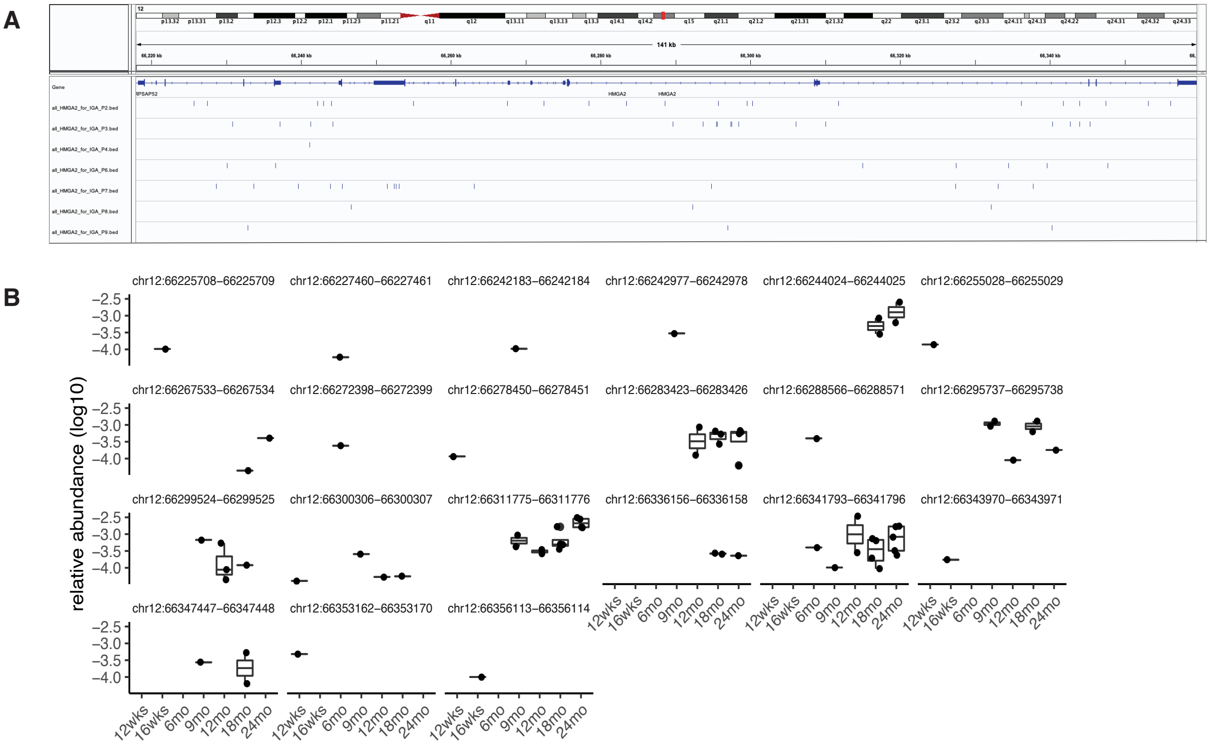


Fig. S5: Vector integration sites near the gene *HMGA2*. (**A**) Genome browser view of vector integration sites near the gene *HMGA2* for individual patients. (**B**) Relative abundance of 21 VISs of patient 2 within the gene *HMGA2* across all time points. Multiple samples at a single time point were displayed in a boxplot. There is no evidence of clonal proliferation in all the 21 VISs.

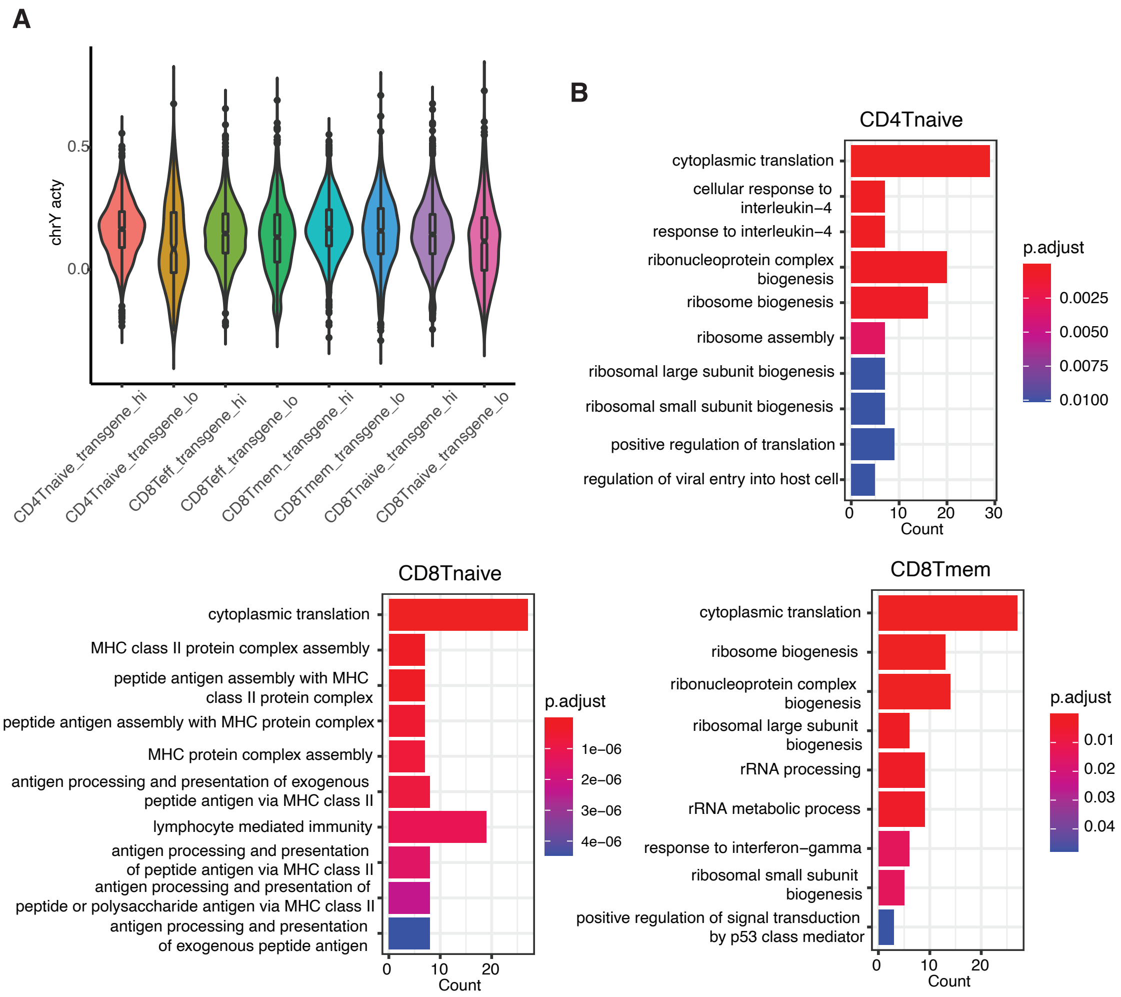
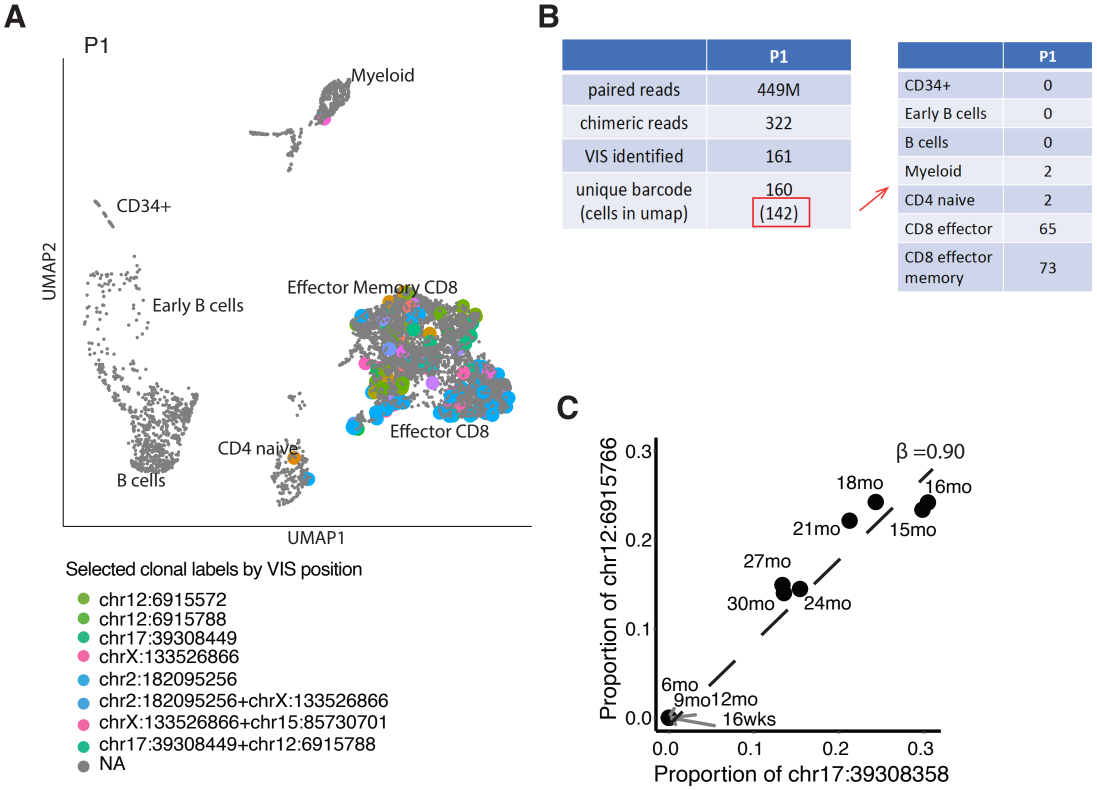


Fig. S6: Clusters of single cells separated by high and low transgene expression (**A**) Average expression of genes from chromosome Y. In multiple types of T cells, there is no difference in chromosome Y activity between clusters with high and low transgene expression, indicating the separation is not a result of maternal cells. (**B**) GO terms enrichment analysis for genes that are up-regulated in the transgene-high cluster as compared to the corresponding transgene-low cluster.



	P1
paired reads	449M
chimeric reads	322
VIS identified	161
unique barcode (cells in umap)	160 (142)

	P1
CD34+	0
Early B cells	0
B cells	0
Myeloid	2
CD4 naive	2
CD8 effector	65
CD8 effector memory	73

Fig. S7: Mapping VISs with single-cell resolution. (**A**) Cells with VISs mapped by scATAC-seq data for patient 1. Cells with VISs profiled simultaneously are marked with larger circles whose colors represent the corresponding clones. Only selected clones are labeled, mainly those with multiple cells mapped. Some clones have double and distant VISs, e.g., chr17:39308449 and chr12:6915788. (**B**) Detailed statistics of the EpiVIA pipeline. Only 142 cells had VISs sequenced by the EpiVIA pipeline. (**C**) Relative abundance of the two VIS clones in CD3+ samples from patient 1 across time. The two distant integration sites (chr12:6915766, chr17:39308358) are highly correlated. Linear regression shows their ratio to be roughly 1:1 (linear coefficient: β = 0.90).

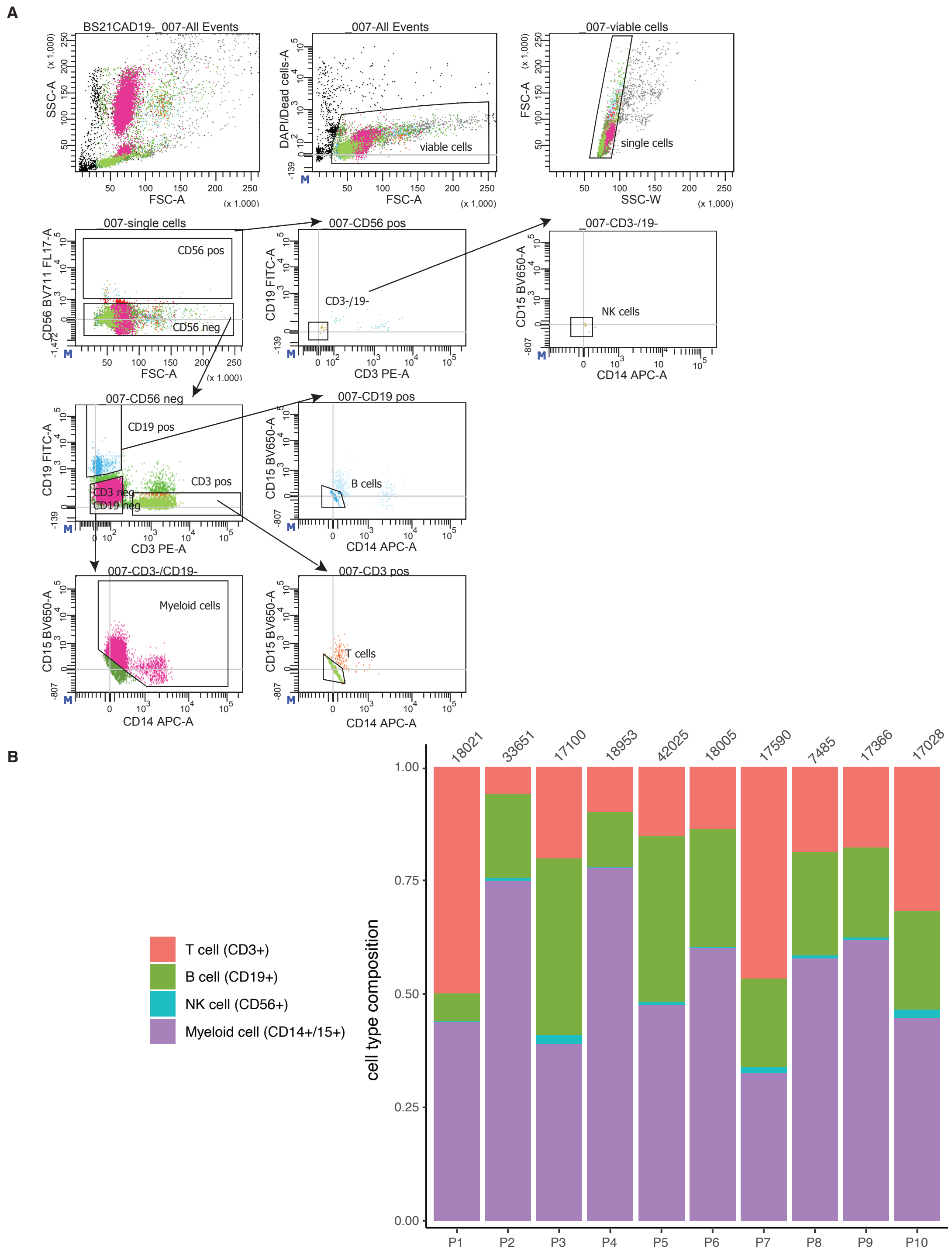


Fig. S8: Cell type composition in patients' bone marrow samples via flow cytometry. (A) Flow cytometry workflow for bone marrow samples of P1. **(B)** Composition of T cells. B cells, NK cells and Myeloid cells in patients' bone marrow. The numbers at the top of the bars represent the total number of cells.

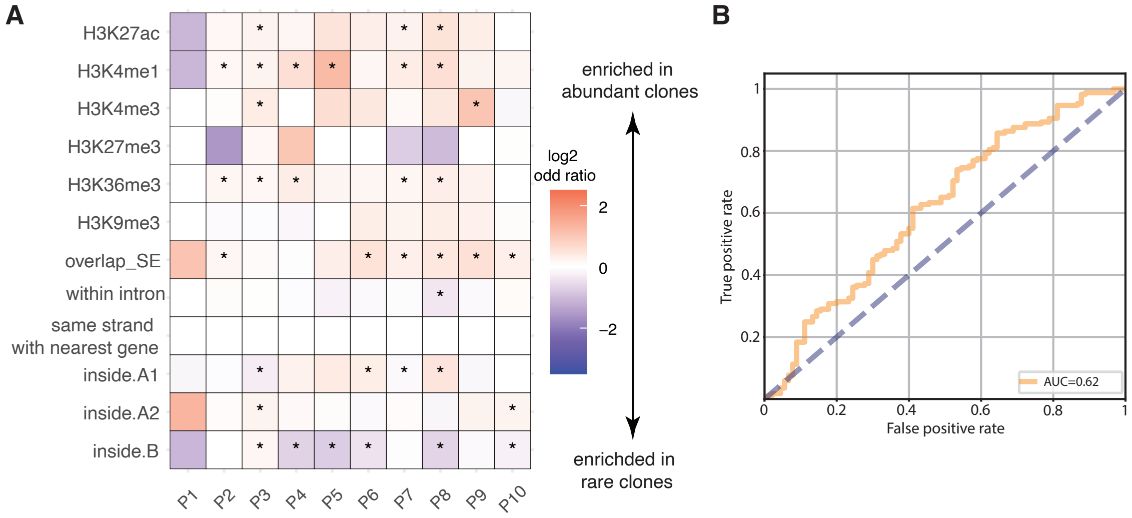


Fig. S9: LV integrome signatures have limited predictive power on clonal population. (**A**) Some integrome signatures weakly distinguish clones with high and low abundance. The features listed are binary (present or absent for a given VIS). High- and low-abundance clones with or without a particular feature were counted, and the log2 odds ratio was calculated. Features enriched among the high-abundance clones have a positive log odds ratio (red), and features enriched among the low-abundance clones have a negative log odds ratio (blue). Log-odd ratio with $P <$ 0.05 (Fisher's exact test) are marked by *. **B**) ROC curve of the classification model. The model was trained to classify abundant and rare clones by using integrome signatures. The predictive power of signatures is rather limited (AUC=0.62).

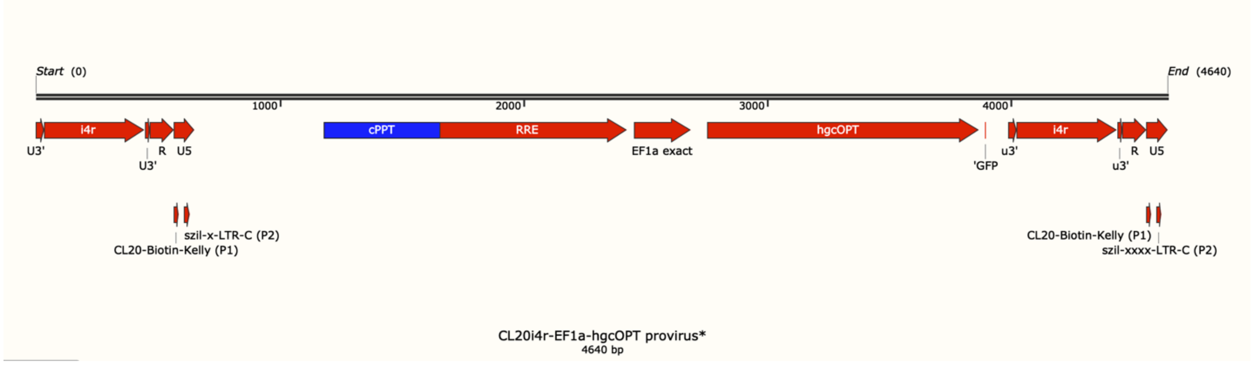
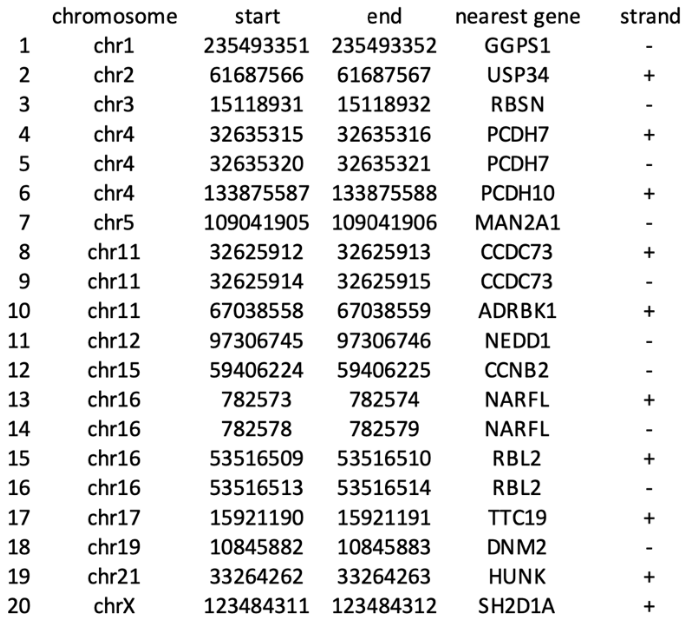


Fig. S10. Structure of CL20-i4-EF1α-hγc-OPT vector. The *IL2RG* transgene in located from 2758 to 3867 bp.



	chromosome	start	end	nearest gene	strand
1	chr1	235493351	235493352	GGPS1	-
2	chr2	61687566	61687567	USP34	+
3	chr3	15118931	15118932	RBSN	-
4	chr4	32635315	32635316	PCDH7	+
5	chr4	32635320	32635321	PCDH7	-
6	chr4	133875587	133875588	PCDH10	+
7	chr5	109041905	109041906	MAN2A1	-
8	chr11	32625912	32625913	CCDC73	+
9	chr11	32625914	32625915	CCDC73	-
10	chr11	67038558	67038559	ADRBK1	+
11	chr12	97306745	97306746	NEDD1	-
12	chr15	59406224	59406225	CCNB2	-
13	chr16	782573	782574	NARFL	+
14	chr16	782578	782579	NARFL	-
15	chr16	53516509	53516510	RBL2	+
16	chr16	53516513	53516514	RBL2	-
17	chr17	15921190	15921191	TTC19	+
18	chr19	10845882	10845883	DNM2	-
19	chr21	33264262	33264263	HUNK	+
20	chrX	123484311	123484312	SH2D1A	+

Fig. S11. List of 20 vector integration sites in the Jurkat cell clone for pipeline calibration.

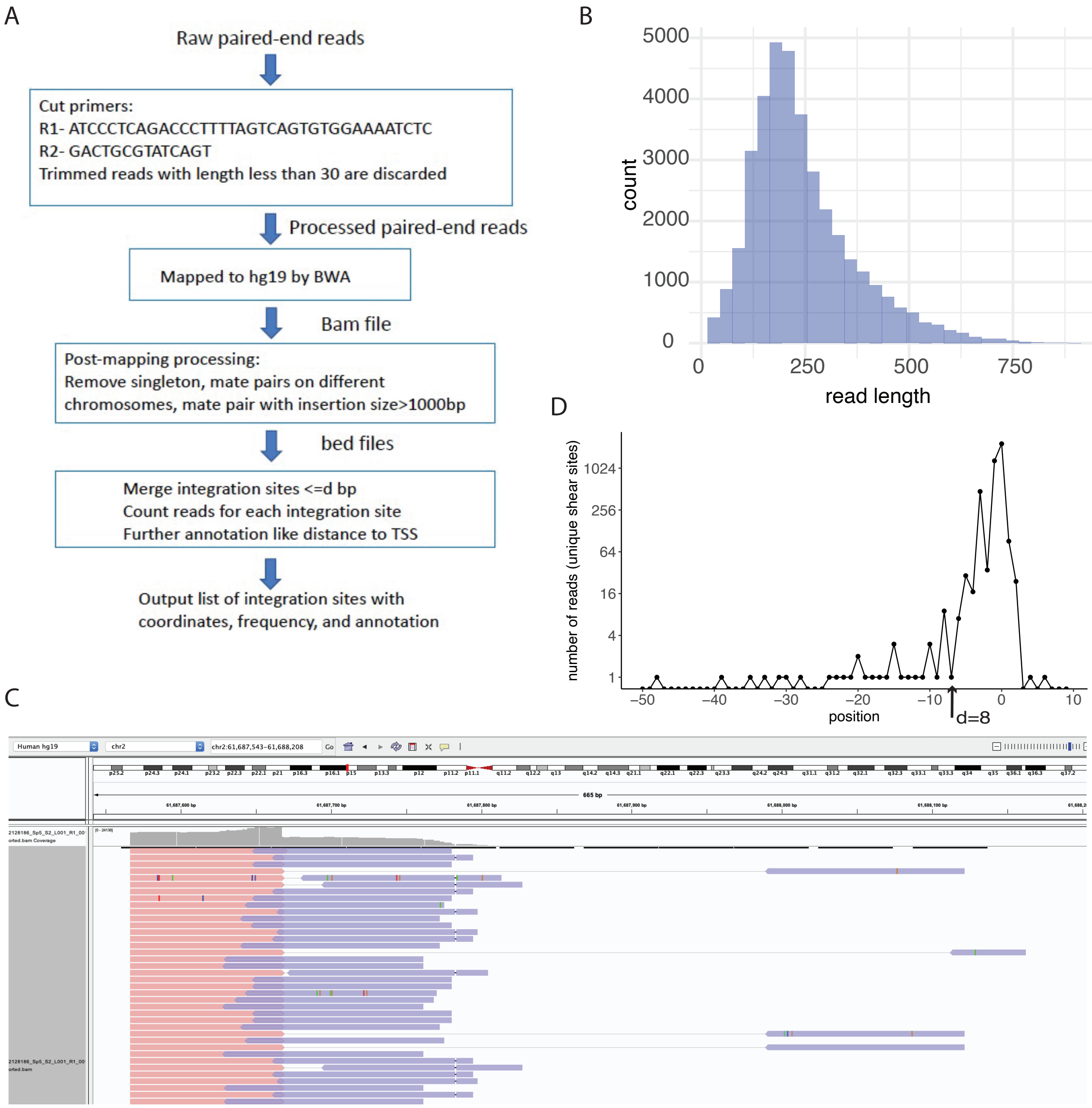


Fig. S12. Bioinformatics pipeline. (A) Overview of the bioinformatics pipeline. **(B)** Distribution of read length after post-mapping processing. Reads were collected from profiling of multiple samples of Jurkat cell line containing the DNA from the transduced clone. Only non-duplicated reads are shown. **(C)** IGV screenshot showing reads corresponding to a vector integration site. PE reads sharing the same vector integration site (chr2:61687566+), with R1 (red) all aligned at the same position. R2 (blue) ended at different positions because of random shear sites. **(D)** Starting position of reads relative to the known integration sites. Majority of reads match precisely the known position, whereas most of the rest fall closely. Based on this result, '8' was chosen as the merging parameter *d*.

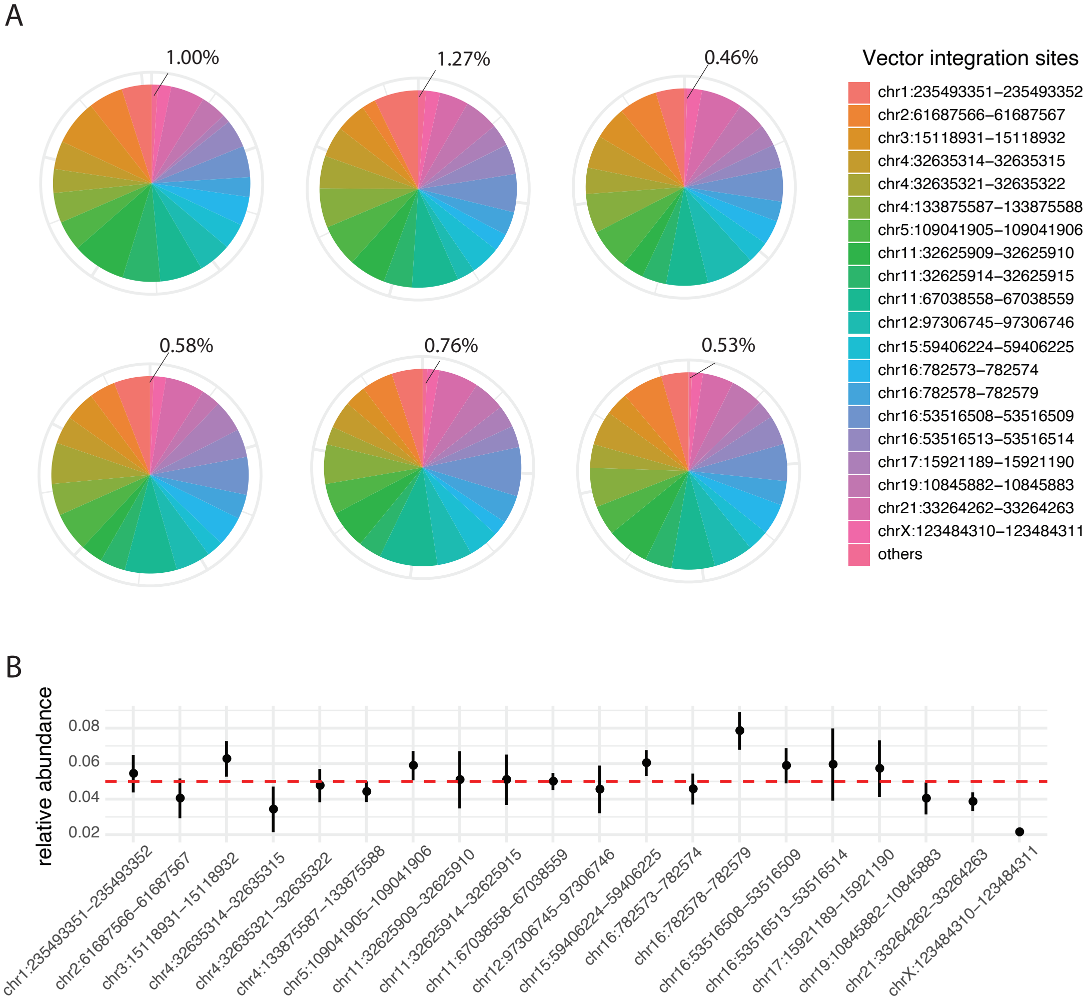
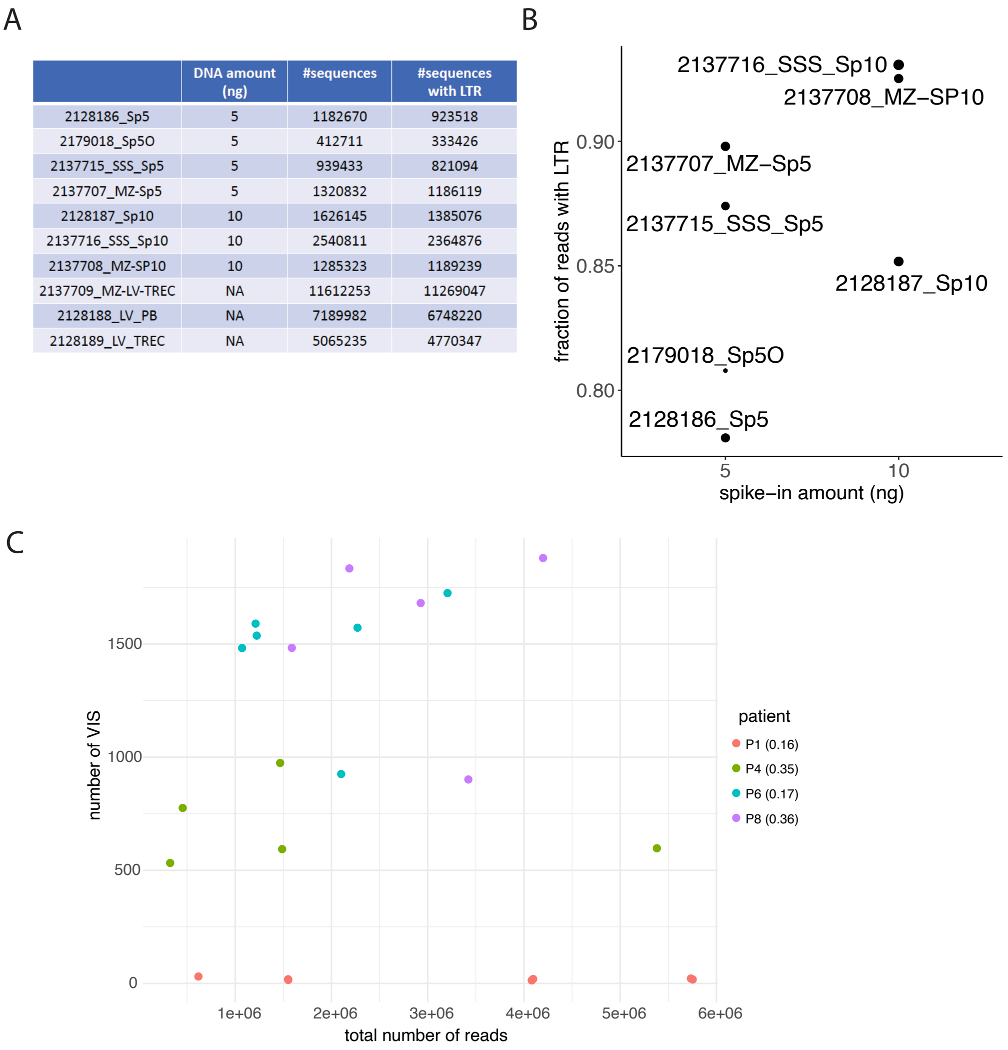


Fig. S13. Signal and noise in qsLAM assay. (A) Relative abundance of the 20 known vector integration sites in 6 samples of Jurkat cell line containing the transduced clone. The fraction of reads correspond to false discovery is around or less than 1%. **(B)** The relative abundance of the 20 sites. The dots and the error bars show the mean and standard deviation of each site across 6 samples. The theoretical mean is 0.05 (red line).



	DNA amount (ng)	#sequences	#sequences with LTR
2128186_Sp5	5	1182670	923518
2179018_Sp5O	5	412711	333426
2137715_SSS_Sp5	5	939433	821094
2137707_MZ-Sp5	5	1320832	1186119
2128187_Sp10	10	1626145	1385076
2137716_SSS_Sp10	10	2540811	2364876
2137708_MZ-SP10	10	1285323	1189239
2137709_MZ-LV-TREC	NA	11612253	11269047
2128188_LV_PB	NA	7189982	6748220
2128189_LV_TREC	NA	5065235	4770347

Fig. S14. Effects of read coverage and vector copy number. (A) Coverage depends on the amount of spike-in DNA (from the 20-sites Jurkat clone). The raw number of reads and the number of reads with LTR are shown for 10 samples. The top 7 are spike-in samples. The last 3 rows are patient samples with no spike-in. **(B)** The fraction of reads with LTR sequences depend on the amount of spike-in DNA. As the fraction is higher for the 3 patient samples, the amount of DNA corresponding to VIS is higher as compared spike-in samples. **(C)** The number of reads versus the number of VIS profiled in samples of 4 selected patients. P1 and P6 have very similar VCN (shown inside brackets), and the same for P4 and P6. All samples are PB samples, without sorting, but obtained at different time points.

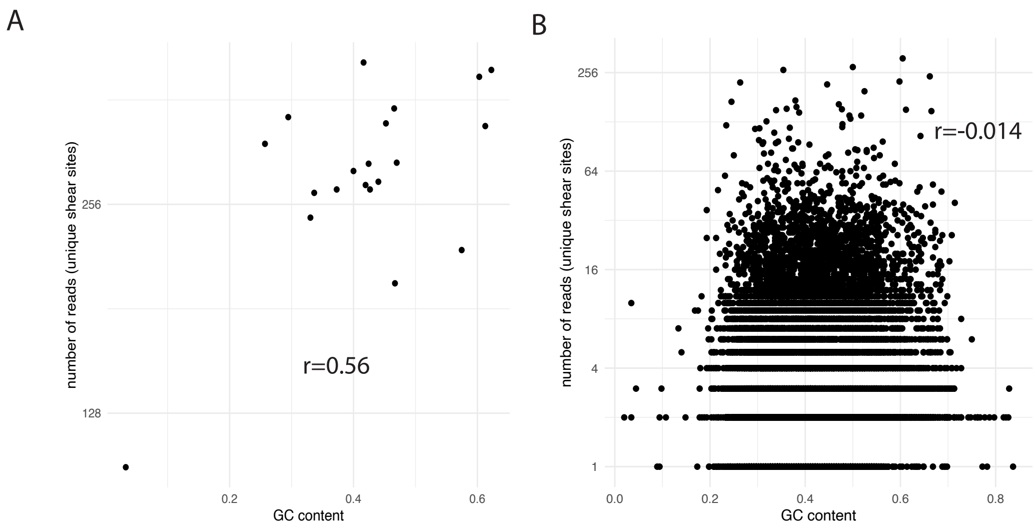


Fig. S15. Effects of GC content. (A) The number of reads associated with the 20 VIS versus their nearby GC content. Despite a strong positive correlation, it is merely caused by the outliner chrX:123484310-123484311. Without it, the correlation is only -0.18. **(B)** The number of reads associated with VIS versus their nearby GC content in a sample of P2. There is only a weak correlation of -0.014.

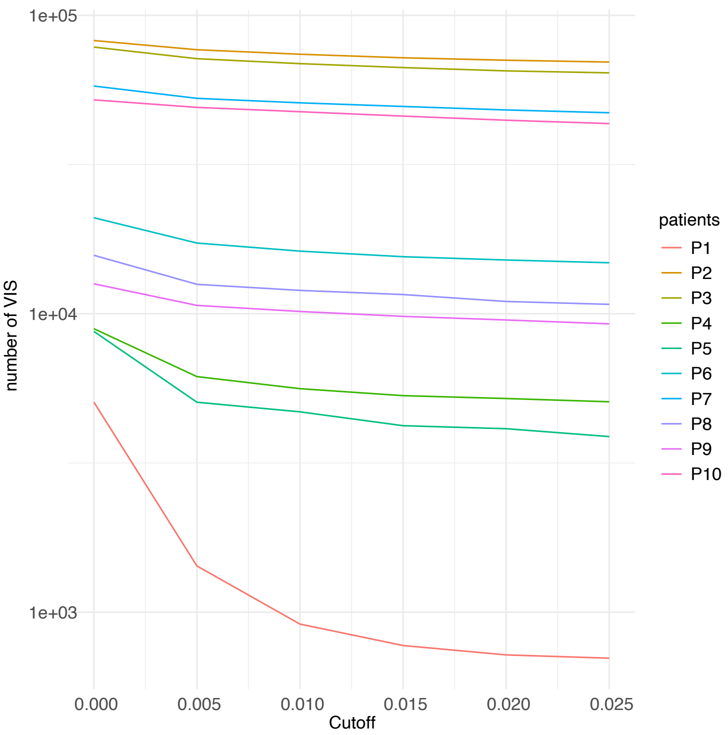


Fig. S16. Effect on the number of VIS in each patient by filtering the rare sites. The cutoff value 1.5% was employed.

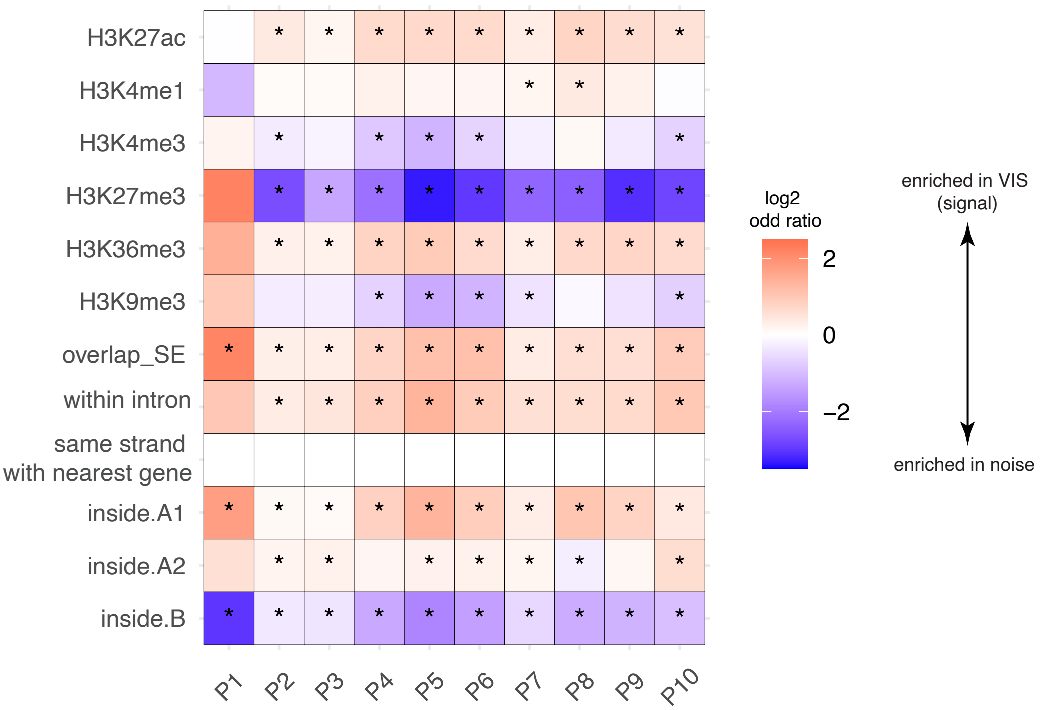


Fig. S17. **Integrome signatures distinguish potential false discoveries from real vector integration sites.** For each patient, false discoveries, for instance PCR artifacts, are filtered by the VIS calling pipeline based on the relative abundance. Integrome signatures distinguish the false discoveries from true vector integration sites. The features listed are binary (present or absent for a given VIS). True VIS and false discoveries with or without a particular feature were counted, and the log2 odds ratio was calculated. Features enriched among the true sites have a positive log odds ratio (red), and features enriched among the noise have a negative log odds ratio (blue). Log-odd ratio with $P < 0.05$ (Fisher's exact test) are marked by *. Most of the signatures are significant in most patients.

ADDITIONAL MATERIALS

Supplementary Table 1: Recurrent integration genes in all SCID-X1 patients

Supplementary Table 2: List of all ENCODE datasets used in the analysis

Supplementary Data 1: Vector integration sites and their abundance in SCID-X1 patients