Analysis of integration site distributions and relative clonal abundance for subject pin32

January 23, 2025

Contents

Sample Overview

The table below summarizes the samples analyzed in this report. “GTSP” indicates our accession numbers. The results are discussed in detail below.

GTSP	refGenome	Timepoint	CellType	TotalReads	InferredCells	UniqueSites
LS-13	hg38	D-1	TCELL	9,144	1,759	1,759
GTSP4698	hg38	D1	PBMC	187,157	58	53
GTSP4707	hg38	W2	PBMC	202,534	162	157
GTSP3211	hg38	W8	PBMC	245,025	36	21
GTSP3212	hg38	W8	PBMC	34	18	9
GTSP4716	hg38	W8	PBMC	318,683	324	312
LS-14	hg38	M2	TCELL	3,302	42	32
GTSP4725	hg38	Y4	PBMC	221,426	47	40
LS-15	hg38	Y7	TCELL	2,642	7	5
GTSP4734	hg38	Y8	PBMC	91,603	47	46
GTSP3213	hg38	Y8.5	PBMC	110,212	14	8
GTSP4741	hg38	Y15	PBMC	256,575	28	24
GTSP3214	NA	y8.5	PBMC	NA	NA	NA

Clonal expansion summary

Is there a rich population of progenitor cells delivering mature cells to the periphery?

To provide a simple measure, we ask whether there are ≥ 1000 descendants of independent progenitors (i.e. unique integration sites) in minimally fractionated cell specimens (PBMC and WHOLE BLOOD). Cell specimens that pass these criteria are operationally designated Rich.

Time point	PBMC	Rich
D1	53	No
W2	157	No
W8	312	No
Y4	40	No
Y8	46	No
Y8.5	8	No
Y15	24	No

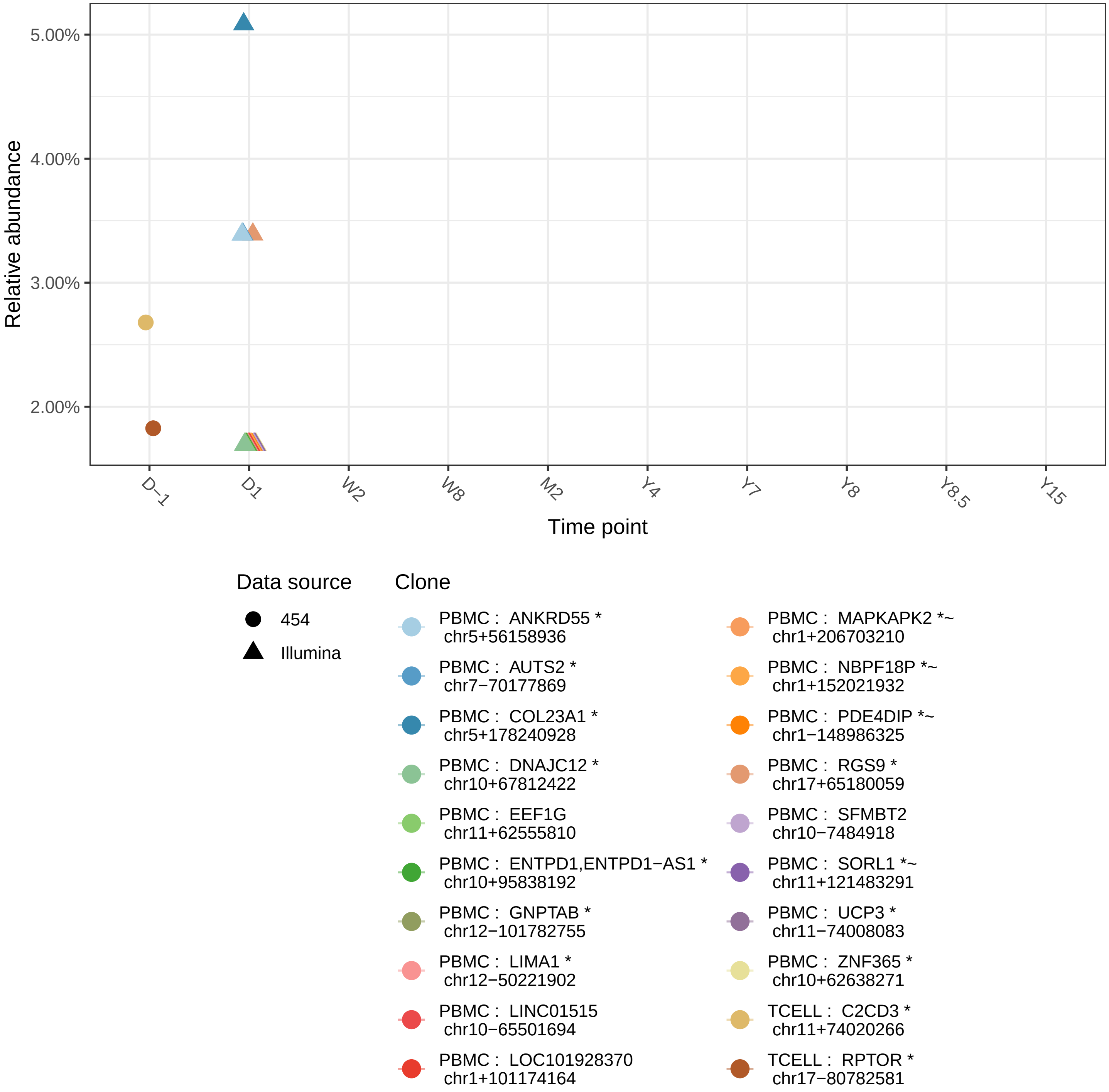
Do any cell clones account for more than 20% of all clones?

For some trials, a reporting criteria is whether any cell clones expand to account for greater than 20% of all clones. The table below highlights samples with relative abundances $\geq$ 20% considering only samples with 50 or more inferred cells.

No clones exceed 20% in any samples.

Are any cell clones increasing in proportion over time?

The plot below details the longitudinal sample relative abundances of the most abundant 20 clones where only samples with 50 or more inferred cells are considered.



Introduction

The attached report describes results of analysis of integration site distributions and relative abundance for samples from gene therapy trials. For cases of gene correction in circulating blood cells, it is possible to harvest cells sequentially from blood to monitor cell populations. Frequency of isolation information can provide information on the clonal structure of the population. This report summarizes results for subject pin32 over time points D-1, D1, M2, W2, W8, Y4, Y7, Y8, Y8.5, Y15 in UCSC genome draft .

The samples studied in this report, the numbers of sequence reads, recovered integration vectors, and unique integration sites available for this subject are shown below. We quantify population clone diversity using Gini coefficients, Shannon index, and UC50. The Gini coefficient provides a measure of inequality in clonal abundance in each sample. The coefficient equals zero when all sites are equally abundant (polyclonal) and increases as fewer sites account for more of the total (oligoclonal). Shannon index is another widely used measure of diversity and it accounts for both abundance and evenness of the integration events. Alternatively, the UC50 is the number of unique clones which make up the top 50% of the sample's abundance. For polyclonal samples, one may expect a low Gini coefficient, high Shannon Index, and high UC50 (proportional to the total number of unique sites identified in the sample).

Under most circumstances only a subset of sites will be sampled. We thus include an estimate of sample size based on frequency of isolation information from the SonicLength method (Berry, 2012). The 'S.chao1' column denotes the estimated lower bound for population size derived using Chao estimate (Chao, 1987). If sample replicates were present then estimates were subjected to jackknife bias correction.

We estimate the numbers of cell clones sampled using the SonicLength method (Berry, 2012); this is summarized in the column "Inferred cells". Integration sites were recovered using ligation mediated PCR after random fragmentation of genomic DNA, which reduces recovery biases compared with restriction enzyme cleavage. Relative abundance was not measured from read counts, which are known to be inaccurate, but from marks introduced into DNA specimens prior to PCR amplification using the SonicLength method PMID:22238265.

We quantify population diversity using Gini coefficients, Shannon index, and UC50. The Gini coefficient provides a measure of inequality in clonal abundance in each sample. The coefficient equals zero when all sites are equally abundant (polyclonal) and increases as fewer sites account for more of the total (oligoclonal). Shannon index is another widely used measure of diversity and it accounts for both abundance and evenness of the integration events. UC50 is the number of clones which make up the top 50% of the sample's abundance. For polyclonal samples, one may expect a low Gini coefficient, high Shannon Index, and high UC50 (proportional to the total number of unique sites identified in the sample).

Integration positions are reported with the format (nearest gene, chromosome, +/-, genomic position) where the nearest gene is the nearest transcriptional boundary to the integration position, '+' refers to integration in the positive orientation and '-' refers to integration in the reverse orientation. Reported distances are signed where where the sign indicates if integrationsare upstream (-) or downstream (+, no sign) of the nearest gene. Nearest genes possess additional annotations described in the table below.

Symbol	Meaning
*	site is within a transcription unit
~	site is within 50kb of a cancer related gene
!	nearest gene was assocaited with lymphoma in humans

Sample Summary

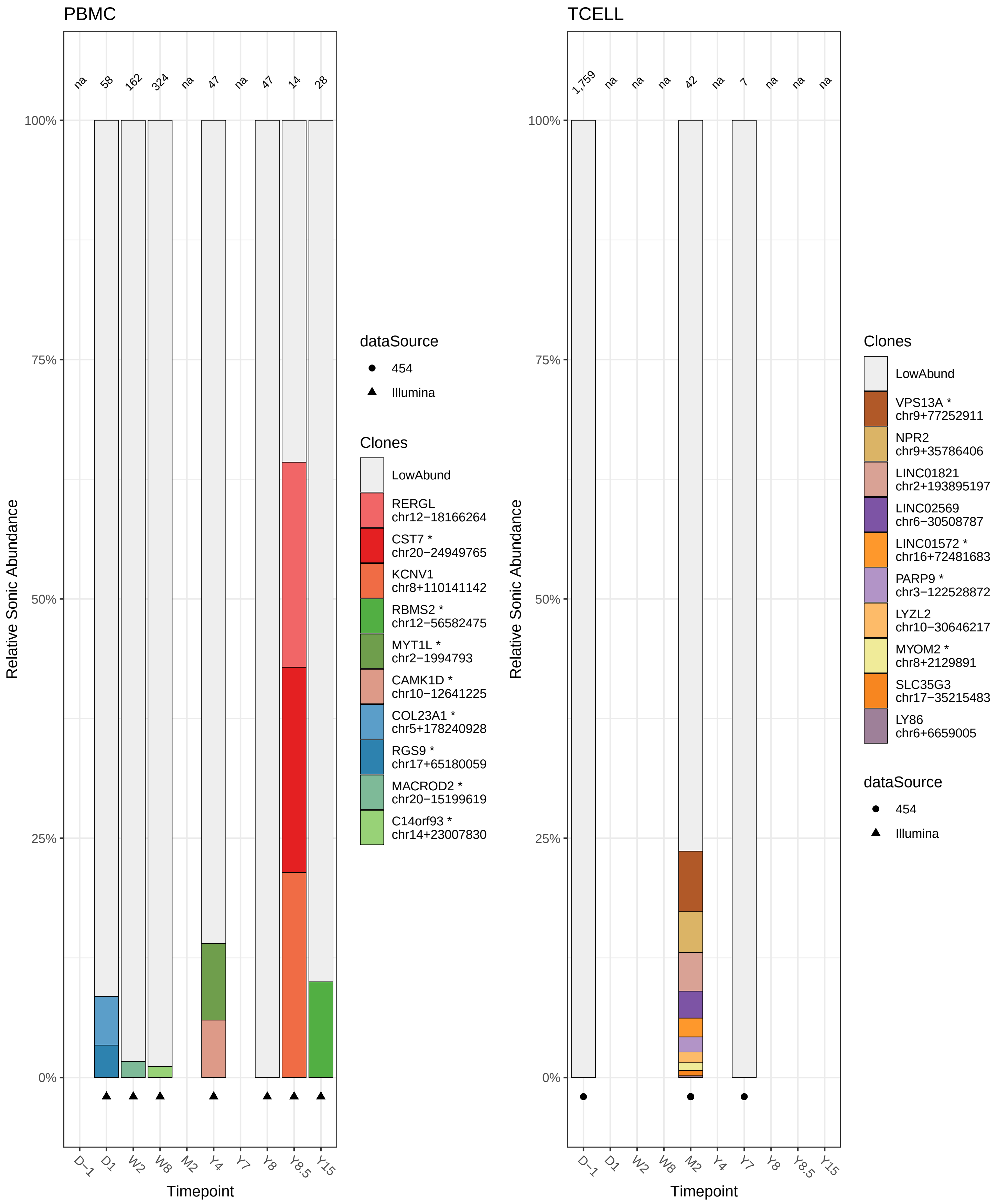
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	refGenome	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
LS-13	454	hg38	D-1	TCELL	9,144	1,759	1,759	0.000	1,547,920	7.47	1.000	880	yes	NA	NA
GTSP4698	Illumina	hg38	D1	PBMC	187,157	58	53	0.081	347	3.93	0.990	25	yes	2022-04-05	NA
GTSP4707	Illumina	hg38	W2	PBMC	202,534	162	157	0.030	3,064	5.04	0.997	77	yes	2022-04-05	NA
GTSP3211	Illumina	hg38	W8	PBMC	245,025	36	21	0.294	40	2.89	0.948	6	no	2021-01-24	NA
GTSP3212	Illumina	hg38	W8	PBMC	34	18	9	0.284	11	2.06	0.938	3	no	2021-01-24	NA
GTSP4716	Illumina	hg38	W8	PBMC	318,683	324	312	0.036	4,857	5.73	0.997	151	yes	2023-03-01	NA
LS-14	454	hg38	M2	TCELL	3,302	42	32	0.164	53	3.41	0.983	12	yes	NA	NA
GTSP4725	Illumina	hg38	Y4	PBMC	221,426	47	40	0.138	250	3.60	0.977	17	yes	2022-04-05	NA
LS-15	454	hg38	Y7	TCELL	2,642	7	5	0.171	6	1.55	0.963	2	yes	NA	NA
GTSP4734	Illumina	hg38	Y8	PBMC	91,603	47	46	0.021	541	3.82	0.998	23	yes	2023-01-26	NA
GTSP3213	Illumina	hg38	Y8.5	PBMC	110,212	14	8	0.268	18	1.93	0.929	3	yes	2021-01-24	NA
GTSP4741	Illumina	hg38	Y15	PBMC	256,575	28	24	0.128	94	3.12	0.980	11	yes	2022-04-05	NA
GTSP3214	NA	NA	y8.5	PBMC	NA	NA	NA	NA	NA	NA	NA	NA	NA	2021-01-24	NA

Tracking of clonal abundances

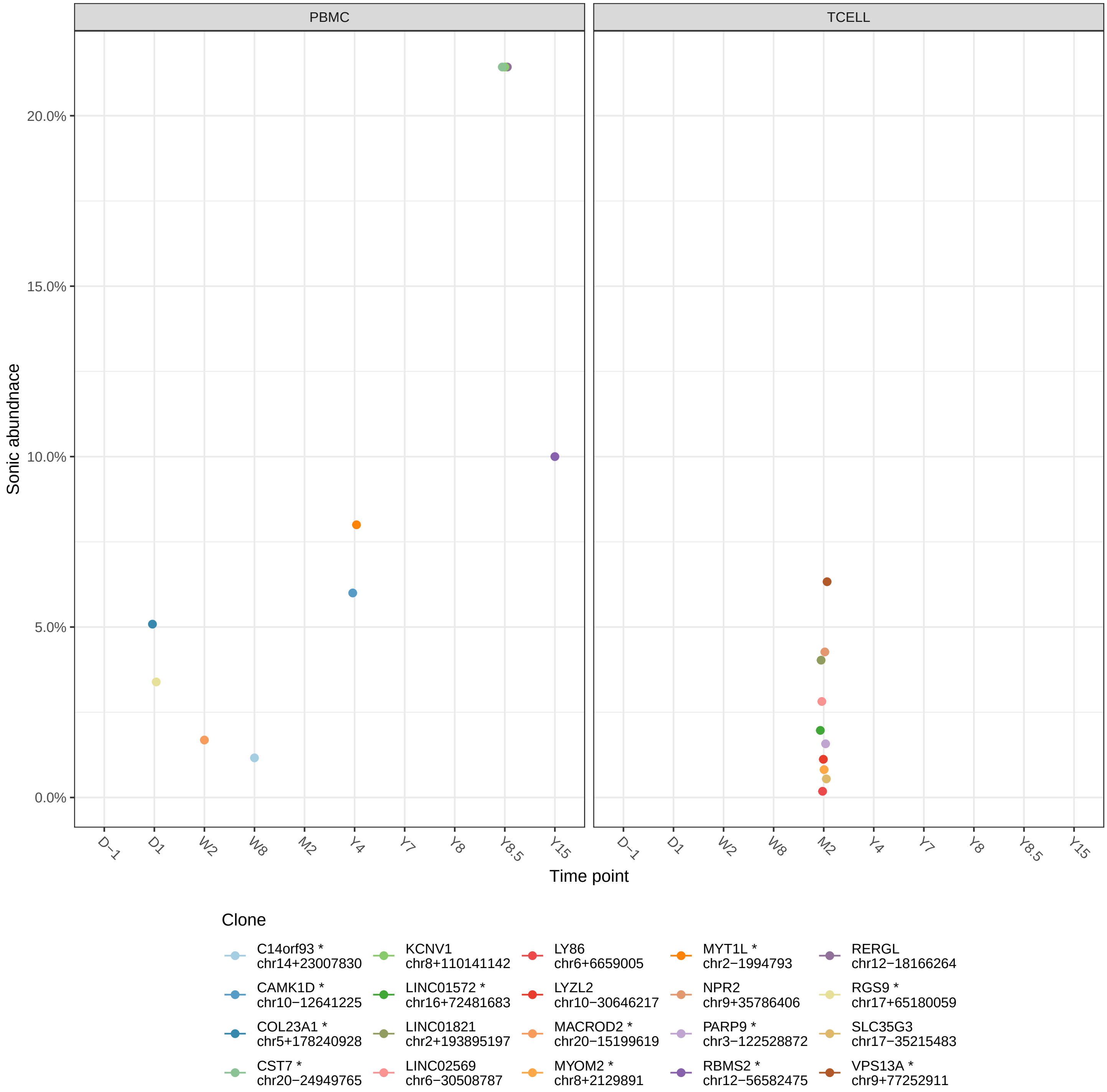
Relative abundance of cell clones

The relative abundances of cell clones is summarized in the stacked bar plots below. The cell fraction studied is named at the top of each plot and the time points are marked at the bottom. The different bars in each panel show the major cell clones, as marked by integration sites where the x-axis indicates time points and the y-axis is scaled by proportion of the total cells sampled. The top 10 most abundant clones from each cell type have been named by the nearest gene while the remaining sites are binned as low abundance (LowAbund; grey). The total number of genomic fragments used to identify integration sites are listed atop of each plot. These fragments are generated by restriction endonucleases in 454 sequencing experiments and by sonic shearing in Illumina sequencing experiments. Relative abundances are calculated using the total number of reads associated with clones in 454 sequencing experiments while the number of unique sonic breaks is used in Illumina sequencing experiments.


PBMC
TCELL
na
58
162
324
na
47
na
47
14
28
1,759
na
na
na
42
na
7
na
na
na
100%
75%
50%
25%
0%
Relative Sonic Abundance
D−1
D1
W2
W8
M2
Y4
Y7
Y8
Y8.5
Y15
Timepoint
dataSource
● 454
▲ Illumina
Clones
LowAbund
RERGL chr12−18166264
CST7 * chr20−24949765
KCNV1 chr8+110141142
RBMS2 * chr12−56582475
MYT1L * chr2−1994793
CAMK1D * chr10−12641225
COL23A1 * chr5+178240928
RGS9 * chr17+65180059
MACROD2 * chr20−15199619
C14orf93 * chr14+23007830
Clones
LowAbund
VPS13A * chr9+77252911
NPR2 chr9+35786406
LINC01821 chr2+193895197
LINC02569 chr6−30508787
LINC01572 * chr16+72481683
PARP9 * chr3−122528872
LYZL2 chr10−30646217
MYOM2 * chr8+2129891
SLC35G3 chr17−35215483
LY86 chr6+6659005
dataSource
● 454
▲ Illumina

Longitudinal behavior of major clones

When multiple time points are available, it is of interest to track the behavior of the most abundant clones across different cell types. A plot of the relative abundances of the most abundant 20 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



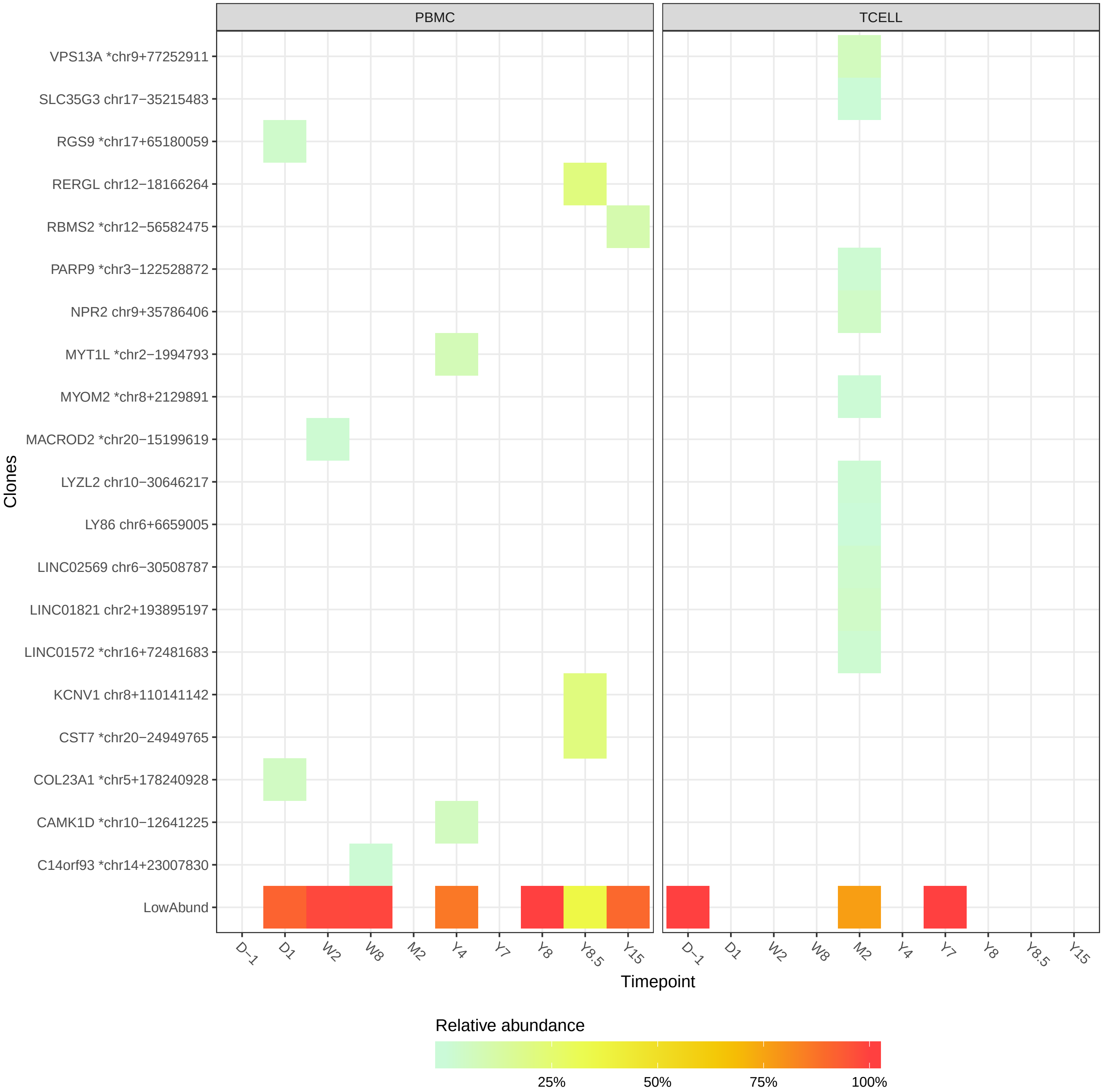
Integration sites near particular genes of interest

Integration sites near genes that have been associated with adverse events are of particular interest. Below are longitudinal relative abundance plots that focus on the most abundant 5 clones whoes nearest genes are LMO2, IKZF1, CCND2, HMGA2, and MECOM.

No integration sites were found near LMO2, IKZF1, CCND2, HMGA2 or MECOM

Sample relative abundance heatmap

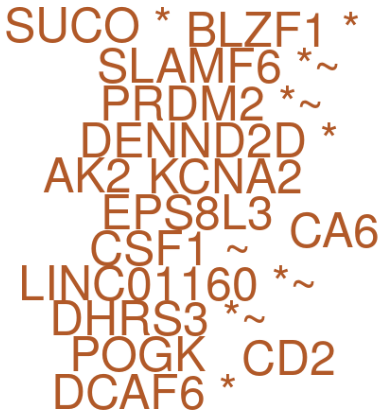
Alternatively, the relative abundances of the most abundant 10 clones from each cell sampled type can be visualized as a heat map.



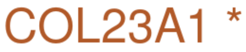
What are the most frequently occuring gene types in the subject?

The word clouds below illustrate the nearest genes of the most abundant clones from each sample where the numeric ranges represent the upper and lower clonal abundances.

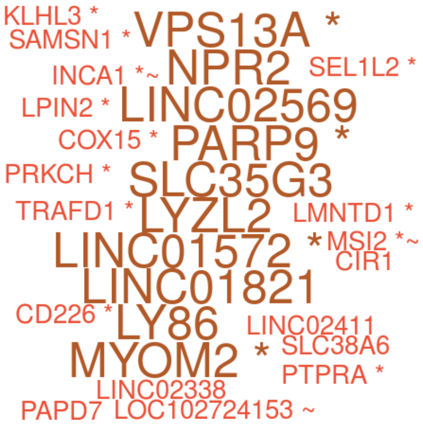
TCELL
D-1 1:1



PBMC
D1 1:3



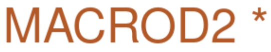
TCELL
M2 1:2

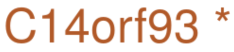


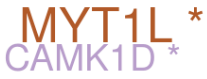
PBMC
W2 1:3

PBMC
W8 1:4

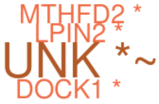
PBMC
Y4 1:4



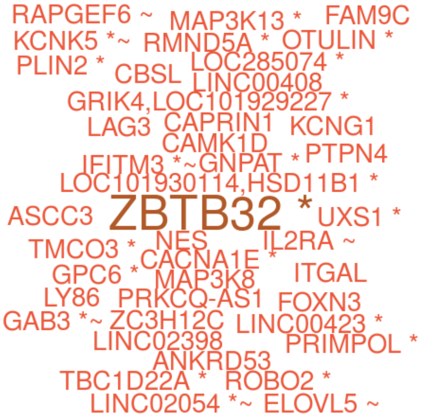




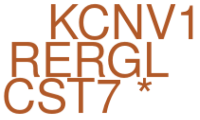
TCELL
Y7 1:2



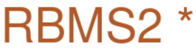
PBMC
Y8 1:2



PBMC
Y8.5 1:3



PBMC
Y15 1:3



Multihits

This analysis has been looking at integration sites that can be uniquely mapped. But it is also helpful to look at reads finding multiple equally good alignments in the genome which can be reffered to as ‘Multihits’. If an integration site occurred within a repeat element (i.e. Alus, LINE, SINE, etc), then it might be helpful to access those sites for potential detrimental effects. These collection of sequences are analyzed separately due to their ambiguity.

Sample multihits groupings with relative abundances > 20%.

Replicate	Multihit id	Total cells in replicate	Multihit relative Abundance	Celltype	Timepoint
GTSP3211-5	100696440	5	20.0%	PBMC	w8

Methods

Detailed methods can be found these publications:
- Bioinformatics. 2012 Mar 15; 28(6): 755–762.
- Mol Ther Methods Clin Dev. 2017 Mar 17; 4: 17–26.
- Mol Ther Methods Clin Dev. 2017 Mar 17; 4: 39–49.

Analysis software:
- INSPIIRED v1.1 (http://github.com/BushmanLab/INSPIIRED)

Supplementary table 1.

Greatest sample relative abundance values.

GTSP	timePoint	cellType	relAbund	posid	nearestFeature
GTSP3211	W8	PBMC	11.11%	chr11-10456605	AMPD3 *
GTSP3212	W8	PBMC	22.22%	chr12-68482061	LINC02384
GTSP3213	Y8.5	PBMC	21.43%	chr12-18166264	RERGL
GTSP4698	D1	PBMC	5.08%	chr5+178240928	COL23A1 *
GTSP4707	W2	PBMC	1.69%	chr20-15199619	MACROD2 *
GTSP4716	W8	PBMC	1.16%	chr14+23007830	C14orf93 *
GTSP4725	Y4	PBMC	8.00%	chr2-1994793	MYT1L *
GTSP4734	Y8	PBMC	4.08%	chr19-35710133	ZBTB32 *
GTSP4741	Y15	PBMC	10.00%	chr12-56582475	RBMS2 *
LS-13	D-1	TCELL	2.68%	chr11+74020266	C2CD3 *
LS-14	M2	TCELL	24.47%	chr20-3011575	PTPRA *
LS-15	Y7	TCELL	90.20%	chr8-24387356	LOC101929294,ADAMDEC1 *