Analysis of integration site distributions and relative clonal abundance for subject pin22

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
GTSP3898	hg38	D0	PBMC	76	71	64
GTSP4693	hg38	D1	PBMC	80,968	15	13
GTSP4702	hg38	W2	PBMC	346,565	91	83
GTSP4711	hg38	W8	PBMC	303,850	133	112
LS-1	hg38	M2	TCELL	2,713	14	14
GTSP4720	hg38	Y4	PBMC	233,810	12	11
GTSP4729	hg38	Y8	PBMC	266,409	23	20
GTSP4738	hg38	Y15	PBMC	51,751	9	8

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
D0	64	No
D1	13	No
W2	83	No
W8	112	No
Y4	11	No
Y8	20	No
Y15	8	No

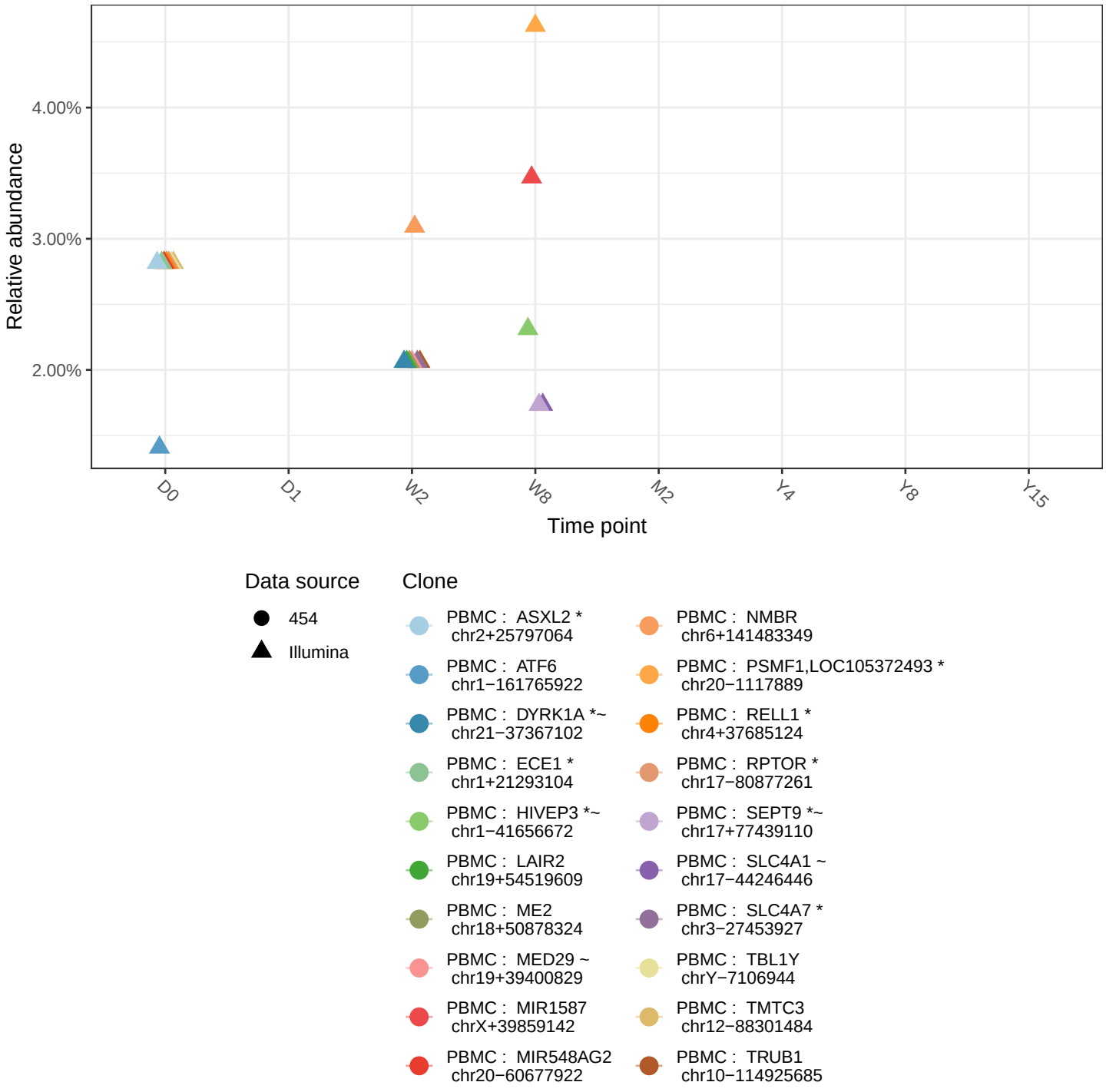
Do any cell clones account for more than 20% of all clones?

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 pin22 over time points D0, D1, M2, W2, W8, Y4, Y8, 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

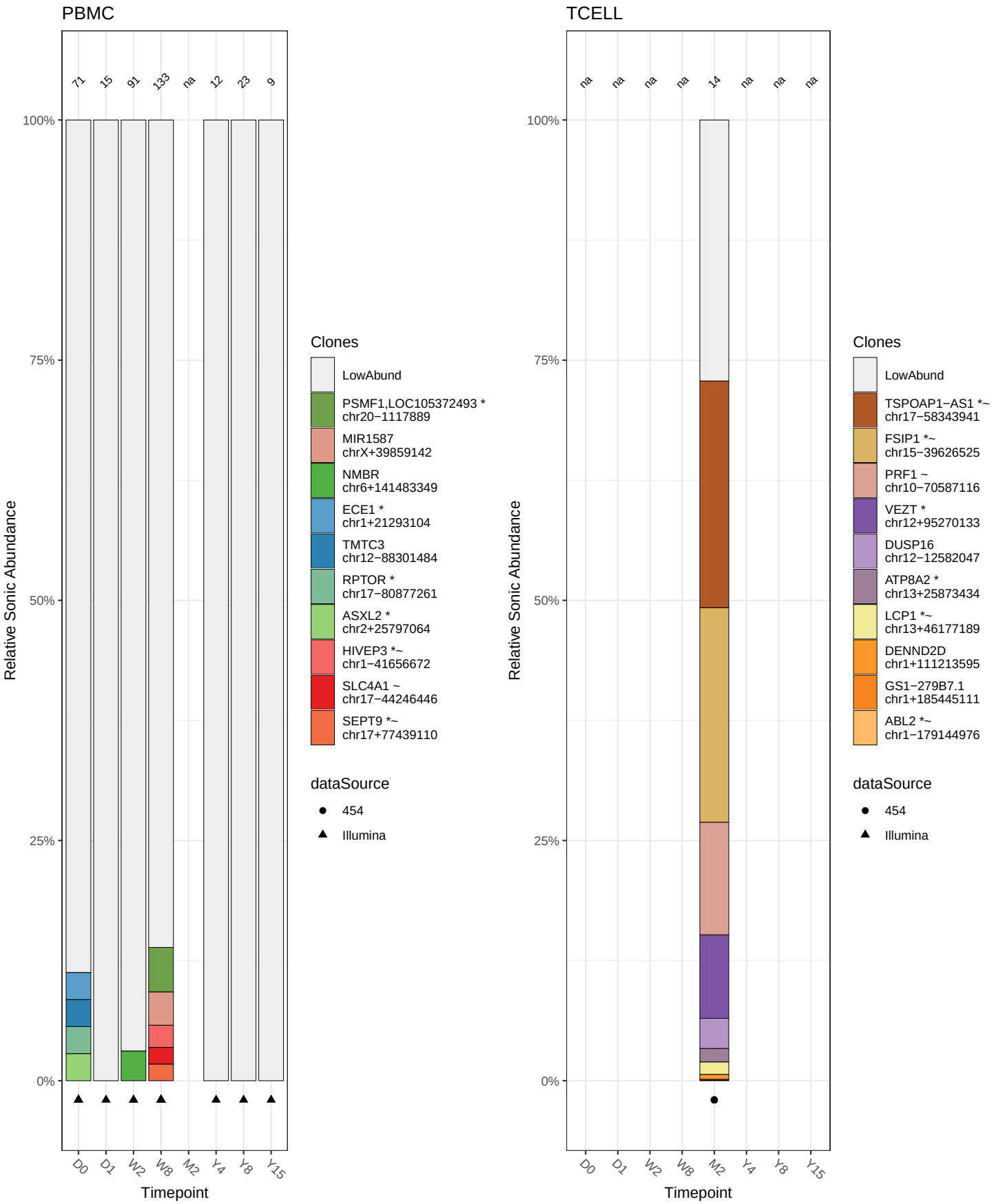
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, 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
GTSP3898	Illumina	hg38	D0	PBMC	76	71	64	0.088	264	4.13	0.992	29	yes	2021-01-24	NA
GTSP4693	Illumina	hg38	D1	PBMC	80,968	15	13	0.113	31	2.52	0.984	6	yes	2022-03-03	NA
GTSP4702	Illumina	hg38	W2	PBMC	346,565	91	83	0.081	490	4.38	0.992	38	yes	2023-01-26	NA
GTSP4711	Illumina	hg38	W8	PBMC	303,850	133	112	0.152	1,932	4.57	0.969	46	yes	2023-03-01	NA
LS-1	454	hg38	M2	TCELL	2,713	14	14	0.000	105	2.64	1.000	8	yes	NA	NA
GTSP4720	Illumina	hg38	Y4	PBMC	233,810	12	11	0.076	34	2.37	0.988	6	yes	2022-03-03	NA
GTSP4729	Illumina	hg38	Y8	PBMC	266,409	23	20	0.111	54	2.95	0.986	9	yes	2022-03-03	NA
GTSP4738	Illumina	hg38	Y15	PBMC	51,751	9	8	0.097	18	2.04	0.983	4	yes	2022-03-11	NA

Tracking of clonal abundances

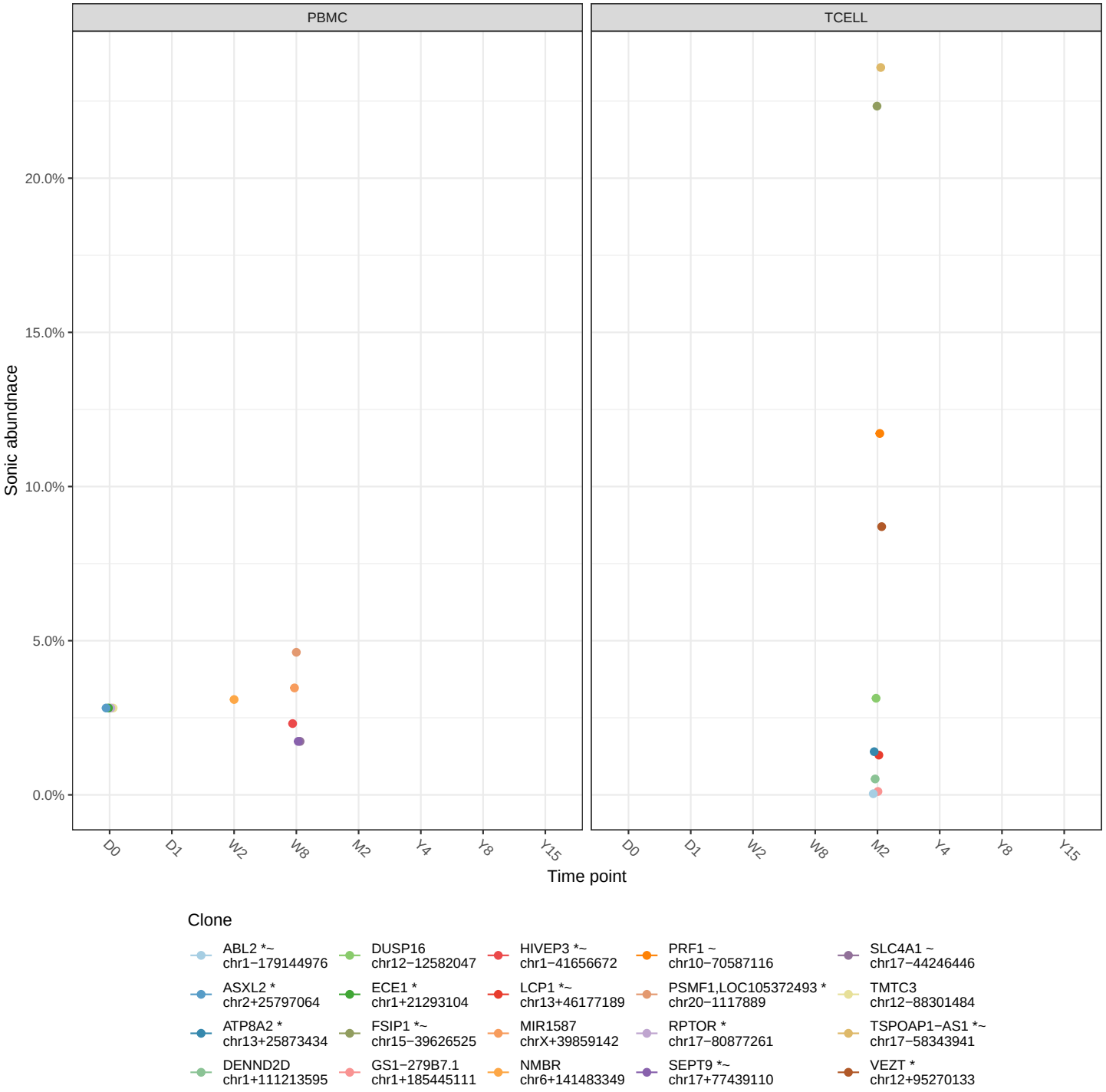
Relative abundance of cell clones

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
71
15
91
133
na
12
23
9
100%
75%
50%
25%
0%
Relative Sonic Abundance
D0
D1
W2
W8
M2
Y4
Y8
Y15
Timepoint
Clones
LowAbund
PSMF1,LOC105372493 *
chr20−1117889
MIR1587
chrX+39859142
NMBR
chr6+141483349
ECE1 *
chr1+21293104
TMTC3
chr12−88301484
RPTOR *
chr17−80877261
ASXL2 *
chr2+25797064
HIVEP3 *~
chr1−41656672
SLC4A1 ~
chr17−44246446
SEPT9 *~
chr17+77439110
dataSource
454
Illumina
TCELL
na
na
na
na
14
na
na
na
100%
75%
50%
25%
0%
Relative Sonic Abundance
D0
D1
W2
W8
M2
Y4
Y8
Y15
Timepoint
Clones
LowAbund
TSPOAP1−AS1 *~
chr17−58343941
FSIP1 *~
chr15−39626525
PRF1 ~
chr10−70587116
VEZT *
chr12+95270133
DUSP16
chr12−12582047
ATP8A2 *
chr13+25873434
LCP1 *~
chr13+46177189
DENND2D
chr1+111213595
GS1−279B7.1
chr1+185445111
ABL2 *~
chr1−179144976
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.



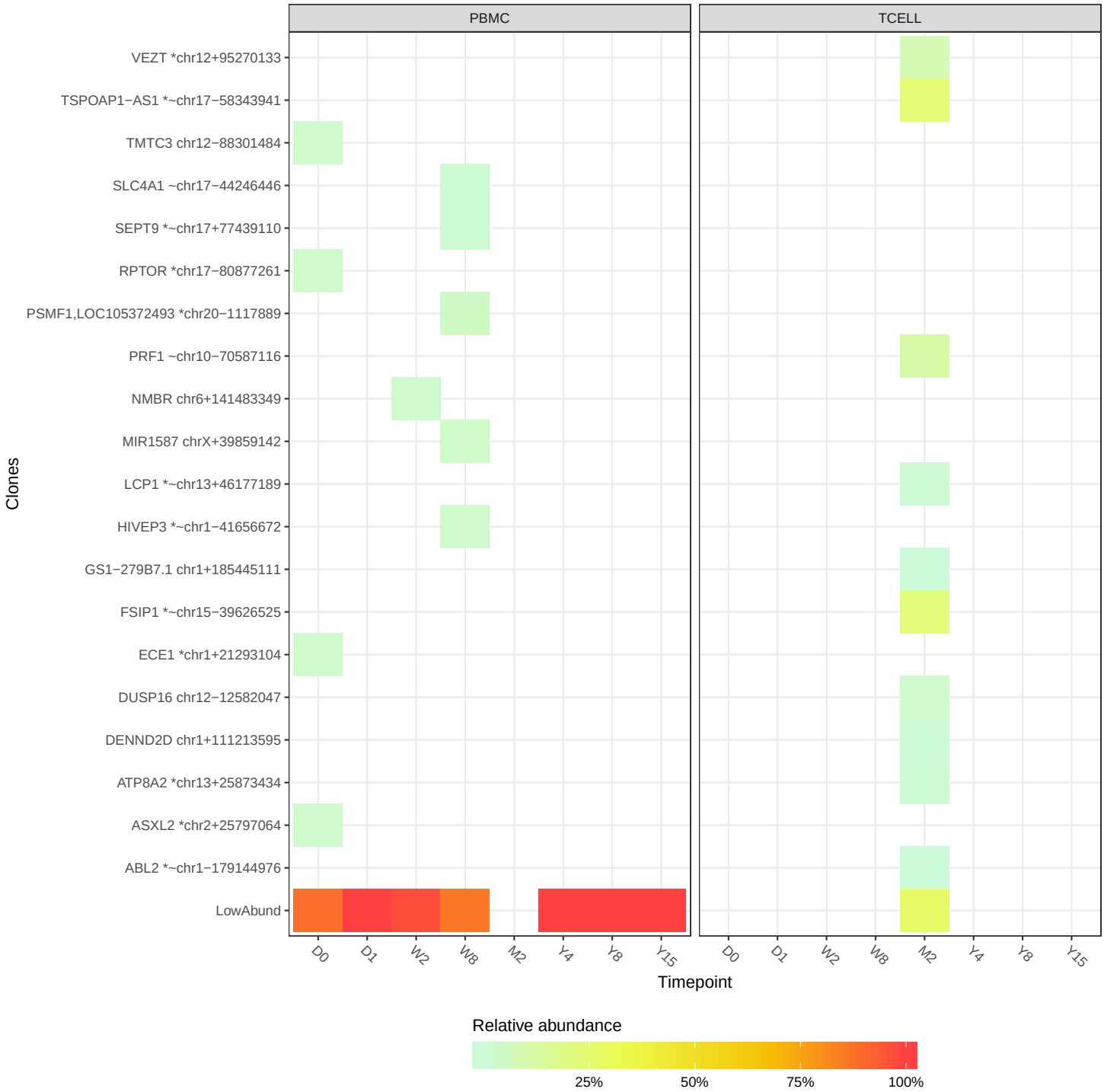
Integration sites near particular genes of interest

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.



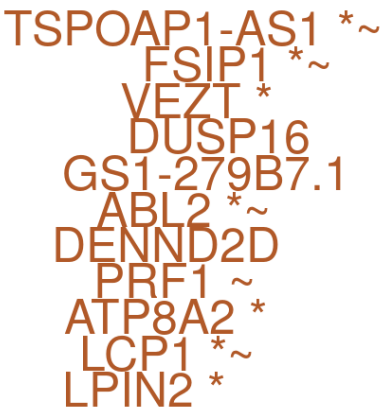
What are the most frequently occuring gene types in the subject?

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.

PBMC
D0 1:2

PBMC
D1 1:2

TCELL
M2 1:1



PBMC
W2 1:3

PBMC
W8 1:8

PBMC
Y4 1:2

PBMC
Y8 1:2

PBMC
Y15 1:2

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
GTSP4720-3	100837381	2	50.0%	PBMC	y4
GTSP4738-3	100837494	3	33.3%	PBMC	y15
GTSP4720-4	100837382	13	23.1%	PBMC	y4
GTSP4693-1	100837321	20	20.0%	PBMC	d1

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
GTSP3898	D0	PBMC	2.82%	chr1+21293104	ECE1 *
GTSP4693	D1	PBMC	13.33%	chr21+14590740	SAMSN1-AS1 *
GTSP4702	W2	PBMC	3.09%	chr6+141483349	NMBR
GTSP4711	W8	PBMC	4.62%	chr20-1117889	PSMF1,LOC105372493 *
GTSP4720	Y4	PBMC	16.67%	chr20+48826152	PREX1 *~
GTSP4729	Y8	PBMC	8.33%	chr11+96065187	MAML2 *~
GTSP4738	Y15	PBMC	22.22%	chr3-149225722	CP
LS-1	M2	TCELL	23.59%	chr17-58343941	TSPOAP1-AS1 *~