Analysis of integration site distributions and relative clonal abundance for subject pin26

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-8	hg38	D-1	TCELL	5,027	580	522
GTSP3466	hg38	D0	PBMC	574,645	1,534	1,483
GTSP4695	hg38	D1	PBMC	163,138	44	41
GTSP4704	hg38	W2	PBMC	628,753	737	714
GTSP4713	hg38	W8	PBMC	421,081	409	397
LS-9	hg38	M2	TCELL	4,383	169	169
GTSP4722	hg38	Y4	PBMC	258,765	94	77
GTSP4731	hg38	Y8	PBMC	94,405	82	47
GTSP4740	hg38	Y15	PBMC	360,381	217	192

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 $\geq$ 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	1,483	Yes
D1	41	No
W2	714	No
W8	397	No
Y4	77	No
Y8	47	No
Y15	192	No

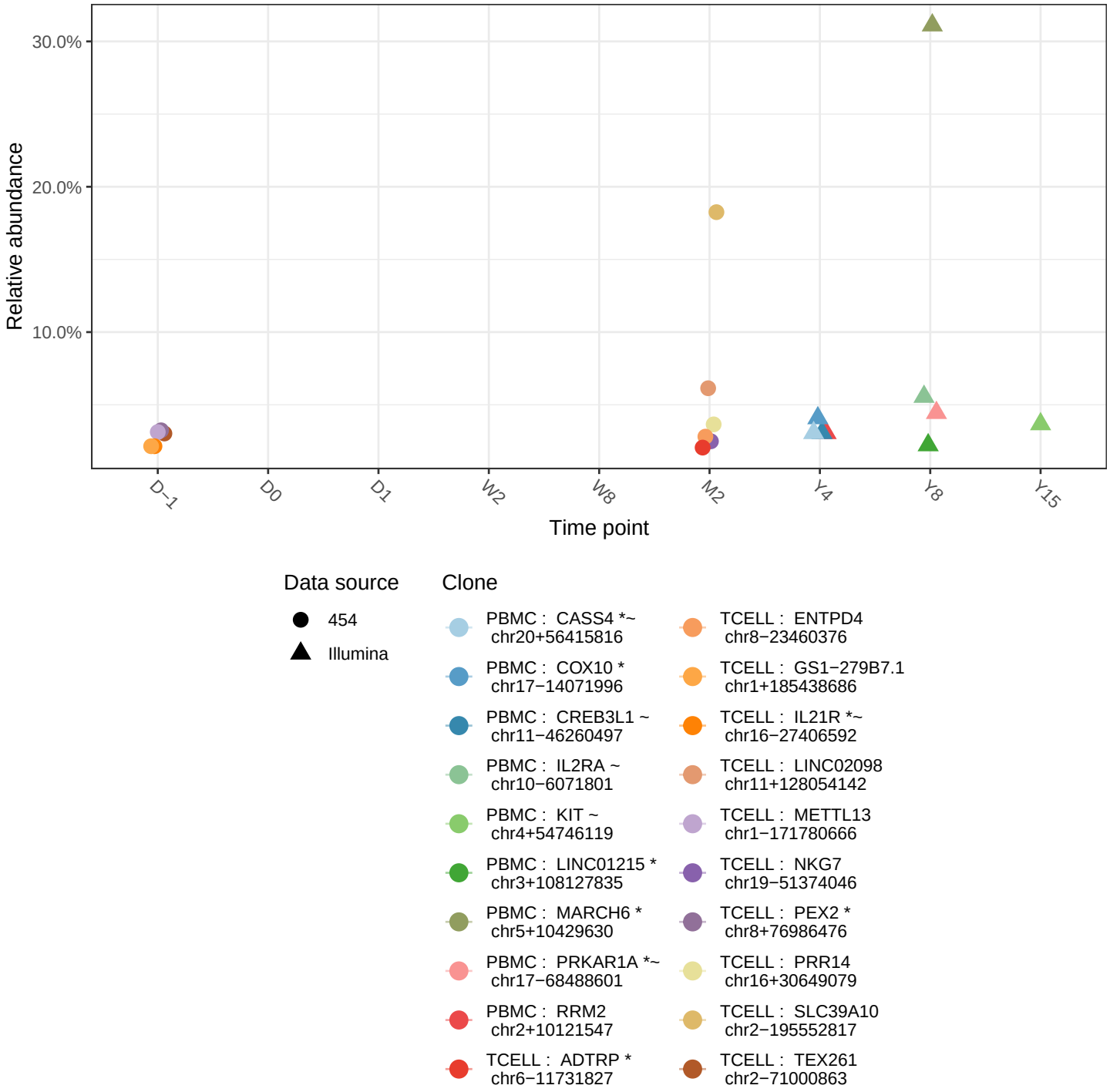
Do any cell clones account for more than 20% of all clones?

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.

IntSite	Abundance	Relative abudance	time point	Cell type	Nearest gene	Distance (KB)	Nearest oncogene	Distance (KB)
chr5+10429630	28	31.1%	Y8	PBMC	MARCH6	0.00	CTNND2	542.20

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 pin26 over time points D-1, 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

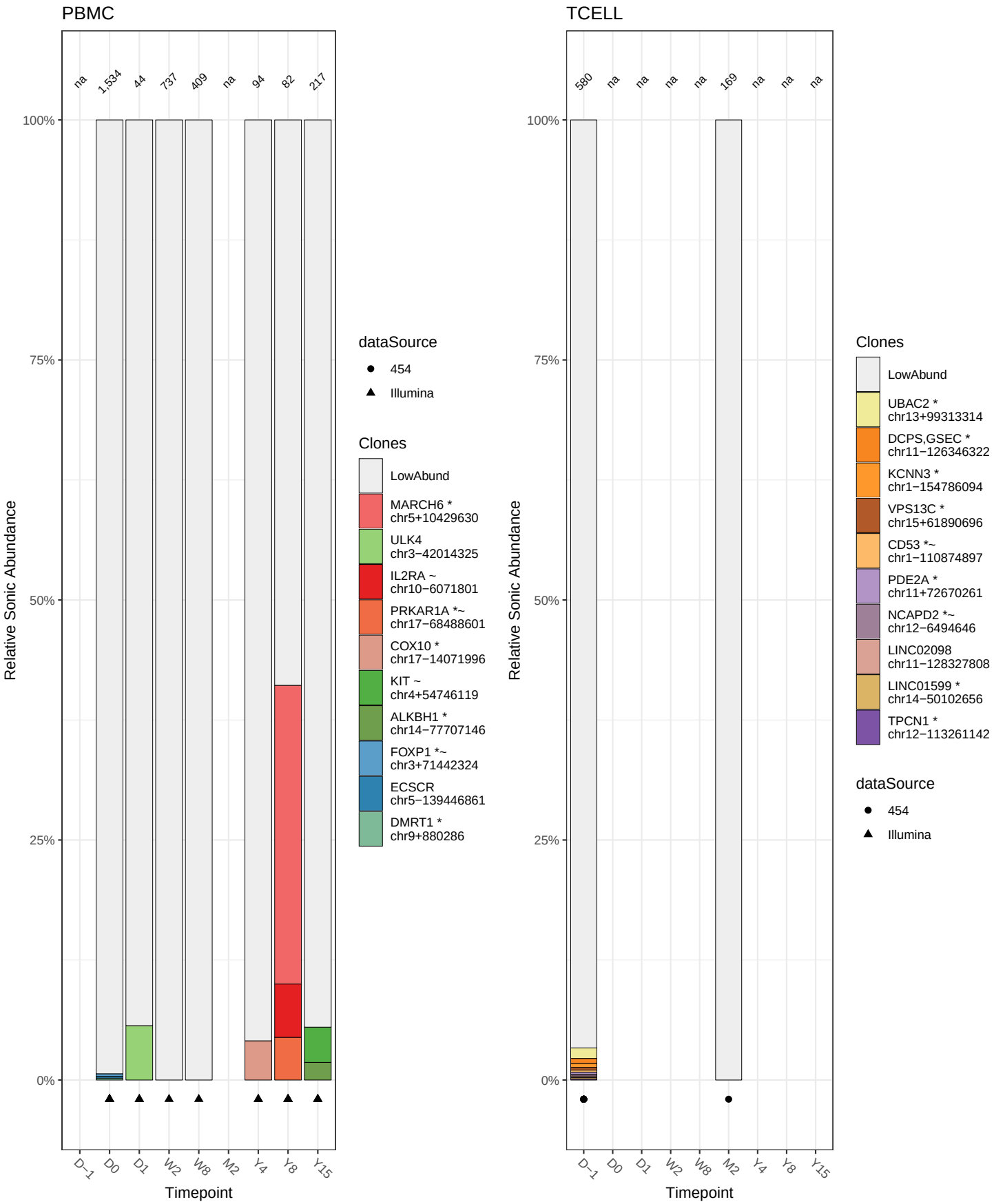
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, 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
LS-8	454	hg38	D-1	TCELL	5,027	580	522	0.089	2,343	6.22	0.995	233	yes	NA	NA
GTSP3466	Illumina	hg38	D0	PBMC	574,645	1,534	1,483	0.032	24,379	7.29	0.998	717	yes	2023-03-01	NA
GTSP4695	Illumina	hg38	D1	PBMC	163,138	44	41	0.065	412	3.68	0.990	20	yes	2022-03-11	NA
GTSP4704	Illumina	hg38	W2	PBMC	628,753	737	714	0.030	10,647	6.56	0.998	346	yes	2023-03-01	NA
GTSP4713	Illumina	hg38	W8	PBMC	421,081	409	397	0.028	6,083	5.97	0.998	193	yes	2023-03-01	NA
LS-9	454	hg38	M2	TCELL	4,383	169	169	0.000	14,365	5.13	1.000	85	yes	NA	NA
GTSP4722	Illumina	hg38	Y4	PBMC	258,765	94	77	0.159	308	4.26	0.981	31	yes	2022-03-11	NA
GTSP4731	Illumina	hg38	Y8	PBMC	94,405	82	47	0.411	498	3.09	0.802	7	yes	2022-03-11	NA
GTSP4740	Illumina	hg38	Y15	PBMC	360,381	217	192	0.109	1,490	5.18	0.985	84	yes	2023-03-01	NA

Tracking of clonal abundances

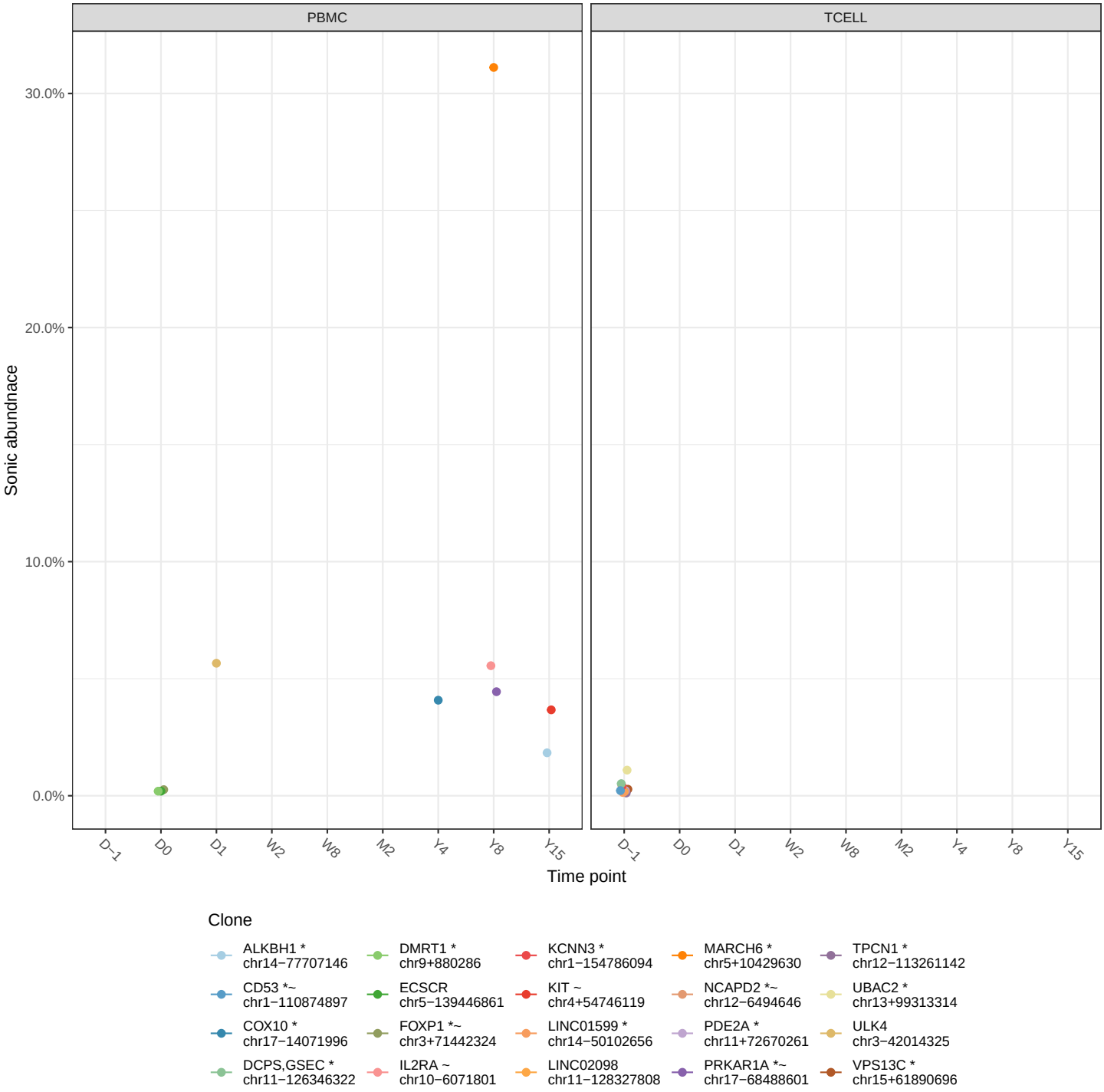
Relative abundance of cell clones

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
na
1,534
44
737
409
na
94
82
217
100%
75%
50%
25%
0%
Relative Sonic Abundance
D−1
D0
D1
W2
W8
M2
Y4
Y8
Y15
Timepoint
dataSource
454
Illumina
Clones
LowAbund
MARCH6 * chr5+10429630
ULK4 chr3−42014325
IL2RA ~ chr10−6071801
PRKAR1A *~ chr17−68488601
COX10 * chr17−14071996
KIT ~ chr4+54746119
ALKBH1 * chr14−77707146
FOXP1 *~ chr3+71442324
ECSCR chr5−139446861
DMRT1 * chr9+880286
TCELL
580
na
na
na
na
169
na
na
na
100%
75%
50%
25%
0%
Relative Sonic Abundance
D−1
D0
D1
W2
W8
M2
Y4
Y8
Y15
Timepoint
Clones
LowAbund
UBAC2 * chr13+99313314
DCPS,GSEC * chr11−126346322
KCNN3 * chr1−154786094
VPS13C * chr15+61890696
CD53 *~ chr1−110874897
PDE2A * chr11+72670261
NCAPD2 *~ chr12−6494646
LINC02098 chr11−128327808
LINC01599 * chr14−50102656
TPCN1 * chr12−113261142
dataSource
454
Illumina

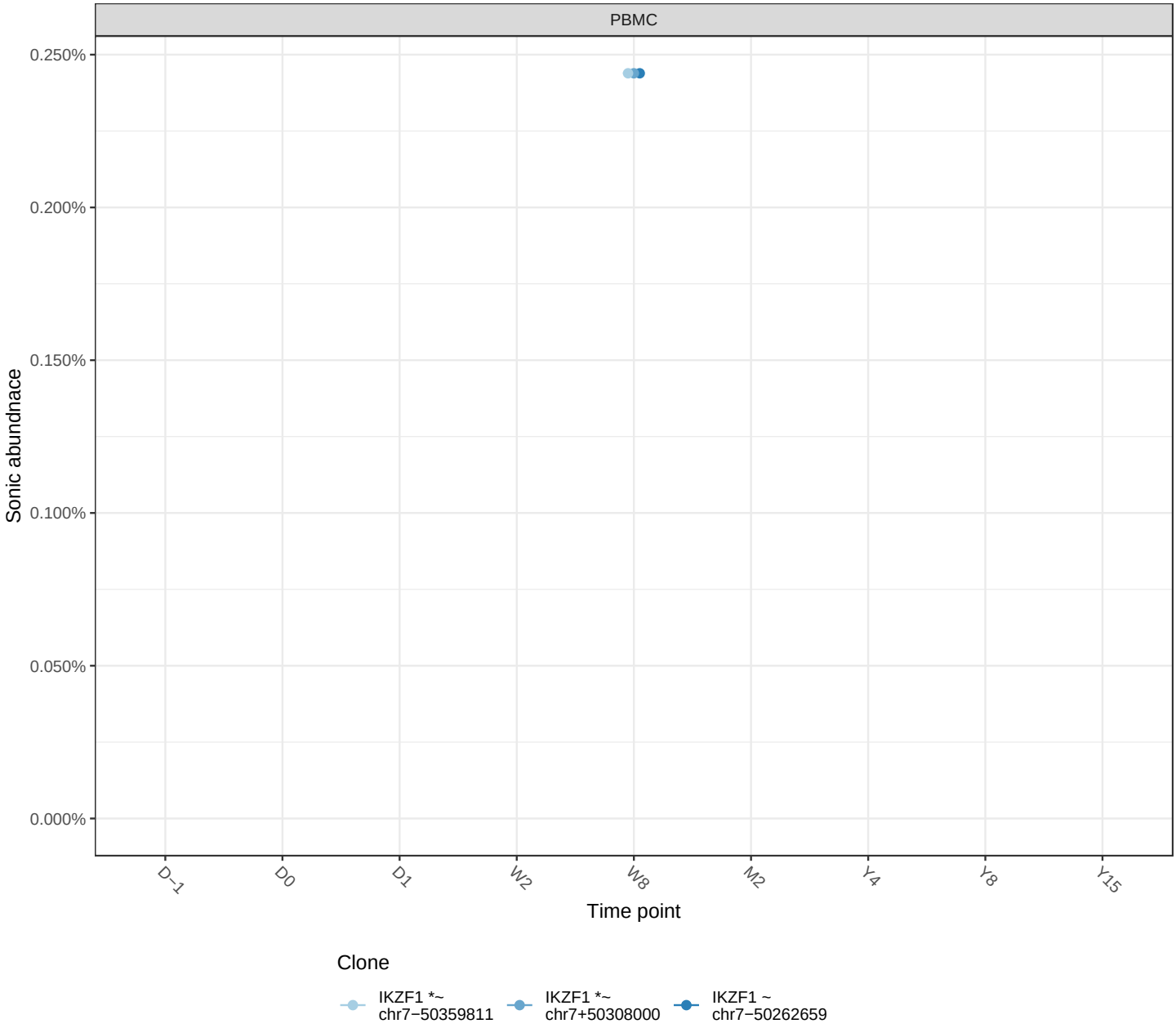
Longitudinal behavior of major clones

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.



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 1:2

PBMC
D0 1:4

PBMC
D1 1:3

TCELL
M2 1:1

PBMC
W2 1:2

PBMC
W8 1:2

PBMC
Y4 1:4

PBMC
Y8 1:28

PBMC
Y15 1:8

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
GTSP4740-12	100904476	5	40.0%	PBMC	y15

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
GTSP3466	D0	PBMC	0.26%	chr3+71442324	FOXP1 *~
GTSP4695	D1	PBMC	5.66%	chr3-42014325	ULK4
GTSP4704	W2	PBMC	0.27%	chr10-30684900	SVILP1
GTSP4713	W8	PBMC	0.49%	chr1-149850490	HIST2H2BC *
GTSP4722	Y4	PBMC	4.08%	chr17-14071996	COX10 *
GTSP4731	Y8	PBMC	31.11%	chr5+10429630	MARCH6 *
GTSP4740	Y15	PBMC	3.67%	chr4+54746119	KIT ~
LS-8	D-1	TCELL	3.24%	chr8+76986476	PEX2 *
LS-9	M2	TCELL	18.25%	chr2-195552817	SLC39A10