Analysis of integration site distributions and relative clonal abundance for subject pin30

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-10	hg38	D-1	TCELL	7,785	2,635	2,635
GTSP4697	hg38	D1	PBMC	161,928	59	55
GTSP4706	hg38	W2	PBMC	214,684	195	183
GTSP4715	hg38	W8	PBMC	311,254	156	148
LS-11	hg38	M2	TCELL	1,406	69	69
GTSP4724	hg38	Y4	PBMC	226,221	190	178
GTSP4733	hg38	Y8	PBMC	338,522	166	153
LS-12	hg38	Y8	TCELL	3,918	20	17
GTSP4739	hg38	Y15	PBMC	169,263	169	151
GTSP3209	NA	w8	PBMC	NA	NA	NA
GTSP3210	NA	w8	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	55	No
W2	183	No
W8	148	No
Y4	178	No
Y8	153	No
Y15	151	No

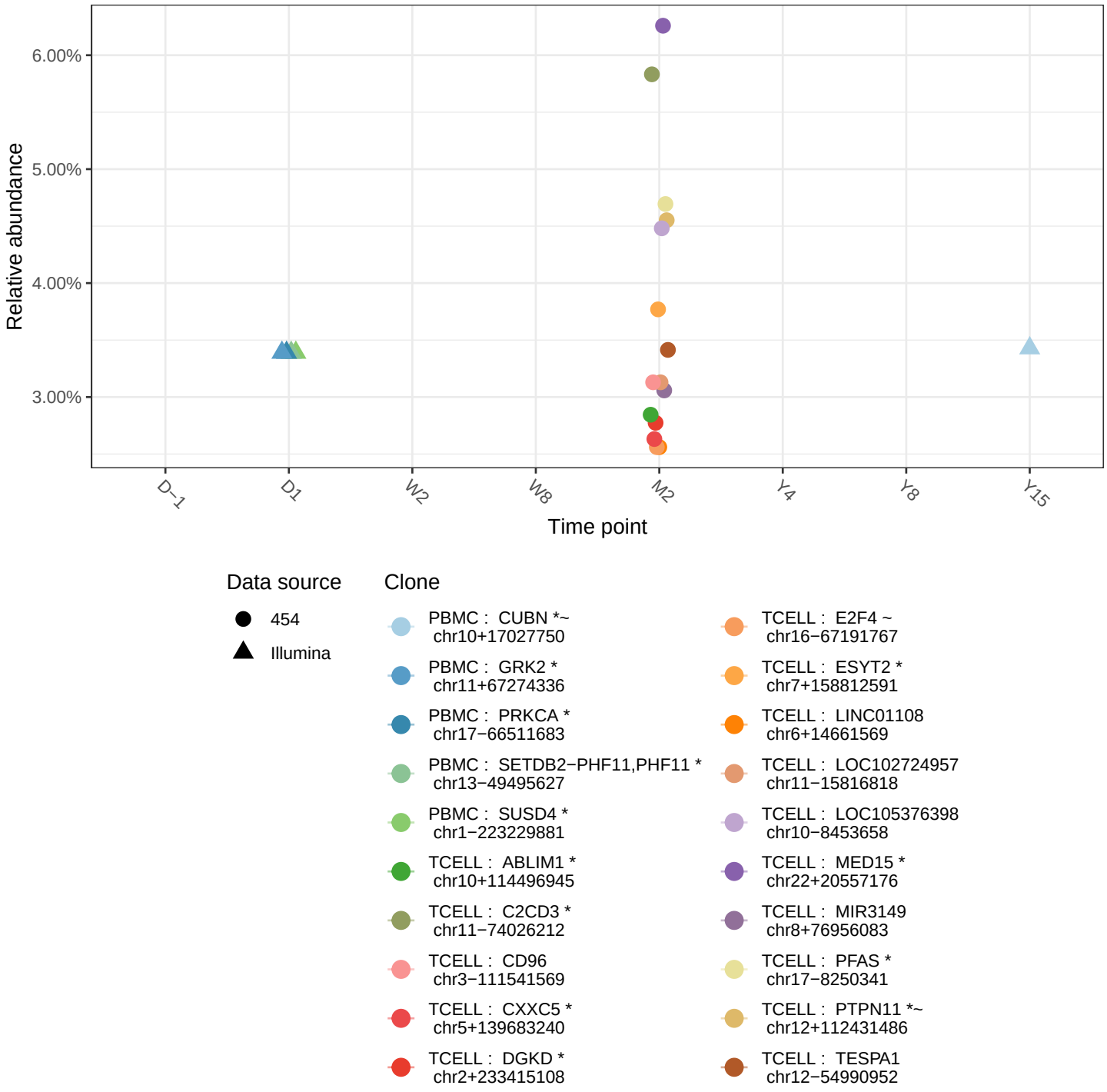
Do any cell clones account for more than 20% of all clones?

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 pin30 over time points D-1, 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

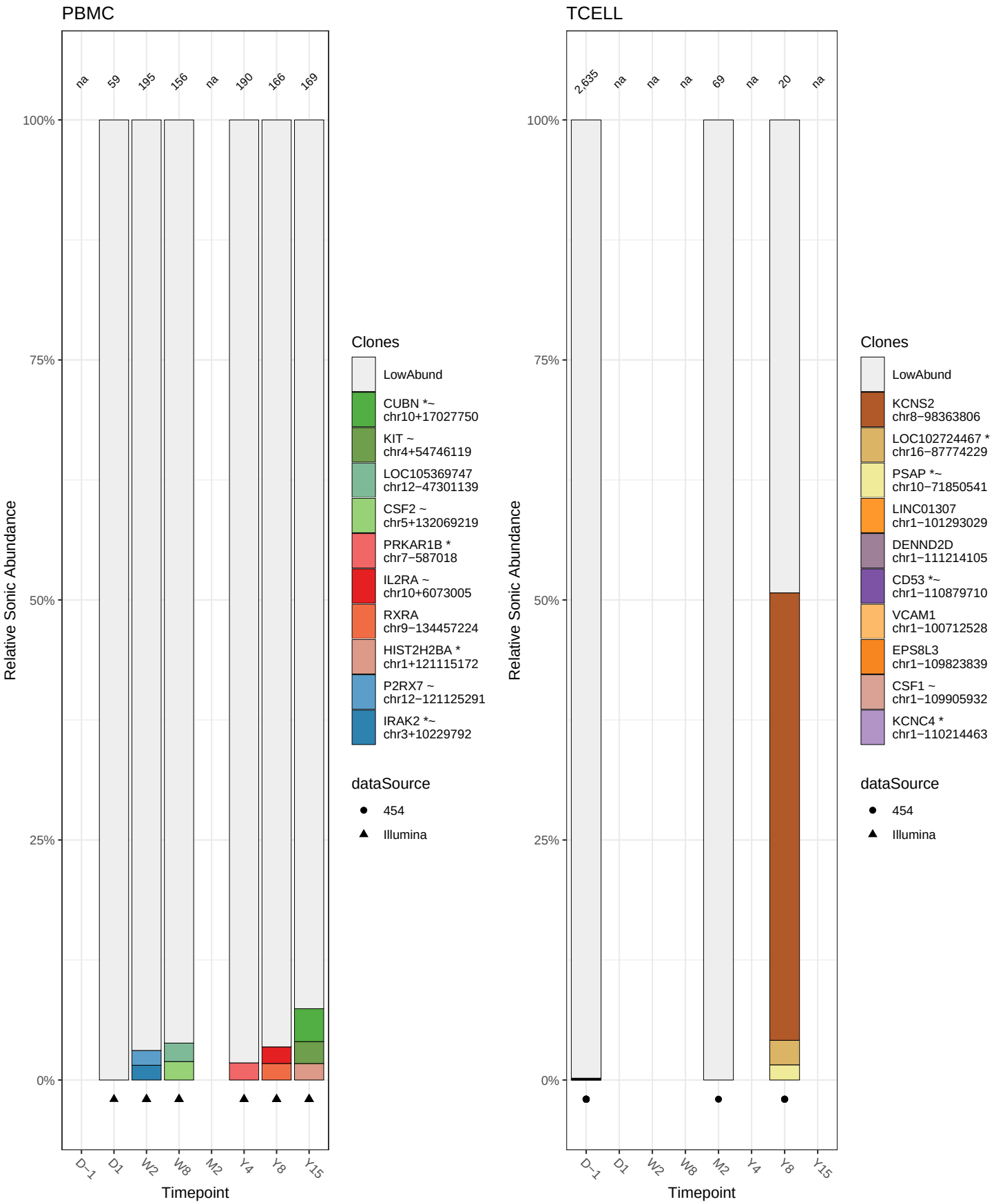
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, 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-10	454	hg38	D-1	TCELL	7,785	2,635	2,635	0.000	3,472,930	7.88	1.000	1,318	yes	NA	NA
GTSP4697	Illumina	hg38	D1	PBMC	161,928	59	55	0.063	310	3.98	0.994	26	yes	2022-03-24	NA
GTSP4706	Illumina	hg38	W2	PBMC	214,684	195	183	0.059	1,836	5.18	0.995	86	yes	2022-03-24	NA
GTSP4715	Illumina	hg38	W8	PBMC	311,254	156	148	0.050	2,150	4.97	0.995	71	yes	2022-03-24	NA
LS-11	454	hg38	M2	TCELL	1,406	69	69	0.000	2,415	4.23	1.000	35	yes	NA	NA
GTSP4724	Illumina	hg38	Y4	PBMC	226,221	190	178	0.060	1,581	5.15	0.994	84	yes	2023-01-26	NA
GTSP4733	Illumina	hg38	Y8	PBMC	338,522	166	153	0.073	1,154	5.00	0.993	71	yes	2023-03-01	NA
LS-12	454	hg38	Y8	TCELL	3,918	20	17	0.124	40	2.79	0.984	8	yes	NA	NA
GTSP4739	Illumina	hg38	Y15	PBMC	169,263	169	151	0.102	1,561	4.95	0.986	67	yes	2022-03-24	NA
GTSP3209	NA	NA	w8	PBMC	NA	NA	NA	NA	NA	NA	NA	NA	NA	2019-09-10	NA
GTSP3210	NA	NA	w8	PBMC	NA	NA	NA	NA	NA	NA	NA	NA	NA	2019-09-10	NA

Tracking of clonal abundances

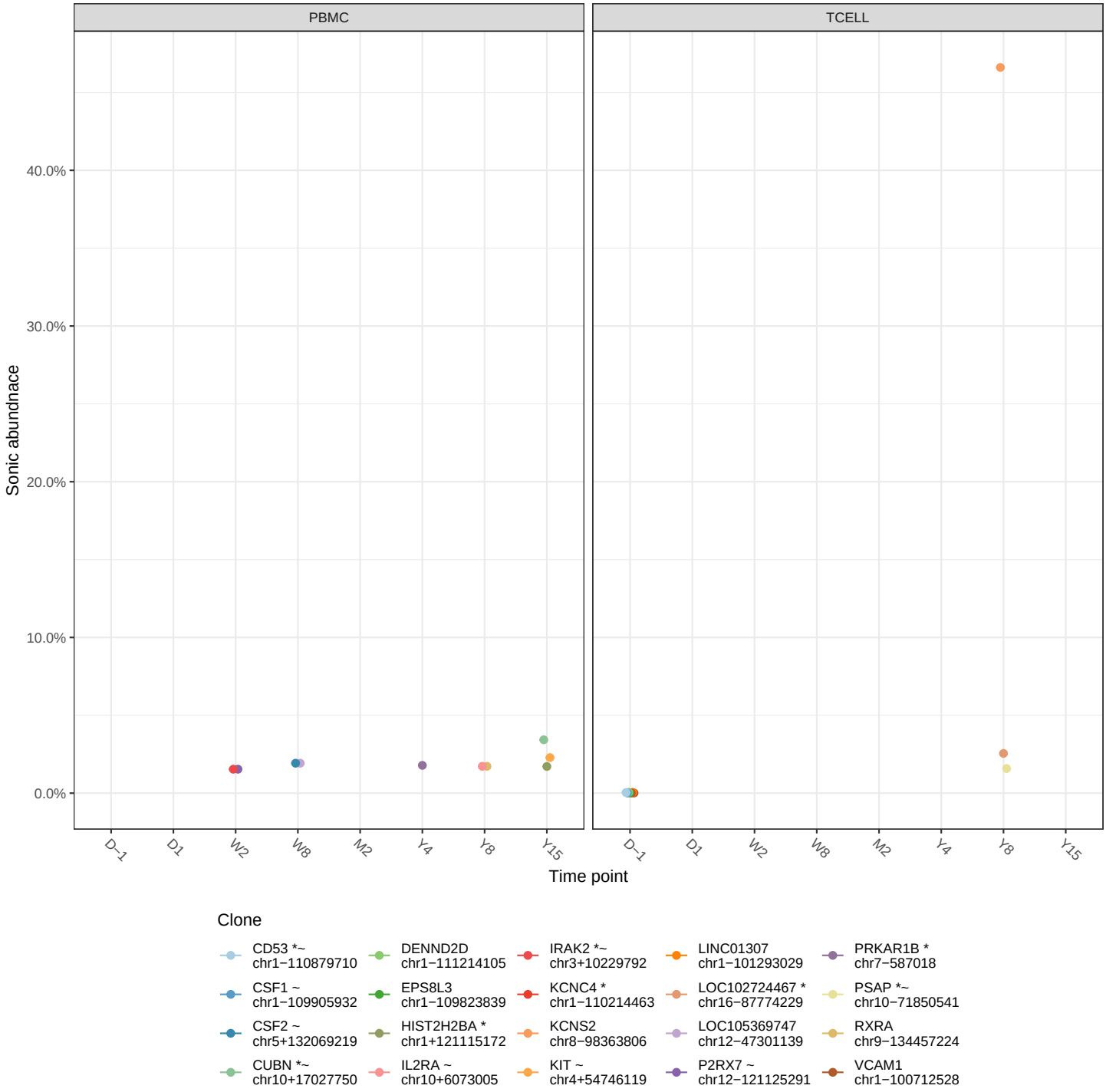
Relative abundance of cell clones

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
59
195
156
na
190
166
169
100%
75%
50%
25%
0%
Relative Sonic Abundance
D−1
D1
W2
W8
M2
Y4
Y8
Y15
Timepoint
Clones
LowAbund
CUBN *~ chr10+17027750
KIT ~ chr4+54746119
LOC105369747 chr12−47301139
CSF2 ~ chr5+132069219
PRKAR1B * chr7−587018
IL2RA ~ chr10+6073005
RXRA chr9−134457224
HIST2H2BA * chr1+121115172
P2RX7 ~ chr12−121125291
IRAK2 *~ chr3+10229792
dataSource
454
Illumina
TCELL
2,635
na
na
na
69
na
20
na
100%
75%
50%
25%
0%
Relative Sonic Abundance
D−1
D1
W2
W8
M2
Y4
Y8
Y15
Timepoint
Clones
LowAbund
KCNS2 chr8−98363806
LOC102724467 * chr16−87774229
PSAP *~ chr10−71850541
LINC01307 chr1−101293029
DENND2D chr1−111214105
CD53 *~ chr1−110879710
VCAM1 chr1−100712528
EPS8L3 chr1−109823839
CSF1 ~ chr1−109905932
KCNC4 * chr1−110214463
dataSource
454
Illumina

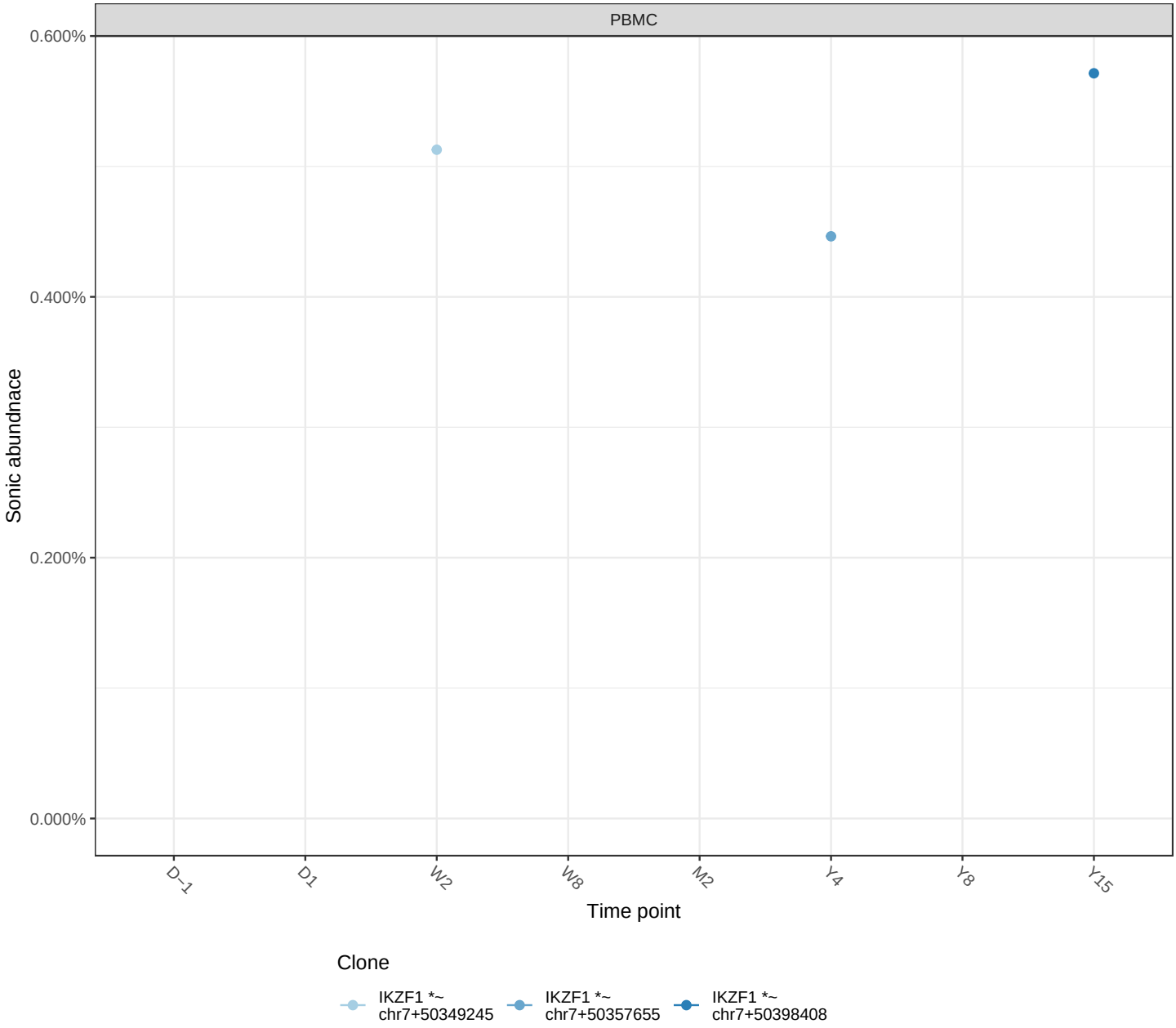
Longitudinal behavior of major clones

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.



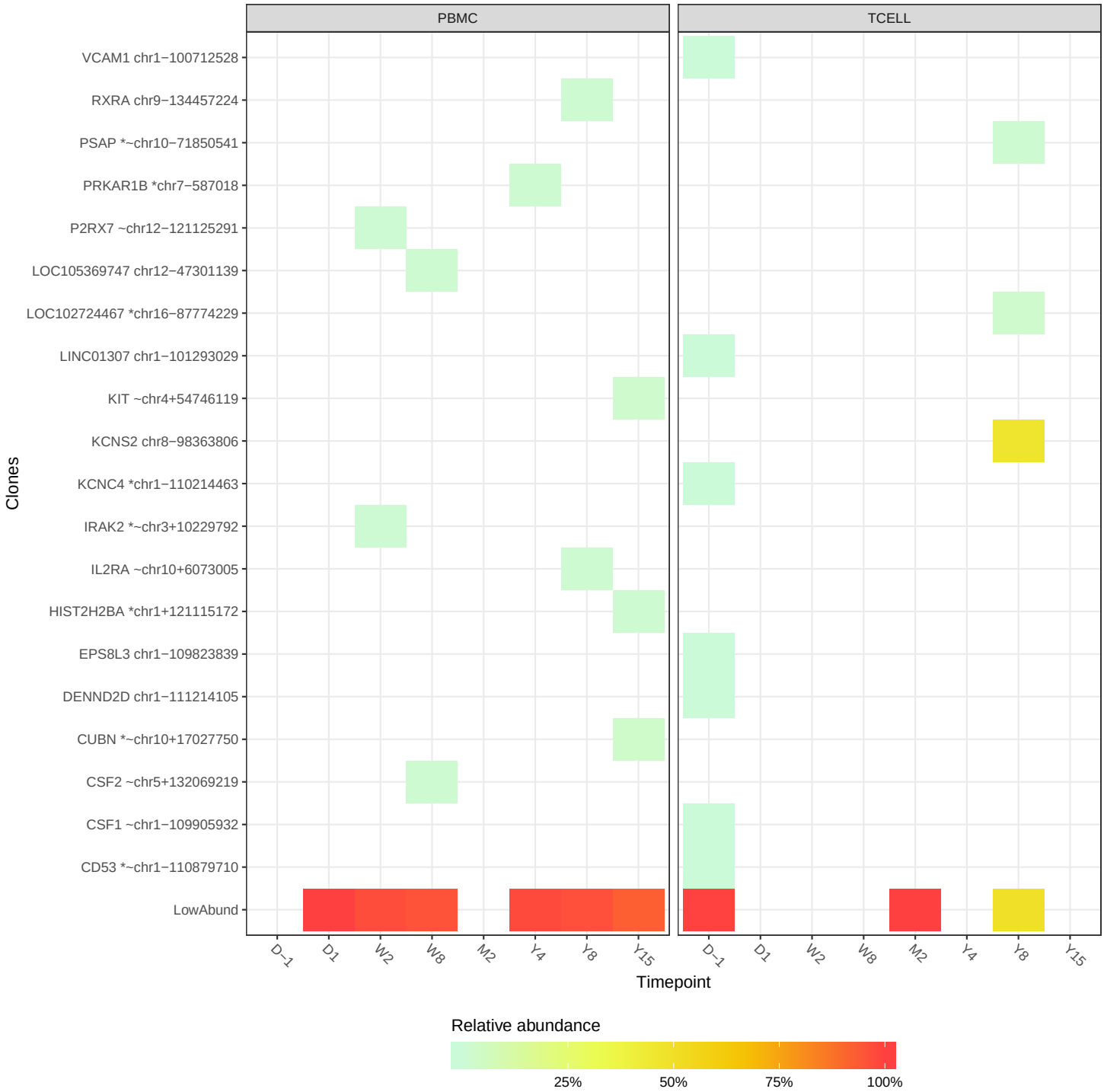
Integration sites near particular genes of interest

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 1:1

PBMC
D1 1:2

TCELL
M2 1:1

PBMC
W2 1:3

PBMC
W8 1:3

PBMC
Y4 1:4

PBMC
Y8 1:3

TCELL
Y8 1:2

PBMC
Y15 1:6

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.

No sample contained a multihit grouping which exceeded 20% of the sample's inferred cells.

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
GTSP4697	D1	PBMC	3.39%	chr1-223229881	SUSD4 *
GTSP4706	W2	PBMC	1.54%	chr12-121125291	P2RX7 ~
GTSP4715	W8	PBMC	1.92%	chr12-47301139	LOC105369747
GTSP4724	Y4	PBMC	1.79%	chr7-587018	PRKAR1B *
GTSP4733	Y8	PBMC	1.72%	chr10+6073005	IL2RA ~
GTSP4739	Y15	PBMC	3.43%	chr10+17027750	CUBN *~
LS-10	D-1	TCELL	0.44%	chrX-150470485	MAMLD1 *
LS-11	M2	TCELL	6.26%	chr22+20557176	MED15 *
LS-12	Y8	TCELL	46.61%	chr8-98363806	KCNS2