Analysis of integration site distributions and relative clonal abundance for subject pin25

January 23, 2025

Contents

Sample Overview **2**

Clonal expansion summary **2**
Is there a rich population of progenitor cells delivering mature cells to the periphery? 2
Do any cell clones account for more than 20% of all clones? 2
Are any cell clones increasing in proportion over time? 3

Introduction **4**

Sample Summary **5**

Tracking of clonal abundances **6**
Relative abundance of cell clones 6
Longitudinal behavior of major clones 8
Integration sites near particular genes of interest 9
Sample relative abundance heatmap 10

What are the most frequently occuring gene types in the subject? **11**

Multihits **12**

Methods **13**

Supplementary table 1. **14**
Greatest sample relative abundance values. 14

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-6	hg38	D-1	TCELL	5,366	1,397	1,397
GTSP4694	hg38	D1	PBMC	261,619	34	30
GTSP4703	hg38	W2	PBMC	188,290	454	447
GTSP4712	hg38	W8	PBMC	645,075	455	421
GTSP4721	hg38	Y4	PBMC	318,335	101	76
GTSP4730	hg38	Y8	PBMC	167,451	87	84
LS-7	hg38	Y8	TCELL	4,703	24	20
GTSP4744	hg38	Y15	PBMC	185,380	150	134

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	30	No
W2	447	No
W8	421	No
Y4	76	No
Y8	84	No
Y15	134	No

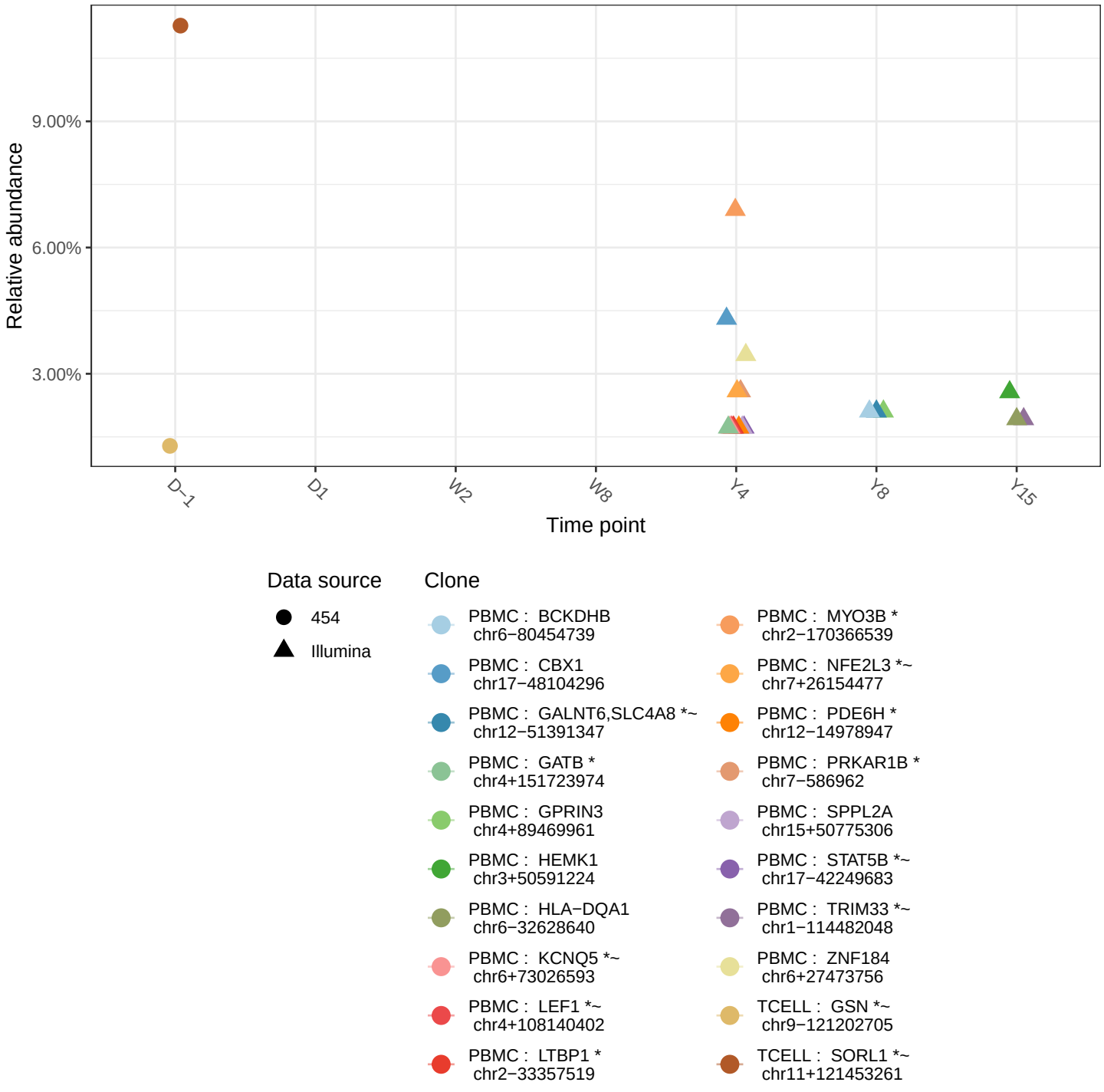
Do any cell clones account for more than 20% of all clones?

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 pin25 over time points D-1, D1, 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

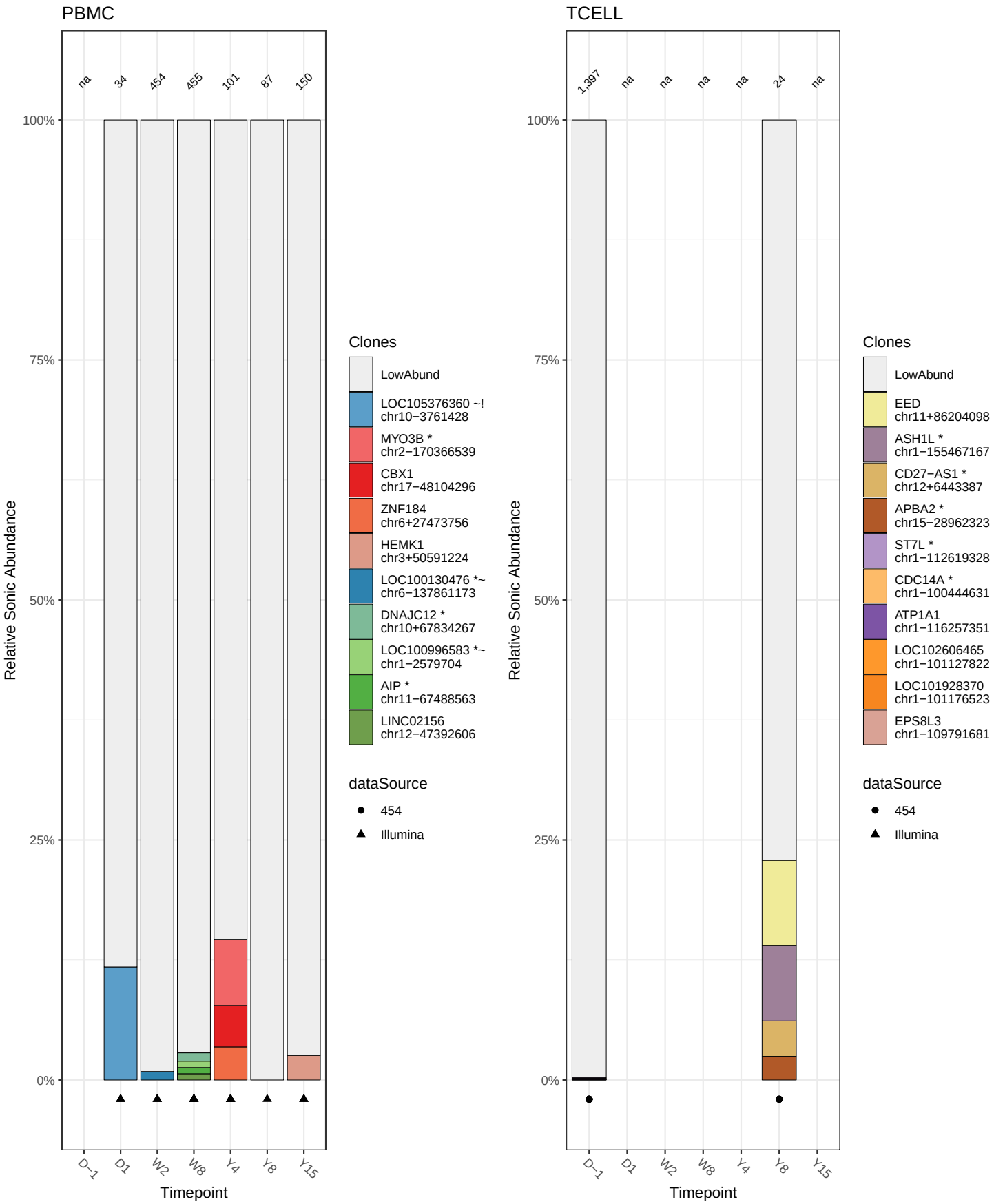
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, 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-6	454	hg38	D-1	TCELL	5,366	1,397	1,397	0.000	976,503	7.24	1.000	699	yes	NA	NA
GTSP4694	Illumina	hg38	D1	PBMC	261,619	34	30	0.112	219	3.32	0.977	14	yes	2022-03-03	NA
GTSP4703	Illumina	hg38	W2	PBMC	188,290	454	447	0.015	19,939	6.09	0.999	221	yes	2022-03-03	NA
GTSP4712	Illumina	hg38	W8	PBMC	645,075	455	421	0.071	5,334	6.00	0.994	194	yes	2023-03-01	NA
GTSP4721	Illumina	hg38	Y4	PBMC	318,335	101	76	0.223	328	4.15	0.959	26	yes	2022-03-03	NA
GTSP4730	Illumina	hg38	Y8	PBMC	167,451	87	84	0.033	894	4.42	0.997	41	yes	2022-03-03	NA
LS-7	454	hg38	Y8	TCELL	4,703	24	20	0.133	44	2.95	0.984	9	yes	NA	NA
GTSP4744	Illumina	hg38	Y15	PBMC	185,380	150	134	0.099	872	4.85	0.990	60	yes	2022-03-11	NA

Tracking of clonal abundances

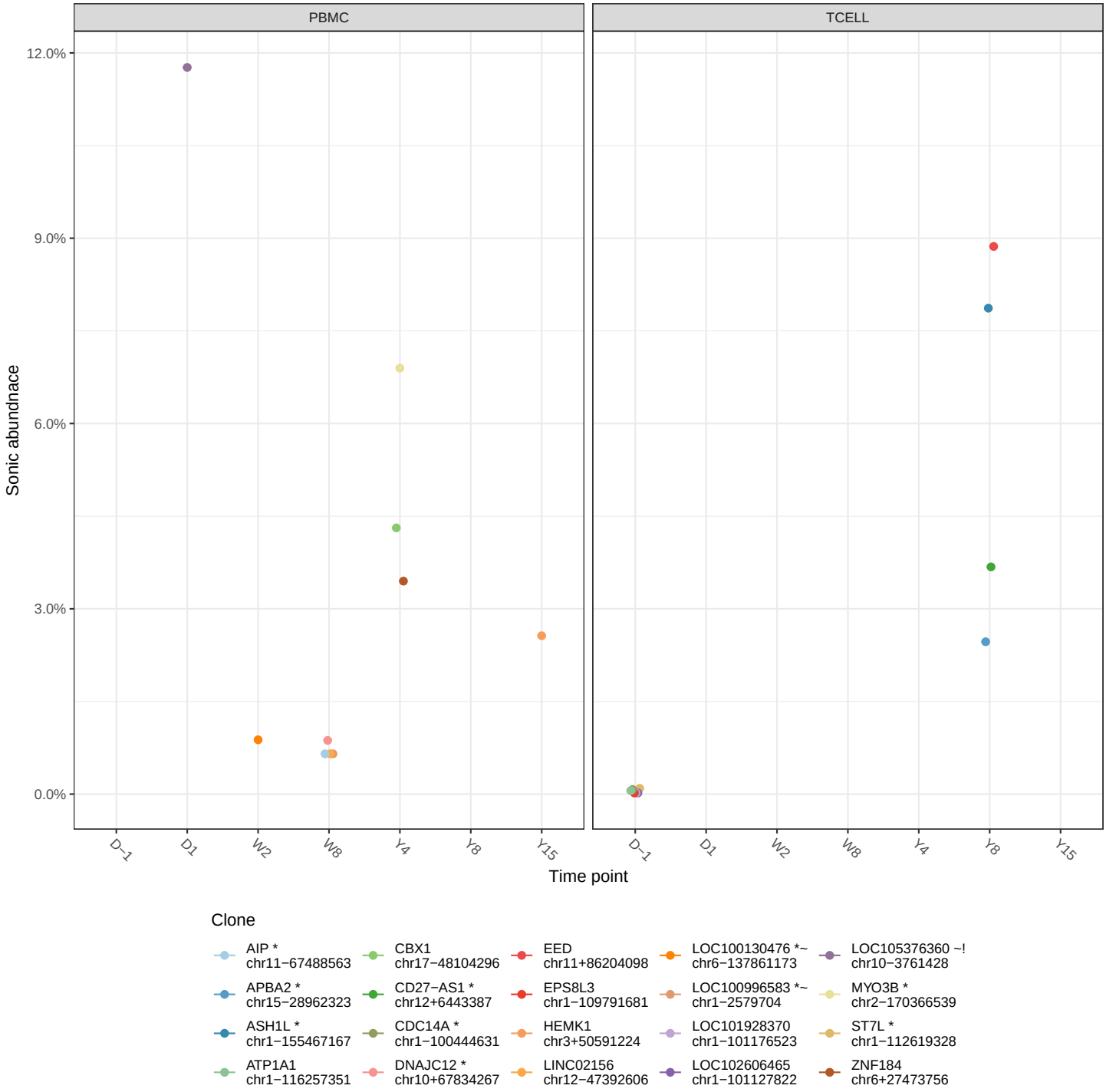
Relative abundance of cell clones

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
TCELL
na
34
454
455
101
87
150
1,397
24
100%
75%
50%
25%
0%
Relative Sonic Abundance
D−1
D1
W2
W8
Y4
Y8
Y15
Timepoint
Clones
LowAbund
LOC105376360 ~!
chr10−3761428
MYO3B *
chr2−170366539
CBX1
chr17−48104296
ZNF184
chr6+27473756
HEMK1
chr3+50591224
LOC100130476 *~
chr6−137861173
DNAJC12 *
chr10+67834267
LOC100996583 *~
chr1−2579704
AIP *
chr11−67488563
LINC02156
chr12−47392606
EED
chr11+86204098
ASH1L *
chr1−155467167
CD27−AS1 *
chr12+6443387
APBA2 *
chr15−28962323
ST7L *
chr1−112619328
CDC14A *
chr1−100444631
ATP1A1
chr1−116257351
LOC102606465
chr1−101127822
LOC101928370
chr1−101176523
EPS8L3
chr1−109791681
dataSource
454
Illumina

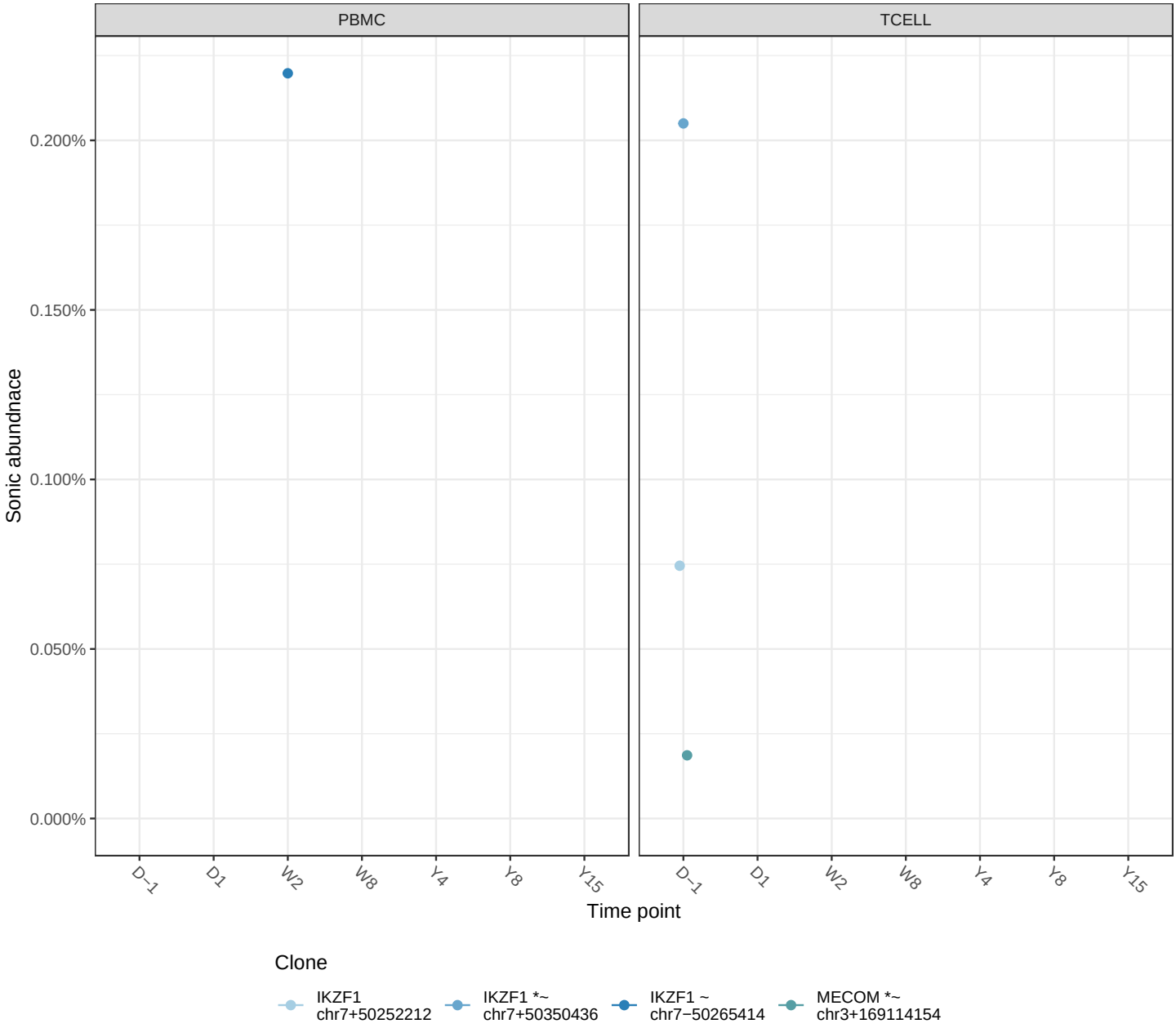
Longitudinal behavior of major clones

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 1:4

PBMC
W2 1:4

PBMC
W8 1:4

PBMC
Y4 1:8

PBMC
Y8 1:2

TCELL
Y8 1:2

PBMC
Y15 1:4

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
GTSP4694	D1	PBMC	11.76%	chr10-3761428	LOC105376360 ~!
GTSP4703	W2	PBMC	0.88%	chr6-137861173	LOC100130476 *~
GTSP4712	W8	PBMC	0.87%	chr10+67834267	DNAJC12 *
GTSP4721	Y4	PBMC	6.90%	chr2-170366539	MYO3B *
GTSP4730	Y8	PBMC	2.11%	chr12-51391347	GALNT6,SLC4A8 *~
GTSP4744	Y15	PBMC	2.56%	chr3+50591224	HEMK1
LS-6	D-1	TCELL	11.27%	chr11+121453261	SORL1 *~
LS-7	Y8	TCELL	42.93%	chr5+394118	AHRR *