Analysis of integration site distributions and relative clonal abundance for subject p04409-10

July 02, 2019

Contents

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 (Whole blood, T cells, B cells, NK cells, Neutrophils, Monocytes and PBMC). Cell specimens that pass these criteria are operationally designated Rich.

Time point	PBMC	T cells	Rich
D0	na	1,046	Yes
D10	23	na	No
D28	387	na	No
D63	88	na	No
D92	14	na	No
D120	21	na	No
D121	4	na	No
D147	27	na	No
D204	14	na	No
D442	53	na	No
D801	57	na	No

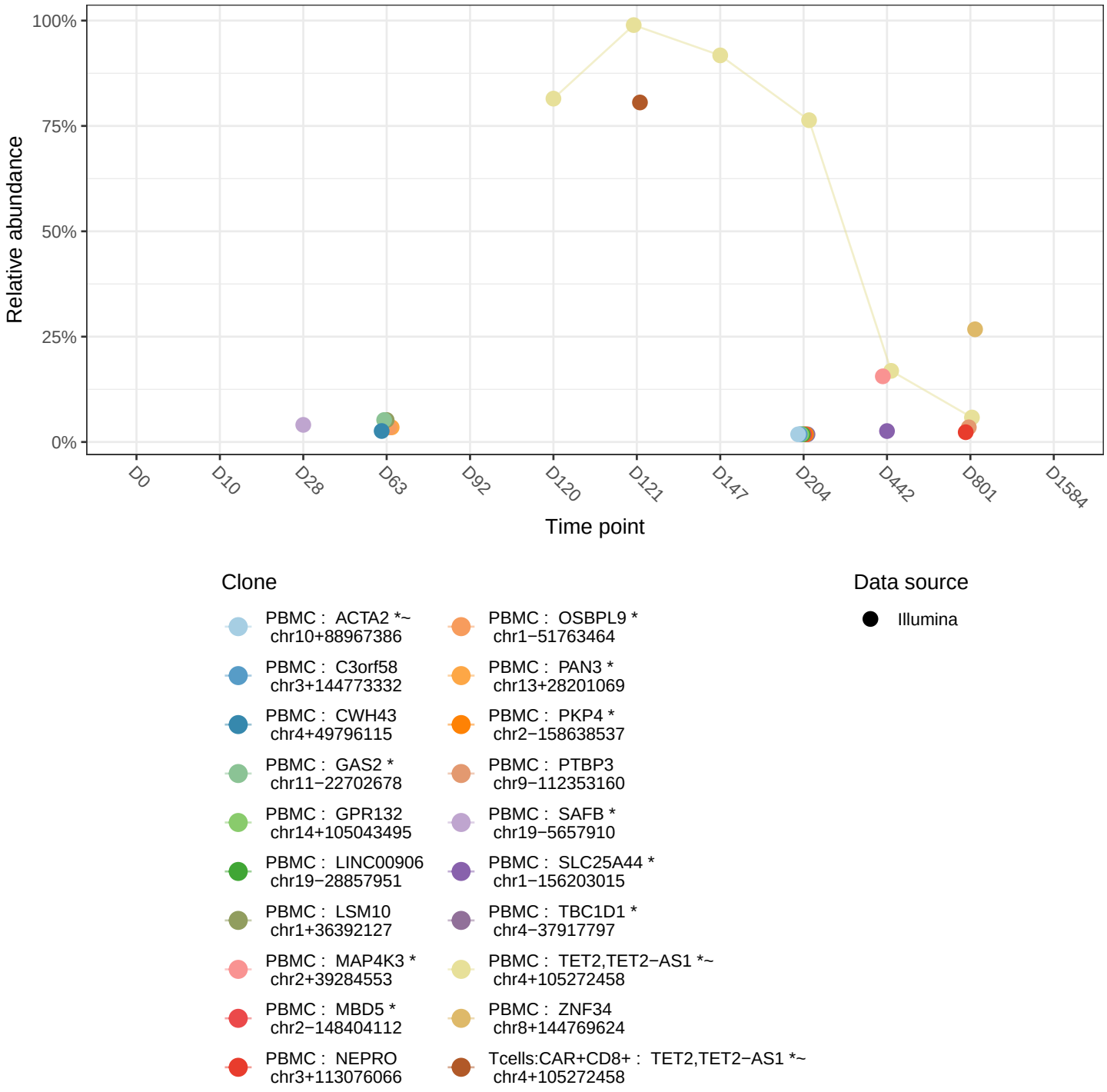
Do any cell clones account for more than 20% of all clones?

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)
chr4+105272458	88	81.5%	D120	PBMC	TET2,TET2-AS1	0.00	TET2	0.00
chr4+105272458	448	80.6%	D121	Tcells:CAR+CD8+	TET2,TET2-AS1	0.00	TET2	0.00
chr4+105272458	278	98.9%	D121	PBMC	TET2,TET2-AS1	0.00	TET2	0.00
chr4+105272458	289	91.7%	D147	PBMC	TET2,TET2-AS1	0.00	TET2	0.00
chr4+105272458	42	76.4%	D204	PBMC	TET2,TET2-AS1	0.00	TET2	0.00
chr8+144769624	23	26.7%	D801	PBMC	ZNF34	2.60	RECQL4	-251.80

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 p04409-10 over time points D0, D10, D28, D63, D92, D120, D121, D147, D204, D442, D801, D1584 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

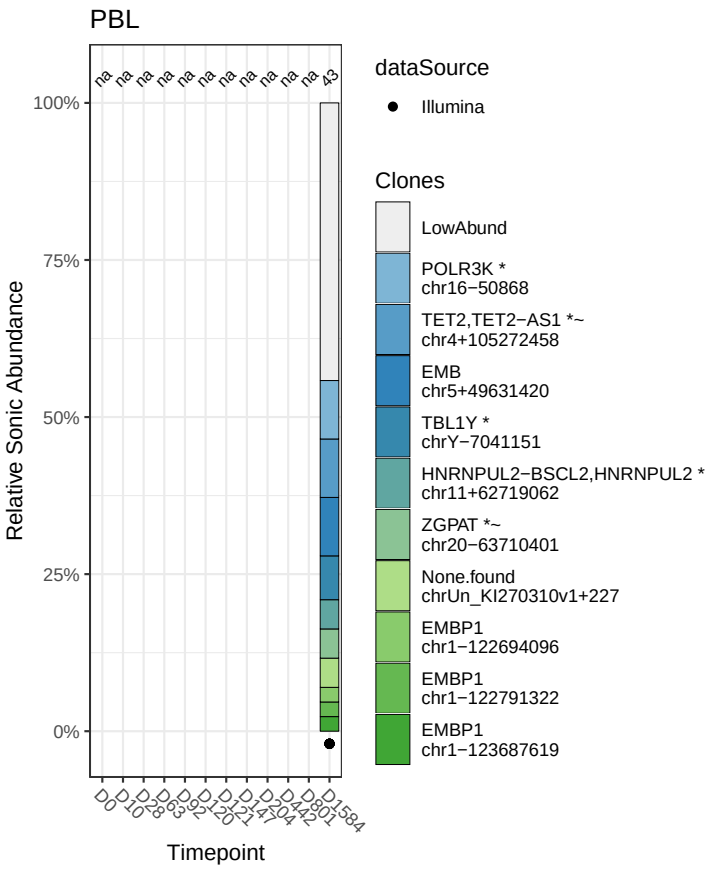
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report.

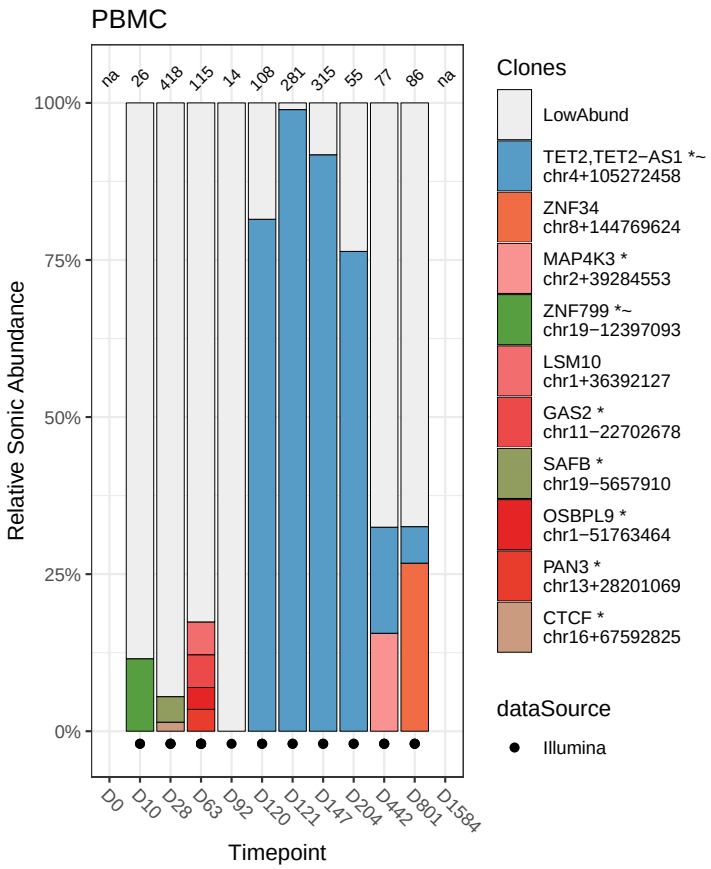
GTSP	dataSource	Patient	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	VCN
GTSP0605	Illumina	p04409-10	D0	T cells	315,963	1,311	1,046	0.179	4,329	6.84	0.984	391	yes	0.16000
GTSP0606	Illumina	p04409-10	D10	PBMC	1,379	26	23	0.107	128	3.08	0.982	11	yes	0.00031
GTSP0607	Illumina	p04409-10	D28	PBMC	289	11	10	0.082	28	2.27	0.987	5	no	0.00298
GTSP1603	Illumina	p04409-10	D28	PBMC	435,384	418	387	0.073	10,512	5.86	0.983	179	yes	NA
GTSP1604	Illumina	p04409-10	D63	PBMC	865,692	115	88	0.211	358	4.32	0.966	31	yes	NA
GTSP1605	Illumina	p04409-10	D92	PBMC	33,434	14	14	0.000	105	2.64	1.000	8	yes	NA
GTSP0608	Illumina	p04409-10	D120	PBMC	52,265	108	21	0.767	211	1.03	0.340	1	yes	0.00170
GTSP0560	Illumina	p04409-10	D121	Tcells:CAR+CD8+	685,760	556	107	0.800	1,892	1.40	0.299	1	yes	0.00000
GTSP0746	Illumina	p04409-10	D121	PBMC	138,379	281	4	0.739	7	0.07	0.051	1	yes	0.00000
GTSP1606	Illumina	p04409-10	D147	PBMC	735,867	315	27	0.880	352	0.55	0.168	1	yes	NA
GTSP0609	Illumina	p04409-10	D204	PBMC	23,395	55	14	0.692	92	1.15	0.437	1	yes	0.01393
GTSP0610	Illumina	p04409-10	D442	PBMC	5,089	17	12	0.270	67	2.20	0.886	4	no	0.00180
GTSP1607	Illumina	p04409-10	D442	PBMC	303,576	77	53	0.299	666	3.51	0.883	15	yes	NA
GTSP0611	Illumina	p04409-10	D801	PBMC	35	20	13	0.323	79	2.16	0.844	4	no	0.00073
GTSP1608	Illumina	p04409-10	D801	PBMC	646,215	86	57	0.327	746	3.47	0.858	15	yes	NA
GTSP1609	Illumina	p04409-10	D1584	PBL	234,480	43	29	0.266	87	3.20	0.951	8	yes	NA

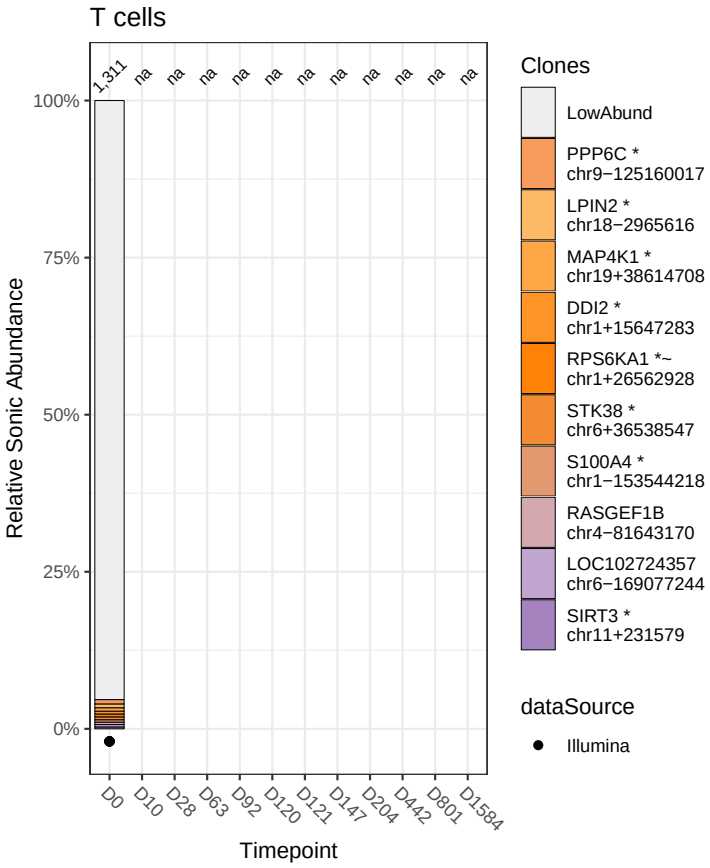
Tracking of clonal abundances

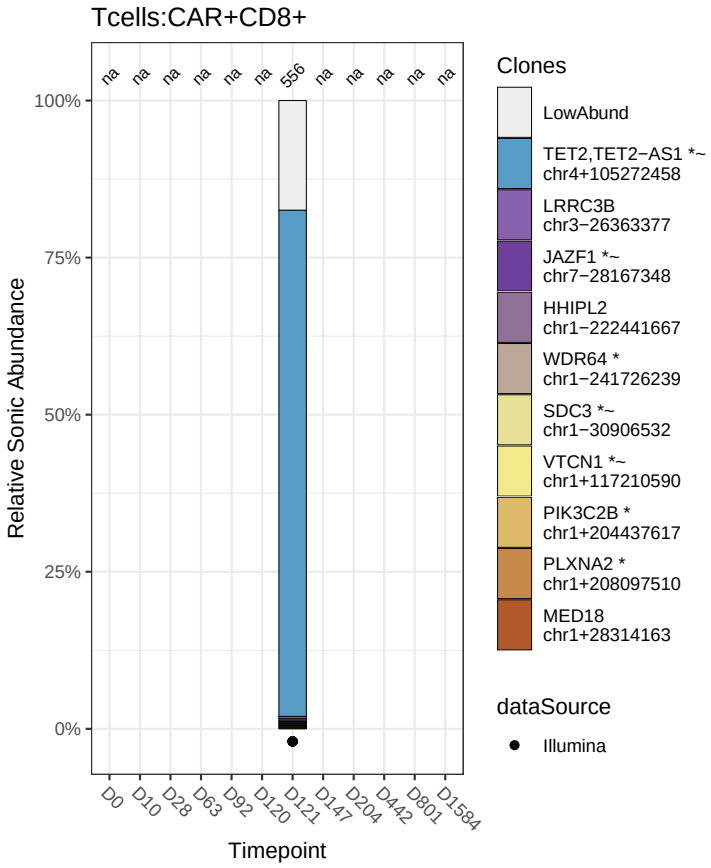
Relative abundance of cell clones

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.


PBL
Relative Sonic Abundance
100%
75%
50%
25%
0%
na na na na na na na na na na 43
D0 D10 D28 D63 D92 D120 D121 D147 D204 D442 D801 D1584
Timepoint
dataSource
Illumina
Clones
LowAbund
POLR3K * chr16−50868
TET2,TET2−AS1 *~ chr4+105272458
EMB chr5+49631420
TBL1Y * chrY−7041151
HNRNPUL2−BSCL2,HNRNPUL2 * chr11+62719062
ZGPAT *~ chr20−63710401
None.found chrUn_KI270310v1+227
EMBP1 chr1−122694096
EMBP1 chr1−122791322
EMBP1 chr1−123687619

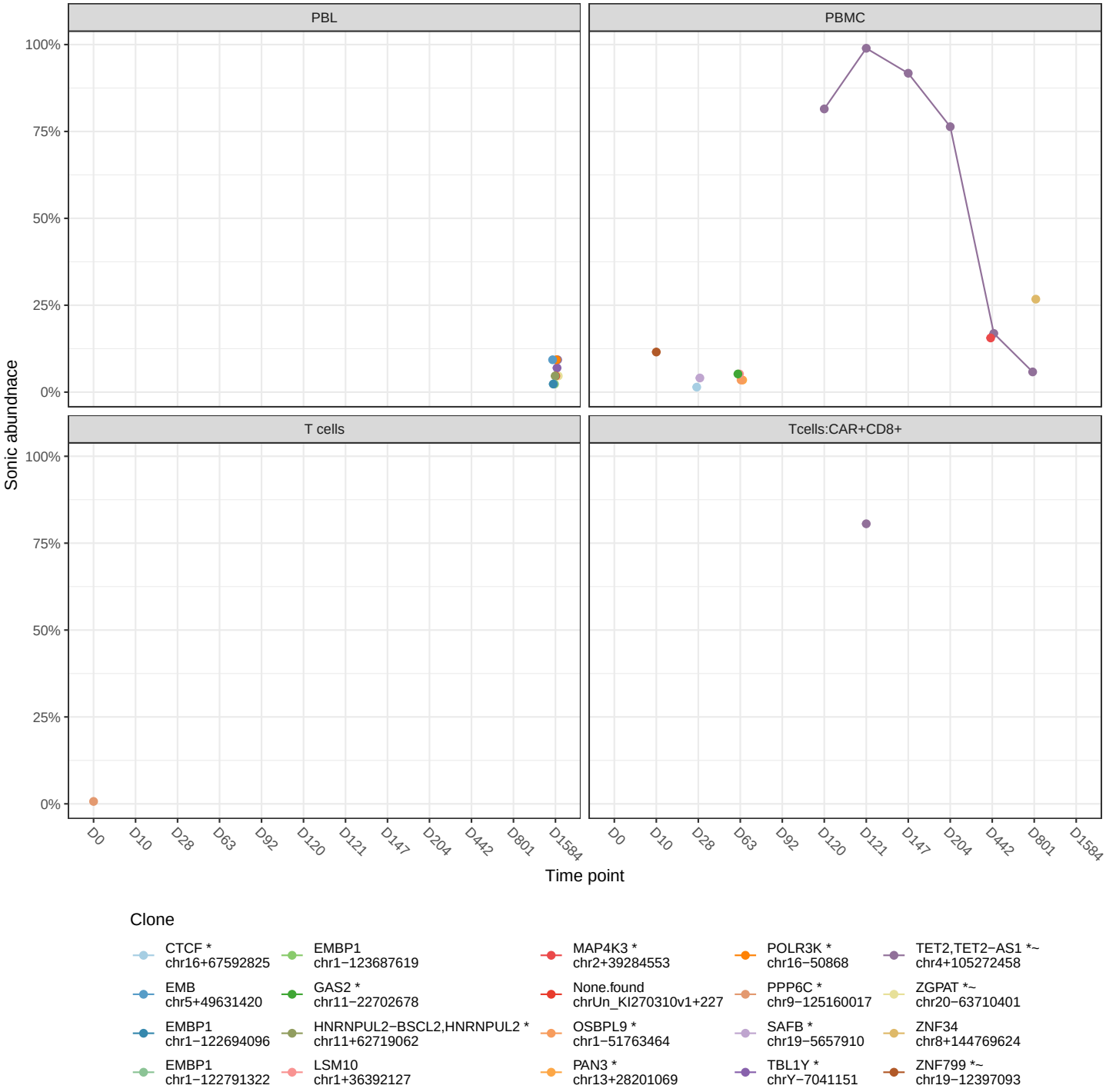

PBMC
Relative Sonic Abundance
100%
75%
50%
25%
0%
na 26 418 115 14 108 281 315 55 77 86 na
D0 D10 D28 D63 D92 D120 D121 D147 D204 D442 D801 D1584
Timepoint
Clones
LowAbund
TET2,TET2−AS1 *~ chr4+105272458
ZNF34 chr8+144769624
MAP4K3 * chr2+39284553
ZNF799 *~ chr19−12397093
LSM10 chr1+36392127
GAS2 * chr11−22702678
SAFB * chr19−5657910
OSBPL9 * chr1−51763464
PAN3 * chr13+28201069
CTCF * chr16+67592825
dataSource
Illumina


T cells
Relative Sonic Abundance
100%
75%
50%
25%
0%
1,311 na na na na na na na na na na na
D0 D10 D28 D63 D92 D120 D121 D147 D204 D442 D801 D1584
Timepoint
Clones
LowAbund
PPP6C * chr9−125160017
LPIN2 * chr18−2965616
MAP4K1 * chr19+38614708
DDI2 * chr1+15647283
RPS6KA1 *~ chr1+26562928
STK38 * chr6+36538547
S100A4 * chr1−153544218
RASGEF1B chr4−81643170
LOC102724357 chr6−169077244
SIRT3 * chr11+231579
dataSource
Illumina


Tcells:CAR+CD8+
Relative Sonic Abundance
100%
75%
50%
25%
0%
na na na na na na 556 na na na na na
D0 D10 D28 D63 D92 D120 D121 D147 D204 D442 D801 D1584
Timepoint
Clones
LowAbund
TET2,TET2−AS1 *~ chr4+105272458
LRRC3B chr3−26363377
JAZF1 *~ chr7−28167348
HHIPL2 chr1−222441667
WDR64 * chr1−241726239
SDC3 *~ chr1−30906532
VTCN1 *~ chr1+117210590
PIK3C2B * chr1+204437617
PLXNA2 * chr1+208097510
MED18 chr1+28314163
dataSource
Illumina

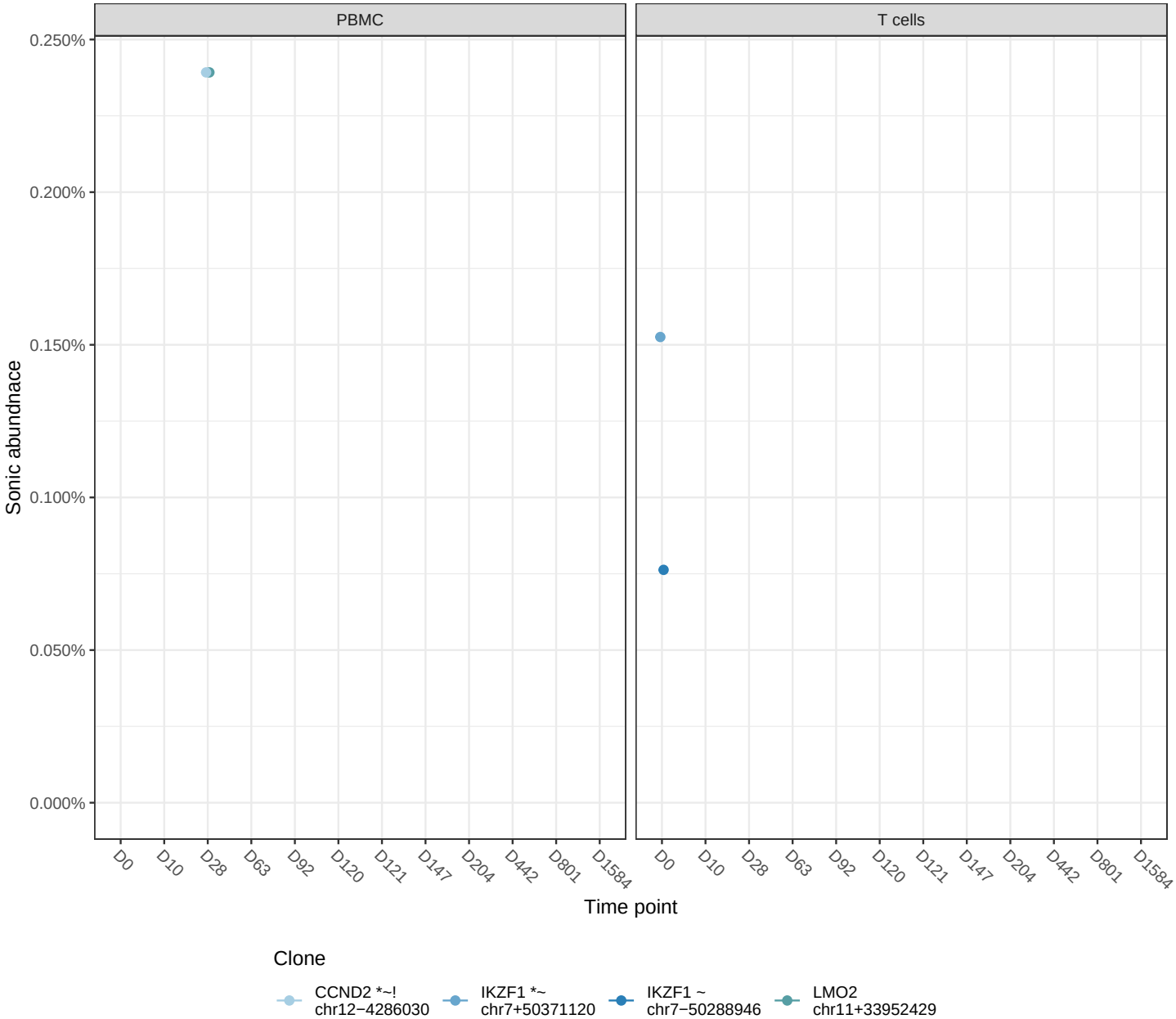
Longitudinal behavior of major clones

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.



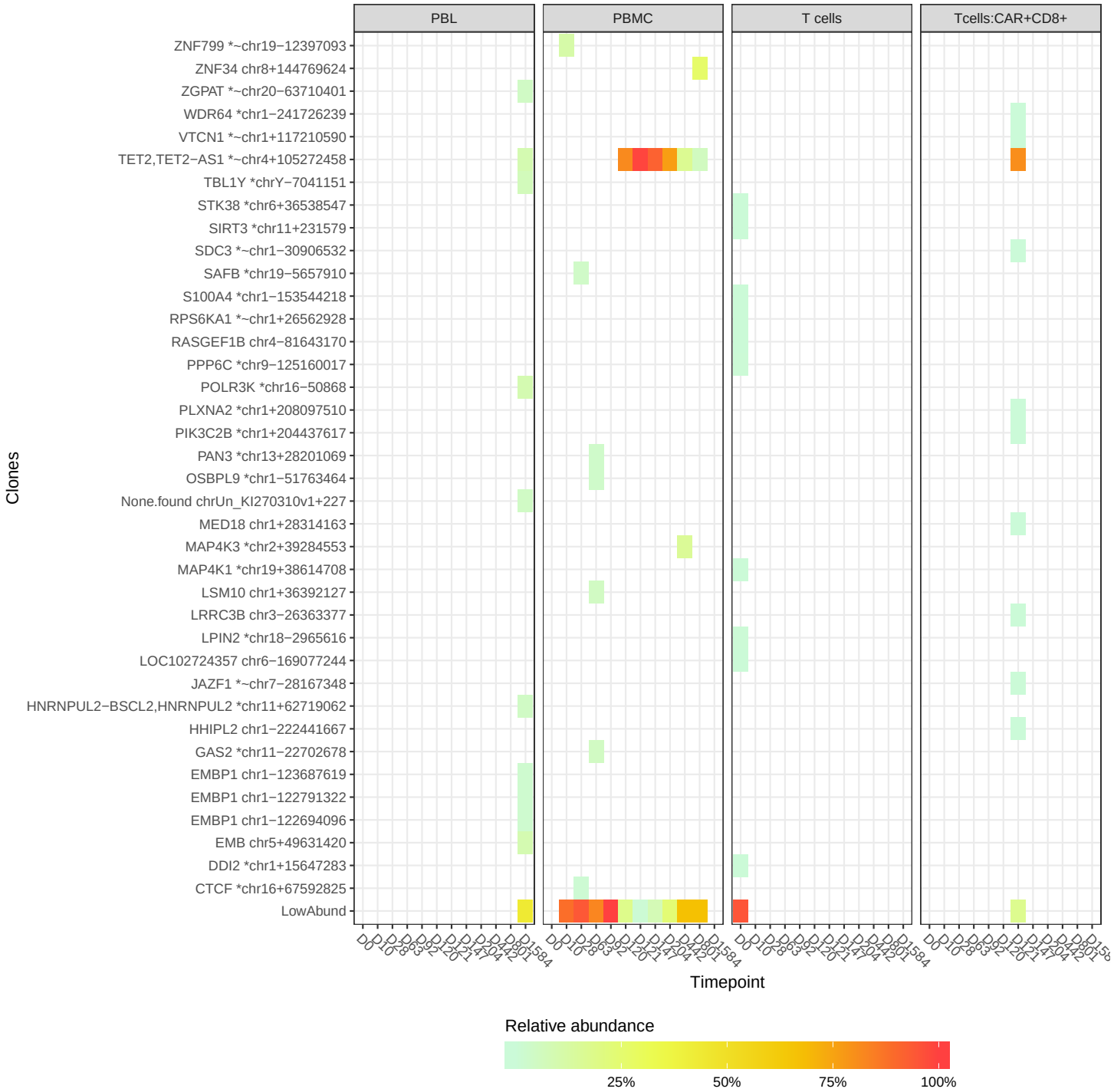
Integration sites near particular genes of interest

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.

T cells
D0 2:9

PBMC
D10 1:3

ZNF799 *~

PBMC
D28 1:17

PBMC
D63 1:6

PBMC
D92 1:1

PBMC
D120 1:88

TET2,TET2-AS1 *~

PBMC
D121 1:278

Tcells:CAR+CD8+
D121 1:448

PBMC
D147 1:289

PBMC
D204 1:42

PBMC
D442 1:13

PBMC
D801 1:23

PBL
D1584 1:4

Methods

All coordinates are on human genome draft hg38.

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)

Report generation software:
- subjectReport v0.1 (http://github.com/everettJK/geneTherapySubjectReport)