REINHARDT, HABIB ET AL. SUPPLEMENTAL TABLES AND FIGURES

SUPPLEMENTAL FIGURES

Supplemental Figure 1. *Growth Curves.*

Supplemental Figure 2. *Correlation of Shannon Diversity Index and mean granulocyte vector copy number (VCN).*

Supplemental Figure 3. *Unique Vector Integration Sites per Patient*

Supplemental Figure 4. *Analysis of clones with MECOM-adjacent integrants.*

Supplemental Figure 5. *Median absolute lymphocyte counts (ALC) versus (A) median PBMC ADA enzyme activity and (B*) age at gene therapy.

Supplemental Figure 6. *Correlation of TCR Repertoire Diversity at t1 vs. t2.*

Supplemental Figure 7. *Correlation of TCR Repertoire Diversity and Gene Marking (VCN) in PBMC and Granulocytes.*

Supplemental Figure 8. *Correlation between Subject Age and Clone Count.*

SUPPLEMENTAL TABLES

Supplemental Table 1. MND ADA Vector Integration Site Data (Gene Details)

Supplemental Table 2. Coefficient of Variation (sd/mean) for Lymphocyte Subsets.

Supplemental Table 3. Vaccine administrations and titer responses

Supplemental Table 4. *TRBV* Sequencing Meta-Data

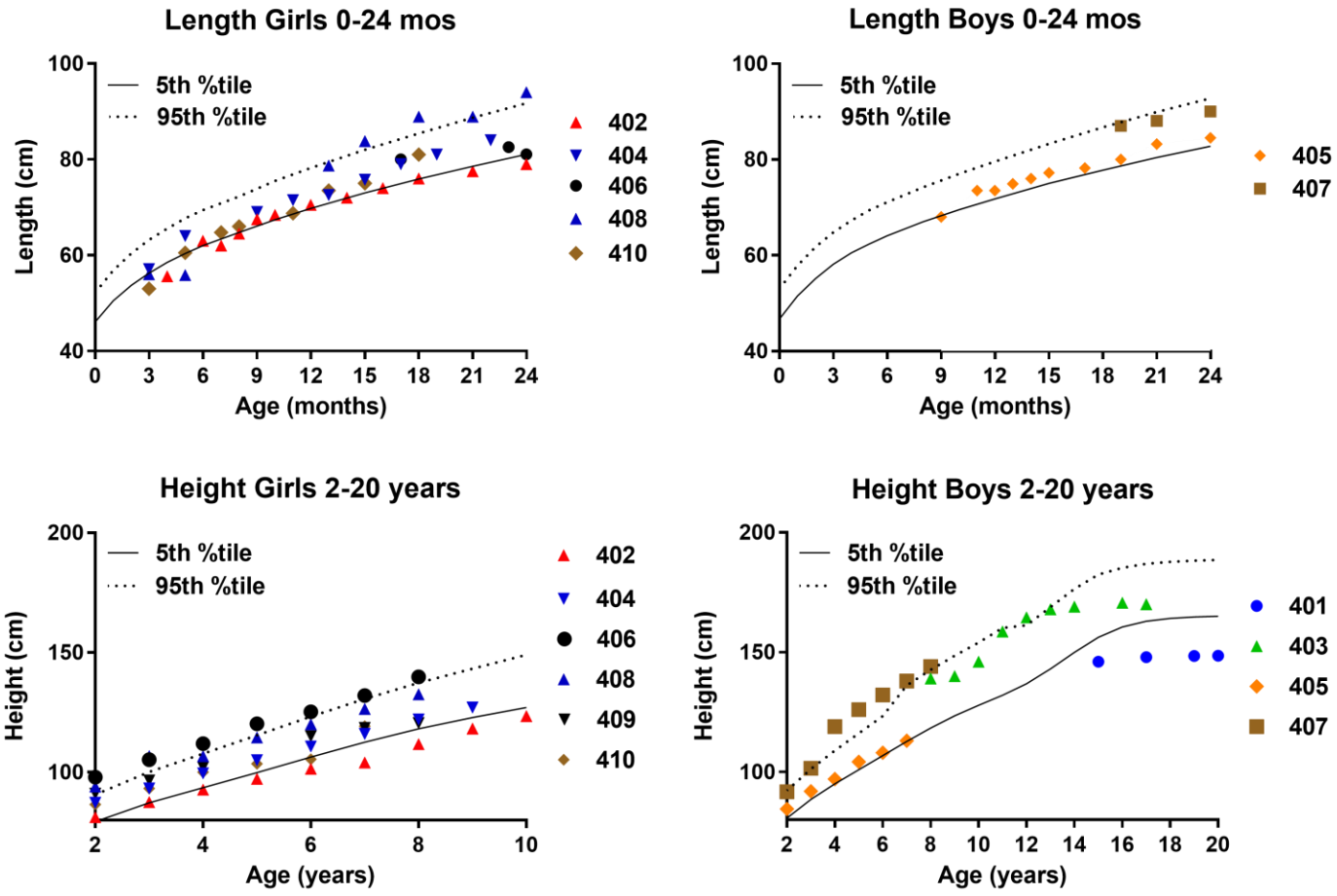
Supplemental Table 5. CD34+ cell dose, cell product vector copy number, and busulfan conditioning intensity.

Supplemental Table 6. Correlations between cell dose and AUC with granulocyte VCN.

Analysis of integration site distributions and relative clonal abundance

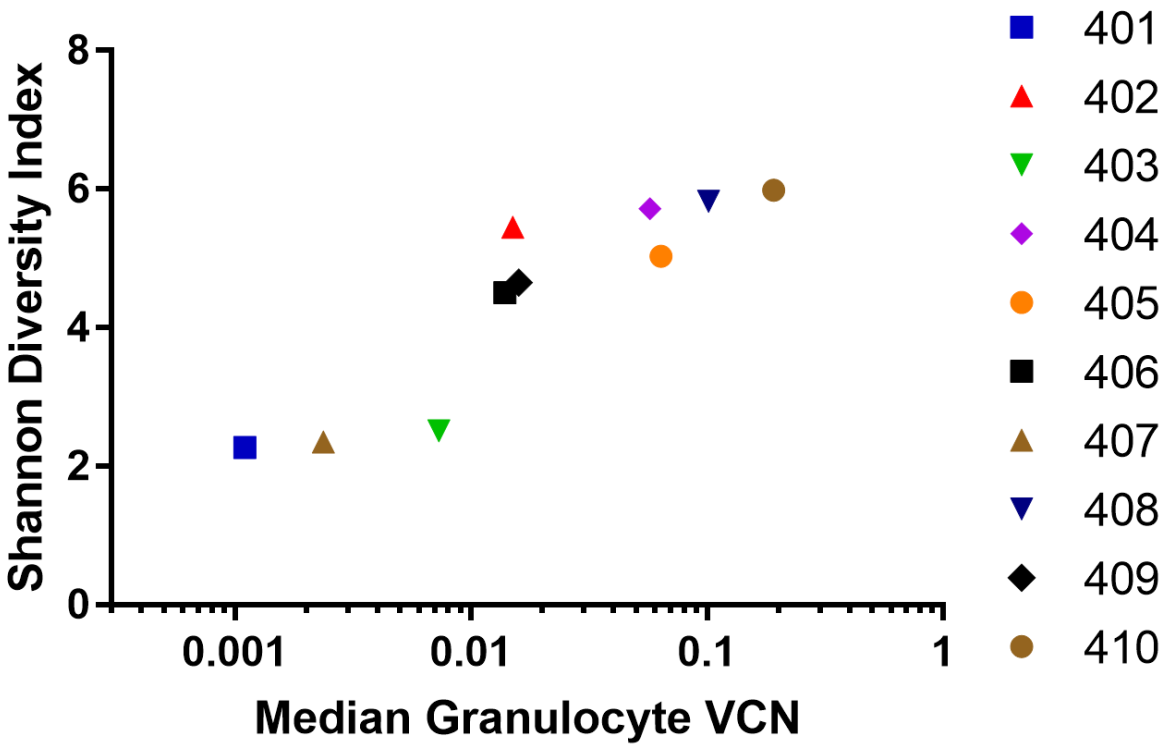
Vector Integration Site Analysis - full report

Supplemental Figure 1. Growth Curves.



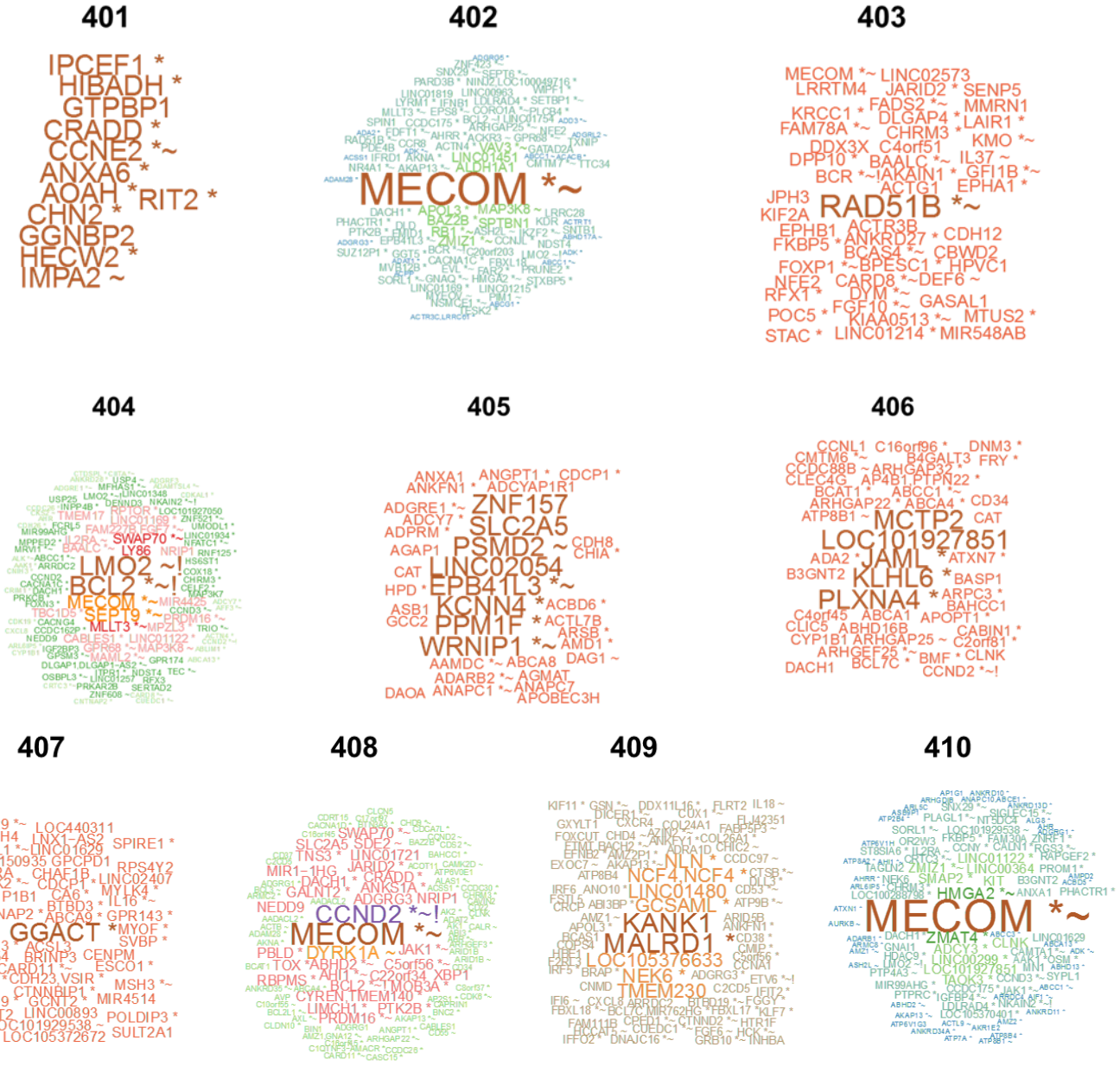
Supplemental Figure 1. *Growth Curves.* Patient length (0-24 months) or height (2-20 years) measured at clinic visits are shown compared to age-related normal ranges. Black line is 5th and dashed line is 95th percentiles, based on Centers for Disease Control and Prevention Clinical Growth Charts.

Supplemental Figure 2.
Granulocyte VCN vs Shannon Diversity Index



Supplemental Figure 2. Granulocyte VCN vs. Shannon Diversity Index. The Shannon Diversity Index for the unique vector integration sites in each subject are plotted against the mean of the vector copy number (VCN) values in their granulocyte samples from 24 months after gene therapy through follow-up.

Supplemental Figure 3. *Unique Vector Integration Sites in each Patient*



Supplemental Figure 3. *Unique Vector Integration Sites in each Patient.* The relative sizes of the gene names indicate their relative abundance. The asterisk indicates integration within the transcription unit of the named gene. The tilde indicates present in a broad list of cancer-associated gene (allOnco). The exclamation mark indicates gene associated with a list of 38 human lymphoma-associated genes. Gene lists: http://www.bushmanlab.org/links/genelists

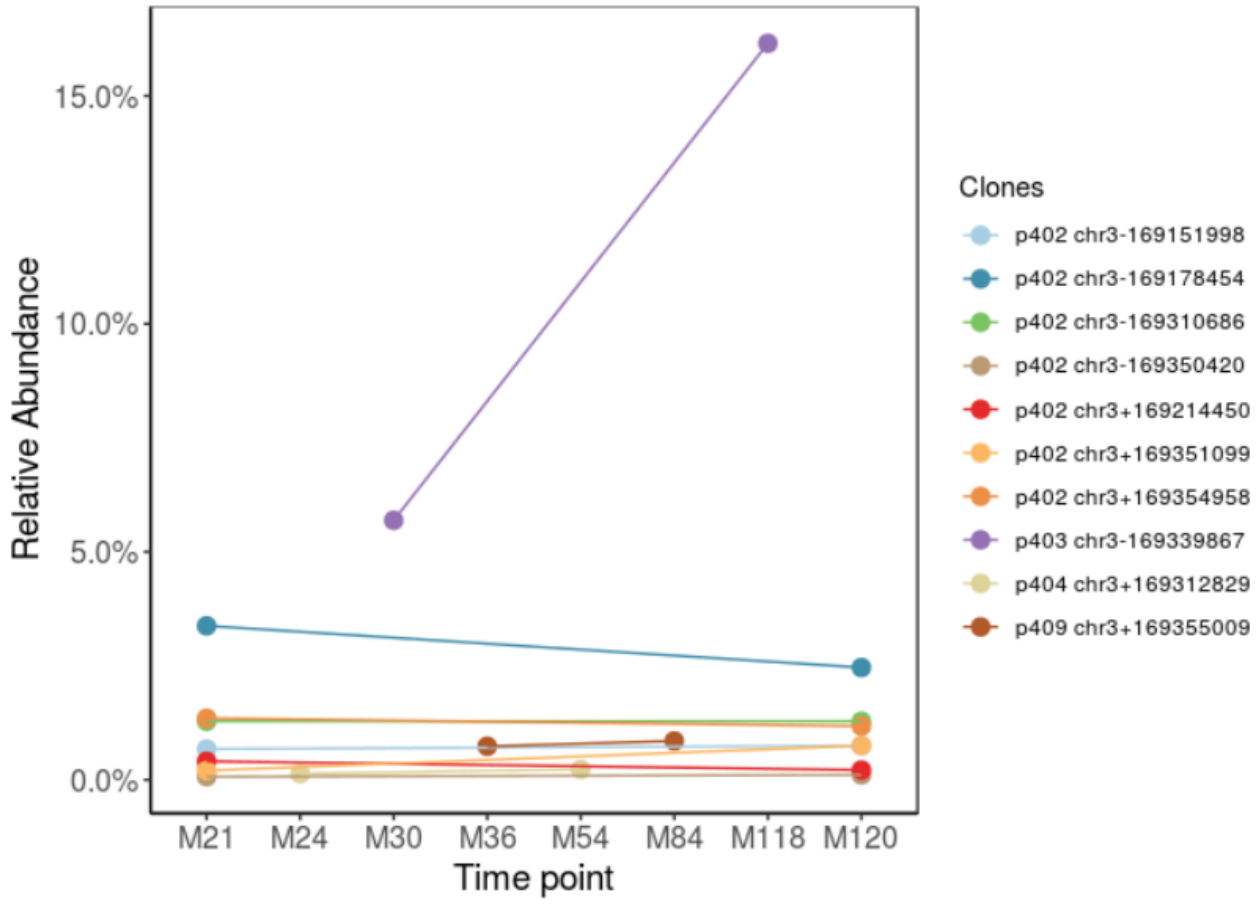
Supplemental Figure 4. *Analysis of clones with MECOM-adjacent integrants.*

Each patient in the study has two PBMC time points which we will call early and late. The matrix below shows the number of early and late sites near MECOM.

	Early	Late
Not MECOM	3172	2384
MECOM	23	29

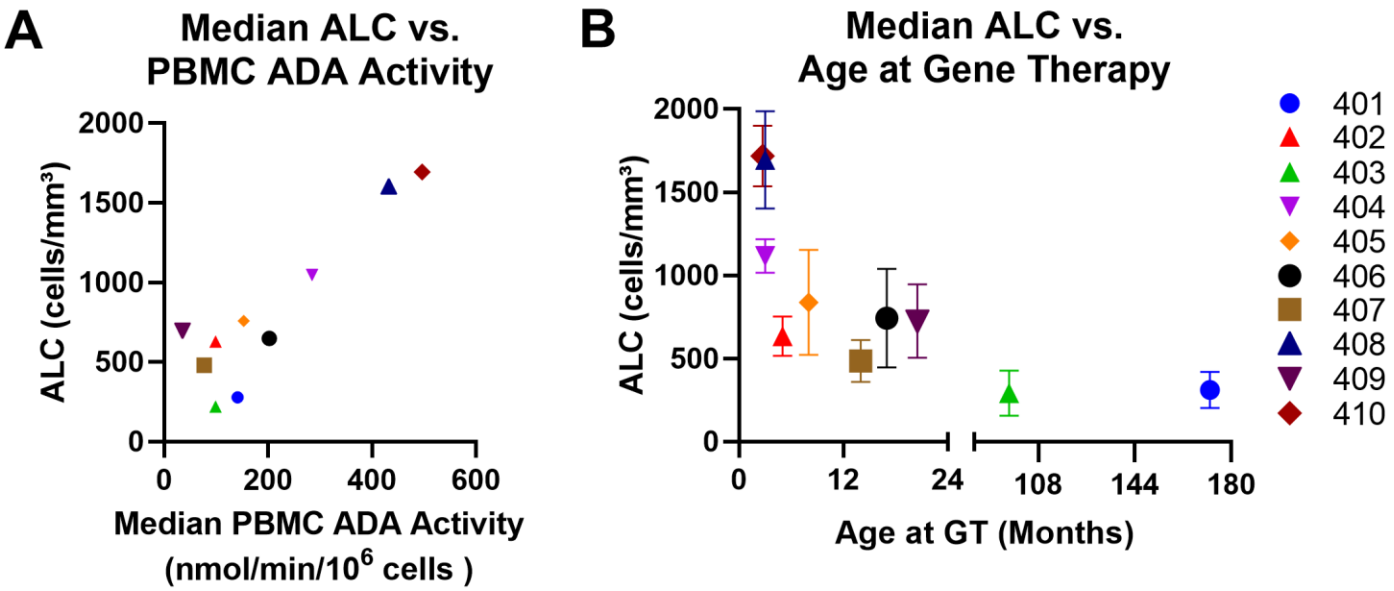
A Fishers two-sided test of this matrix yields a p-value of 0.068 and a one-sided test yields a p-value of 0.043 (later time points have more MECOM sites).

One MECOM clone expanded in frequency between the early and late PBMC samples. This can be seen graphically below. This version considers all MECOM clones with sufficient replicate level data (> 50 cells per replicate). Only samples with more than 50 inferred cells (all replicates combined) were considered.

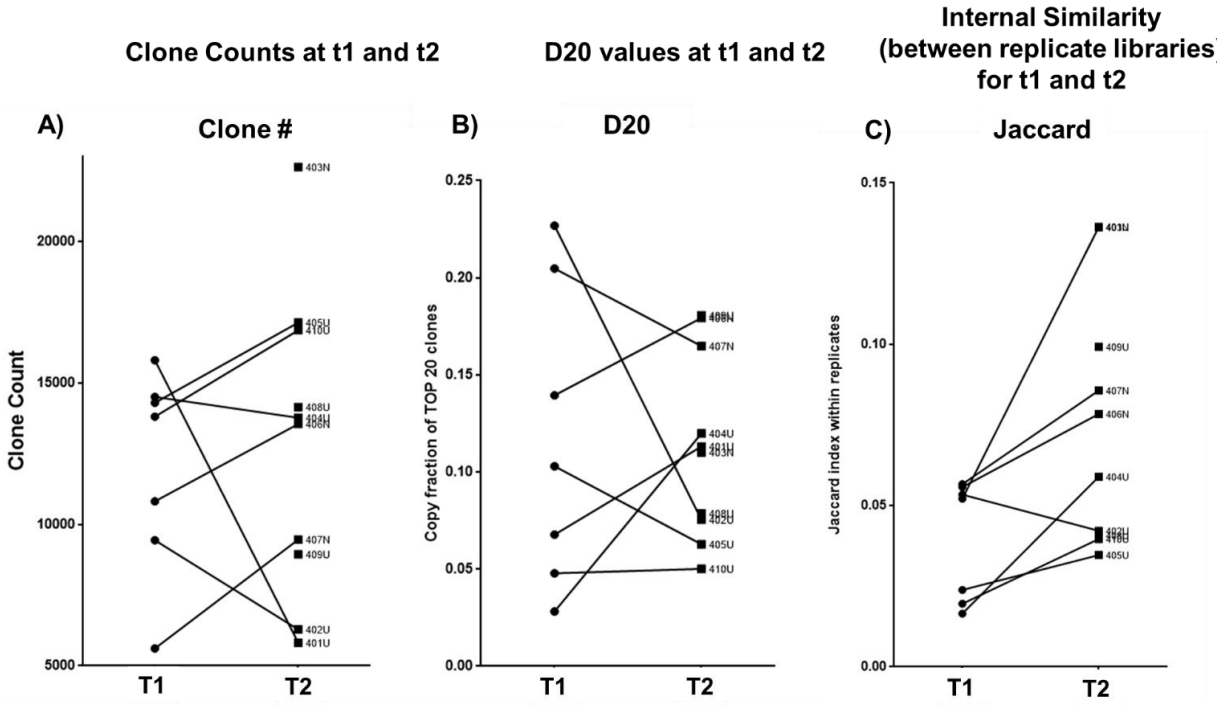


This expansion is not due to a single replicate based on analysis which compares the means of replicate relative abundances between early and later time points. The expanding clone in p403 yielded a significant Wilcox p-value (0.012).

Supplemental Figure 5. *Median absolute lymphocyte counts (ALC) versus (A) median PBMC ADA enzyme activity and (B*) age at gene therapy. (A) The median absolute lymphocyte counts measured in patients beyond two years after GT are plotted against their median PBMC ADA activity. (B) The median absolute lymphocyte counts measured in patients beyond two years after GT are plotted against the patients' ages at the time of GT.



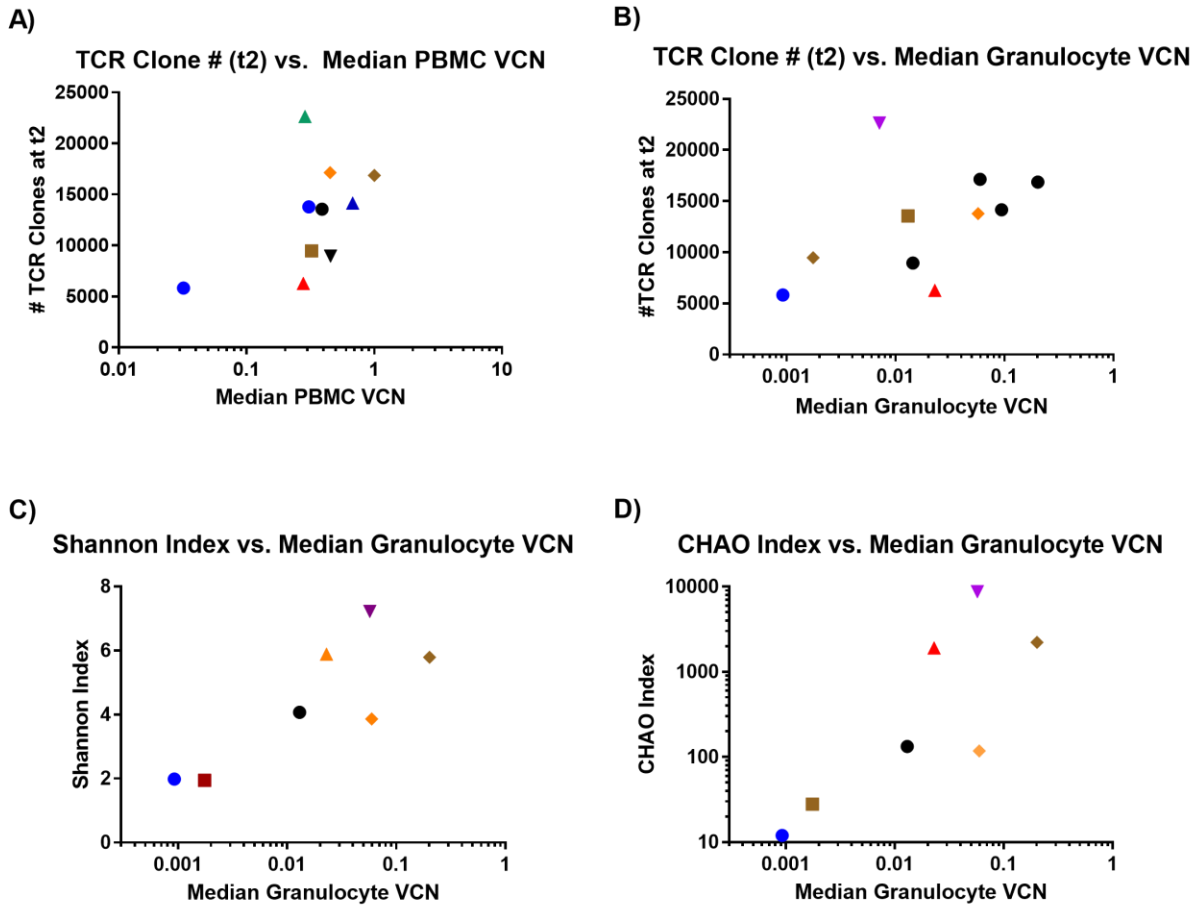
Supplemental Figure 6. *Lack of correlation of TCR repertoire diversity at t1 vs. t2.*



TCR data analysis methods. Raw reads from the Illumina MiSeq were processed as described previously.[24] IgBLAST v1.17.0 [10.1093/nar/gkt382] was used to annotate TRBV- and TRBJ-genes which were then imported into ImmuneDB v0.29.9 [10.3389/fimmu.2018.02107] for clonal inference and downstream analysis. Sequences sharing the same V-gene, J-gene, CDR3 length, and CDR3 amino-acid sequence were grouped into clones. Sequences with only one copy in each subject were excluded from clones to avoid the inference of spurious clones due to sequencing errors. Raw sequencing data are available on SRA under accession number PRJNA716857.

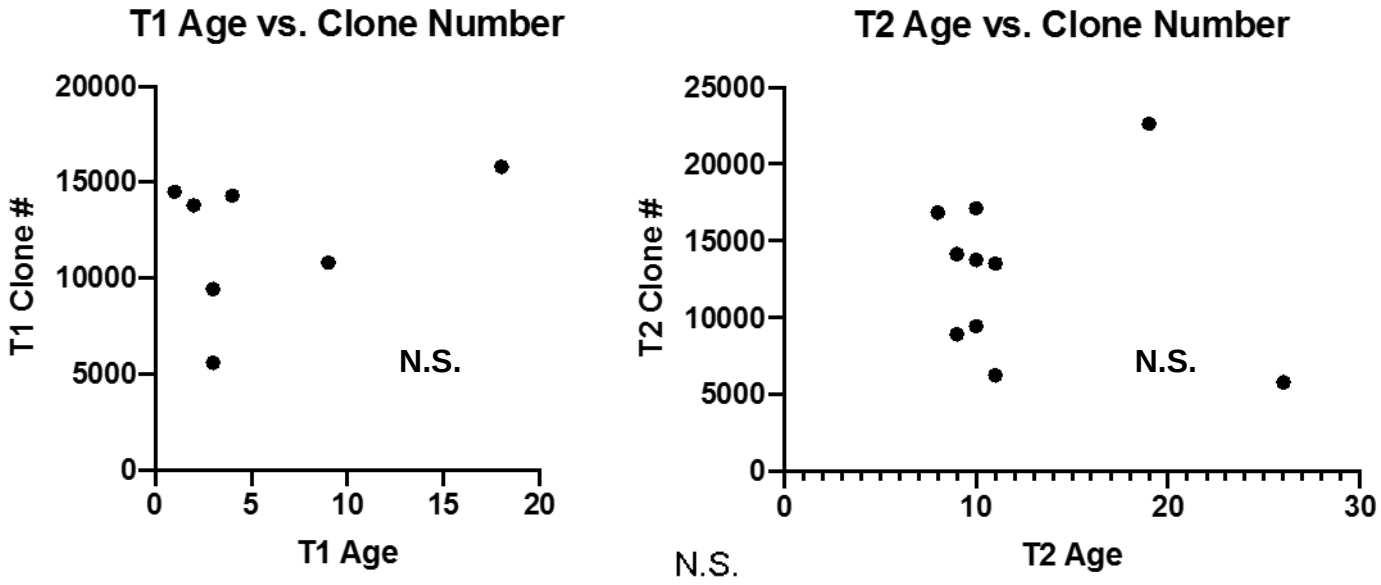
Results: To determine if diversity of the repertoire at t1 correlated with engraftment diversity at t2, we studied seven patients in whom paired data were available in detail. Lines connect samples from different time points from the same subject. **(A).** Clone count per sample vs. time point. **(B).** The D20 index is the sum of sequence copies in the top twenty ranked clones divided by all of the sequence copies in the library. The D20 serves as a metric of the contribution of very large clones to the overall repertoire. (**C**). Jaccard Index (fraction of overlapping clones divided by the total unique clones) in replicate sequencing libraries generated from each sample, which are separately amplified from gDNA. The Jaccard index between replicate libraries from the same sample correlates with the fraction of large clones that are readily resampled.

Supplemental Figure 7. *Correlation of TCR Repertoire Diversity and Gene Marking (VCN) in PBMC and Granulocytes.* The diversity of TCR repertoire was examined based on the number of unique TCR sequence clones or using the Shannon Diversity Index or the Chao estimator of diversity and correlated to the median VCN in each subject's PBMC or granulocytes 2 years and later after gene therapy. While the PBMC VCN did not correlate (A), there were non-significant trends between the granulocyte VCN and the clone # (B), the Shannon Diversity Index (C) and the Chao Diversity Index (D).



Supplemental Figure 8. *Correlation between Subject Age and Clone Count.*

The number of TCR clones per sample did not correlate with age at either time point. (p> 0.05 by Spearman correlation; N.S. = not significant).



SUPPLEMENTAL TABLES

Supplemental Table 1. MND ADA Vector Integration Site Data (Gene Details)

A. Clones at >20% frequency.

B. Clones with integrants near retroviral common integration site proto-oncogenes (LMO2, IKZF1,CCND2, HMGA2 or MECOM).

Supplemental Table 2. Coefficient of Variation (sd/mean) for Lymphocyte Subsets.

Supplemental Table 3. Vaccine administrations and titer responses

Supplemental Table 4. CD34+ cell dose, cell product vector copy number, and busulfan conditioning intensity.

Supplemental Table 5. Statistical table on cell dose and AUC correlations with granulocyte VCN.

Supplemental Table 1A. MND ADA Vector Integration Site Data (Gene Details)

Patient	Time (Mo)	VCN	Unique Int	Chao 1	Shannon	>20% Frequency Clones		
401	24	0.010	9	12	1.98	CRADD* chr12-93739144	MEX3B* chr15-82018028	
	82	0.105	27	50	2.27			
402	21	0. 742	729	1,909	5.89			
	120	0.476	430	1,100	5.45			
403	30	0.195	26	33	2.54	PPP1R1A* chr12+54304228		
	118	0.545	47	72	2.51			
404	24	0.686	1,725	8,750	7.22			
	54	0.540	353	1,573	5.72			
405	42	0.490	67	118	3.86			
	60	0.496	225	675	5.03			
406	24	0.172	74	133	4.07			
	84	0.483	196	384	4.58			
407	24	0.401	25	28	1.94	CA6* chr1+8947510	FGF14* chr13+100586636	ZDHHC3* chr3+45106737
	96	0.519	50	221	2.35		FGF14* chr13+100586636	
408	42	1.240	114	465	4.65			
	84	1.280	373	1113	5.82			
409	36	0.287	73	191	4.02			
	84	0.627	247	701	4.59			
410	24	1.130	342	2,220	5.79			
	72	0.91	442	2,068	5.98			

Supplemental Table 1B. MND ADA Vector Integration Site Data (Gene Details)

Patient	Time (Mo)	VCN	Near Retroviral Common Integration Site Proto-Oncogenes (LMO2, IKZF1, CCND2, HMGA2 or MECOM)					
401	24	0.010	No					
	82	0.105	No					
402	21	0. 742	MECOM* (1) 3.6%	MECOM* (2) 1.25%	MECOM* (3) 1.27%			
	120	0.476	MECOM* (1) 2.5%	MECOM* (2) 1.25%	MECOM* (3) 1.24%			
403	30	0.195	MECOM* 5%					
	118	0.545	MECOM* 16.5%					
404	24	0.686	No					
	54	0.540	No					
405	42	0.490	No					
	60	0.496	LMO2* 0.25%					
406	24	0.172	No					
	84	0.483	No					
407	24	0.401	No					
	96	0.519	No					
408	42	1.240	LMO2* (1) 0.74%	MECOM* (1) 1.4%	MECOM* (2) 0.74%	No	No	
	84	1.280	No	No	MECOM* (2) 0.8% MECOM* (3) 0.24% MECOM* (4, 5) 0.24%	LMO2* (2) 0.4%	HMGA2* 0.24%	
409	36	0.287	LMO2* (1) 0.73%	No	MECOM* 0.73%			
	84	0.627	No	LMO2* (2) 0.12%	MECOM* 0.85%			
410	24	1.130	LMO2* (1) 0.52%	LMO2* (2) 0.26%	HMGA2* (1) 0.26%	No	No	No
	72	0.91	No	No	No	HMGA2* (2) 0.75% HMGA2* (3) 019% HMGA2* (4) 0.38%	MECOM* (1,2) 0.56% MECOM* (3,4,5) 0.38%	LMO2* (3) 0.19%

Supplemental Table 2. Coefficient of Variation (sd/mean) for Lymphocyte Subsets.

Patient	401	402	403	404	405	406	407	408	409	410	Krishnamoorthy and.Lee test
CD3+	0.399	0.172	0.553	0.486	0.205	0.481	0.359	0.194	0.845	0.226	0.099
CD4+	0.340	0.135	0.544	0.241	0.194	0.372	0.287	0.159	NA	0.255	0.023
CD8+	0.637	0.223	0.573	0.318	0.356	0.542	0.378	0.212	1.139	0.321	0.222
CD19+	0.413	0.359	0.890	0.397	0.247	0.709	0.509	0.291	0.256	0.259	0.106
CD16/56 0.150	0.473	0.570	0.421	0.592	0.366	0.499	0.317	0.561	0.330	0.381	
CD4+CD45RA+	0.471	NA	0.943	0.170	NA	0.447	0.472	0.051	0.000	0.500	0.298

Supplemental Table 3

Vaccination results

After cessation of immunoglobulin replacement therapy (IgRT), patients were administered multiple vaccines, including Diphtheria/Tetanus/Pertussis (DTaP), *Haemophilus influenza* (Hib), meningococcus, pneumococcus, Varicella-Zoster Virus and annual influenza A. Patients 404 and 410 developed positive immune responses to immunizations, whereas 405 and 408 were not tested (**Supplemental Table 3**, below) . Patient 403 was administered three rabies vaccinations while continuing IgRT under a research trial (NCT #00023504) which was testing whether the rabies vaccine as a neo-antigen could assess humoral immunity for patients receiving immunoglobulin replacement. The antibody titer responses were below the detection limit (<0.1 IU/mL) and IgRT was continued. Patient 402 received a DTaP vaccination during an unsuccessful trial off IgRT. All the patients on IgRT received annual influenza vaccination after GT.

Supplemental Table 3. Vaccine administrations and titer responses

UPN	Vaccinations	Administration Date(s)	Titer Responses
402	DTaP	Unknown	
403	3 x Rabies	2017	5/22/17 - Rabies: <0.1 IU/mL
404	3 x DTaPs 3 x IPV 1 x Haemophilus B Influenza 2 x Hepatitis B 2 x Prevnar 13 Varicella and MMR Influenza (yearly)	2012 2012 2012 2012 2012 2014 2014, -15, -16, -17, -18,	9/16/14 - Tetanus IgG Ab: 5.20
405	Influenza (yearly) Influenza Hepatitis A Hepatitis B HPV MMR Pneumococcal Polio TDAP/TD Varicella	9/18/14, 9/15/15, 2016, 2/23/17, 3/14/18, 11/19/19 9/18/14 7/29/19, 2/21/20 7/29/19, 8/30/19, 7/7/20 7/29/19, 9/20/19, 7/7/20 9/20/19, 2/21/20 11/19/19, 2/21/20, 7/7/20 7/29/19, 8/30/19, 11/19/19, 7/7/20 7/29/19, 8/30/19, 7/7/20 9/20/19, 2/21/20	Not done
408	DTaP IPV Influenza (yearly)	2016 2016 10/12/17, 1/3/19	Not done.
409	Varicella	2015	Not done. Varicella followed post-vaccination. Hospitalized for IV acyclovir. Resolved.
410	DTaP IPV Hepatitis A Hepatitis B PCV13 Polio Varicella Influenza (yearly) Measles	9/14, 10/16 9/14 2014 2014 2014 10/16 5/18 2014, -15, -16, -17,-18, -19, 2020 2018	8/3/15 – Streptococcus Pneumoniae IgG Ab: >2.0 mcg/mL (14/23 serotypes) Haemophilus Influenza B IgG Ab: 0.82 mg/L Tetanus IgG Ab: 0.47 Poliovirus Type 1,2,3 Ab ≥1:8 Not tested IgG antibody index-7.1

Supplementary Table 4. *TRBV Sequencing Metadata*

sample_id	# reps	input DNA (ng) per rep	copies	clones	cdr3_num _nts	in_frame
401U-11yr	2	200	319219	5813	44.84	0.81
401U-36mo	2	200	261000	15807	44.27	0.83
402U-10yr	2	200	336336	6281	44.07	0.78
402U-30mo	2	200	269031	9438	44.00	0.80
403N-11yr	2	200	316299	22641	43.97	0.81
404U-21m	2	200	255359	14512	43.90	0.78
404U-9yr	2	200	289671	13773	44.03	0.78
405U-30mo	2	200	286665	14303	44.32	0.83
405U-8yr	2	200	288019	17136	44.23	0.83
406N-30mo	2	200	291603	10817	43.70	0.79
406N-9yr	2	200	285613	13554	43.75	0.79
407N-21mo	2	200	256170	5614	44.56	0.76
407N-9yr	2	200	295548	9464	44.40	0.79
408U-9yr	2	200	257551	14148	44.17	0.80
409U-8yr	2	200	337529	8938	44.33	0.79
410U-21mo	2	200	288414	13811	44.37	0.78
410U-7yr	2	200	311300	16869	44.17	0.80

Shown are the sample IDs which include the subject ID and time point. The earlier time point after gene therapy is t1 and the most recent sample corresponds to t2. T cell receptor rearrangements (beta chain) were amplified from genomic DNA extracted from bulk PBMCs. Data are pooled from two separately amplified replicates (reps). The amount of input DNA was 200 nanograms per replicate. Copies indicate the number of valid reads, clones are sequences that share the same TRBV, TRBJ, have the same length third complementarity determining region (CDR3) and have identical CDR3 amino acid sequences. Each clone is counted only once if it is present in more than one replicate. Sequences with fewer than two copies at the subject level are excluded from the clone count to reduce the effects of sequencing error. The CDR3 length is given in nucleotides (num_nts). The final column, in_frame, indicates the fraction of unique sequences with productive rearrangements (those that are in the correct reading frame and have no termination codons.)

Supplemental Tables 5 and 6

Correlation of levels of engraftment of gene-corrected cells (granulocyte VCN) with transplant factors – drug product cell dose and VCN, busulfan intensity

To assess the effects of important parameters of the GT transplant on the resulting granulocyte VCN, we examined CD34+ cells dose/kg, VCN in the CD34+ cell product, and the intensity of cytoreductive conditioning, based on measured busulfan AUC, as reported by Shaw et al (13) (**Supplemental Table 4**, below). Median VCN in granulocytes per subject correlated significantly with the CD34+ cell dose delivered (r_s= 0.67, p=0.04) (**Supplemental Table 5**, below**)**. To model the dose of *gene-modified* CD34+ cells delivered, the CD34+ cell dose/kg was multiplied by the VCN in each cell product. This computed dose of gene-modified cells also correlated with the granulocyte VCN, but did not reach statistical significance. (r_s= 0.61, p= 0.07). To determine if the intensity of cytoreductive conditioning the patients received, based on the measured busulfan levels, influenced engraftment of the gene-modified cells, the dose of transduced cells delivered was multiplied by the busulfan AUC for that patient. This composite index of cell and conditioning intensity correlated significantly with the median neutrophil VCN (r_s= 0.80, p= 0.01)**.** These findings suggest that cell dose, VCN in the drug product and intensity of cytoreductive conditioning are each important factors to achieve optimal engraftment of gene-corrected HSC.

Supplemental Table 5. CD34+ cell dose, cell product vector copy number, and busulfan conditioning intensity.

Patient	CD34+ cells/kg*	VCN*	CD34+ Cells/kg x VCN	Busulfan AUC* (µmol/L*min)	CD34+ Cells/kg x VCN x Busulfan AUC
401	0.6	0.6	0.36	n.d.	n.d.
402	1.7	0.71	1.21	5437	6579
403	0.92	0.18	0.16	3871	619
404	7.1	2.0	14.2	3532	50,154
405	7.62	2.0	15.24	5469	83,348
406	8	2.6	20.8	2427	50,482
407	5.6	2.31	12.94	3232	41,822
408	6.8	2.68	18.2	5344	97,261
409	2.9	1.22	3.54	5608	19,852
410	8.4	2.38	20.0	6714	134,280
Median	**4.964**	**1.668**	**10.665**	**4626**	**53091.08**

*From Table2, Shaw et al JCI 2017.

Supplemental Table 6. *Correlations between cell dose and AUC with granulocyte VCN.*

Median VCN Correlates	Spearman Correlation	P Value
Median CD34+/kg	0.673	0.039
Median CD34+/kg x VCN	0.612	0.066
Median CD34+/kg x VCN x AUC	0.088	0.014

Analysis of integration site distributions and relative clonal abundance for subject p401

November 02, 2020

Contents

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**

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	Rich
M24	9	No
M84	27	No

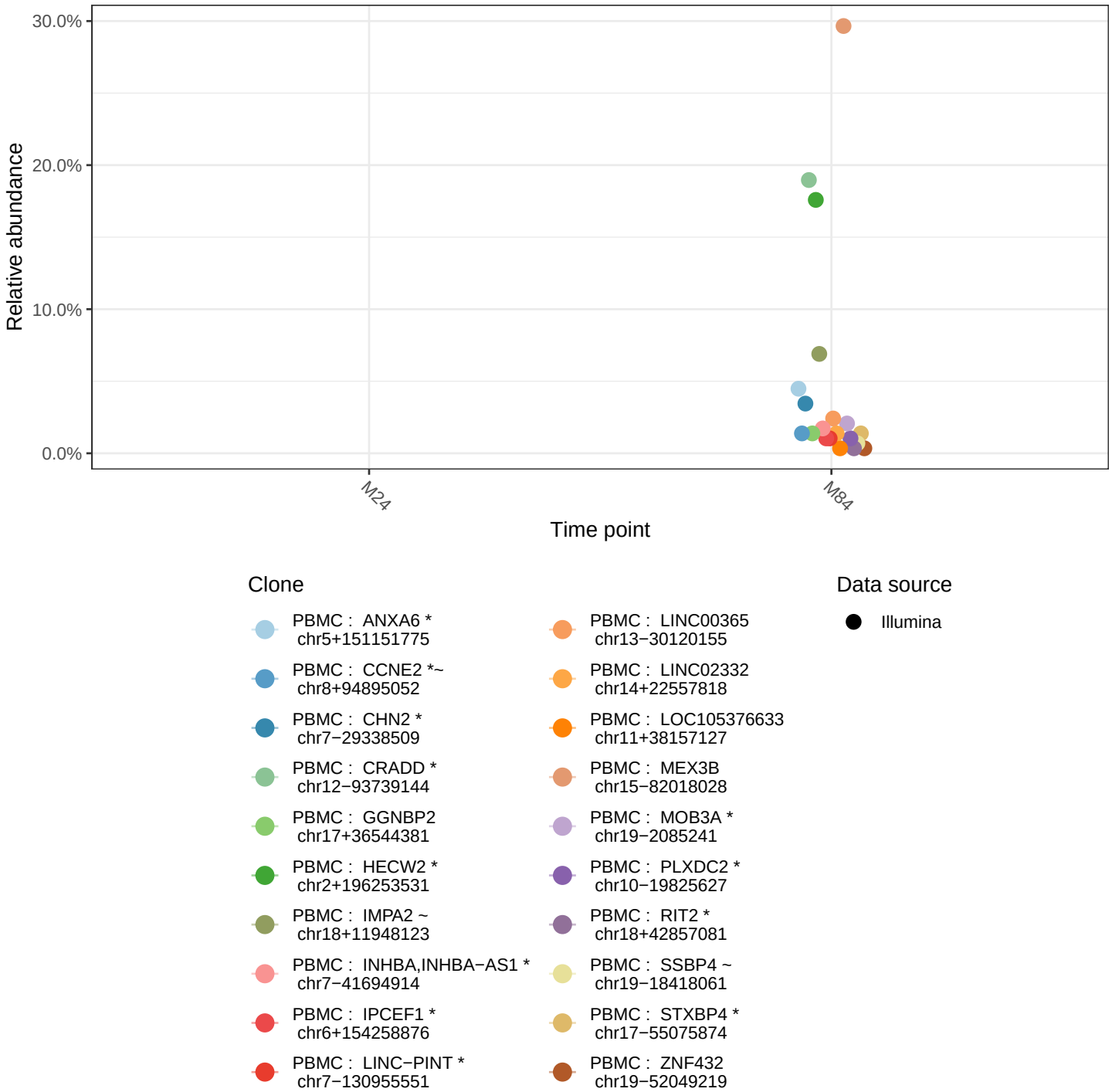
Do any cell clones account for more than 20% of all clones?

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)
chr15-82018028	86	29.7%	M84	PBMC	MEX3B	23.70	IL16	705.30

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 p401 over time points M24, M84 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

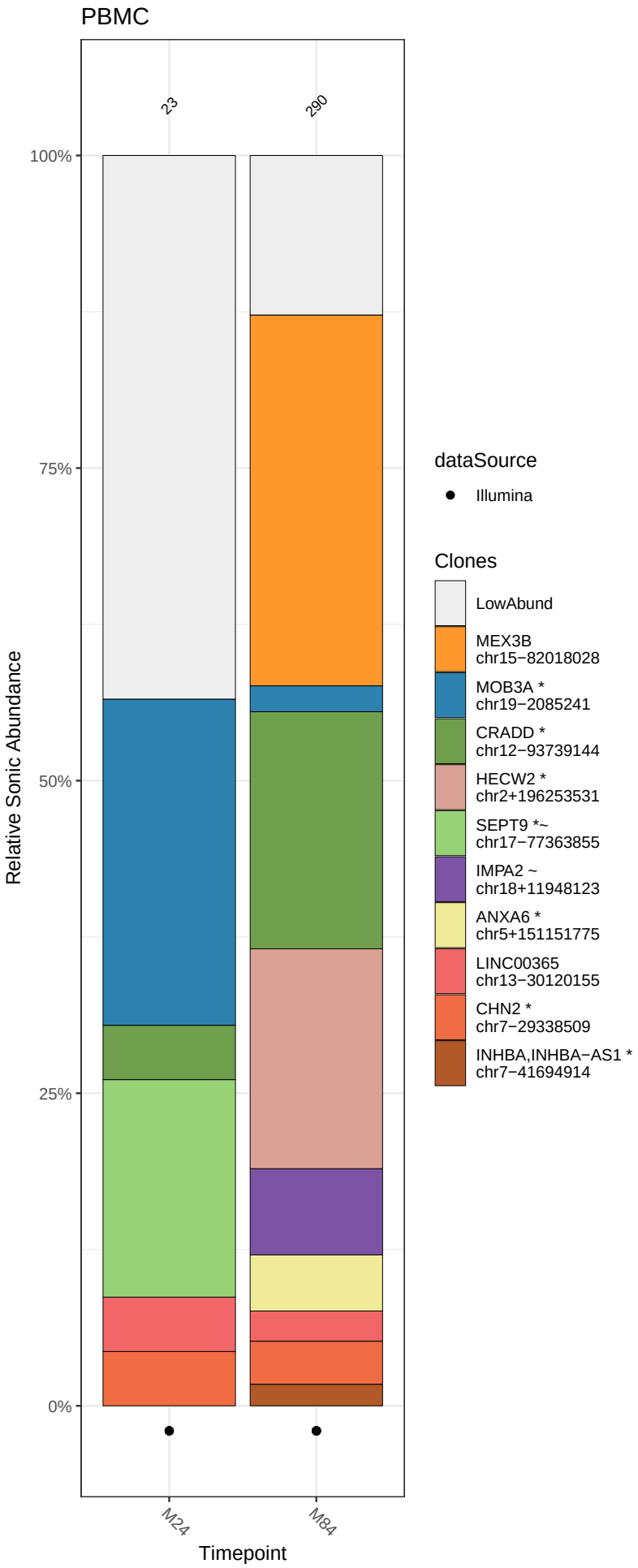
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3597	Illumina	M24	PBMC	222,308	23	9	0.357	12	1.98	0.902	3	yes	2020-10-28	0.010
GTSP3598	Illumina	M84	PBMC	1,001,982	290	27	0.712	50	2.27	0.689	3	yes	2020-10-28	0.105

Tracking of clonal abundances

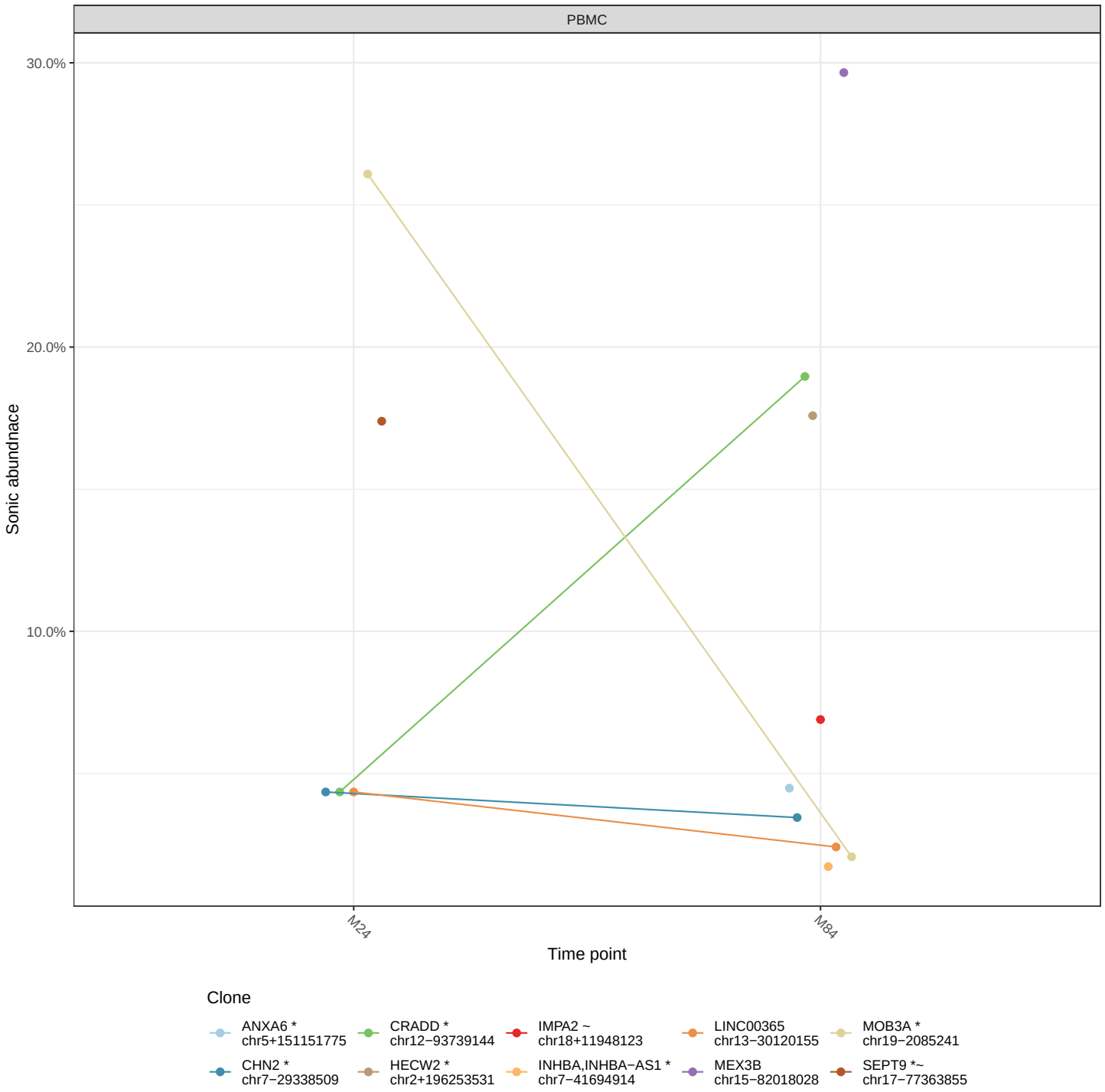
Relative abundance of cell clones

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
23
290
100%
75%
50%
25%
0%
Relative Sonic Abundance
M24
M84
Timepoint
dataSource
Illumina
Clones
LowAbund
MEX3B
chr15−82018028
MOB3A *
chr19−2085241
CRADD *
chr12−93739144
HECW2 *
chr2+196253531
SEPT9 *~
chr17−77363855
IMPA2 ~
chr18+11948123
ANXA6 *
chr5+151151775
LINC00365
chr13−30120155
CHN2 *
chr7−29338509
INHBA,INHBA−AS1 *
chr7−41694914

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



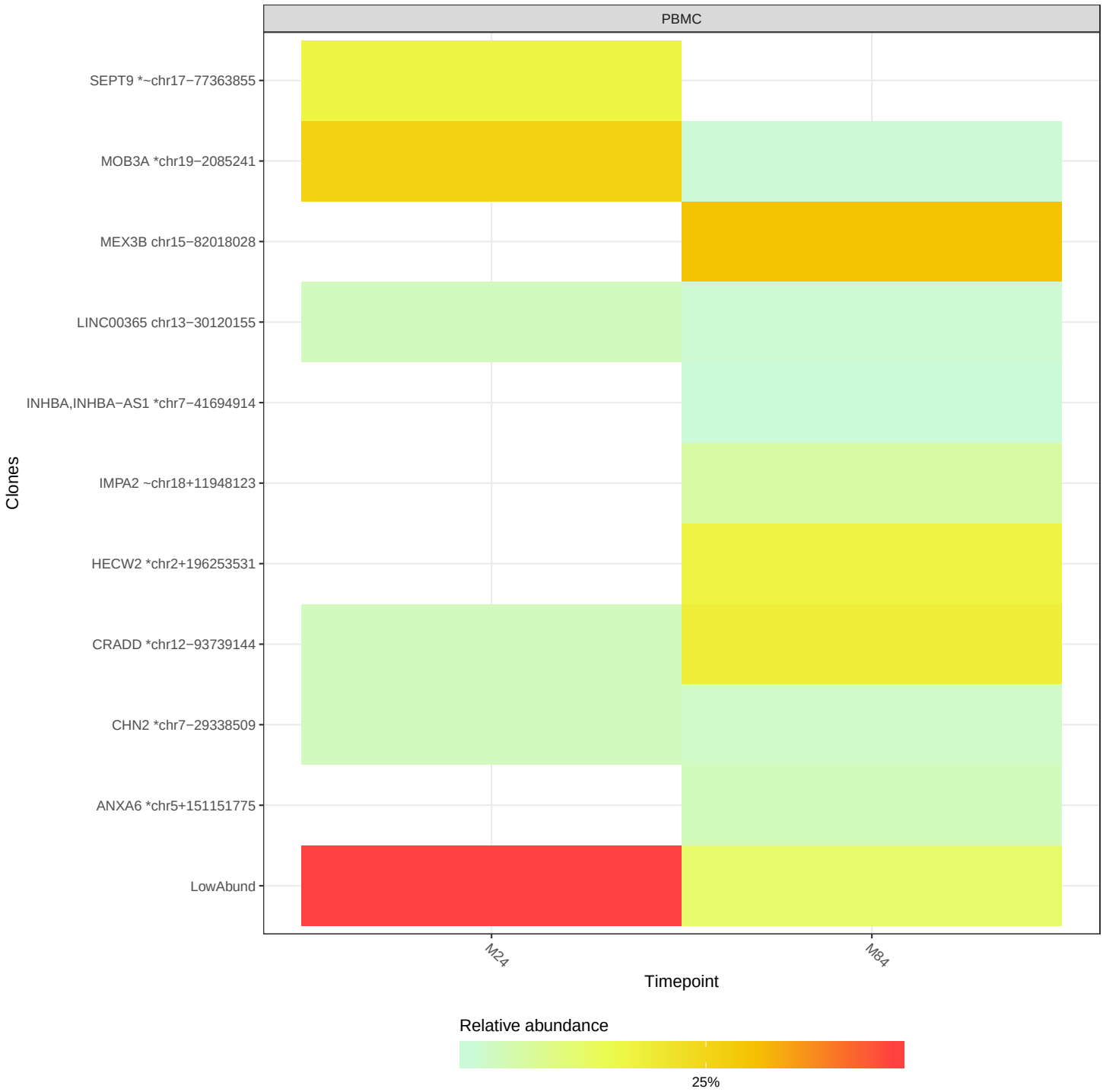
Integration sites near particular genes of interest

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
M24 1:6

PBMC
M84 1:86

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

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)

Analysis of integration site distributions and relative clonal abundance for subject p402

November 02, 2020

Contents

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**

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 (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	Rich
M21	729	No
M120	430	No

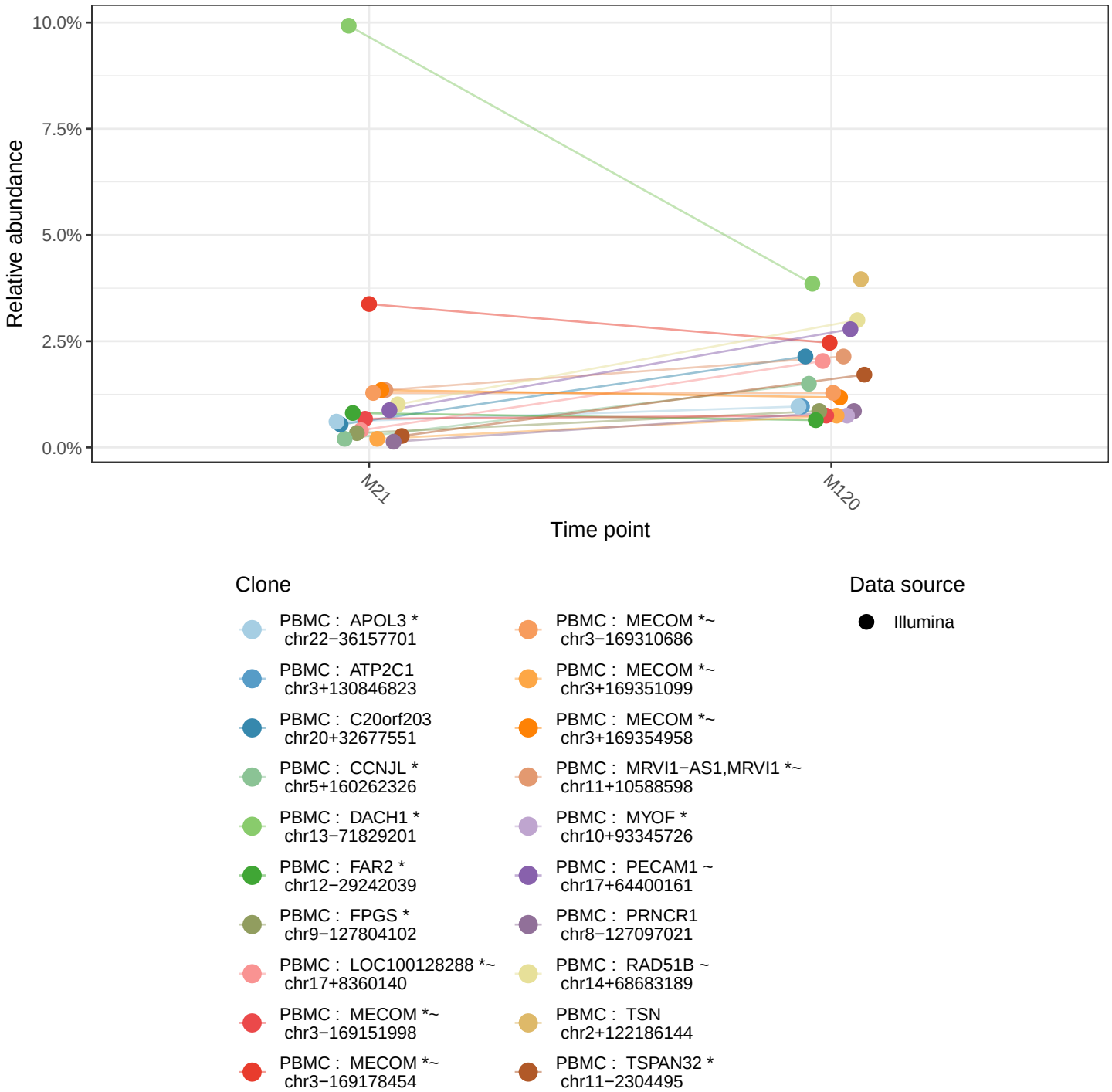
Do any cell clones account for more than 20% of all clones?

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 p402 over time points M21, M120 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

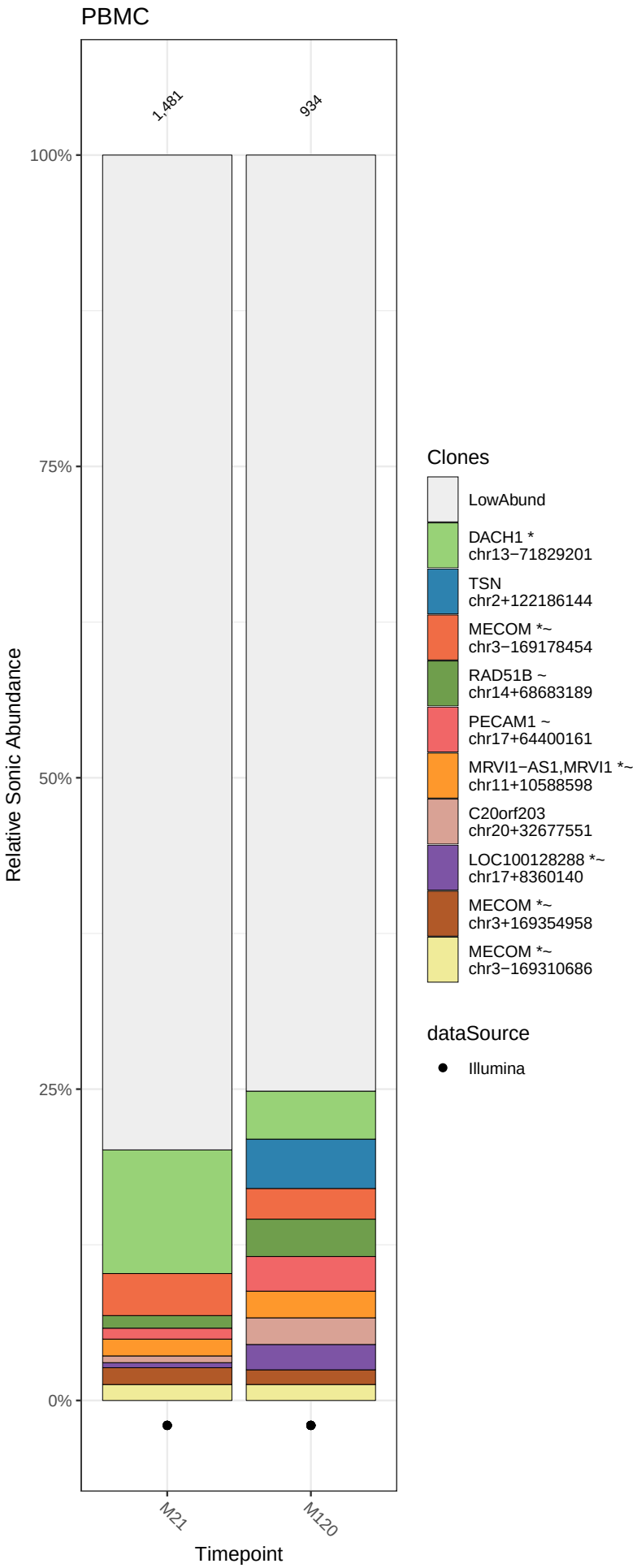
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3599	Illumina	M21	PBMC	763,881	1,481	729	0.444	1,909	5.89	0.894	106	yes	2020-10-12	0.742
GTSP3600	Illumina	M120	PBMC	601,813	934	430	0.471	1,100	5.45	0.899	55	yes	2020-10-12	0.476

Tracking of clonal abundances

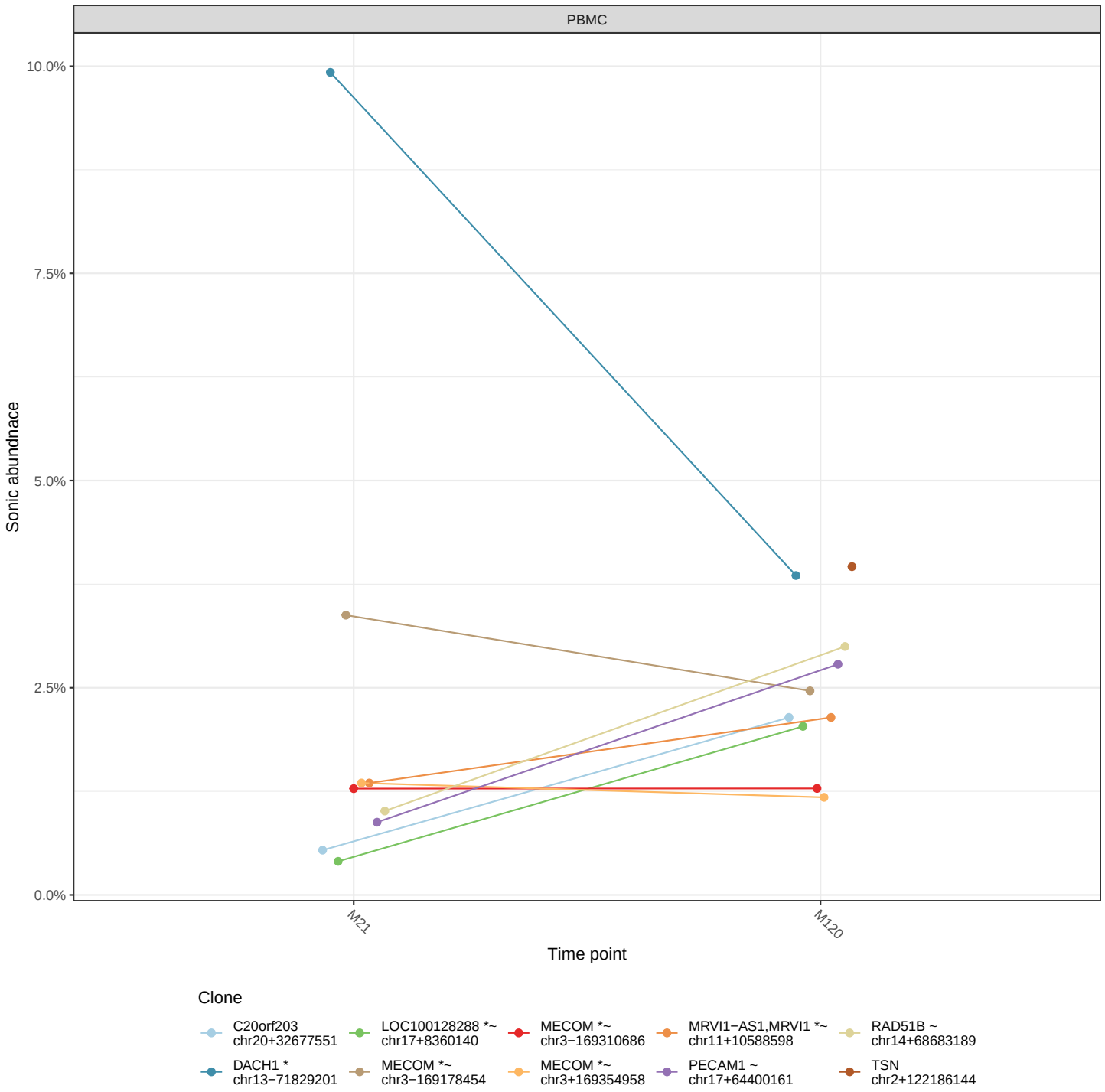
Relative abundance of cell clones

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
1,481
934
100%
75%
50%
25%
0%
Relative Sonic Abundance
M21
M120
Timepoint
Clones
LowAbund
DACH1 *
chr13−71829201
TSN
chr2+122186144
MECOM *~
chr3−169178454
RAD51B ~
chr14+68683189
PECAM1 ~
chr17+64400161
MRVI1−AS1,MRVI1 *~
chr11+10588598
C20orf203
chr20+32677551
LOC100128288 *~
chr17+8360140
MECOM *~
chr3+169354958
MECOM *~
chr3−169310686
dataSource
Illumina

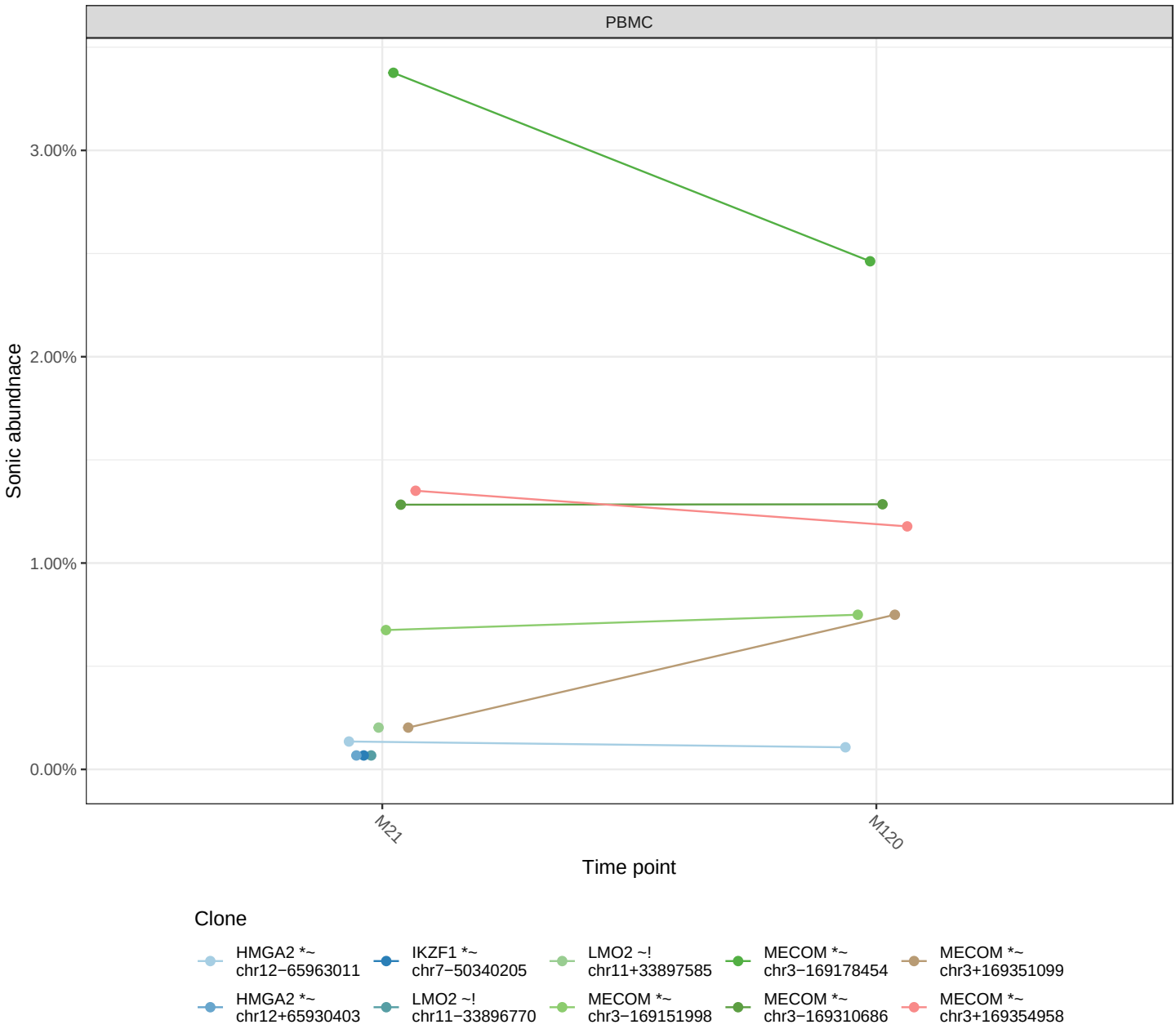
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



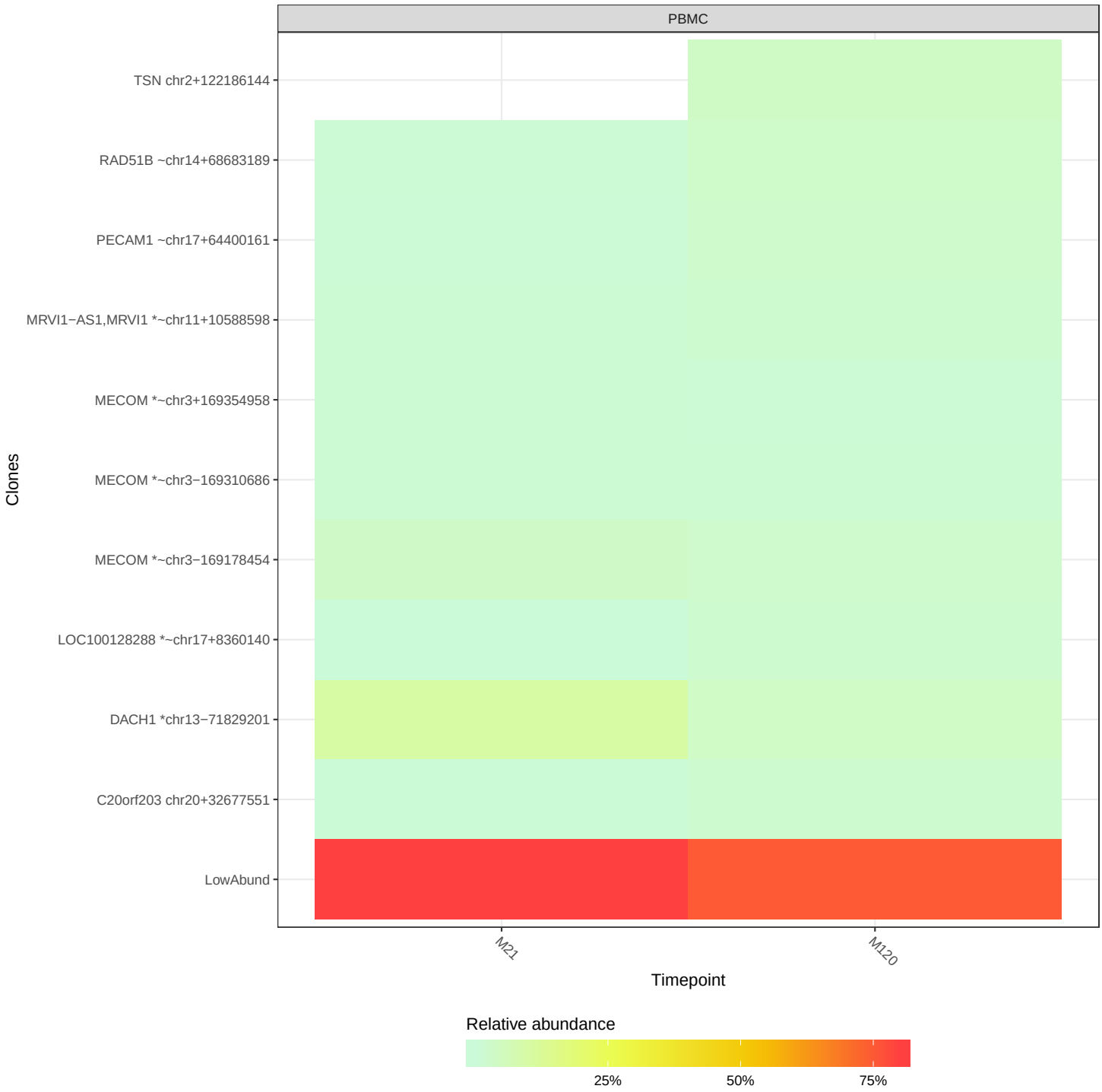
Integration sites near particular genes of interest

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.

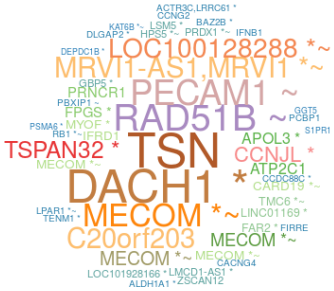


What are the most frequently occuring gene types in the subject?

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.

PBMC
M21 3:147

PBMC
M120 2:37



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

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)

Analysis of integration site distributions and relative clonal abundance for subject p403

November 02, 2020

Contents

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**

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 (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	Rich
M30	38	No
M118	63	No

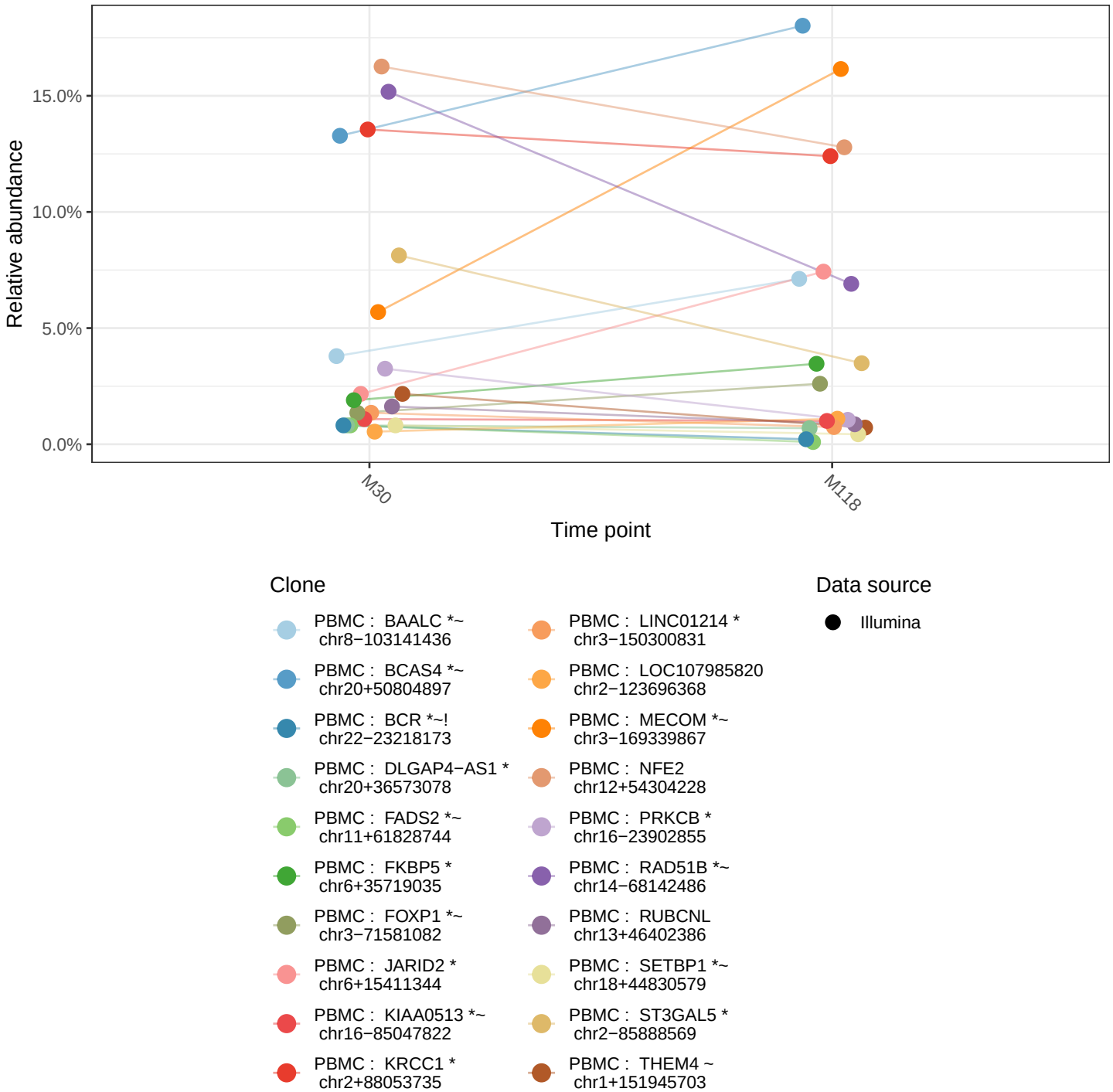
Do any cell clones account for more than 20% of all clones?

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 p403 over time points M30, M118 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

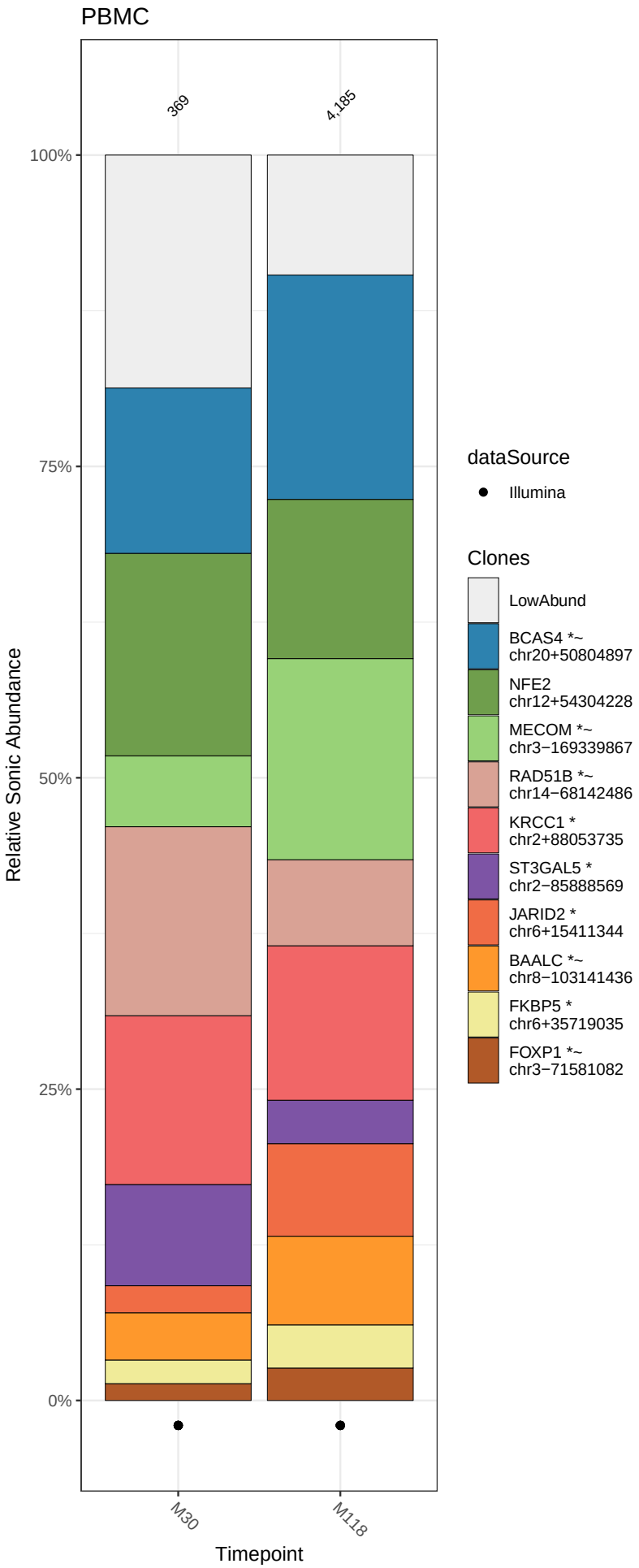
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3601	Illumina	M30	PBMC	1,251,244	369	38	0.697	68	2.69	0.741	4	yes	2020-10-28	0.195
GTSP3602	Illumina	M118	PBMC	991,798	4,185	63	0.847	131	2.54	0.613	4	yes	2020-10-28	0.545

Tracking of clonal abundances

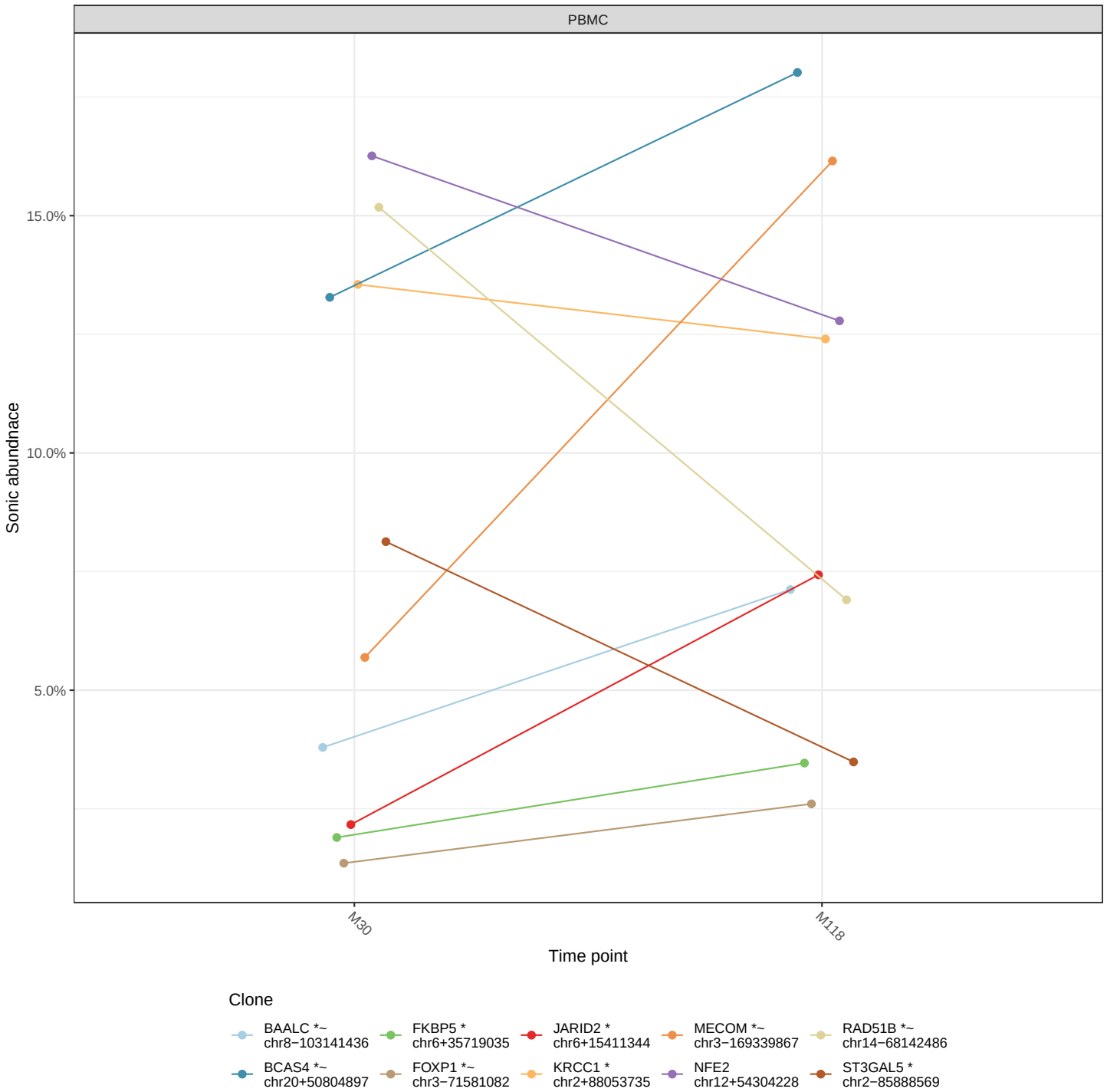
Relative abundance of cell clones

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
369
4,185
100%
75%
50%
25%
0%
Relative Sonic Abundance
M30
M118
Timepoint
dataSource
Illumina
Clones
LowAbund
BCAS4 *~
chr20+50804897
NFE2
chr12+54304228
MECOM *~
chr3−169339867
RAD51B *~
chr14−68142486
KRCC1 *
chr2+88053735
ST3GAL5 *
chr2−85888569
JARID2 *
chr6+15411344
BAALC *~
chr8−103141436
FKBP5 *
chr6+35719035
FOXP1 *~
chr3−71581082

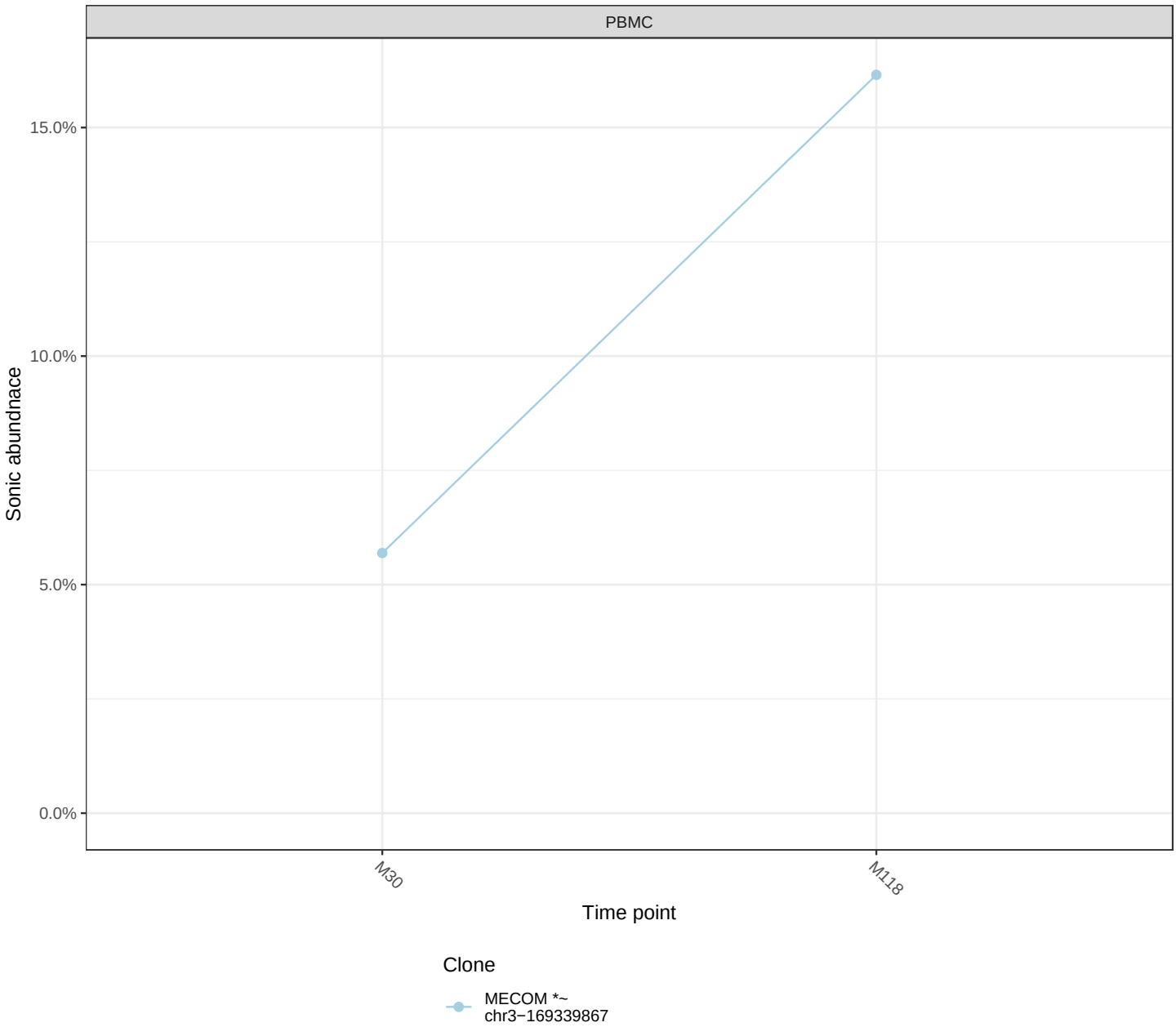
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



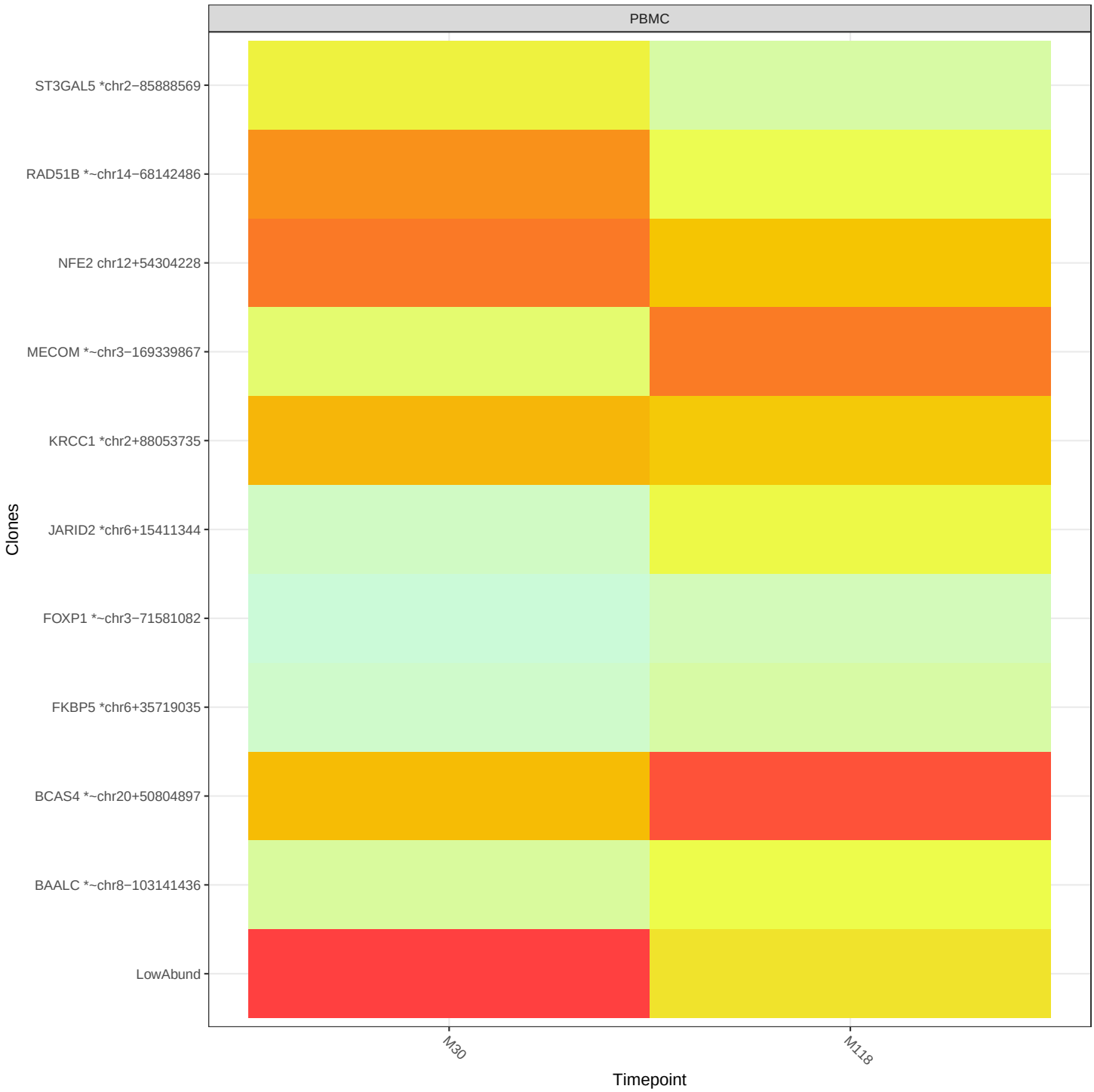
Integration sites near particular genes of interest

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.

PBMC
M30 1:60

PBMC
M118 1:754

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

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)

Analysis of integration site distributions and relative clonal abundance for subject p404

November 02, 2020

Contents

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**

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 (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	Rich
M24	1,725	Yes
M54	353	No

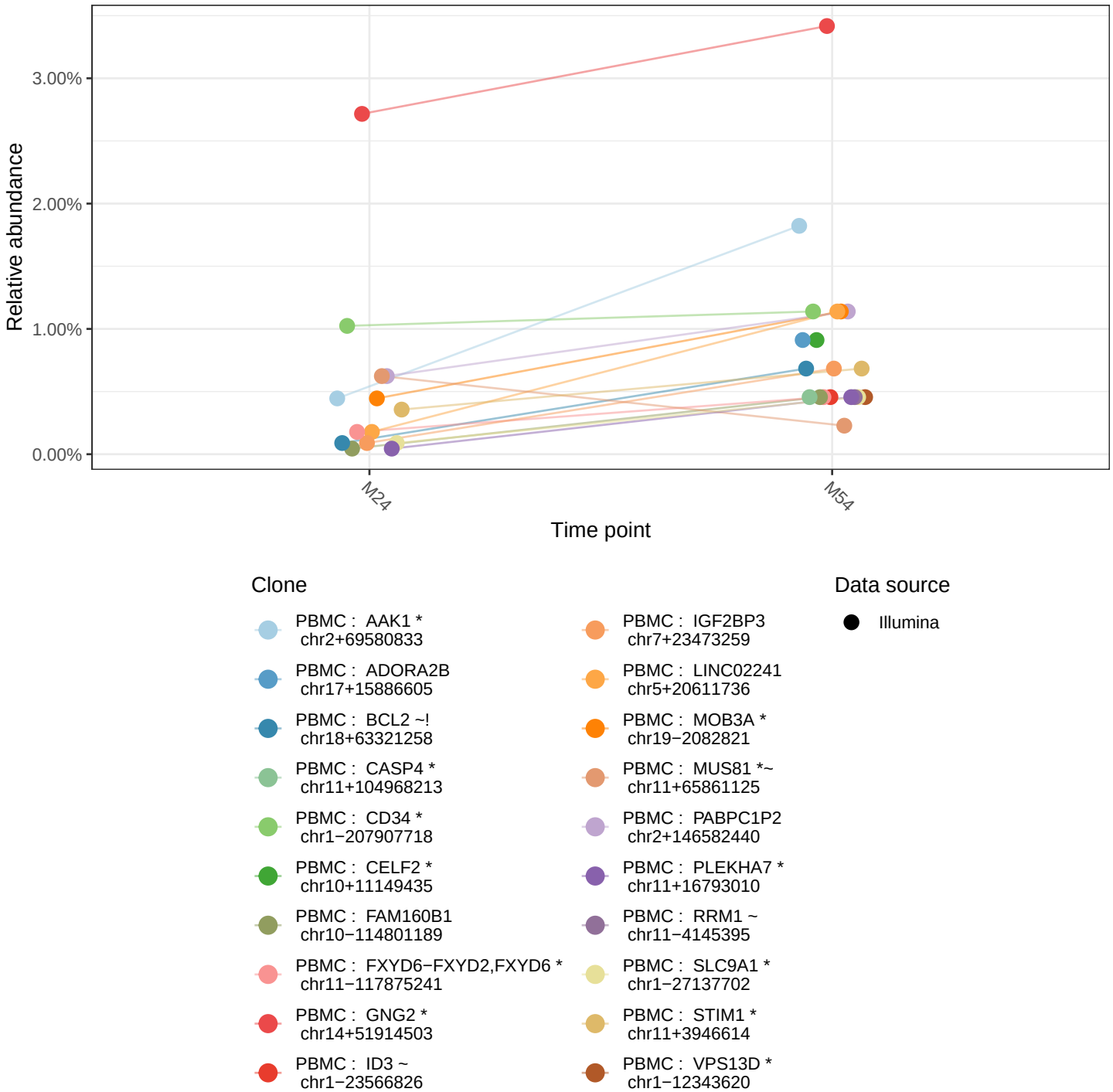
Do any cell clones account for more than 20% of all clones?

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 p404 over time points M24, M54 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

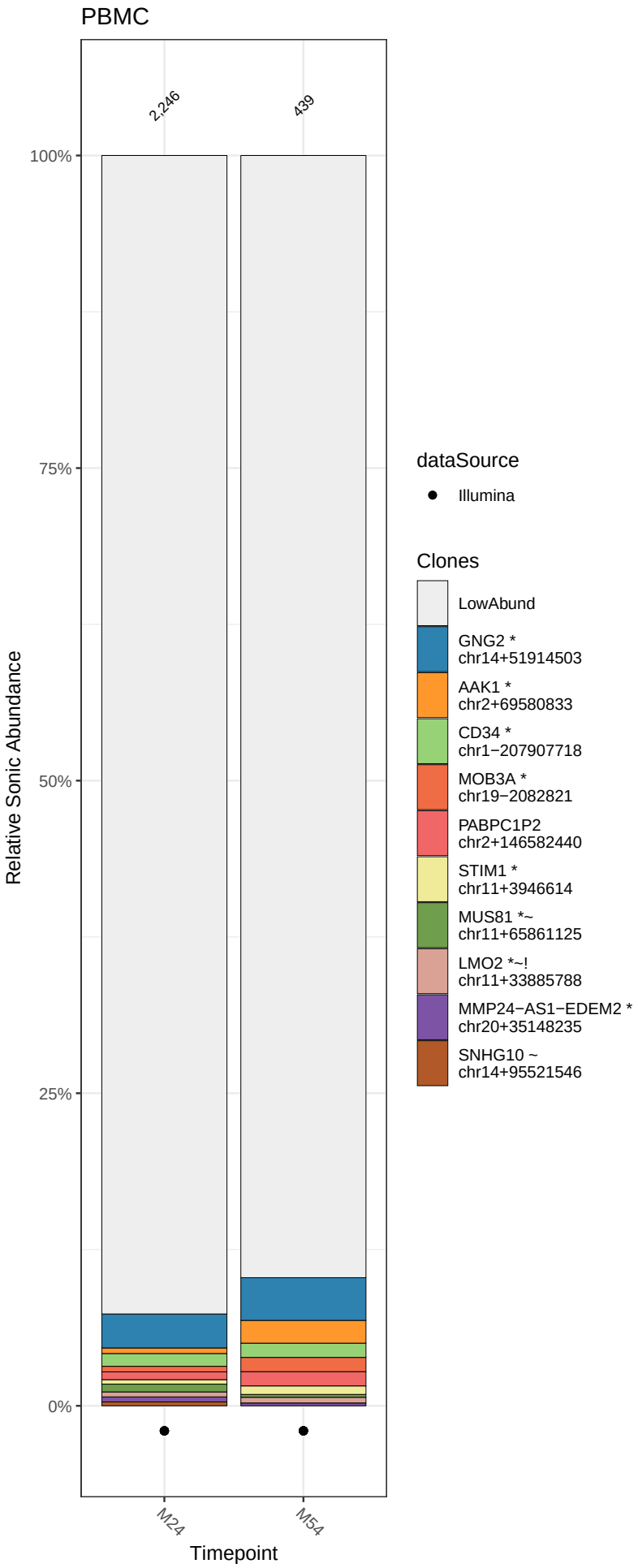
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3603	Illumina	M24	PBMC	317,269	2,246	1,725	0.214	8,750	7.22	0.969	603	yes	2020-10-28	0.686
GTSP3604	Illumina	M54	PBMC	263,974	439	353	0.179	1,573	5.72	0.974	134	yes	2020-10-28	0.540

Tracking of clonal abundances

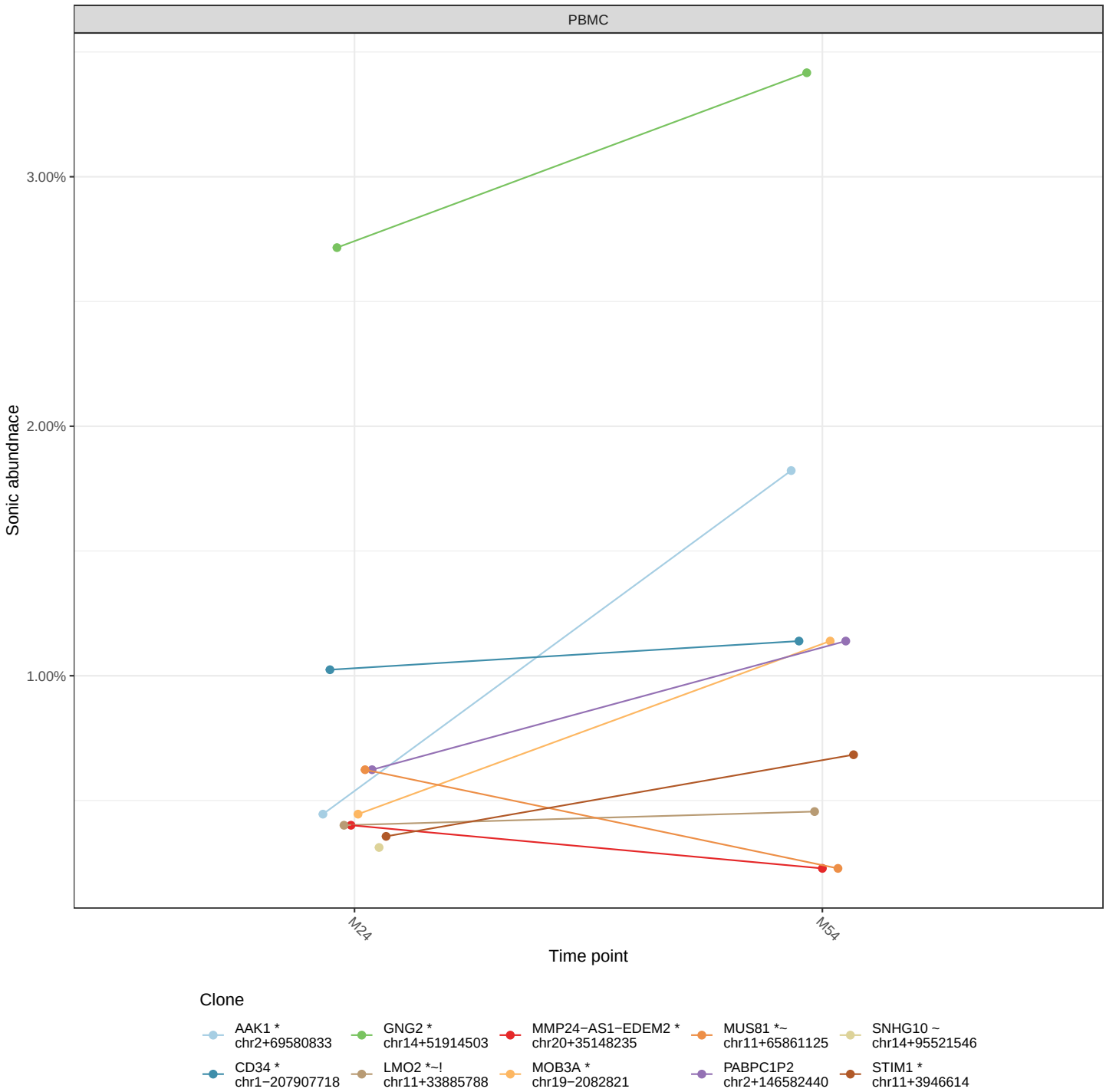
Relative abundance of cell clones

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
2,246
439
100%
75%
50%
25%
0%
Relative Sonic Abundance
M24
M54
Timepoint
dataSource
Illumina
Clones
LowAbund
GNG2 *
chr14+51914503
AAK1 *
chr2+69580833
CD34 *
chr1−207907718
MOB3A *
chr19−2082821
PABPC1P2
chr2+146582440
STIM1 *
chr11+3946614
MUS81 *~
chr11+65861125
LMO2 *~!
chr11+33885788
MMP24−AS1−EDEM2 *
chr20+35148235
SNHG10 ~
chr14+95521546

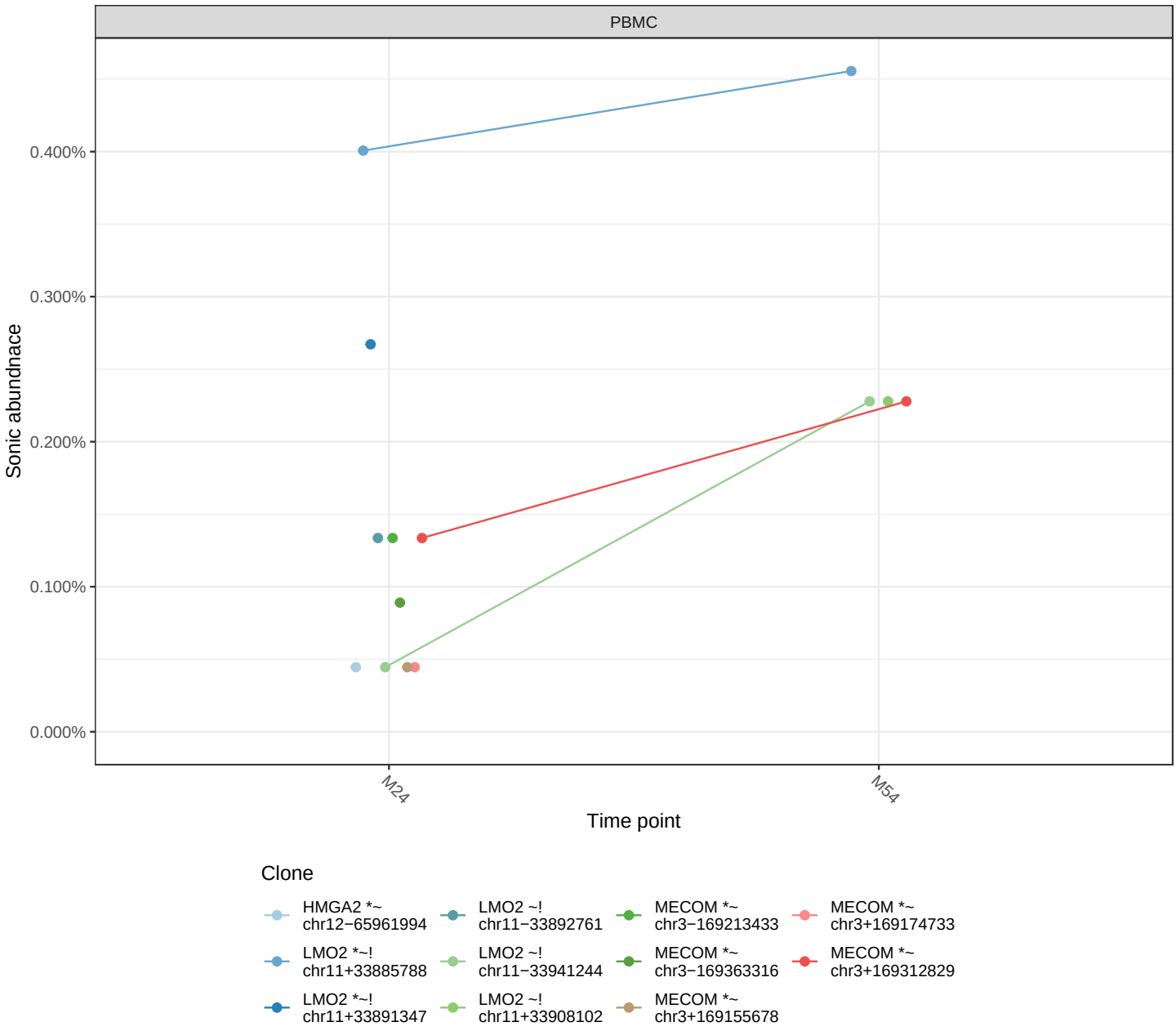
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



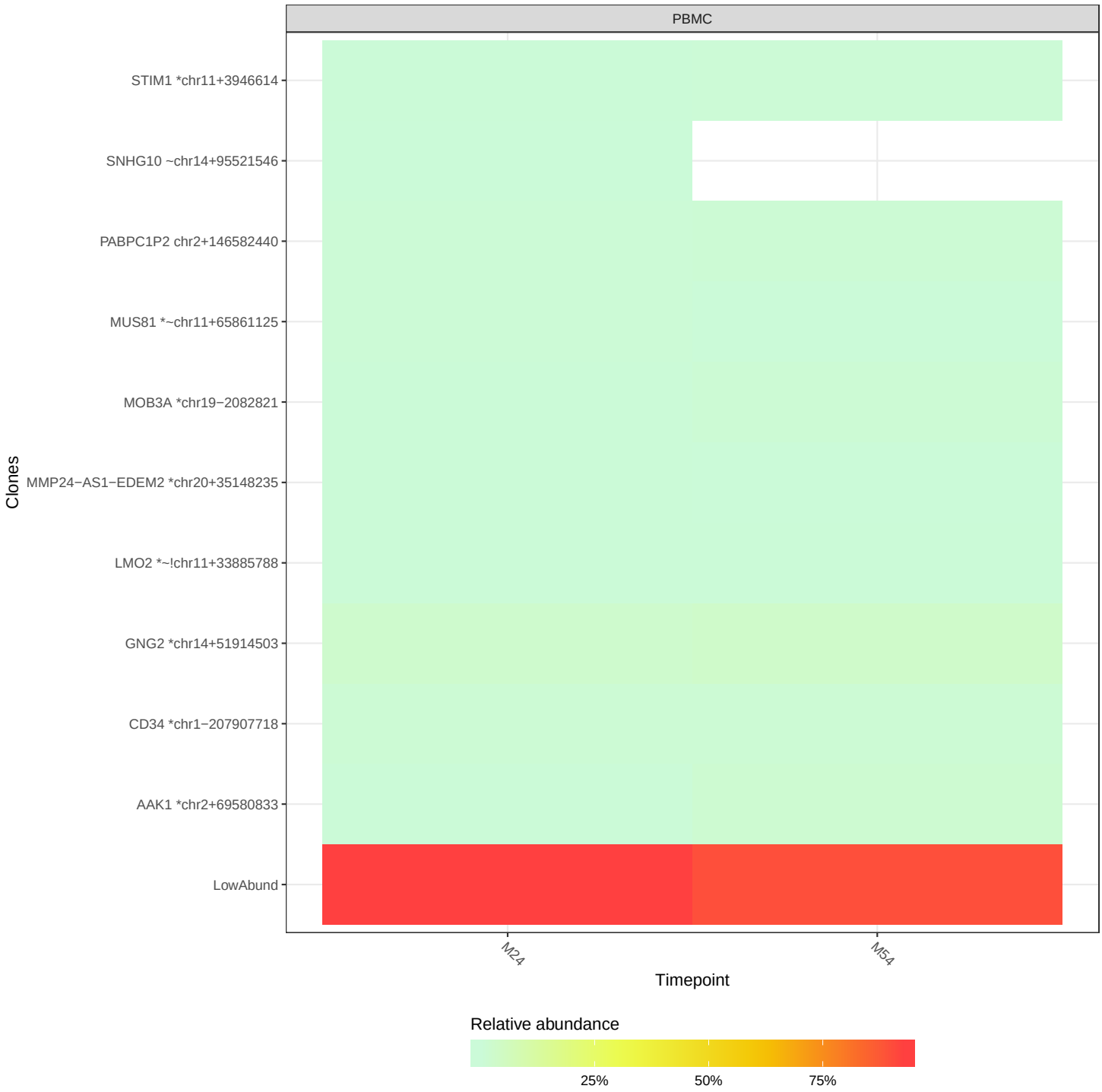
Integration sites near particular genes of interest

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.

PBMC
M24 2:61

PBMC
M54 1:15

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

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)

Analysis of integration site distributions and relative clonal abundance for subject p405

November 02, 2020

Contents

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**

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 (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	Rich
M42	67	No
M60	225	No

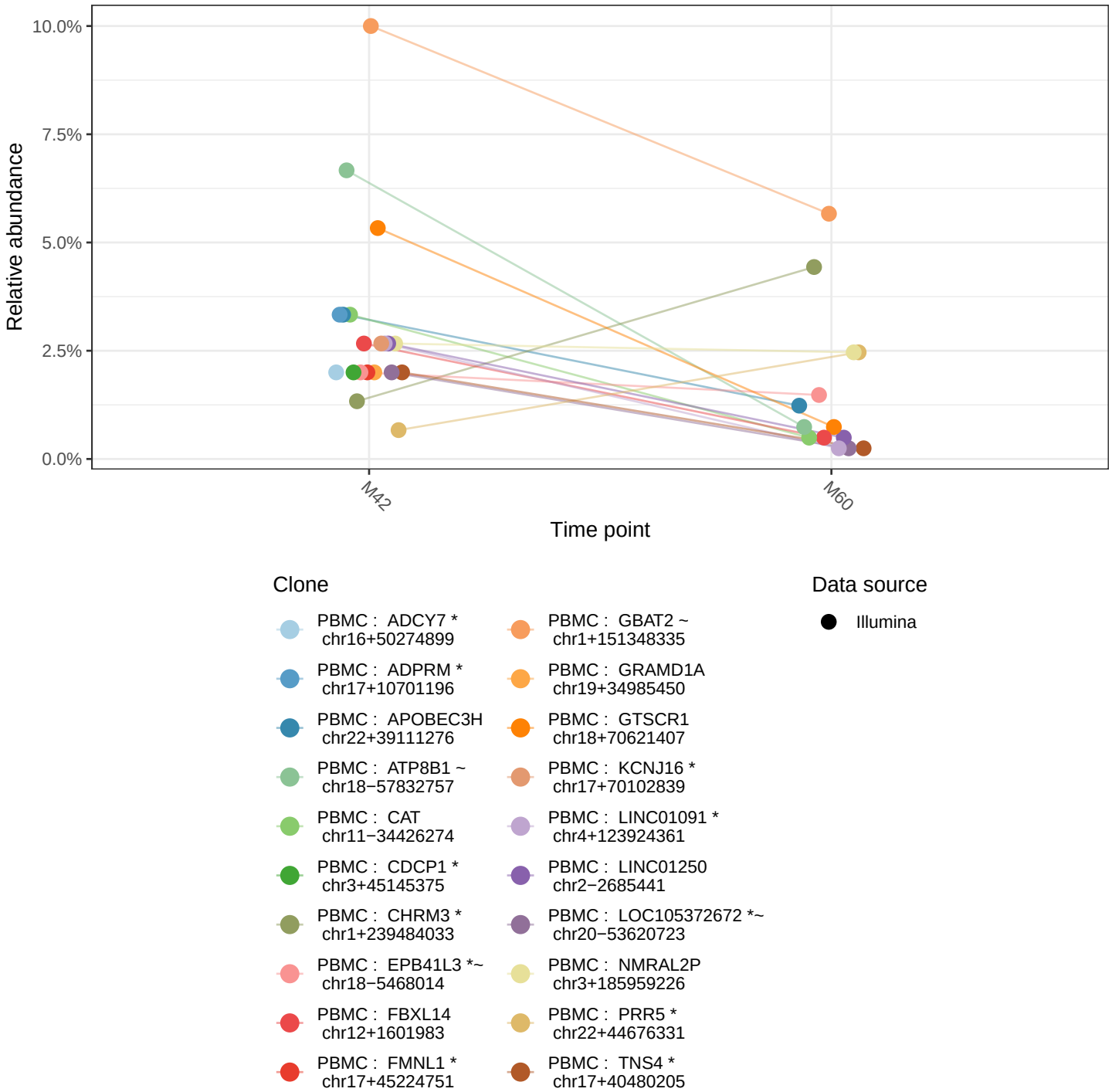
Do any cell clones account for more than 20% of all clones?

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 p405 over time points M42, M60 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

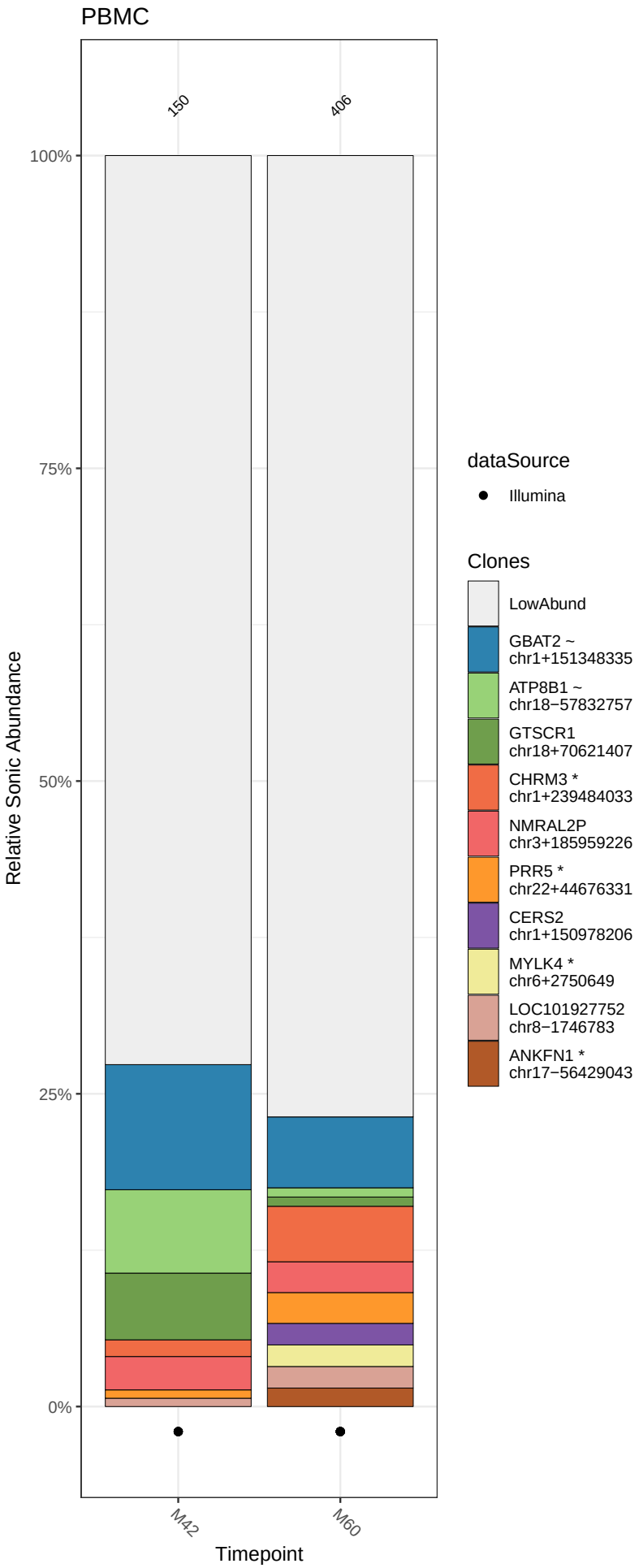
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3605	Illumina	M42	PBMC	1,190,436	150	67	0.412	118	3.86	0.919	14	yes	2020-10-12	0.490
GTSP3606	Illumina	M60	PBMC	665,842	406	225	0.380	675	5.03	0.930	43	yes	2020-10-12	0.496

Tracking of clonal abundances

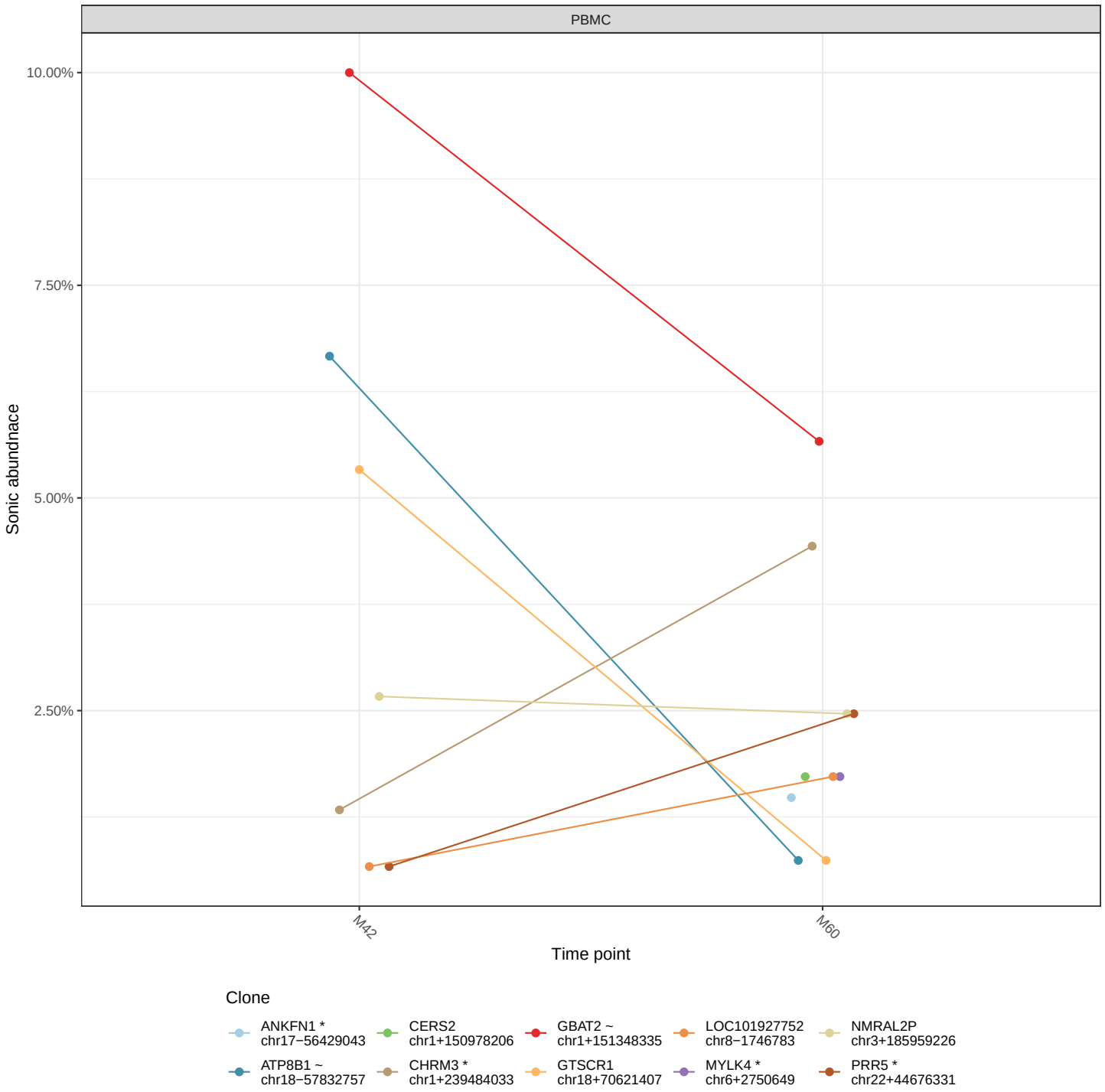
Relative abundance of cell clones

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
150
406
100%
75%
50%
25%
0%
Relative Sonic Abundance
M42
M60
Timepoint
dataSource
Illumina
Clones
LowAbund
GBAT2 ~ chr1+151348335
ATP8B1 ~ chr18−57832757
GTSCR1 chr18+70621407
CHRM3 * chr1+239484033
NMRAL2P chr3+185959226
PRR5 * chr22+44676331
CERS2 chr1+150978206
MYLK4 * chr6+2750649
LOC101927752 chr8−1746783
ANKFN1 * chr17−56429043

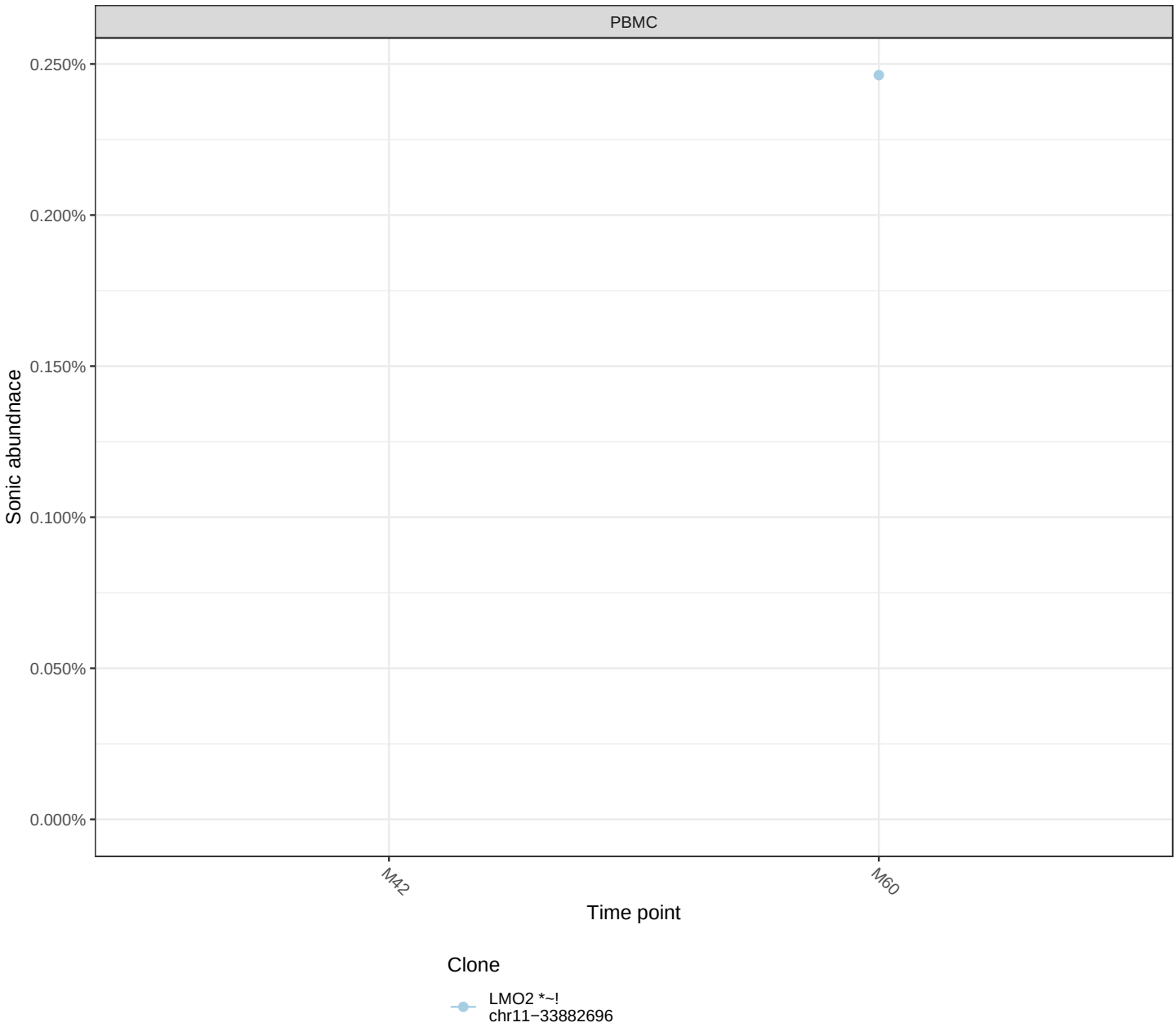
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



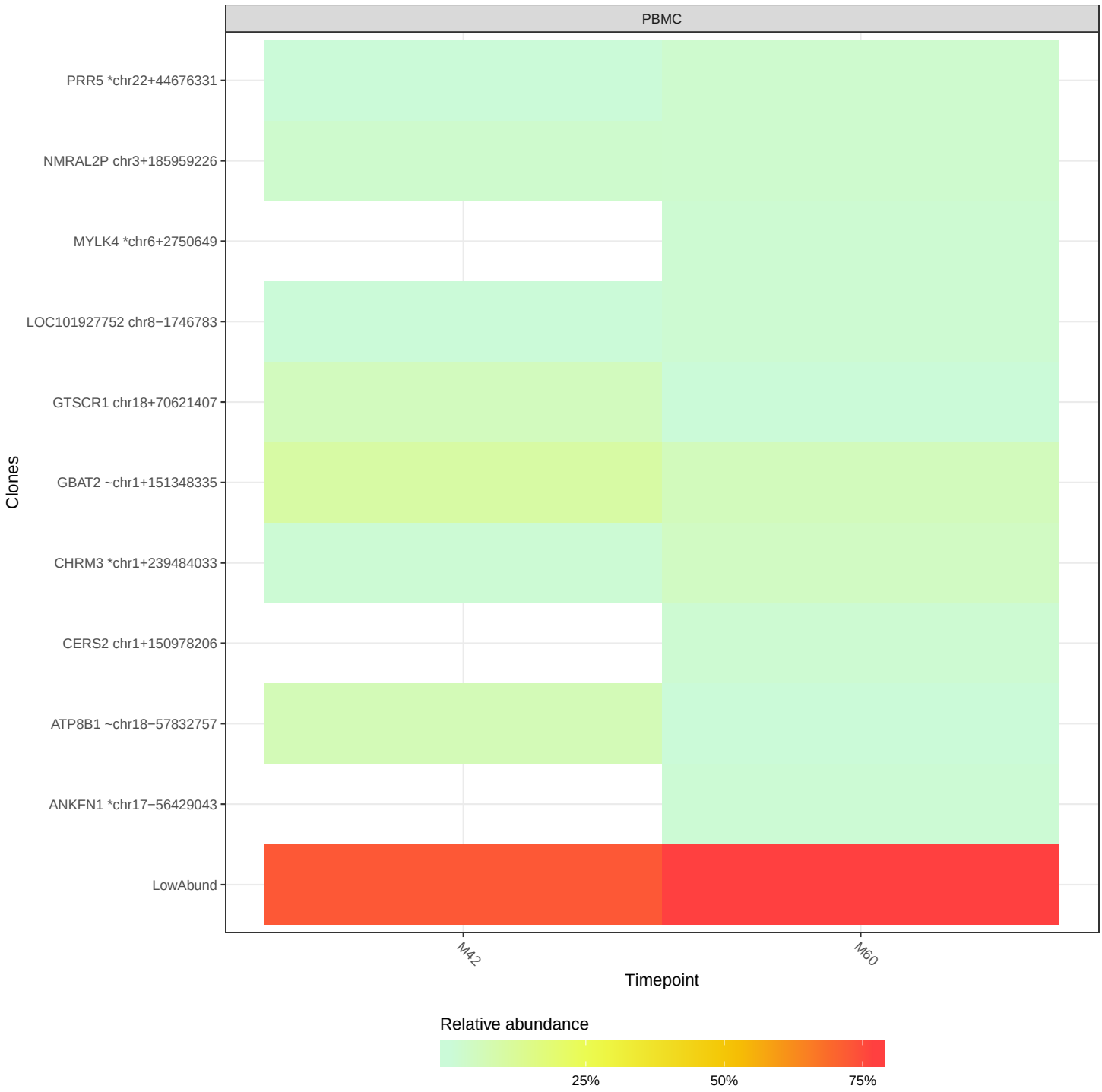
Integration sites near particular genes of interest

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.

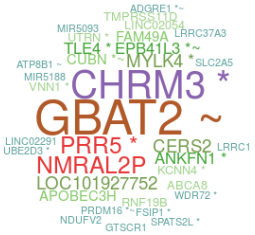


What are the most frequently occuring gene types in the subject?

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.

PBMC
M42 1:15

PBMC
M60 1:23



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

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)

Analysis of integration site distributions and relative clonal abundance for subject p406

November 02, 2020

Contents

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**

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 (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	Rich
M24	74	No
M84	196	No

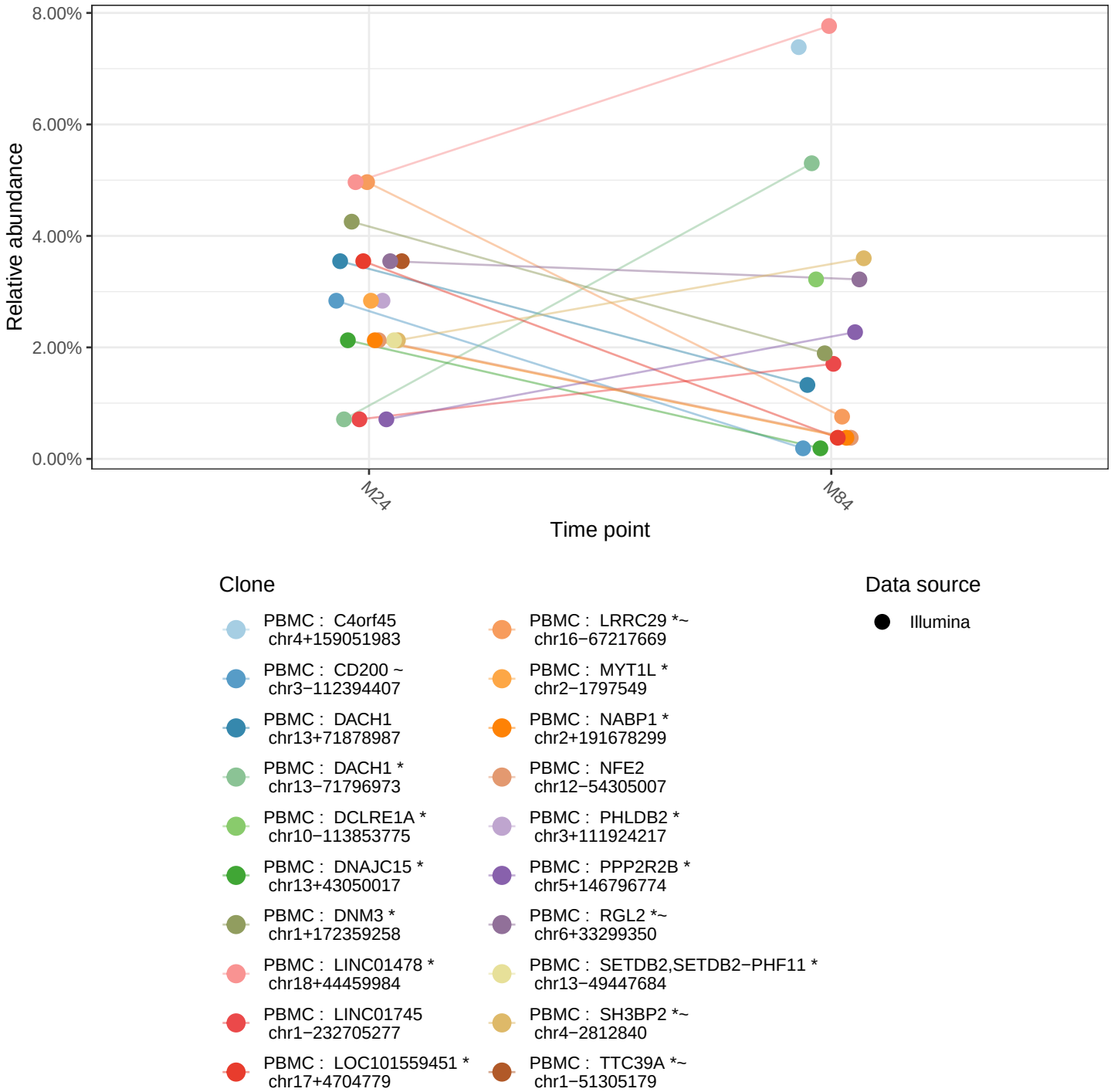
Do any cell clones account for more than 20% of all clones?

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 p406 over time points M24, M84 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

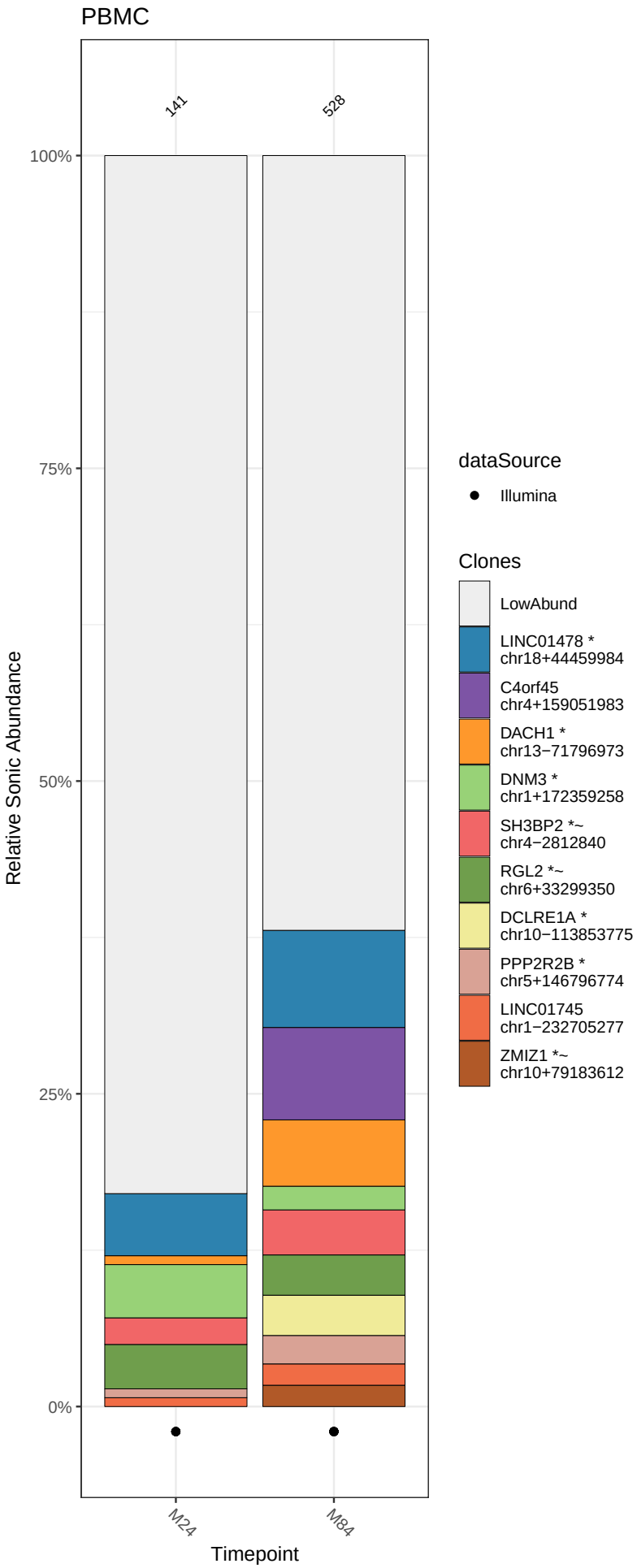
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3607	Illumina	M24	PBMC	202,923	141	74	0.352	133	4.07	0.945	17	yes	2020-10-14	0.172
GTSP3608	Illumina	M84	PBMC	289,008	528	196	0.532	384	4.58	0.867	21	yes	2020-10-14	0.483

Tracking of clonal abundances

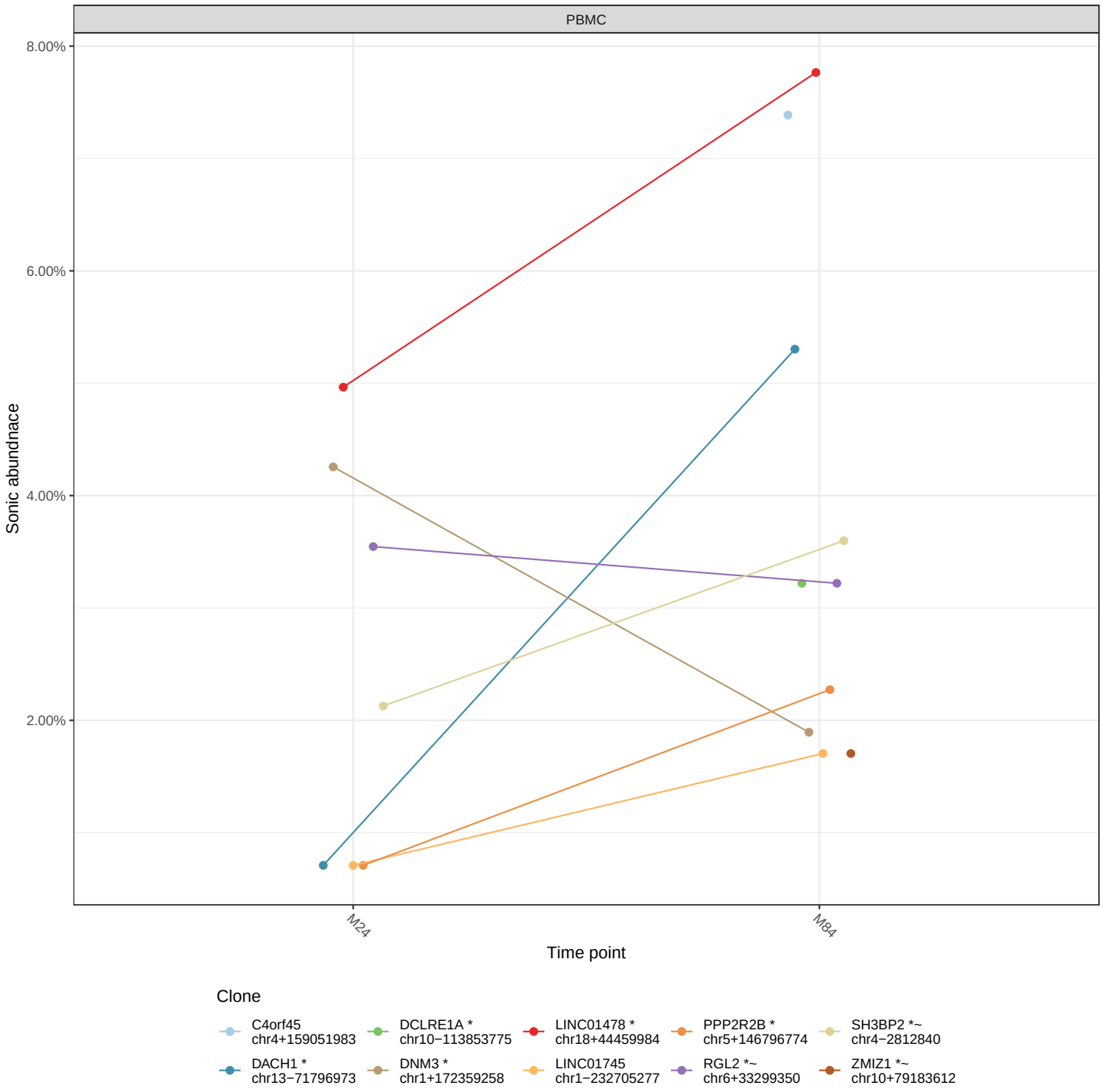
Relative abundance of cell clones

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
141
528
100%
75%
50%
25%
0%
Relative Sonic Abundance
M24
M84
Timepoint
dataSource
Illumina
Clones
LowAbund
LINC01478 *
chr18+44459984
C4orf45
chr4+159051983
DACH1 *
chr13−71796973
DNM3 *
chr1+172359258
SH3BP2 *~
chr4−2812840
RGL2 *~
chr6+33299350
DCLRE1A *
chr10−113853775
PPP2R2B *
chr5+146796774
LINC01745
chr1−232705277
ZMIZ1 *~
chr10+79183612

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



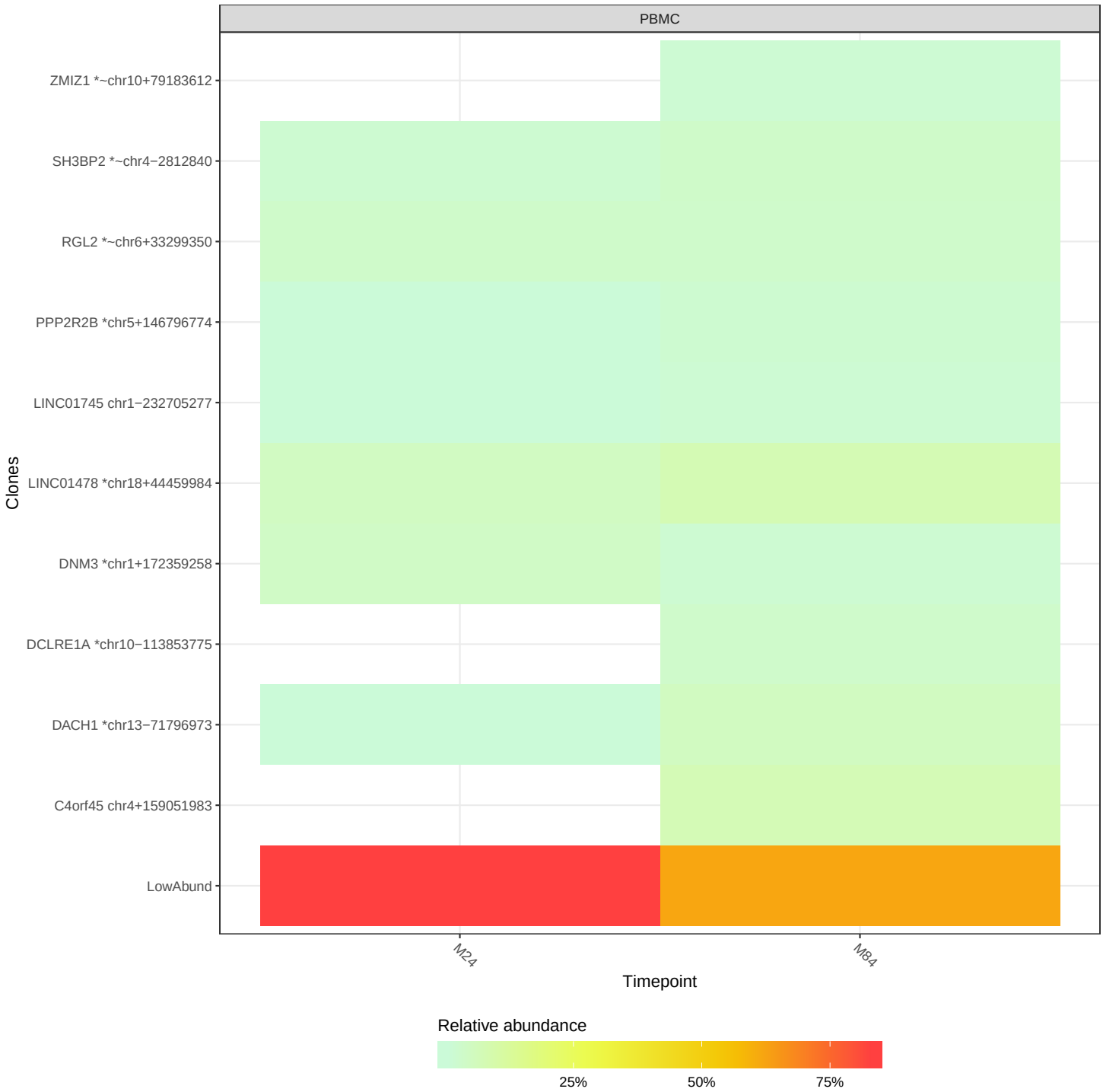
Integration sites near particular genes of interest

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
M24 1:7

PBMC
M84 1:41

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

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)

Analysis of integration site distributions and relative clonal abundance for subject p407

November 02, 2020

Contents

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**

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	Rich
M24	25	No
M96	57	No

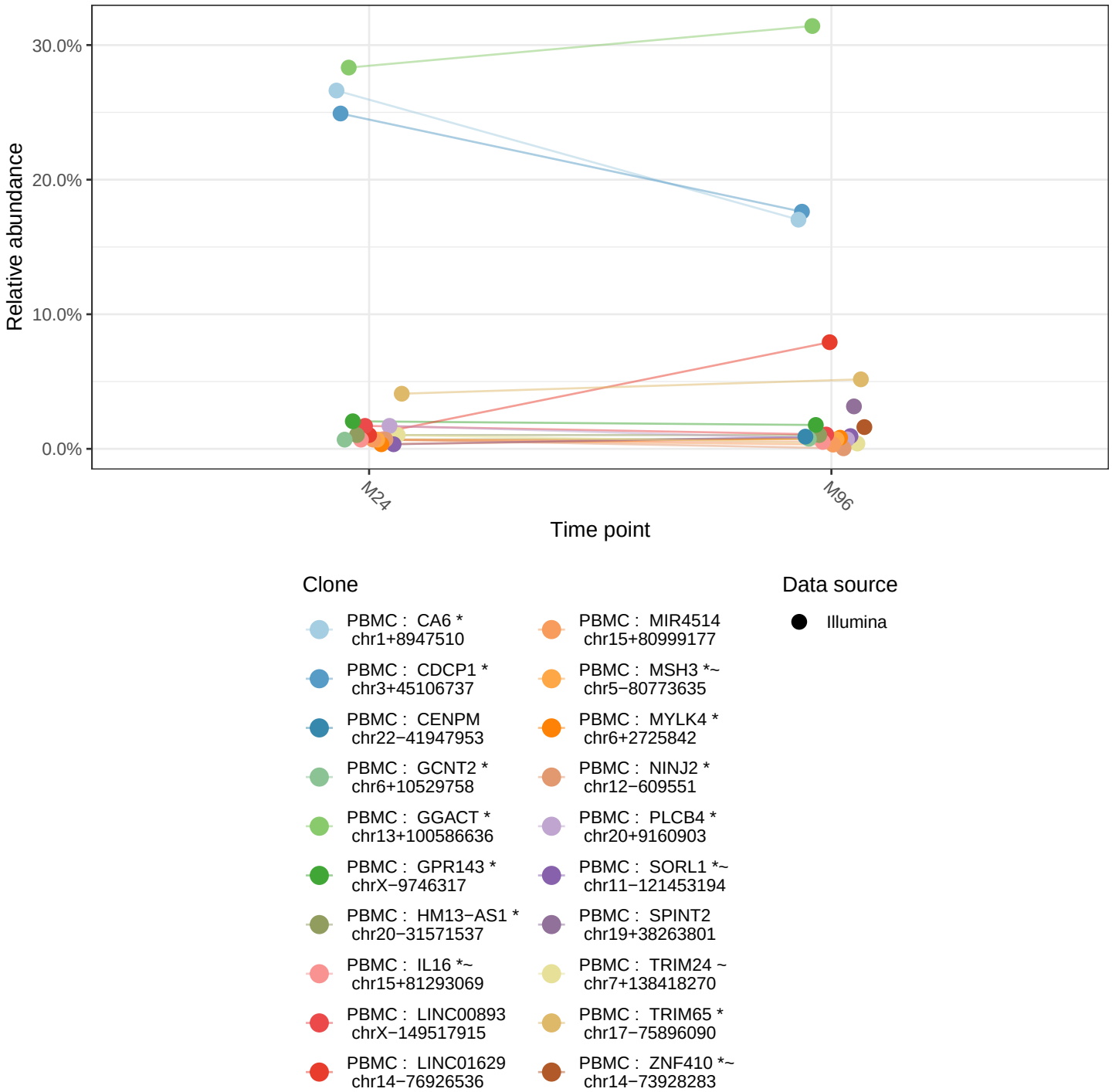
Do any cell clones account for more than 20% of all clones?

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)
chr1+8947510	78	26.6%	M24	PBMC	CA6	0.00	ENO1	-68.80
chr13+100586636	83	28.3%	M24	PBMC	GGACT	0.00	FGF14	1134.20
chr3+45106737	73	24.9%	M24	PBMC	CDCP1	0.00	ZDHHC3	-130.60
chr13+100586636	797	31.4%	M96	PBMC	GGACT	0.00	FGF14	1134.20

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 p407 over time points M24, M96 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

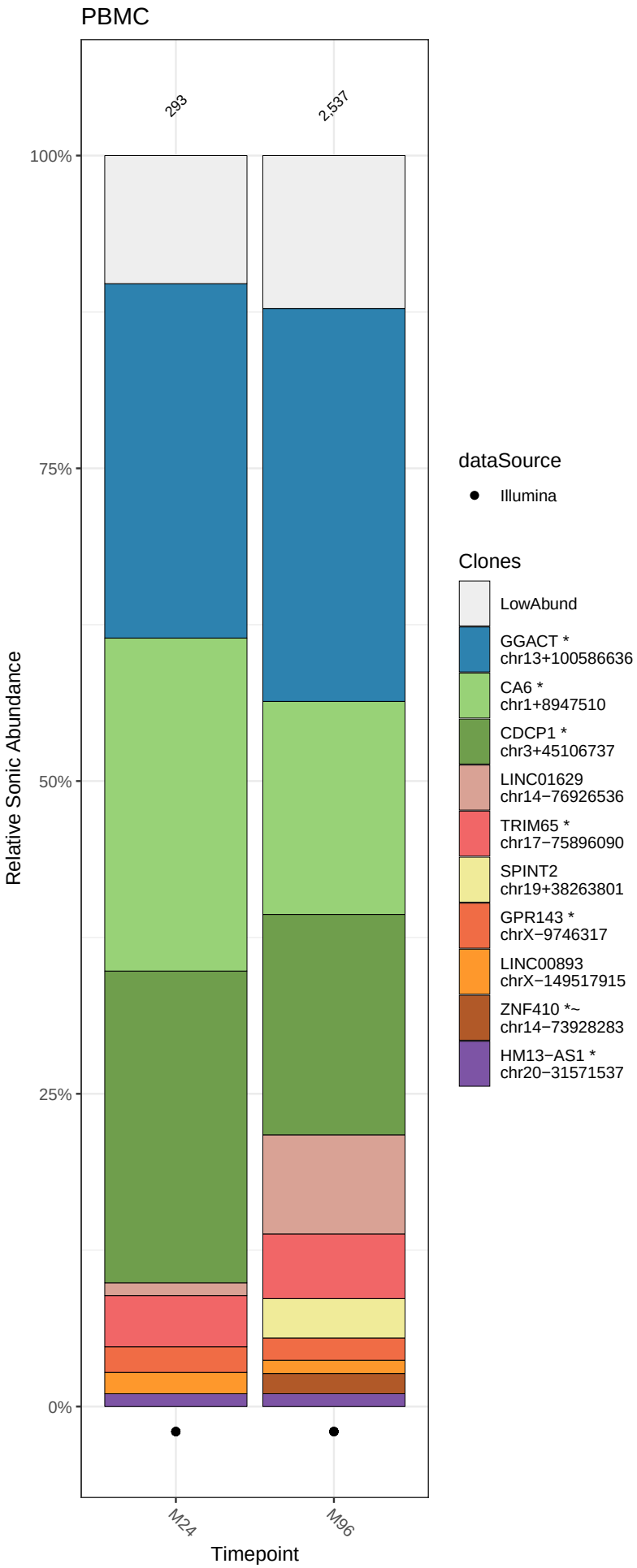
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3609	Illumina	M24	PBMC	233,479	293	25	0.753	28	1.94	0.603	2	yes	2020-10-14	0.401
GTSP3610	Illumina	M96	PBMC	482,939	2,537	57	0.844	92	2.33	0.576	3	yes	2020-10-28	0.519

Tracking of clonal abundances

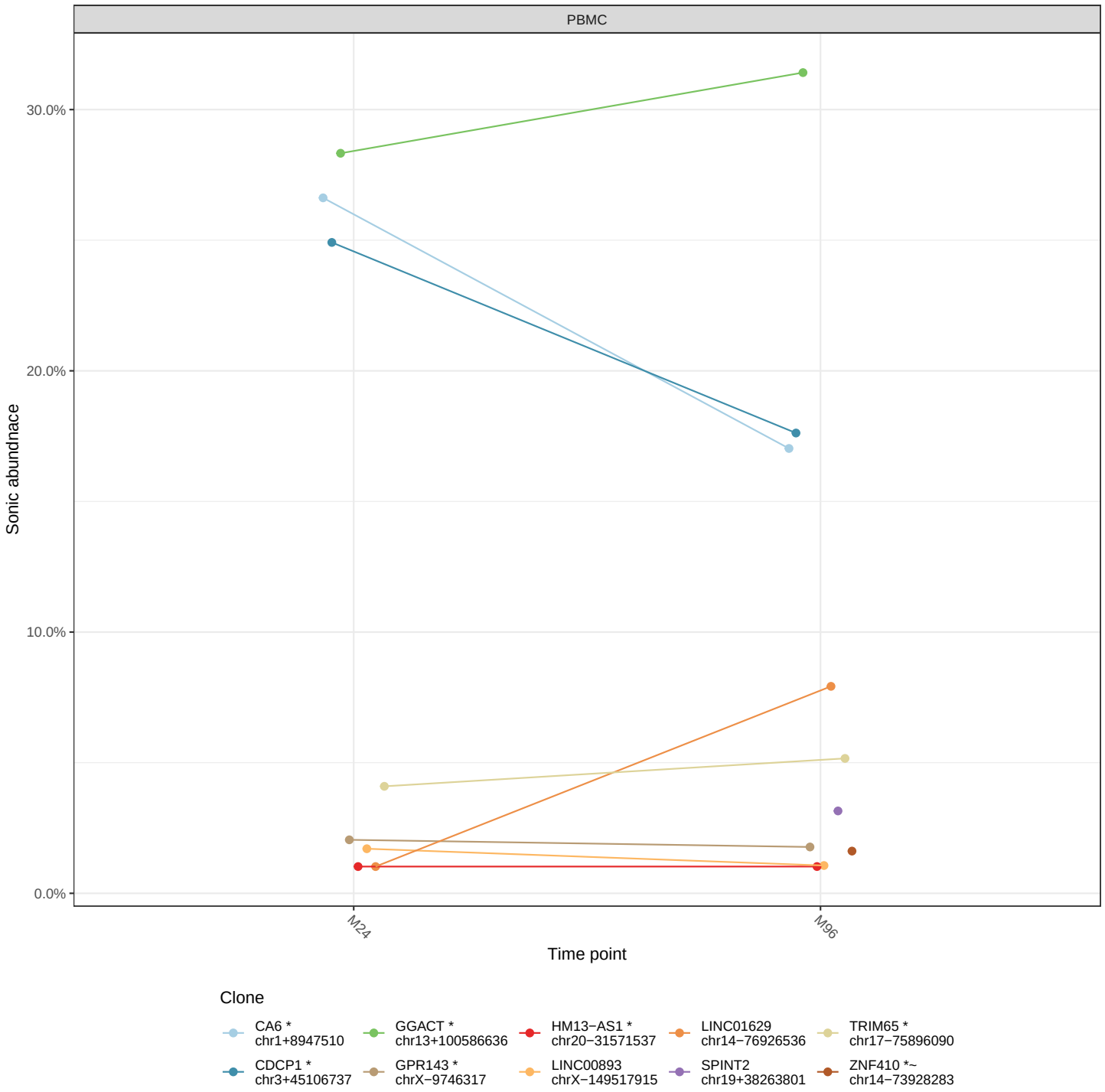
Relative abundance of cell clones

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
293
2,537
100%
75%
50%
25%
0%
Relative Sonic Abundance
M24
M96
Timepoint
dataSource
Illumina
Clones
LowAbund
GGACT *
chr13+100586636
CA6 *
chr1+8947510
CDCP1 *
chr3+45106737
LINC01629
chr14−76926536
TRIM65 *
chr17−75896090
SPINT2
chr19+38263801
GPR143 *
chrX−9746317
LINC00893
chrX−149517915
ZNF410 *~
chr14−73928283
HM13−AS1 *
chr20−31571537

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



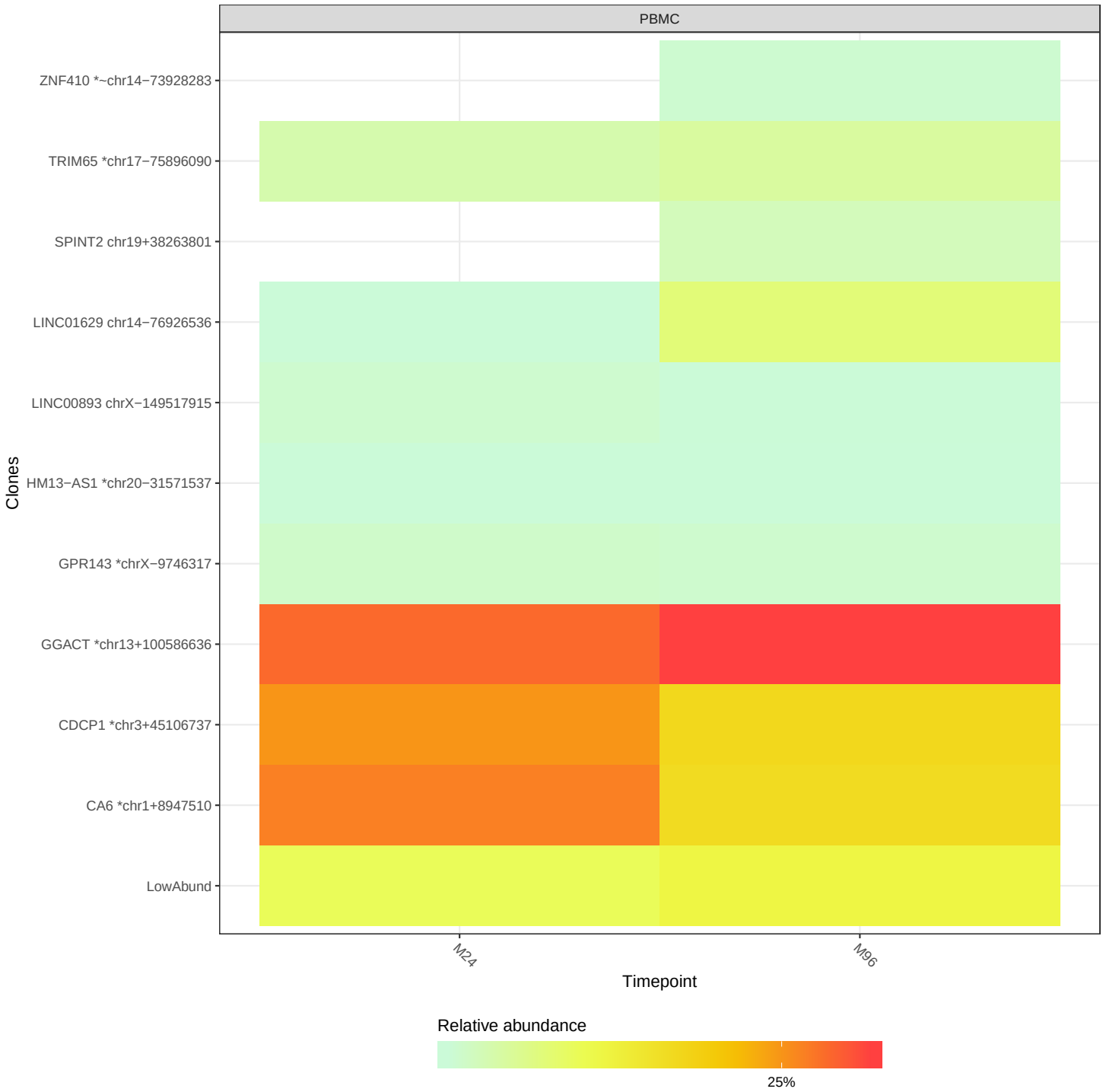
Integration sites near particular genes of interest

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
M24 1:83

PBMC
M96 1:797

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

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)

Analysis of integration site distributions and relative clonal abundance for subject p408

November 02, 2020

Contents

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**

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 (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	Rich
M42	114	No
M84	373	No

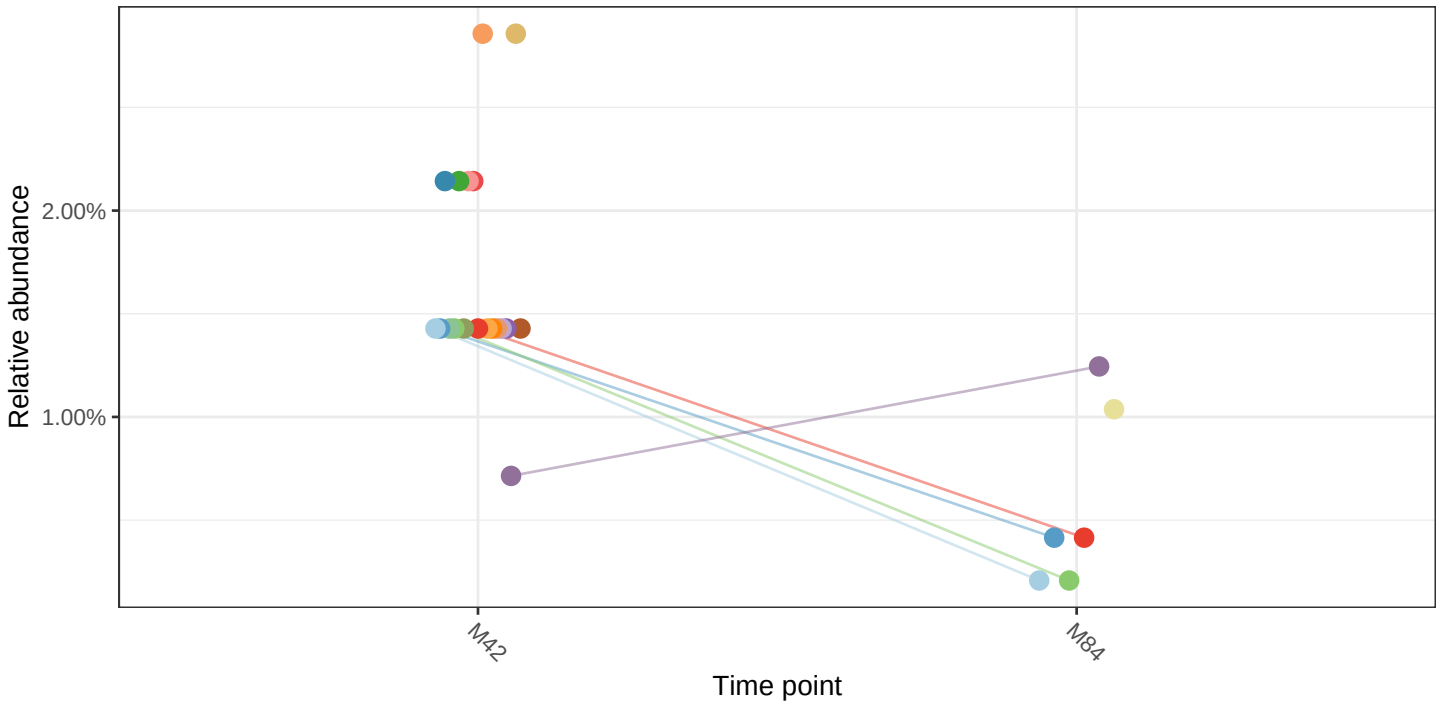
Do any cell clones account for more than 20% of all clones?

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.



Clone

- PBMC : ABHD2 *~ chr15+89090726
- PBMC : ABHD2 *~ chr15+89133212
- PBMC : CYP1B1−AS1 chr2−38107173
- PBMC : EOMES chr3−27666257
- PBMC : EPS8L2 * chr11+726580
- PBMC : FER1L6,FER1L6−AS1 * chr8−124007974
- PBMC : HSH2D *~ chr19+16148391
- PBMC : IL1R2 chr2−102032722
- PBMC : LINC02068 * chr3+172577454
- PBMC : LIX1L−AS1 * chr1−145928118
- PBMC : LRRC8C * chr1+89719152
- PBMC : MECOM *~ chr3−169351631
- PBMC : MIR1−1HG chr20+62604504
- PBMC : MIR924HG * chr18+39365373
- PBMC : PRDM16 *~ chr1−3184351
- PBMC : PTK2B * chr8+27377960
- PBMC : SMIM20 * chr4−25914814
- PBMC : STON2 * chr14−81417017
- PBMC : TTC34 chr1−2887842
- PBMC : ZSWIM2 chr2+186883495

Data source

- Illumina

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 p408 over time points M42, M84 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

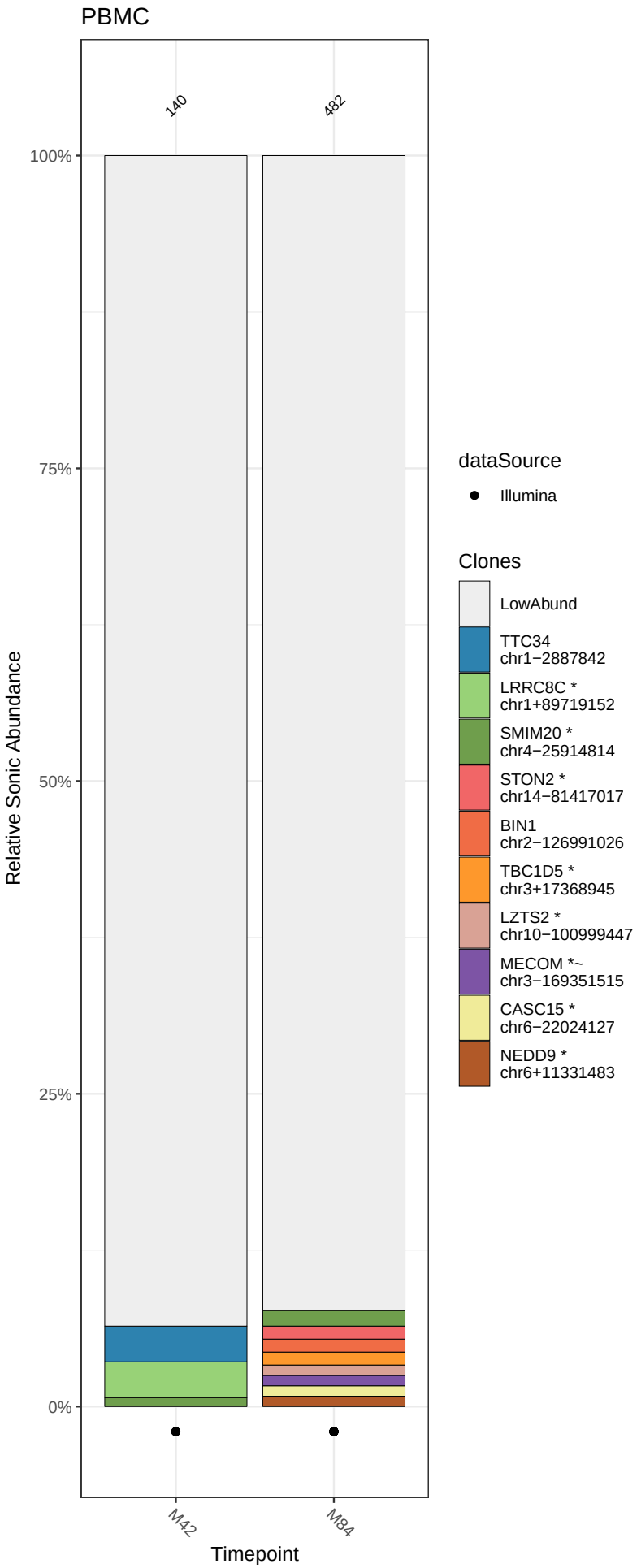
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3611	Illumina	M42	PBMC	268,031	140	114	0.163	465	4.65	0.982	45	yes	2020-10-14	1.24
GTSP3612	Illumina	M84	PBMC	254,608	482	373	0.191	1,113	5.82	0.982	133	yes	2020-10-14	1.28

Tracking of clonal abundances

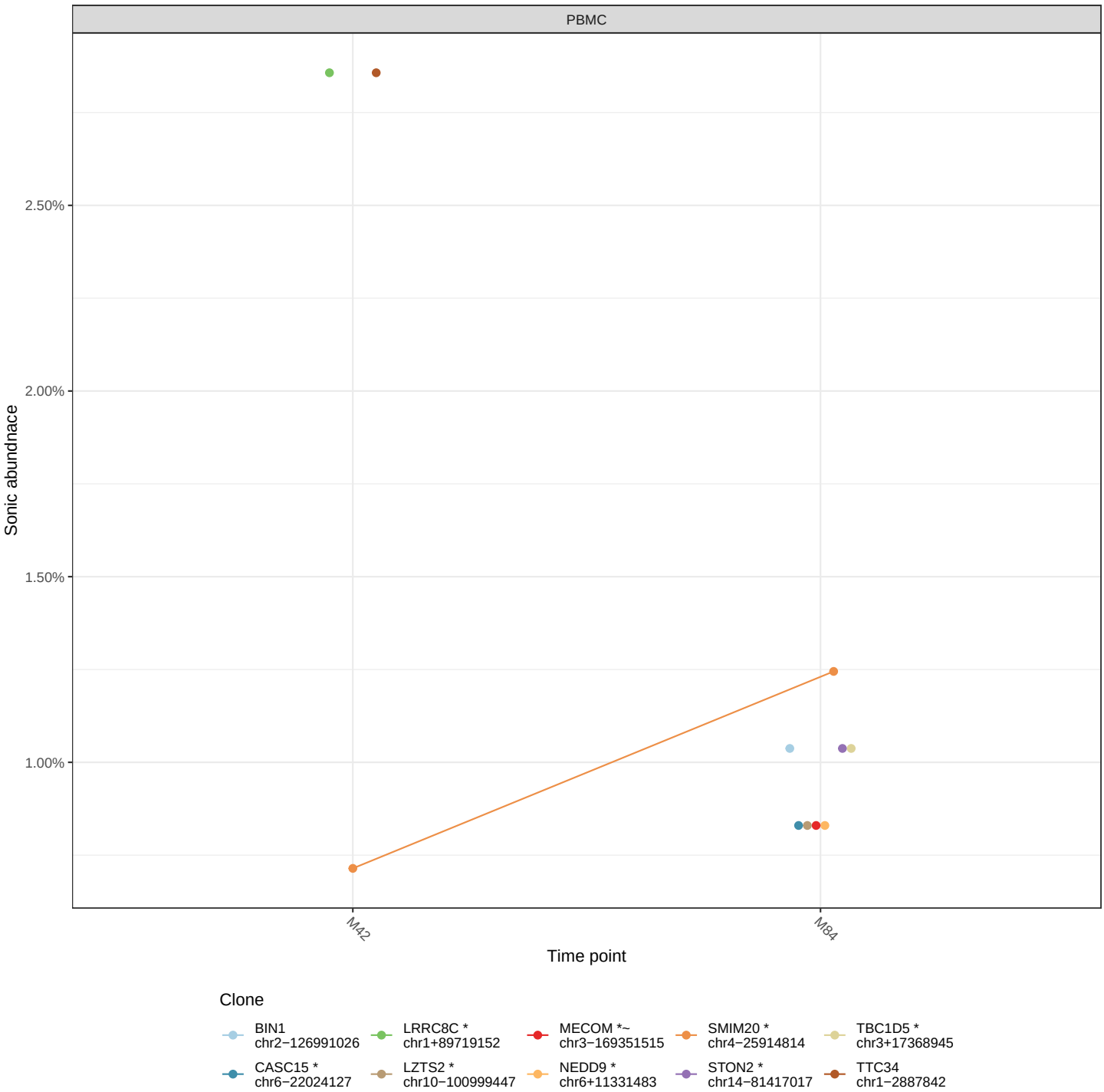
Relative abundance of cell clones

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
140
482
100%
75%
50%
25%
0%
Relative Sonic Abundance
M42
M84
Timepoint
dataSource
Illumina
Clones
LowAbund
TTC34
chr1−2887842
LRRC8C *
chr1+89719152
SMIM20 *
chr4−25914814
STON2 *
chr14−81417017
BIN1
chr2−126991026
TBC1D5 *
chr3+17368945
LZTS2 *
chr10−100999447
MECOM *~
chr3−169351515
CASC15 *
chr6−22024127
NEDD9 *
chr6+11331483

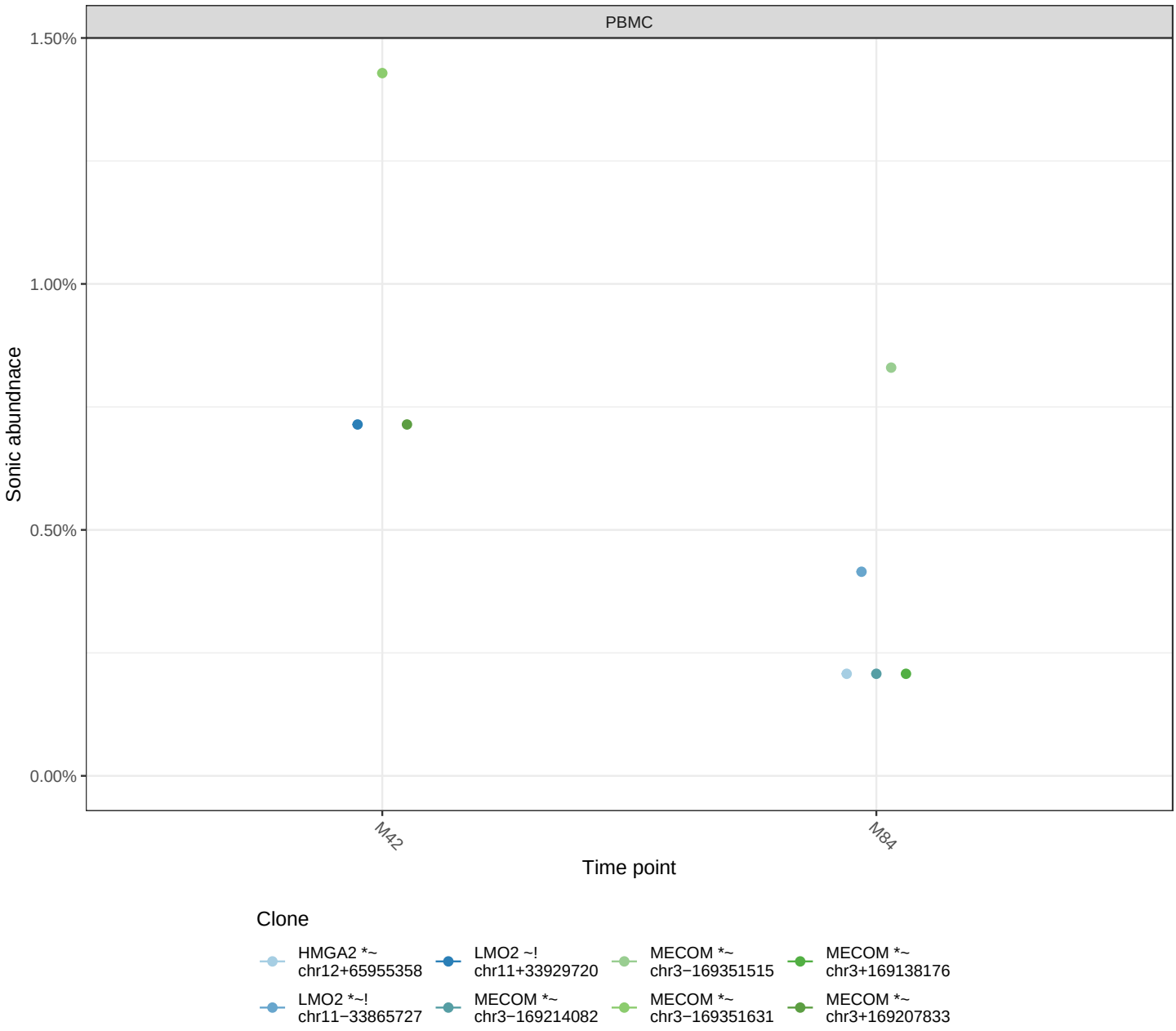
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



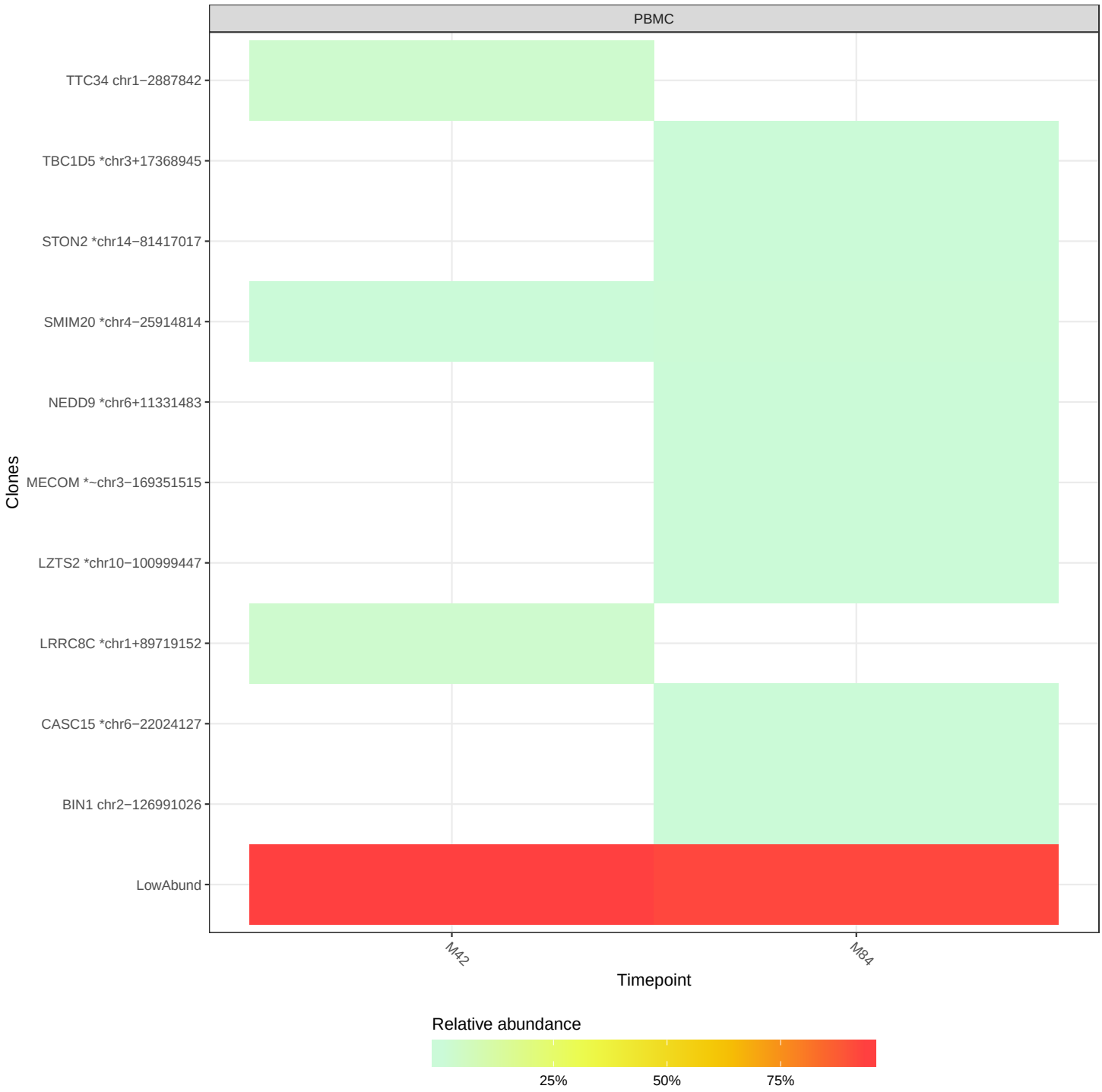
Integration sites near particular genes of interest

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.

PBMC
M42 1:4

PBMC
M84 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

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)

Analysis of integration site distributions and relative clonal abundance for subject p409

November 02, 2020

Contents

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**

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 (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	Rich
M36	73	No
M84	247	No

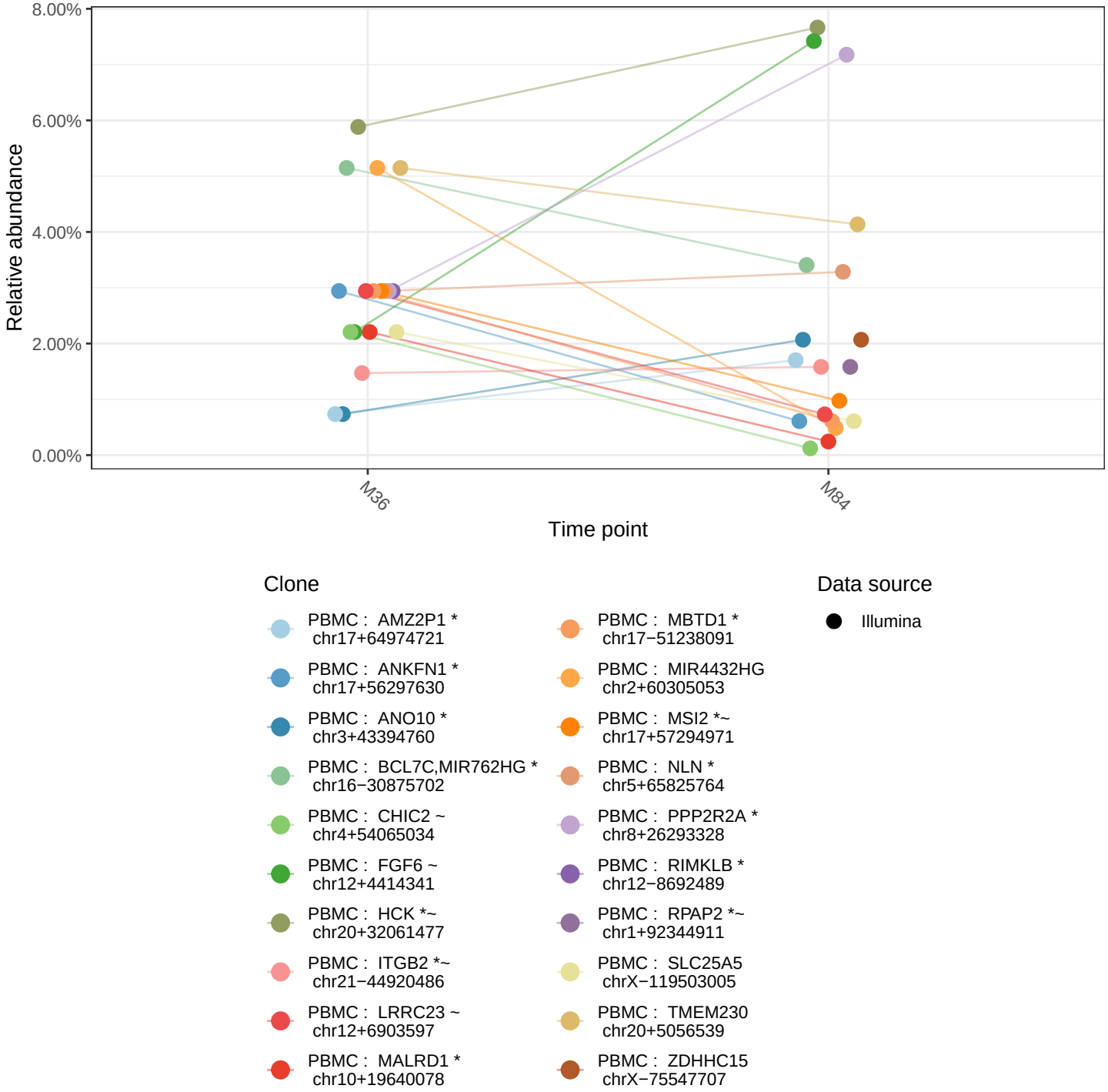
Do any cell clones account for more than 20% of all clones?

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 p409 over time points M36, M84 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

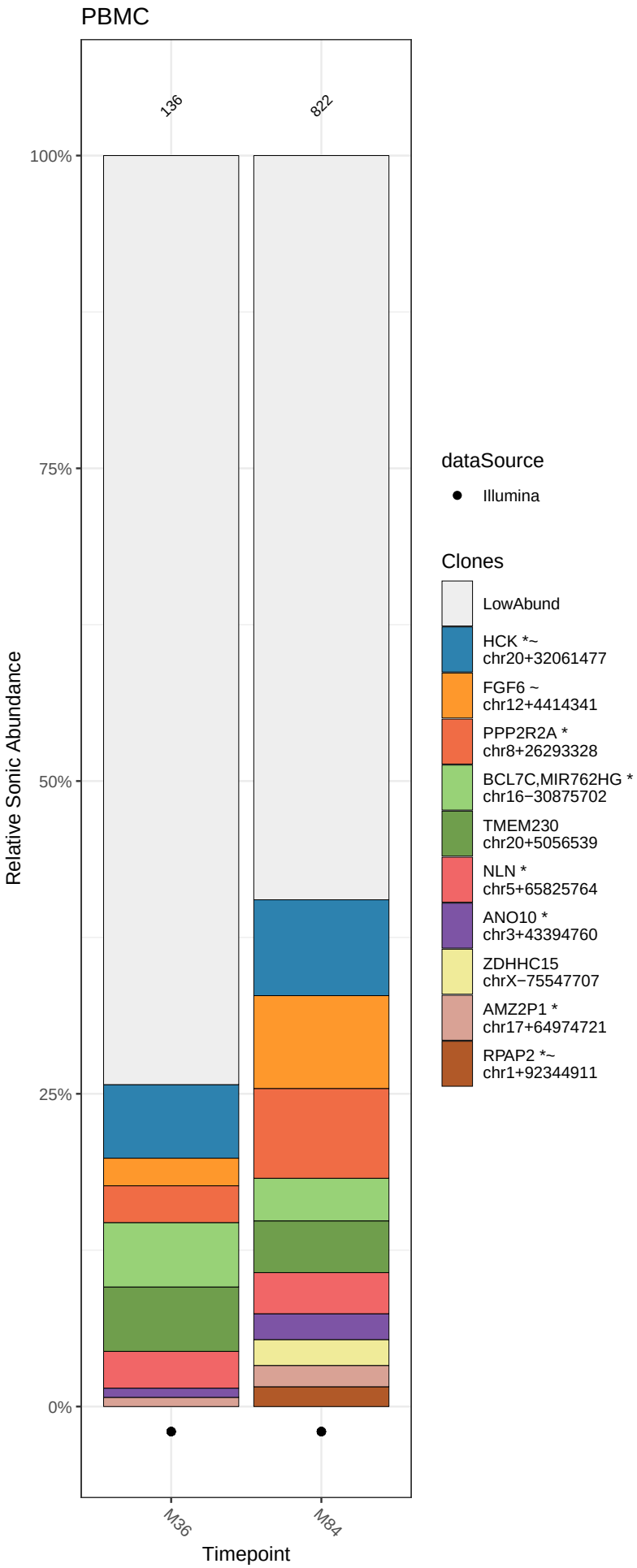
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3613	Illumina	M36	PBMC	207,626	136	73	0.365	191	4.02	0.936	15	yes	2020-10-14	0.287
GTSP3614	Illumina	M84	PBMC	241,890	822	247	0.605	701	4.59	0.832	18	yes	2020-10-14	0.627

Tracking of clonal abundances

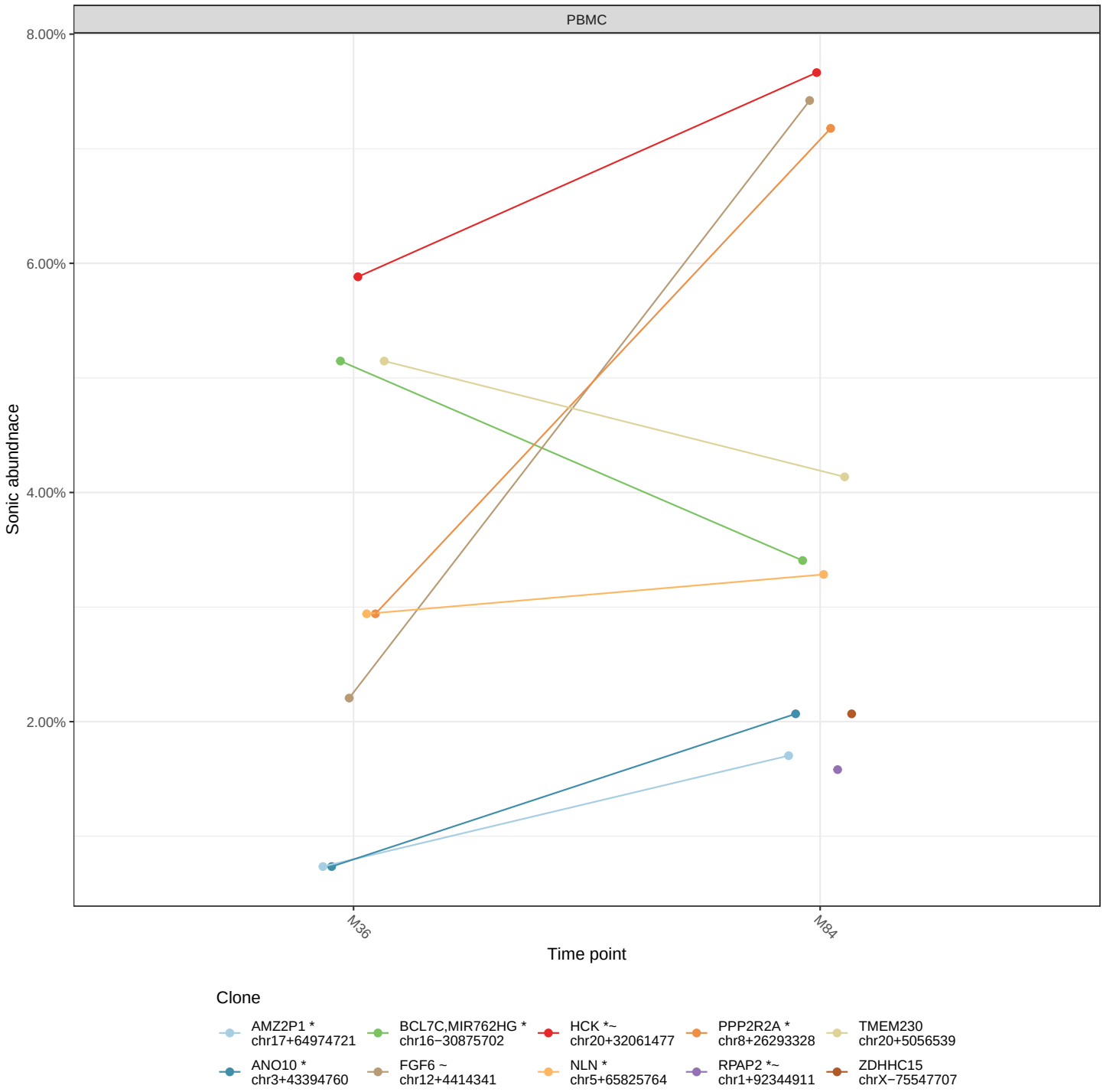
Relative abundance of cell clones

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
136
822
100%
75%
50%
25%
0%
Relative Sonic Abundance
M36
M84
Timepoint
dataSource
Illumina
Clones
LowAbund
HCK *~ chr20+32061477
FGF6 ~ chr12+4414341
PPP2R2A * chr8+26293328
BCL7C,MIR762HG * chr16−30875702
TMEM230 chr20+5056539
NLN * chr5+65825764
ANO10 * chr3+43394760
ZDHHC15 chrX−75547707
AMZ2P1 * chr17+64974721
RPAP2 *~ chr1+92344911

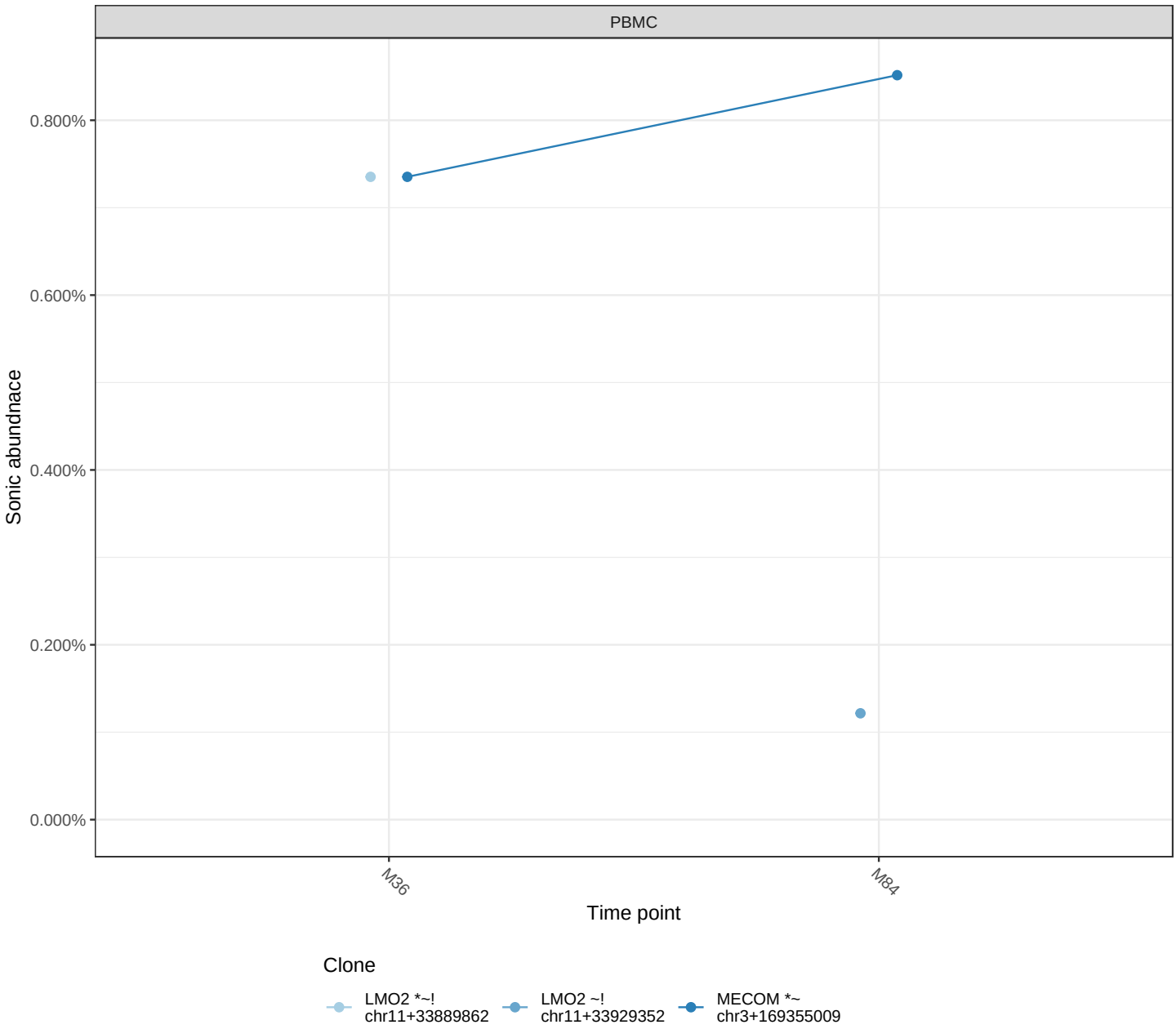
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



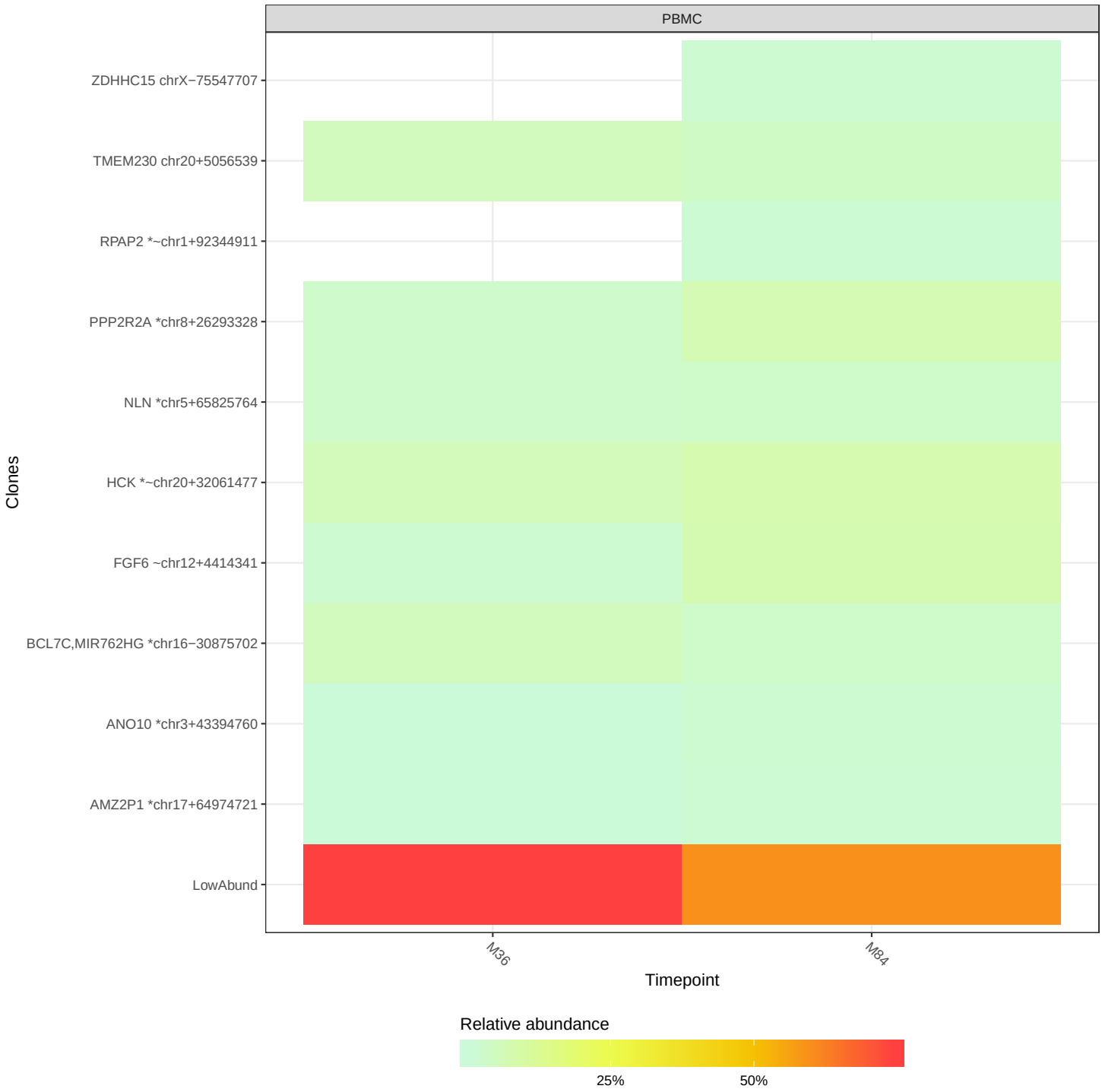
Integration sites near particular genes of interest

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.

PBMC
M36 1:8

PBMC
M84 1:63

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

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)

Analysis of integration site distributions and relative clonal abundance for subject p410

November 02, 2020

Contents

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**

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 (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	Rich
M24	342	No
M72	442	No

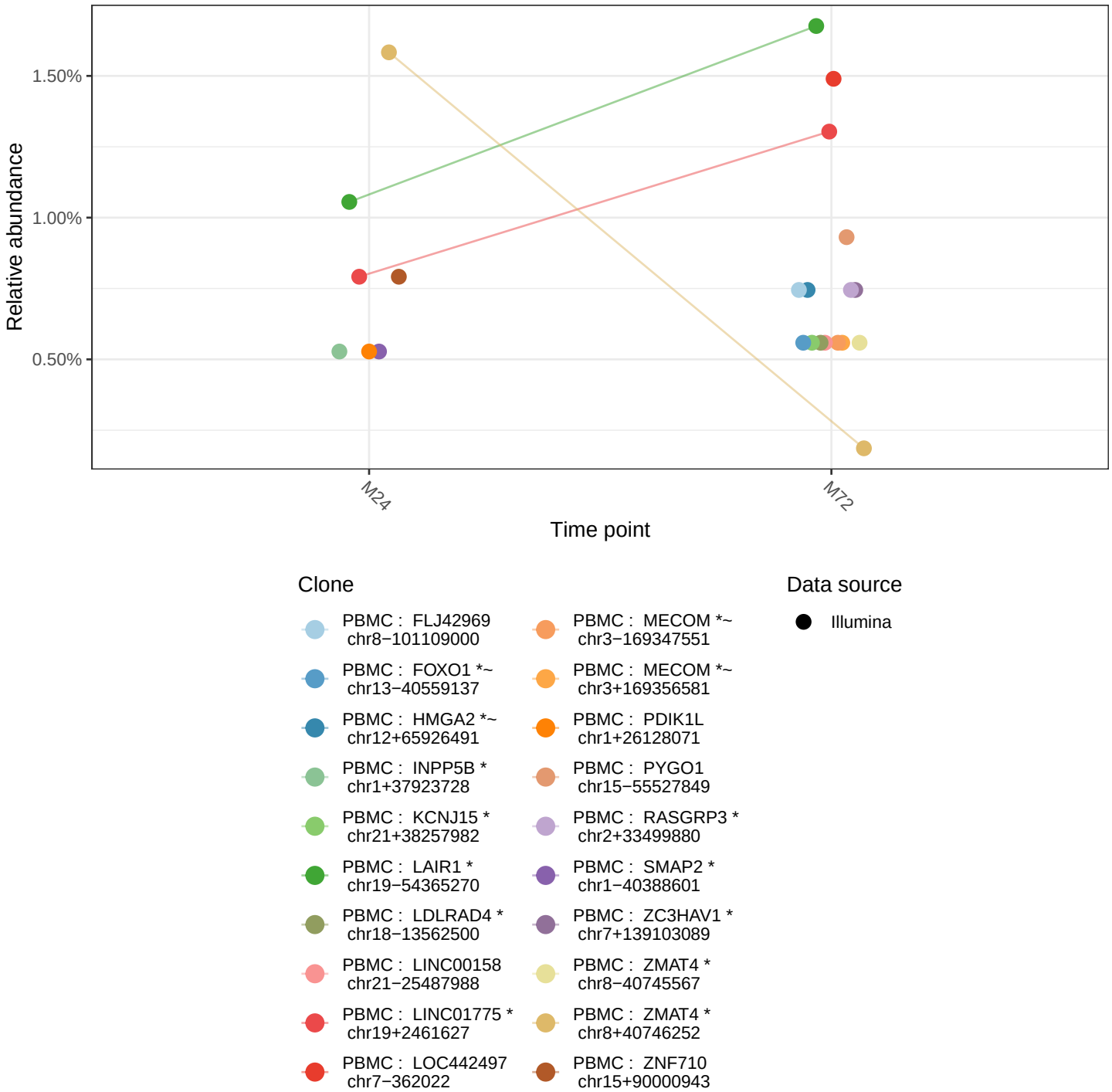
Do any cell clones account for more than 20% of all clones?

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 p410 over time points M24, M72 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

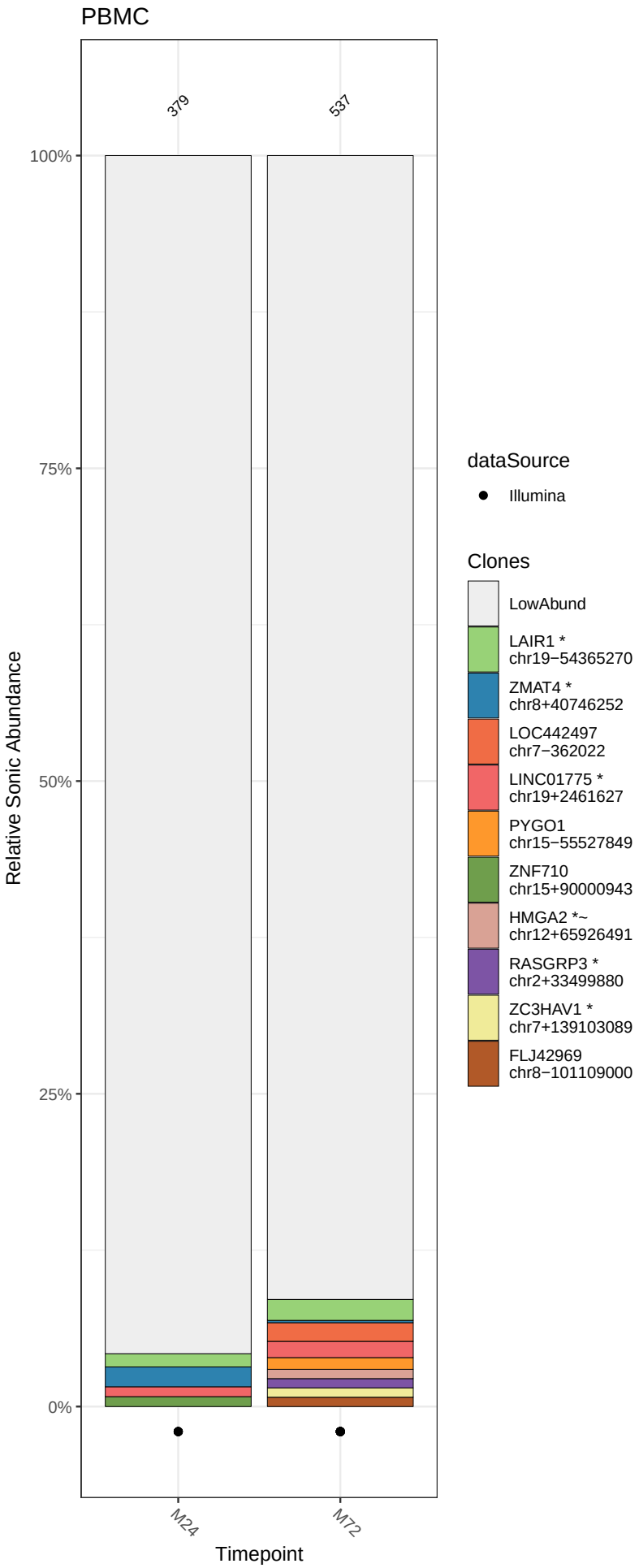
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells, UniqueSite and other columns represent samples which were analyzed but no integration sites were identified.

GTSP	dataSource	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50	Included	runDate	VCN
GTSP3615	Illumina	M24	PBMC	185,874	379	342	0.091	2,220	5.79	0.992	153	yes	2020-10-14	1.130
GTSP3616	Illumina	M72	PBMC	140,628	537	442	0.161	2,068	5.98	0.982	174	yes	2020-10-14	0.905

Tracking of clonal abundances

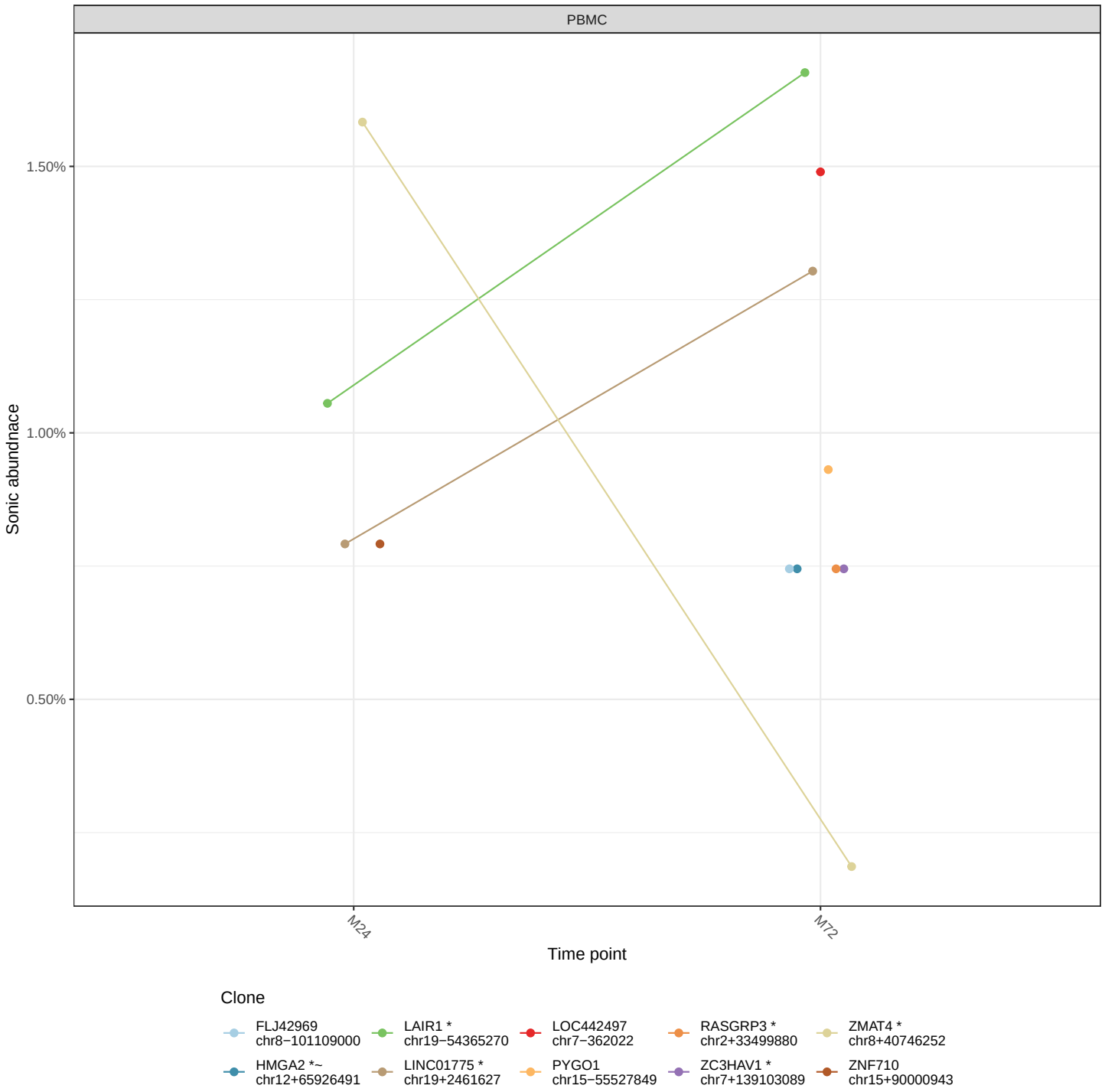
Relative abundance of cell clones

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
379
537
100%
75%
50%
25%
0%
Relative Sonic Abundance
M24
M72
Timepoint
dataSource
Illumina
Clones
LowAbund
LAIR1 *
chr19−54365270
ZMAT4 *
chr8+40746252
LOC442497
chr7−362022
LINC01775 *
chr19+2461627
PYGO1
chr15−55527849
ZNF710
chr15+90000943
HMGA2 *~
chr12+65926491
RASGRP3 *
chr2+33499880
ZC3HAV1 *
chr7+139103089
FLJ42969
chr8−101109000

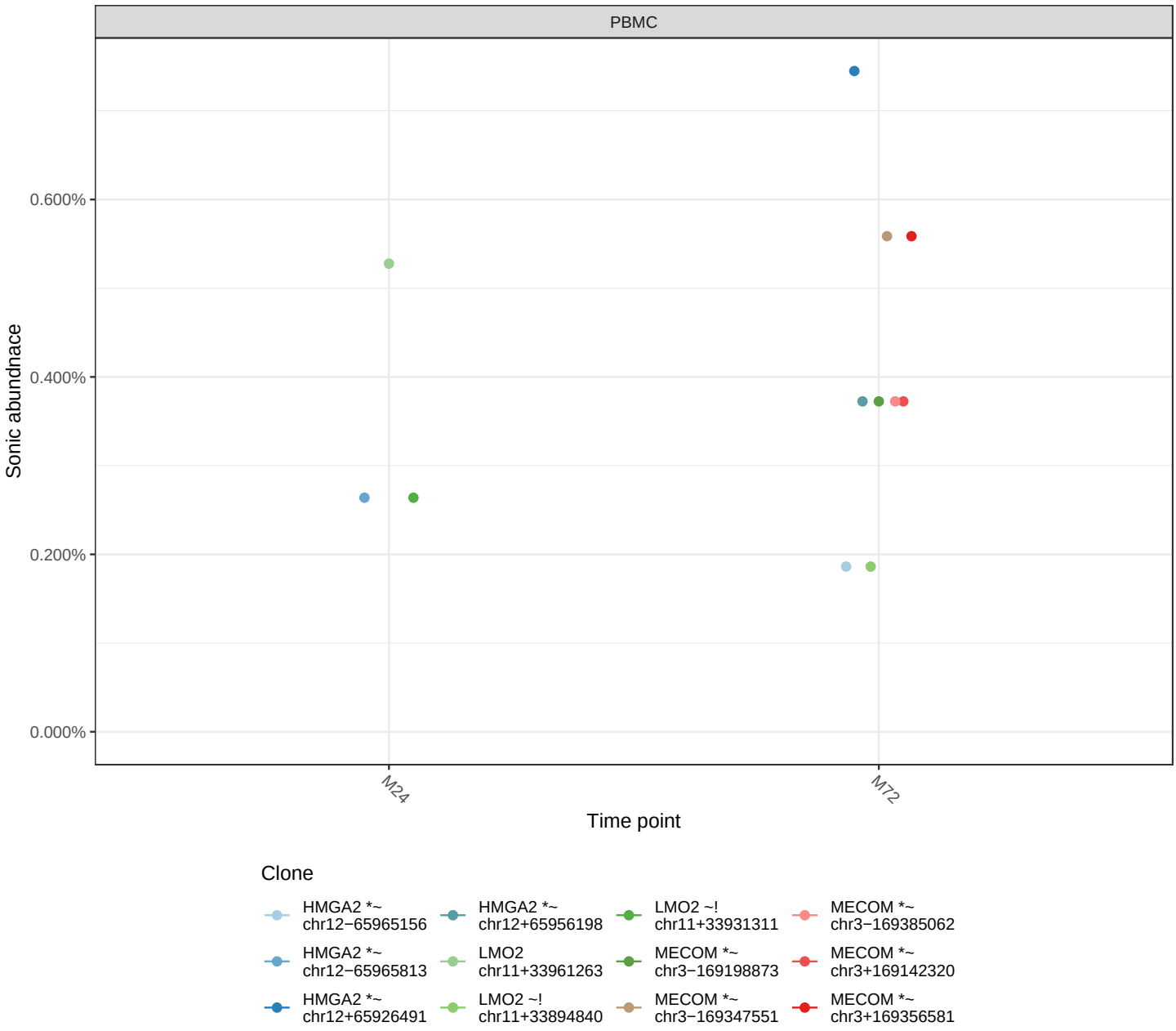
Longitudinal behavior of major clones

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 10 clones is shown below. For cases where only a single time point is available, the data is plotted as unlinked points.



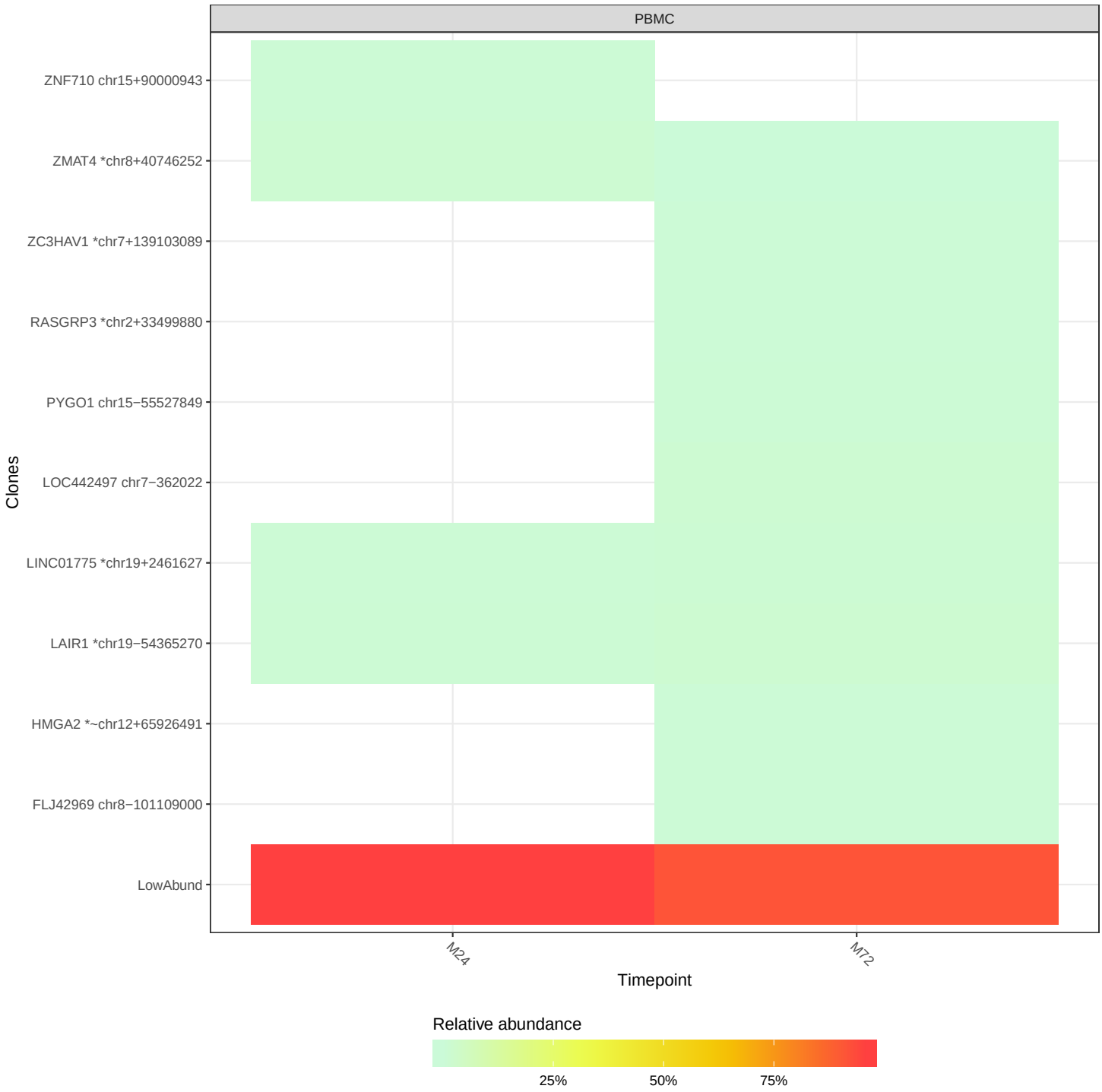
Integration sites near particular genes of interest

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.

PBMC
M24 1:6

PBMC
M72 1:9

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

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)