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

November 02, 2020

Contents

Summary

Is there a rich population of progenitor cells delivering mature cells to the periphery?

To provide a simple measure, we ask whether there are $\geq$ 1000 descendants of independent progenitors (i.e. unique integration sites) in minimally fractionated cell specimens (Whole blood, T cells, B cells, NK cells, Neutrophils, Monocytes and PBMC). Cell specimens that pass these criteria are operationally designated Rich.

Time point	PBMC	Rich
M24	9	No
M84	27	No

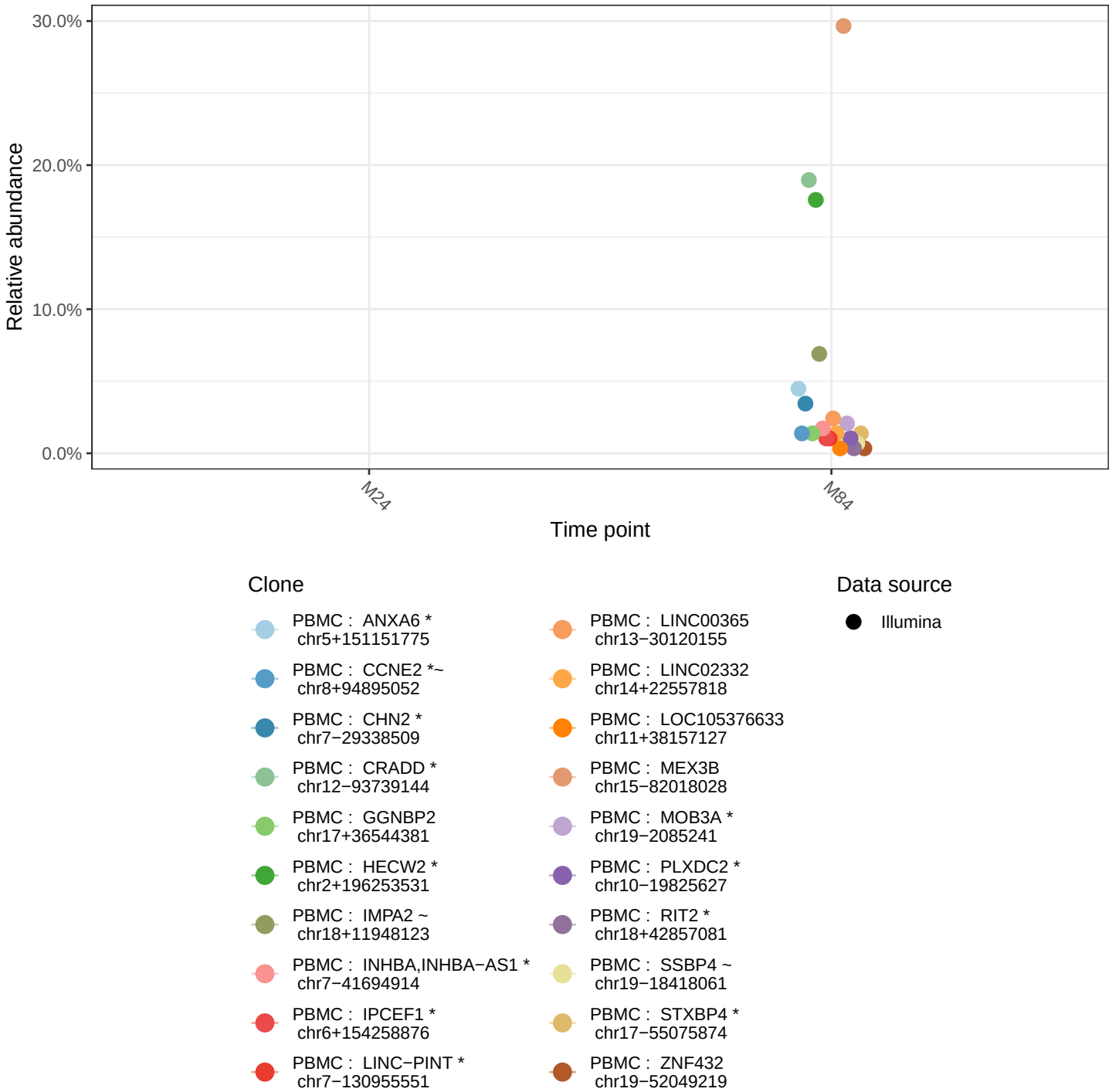
Do any cell clones account for more than 20% of all clones?

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

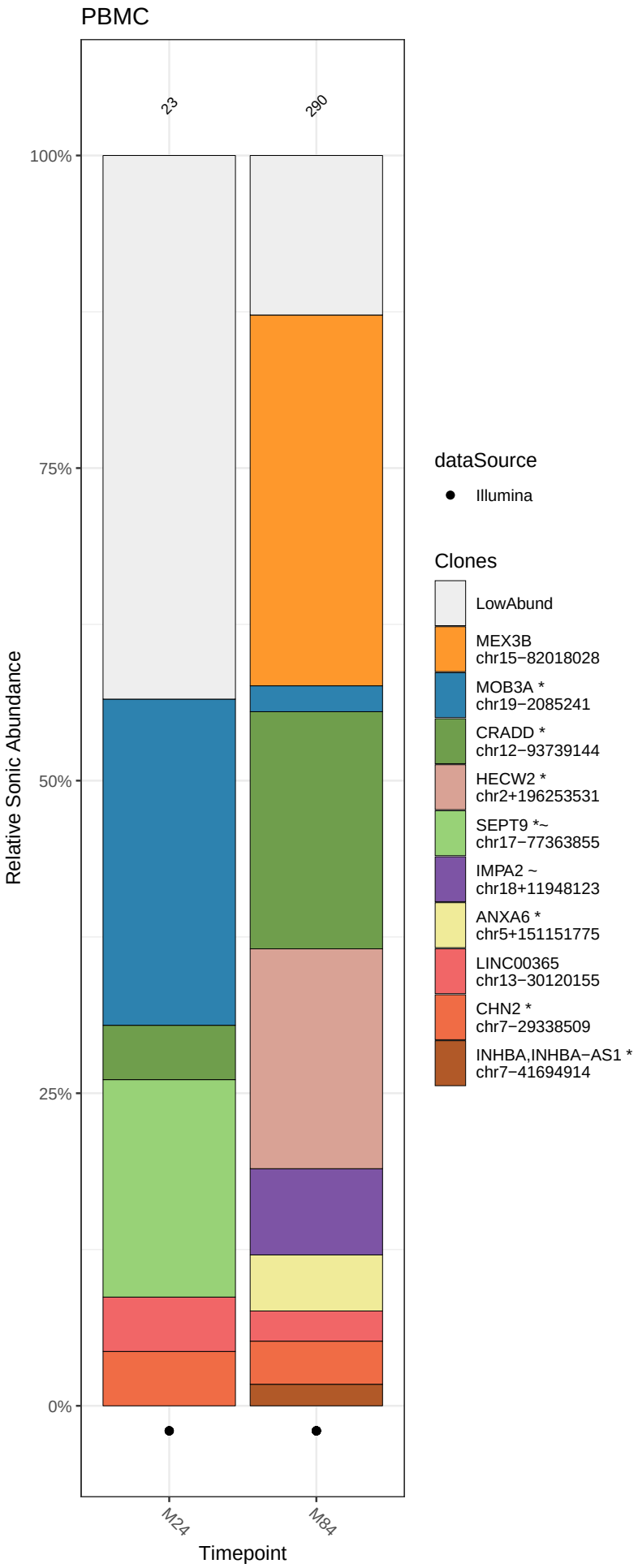
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

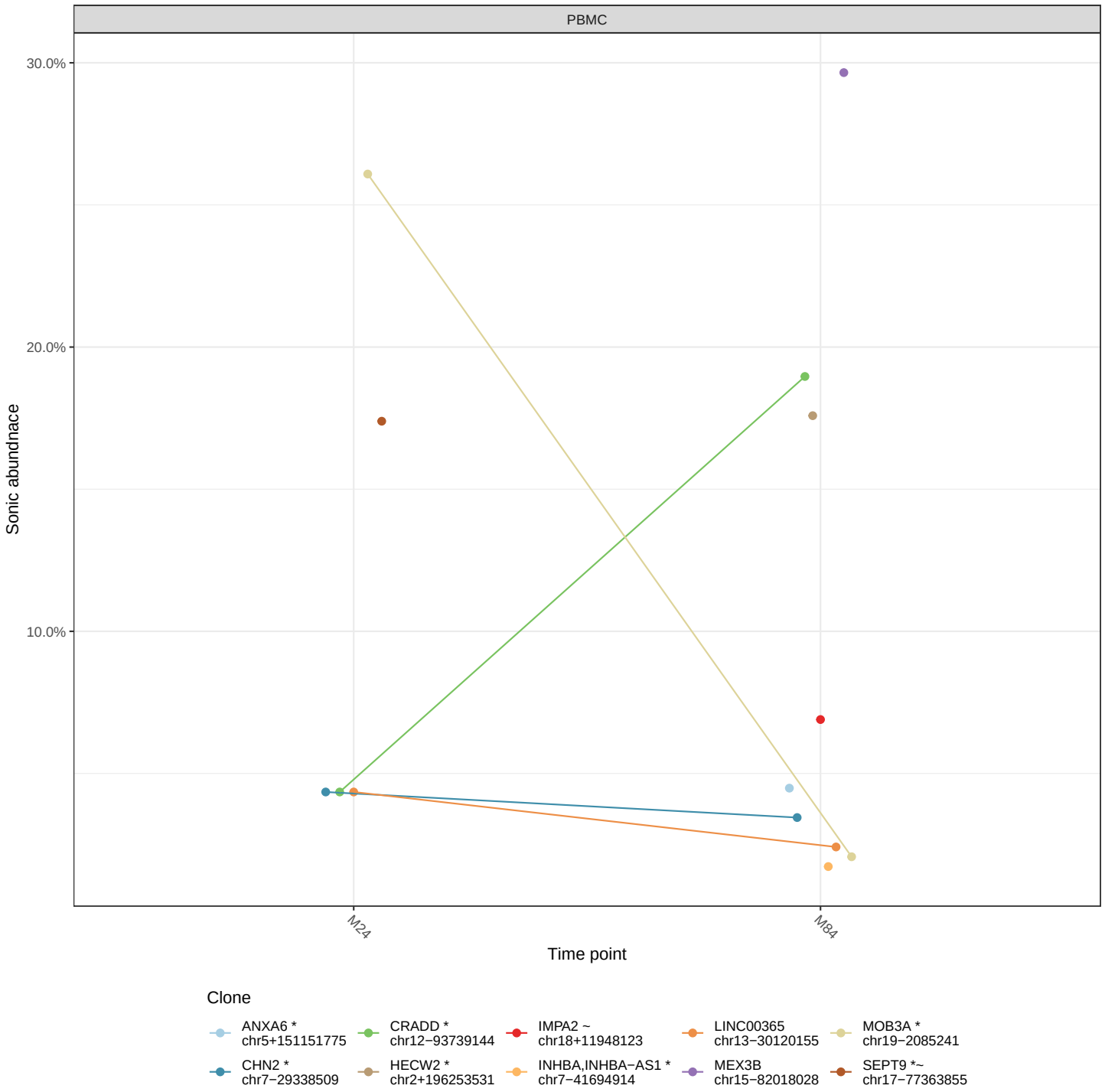
Relative abundance of cell clones

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.



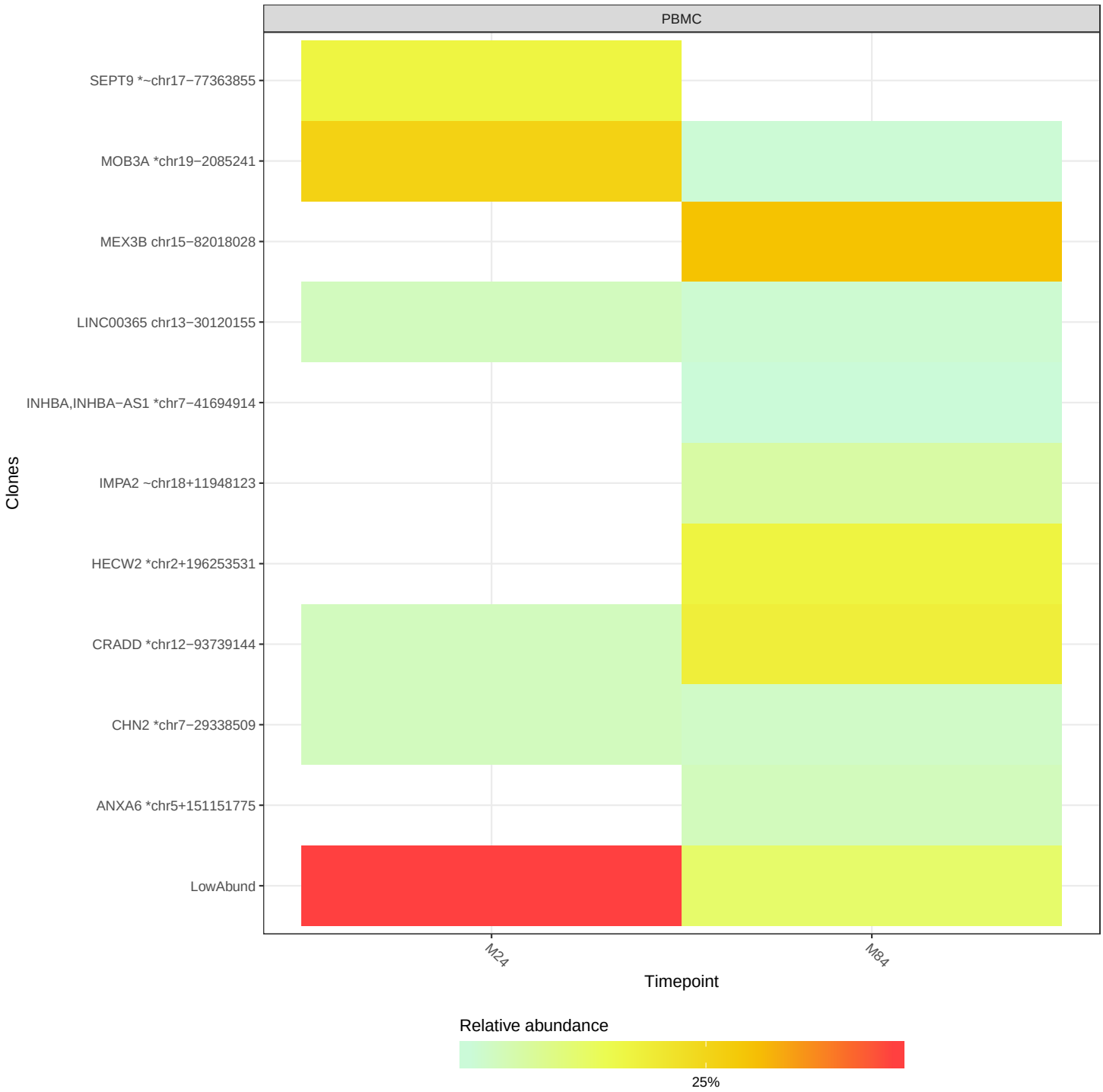
Integration sites near particular genes of interest

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

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

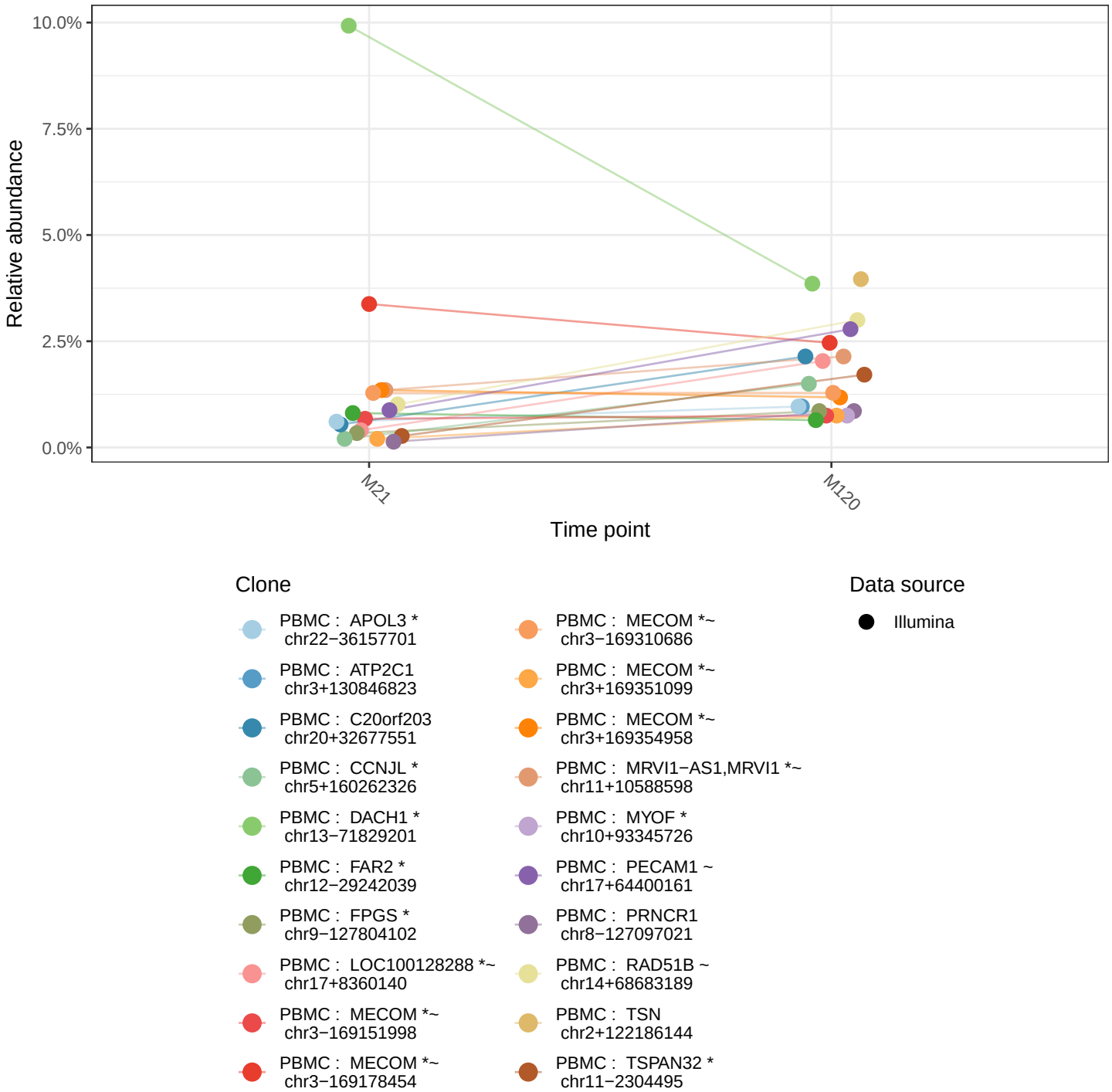
Do any cell clones account for more than 20% of all clones?

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

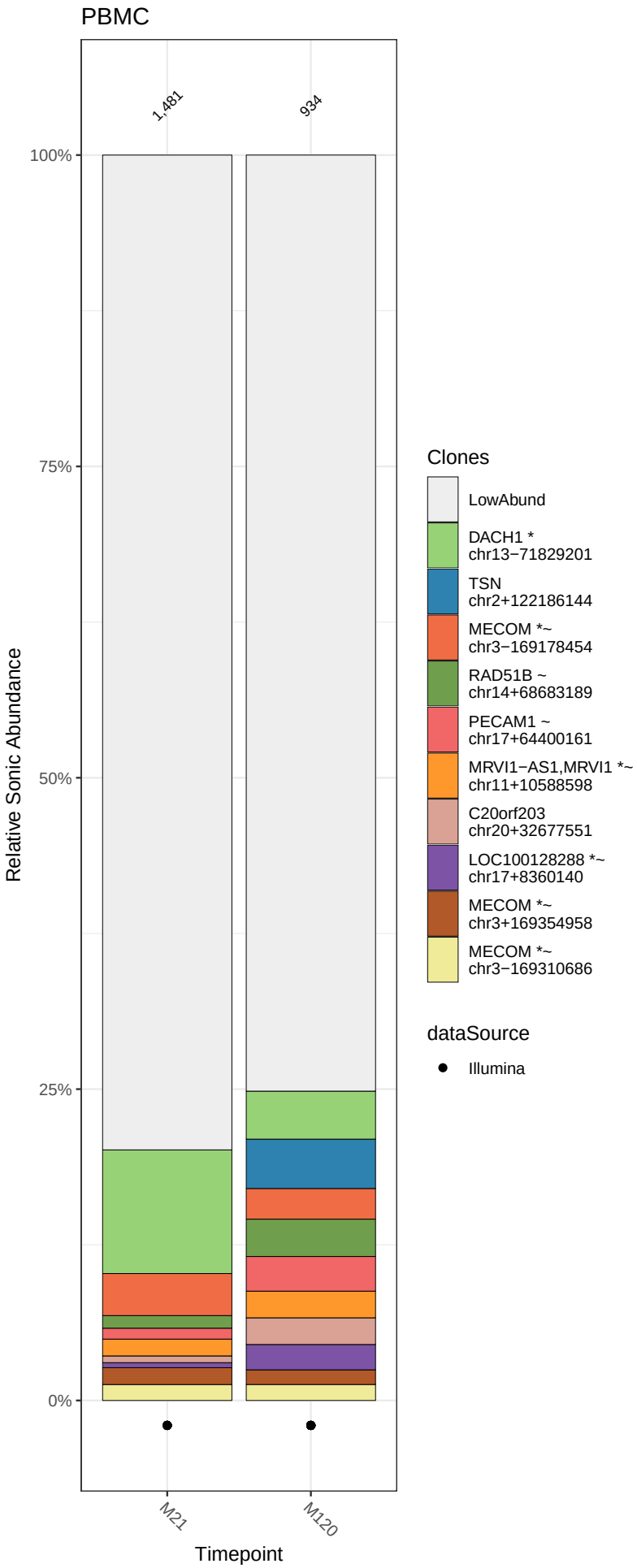
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

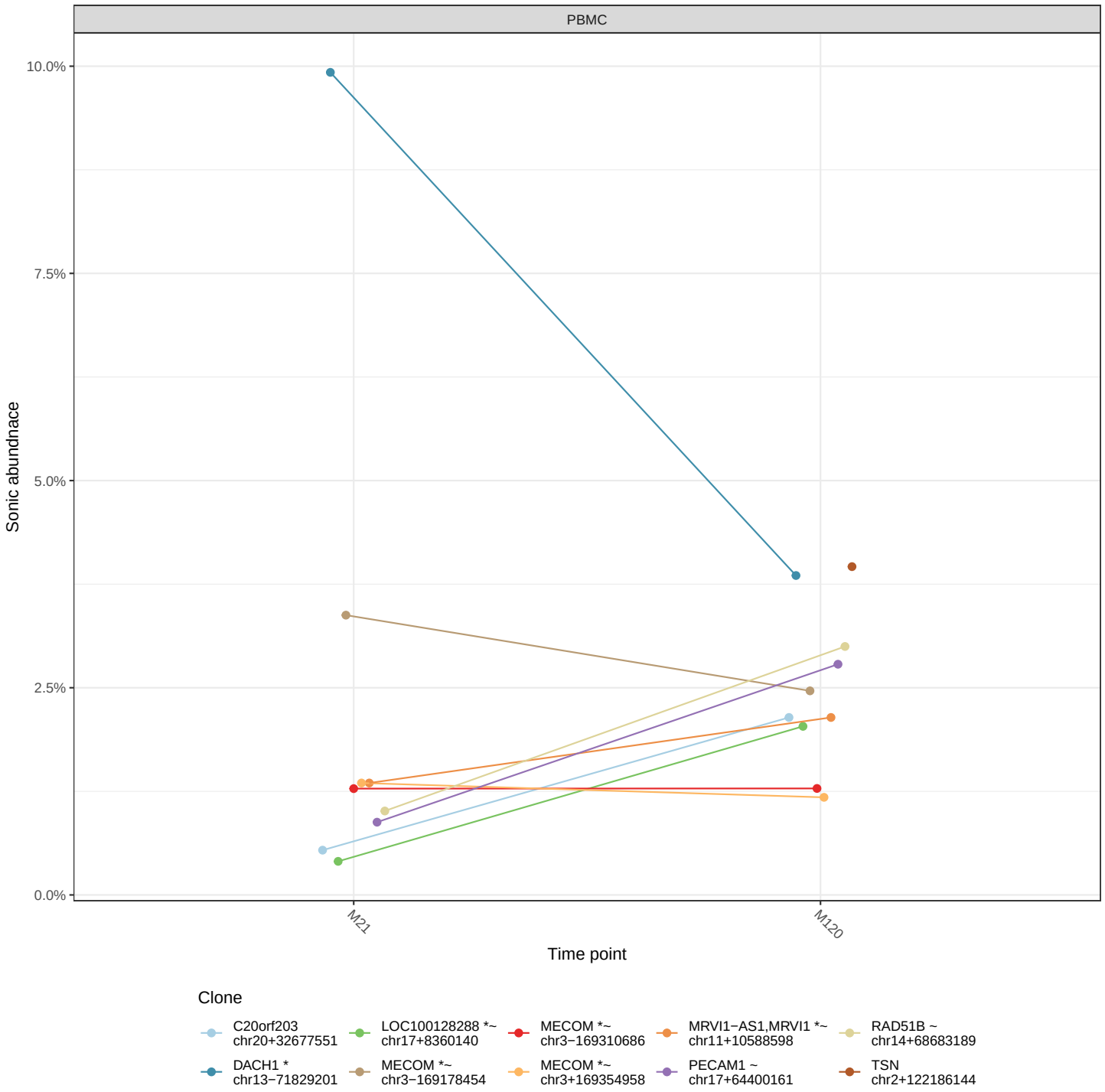
Relative abundance of cell clones

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

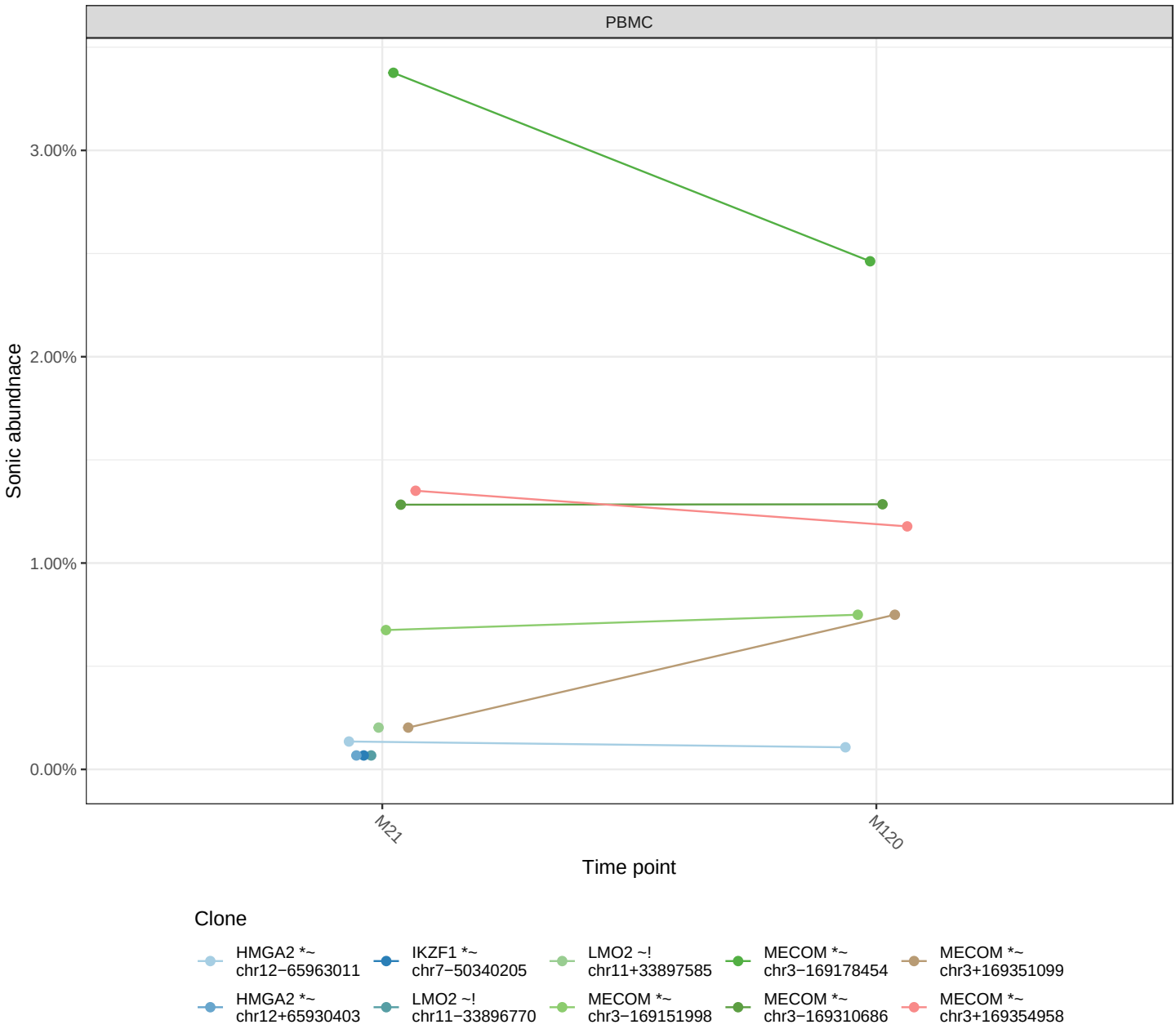
Longitudinal behavior of major clones

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.



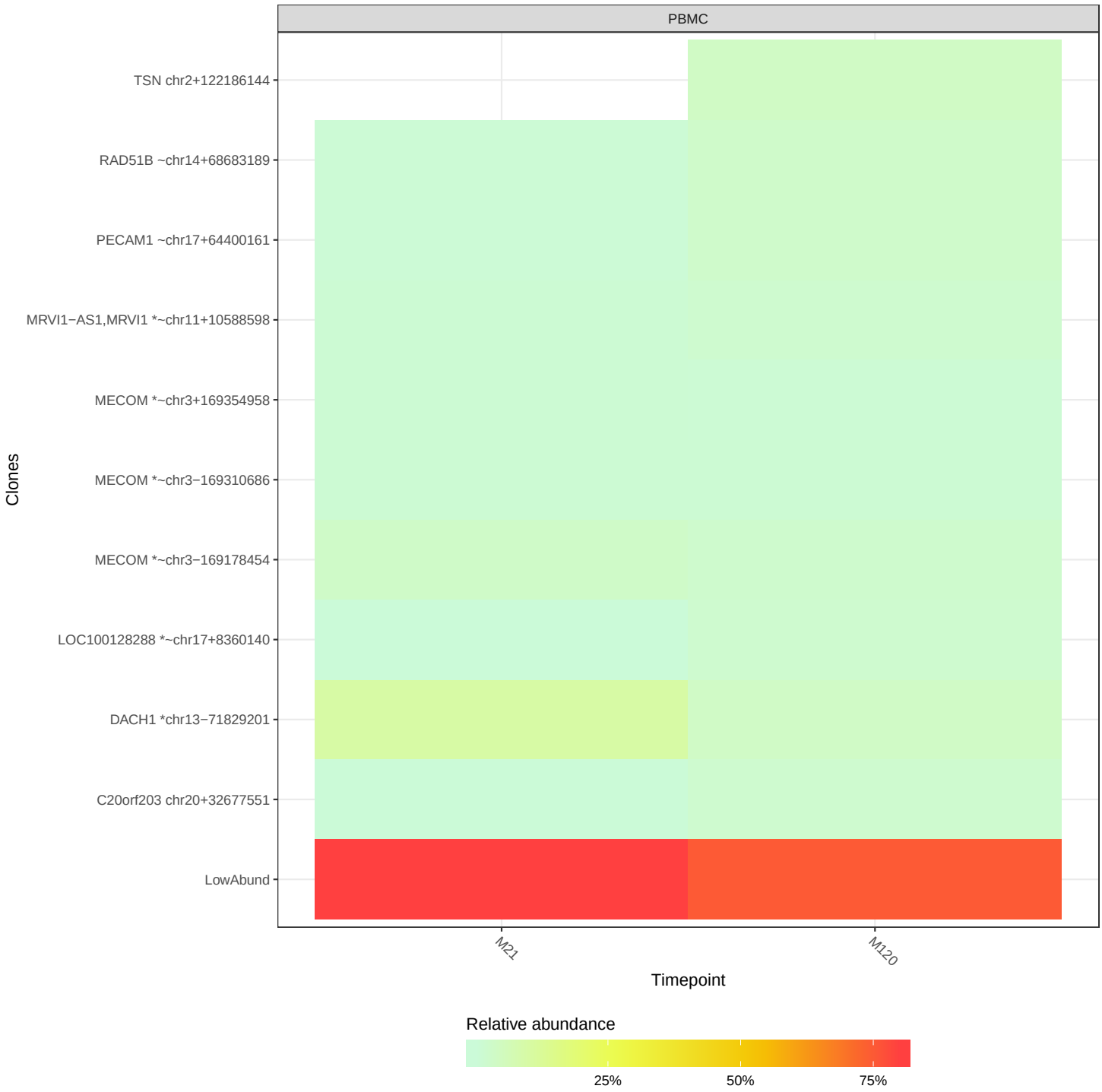
Integration sites near particular genes of interest

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.

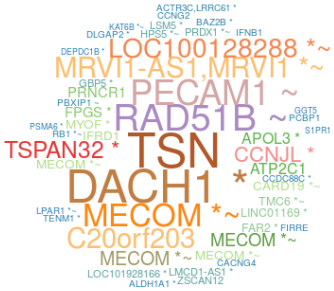


What are the most frequently occuring gene types in the subject?

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

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

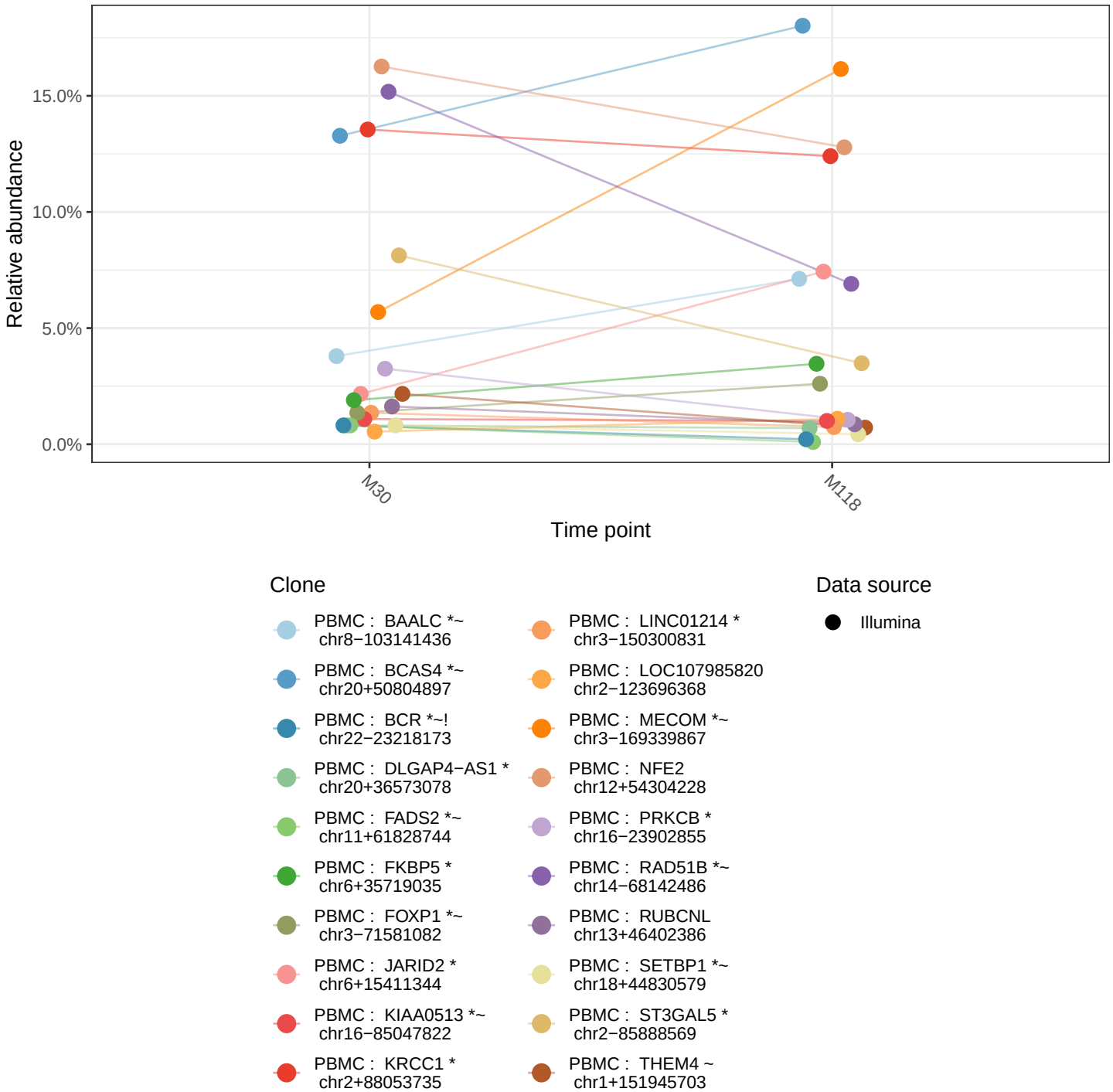
Do any cell clones account for more than 20% of all clones?

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

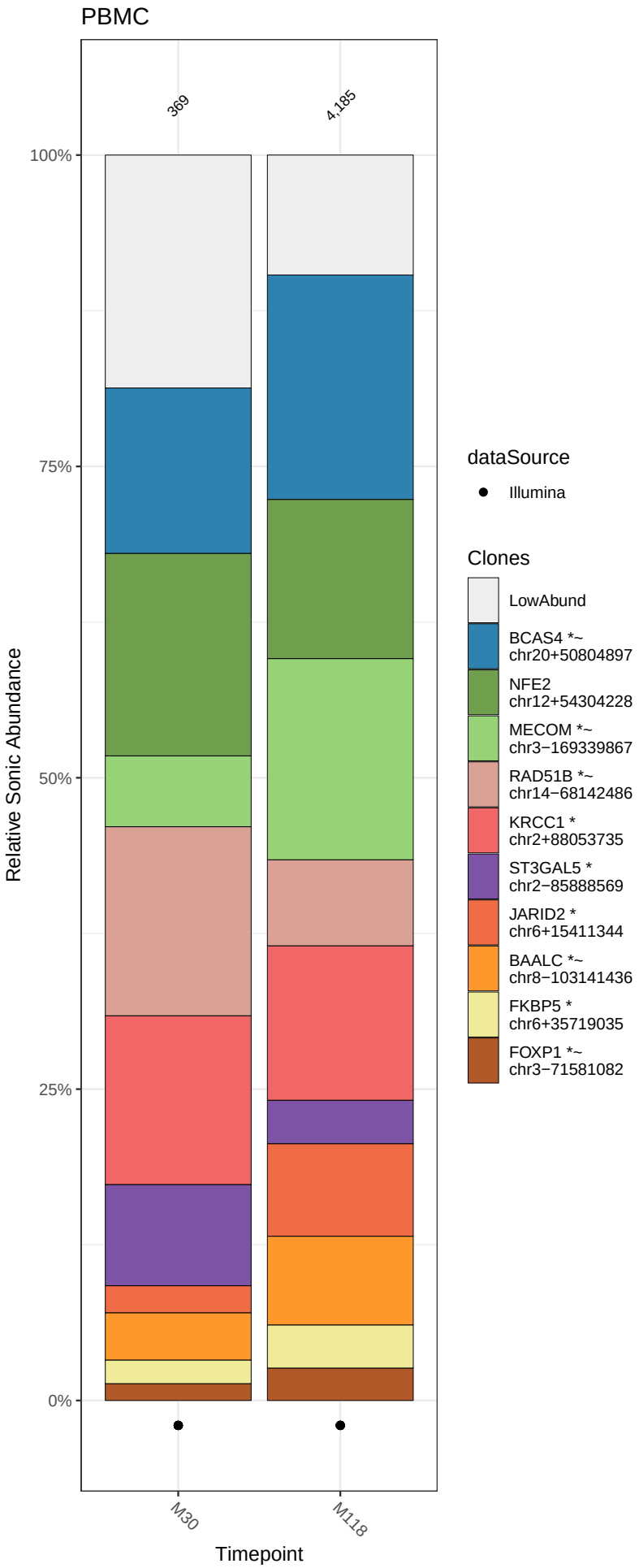
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

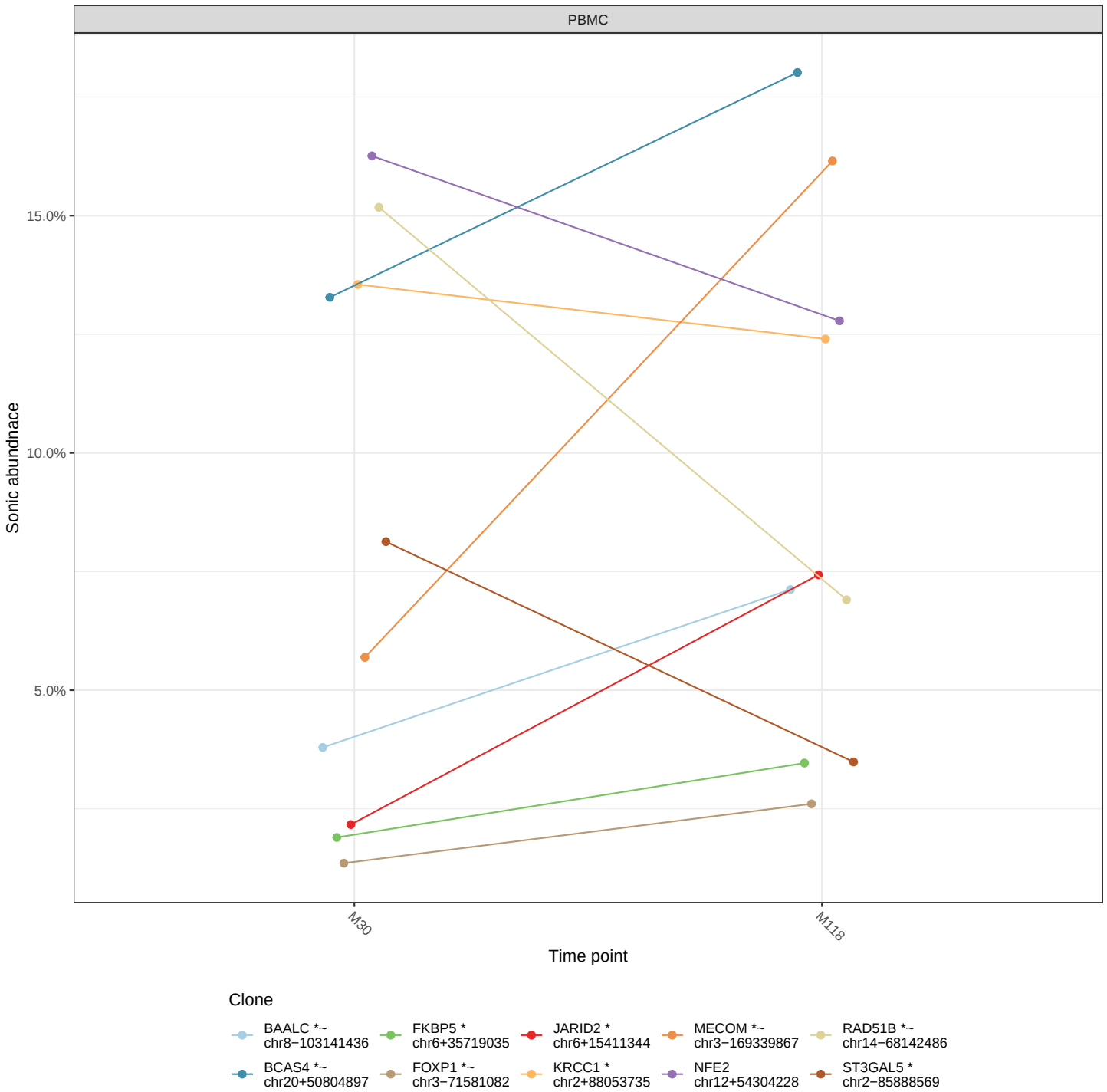
Relative abundance of cell clones

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

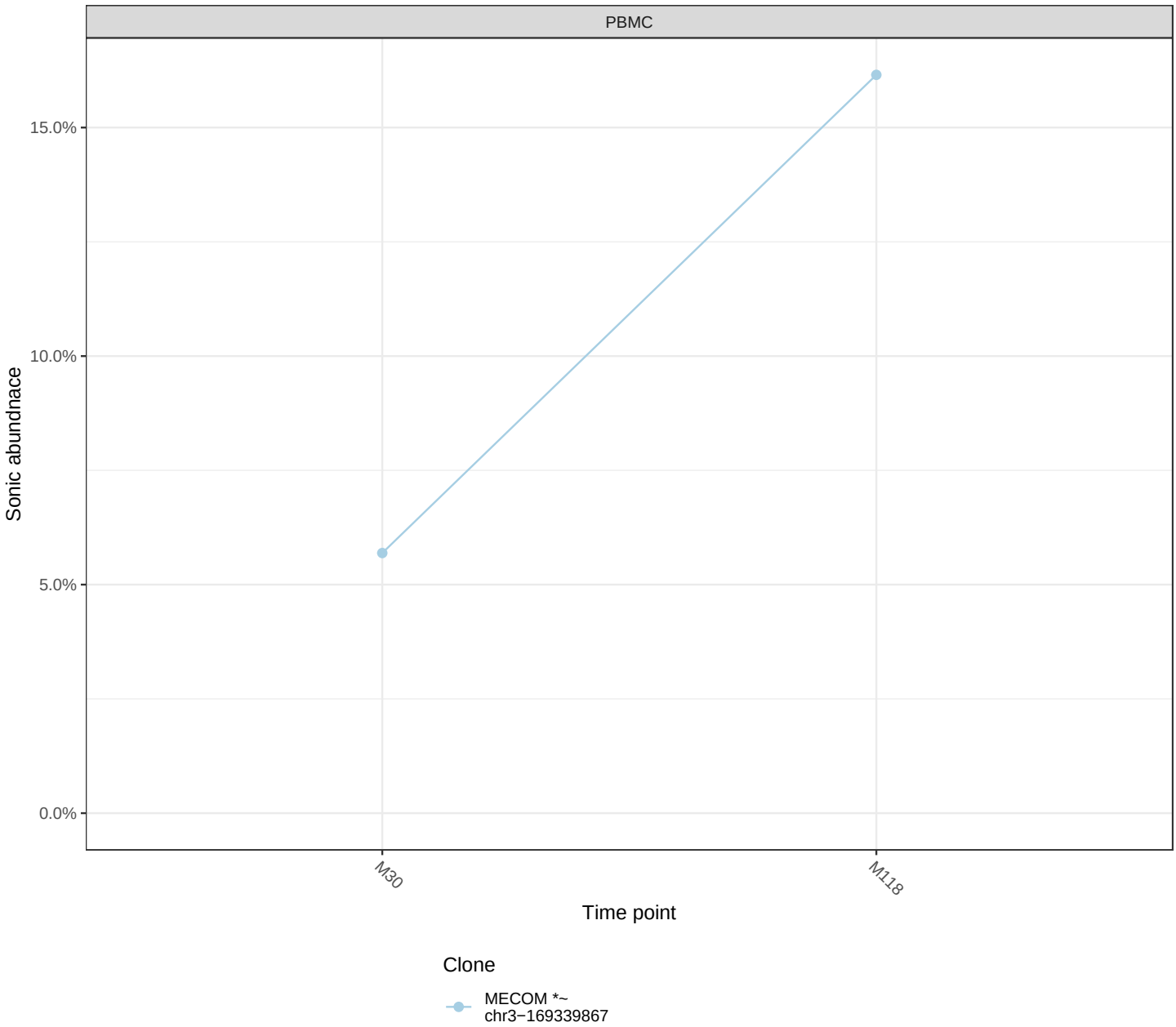
Longitudinal behavior of major clones

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.



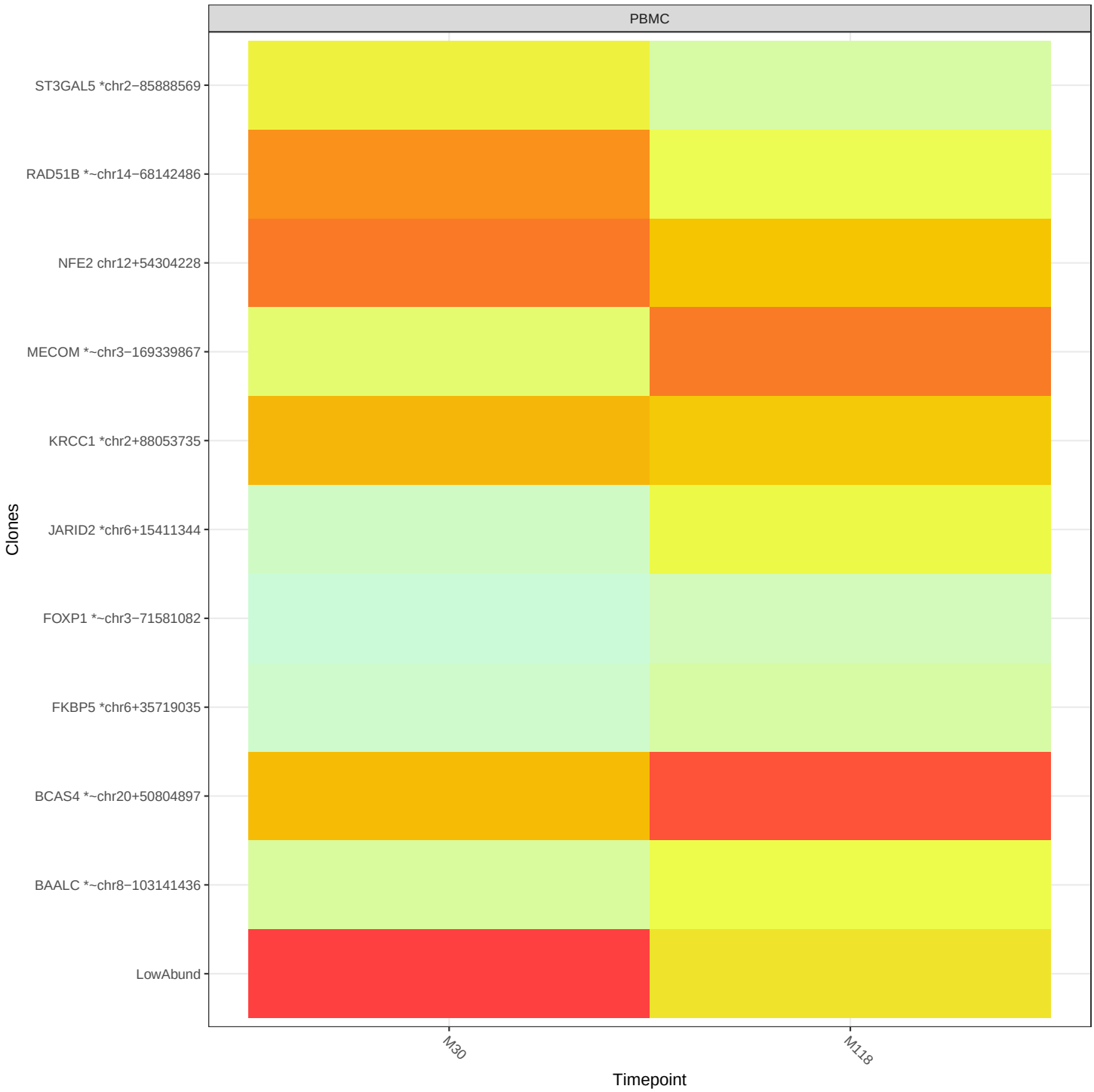
Integration sites near particular genes of interest

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

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

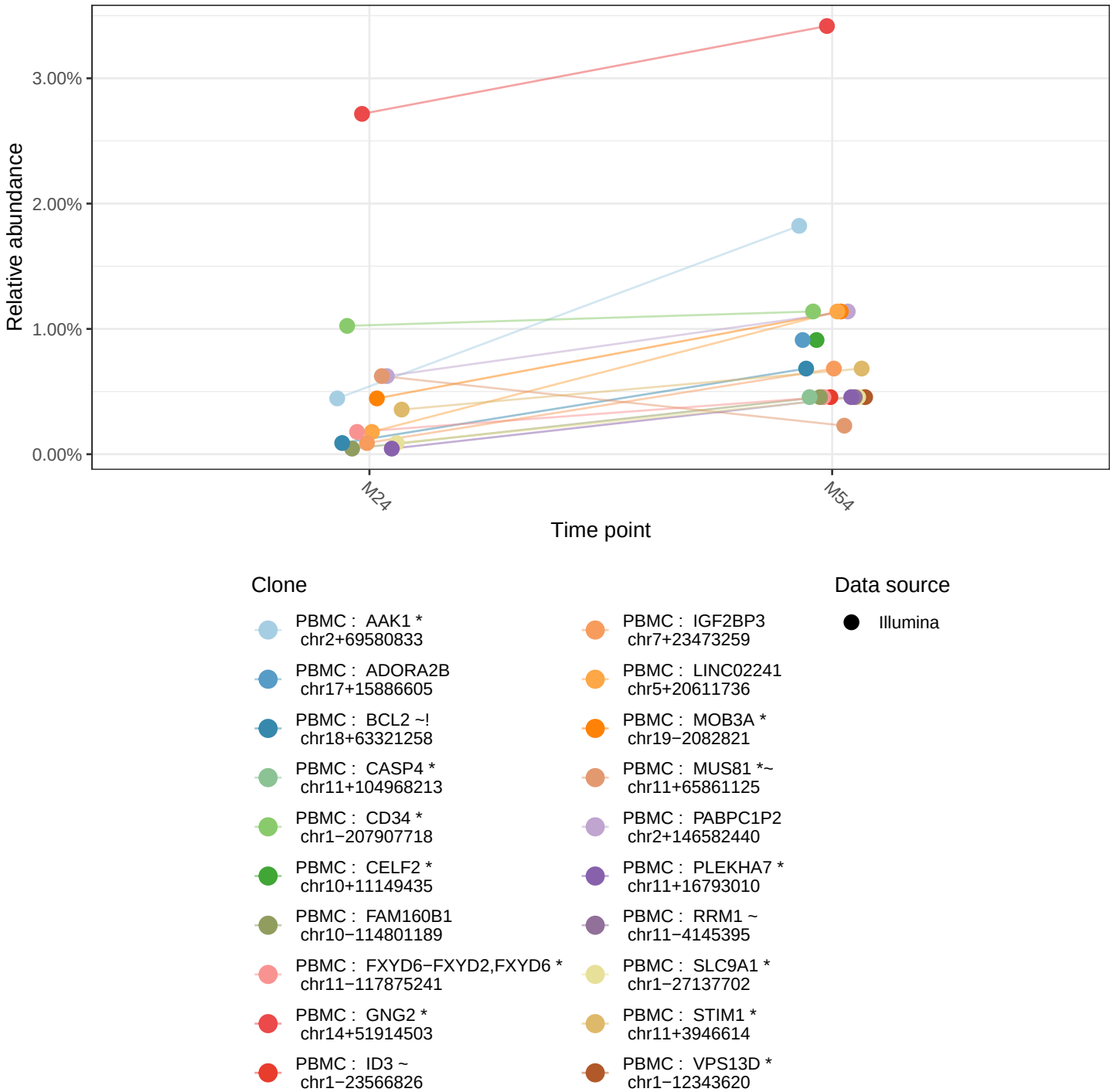
Do any cell clones account for more than 20% of all clones?

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

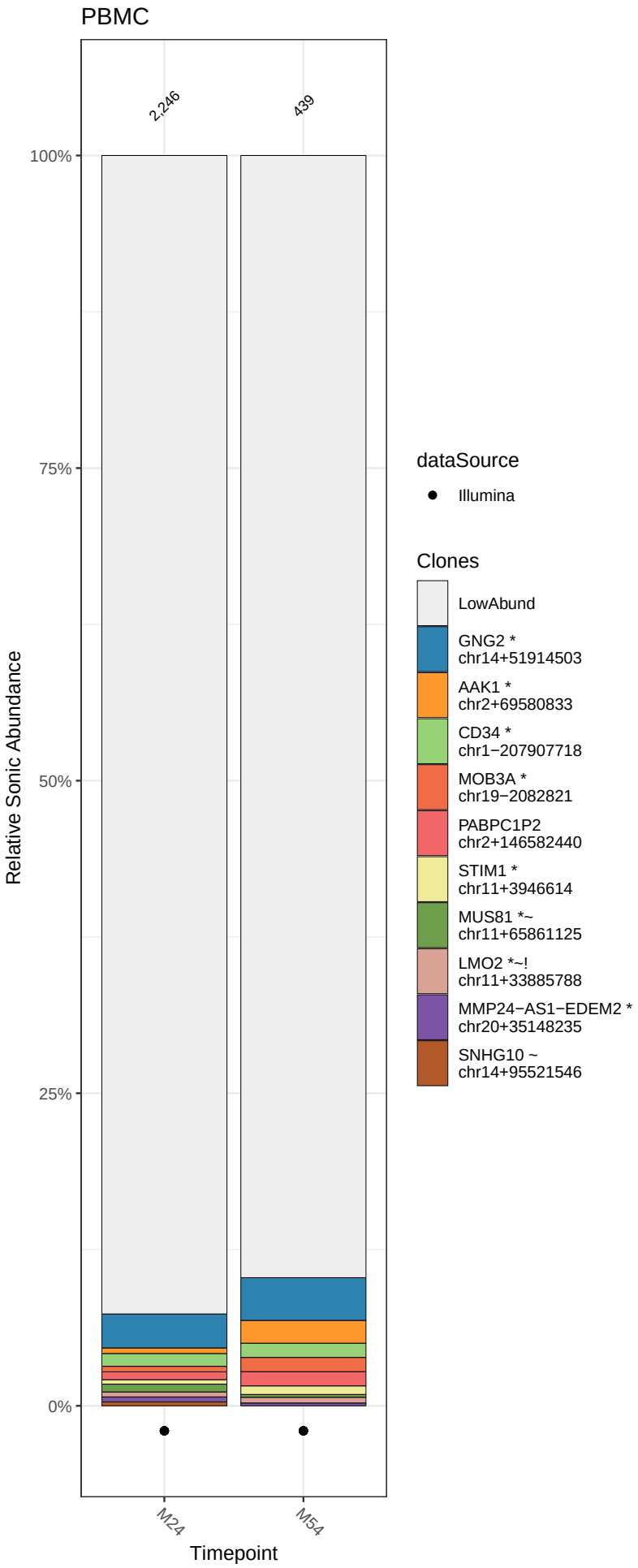
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

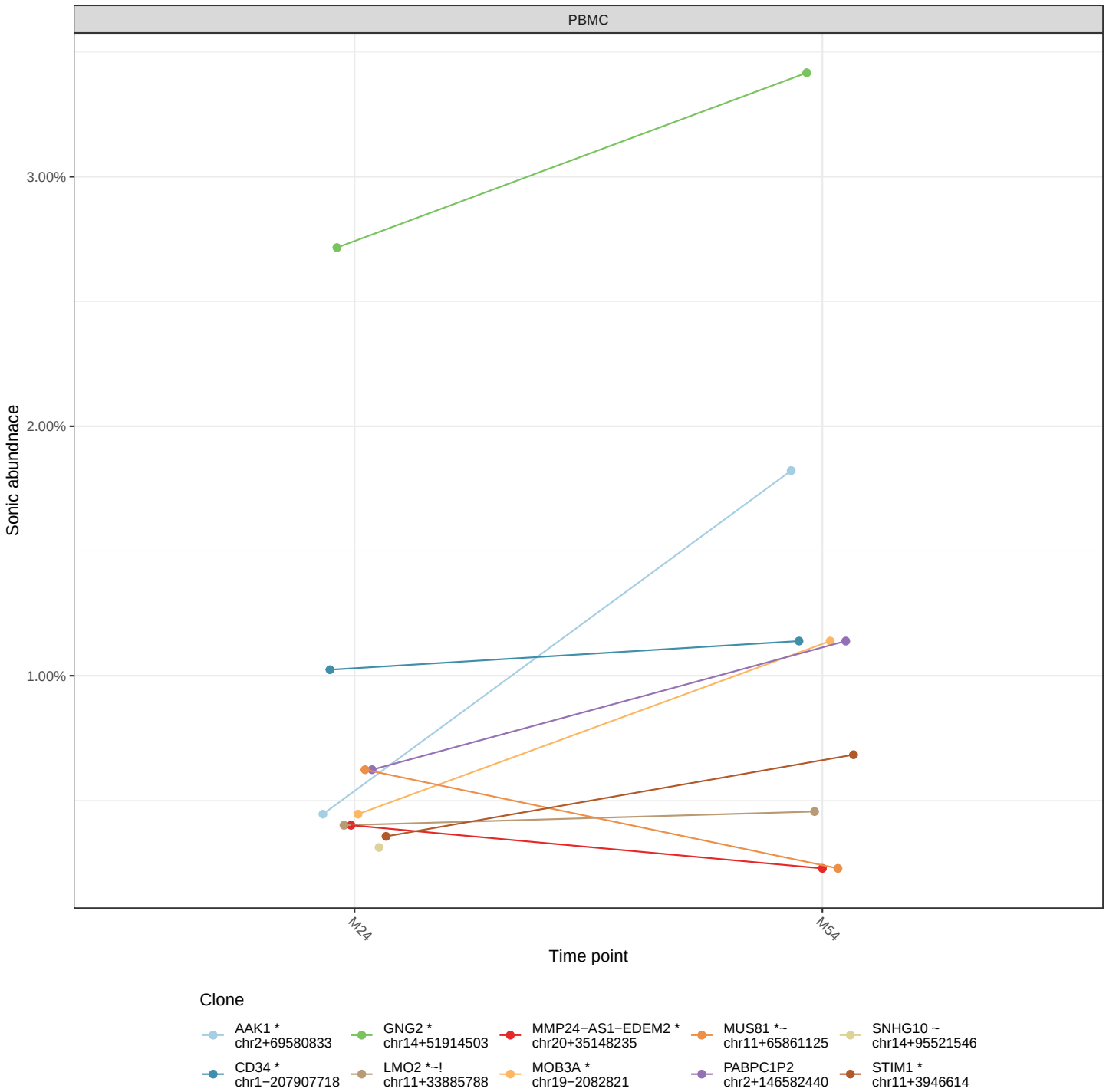
Relative abundance of cell clones

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

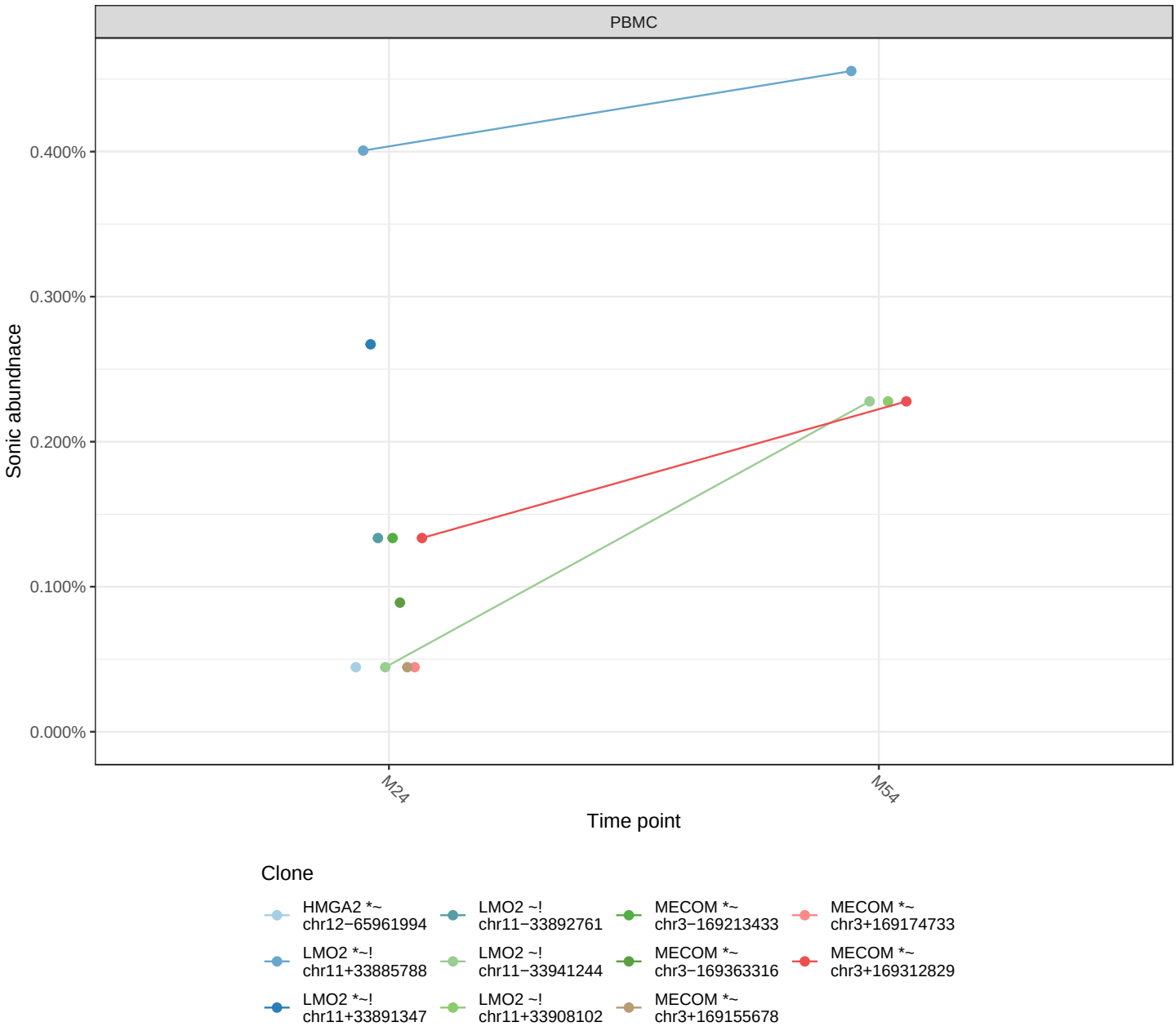
Longitudinal behavior of major clones

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.



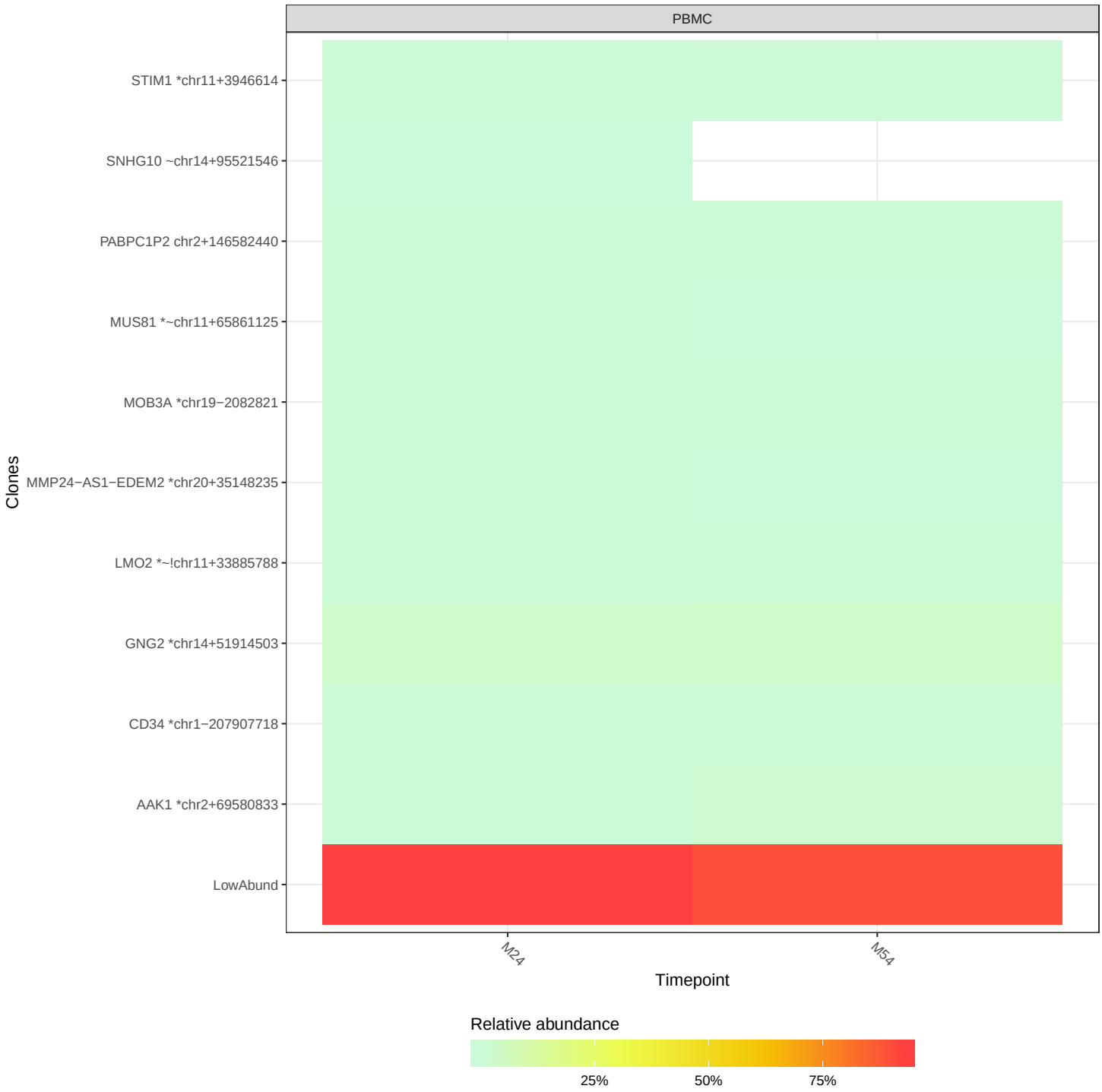
Integration sites near particular genes of interest

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

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

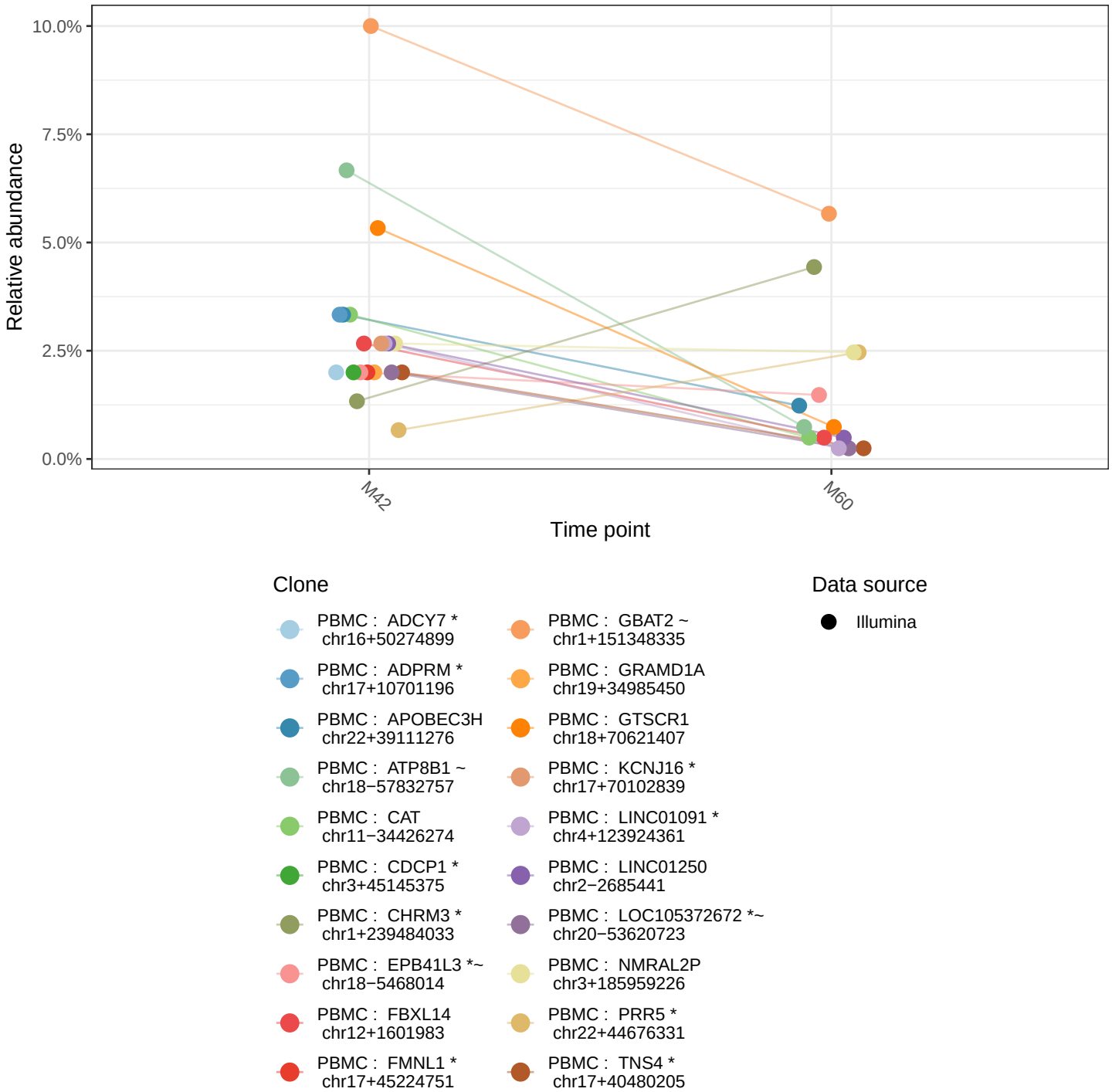
Do any cell clones account for more than 20% of all clones?

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

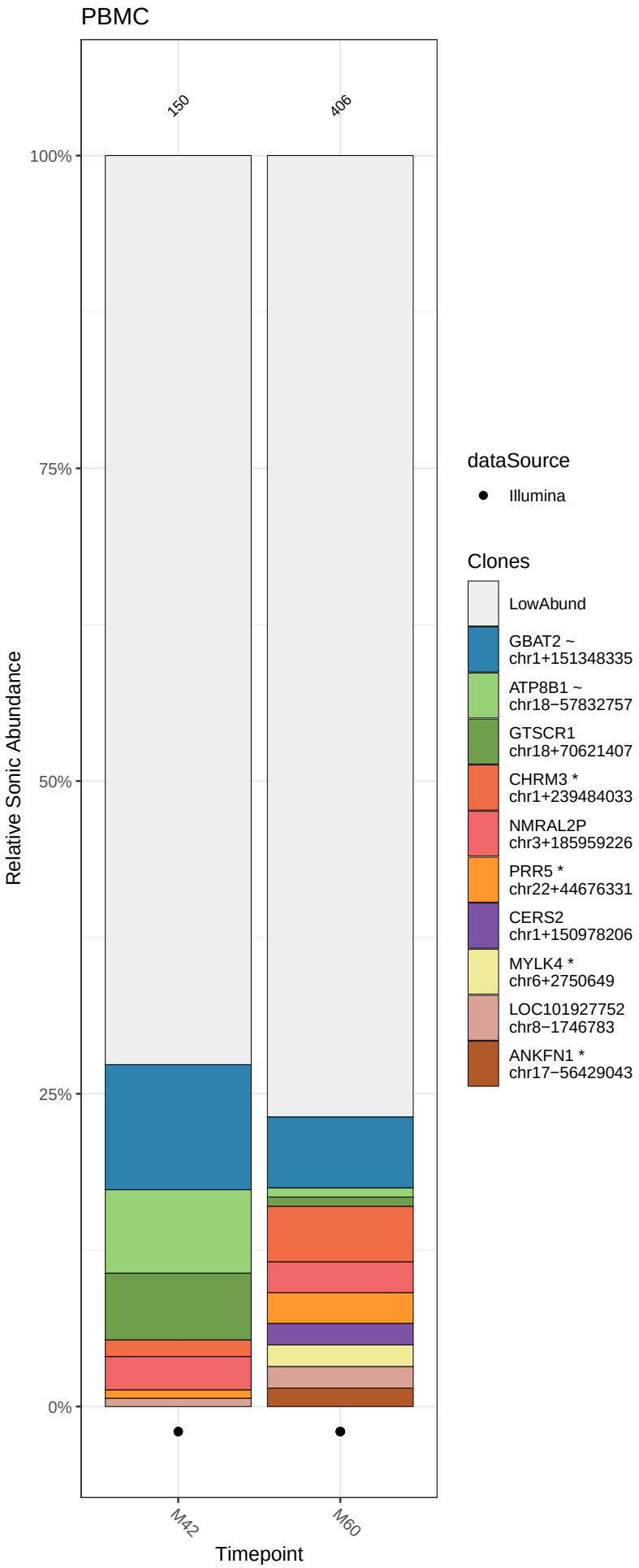
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

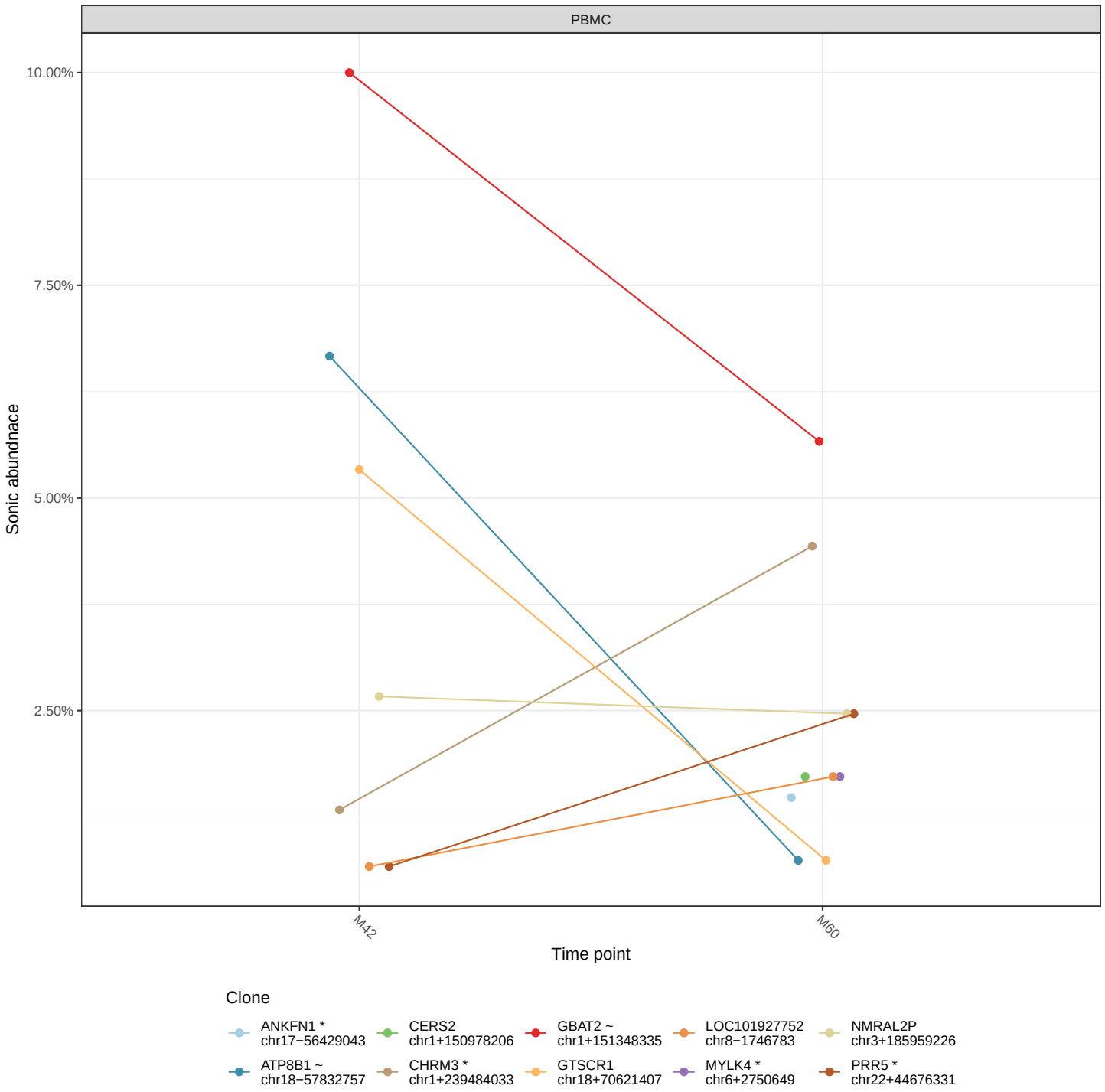
Relative abundance of cell clones

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

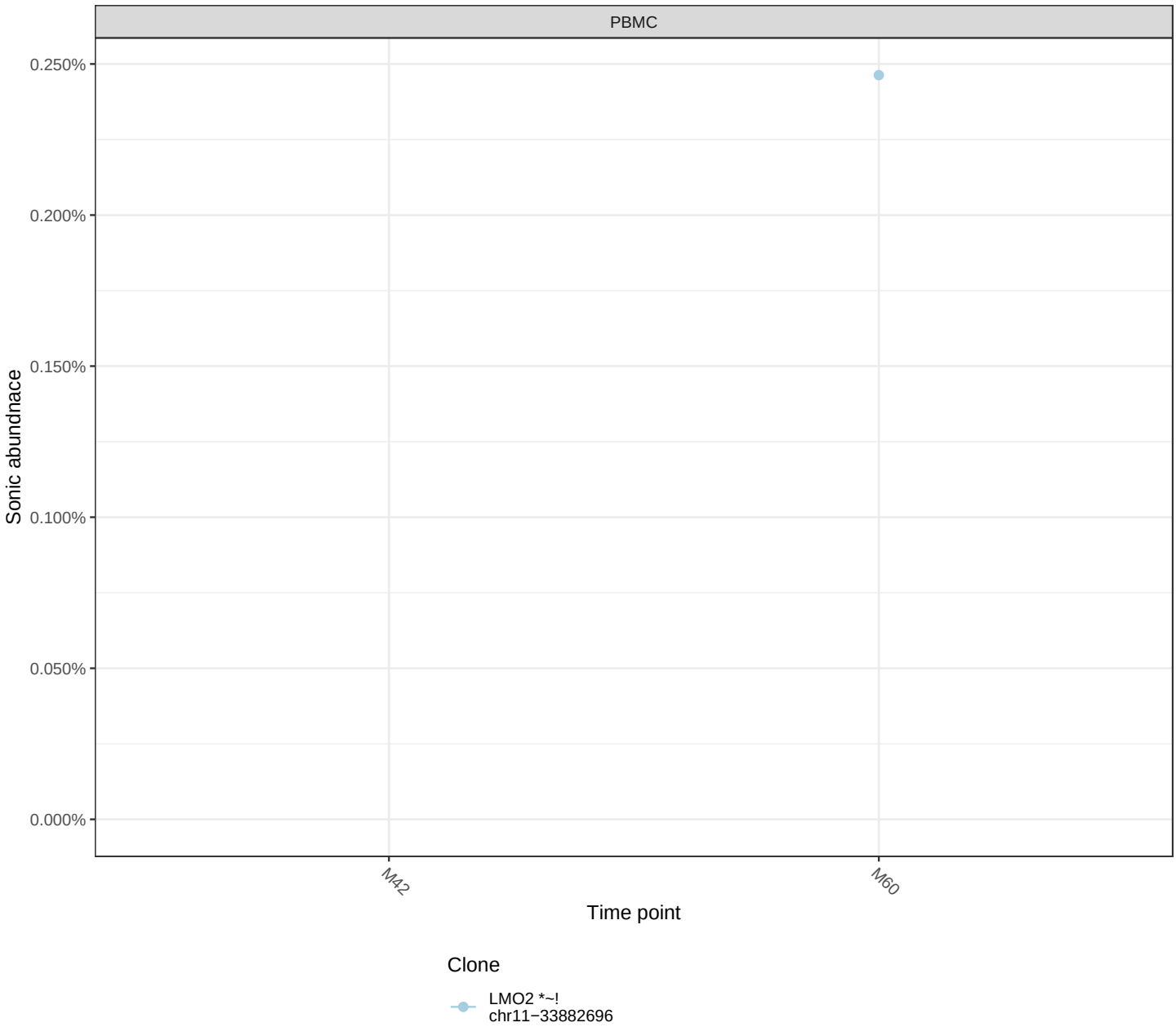
Longitudinal behavior of major clones

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.



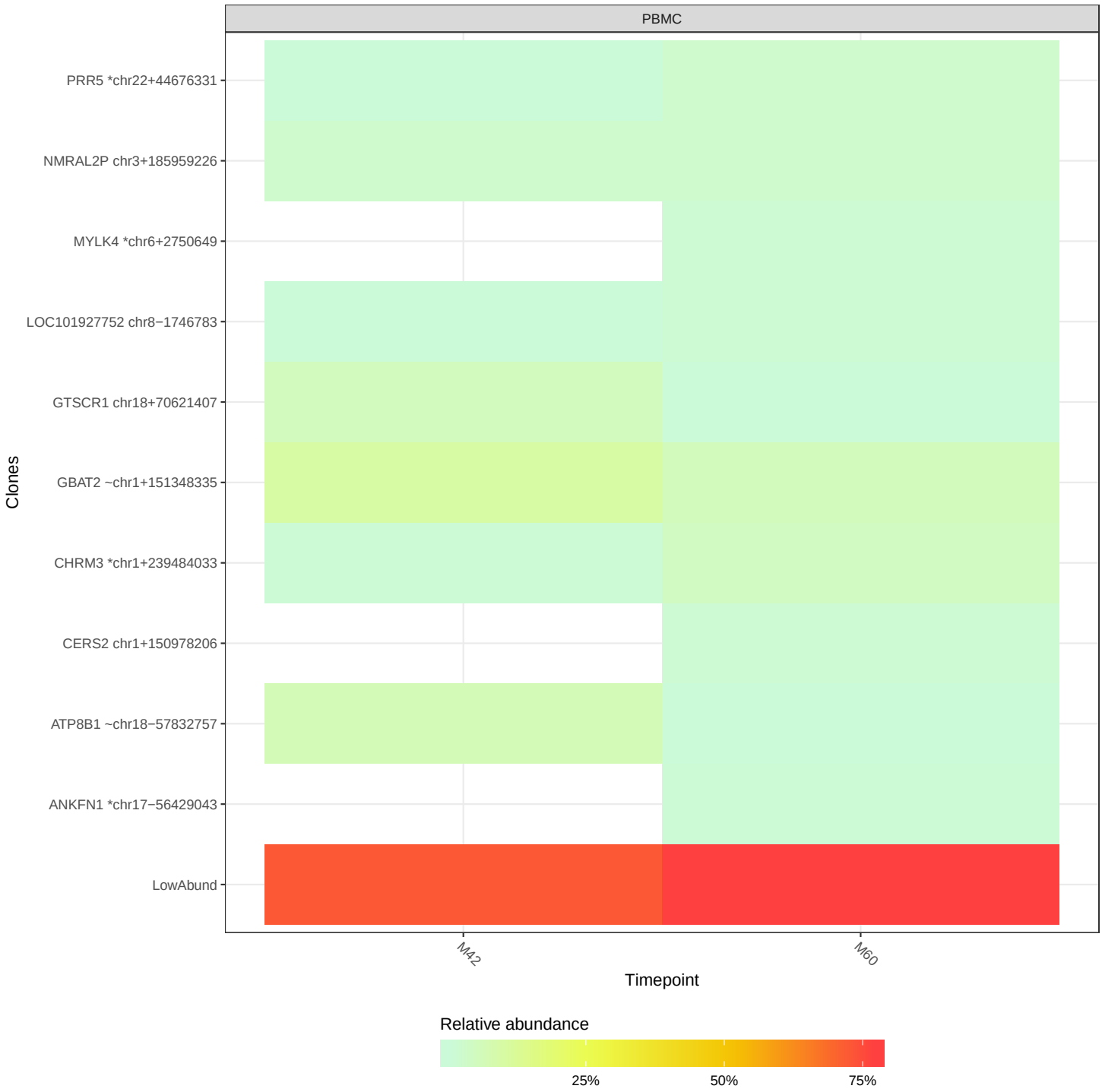
Integration sites near particular genes of interest

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.

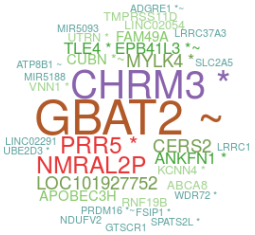


What are the most frequently occuring gene types in the subject?

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

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

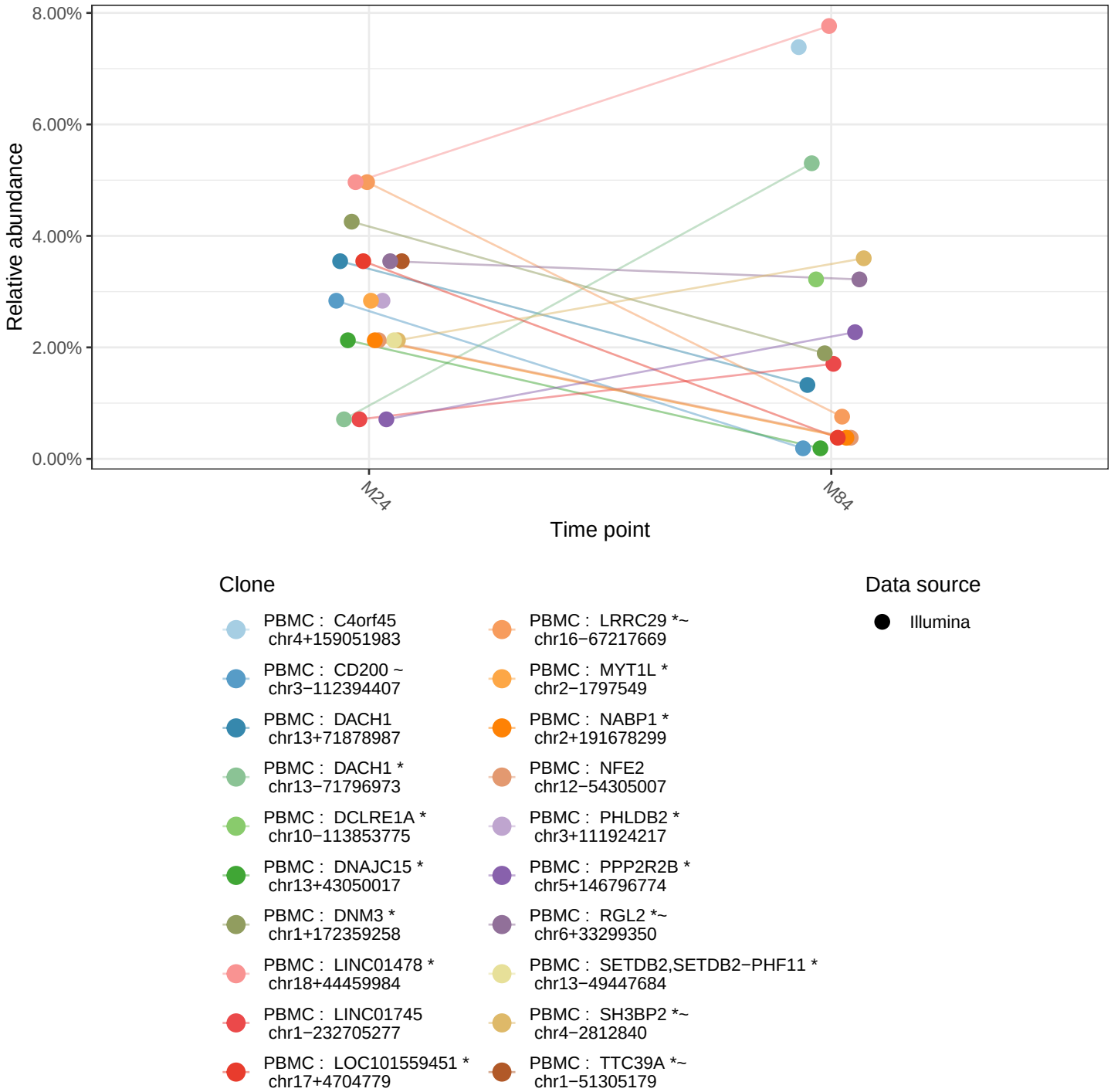
Do any cell clones account for more than 20% of all clones?

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

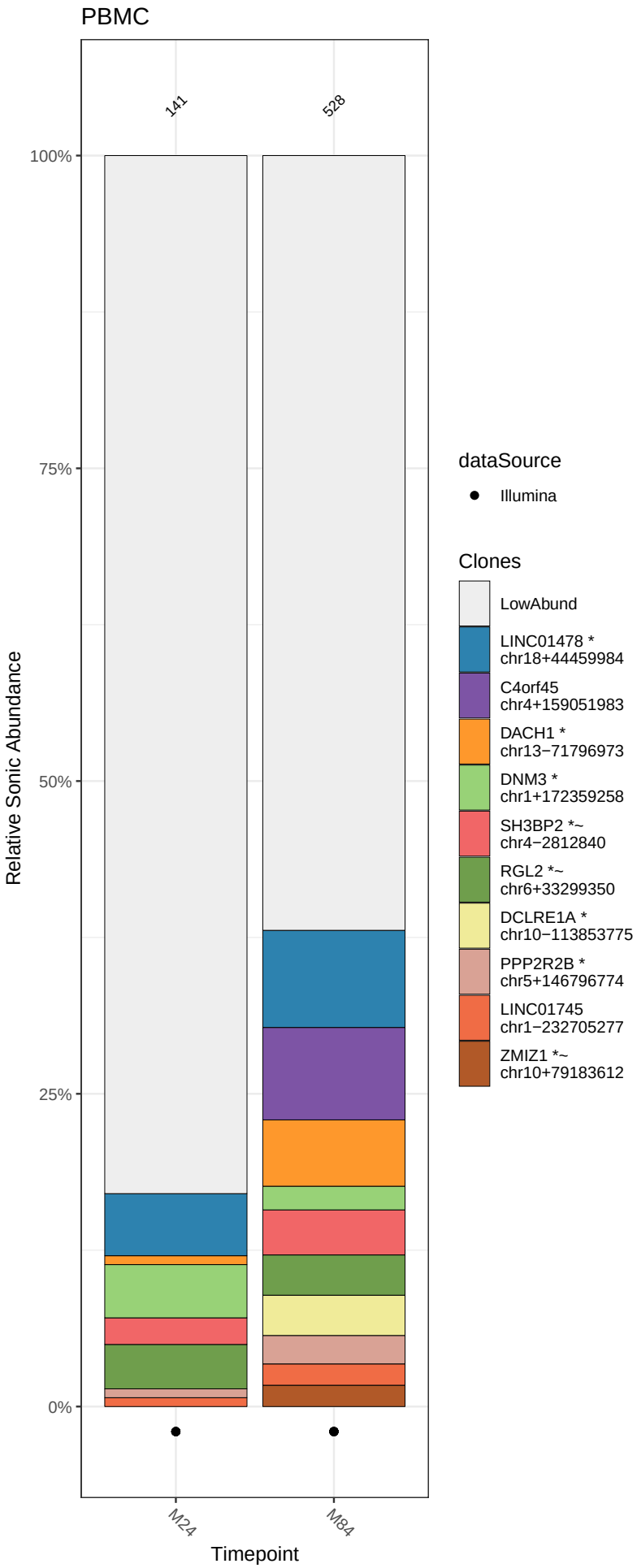
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

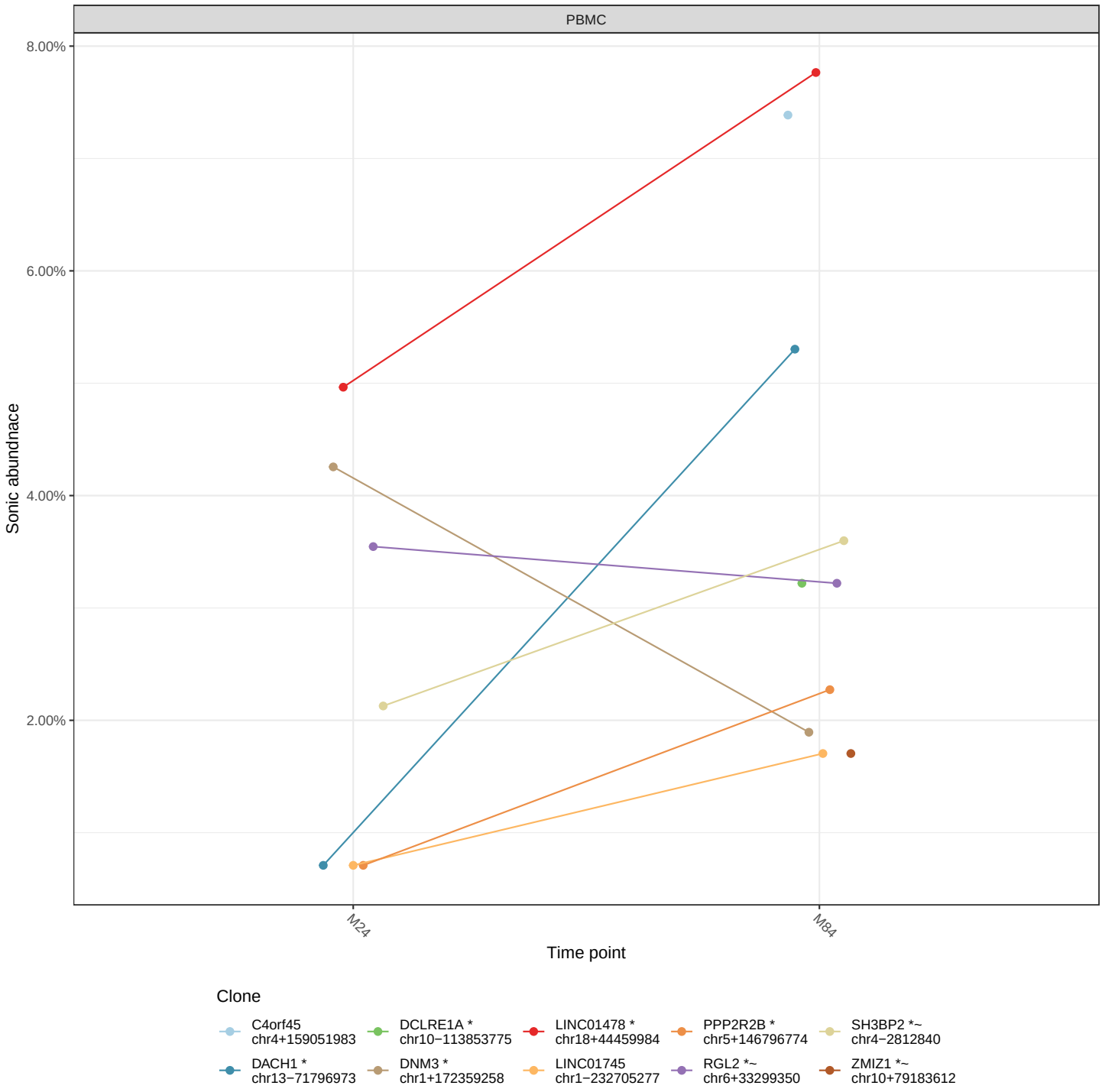
Relative abundance of cell clones

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.



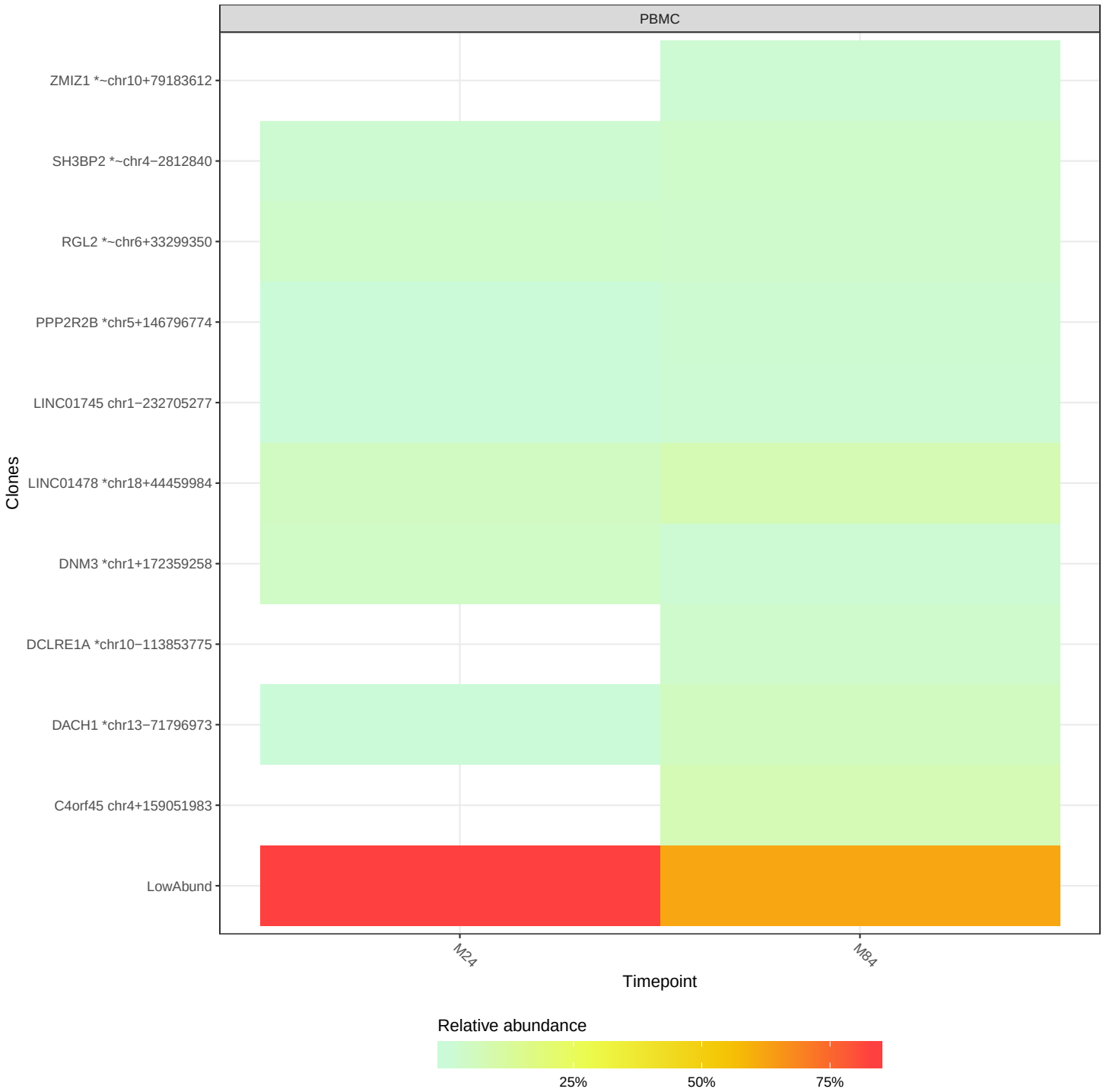
Integration sites near particular genes of interest

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

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

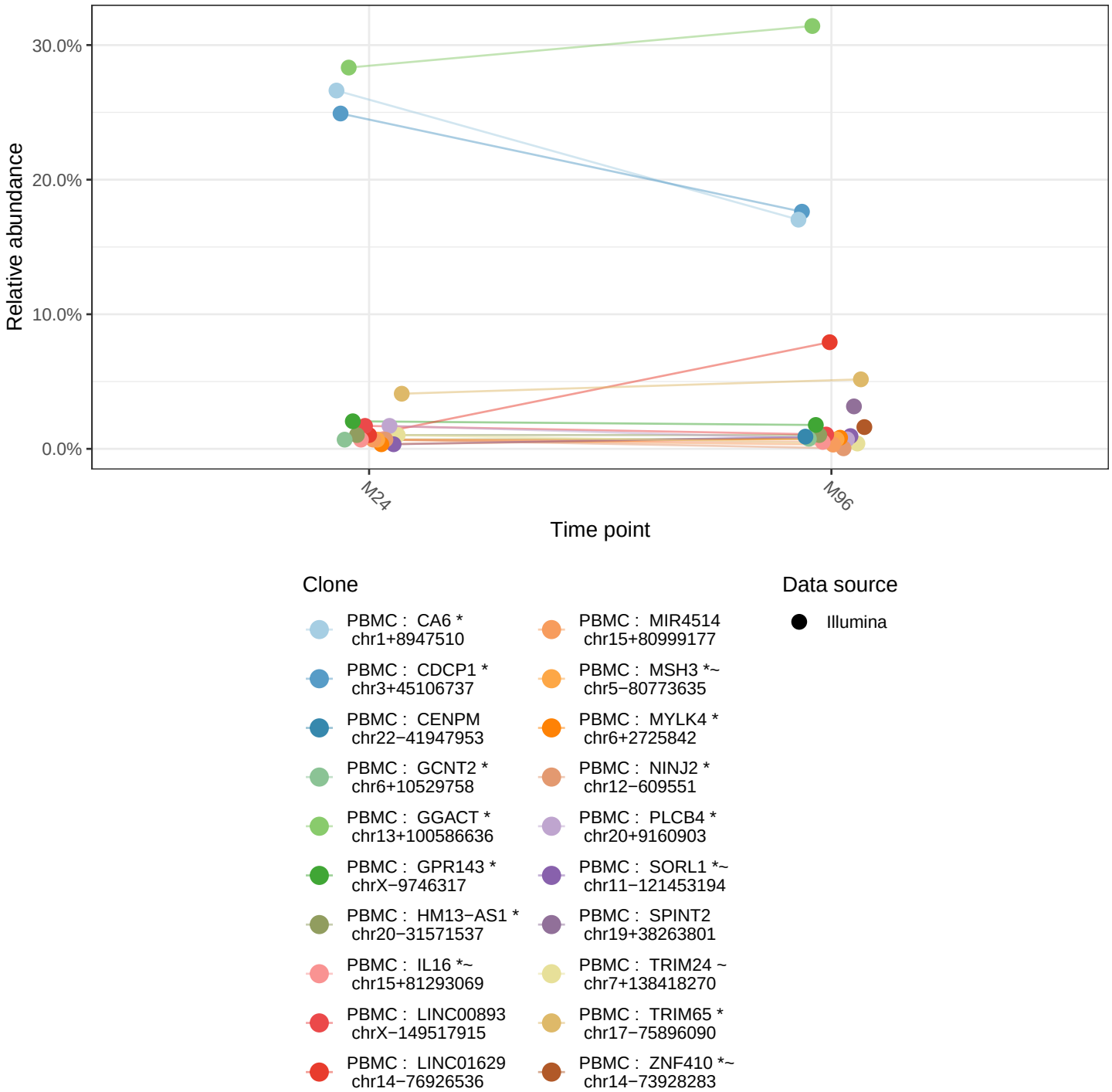
Do any cell clones account for more than 20% of all clones?

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

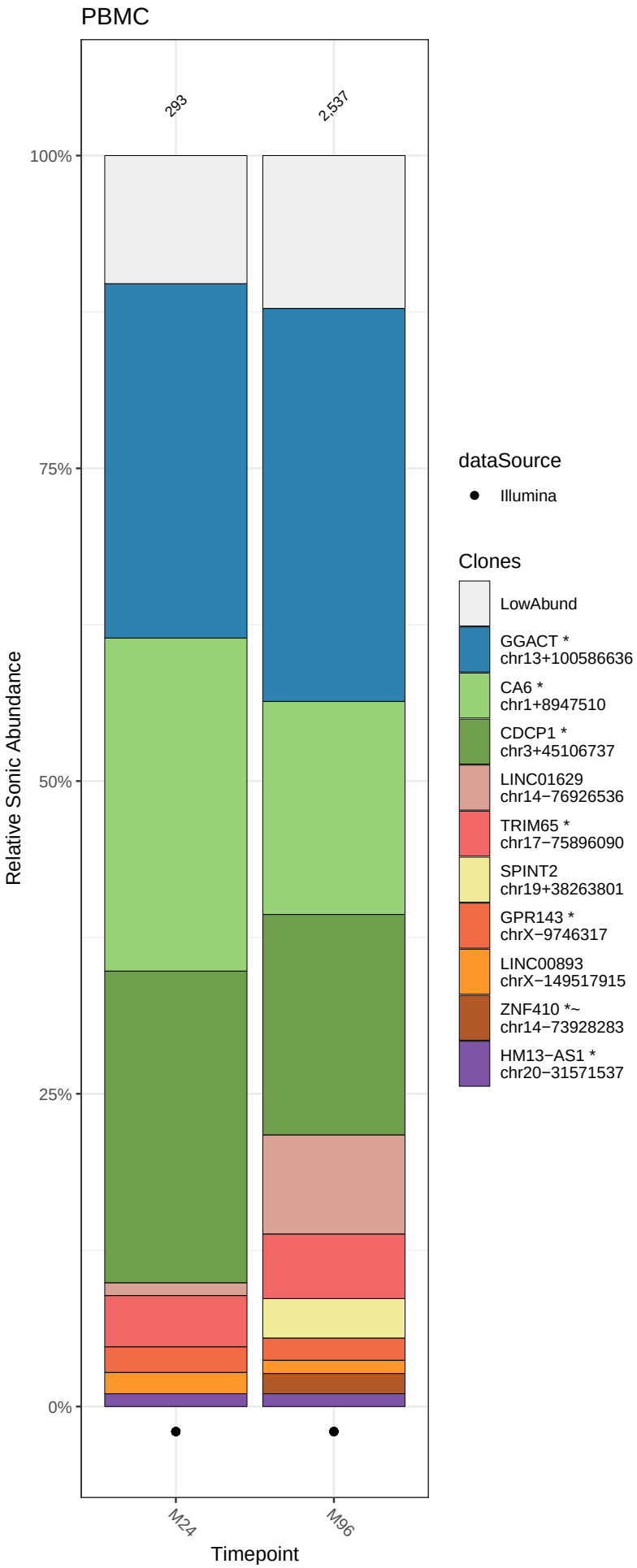
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

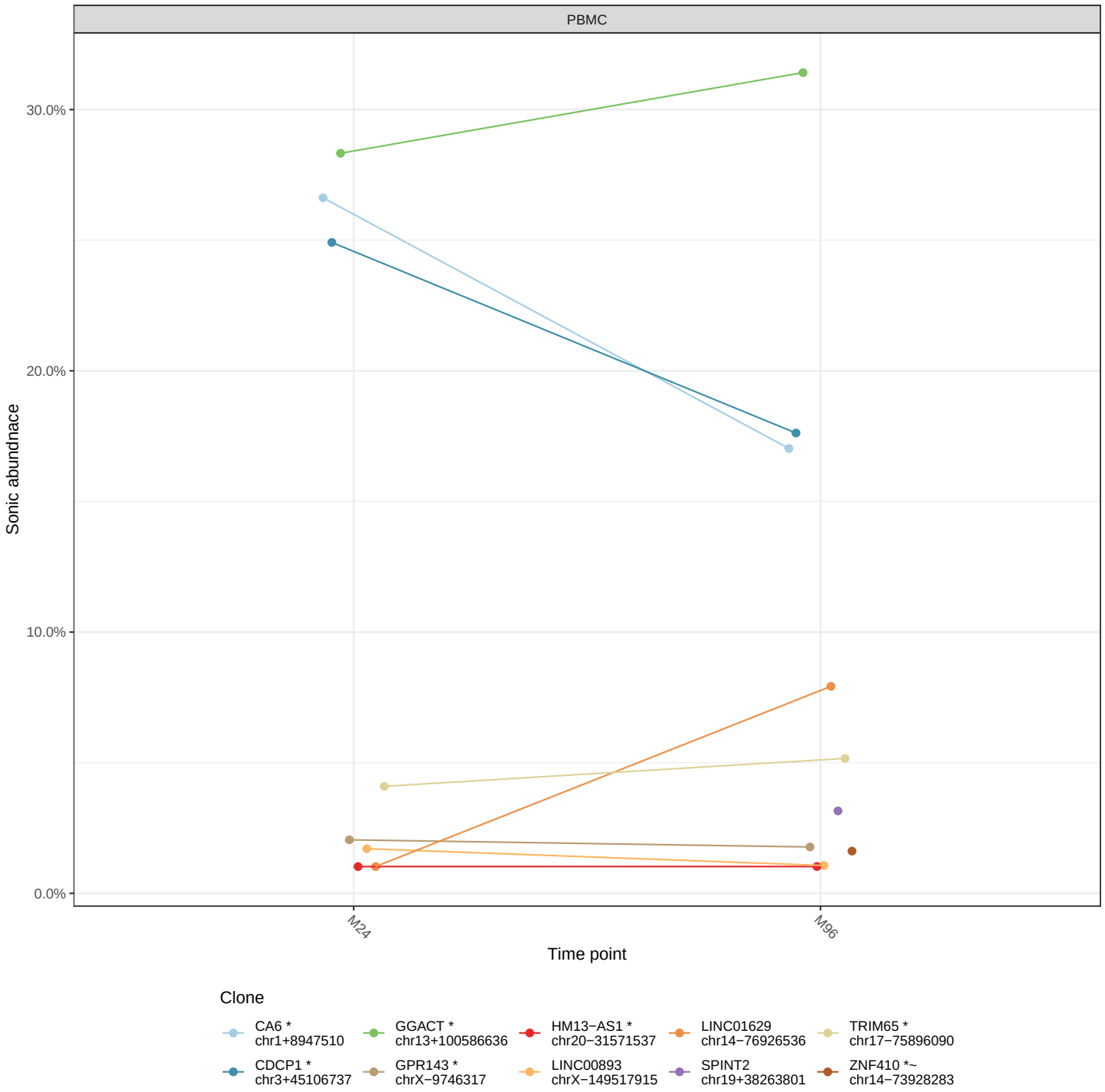
Relative abundance of cell clones

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.



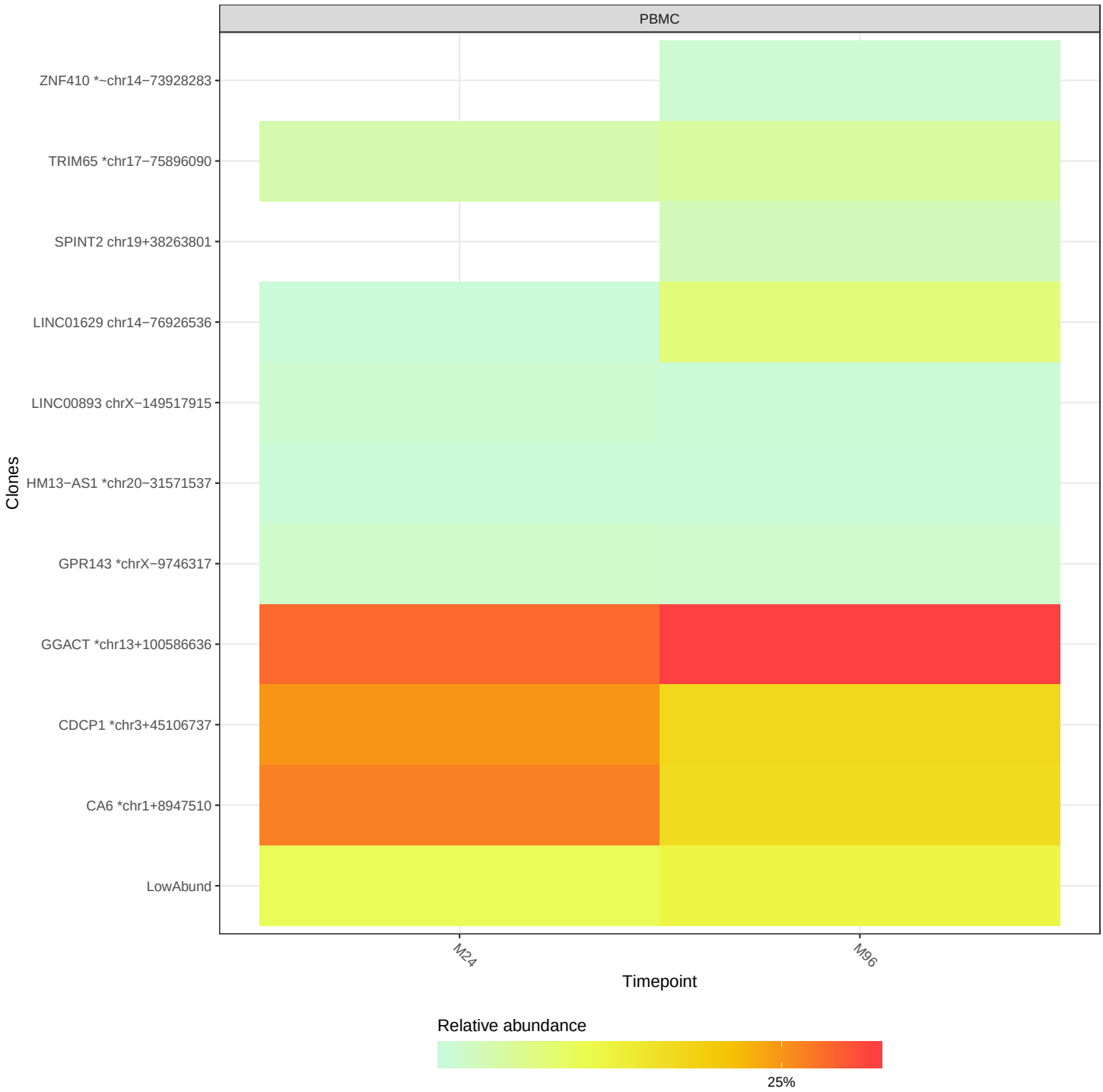
Integration sites near particular genes of interest

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

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

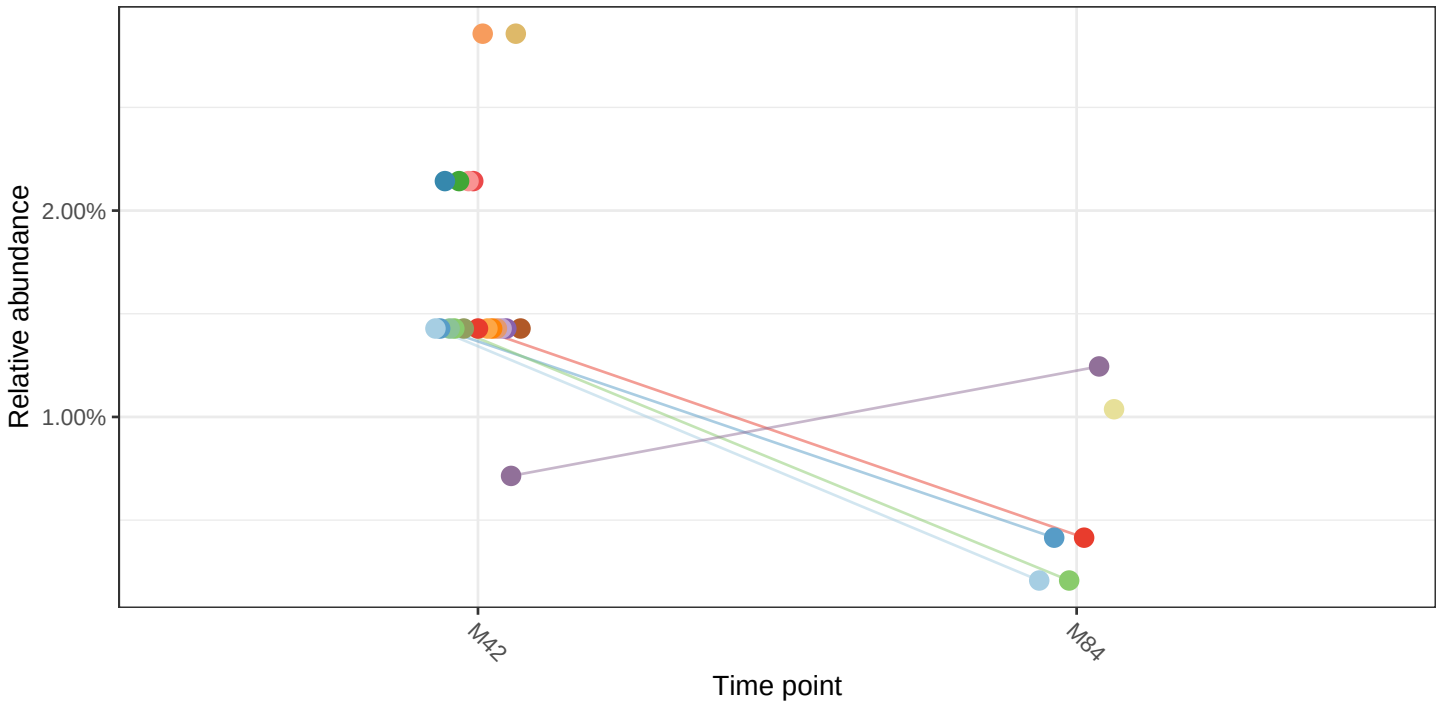
Do any cell clones account for more than 20% of all clones?

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

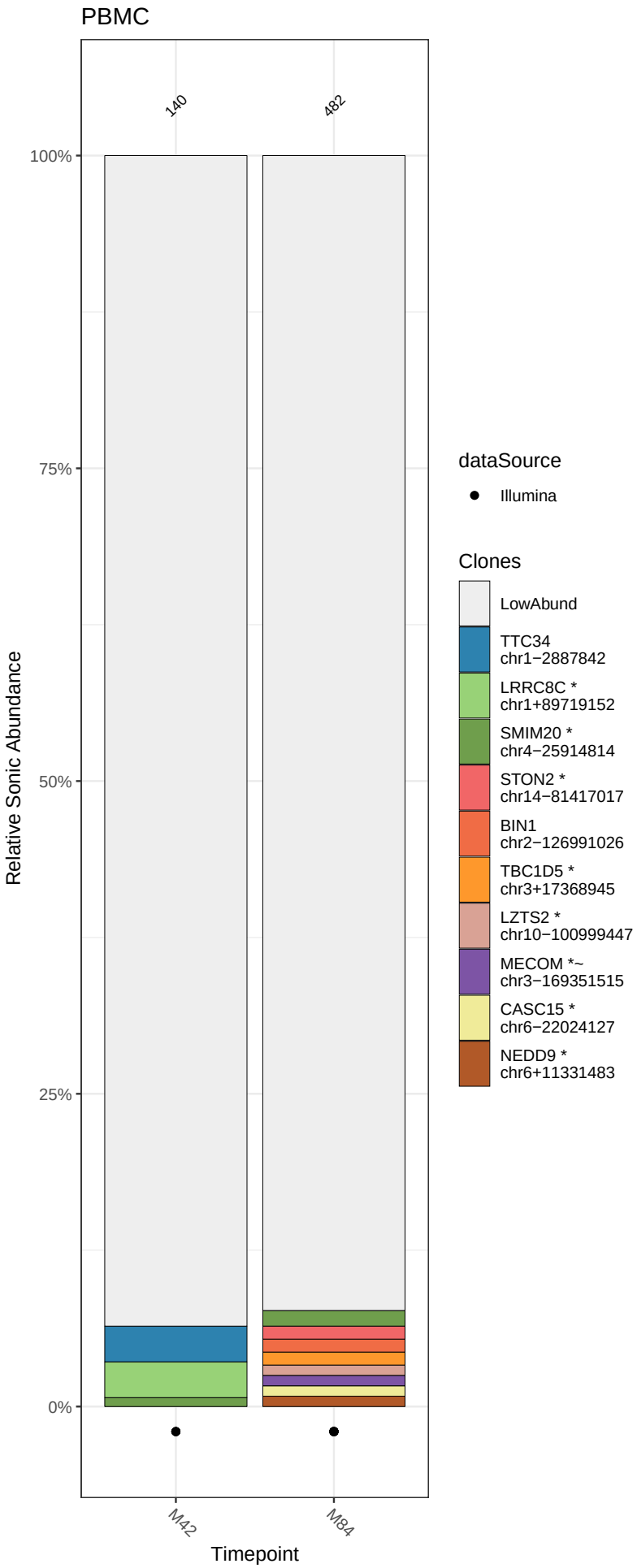
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

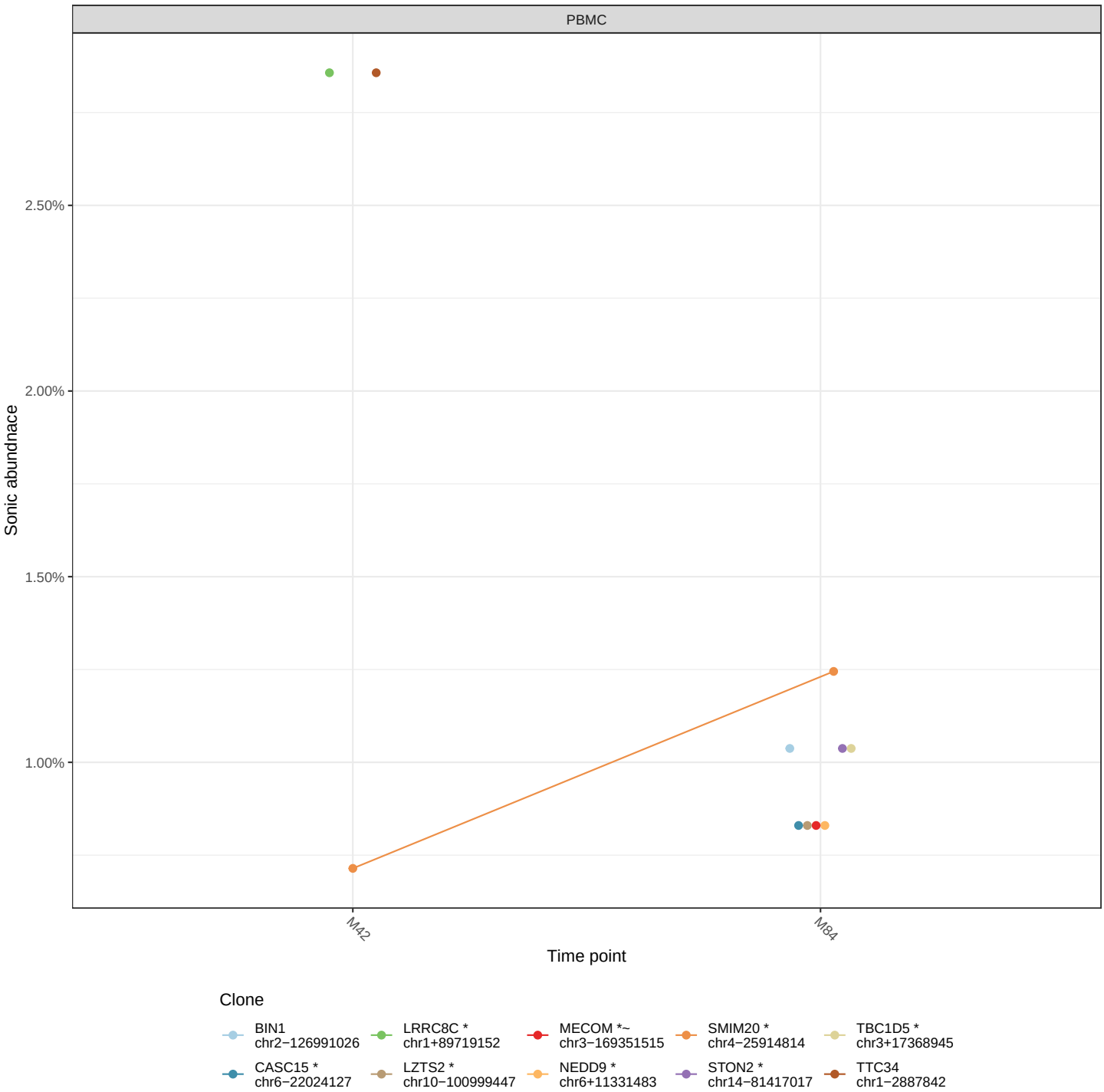
Relative abundance of cell clones

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

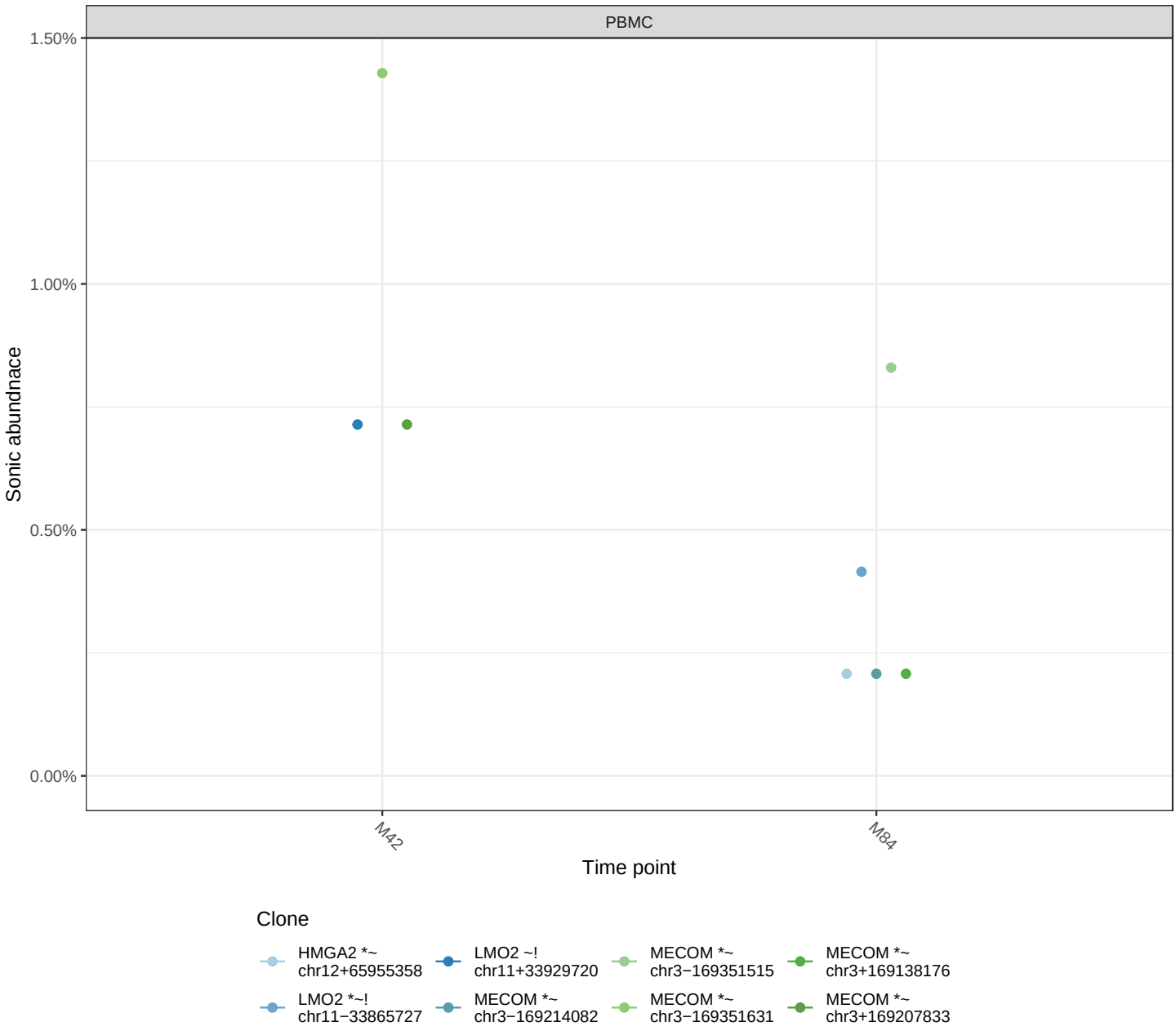
Longitudinal behavior of major clones

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.



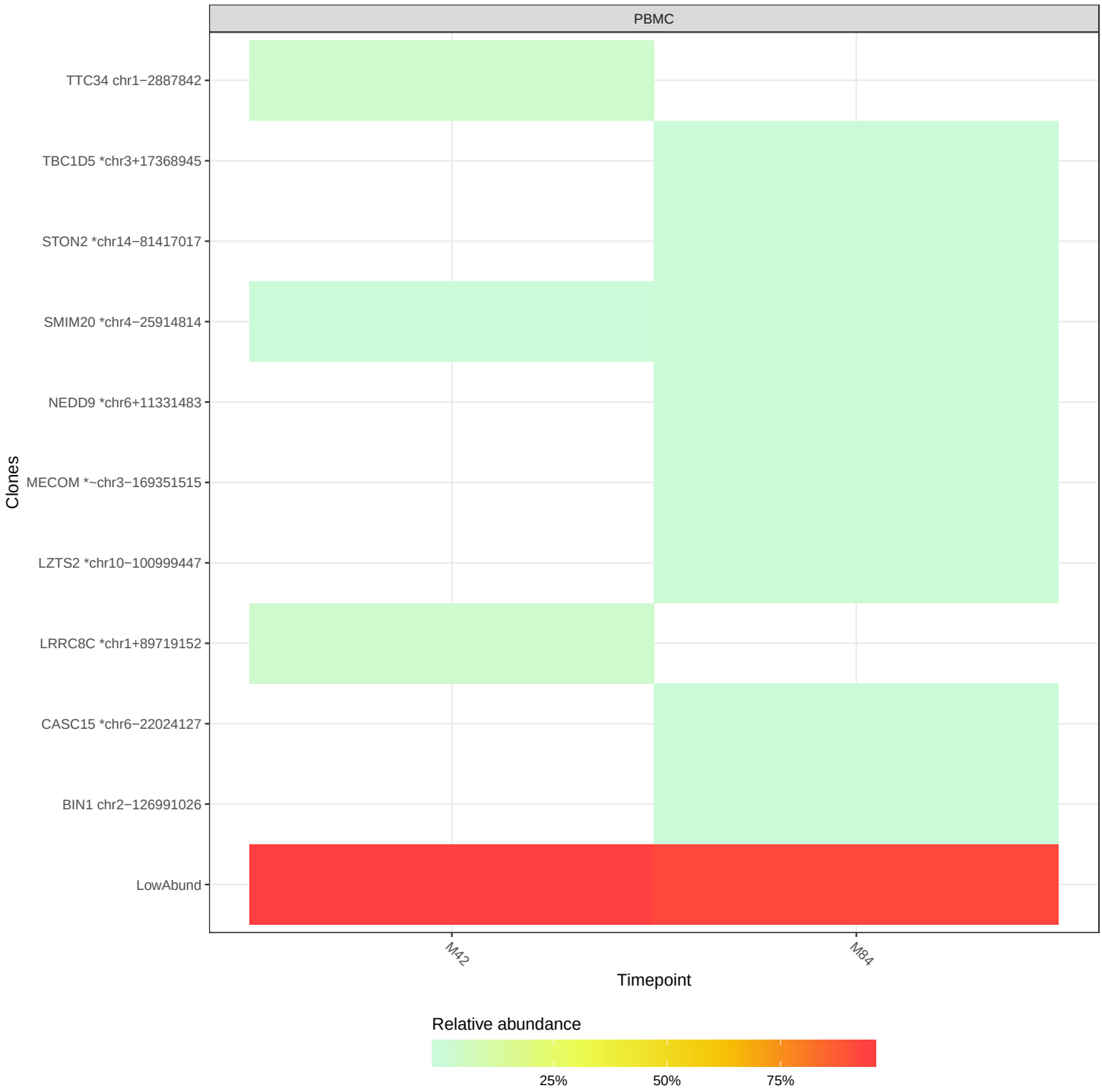
Integration sites near particular genes of interest

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

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

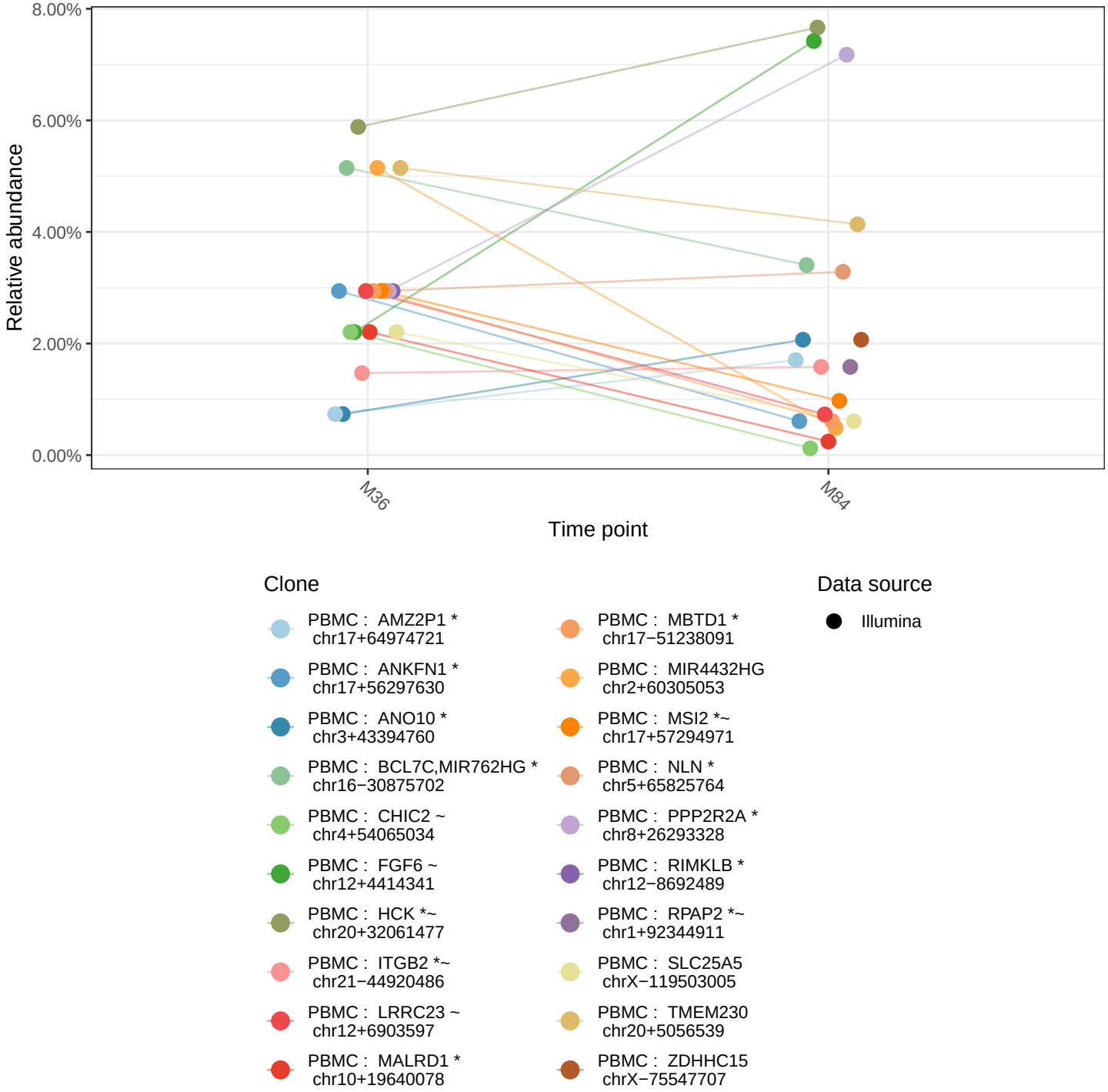
Do any cell clones account for more than 20% of all clones?

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

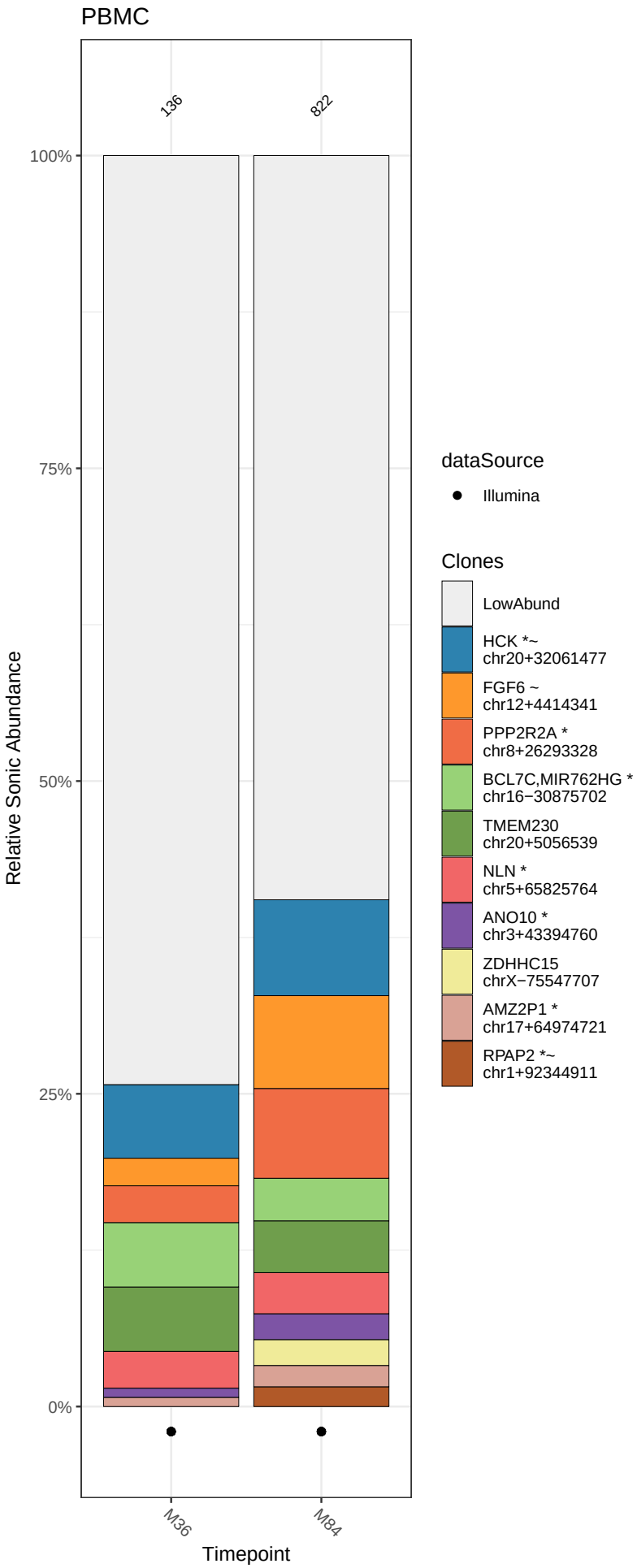
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

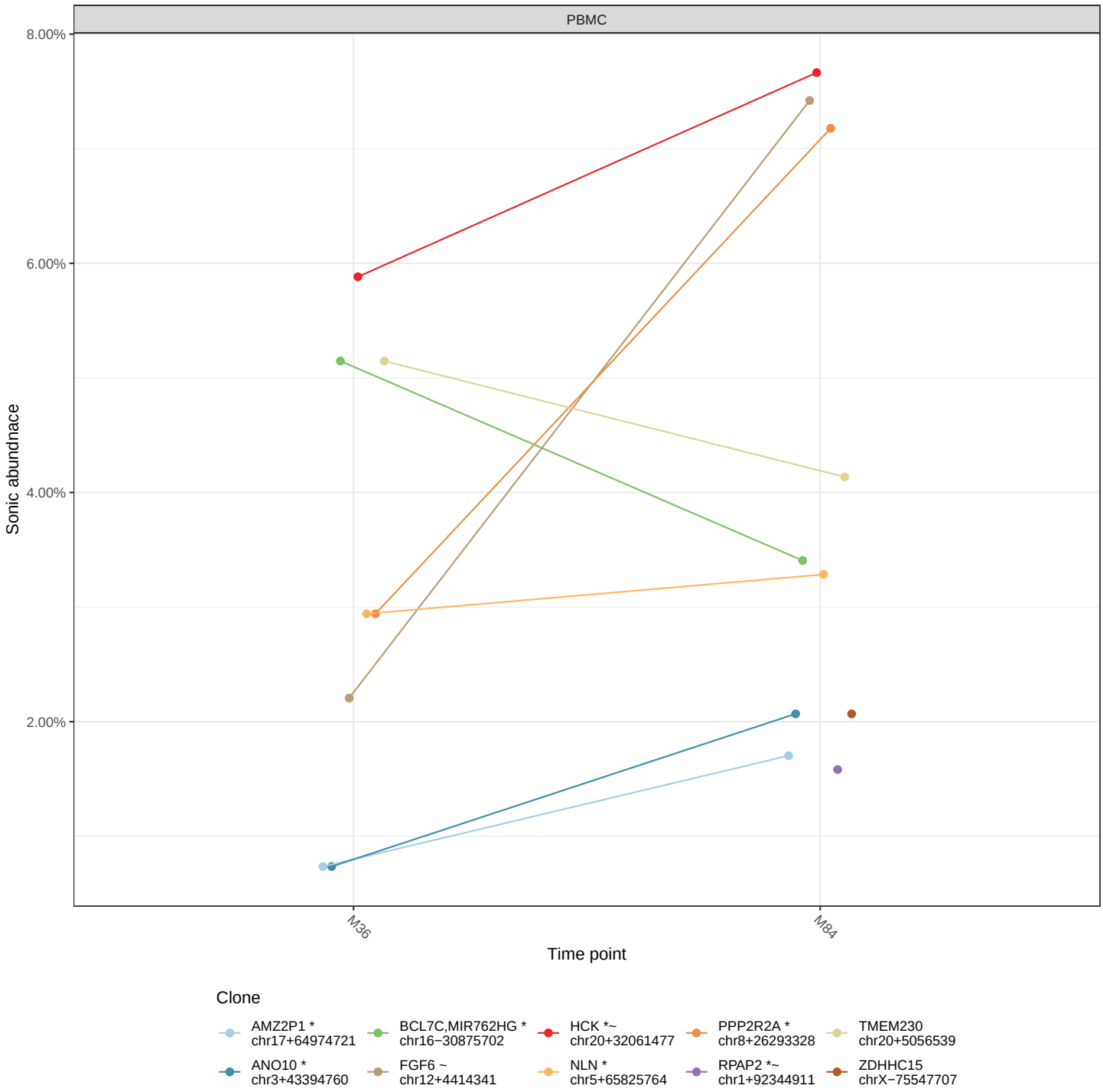
Relative abundance of cell clones

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

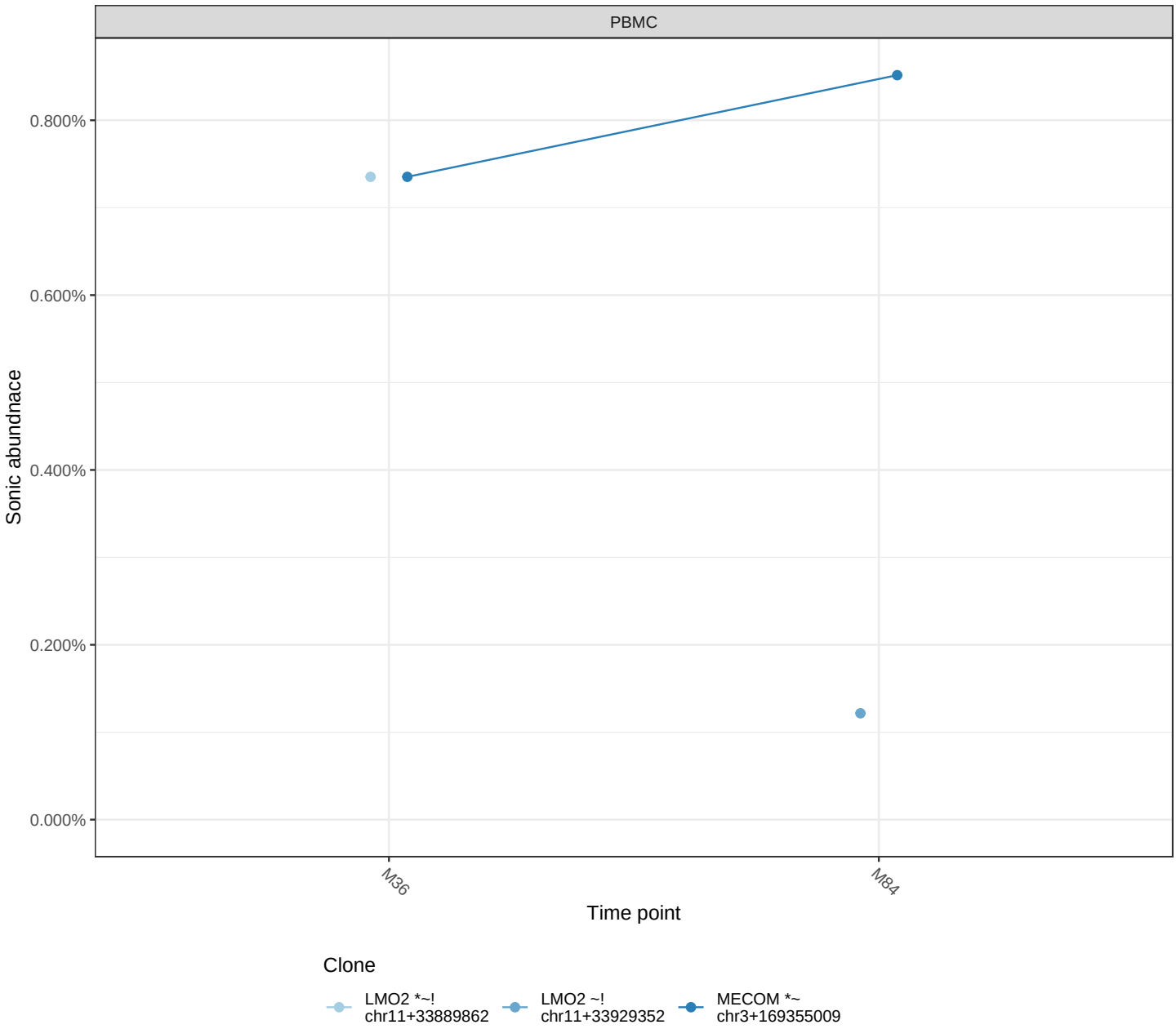
Longitudinal behavior of major clones

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.



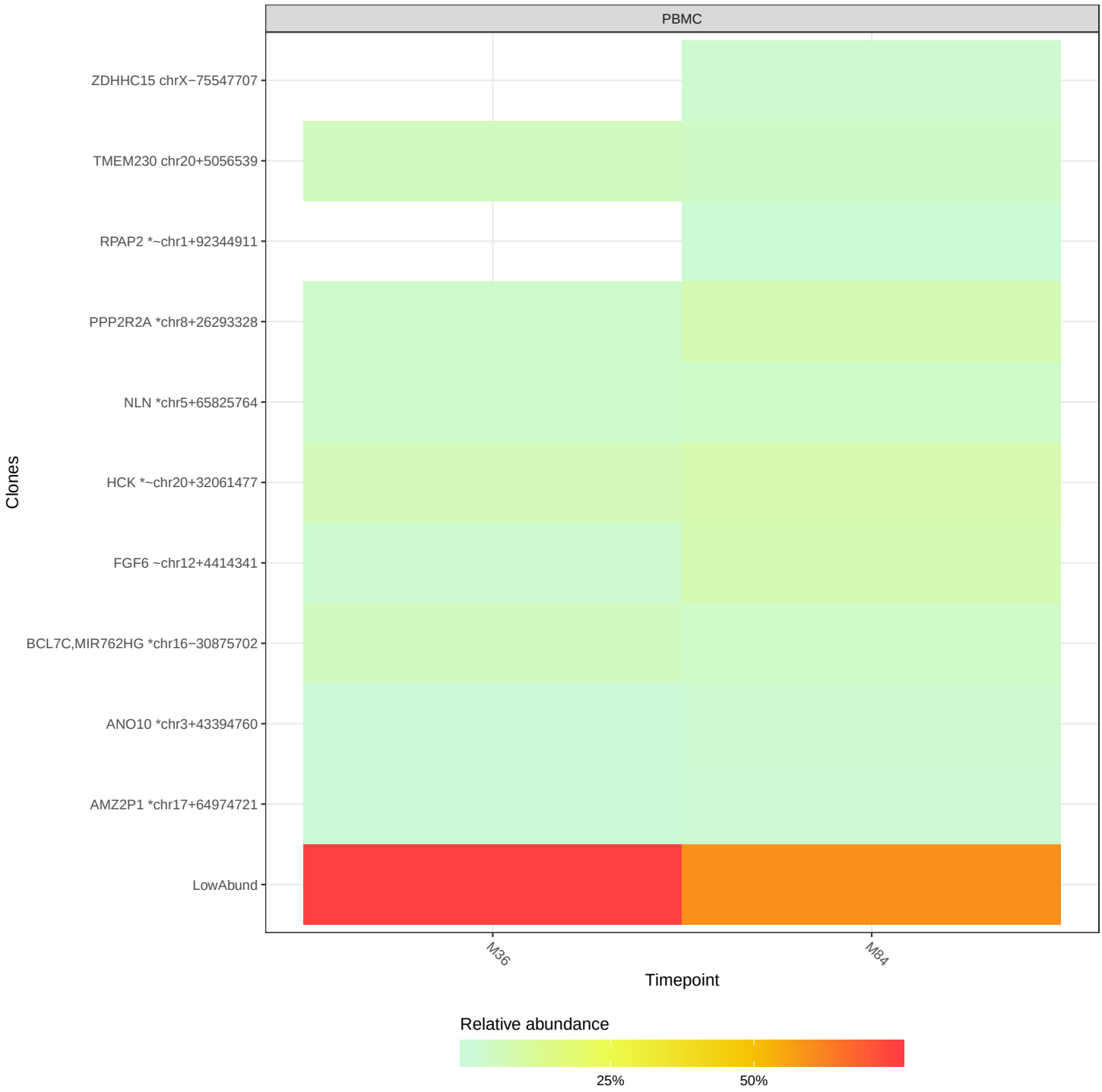
Integration sites near particular genes of interest

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

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

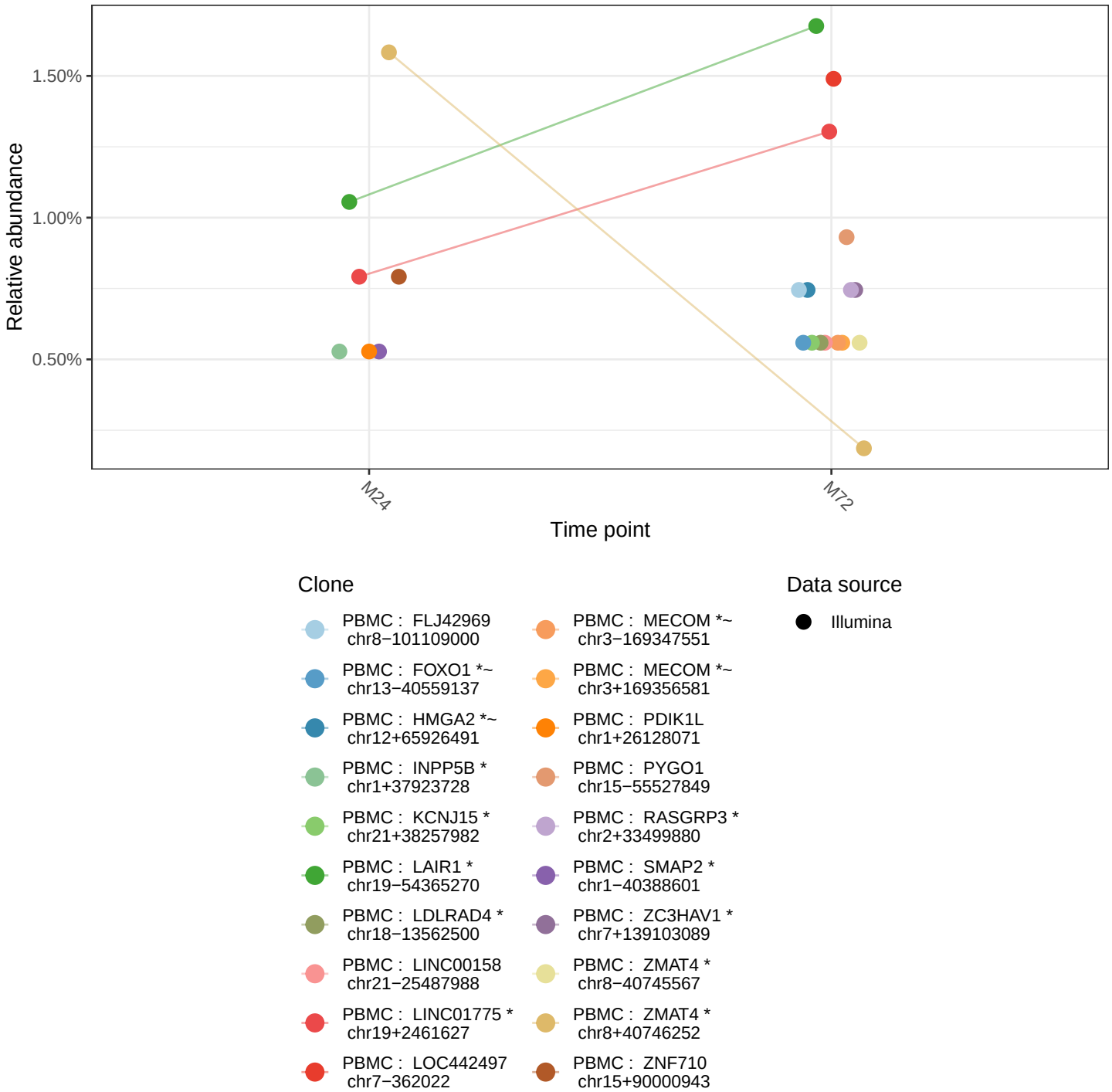
Do any cell clones account for more than 20% of all clones?

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

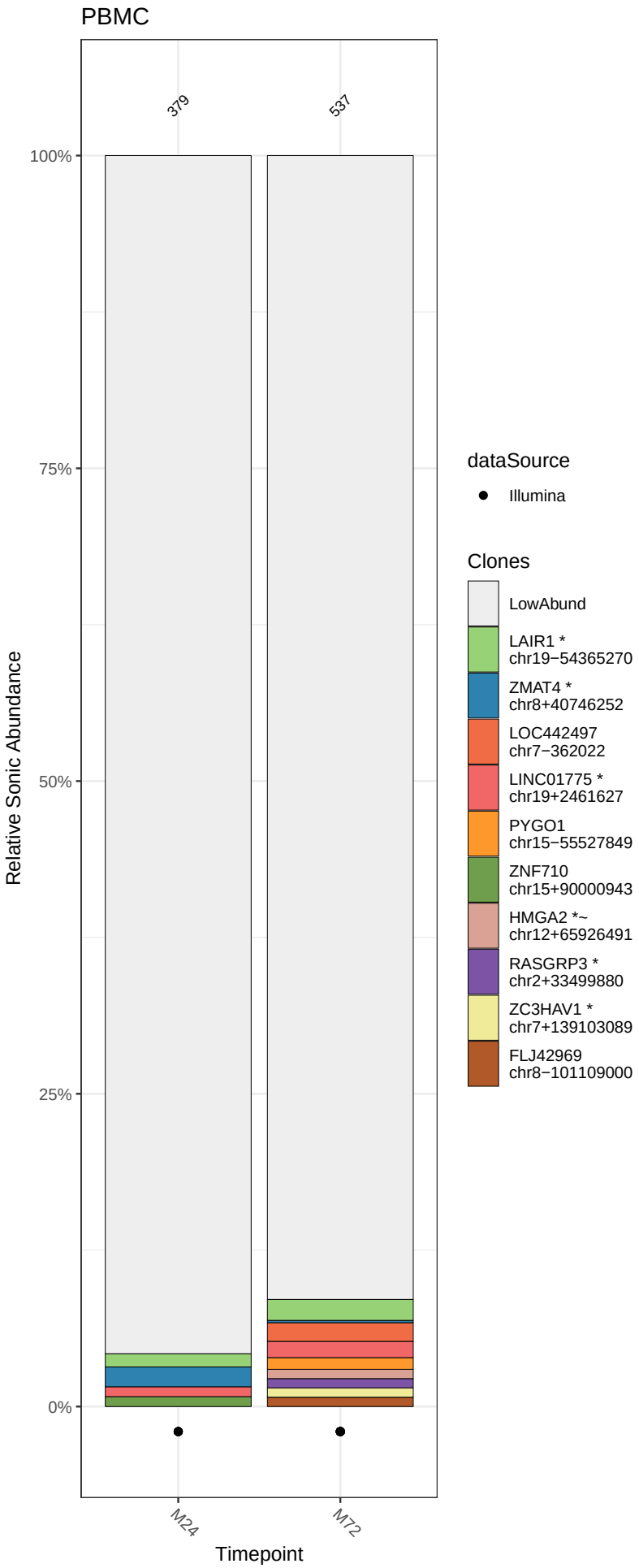
The table below provides population statistics for each analyzed sample. Occasionally multiple samples from the same cell fraction and time point are analyzed where only the sample with greatest number of inferred cells is considered in this report. Sample rows with NA listed in the TotalReads, InferredCells,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

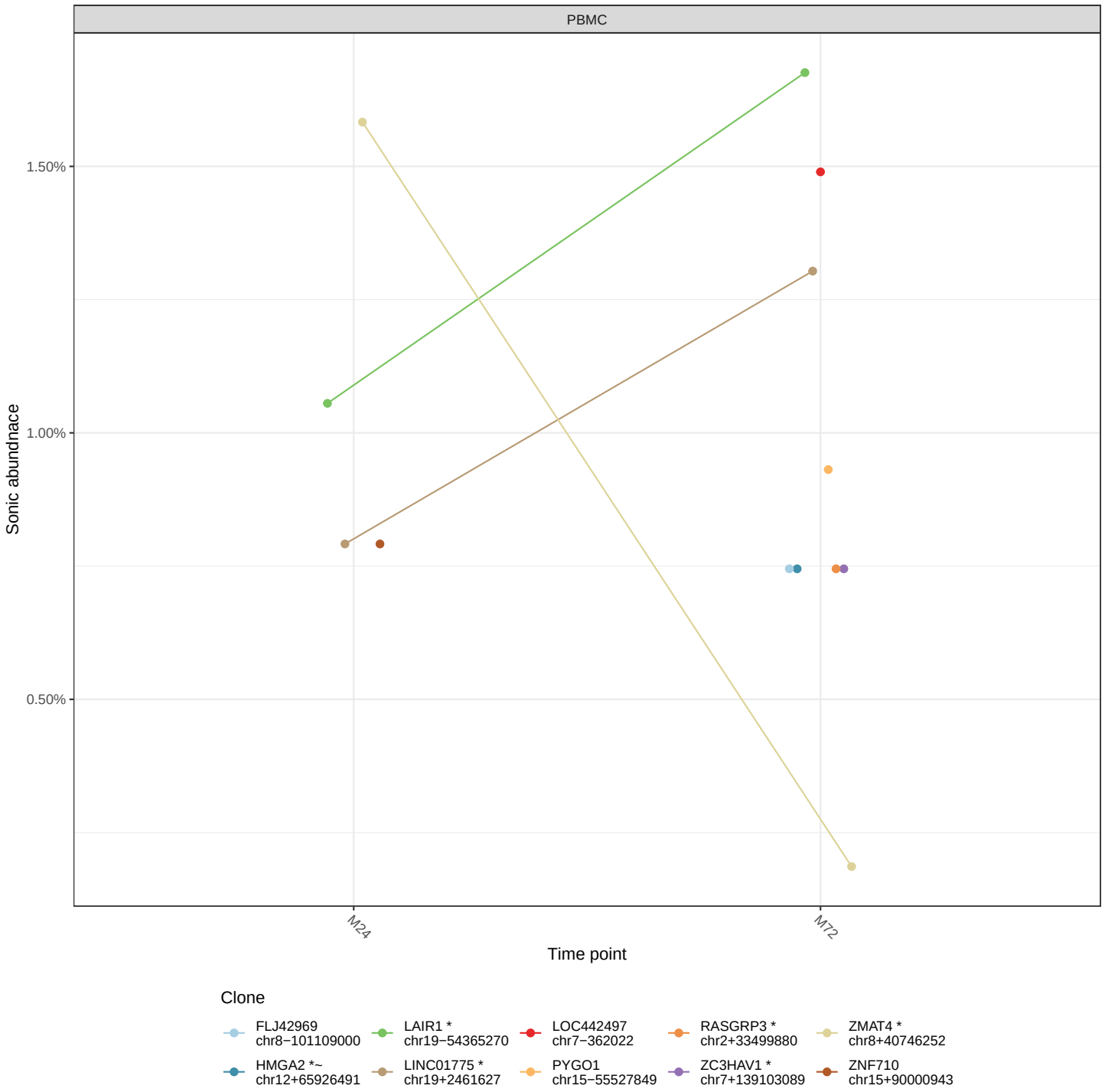
Relative abundance of cell clones

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

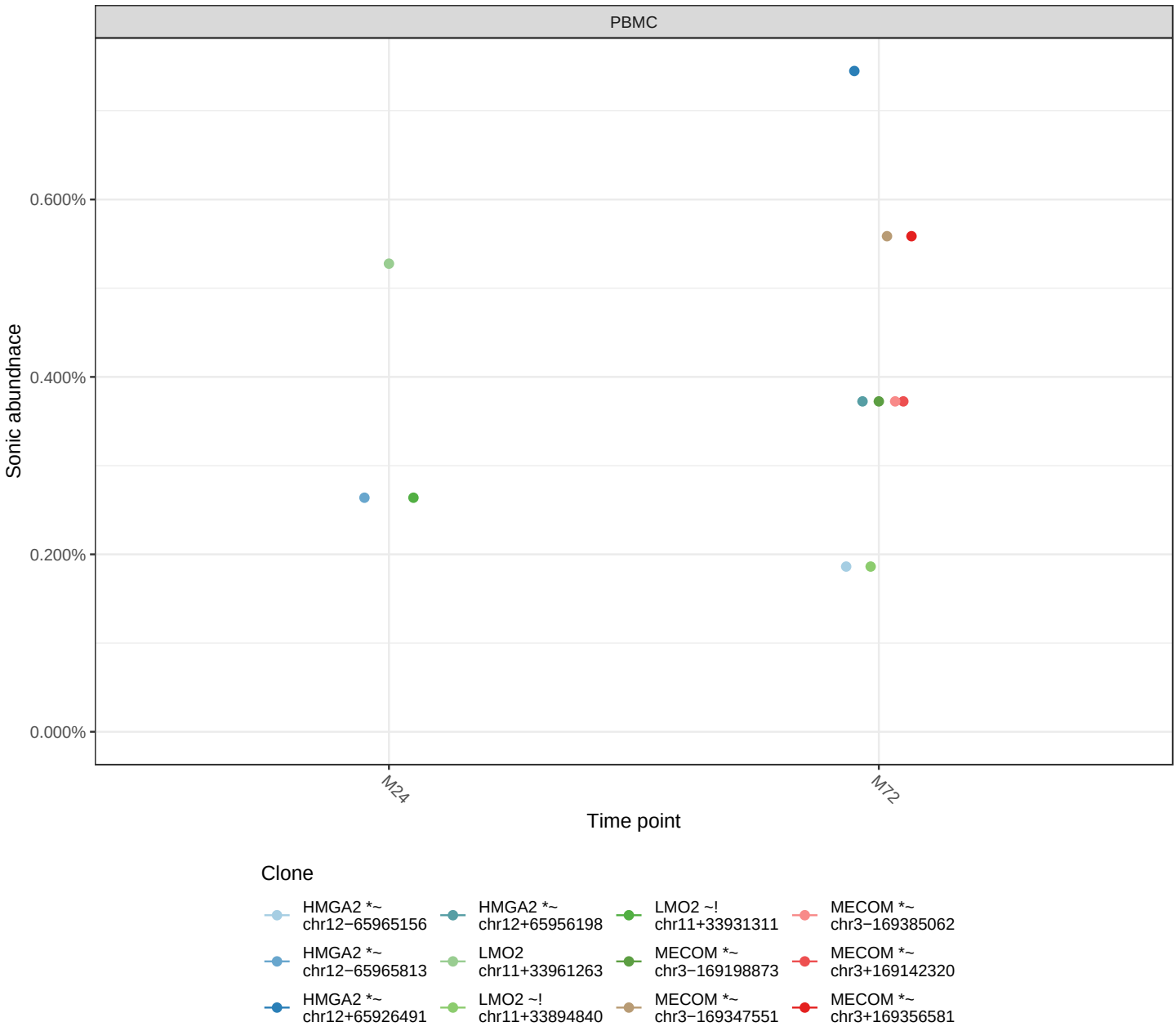
Longitudinal behavior of major clones

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.



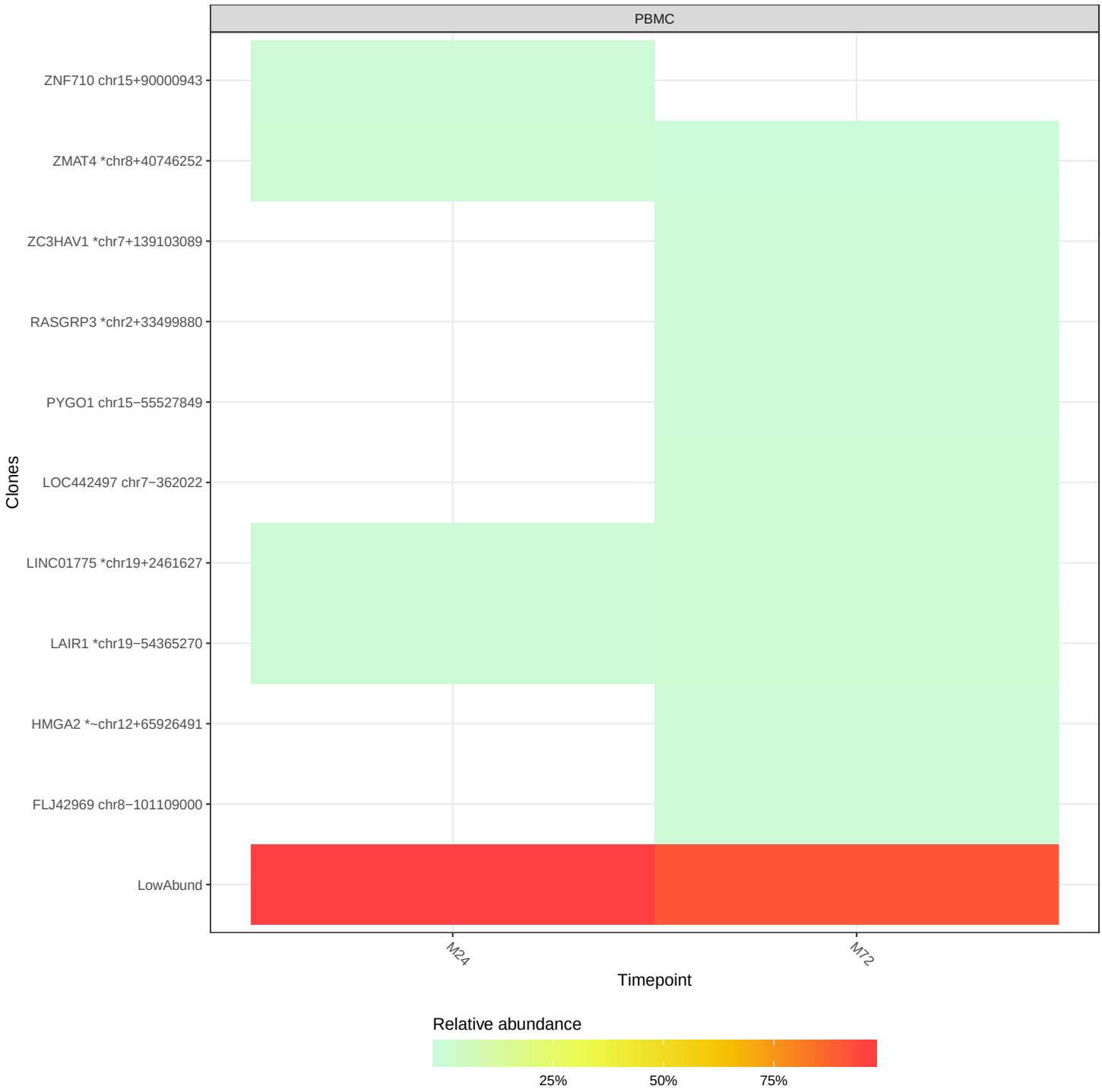
Integration sites near particular genes of interest

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)