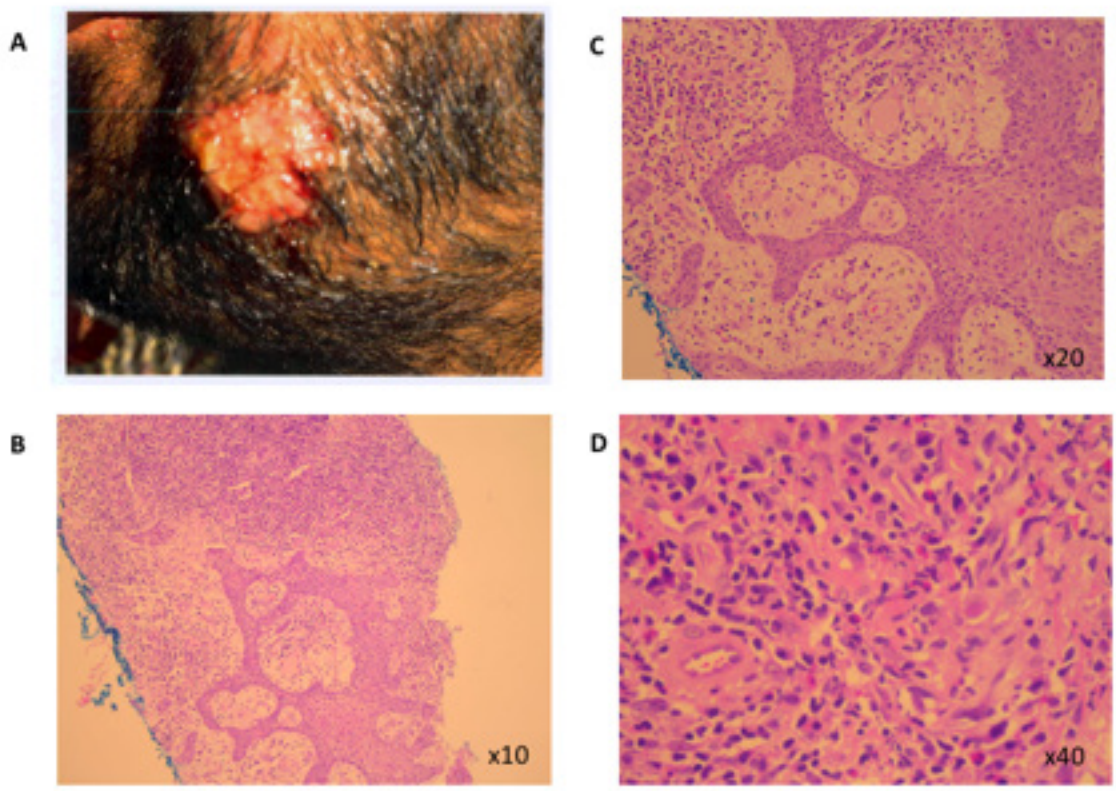


Figure S1. Inflammatory facial ulcer 11 months post gene therapy.

(A) Photograph of lesion; Histological H&E staining (B, C and D).

FIGURE LEGEND

Figure S1. Inflammatory facial ulcer 11 months post gene therapy.

(A) Photograph of lesion; (B and C) H&E histological appearance demonstrating pseudoepitheliomatous hyperplasia at the edge of the ulcer at x10 and x20 magnification, respectively; (D) H&E histological appearance demonstrating a dense inflammatory infiltrate of lymphocytes, plasma cells, eosinophils and neutrophils at x40 magnification.

Analysis of integration site distributions and relative clonal abundance for subject pWAS_UK11

Apr 26 2017

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 pWAS_UK11 over time points m6, m12, m20 in UCSC genome draft hg38.

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.

Sample Summary Table

Trial	GTSP	Replicates	Patient	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	FragMethod	VCN	S.chao1	Gini	Shannon	UC50
WAS	GTSP1277	4	pWAS_UK11	m6	Bcells	1015853	4543	3803	Shearing	NA	20148	0.1443	8.1681	1532
WAS	GTSP1279	4	pWAS_UK11	m6	Monocytes	212871	314	237	Shearing	NA	985	0.1991	5.3693	81
WAS	GTSP1275	4	pWAS_UK11	m6	Neutrophils	764542	1963	1490	Shearing	NA	7231	0.2063	7.1815	509
WAS	GTSP1278	4	pWAS_UK11	m6	NKcells	1088799	2949	2090	Shearing	NA	7757	0.2371	7.5072	616
WAS	GTSP1274	4	pWAS_UK11	m6	PBMC	665356	2061	1357	Shearing	NA	4628	0.2880	6.9892	337
WAS	GTSP1276	4	pWAS_UK11	m6	Tcells	1016780	5570	2996	Shearing	NA	10603	0.4152	7.3942	488
WAS	GTSP1121	4	pWAS_UK11	m12	Bcells	898666	2159	1713	Shearing	NA	7071	0.1757	7.3530	634
WAS	GTSP1123	4	pWAS_UK11	m12	Monocytes	355039	344	233	Shearing	NA	2089	0.2627	5.2909	62
WAS	GTSP1119	4	pWAS_UK11	m12	Neutrophils	541561	707	476	Shearing	NA	1803	0.2653	5.9920	127
WAS	GTSP1122	4	pWAS_UK11	m12	NKcells	498233	688	466	Shearing	NA	1850	0.2575	5.9939	127
WAS	GTSP1118	4	pWAS_UK11	m12	PBMC	456621	2047	1721	Shearing	NA	7899	0.1468	7.3394	698
WAS	GTSP1120	4	pWAS_UK11	m12	Tcells	960593	3050	1999	Shearing	NA	8520	0.2988	7.3013	477
WAS	GTSP1491	4	pWAS_UK11	m20	Bcells	693499	975	562	Shearing	NA	2649	0.3077	6.1524	143
WAS	GTSP1493	4	pWAS_UK11	m20	Monocytes	791066	127	60	Shearing	NA	903	0.4867	3.3602	5
WAS	GTSP1489	4	pWAS_UK11	m20	Neutrophils	764120	1925	1226	Shearing	NA	4433	0.2796	6.9410	333
WAS	GTSP1492	4	pWAS_UK11	m20	NKcells	456570	523	361	Shearing	NA	6729	0.2813	5.6314	100
WAS	GTSP1488	4	pWAS_UK11	m20	PBMC	707262	6460	4335	Shearing	NA	14817	0.2963	7.9689	1106
WAS	GTSP1490	4	pWAS_UK11	m20	Tcells	749603	1165	559	Shearing	NA	1983	0.4150	5.8776	105

Do any uniquely mapped clones account for greater than 20% of the total?

For some trials, a reporting criteria is whether any cell clones expand to account for greater than 20% of all clones. This is summarized below for subject pWAS_UK11. Abundance is estimated using the SonicLength method. Data such as this must, of course, be interpreted in the context of results from other assays. Distances reported refer to transcription start sites (5').

No sites found in this patient which are greater than 20% of the total data.

Do any multihit event account for greater than 20% of the total?

Up until now, all the 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. To make some sense of these multihits, we bin any sequence(s) which share 1 or more genomic locations hence forming psuedo-collections which can be reffered to as OTUs (operation taxonomic units). Once the OTUs are formed, we compare breakpoints of unique sites and multihits. The idea is to see if there are any multihits which higher in abundance than a unique site in a given sample. Below is a table similar to the one shown previously except we show any events instead of genomic locations which might account for greater than 20% of all clones in the data.

No multihits sites found in this patient which are greater than 20% of the total data.

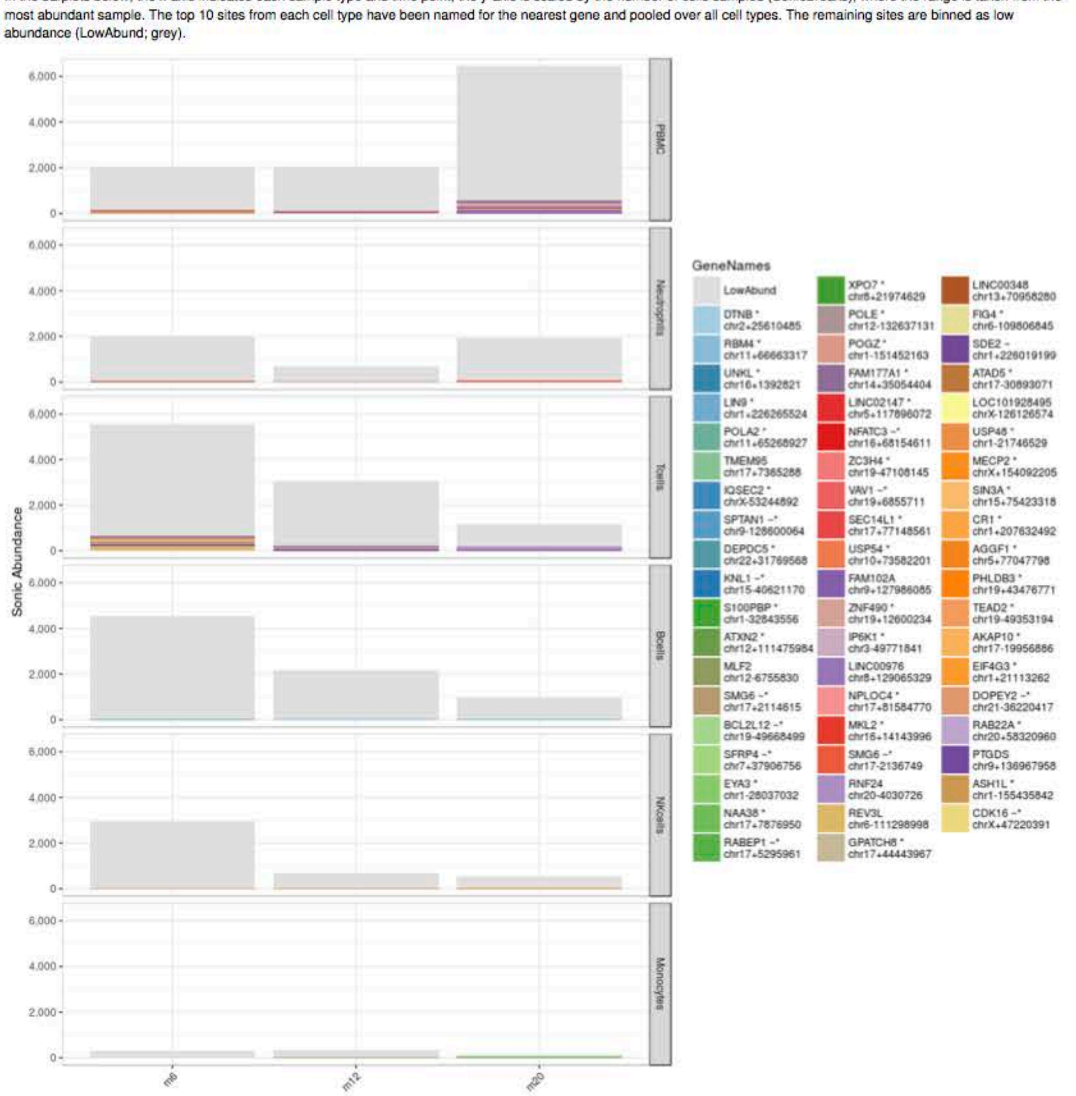
Relative abundance of cell clones

The relative abundance of cell clones is summarized in the attached stacked bar graphs. The cell fraction studied is named at the top, the time points are marked at the bottom. The different bars in each panel show the major cell clones, as marked by integration sites. A key to the sites is shown at the right. Throughout the whole report, each integration site is assigned a GeneName given by:

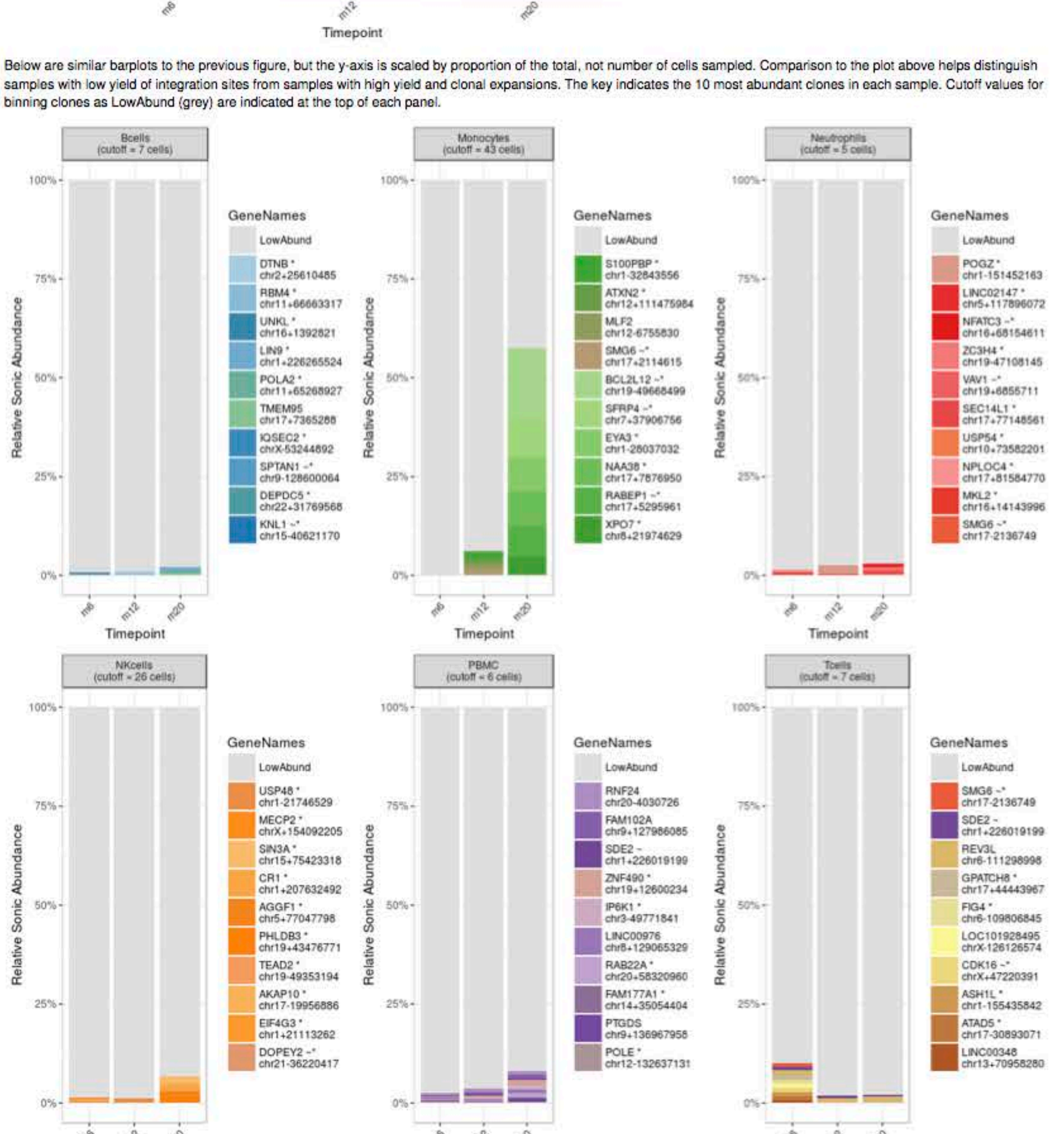
```
- GeneName refers to the closest gene to either end and strand,
- * indicates the site is within a transcription unit,
- ~ indicates the site is within 50kb of a cancer related gene,
- ! indicates the gene was assocaited with lymphoma in humans.
```

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.

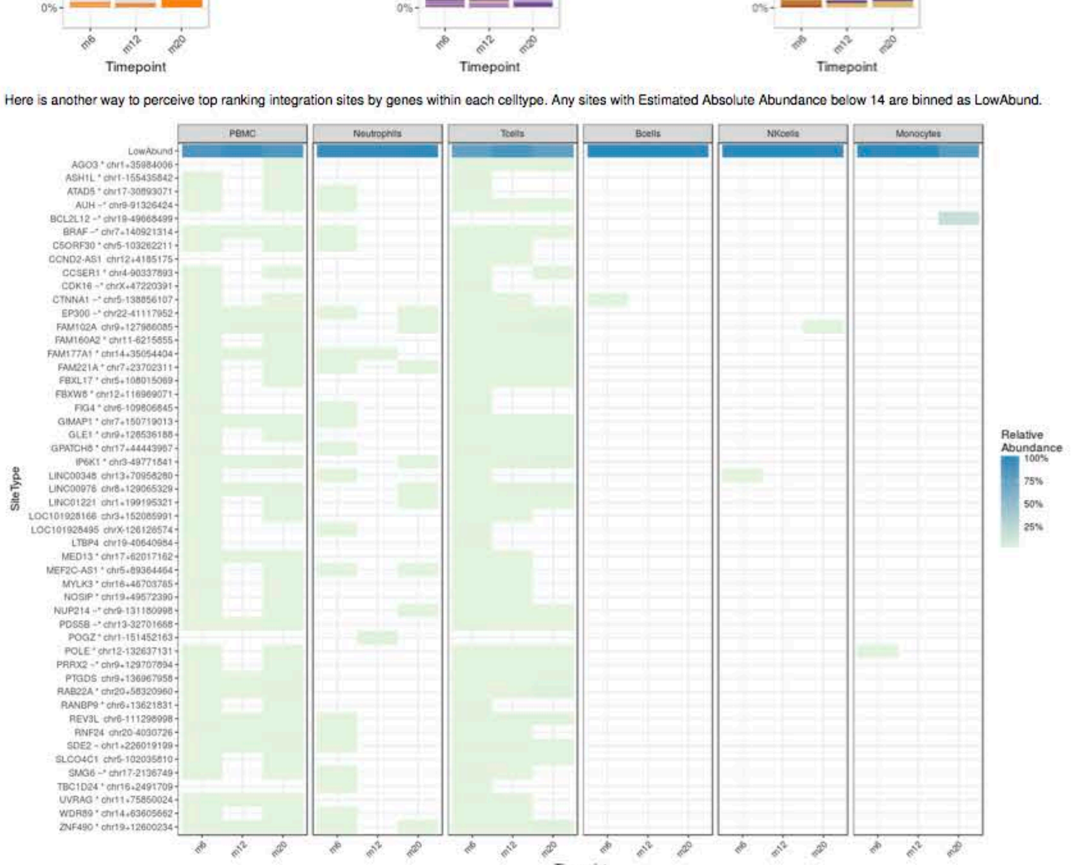
In the barplots below, the x-axis indicates each sample type and time point, the y-axis is scaled by the number of cells sampled (SonicBreaks), where the range is taken from the most abundant sample. The top 10 sites from each cell type have been named for the nearest gene and pooled over all cell types. The remaining sites are binned as low abundance (LowAbund; grey).



Below are similar barplots to the previous figure, but the y-axis is scaled by proportion of the total, not number of cells sampled. Comparison to the plot above helps distinguish samples with low yield of integration sites from samples with high yield and clonal expansions. The key indicates the 10 most abundant clones in each sample. Cutoff values for binning clones as LowAbund (grey) are indicated at the top of each panel.

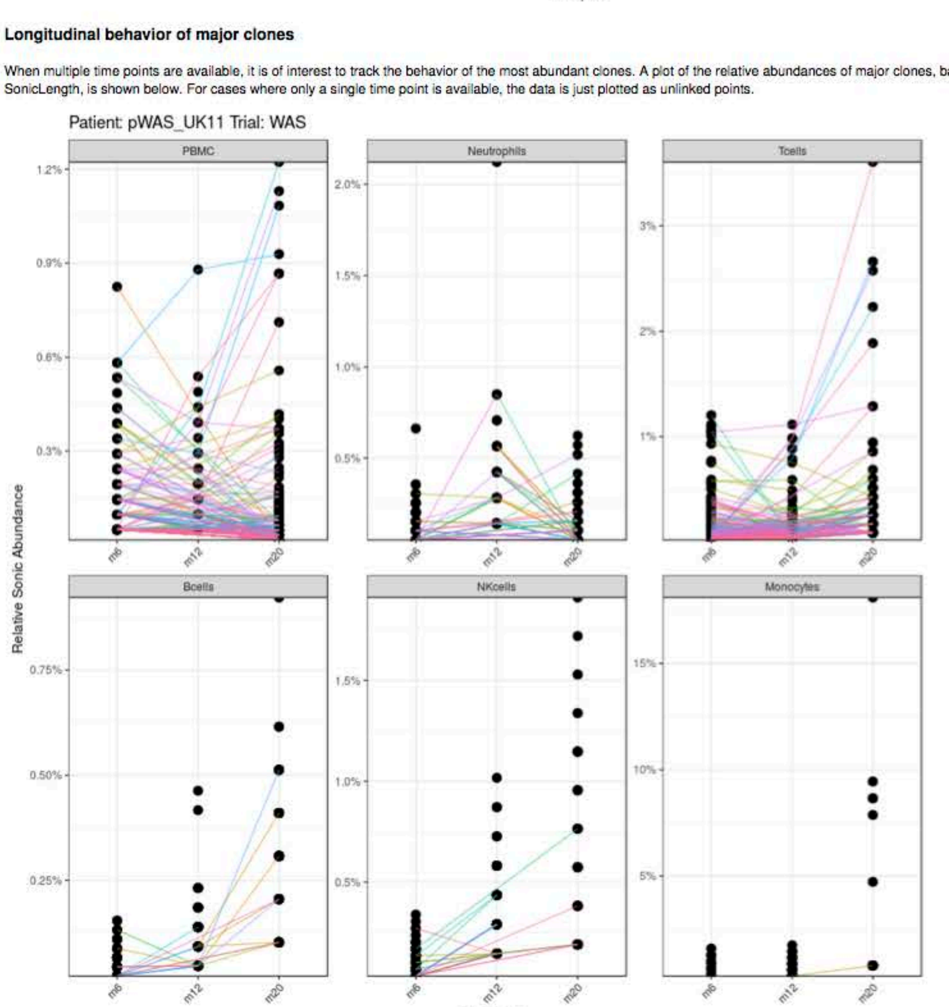


Here is another way to perceive top ranking integration sites by genes within each celltype. Any sites with Estimated Absolute Abundance below 14 are binned as LowAbund.



Longitudinal behavior of major clones

When multiple time points are available, it is of interest to track the behavior of the most abundant clones. A plot of the relative abundances of major clones, based on output from SonicLength, is shown below. For cases where only a single time point is available, the data is just plotted as unlinked points.

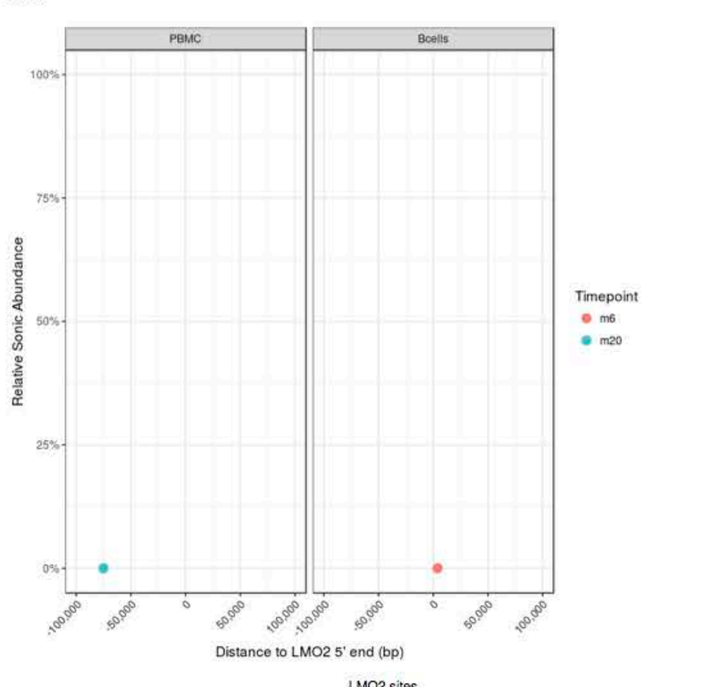


Integration sites near particular genes of interest

Integration sites near genes that have been associated with adverse events are of particular interest. Thus, we have cataloged all integration sites for which a gene of interest is the nearest cancer-related gene. Results are summarized below as a scatter plot where the y-axis shows relative abundance of sites and x-axis is distance to the nearest onconogene 5' end.

Negative distances indicate that the integration site is downstream from (i.e. after) the TSS. Positive distances indicate that the integration site is upstream from (i.e. before) the TSS. Note that all RefSeq splicing isoforms are used for this analysis, so the reference TSS may not be the same for each listed integration site.

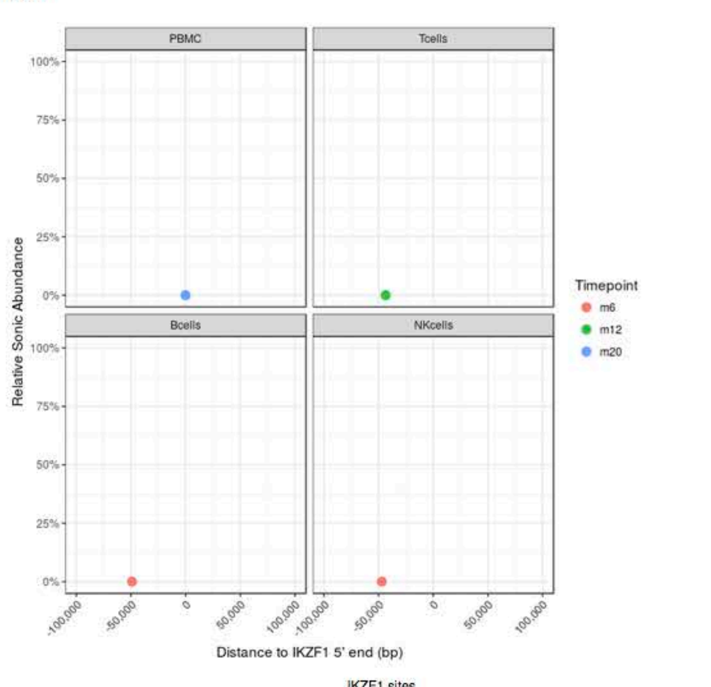
LMO2



LMO2 sites

Timepoint	CellType	Integration Site	nearestOnco	distToNearestOnco	sonicAbundance	sonicAbundanceRank
m6	Bcells	chr11+33888327	LMO2	3962	1	3803
m20	PBMC	chr11+33967325	LMO2	-75036	1	4335

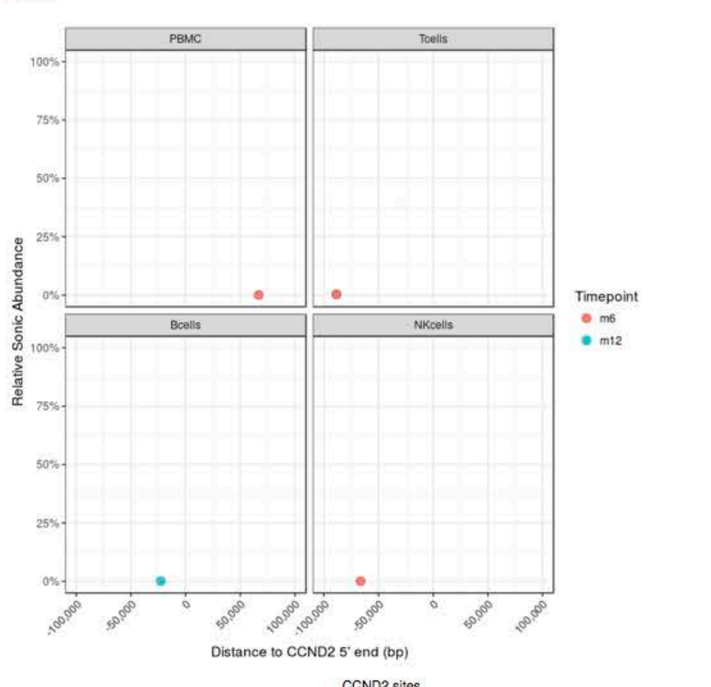
IKZF1



IKZF1 sites

Timepoint	CellType	Integration Site	nearestOnco	distToNearestOnco	sonicAbundance	sonicAbundanceRank
m12	Tcells	chr7+50260618	IKZF1	-43464	1	1999
m6	Bcells	chr7+50255087	IKZF1	-48995	2	549
m6	NKcells	chr7+50256966	IKZF1	-47116	1	2090
m20	PBMC	chr7+50319109	IKZF1	62	1	4335

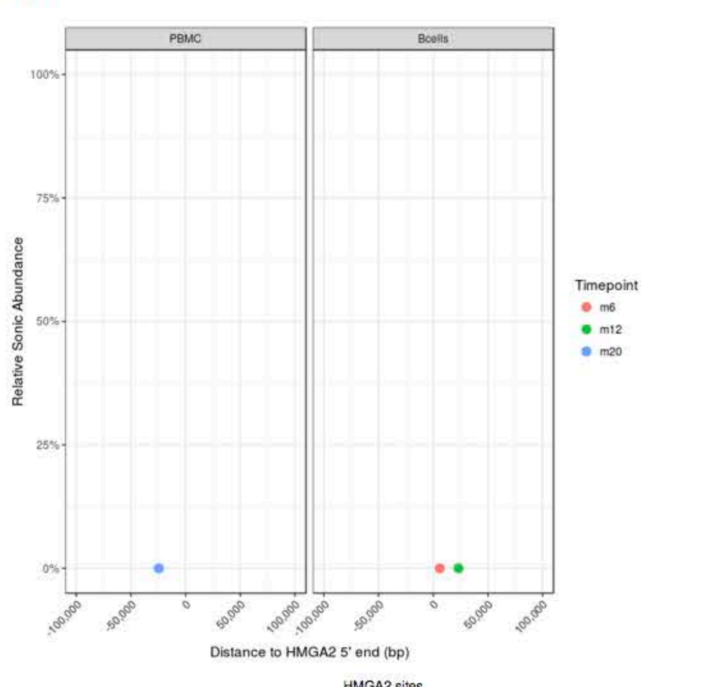
CCND2



CCND2 sites

Timepoint	CellType	Integration Site	nearestOnco	distToNearestOnco	sonicAbundance	sonicAbundanceRank
m12	Tcells	chr12+4185175	CCND2	-88560	10	14
m12	Bcells	chr12-4251002	CCND2	-22733	1	1713
m6	PBMC	chr12-4340473	CCND2	66738	1	1357
m6	Tcells	chr12+4185175	CCND2	-88560	17	29
m6	NKcells	chr12+4207281	CCND2	-66454	1	2090

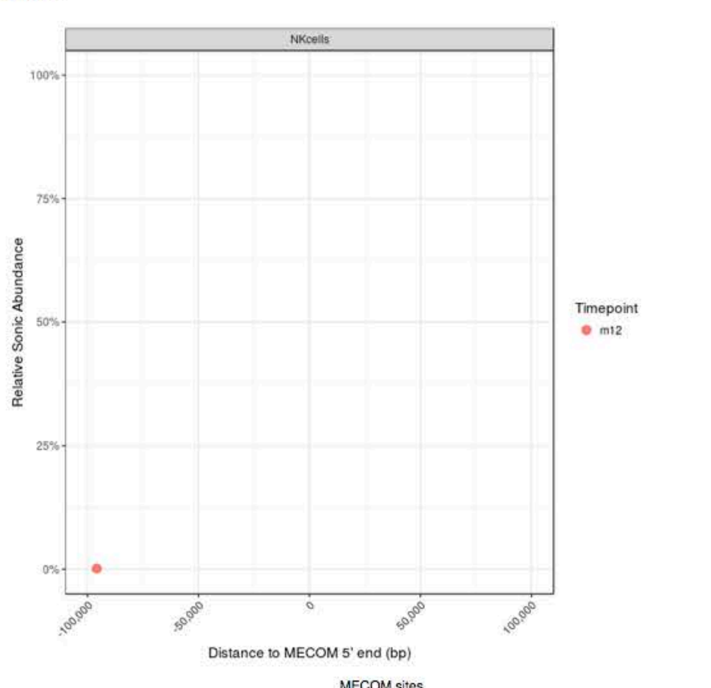
HMGA2



HMGA2 sites

Timepoint	CellType	Integration Site	nearestOnco	distToNearestOnco	sonicAbundance	sonicAbundanceRank
m12	Bcells	chr12-65847548	HMGA2	23089	1	1713
m6	Bcells	chr12+65830416	HMGA2	5957	1	3803
m20	PBMC	chr12-65799922	HMGA2	-24537	2	897

MECOM

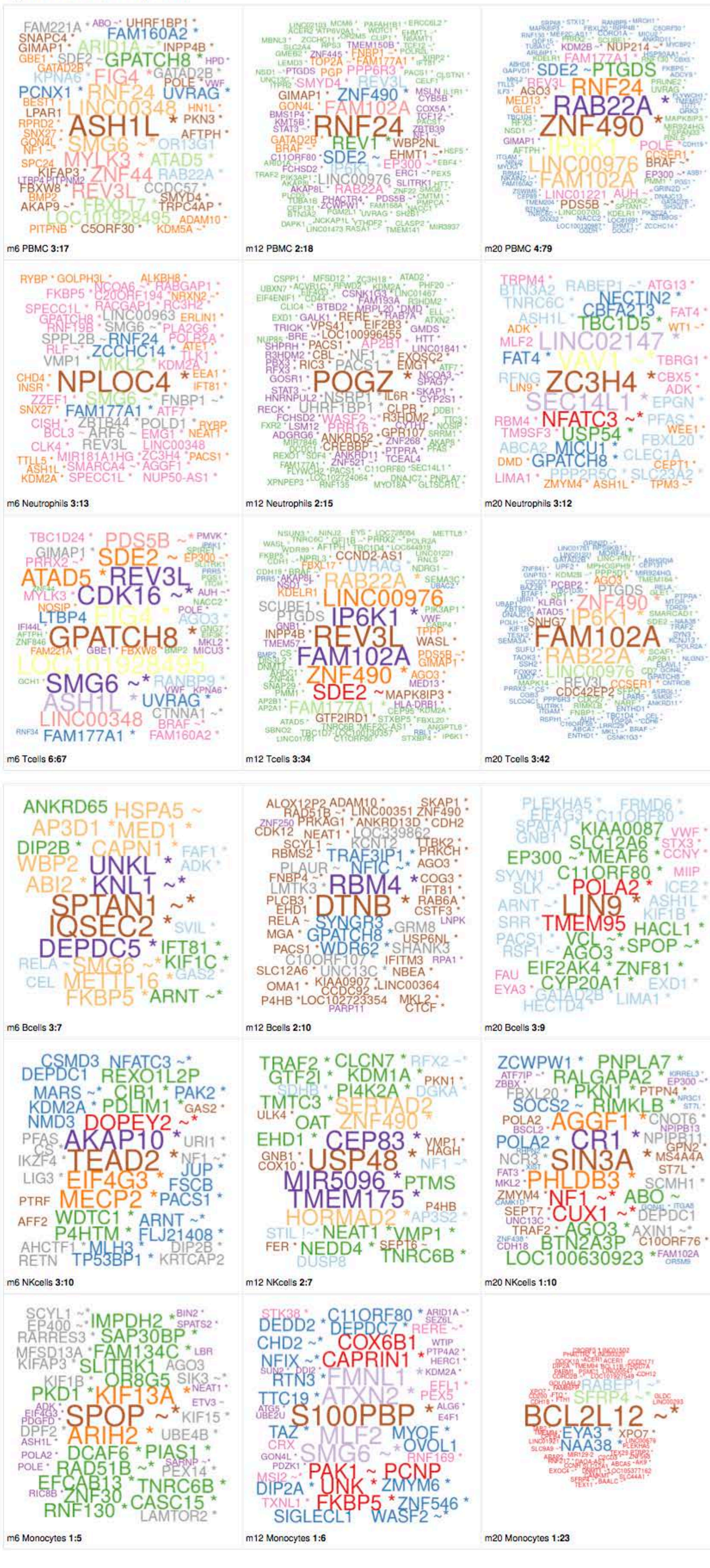


MECOM sites

Timepoint	CellType	Integration Site	nearestOnco	distToNearestOnco	sonicAbundance	sonicAbundanceRank
m12	NKcells	chr3-169243565	MECOM	-95831	1	466

What are the most frequently occuring gene types in subject pWAS_UK11?

The word clouds below show the abundances of integration sites as designated by genenames for each sample. Each word cloud is labled by the timepoint, celltype, and the range of sonic abundance for the top 100 sites.



m6 PBMC 3:17 | m12 PBMC 2:18 | m20 PBMC 4:79

m6 Neutrophils 3:13 | m12 Neutrophils 2:15 | m20 Neutrophils 3:12

m6 Tcells 6:67 | m12 Tcells 3:34 | m20 Tcells 3:42

m6 Bcells 3:7 | m12 Bcells 2:10 | m20 Bcells 3:9

m6 NKcells 3:10 | m12 NKcells 2:7 | m20 NKcells 1:10

m6 Monocytes 1:5 | m12 Monocytes 1:6 | m20 Monocytes 1:23

Integration sites shared between cell types.

Month 6

	Bcells	Monocytes	Neutrophils	NKcells	PBMC	Tcells
Bcells	3821	0	2	5	8	5
Monocytes	0	237	5	2	2	1
Neutrophils	2	5	1499	14	43	63
NKcells	5	2	14	2104	6	3
PBMC	8	2	43	6	1362	317
Tcells	5	1	63	3	317	3016

Month 12

	Bcells	Monocytes	Neutrophils	NKcells	PBMC	Tcells
Bcells	1725	4	6	5	17	3
Monocytes	4	238	7	3	6	0
Neutrophils	6	7	478	8	16	2
NKcells	5	3	8	468	14	0
PBMC	17	6	16	14	1732	42
Tcells	3	0	2	0	42	2014

Month 20

	Bcells	Monocytes	Neutrophils	NKcells	PBMC	Tcells
Bcells	583	0	8	2	10	0
Monocytes	0	60	0	0	1	0
Neutrophils	8	0	1237	9	112	13
NKcells	2	0	9	362	8	2
PBMC	10	1	112	8	4370	110
Tcells	0	0	13	2	110	564