Supplementary Report 3:
Genes of interest marked by vector integration
ALL Patients only (CR/PRtd & PR/NR Response Groups)

Contents

Summary

Lentiviral vectors integrate into genomes of targeted host cells (Tcells). These genomic locations of vector integrations are identifiable through integration site sequencing. Abundances of individual cell clones can be inferred by the sonicLength method (**Berry *et al.* 2012**).

In this report, we mined the data collected from integration site sequencing for 11 CART treated subjects. We constructed 4 gene lists based on: 1 & 2) increased / decreased integration site occurrence in patient samples relative to the initial transduction product, 3) peak clonal abundance, and 4) longitudinal clonal persistence. More about each of these criteria is below:

- **Integration Frequency** is the rate at which integration sites are observed within a gene. This is compared between patient samples and the initial transduction product to score enrichment or depletion during growth in patients. The top of genes with higher patient sample integration frequency over transduction samples were chosen for study (p-value <= 0.05 after exclusion of genes with clones from less than 2 patients and less than 10 observed clones).
- **Clonal Abundance** can be determined during analysis by quantifying the number of sites of linker ligation associated with each unique integration site. This method is further described in **Berry *et al.* 2012**. This allows clonal expansion to be quantified. The top 1% of the genes were selected for study based on their maximal peak clonal abundance.
- **Longitudinal Observation** of clones is the quantification of observed timespans and last observed timepoints. The maximum value for clones within a gene were considered for characterization of the gene in this analysis. Genes were only considered if there were 10 or more integration sites isolated from at least two different patient samples. Genes were also not considered if they only consited of clones which were observed once or the last observed timepoint was less than 90 days from initial infusion.

A point to keep in mind through all this analysis is that integration sites are sampled from a larger population. It would be rare for all integration sites in a sample to be represented in the sequence data.

Table 1: Summary of each filtering criteria.

Criteria	Gene Count	Onco Related1 (%)	Tumor Suppressors (%)	Lymphoma Related2 (%)	COSMIC Related3 (%)	TCGA Related4 (%)	Clonal Hema. Related5 (%)
Enrichment	2	*/* 100.0	/ 50.0	/ 0	*/* 100.0	/ 50.0	*/* 50.00
Depletion	0	NA/NA	NA/NA	NA/NA	NA/NA	NA/NA	NA/NA
Abundance	40	*/* 22.5	/ 12.5	/ 0	*/* 15.0	*/* 12.5	/ 0.00
Longitudinal	0	NA/NA	NA/NA	NA/NA	NA/NA	NA/NA	NA/NA
Composite	41	*/* 24.4	*/ 14.6	/ 0	*/* 17.1	*/* 14.6	/ 2.44

Table 1 summarizes the size and contents of each criteria gene list identified by the various methods. Significance of overlap between lists are displayed by asterisks before the percent of genes identifed from the criteria list which overlap with the column specified group. The asterisk to the left of the "/" indicates a p-value below 0.05 *before* multiple comparison corrections, while an asterisk to the right of the "/" indicates a p-value below 0.05 *after* multiple comparison corrections. Significance was tested using Fishers Exact test and multiple comparison corrections were made using a Benjamini-Hochberg (FDR) method for each criteria based list.

Percent of all analyzed transcription units associated with each list as as follows:

- Onco Related: 10.52%
- Tumor Suppressors: 5.13%
- Lymphoma Related: 0.21%
- COSMIC Related: 4.32%
- TCGA Related: 3.19%
- Clonal Hematopoiesis Related: 0.22%

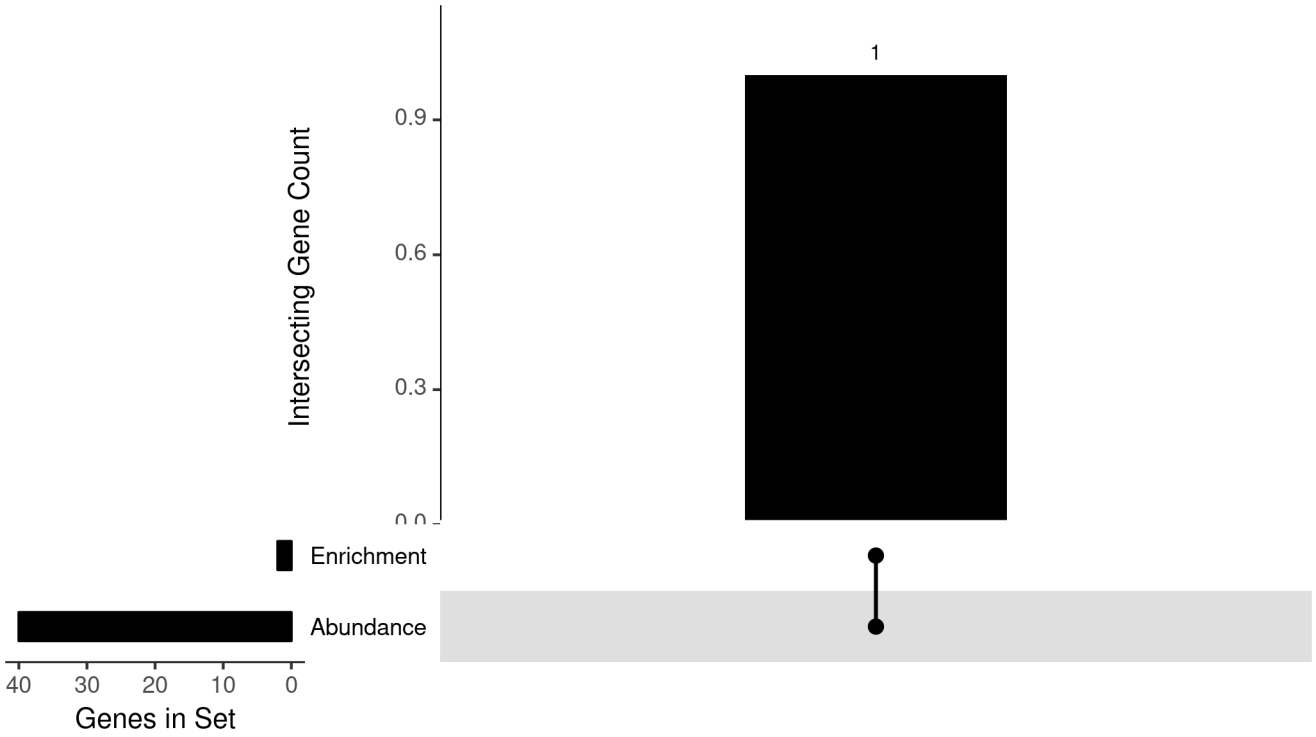

Figure 1: Intersecting gene lists identified through the various selection criteria.

Table 2: The most consistently observed genes from filtering by various criteria. The 'Criteria.' column is a count of how many times the gene was identified by these methods, while the 'Patients' column notes how many specimens collected from patients have had integration sites within the noted gene.

Gene	Patients	Freq. Change (%)	Peak Abund.	Long. Obs.	Criteria
RNF213	3	136.9	5	28	EA

Table 3: GO Biological Process. Top 1 per group. Total genes considered: 35

Group	GO ID	GO Term	Term Size	Gene Count	Adjusted P-value
1	GO:0006325	chromatin organization	481	8	0.0169853
2	GO:0048534	hematopoietic or lymphoid organ development	471	7	0.0169853

Integration Frequency (Enrichment)

Table 4: Table of top 2 genes with the most frequent clonal enrichment.

Gene	Num. Patients	TDN Sites	Patient Sites	Onco-Related	Frequency Increase (%)
CREBBP	4	36	11	TRUE	218.6
RNF213	3	44	10	TRUE	136.9

Genes with the Most Abundant Clones

Table 5: Table of top 40 Genes containing the highest abundant clones.

Gene	Num. Patients	Peak Abundance	Peak Rel. Abund.	Clonal Gini Index	Onco-Related
ATAT1	1	13	0.520	0.000	FALSE
KIF1B	3	12	0.038	0.550	FALSE
PIAS2	1	10	0.054	0.000	FALSE
DSE	2	9	0.029	0.400	FALSE
PGD	1	9	0.030	0.400	FALSE
DUSP16	1	8	0.121	0.000	FALSE
AHRR	2	7	0.042	0.444	FALSE
BAZ2A	1	7	0.023	0.000	FALSE
HNRNPUL2-BSCL2	2	7	0.023	0.400	FALSE
PPP6R3	3	7	0.022	0.408	FALSE
ERP44	1	6	0.019	0.000	FALSE
FANCA	5	6	0.009	0.301	TRUE
HOPX	1	6	0.036	0.000	FALSE
LYPLAL1	1	6	0.037	0.000	FALSE
MRE11	1	6	0.037	0.000	TRUE
PPP3CA	3	6	0.013	0.400	FALSE
UBR2	4	6	0.019	0.357	FALSE
ZNF354B	1	6	0.019	0.000	FALSE
ABHD17A	1	5	0.031	0.000	FALSE
AP3B1	2	5	0.006	0.381	FALSE
APOF	1	5	0.031	0.000	FALSE
ARNT	1	5	0.005	0.333	TRUE
C3orf58	2	5	0.016	0.333	FALSE
DNMT1	3	5	0.007	0.311	TRUE
EML5	1	5	0.030	0.000	FALSE
IPO7	2	5	0.139	0.333	FALSE
KMT2C	3	5	0.016	0.333	TRUE
MAN1A2	2	5	0.027	0.333	FALSE
NLRC3	1	5	0.030	0.000	FALSE
PHF20	1	5	0.006	0.333	TRUE
PIP4K2A	1	5	0.030	0.000	FALSE
POT1	2	5	0.016	0.333	FALSE
PSMD13	2	5	0.005	0.375	FALSE
RCAN3	2	5	0.016	0.333	FALSE
RNF157	3	5	0.076	0.257	TRUE
RNF213	3	5	0.016	0.257	TRUE
SET	1	5	0.016	0.000	TRUE
SLC44A2	2	5	0.017	0.346	FALSE
SNX13	3	5	0.007	0.333	FALSE
UTY	2	5	0.016	0.361	FALSE

Reference Data

The NCBI RefGenes data set was used to identify gene regions (hg38) while genes identified as onco-related were from the Bushman Lab curated list of **onco-related genes**.

Gene Ontologies were extracted from the `GO.db` R-package (v3.4.1). KEGG pathways were acquired via interfacing with the KEGG web-server API through the `KEGGREST` R-package (v1.16.1). Gene lists, including RefSeq genes used for annotation of integration sites, were standardized to HGNC gene symbols (date: 2018-02-07). Groups identified in GO and KEGG analyses were determined from Jaccard distances between identified terms, followed by modularity-optimizing clustering from a weighted-undirected graph using a Louvain algorithm (**Blondel *et al.* 2008**). Terms within groups of GO or KEGG terms have greater overlap of gene lists between themselves that between terms found in other groups. This method was implemented to help reduce the functional redundancy commonly observed in GO and overlapping pathways observed with KEGG.

Comprehensive Genes of Interest Table

Table 6: Table of all genes identified within analysis.

Gene	Chromosome	Start Pos.	End Pos.	Patients	Freq. Change (%)	Peak Abund.	Long. Obs.	Criteria
RNF213	chr17	80,255,860	80,403,781	3	136.9	5	0	EA
FANCA	chr16	89,732,550	89,821,657	5	18.5	6	0	A
CREBBP	chr16	3,720,054	3,885,120	4	218.6	2	0	E
UBR2	chr6	42,559,021	42,698,505	4	32.7	6	0	A
DNMT1	chr19	10,128,343	10,200,135	3	0.9	5	0	A
KIF1B	chr1	10,205,705	10,386,603	3	189.6	12	0	A
KMT2C	chr7	152,129,924	152,441,005	3	247.5	5	0	A
PPP3CA	chr4	101,018,429	101,352,471	3	178.0	6	0	A
PPP6R3	chr11	68,455,717	68,620,333	3	69.7	7	0	A
RNF157	chr17	76,137,452	76,245,311	3	-20.4	5	0	A
SNX13	chr7	17,785,760	17,945,508	3	468.7	5	0	A
AHRR	chr5	299,175	443,290	2	-13.1	7	0	A
AP3B1	chr5	77,997,325	78,299,755	2	247.5	5	0	A
ATAT1	chr6	30,621,841	30,651,823	2	197.9	13	0	A
DSE	chr6	116,249,151	116,443,291	2	1985.1	9	0	A
IPO7	chr11	9,379,621	9,453,127	2	131.7	5	0	A
MAN1A2	chr1	117,362,462	117,530,698	2	247.5	5	0	A
POT1	chr7	124,817,385	124,934,983	2	131.7	5	0	A
PSMD13	chr11	231,807	257,984	2	-19.8	5	0	A
RCAN3	chr1	24,497,350	24,542,020	2	4.3	5	0	A
SLC44A2	chr19	10,597,444	10,649,559	2	78.7	5	0	A
UTY	chrY	13,243,378	13,485,670	2	681.9	2	0	A
ARNT	chr1	150,804,704	150,881,768	1	-9.3	5	0	A
BAZ2A	chr12	56,590,595	56,641,379	1	-25.5	7	0	A
DUSP16	chr12	12,468,281	12,567,514	1	-19.8	8	0	A
EML5	chr14	88,609,829	88,797,752	1	421.3	5	0	A
ERP44	chr9	99,974,180	100,104,052	1	108.5	6	0	A
HOPX	chr4	56,642,987	56,686,706	1	Inf	6	0	A
LYPLAL1	chr1	219,168,830	219,217,865	1	73.8	6	0	A
MRE11	chr11	94,412,300	94,498,908	1	73.8	6	0	A
NLRC3	chr16	3,534,035	3,582,404	1	48.9	5	0	A
PGD	chr1	10,393,991	10,425,511	1	Inf	9	0	A
PHF20	chr20	35,767,000	35,955,366	1	22.7	5	0	A
PIAS2	chr18	46,798,224	46,925,167	1	48.9	10	0	A
PIP4K2A	chr10	22,529,836	22,719,574	1	108.5	5	0	A
SET	chr9	128,678,654	128,701,396	1	15.8	5	11	A
ZNF354B	chr5	178,854,952	178,889,423	1	247.5	6	0	A