Ghassemi Report

Introduction

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 et al. (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 et al. (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 Berry et al. (2012).

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 integrations are upstream (-) or downstream (+, no sign) of the nearest gene. Nearest genes possess additional annotations described in the table below.

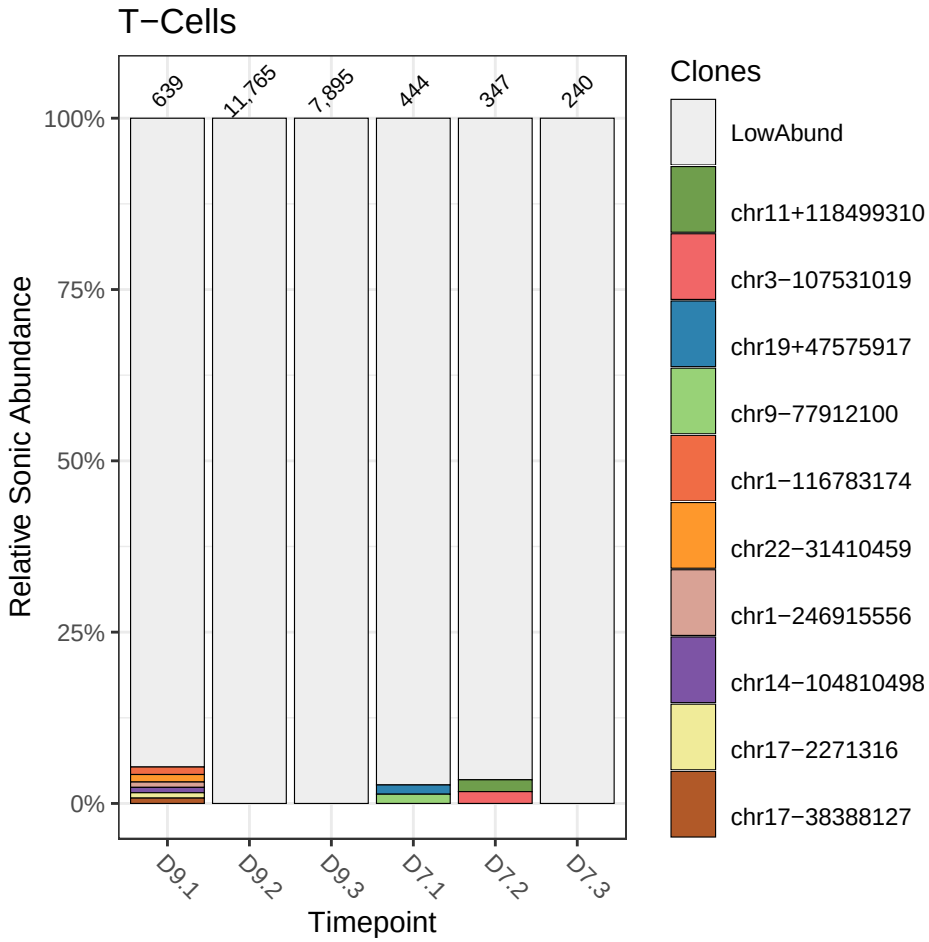
Symbol	Meaning
*	site is within a transcription unit
~	site is within 50kb of a cancer related gene
!	nearest gene was assocaited with lymphoma in humans

Samples

GTSP	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	Gini	Chao1	Shannon	Pielou	UC50
GTSP3801	D7.1	T-Cells	445,904	444	314	0.244	981	5.60	0.974	93
GTSP3802	D7.2	T-Cells	569,295	347	285	0.165	2,293	5.53	0.979	112
GTSP3803	D7.3	T-Cells	313,069	240	162	0.263	467	4.93	0.969	43
GTSP3798	D9.1	T-Cells	1,235,174	639	451	0.247	1,672	5.96	0.975	132
GTSP3799	D9.2	T-Cells	1,303,789	11,765	11,168	0.049	129,421	9.30	0.998	5,286
GTSP3800	D9.3	T-Cells	1,341,729	7,895	6,644	0.135	22,310	8.74	0.994	2,697

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.



Scan stats

We performed a Scan Statistics analysis as described in Berry et al. (2014). This method looks for clusters of insertion sites that differentiate two samples in a way that prevents doing multiple comparisons and reducing the significance of the test. Genes associated with the Scan intervals were retrieved using two methods. genesIntsites uses the closest gene to each insertion site. genesEntrez retrieve all genes from the Entrez database that intersects that interval. Those tables are also available in "Scan_stats.xlsx".

Table 1: scan statistics

seqnames	start	end	width	countD7	countD9	target.min	clusterSource	genesIntsites	genesEntrez
chr3	197,977,521	198,005,211	27,691	6	1	0.0000772	D7	LMLN	LMLN
chr9	43,401,923	43,447,894	45,972	4	0	0.0483634	D7	XLOC_007697	LOC105379443
chr12	132,653,197	132,654,034	838	3	1	4.6910002	D7	POLE	POLE
chr19	47,504,800	47,575,928	71,129	3	1	4.6910002	D7	ZNF541 NAPA	NAPA ZNF541
chr22	50,391,250	50,461,381	70,132	7	14	0.6426809	D9	PPP6R2 SBF1	SBF1 PPP6R2

hot ROCs

The ROC curve heatmaps were generated as described in Berry et al. (2014). They show how much the distribution of insertion sites across several genomic and epigenomic features differs from a random distribution. The heatmaps include D7 and D9 samples as well as some preinfusion samples from Nobles et al. (2020). All samples fallow roughly the same pattern for all features.

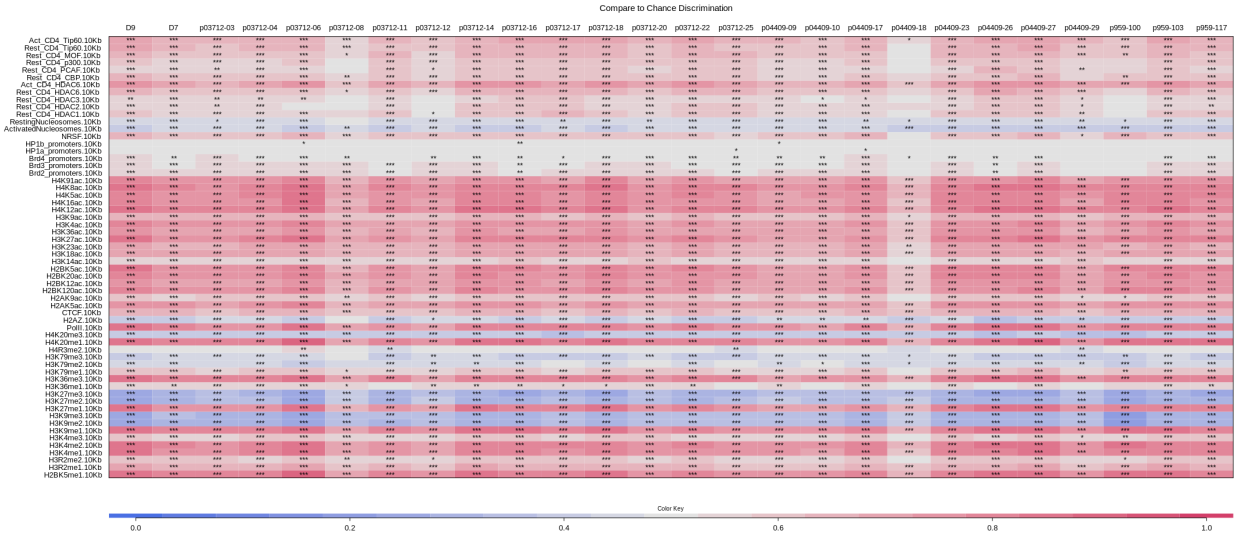


Figure 1: Metagenomic features

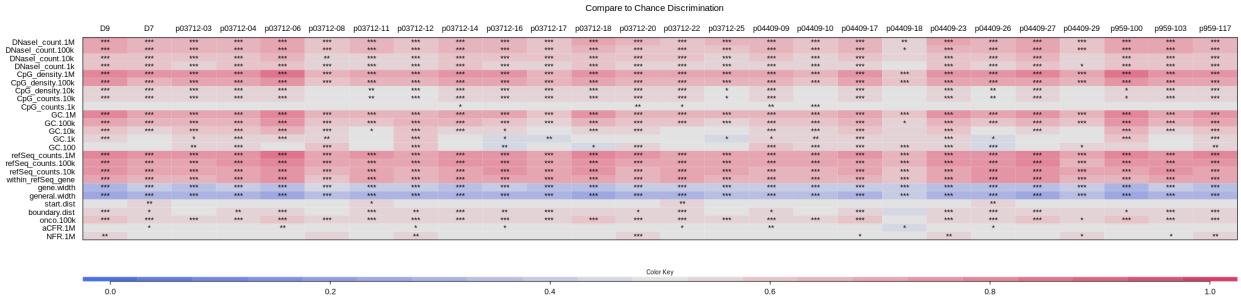


Figure 2: Genomic features

Tracking of clonal abundances

code availability

all the code used to generate this report is available at https://github.com/Adrian-Cantu/Ghassemi_CART

References

Charles C. Berry, Nicolas A. Gillet, Anat Melamed, Niall Gormley, Charles R. M. Bangham, and Frederic D. Bushman. Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics*, 28(6):755–762, March 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts004.

Charles C. Berry, Karen E. Ocwieja, Nirav Malani, and Frederic D. Bushman. Comparing DNA integration site clusters with scan statistics. *Bioinformatics*, 30(11):1493–1500, June 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu035.

Anne Chao. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, 43(4):783–791, 1987. ISSN 0006-341X. doi: 10.2307/2531532.

Christopher L. Nobles, Scott Sherrill-Mix, John K. Everett, Shantan Reddy, Joseph A. Fraietta, David L. Porter, Noelle Frey, Saar I. Gill, Stephan A. Grupp, Shannon L. Maude, Donald L. Siegel, Bruce L. Levine, Carl H. June, Simon F. Lacey, J. Joseph Melenhorst, and Frederic D. Bushman. CD19-targeting CAR T cell immunotherapy outcomes correlate with genomic modification by vector integration. *The Journal of Clinical Investigation*, 130(2):673–685, February 2020. ISSN 1558-8238. doi: 10.1172/JCI130144.