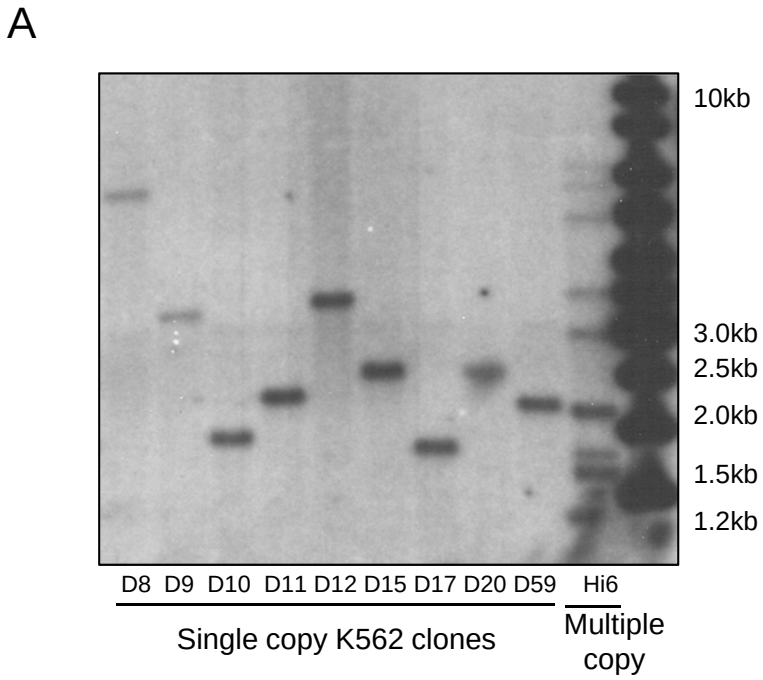
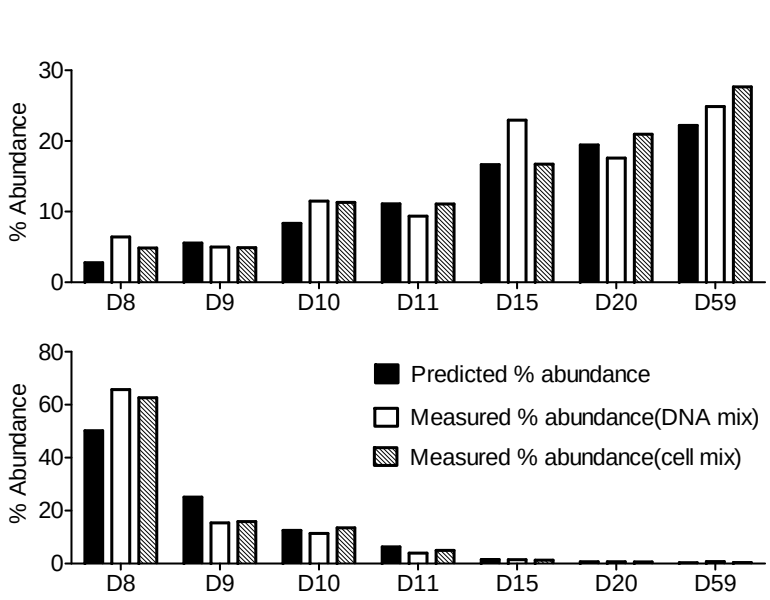
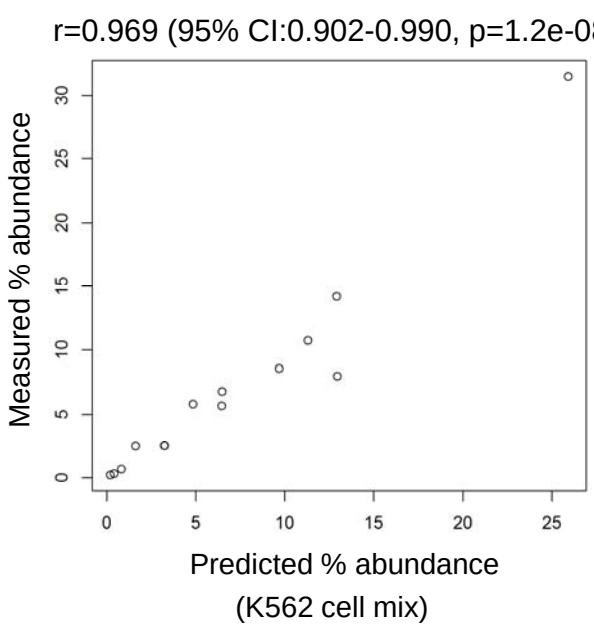
Figure S1


A
10kb
3.0kb
2.5kb
2.0kb
1.5kb
1.2kb
D8 D9 D10 D11 D12 D15 D17 D20 D59
Hi6
Single copy K562 clones
Multiple copy


B
% Abundance
D8 D9 D10 D11 D15 D20 D59
Predicted % abundance
Measured % abundance(DNA mix)
Measured % abundance(cell mix)


C
r=0.969 (95% CI:0.902-0.990, p=1.2e-08)
Measured % abundance
Predicted % abundance
(K562 cell mix)

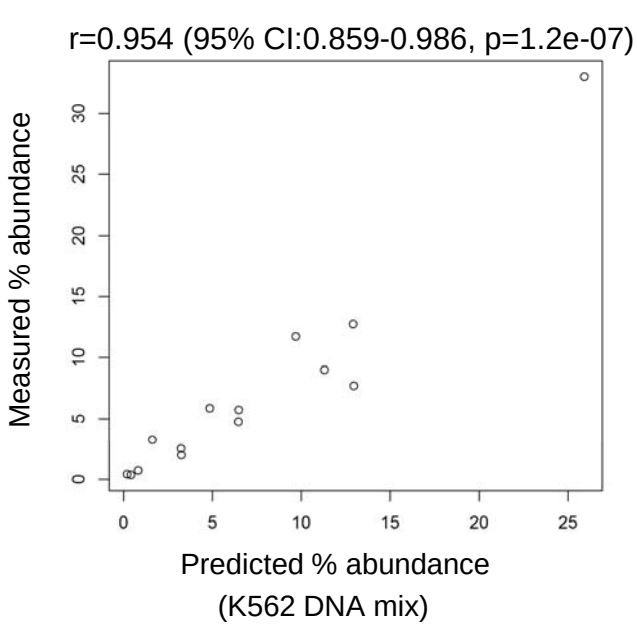

D
r=0.954 (95% CI:0.859-0.986, p=1.2e-07)
Measured % abundance
Predicted % abundance
(K562 DNA mix)

Figure S2

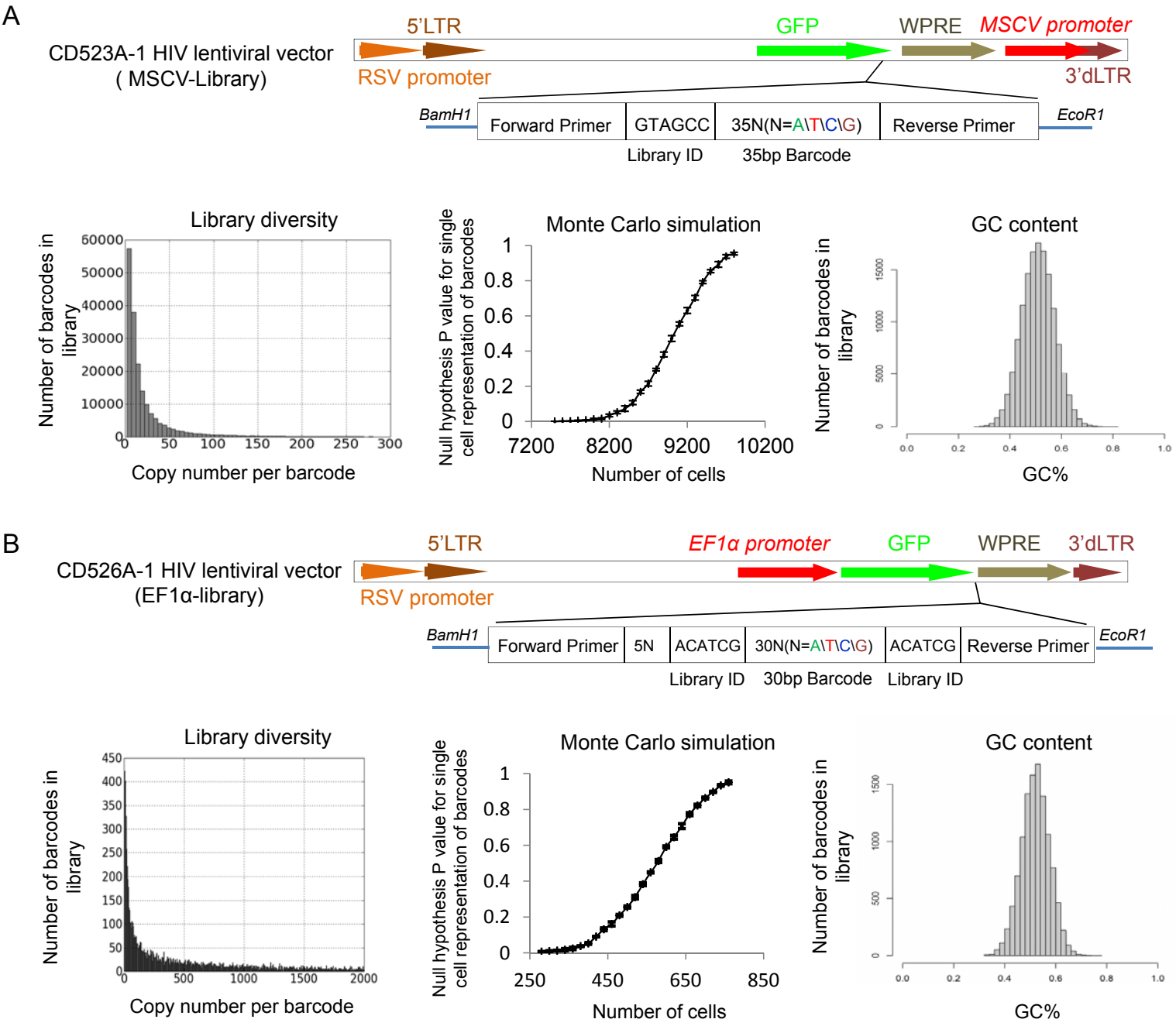

A
CD523A-1 HIV lentiviral vector
(MSCV-Library)
5'LTR
GFP
WPRE
MSCV promoter
RSV promoter
3'dLTR
BamH1
Forward Primer
GTAGCC
35N(N=A\T\C\G)
Reverse Primer
EcoR1
Library ID
35bp Barcode
Library diversity
Number of barcodes in library
Copy number per barcode
Monte Carlo simulation
Null hypothesis P value for single cell representation of barcodes
Number of cells
GC content
Number of barcodes in library
GC%
B
CD526A-1 HIV lentiviral vector
(EF1α-library)
5'LTR
EF1α promoter
GFP
WPRE
3'dLTR
RSV promoter
BamH1
Forward Primer
5N
ACATCG
30N(N=A\T\C\G)
ACATCG
Reverse Primer
EcoR1
Library ID
30bp Barcode
Library ID
Library diversity
Number of barcodes in library
Copy number per barcode
Monte Carlo simulation
Null hypothesis P value for single cell representation of barcodes
Number of cells
GC content
Number of barcodes in library
GC%

Figure S3

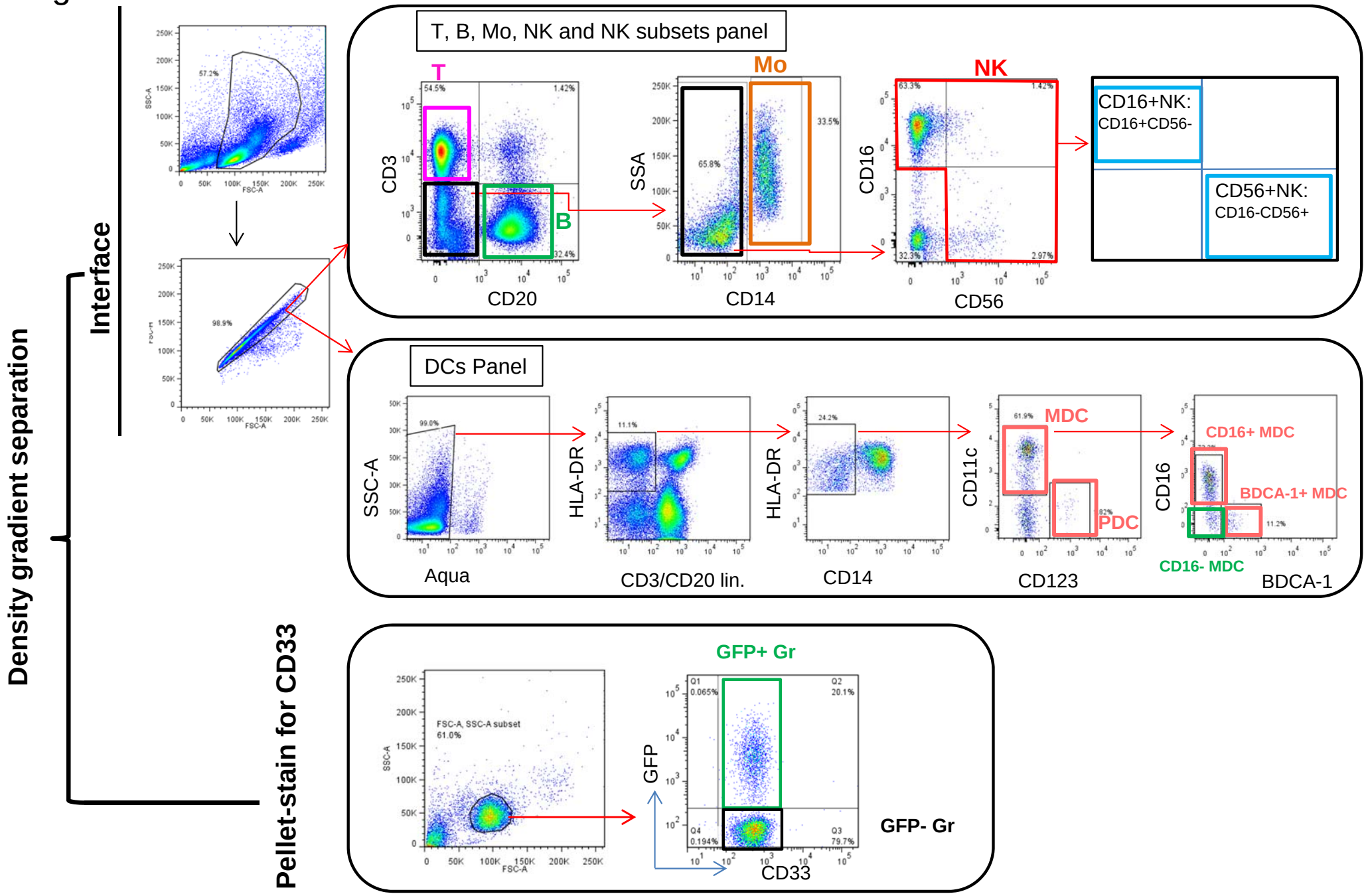
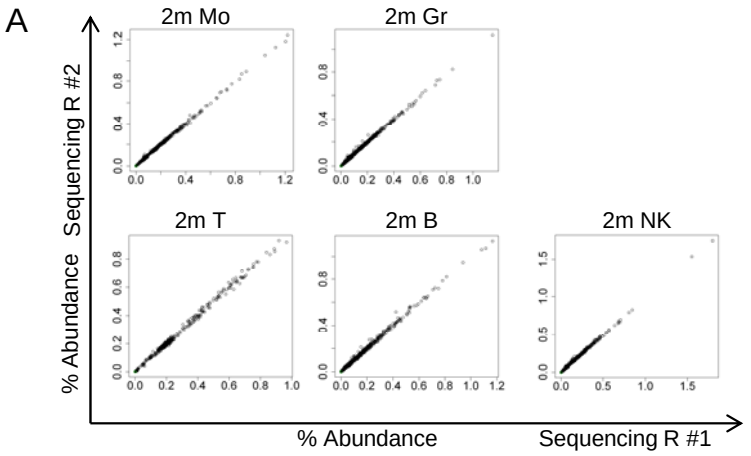
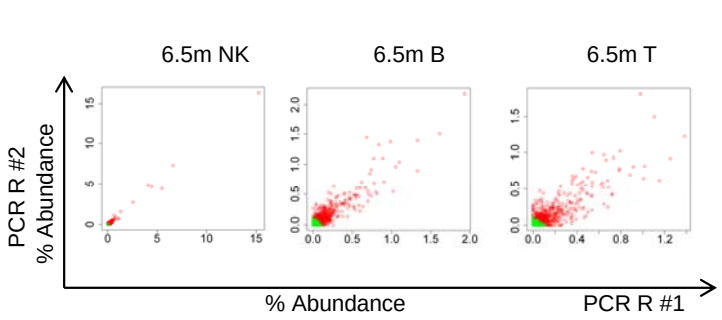
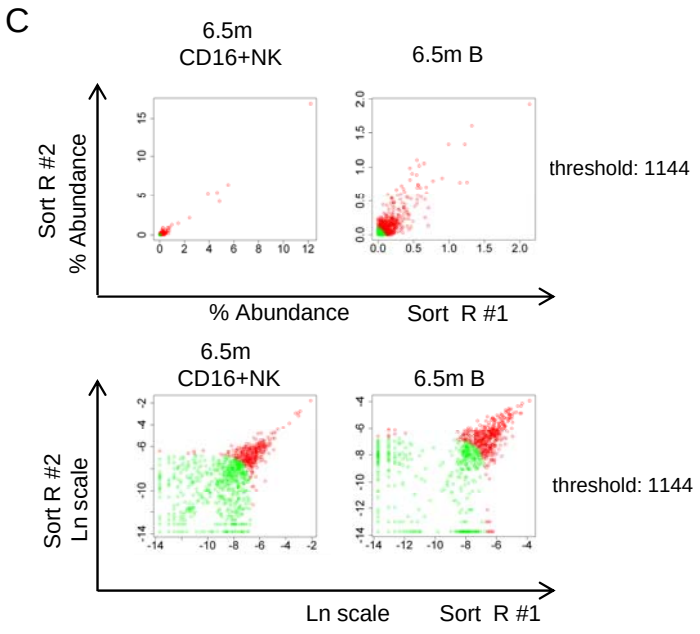

Density gradient separation
Interface
Pellet-stain for CD33
T, B, Mo, NK and NK subsets panel
T
B
Mo
NK
CD3
CD20
SSA
CD14
CD16
CD56
CD16+NK:
CD16+CD56-
CD56+NK:
CD16-CD56+
DCs Panel
SSC-A
Aqua
HLA-DR
CD3/CD20 lin.
HLA-DR
CD14
CD11c
CD123
MDC
PDC
CD16
CD16+ MDC
BDCA-1+ MDC
CD16- MDC
BDCA-1
GFP+ Gr
GFP- Gr
GFP
CD33
FSC-A, SSC-A subset
61.0%

Figure S4


A
2m Mo
2m Gr
2m T
2m B
2m NK
% Abundance Sequencing R #2
% Abundance
Sequencing R #1


B
6.5m NK
6.5m B
6.5m T
PCR R #2
% Abundance
% Abundance
PCR R #1


C
6.5m
CD16+NK
6.5m B
Sort R #2
% Abundance
threshold: 1144
% Abundance
Sort R #1
6.5m
CD16+NK
6.5m B
Sort R #2
Ln scale
threshold: 1144
Ln scale
Sort R #1

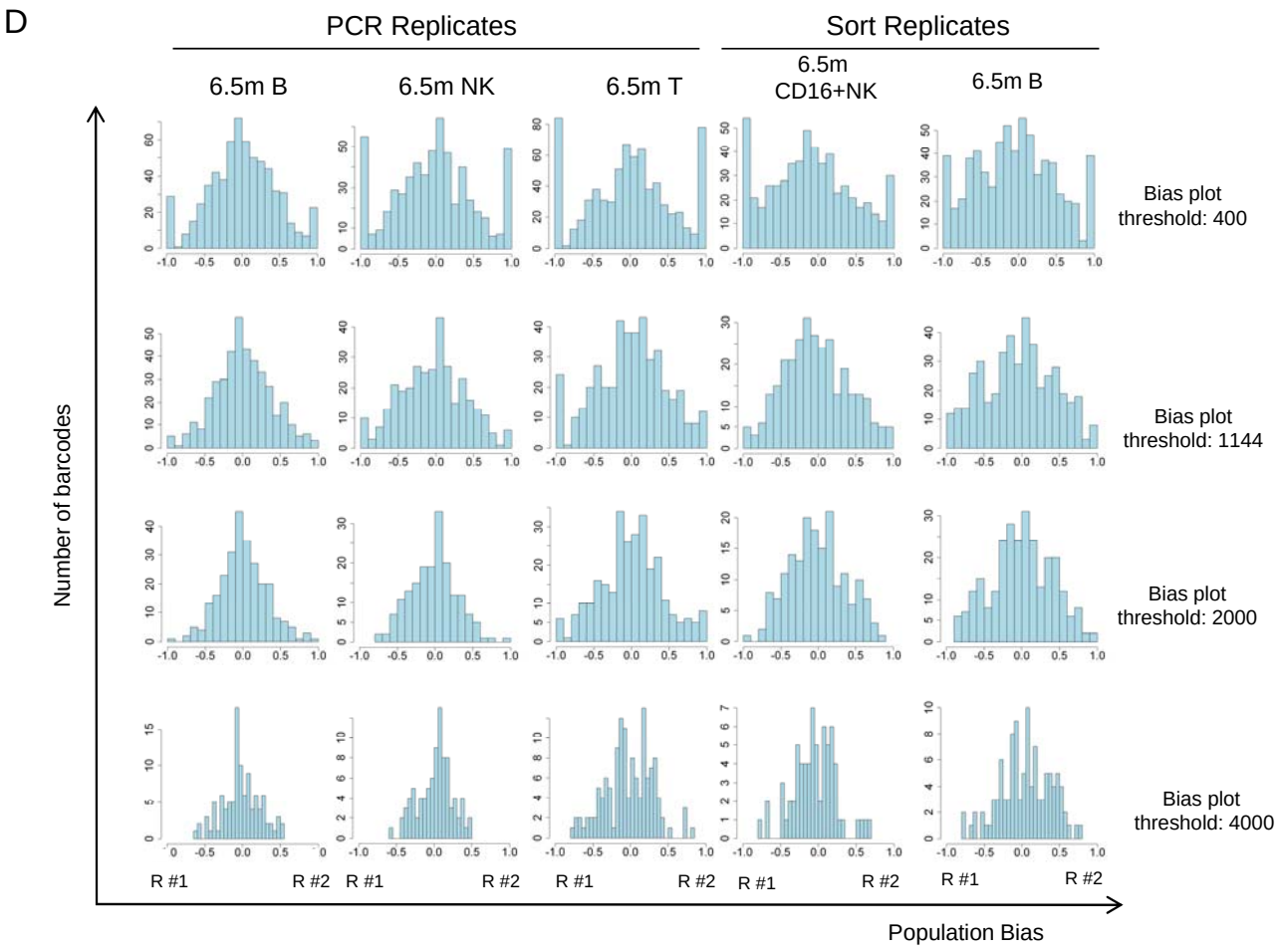

D
PCR Replicates
Sort Replicates
6.5m B
6.5m NK
6.5m T
6.5m
CD16+NK
6.5m B
Bias plot
threshold: 400
Bias plot
threshold: 1144
Bias plot
threshold: 2000
Bias plot
threshold: 4000
Number of barcodes
R #1
R #2
Population Bias

Figure S5

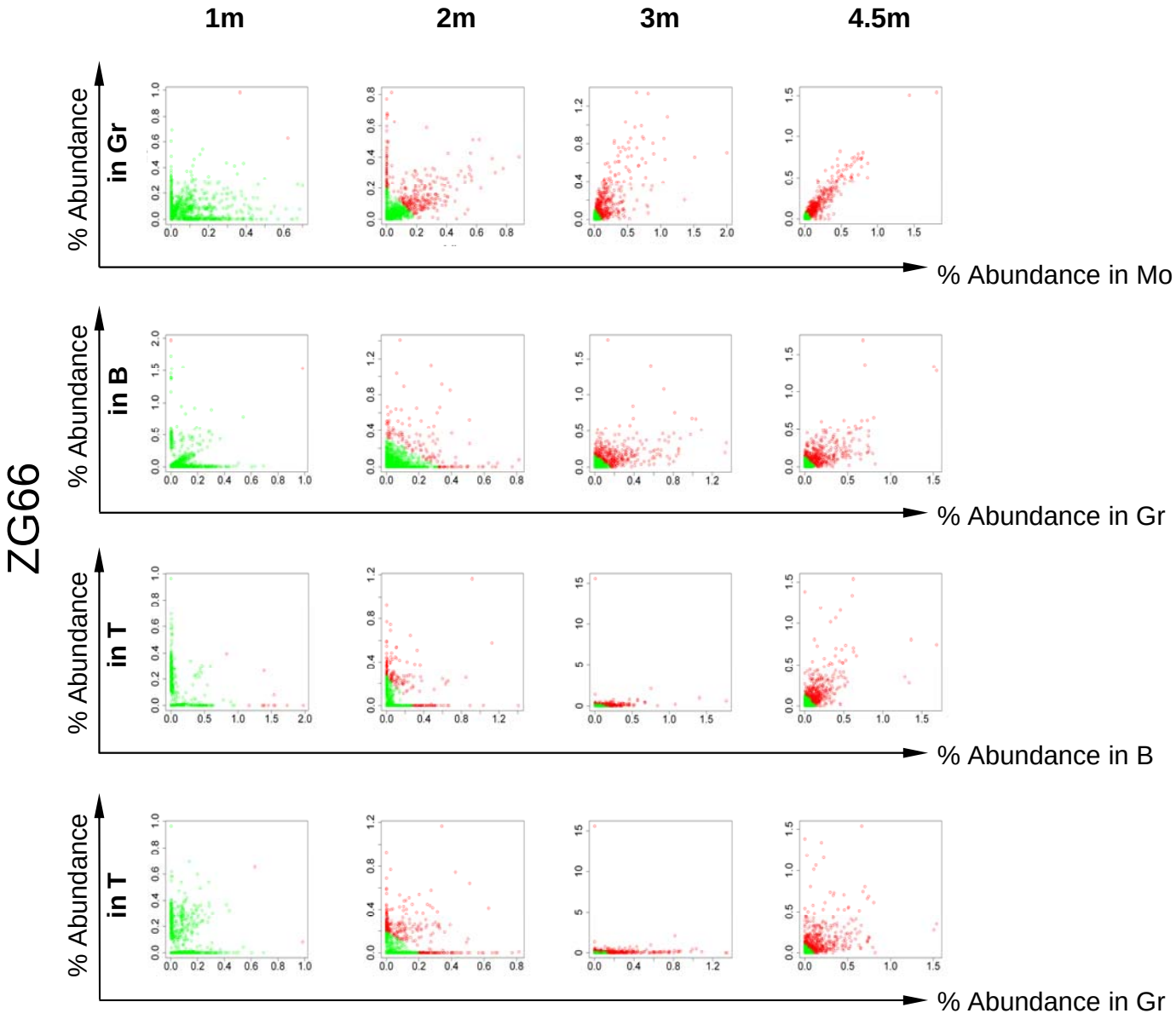

ZG66
1m
2m
3m
4.5m
% Abundance in Gr
% Abundance in Mo
% Abundance in B
% Abundance in Gr
% Abundance in T
% Abundance in B
% Abundance in T
% Abundance in Gr

Figure S6

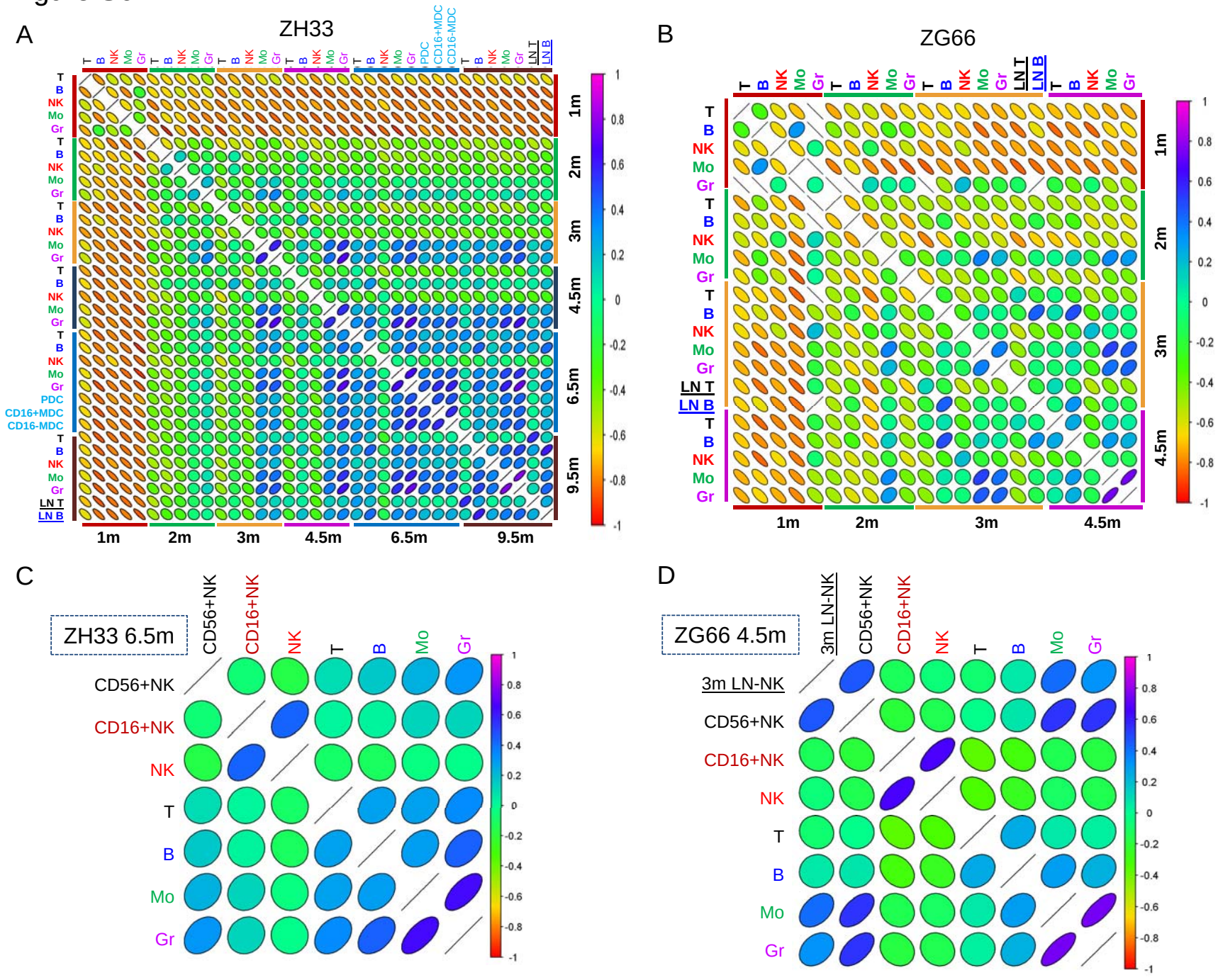
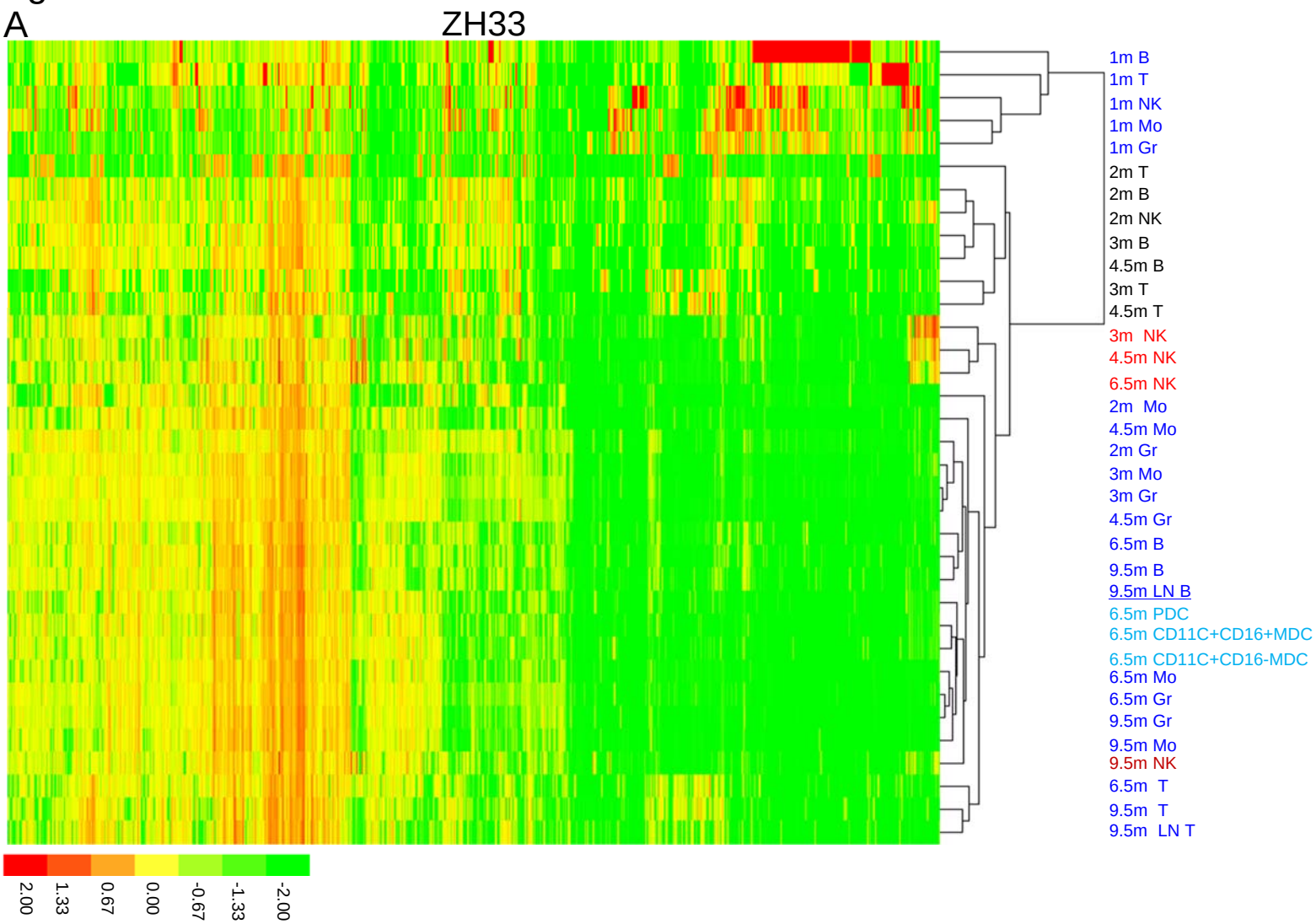
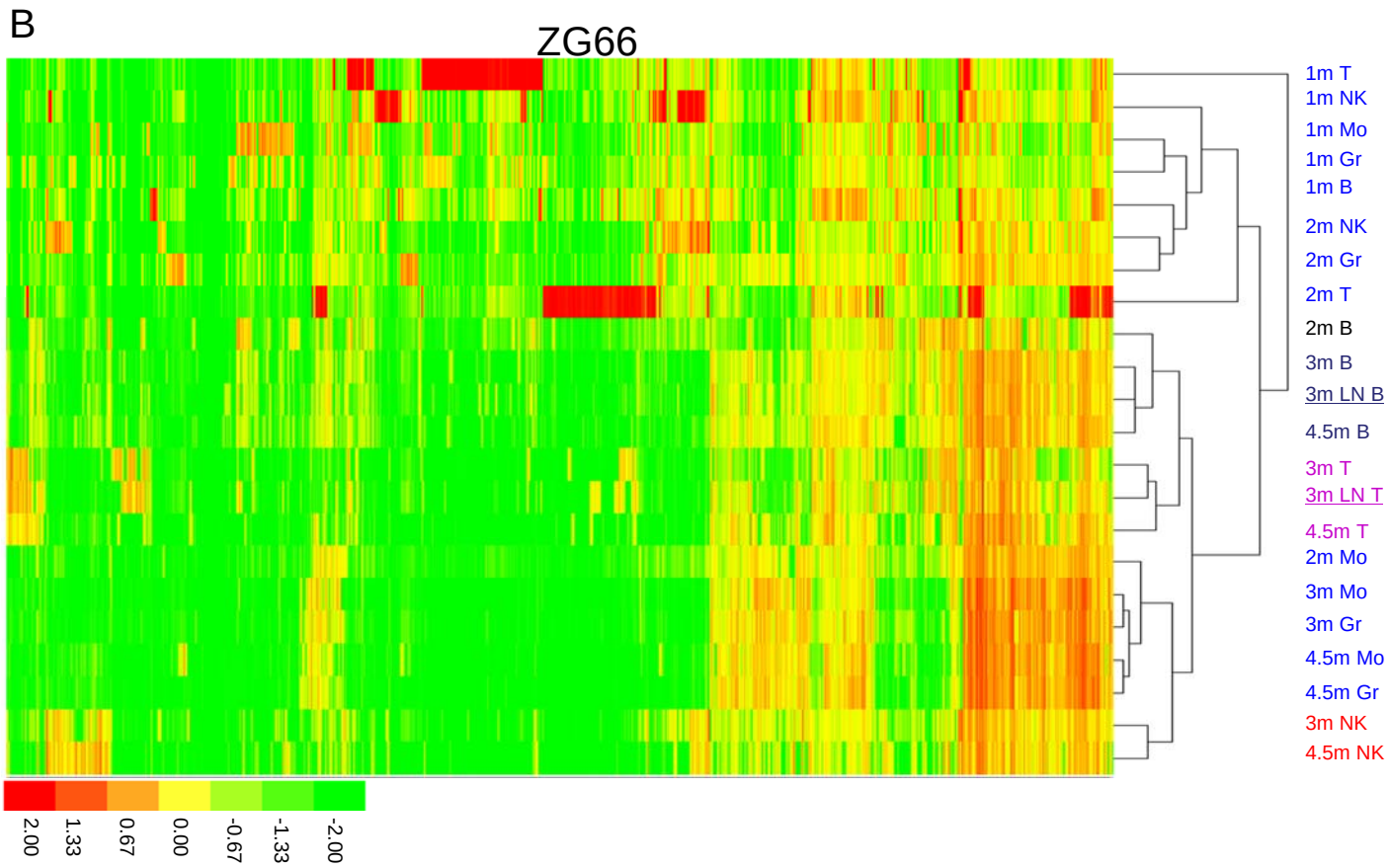

A
ZH33
T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr PDC CD16+MDC CD16-MDC T B NK Mo Gr LN T LN B
T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr PDC CD16+MDC CD16-MDC LN T LN B
1m 2m 3m 4.5m 6.5m 9.5m
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
B
ZG66
T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr LN T LN B T B NK Mo Gr
T B NK Mo Gr T B NK Mo Gr T B NK Mo Gr LN T LN B T B NK Mo Gr
1m 2m 3m 4.5m
C
ZH33 6.5m
CD56+NK CD16+NK NK T B Mo Gr
D
ZG66 4.5m
3m LN-NK CD56+NK CD16+NK NK T B Mo Gr

Figure S7


A
ZH33
1m B
1m T
1m NK
1m Mo
1m Gr
2m T
2m B
2m NK
3m B
4.5m B
3m T
4.5m T
3m NK
4.5m NK
6.5m NK
2m Mo
4.5m Mo
2m Gr
3m Mo
3m Gr
4.5m Gr
6.5m B
9.5m B
9.5m LN B
6.5m PDC
6.5m CD11C+CD16+MDC
6.5m CD11C+CD16-MDC
6.5m Mo
6.5m Gr
9.5m Gr
9.5m Mo
9.5m NK
6.5m T
9.5m T
9.5m LN T
2.00
1.33
0.67
0.00
-0.67
-1.33
-2.00


B
ZG66
1m T
1m NK
1m Mo
1m Gr
1m B
2m NK
2m Gr
2m T
2m B
3m B
3m LN B
4.5m B
3m T
3m LN T
4.5m T
2m Mo
3m Mo
3m Gr
4.5m Mo
4.5m Gr
3m NK
4.5m NK
2.00
1.33
0.67
0.00
-0.67
-1.33
-2.00

Supplemental Figure Legends

Figure S1: Validation of barcode retrieval as a quantitative reflection of clonal contributions. Individual transduced K562 clones were isolated by single cell sorting of GFP+ cells, and single barcode copy clones were identified via single bands on Southern blot (A) following digestion with EcoR1, a single cutter within the vector, and hybridization with a WPRE probe. Mixtures of different ratios of the K562 clones, either cells or DNA, were generated, and low cycle PCR followed by Illumina sequencing and quantitation of individual barcode reads was performed on DNA from these mixtures. The expected versus actual retrieved % contribution of each barcode in the DNA or cell mixtures are summarized in (B). Pearson correlation coefficients for the relationship between expected and actual barcode fractions are shown in (C) and (D) for cell and DNA mixtures along with the 95% confidence intervals and p values. Related to Fig 1.

Figure S2: Analysis of lentiviral vector barcode library diversity. (A) CD523A-1 HIV-derived lentiviral vector backbone showing the location of the inserted 6bp library ID, 35bp barcode, MSCV promoter and GFP marker gene. 3'dLTR=deleted LTR. Library #11 was used for animals ZH33 and ZG66. The rhesus LCL8864 cell line was transduced with barcode library #11 and analyzed via low cycle PCR barcode flanking PCR, Illumina sequencing and data processing. A total of 181,161 unique barcodes were retrieved. The left hand graph shows copy number bins for each barcode on the x axis, and the number of barcodes in falling into each bin on the y axis. The middle graph shows a Monte Carlo simulation run on the on the barcode data from this experiment (python code included in data supplement). For each target cell population size (7500 to 10,000 cells with step size of 100 cells), 1,000 experiments were conducted for each round during which a DNA barcode was randomly assigned for each cell. 5 rounds of simulations were performed in total. The p value is plotted against the starting size of the cell population. The diversity of library #11 is sufficient to transduce 8200 cells simultaneously, ensuring with greater than 95% probability that more than 95% of barcodes represent single cells. The right hand panel shows the GC content of retrieved barcodes: only 0.2% of barcodes had more than 65% GC content and no barcode had GC content less than 12%. (B) CD526-1 HIV-derived lentiviral vector backbone showing the location of the inserted 6bp library IDs, 30bp barcode, EF1**α** promoter and the GFP

marker gene. Library #2 was used for animal ZH17. The diversity of library #2 following in vitro transduction of target cells, Monte Carlo simulation using this data, and GC content of retrieved barcodes are shown, analogous to above. For the total number of clones retrieved from ZH17, there is 95% probability that 88% of barcodes represent single cells. Related to Fig 1.

Figure S3: Gating strategy for flow cytometric sorting of hematopoietic lineages from the peripheral blood. Peripheral blood samples were separated over a density gradient, and stained with the antibodies listed in Table S2. The sequential gating strategy shown was used to purify large granular CD33+/GFP+ or CD33+/GFP- granulocytes from the cell pellet (bottom panel). Smaller mononuclear CD3+/CD20- T cells, CD3-/CD20+ B cells, CD3-/CD20-/CD14+ monocytes, CD3-/CD20-/CD14-/CD56 or CD16+ overall NK cells and NK subsets were purified from the density gradient interface (upper panel) , as well as dendritic cell subsets (DCs) including plasmacytoid dendritic cells (PDCs:CD3-CD20-HLADR+/CD14-/CD11c-CD123+), and myeloid dendritic cells (MDCs) (middle panel):CD16+MDCs(CD3-CD20-HLADR+/CD14-/CD11c+CD123-/BDCA-1-/CD16+), CD16-BACD-1-MDCs(D3-CD20-HLADR+/CD14-/CD11c+CD123-/CD16-/BDCA-1-) , and CD16-BACD-1+MDCs(D3-CD20-HLADR+/CD14-/CD11c+CD123-/CD16-/BDCA-1+). Post sort purities are given in Table S3. Related to Fig 1.

Figure S4: The reproducibility of quantitative barcode retrieval in technical replicates. (A) A multiplex PCR mixture of 5 samples was run through two independent Illumina sequencing replicates. Dot plots are shown with each barcode as a dot, and the fractional abundance of each barcode in replicate 1 versus replicate 2 is plotted. The average correlation was r=0.990 between replicates. (B) Replicate samples of DNA independently underwent PCR followed by sequencing and analysis, and the dot plots are shown, with an average correlation between replicates of r=0.908. Standard deviation of read fraction represented by each clone was 5.97×10^{-4} for T and 7.01×10^{-4} for NK cells in ZH33 triplicate PCRs. Standard deviation was computed individually for each clone and pooled across barcodes. ZH33 had median read counts per sample 1.07×10^{6}, with interquartile ranges 6.24×10^{5}-1.17×10^{6}. (C) Blood samples from ZH33 at 6.5m were collected, and two independent flow cytometric sorts were performed. The purified cell replicates were then used for independent DNA extraction, PCR, sequencing and analysis. The replicate results were highly correlated (mean r=0.909).

(D) A series of bias plots comparing replicate samples shown in (B and C) with different pairwise read sums as lower copy number retrieval thresholds for the two data sets being compared. We established a sampling error threshold of 1144 reads, which is the median pairwise read sum eliminating 95% of 10x biased clones in the replicate samples. In the dot plots shown in B and C, barcodes in red were above the threshold, and barcodes in green were below the threshold, and as shown on the log plot in C (lower panels), elimination of these barcodes removed 95% of the sampling error. Related to Fig 1,3,4,5.

Figure S5. Scatter plots for ZG66. Scatter plots comparing the sequencing read fraction for each barcode detected in Gr versus Mo, T versus B, B versus Gr and T versus Gr over time in ZG66. On scatter plots, all clones on the ZG66 master list were plotted if present in either lineage. Those with a sum of reads over the sampling threshold of 1144 are shown in red and those below the threshold shown in green. Related to Fig 3.

Figure S6. Spearman correlation analysis. (A) Summary of clonal lineage and kinetic relationships over time. Spearman correlation coefficients comparing barcode contributions for all barcodes falling above the sampling threshold of 1144 for pairwise comparisons, between all lineages and time points, for animal ZH33 (left panel) and ZG66 (right panel). Correlation r values and their associated p values and 95% confidence intervals are given in Table S6. (B) Graphical summary of Spearman correlation coefficients (left panel) comparing NK cells and subsets from the peripheral blood (ZH33 6.5m, left panel) and from ZG66 blood (4.5m) and node (3m) (right panel) (Table S7 gives r values, p values and 95% confidence intervals).The color scale bar for r values is on the right. Slope indicates negative or positive correlation; the shape and the color signify the strength of the correlation. MDC-myeloid dendritic cells, PDC-plasmacytoid dendritic cells, LN-lymph node. Related to Fig 4,5.

Figure S7. Clonal global hierarchical clustering analysis. Euclidian distance hierarchical clustering analysis for all master list clones from all samples in ZH33 (A), and ZG66 (B). Read number was normalized by the total valid reads obtained from each sample, a log10 transformation and a 75th percentile normalization crossing all samples. Related to Fig 4.

Supplemental Table legends

Table S1 (related to Figure 1 and Experimental Procedures): The purity of lineage-sorted cell samples utilized for analyses in all animals.

Table S2 (related to Figure 1 and Experimental Procedures): Antibodies used for flow cytometric sorting of blood and lymph node populations.

Table S3 (related to Figure 1 and Experimental Procedures): The primers and probes used for barcode PCR, multiplex sequencing PCR, and Southern blot.

Table S4 (related to Figure 1 and utilized for generation of all analyses included in Figures 3-7): Barcode datasets for ZH33, ZG66 and ZH17 including the barcode master list for all valid clones over the sequencing artifact threshold, identified at any time point, in any sample or lineage, and the read counts for each barcode.

Table S5 (related to Figure 4): The Pearson correlation coefficients of barcode contributions to pairs of lineages or timed samples used to generate Figure 4, along with 95% confidence intervals (high vs. low) and p values for each comparison of lineages and time points.

Table S6 (related to Figure 4 and Figure S6): The Spearman correlation coefficients of barcode contributions to pairs of lineages or timed samples used to generate Figure S6, along with 95% confidence intervals (high vs. low) and p values for each comparison of lineages and time points.

Table S7 (related to Figure 5 and Figure S6): The Pearson and Spearman correlation coefficients of barcode contributions to NK versus other lineage samples for ZH33 and ZG66 (used to generate Figure 5D and Figure S6B), and the 95% confidence intervals (high vs. low) and p values for the comparisons.

Supplemental Methods

Python code scripts, instructions and test file for generating a barcode master list from raw sequencing output, and for performing Monte Carlo simulations to assess barcode library diversity.

Table S1: The purity of lineage-sorted cells

Sample name	ZH33 sorted cell Purity (%)	ZH17 sorted cell Purity (%)	ZG66 sorted cell Purity (%)
1m T cell	N/A	97.8	99.4
1m B cell	N/A	97.0	99.8
1m NK Cell	N/A	98.8	99.9
1m Grans	99.6	90.5	99.9
1m Mono	N/A	95.0	98.4
2m T cell	98.5	98.4	99.5
2m B cell	98.8	97.7	99.4
2m NK Cell	92.4	99.4	99.8
2m Grans	99.5	90.5	N/A
2m Mono	97.7	94.0	98
3m T cell	98.2	99.3	99.5
3m B cell	98.4	98.9	99.6
3m NK Cell	99.4	99.0	99.3
3m Grans	99.2	90.5	98
3m Mono	92.4	97.0	99.8
3m LN T cell	No collection	No collection	N/A
3m LN B cell			N/A
3m LN NK cell			N/A
4.5m T cell	98.9	98.9	98.6
4.5m B cell	94.4	97.8	98.2
4.5m NK Cell	99.7	99.2	99.1
4.5m Grans	99.9	90.5	98.7
4.5m Mono	93.0	98.1	98
4.5m CD16+CD56-NK Cell	No collection	No collection	99.8
4.5m CD16-CD56+ NK Cell			96.3
4.5m CD16+CD56+NK Cell			99.0
6.5m T cell	95.2	99.3	Follow-up on going
6.5m B cell	95.5	98.7	
6.5m NK Cell	99.4	99.1	
6.5m Grans	99.9	91.5	
6.5m Mono	99.1	99.6	
6.5m CD16+CD56-NK Cell	N/A	Follow-up on going	
6.5m CD16-CD56+ NK Cell	N/A		
6.5m CD16+CD56+NK Cell	N/A		
6.5m CD11C-CD123+PDC	N/A		
6.5m CD11C+BDCA1-CD16-MDC	N/A		
6.5m CD11C+BDCA1-CD16+MDC	N/A		
9.5m T cell	99.5		
9.5m B cell	98.6		
9.5m NK Cell	98.6		
9.5m Grans	91.6		
9.5m Mono	99.9		
9.5m LN T cell	N/A		
9.5m LN B cell	N/A		

N/A: Cells were sorted, but the post sort purity was not tested due to limited cell number.

Table S2: Antibodies used for flow cytometric sorting of blood, lymph node and marrow cell populations.

Antigen	Conjugation	Vendor	Catalog number	Clone
CD3	APC-Cy7	BD Pharmingen	557757	SP34-2
CD20	PE-Cy5	BD Pharmingen	555624	2H7
CD14	Pacific blue	Invitrogen	MHCD1428	TUK4
CD16	APC	Biolegend	302012	3G8
CD56	PE	BD Pharmingen	555516	B159
CD33	PE	Miltenyi Biotec	130-091-732	AC104.3E3
CD34	PE	BD Pharmingen	550761	563
CD20	APC Cy7	BD Pharmingen	335794	L27
HLA-DR	PE/Texas Red®	Invitrogen	MHLDR17	TU36
CD14	Pacific Blue	BioLegend	301815	M5E2
CD11c	PE Cy7	BioLegend	301608	3.9
CD123	PerCP Cy5.5	BD Pharmingen	558714	7G3
CD16	Alexa 700	BioLegend	302026	3G8
CD1c(BDCA-1)	APC	Miltenyi	130-090-903	AD5-8E7

Table S3: The primers and probes used for barcode PCR, multiplex sequencing PCR, and Southern blot

Primer name	Primer sequences 5'----------3'
MSCV/EF1α index primer forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
MSCV index primer R6	CAAGCAGAAGACGGCATACGAGATATTGGCACGGCATACGAGCTCTTCCGATCT
MSCV index primer R7	CAAGCAGAAGACGGCATACGAGATGATCTGACGGCATACGAGCTCTTCCGATCT
MSCV index primer R8	CAAGCAGAAGACGGCATACGAGATTCAAGTACGGCATACGAGCTCTTCCGATCT
MSCV index primer R9	CAAGCAGAAGACGGCATACGAGATCTGATCACGGCATACGAGCTCTTCCGATCT
MSCV index primer R10	CAAGCAGAAGACGGCATACGAGATAAGCTAACGGCATACGAGCTCTTCCGATCT
MSCV index primer R11	CAAGCAGAAGACGGCATACGAGATGTAGCCACGGCATACGAGCTCTTCCGATCT
MSCV index primer R12	CAAGCAGAAGACGGCATACGAGATTACAAGACGGCATACGAGCTCTTCCGATCT
MSCV index primer R13	CAAGCAGAAGACGGCATACGAGATTTGACTACGGCATACGAGCTCTTCCGATCT
MSCV index primer R14	CAAGCAGAAGACGGCATACGAGATGGAACTACGGCATACGAGCTCTTCCGATCT
MSCV index primer R15	CAAGCAGAAGACGGCATACGAGATTGACATACGGCATACGAGCTCTTCCGATCT
MSCV index primer R16	CAAGCAGAAGACGGCATACGAGATGGACGGACGGCATACGAGCTCTTCCGATCT
MSCV index primer R17	CAAGCAGAAGACGGCATACGAGATCTCTACACGGCATACGAGCTCTTCCGATCT
MSCV index primer R18	CAAGCAGAAGACGGCATACGAGATGCGGACACGGCATACGAGCTCTTCCGATCT
MSCV index primer R19	CAAGCAGAAGACGGCATACGAGATTTTCACACGGCATACGAGCTCTTCCGATCT
MSCV index primer R20	CAAGCAGAAGACGGCATACGAGATGGCCAGACGGCATACGAGCTCTTCCGATCT
MSCV index primer R21	CAAGCAGAAGACGGCATACGAGATCGAAACACGGCATACGAGCTCTTCCGATCT
MSCV index primer R22	CAAGCAGAAGACGGCATACGAGATCGTACGACGGCATACGAGCTCTTCCGATCT
MSCV index primer R23	CAAGCAGAAGACGGCATACGAGATCCACTCACGGCATACGAGCTCTTCCGATCT
MSCV index primer R24	CAAGCAGAAGACGGCATACGAGATGCTACCACGGCATACGAGCTCTTCCGATCT
MSCV index primer R25	CAAGCAGAAGACGGCATACGAGATATCAGTACGGCATACGAGCTCTTCCGATCT
MSCV index sequencing primer	AGATCGGAAGAGCTCGTATGCCGT
EF1α index primer R1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R13	CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R14	CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index primer R15	CAAGCAGAAGACGGCATACGAGATTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
EF1α index sequencing primer	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Rh Albumin-Forward primer	TCGACATGCTCTTTGGTAGC
Rh Albumin-Reverse primer	GGCACGGAAAGATTGAAGAT
Rh Albumin probe	CCAAACGCATCCATTCTACCAACTTG
RRE Forward primer	GGGAGCAGCAGGAAGCAC
RRE Reverse primer	TTGTCTGGCCTGTACCGTCA
RRE Probe	ATGGGCGCAGCGTCAATGACG