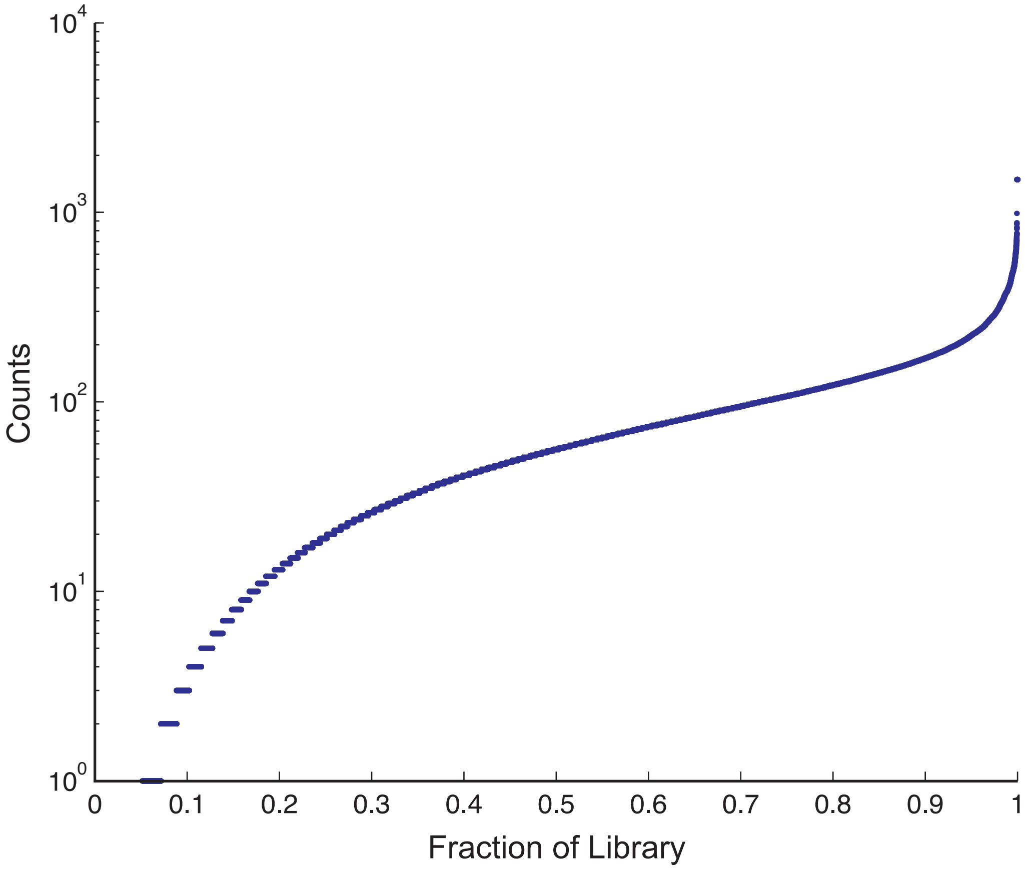
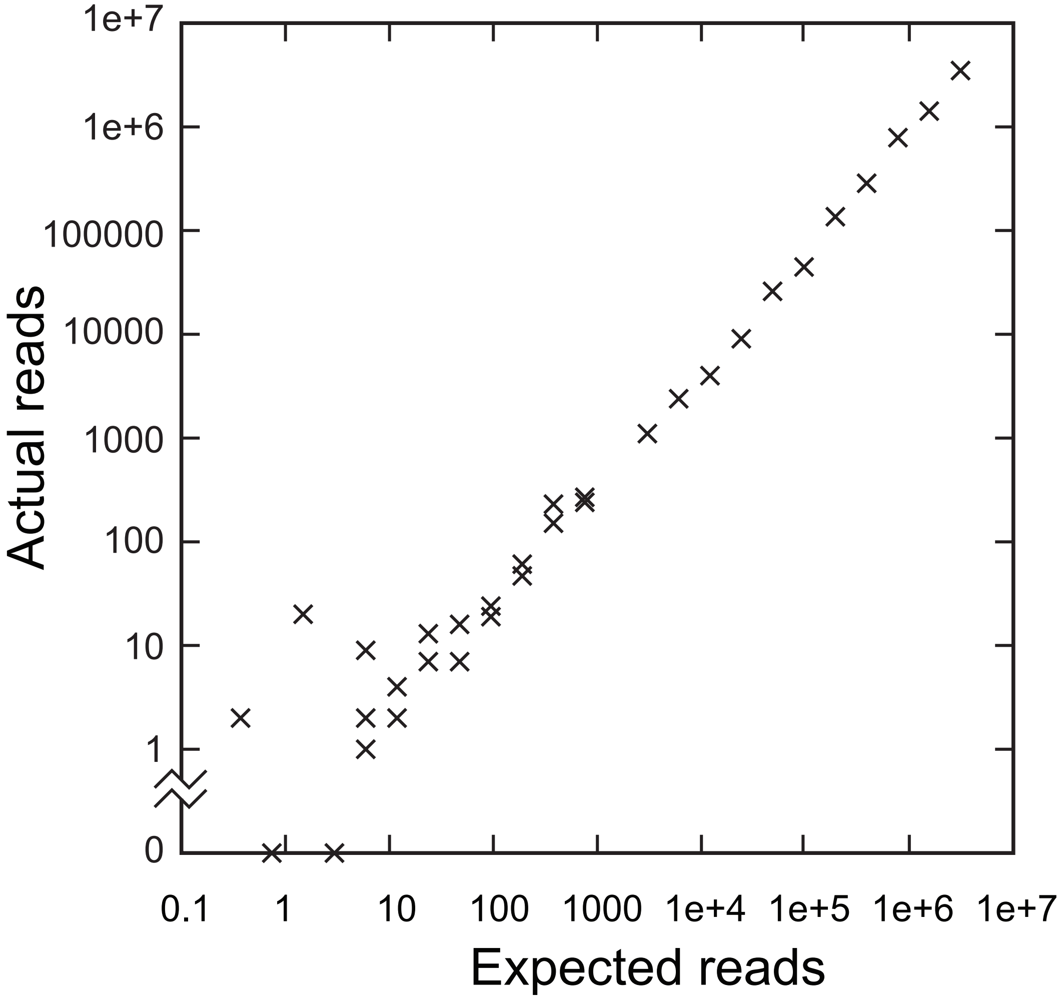
Supplementary Figure 1: Improvements to amplification and determination of library representation by deep sequencing



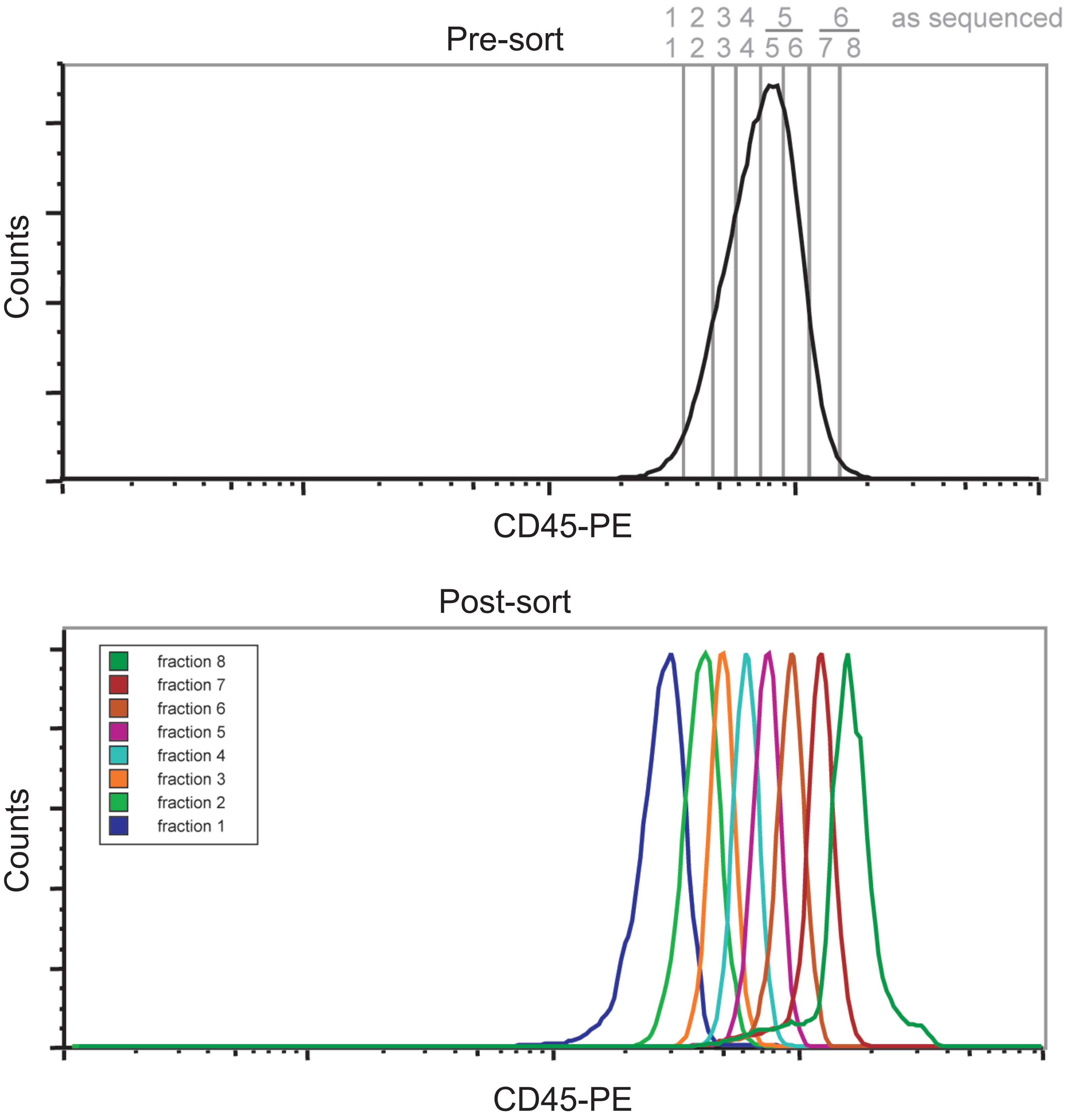
Supplementary Figure 1: Improvements to amplification and determination of library representation by deep sequencing In order to determine the representation of individual elements in the shRNA libraries, we used PCR to add Illumina sequencing adapters and amplify the cloned shRNA libraries. While optimizing this procedure we found that reducing the number of amplification cycles during library creation and using sticky-end restriction sites for cloning greatly reduced the error rate and improved the efficiency of cloning. Further improvements to the analysis of the shRNA libraries after cloning were made by using amplification-Illumina adapter primers which were ~50-100 bp away from the stem of the hairpin. This substantially improved the uniformity of PCR amplification, such that we were able to detect ~95% of expected shRNAs in a single lane of an Illumina flowcell (3.95×10^6 total aligned reads) (Supplementary Fig. 1). In earlier versions of the libraries before optimization of amplification, we only observed ~60% of shRNAs by deep sequencing (~20 shRNAs/gene), although these were still able to function well in selection experiments (CD antigen sorting). It is worth noting that microarray hybridization suggested that early versions of the libraries were quite complete (~95%), and individual 'hits' could easily be picked out when a small number of probes were hybridized, although it is likely that substantial microarray cross-hybridization occurred since these experiments were performed using full hairpin probes which are more likely to self-anneal. The recent use of half-hairpin probes[7-9] should greatly reduce this problem.

Supplementary Figure 2: Accurate counting over a broad concentration range using deep sequencing



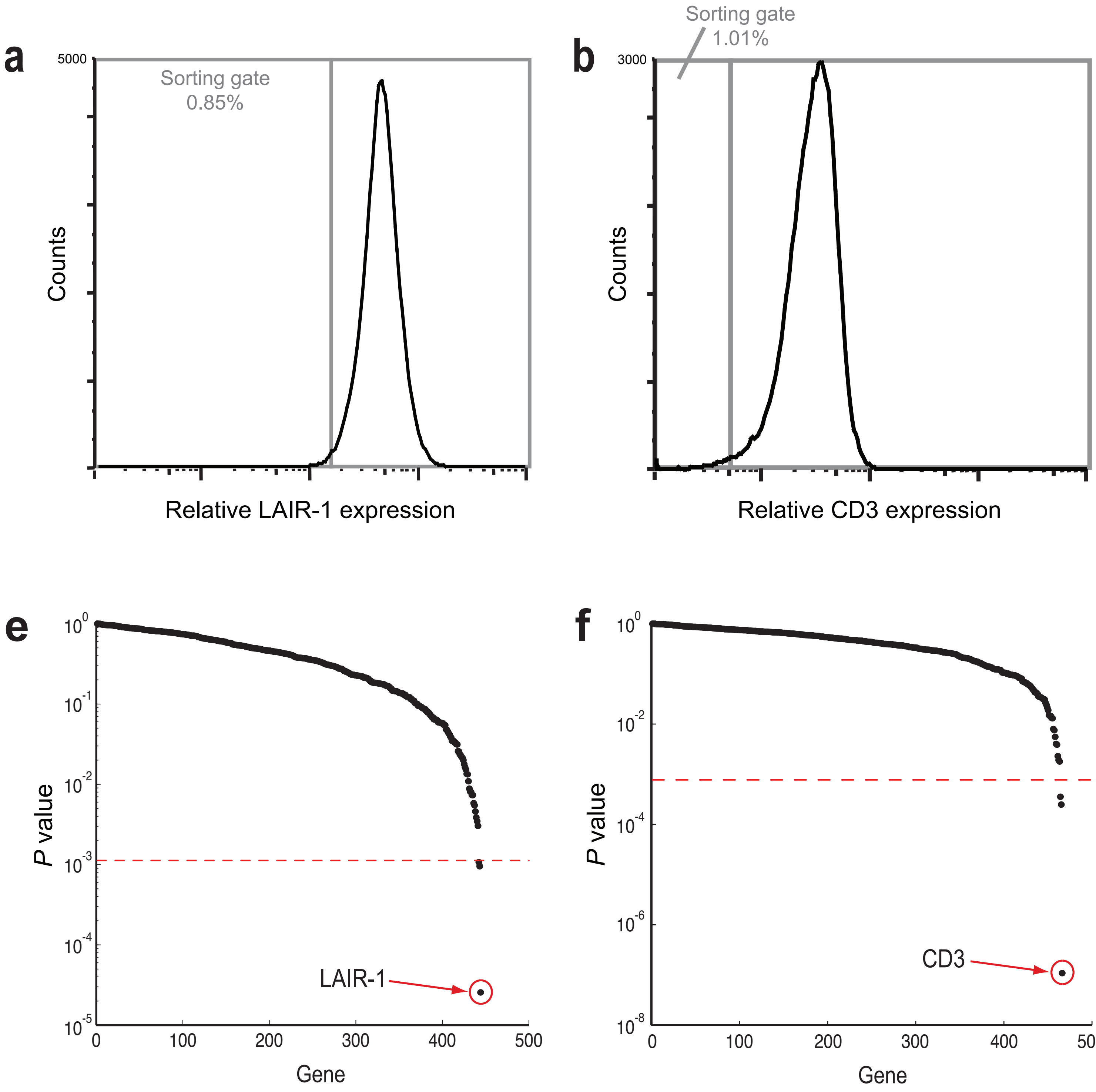
Supplementary Figure 2: Accurate counting over a broad concentration range using deep sequencing
To determine the range of accurate counting by the Illumina Genome Analyzer, 32 unique 28-mer oligos were individually amplified, quantitated, diluted in stepwise series, and subjected to deep sequencing. Note that this analysis was also explored in Supplementary Figure S1 of Ingolia, NT, *et. al*. Genome-Wide Analysis In Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* February 12 (online), 2009.

Supplementary Figure 3: Combination of fractions for sorting, and verification of purity of sorted fractions



Supplementary Figure 3: Combination of fractions for sorting, and verification of purity of sorted fractions
For deep sequencing, Raji B cells were infected at an MOI of ~0.1 with the CD antigen shRNA library, allowed to grow for 7 days, and sorted into the indicated fractions. The sort was performed in 8 fractions, and fractions (5 + 6) were combined, as were (7 + 8), in proportion to the relative cell number collected to fit on a single flow cell with controls. Purity of sorted fractions was verified by re-flowing sorted cells immediately after sorting.

Supplementary Figure 4: Screening for low expression of LAIR1 and CD3



Supplementary Figure 4: Screening for low expression of LAIR1 and CD3
U937 or Jurkat cells were infected with the CD antigen shRNA library, and after 1 week were subjected to flow sorting. mCherry+ cells (expressing the virus) were gated for low LAIR1 expression (**a**) or low CD3 expression (**b**). Percent of mCherry+ cells that were gated is indicated. (**c** and **d**) Following sorting, genomic DNA was prepared from LAIR1-low or CD3-low, mCherry+ cells, and PCR was performed to amplify shRNAs and add the Illumina adapters (see methods). The sorted population was normalized to an unsorted fraction, and the P-value was determined for all genes present in the library. *LAIR1* and *CD3* were the most statistically significant hits in the two independent screens by several orders of magnitude (*LAIR1*, $P = 2.6 \times 10^{-5}$)(**c**), (*CD3*, $P = 1.1 \times 10^{-7}$)(**d**).

Supplementary Figure 5: Enrichment for active shRNAs targeting CD45, and comparison with shRNA prediction algorithm

Percent Knockdown	Algorithm Rank	Sort Fraction (counts)					
		1	2	3	4	5	6
82.74	**9***	**7981**	10	96	2978	3846	1907
80.21	**16**	**0**	0	0	0	0	0
78.86	**17**	**2084**	1	15	281	0	0
71.26	**15***	**19419**	104	1	93	15	158
63.87	**14***	**1139**	10	0	0	0	1
61.19	**4**	**2117**	8	0	0	16	0
59.18	**33**	**276**	2	0	0	0	14
56.90	**1***	**39001**	23	228	1889	1987	1401
48.50	**20**	**608**	0	0	0	0	0
46.23	**26**	**2545**	3	16	465	193	0
41.72	**29**	**354**	3	1	175	142	142
32.50	**24**	**602**	1	0	0	0	0
29.95	**30***	**26515**	332	885	1002	1506	263
27.58	**18**	**363**	2	0	0	0	0
25.56	**22**	**0**	0	0	0	0	0
25.11	**8***	**6472**	63	94	73	264	218
24.74	**6**	**0**	0	0	0	0	184
24.44	**2**	**5579**	444	922	2090	781	931
23.18	**27**	**273**	2	57	16	0	0
16.58	**13**	**87**	3	0	249	0	0
16.03	**10**	**7972**	247	616	532	0	1
15.72	**25**	**1706**	6	13	0	0	0
11.53	**31**	**1043**	168	626	2306	128	697
9.95	**12**	**1126**	0	0	509	0	0
3.85	**21**	**0**	0	0	0	0	0
1.30	**5**	**196**	33	95	177	63	59
0	**3**	**1255**	183	620	1368	0	637
0	**7**	**1741**	285	1874	2070	725	1314
0	**11**	**0**	0	0	0	0	0
0	**28**	**0**	0	0	0	0	0
0	**32**	**669**	15	30	1095	464	579
0	**23**	**445**	3	41	140	42	103
0	**19**	**0**	0	0	0	0	0

Supplementary Figure 5: Enrichment for active shRNAs targeting CD45, and comparison with shRNA prediction algorithm 33 shRNAs targeting the longest isoform of CD45 (CD45R) were designed by a widely-used prediction algorithm (see methods), individually cloned, and tested for activity. shRNAs are ranked by their observed activity (percent knockdown of CD45 expression), and a comparison can be made to the shRNA prediction algorithm rank. Clearly, the shRNAs which were most active were not those at the top of the algorithm ranking, although flow-sorting into fractions based on CD45 expression followed by deep sequencing of shRNAs (see Fig. 3) primarily enriched for the active shRNAs when these were present in the CD antigen shRNA library (compare enrichment for low-CD45 expression in fraction 1 with higher CD45 expression in other fractions). The 6 shRNAs which were identified by single colony sequencing and re-tested (Supplementary Table 3) are highlighted (*). There was no apparent preference for a particular location on the CD45 transcript.

Supplementary Figure 6: Comparison of anti-LAIR1 shRNA activity with shRNA algorithm prediction

Percent LAIR-1 Knockdown	Algorithm Rank
82.31	5
81.28	8
76.00	20
66.19	10
59.32	19
59.11	27
57.47	32
56.60	28
53.16	13
50.72	4
48.68	7
43.45	14
40.96	3
39.86	6
36.71	1
36.05	2
34.64	25
33.90	29
29.83	16
24.58	15
21.67	30
17.01	12
16.25	31
14.24	9
13.93	17
11.19	18
10.52	22
5.78	23
3.59	11
0.2	21
0	24
0	26

Supplementary Figure 6: Comparison of anti-LAIR1 shRNA activity with shRNA algorithm prediction
32 shRNAs targeting the LAIR1/CD305 antigen were individually cloned and tested for activity. shRNAs are ranked by their observed activity (percent knockdown of LAIR1 expression), and a comparison made to the rank given by the shRNA prediction algorithm. As before, there is no clear correlation between predicted rank and efficacy, substantiating the need for a large number of shRNAs in the library.

Supplementary Notes and Tables

Supplementary Note 1: Considerations for performing a pooled screen of a complex library

The ability to quickly construct a high-coverage library that can incorporate changes in vector design, shRNA prediction algorithm, and target preference was a central motivation for the development of our rapid 'clone and use' method. In our hands, PCR, restriction digest, purification, ligation, plating, and maxiprep of the library DNA can be performed in about 1 week. While we are not certain how long the development of commercially available libraries took (e.g., Open Biosystems, Sigma, TRC), the cloning, sequencing, and verifying of each individual element before incorporating it into a 20,000+ element library constitutes an enormous effort of time and resources; these libraries typically have 3-5 shRNAs/gene. Of course, such a library will have the advantage of being free of mutations and distortions. However, our direct use method could be used effectively in several test screens by optimizing amplification and cloning, and normalizing sorted populations to the starting library. Furthermore, readout by deep sequencing allows us to discard imperfect elements from our analysis. The ability to quickly change and adapt these libraries, as well as their greatly expanded size, should make them a useful complementary technology to existing libraries.

In order to maximize representation of the library during our experiments, we maintained > 20× the number of library elements at all stages of use (cloning, library maxiprep, infection, sorting, and re-amplification PCR). Depending on the degree of distortion in the library, this number may need to be increased in order to get accurate quantitation of under-represented species. By normalizing to the starting population (pre-sort experiment), we were able to overcome distortions in the library present after the cloning process and get statistically significant hits primarily for the expected genes.

A number of other considerations may determine choice of library/screening format:

- Microarray analysis costs about $200/array (depending on format), while deep sequencing is currently about $400/lane of an 8-lane flowcell.
- Deep sequencing requires access to an Illumina Genome Analyzer and extensive data processing capacity. The total time for a deep sequencing run and analysis is about 4 days, as compared to 1 day for microarray.
- Different biases may be introduced with sample prep and analysis for each method
- As with any screen, hits will need to be re-confirmed by qPCR and analysis of selected phenotype on a per-gene basis. This will be especially true for genes that are on the edge of statistical signficance.
- Deep sequencing precisely identifies each individual element measured, which is important for determining subtle differences between shRNA species. Although current half-hairpin approaches (see references 7-9) should go a long way to prevent large-scale cross-hybridization, single base

pair mutations and subtle changes in shRNA sequence which may be important for function will be more readily identified using deep sequencing.

- The number of analyzable elements in a deep sequencing run is generally much higher (~80 × 10^6/ flow cell), although number of times an individual shRNA must be counted must be considered. We generally found counts fewer than ~10 to be somewhat less reliable (**Supplementary Figure 2**). This expanded capacity may be important for large-scale screening.

Supplementary Note 2: *P* value determination

The *P* value for CD45 is several orders of magnitude smaller than the *P* values assigned to other genes present in the CD antigen shRNA library. By including a large number of shRNAs (~30) against each gene chosen in our screen we were able to determine a *P* value that describes how likely each gene is to be a hit. The shRNA abundances in fraction 1 were first converted to enrichments by normalizing by the abundances found in the middle fraction, fraction 5. In the case where an shRNA had no reads in the middle fraction, the abundance was normalized by one. This represents a lower bound of how enriched those shRNAs are in fraction 1. We employed a non-parametric statistical test, the Mann-Whitney test, to examine the enrichments of all the shRNAs targeting a given gene in fraction 1 compared to the full collection of enrichments in that fraction. As only a fraction of the shRNAs will effectively knockdown the target gene, the *P* value resulting from this test is diluted by the ineffective shRNA abundances, rendering the test conservative in this case.

The Mann-Whitney test gives a *P* value of 5.2×10^{-7} for the CD45 gene, yet when simulations were performed to determine the frequency of obtaining the abundance ranks that the top three of CD45's 26 shRNAs had obtained or better, the *P* value for the CD45 gene was smaller than 10^{-8} (data not shown). This figure shows the *P* values for all genes in fraction 1 that had more than 5 shRNAs detected in the sequencing run (N = 422). CD45's *P* value is the smallest by orders of magnitude. The dashed red line indicates the *P* value where 0.5 false positives would be expected among the genes analyzed here. Similar tests were performed for the LAIR1/CD305 screen and the CD3 screen by normalizing to an unsorted fraction.

Supplementary Table 1:

			Mismatches:		Deletions:		Insertions:		
Sample	**Total**	**Perfect**	**1-2bp**	**3+bp**	**1-2bp**	**3+bp**	**1-2bp**	**3+bp**	**Perfect (%)**
CD Antigen Library	**122**	**78**	5	12	16	8	3	0	**64**
Library 1	**47**	**38**	1	0	6	2	0	0	**81**
Library 2	**46**	**34**	0	0	11	1	0	0	**74**
Library 3	**277**	**206**	13	10	33	12	3	0	**74**
Library 4	**94**	**68**	5	1	15	5	0	0	**72**
Library 5	**304**	**173**	21	20	76	9	0	5	**57**
Library 6	**216**	**147**	11	6	19	31	2	0	**68**

Supplementary Table 1: Sequence quality table for shRNA libraries. Each library was cloned as described in methods, and the indicated number of individual bacterial colonies were subject to sequencing. Errors were classified as insertions, deletions, or mismatches.

Supplementary Table 2: Sequence quality table for shRNA libraries determined by deep sequencing

Library	Total reads	Perfect reads	Perfect % (26 bp)	Estimated perfect % (52 bp)
CD Antigen	5754912	5058776	87.9	77.3
Library 3	4095374	3632190	88.7	78.6

Supplementary Table 2: Sequence quality table for shRNA libraries determined by deep sequencing

Cloned libraries were amplified with Illumina adapter primers as described in methods and subjected to deep sequencing. For the sequencing run described here, 26 bases of sequence information was obtained, and the number of perfect sequences over this length is reported. Error for the entire hairpin (double the 26 bp length) was then estimated as the square of this error. We should note that this sequencing run itself had an error of measurement of 0.46% per base as estimated by the PhiX control lane (performed according to manufacturer's recommendations), so many of the errors in this table may be due to sequencing error alone. Therefore, the numbers reported here represent an upper bound on the fraction of shRNAs that contain an error consistent with what we observed when sequencing individual shRNAs (Supplementary Table 3).

Supplementary Table 3:

shRNA	Sequence	# Clones	% Knockdown of target
CD45-1	**G**CCCAGAGACTTCCTTCATATAG**TTCAAGAGA**CTATATGAAGGAAGTCTCTGGA	17	82.74
CD45-2	**G**CGGAAATACTCTGGTTAGAAAT**TTCAAGAGA**ATTTCTAACCAGAGTATTTCCA	9	56.90
CD45-3	**G**ATGGCTTAAACTCTTGGCATTT**TTCAAGAGA**AAATGCCAAGAGTTTAAGCCAC	5	63.87
CD45-4	**G**ATGGAAATACTCTGGTTAGAAA**TTCAAGAGA**TTTCTAACCAGAGTATTTCCAG	5	29.95
CD45-5	**G**ATGTCTCCATGACATCAGATAA**TTCAAGAGA**TTATCTGATGTCATGGAGACAG	2	25.11
CD45-6	**G**ATGAATTTGTCTGAATTACATC**TTCAAGAGA**GATGTAATTCAGACAAATTCAC	1	71.26

Supplementary Table 3: anti-*CD45* shRNAs recovered by single-colony sequencing

Genomic DNA was prepared from CD45^{low} cells from the second round of sorting, and the shRNAs from these cells were PCR amplified, cloned, and sequenced. 6 different shRNAs which target *CD45* were identified, and the sequences are depicted. The loop sequence is indicated in bold. The observed frequency for each recovered anti-*CD45* shRNA in single colony sequencing is reported (# clones sequenced), as well as the potency measured for the individual shRNA upon re-testing as a single virus.

Supplementary Table 4: Primers used in this study

Primer Name	Sequence
5' CD antigen oligo	CGCCTGCGAGTCTGGTAT
3' CD antigen oligo	GGAATTCGCCAGCTCGAG
5' CD antigen topo	TGCAGGGGAAAGAATAGTAGAC
3' CD antigen topo	AGTTATGTAACGCGGAACTCC
5' pSicoR-genomic	ATAAATATCCCTTGGAGAAAAGC
3' pSicoR-genomic	GGCGGTAATACGGTTATCCA
5' CD antigen Illumina adaptor	AATGATACGGCGACCACCGACACTCTTTCCCTCCCTTGGAGAAAAGCCTTGTTTG
3' CD antigen Illumina adaptor	CAAGCAGAAGACGGCATACGAATGGATCCTAGTACTCGAG
5' distant CD antigen Illumina adaptor	AATGATACGGCGACCACCGACAGCACAAAAGGAAACTCACC
3' distant CD antigen Illumina adaptor	CAAGCAGAAGACGGCATACGATAATGCATGGCGGTAATACG
Illumina CD antigen sequencing	CACTCTTTCCCTCCCTTGGAGAAAAGCCTTGTTTG
5' Library 3 Illumina adaptor	AATGATACGGCGACCACCGATGGACGAGCTGTACAAGTAA
3' Library 3 Illumina adaptor	CAAGCAGAAGACGGCATACGATGCTCCTAAAGTAGCCCCTTG
Illumina Library 3 sequencing	CACTCTTTCCCTAGTGAAGCCACAGATGTA