

# Rapid creation and quantitative monitoring of high coverage shRNA libraries

Michael C Bassik<sup>1,4</sup>, Robert Jan Lebbink<sup>2,4</sup>, L Stirling Churchman<sup>1</sup>, Nicholas T Ingolia<sup>1</sup>, Weronika Patena<sup>1,2</sup>, Emily M LeProust<sup>3</sup>, Maya Schuldiner<sup>1</sup>, Jonathan S Weissman<sup>1</sup> & Michael T McManus<sup>2</sup>

**Short hairpin RNA libraries are limited by low efficacy of many shRNAs and by off-target effects, which give rise to false negatives and false positives, respectively. Here we present a strategy for rapidly creating expanded shRNA pools (~30 shRNAs per gene) that are analyzed by deep sequencing (EXPAND). This approach enables identification of multiple effective target-specific shRNAs from a complex pool, allowing a rigorous statistical evaluation of true hits.**

Several virus-based shRNA library methods have become valuable tools for conducting RNAi screens (reviewed in refs. 1–3). Microarray synthesis of shRNAs has been used to generate diverse libraries of shRNAs or microRNA-designed shRNAs, which are then cloned, sequence-verified and arrayed into 96-well plate format<sup>2,4,5</sup>. To simplify screening, these barcoded shRNA constructs can be used as pools, and the resulting hits can be identified by recovering and hybridizing the barcodes to a microarray<sup>6–9</sup>.

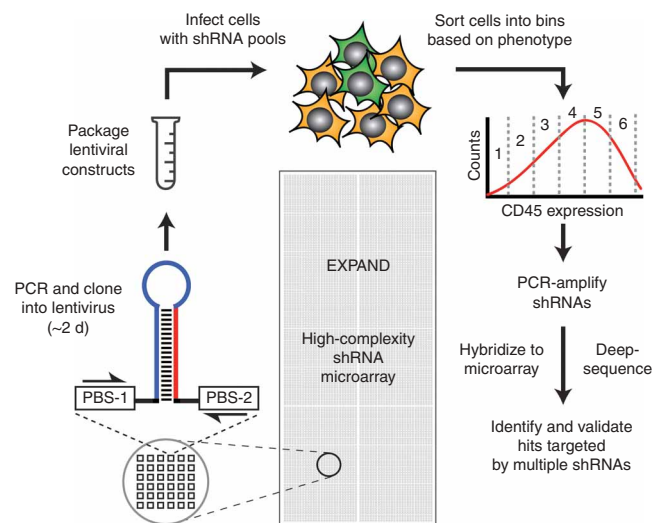
A central shortcoming of existing lentiviral libraries is their low diversity (typically 3–5 shRNAs per gene), which results in high rates of false negatives and false positives. False negatives occur because available algorithms (for example, ref. 10) cannot ensure the presence of effective hairpins specific to a target gene. False positives become problematic because more than one effective shRNA per target must be present to rule out off-target effects. The use of high-diversity libraries would allow identification of multiple potent shRNAs per target (a critical control in RNAi experiments<sup>11</sup>), while increasing sensitivity.

Existing shRNA libraries have used extensive cloning, sequencing and, often, addition of a vector-specific ‘barcode’ sequence to each shRNA in the library, a time-consuming and costly process. Recent improvements in microarray-based oligonucleotide synthesis allow

the production of long oligonucleotides (>100 bp) with an error rate of less than 1/250 bp (data not shown). These oligonucleotides allow a direct clone-and-use strategy that allows easy adoption of changes in RNAi technology, vector choice or assay design.

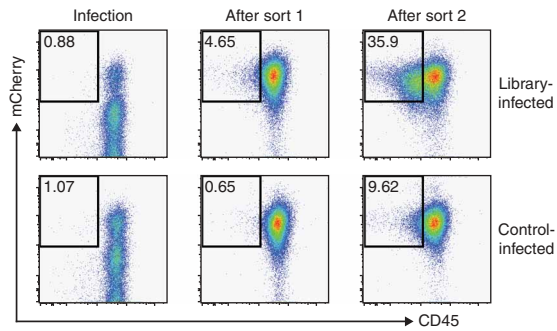
Accordingly, we generated pooled shRNA libraries that (i) have highly expanded per-gene coverage (~30 shRNAs per gene), (ii) are easy to construct and inexpensive to screen, (iii) can be used directly as shRNA pools and (iv) can be readily quantified by microarray and deep sequencing after a screen (Fig. 1). To this end, we designed a pilot library encoding 22,000 shRNAs to target ~600 genes, including nearly all known human CD antigens (CD antigen shRNA library). To maintain the high fidelity and diversity of the input oligonucleotide library, we carefully optimized conditions for PCR amplification, cloning and propagation of the shRNA library (see Online Methods and **Supplementary Note 1** online).

To determine the mutation frequency in the shRNA library, we sequenced 122 random shRNA inserts. Sixty-four percent of the clones were correct (**Supplementary Table 1** online). The errors in the remainder generally consisted of 1- or 2-nucleotide mutations or deletions. Although it is likely that many of the imperfect shRNA sequences retain effectiveness in gene knockdown<sup>12–14</sup>, they can be



**Figure 1** | Schematic for microarray synthesis, cloning and enrichment. Microarrays containing 22,000 shRNAs with ~30 shRNAs per gene are synthesized and are released from the array into a single pool. shRNAs are amplified by PCR using common primer binding sites (PBS-1 and PBS-2), cloned and packaged into lentivirus. Cells infected with the lentiviral shRNA library are sorted by fluorescence-activated cell sorting (FACS) according to phenotype. shRNAs from different fractions are PCR amplified and quantified by either microarray hybridization or deep sequencing.

<sup>1</sup>Department of Cellular and Molecular Pharmacology, California Institute for Quantitative Biomedical Research, and Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, California, USA. <sup>2</sup>Department of Microbiology and Immunology, and University of California San Francisco Diabetes Center, University of California, San Francisco, San Francisco, California, USA. <sup>3</sup>Genomics Solution Unit, Agilent Technologies Inc., Santa Clara, California, USA. <sup>4</sup>These authors contributed equally to this work. Correspondence should be addressed to J.S.W. (weissman@cmp.ucsf.edu) or M.T.M. (michael.mcmanus@ucsf.edu).



**Figure 2** | Raji cells sorted for low CD45 expression are enriched for anti-*CD45* shRNAs. Raji B cells were infected with the CD antigen shRNA library (top panels) or control virus (no insert, lower panels), allowed to grow for 7 d and sorted by FACS for reduced expression of CD45 and expression of mCherry in two rounds (7 d between sorts). Boxes indicate the gates used to sort CD45-downregulated cells. The percentage of mCherry<sup>+</sup> cells showing reduced CD45 expression is indicated in the figure.

identified by deep sequencing and removed from downstream analysis. We created several additional shRNA libraries, with 60–80% correct shRNA sequences (**Supplementary Table 1**).

PCR amplification can lead to a reduction in complexity of an shRNA mixture during the repeated cycling steps. To assess the library complexity, we deep-sequenced PCR-amplified shRNAs. We identified ~95% of the expected shRNAs (**Supplementary Fig. 1** online), with error rates consistent with our previous single-clone measurements (**Supplementary Table 2** online). It is also possible to monitor these libraries using microarrays and improved half-hairpin probes<sup>7–9</sup>. Either approach would allow the direct identification of shRNAs and eliminate the need to independently barcode each vector.

The low error rate and nearly complete shRNA coverage suggested that the libraries could be used directly in RNAi experiments. We infected human Raji B cells with the CD antigen shRNA library and sorted for infected cells (expressing mCherry) that also displayed reduced CD45. The initial comparison of cells 7 d after infection with either the control virus or the shRNA library showed no notable difference during the first sort. However, after two rounds of sorting with cells cultured for 7 d between sorts, we observed a substantial enrichment for mCherry<sup>+</sup>CD45<sup>low</sup> cells as compared to their frequency among cells infected with vector alone (35.9% versus 9.62%) (**Fig. 2**). To identify shRNAs active against *CD45* (*PTPRC*), we PCR-amplified and cloned the lentiviral shRNA inserts from genomic DNA isolated from the CD45<sup>low</sup> sorted cells. Of 83 sequenced shRNA clones, 39 targeted *CD45* (46%). Given that the starting population of the library contained 0.15% *CD45*-targeting shRNAs (33 out of 22,000), this represents an enrichment of > 300-fold.

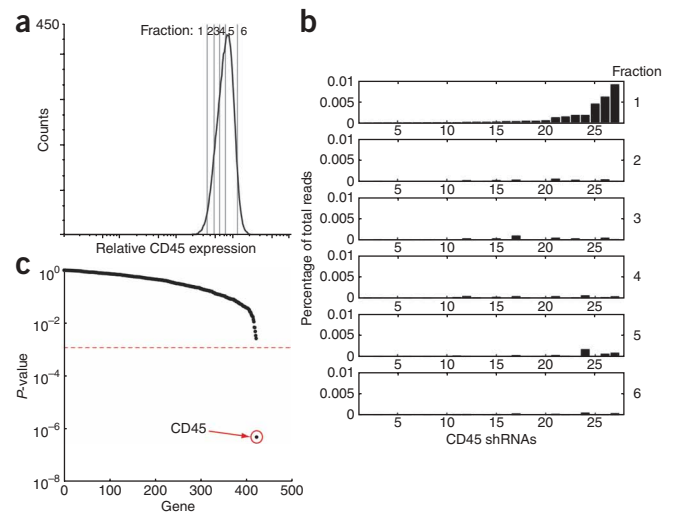
Although other shRNAs were detected in the enriched fraction, we only detected multiple distinct shRNAs for *CD45*, highlighting the power of the expanded library to unambiguously detect ‘hits’ in a single screen. In this initial experiment we recovered six unique *CD45*-targeting shRNAs in the sorted CD45<sup>low</sup> fraction from the total of 33 *CD45*-targeting shRNAs present in the library. These active shRNAs differed markedly in abundance within the sorted fraction (**Supplementary Table 3** online). In general, the shRNAs that were recovered in larger numbers were also more potent when they were individually retested for *CD45* knockdown

(**Supplementary Table 3**). Therefore, the expanded library yielded a diverse population of target-specific shRNAs in a single experiment and allowed us to obtain information about the potency of each individual shRNA from the relative number of each shRNA recovered.

Recent developments in deep sequencing technology make it possible to simultaneously measure the presence of >80 million distinct sequences at a typical read length of ~50–70 nucleotides, ideally suited to monitoring the shRNA sequences in the library described here. A digital readout allows the clear resolution of even very similar shRNA species, which is important for the determination of efficacy of individual shRNAs in high-complexity libraries. In addition, mutant shRNAs can be directly detected and discarded. This should be particularly useful when measuring the loss of shRNAs that cause cell death or slow growth, where the presence of inactive mutant shRNAs would complicate the interpretation of results. Finally, the large capacity of deep sequencing allows the detection of subtle changes in abundance within genome-scale populations of shRNAs.

To evaluate the capacity of deep sequencing to measure sequence abundance over a broad range of concentrations, we performed a dilution series with a known set of 32 oligonucleotides with unique 28-mer sequence tags. We were able to detect a highly linear distribution of oligonucleotide counts over an  $\sim 1 \times 10^6$  concentration range (**Supplementary Fig. 2** online). Sequences that were read less than ~10 times were less reliably measured, but can be accurately measured by simply increasing the sequencing coverage.

The large dynamic range and linearity of this counting approach suggested that deep sequencing could be used to measure the change in abundance of shRNA species at early time points in our test screen. To this end, we used deep sequencing coupled with binned flow-cytometry based sorting to search for shRNAs targeting *CD45*. Human Raji B cells were infected with the CD antigen



**Figure 3** | Binned flow-sorting coupled with deep sequencing can quantitatively resolve active shRNAs after a single sort. (a) Raji B cells infected with the CD-antigen shRNA library were grown for 7 d and sorted into the indicated fractions. (b) Genomic DNA was prepared from cells collected from the fractions in a, and the shRNAs were PCR amplified and deep sequenced. The percentage of total reads for each shRNA in each fraction for *CD45* is depicted; shRNAs are sorted in order of abundance in fraction 1. (c) The abundance of each shRNA in fraction 1 was normalized to its abundance in fraction 5. Only *CD45* gave a significant *P*-value ( $< 2 \times 10^{-7}$ ).

shRNA library and grown for 1 week, after which they were sorted into six fractions representing different levels of CD45 expression (Fig. 3a, Supplementary Fig. 3 online). Genomic DNA was prepared, shRNAs were amplified by PCR, and the abundance of each shRNA in the various fractions was assessed by deep sequencing. Even though a population of CD45<sup>low</sup> cells was undetectable by flow cytometry at this early time point (Fig. 2), we were readily able to measure substantial enrichment of several active anti-CD45 shRNAs in the CD45<sup>low</sup> fractions (Fig. 3b). This included all anti-CD45 shRNAs identified in the previous experiment, which involved two rounds of highly selective sorting performed over several weeks. Moreover, although the shRNAs were present in unequal amounts at the beginning of the experiment, normalization of their abundance in CD45<sup>low</sup> fractions to their abundance in a fraction with high CD45 expression could clearly identify enrichment for active anti-CD45 shRNAs.

The ability to identify multiple active shRNAs specific for each gene is one of the most critical improvements of our approach over existing methodologies, and it is a direct consequence of including 33 shRNAs per gene in the library. Taking into account the full range of shRNAs for each gene, a rigorous statistical test can be performed to differentiate true hits from genes that by chance have one or two off-target shRNAs. This allows the assignment of a *P*-value to every screened gene (rather than to single shRNAs). Using this method, we could distinguish anti-CD45 shRNAs from the rest of the library (Fig. 3c,  $P < 2 \times 10^{-7}$ ; see Supplementary Note 2 online). In contrast, the large majority of shRNAs were not significantly enriched in any fraction. This result was not unique to a particular CD antigen or cell type, as we obtained similar results when sorting for different CD antigens (LAIR1 (CD305) in human U937 monocytic lymphoma cells and CD3 (CD3E) in human Jurkat T lymphoblast cells). LAIR1- and CD3E-specific shRNAs could be clearly resolved from the rest of the library ( $P = 2.6 \times 10^{-5}$  and  $P = 1.1 \times 10^{-7}$ , respectively; Supplementary Fig. 4 online).

To further test the ability of our method to enrich for active target-specific shRNAs, we individually cloned and analyzed the potency of 33 anti-CD45 shRNAs predicted by an algorithm<sup>15</sup> (Supplementary Fig. 5 online). Only ~50% of these had >25% knockdown, a criterion representing minimal activity, and only six had >60% knockdown. In general, although we could see substantial enrichment for active shRNAs after a single sort, there was typically little enrichment for shRNAs with low activity. Nonetheless, the correlation between activity and enrichment was not perfect, possibly because of off-target effects of some shRNAs.

The highly active shRNAs were not restricted to those predicted to be most active by the algorithm; indeed, the most active species were often quite low on the list. We observed similar results when testing shRNAs directed against LAIR1 (Supplementary Fig. 6 online). This analysis illustrates a key advantage of our expanded-coverage library: without including >30 shRNAs per gene, it would have been impossible to predict enough functional shRNAs to corroborate hits. As data accumulate on which hairpins are most active, shRNA prediction algorithms can be improved and library sizes reduced.

In summary, our approach provides an efficient method for rapidly creating and screening shRNA libraries, which addresses both false negative and false positive problems that commonly plague RNAi

screens. With only ~20–30% of predicted hairpins giving >50% knockdown (Supplementary Figs. 5,6), low-complexity libraries will often not have enough shRNAs to corroborate genuine hits in a screen. We show here that high-coverage shRNA libraries can identify many shRNAs targeting a single gene, which increases confidence in hits obtained in RNAi screens. The increased complexity of these libraries can be deconvolved through deep sequencing. We further show that this method allows detection of active shRNA hits in a model screen without extensive selection, sorting and cell proliferation, which will greatly facilitate efforts to identify essential genes whose absence may slow growth or cause cell death. Finally, the direct clone-and-use method provides the flexibility to easily remake libraries to immediately incorporate continual advances in RNAi technology, such as various microRNA contexts for expression and improved algorithms for shRNA prediction.

## METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Note: Supplementary information is available on the Nature Methods website.

## ACKNOWLEDGMENTS

We would like to thank A. Brincat from the Sandler Lentiviral RNAi Core and C. McArthur from the Sandler Asthma Basic Research Center for technical assistance. We would also like to thank Q. Mitrovich and N. Goddard for technical advice, and D. Hirschberg and T. Baxter of Agilent Technologies. This work was supported by a Rubicon grant from The Netherlands Organization for Scientific Research (NWO) to R.J.L. and by a Career Development Fellowship from the Leukemia and Lymphoma Society to M.C.B. M.S. was supported by a postdoctoral fellowship from the Sandler Program in Basic Sciences and is now supported by an US National Institutes of Health K99/R00 (Pathway to Independence) award. N.T.I. is supported by the US National Institutes of Health under a Ruth L. Kirschstein National Research Service Award (GM080853) from the National Institute of General Medical Sciences. This work was supported by a Sandler New Technologies grant to J.S.W. and M.T.M., US National Institutes of Health grant R01 GM80783 to M.T.M. and a grant from the Fight For Mike foundation to J.S.W.

## COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

Published online at <http://www.nature.com/naturemethods/>  
Reprints and permissions information is available online at  
<http://npg.nature.com/reprintsandpermissions/>

- Bernards, R., Brummelkamp, T.R. & Beijersbergen, R.L. *Nat. Methods* **3**, 701–706 (2006).
- Chang, K., Elledge, S.J. & Hannon, G.J. *Nat. Methods* **3**, 707–714 (2006).
- Root, D.E., Hacohen, N., Hahn, W.C., Lander, E.S. & Sabatini, D.M. *Nat. Methods* **3**, 715–719 (2006).
- Cleary, M.A. *et al. Nat. Methods* **1**, 241–248 (2004).
- Silva, J.M. *et al. Nat. Genet.* **37**, 1281–1288 (2005).
- Brummelkamp, T.R. *et al. Nat. Chem. Biol.* **2**, 202–206 (2006).
- Schlabach, M.R. *et al. Science* **319**, 620–624 (2008).
- Silva, J.M. *et al. Science* **319**, 617–620 (2008).
- Luo, B. *et al. Proc. Natl. Acad. Sci. USA* **105**, 20380–20385 (2008).
- Reynolds, A. *et al. Nat. Biotechnol.* **22**, 326–330 (2004).
- Echeverri, C.J. *et al. Nat. Methods* **3**, 777–779 (2006).
- Du, Q., Thonberg, H., Wang, J., Wahlestedt, C. & Liang, Z. *Nucleic Acids Res.* **33**, 1671–1677 (2005).
- Doench, J.G. & Sharp, P.A. *Genes Dev.* **18**, 504–511 (2004).
- Aleman, L.M., Doench, J. & Sharp, P.A. *RNA* **13**, 385–395 (2007).
- Paddison, P.J. *et al. Nat. Methods* **1**, 163–167 (2004).

## ONLINE METHODS

**Amplification, digestion, cloning and preparation of shRNA library.** To design shRNAs against CD antigens, we used the shRNA prediction algorithm<sup>15</sup> freely available through the Hannon lab website (<http://katahdin.cshl.org/homepage/portal/scripts/main2.pl/>). The output shRNA sequences were then modified to have a 9-bp loop (TTCAAGAGA) and common primer binding sites compatible with cloning into the lentiviral vector pSicoR-mCherry, which drives stable expression of the hairpin with a mouse U6 promoter<sup>16</sup>. Primer sequences are listed in **Supplementary Table 4** online. A 22,000-element, 96-bp oligonucleotide pilot library was designed using these sequences and was provided by Agilent Technologies as a pool in a single tube (10 pmol) and dissolved in 200  $\mu$ l water. Improvements to fidelity of long oligonucleotides were made by minimizing the synthetic cycle loss during coupling and deblocking and by minimizing the depurination side reaction.

We identified optimal conditions for amplification of full-length product (96-bp oligonucleotides) by varying the amount of dimethylsulfoxide (DMSO), annealing temperature and extension times for the PCR on a Bio-Rad iCycler. Final conditions were, for a 50- $\mu$ l reaction, 30  $\mu$ l water, 10  $\mu$ l 5 $\times$  Phusion GC buffer (Finnzymes), 1  $\mu$ l 10 mM dNTPs, 5  $\mu$ l 5  $\mu$ M primer mix, 0.025 pmol template, 1  $\mu$ l DMSO, 1 U Hot Start Phusion polymerase (Finnzymes). Cycling parameters were 98  $^{\circ}$ C for 30 s; 15 cycles of 98  $^{\circ}$ C for 10 s and 72  $^{\circ}$ C for 30 s; 72  $^{\circ}$ C for 10 min.

The PCR-amplified oligonucleotides were purified using a nucleotide removal kit (Qiagen) according to manufacturer's recommendations and subjected to restriction digest in a 70- $\mu$ l reaction containing 1  $\mu$ g DNA, 7  $\mu$ l 10 $\times$  NEB buffer 4, 0.7  $\mu$ l of 10 mg ml<sup>-1</sup> BSA and 40 U of *Xho*I and *Mly*I (2  $\mu$ l *Xho*I, 4  $\mu$ l *Mly*I (NEB)), and incubated in a 37  $^{\circ}$ C water bath for 6 h. Digested fragments were verified by electrophoresis on 20% PAGE with 0.5 $\times$  TBE running buffer and purified over a second Qiagen nucleotide removal column before cloning into pSicoR-mCherry.

To prepare the vector for cloning, pSicoR-mCherry was subjected to restriction digest in a 150- $\mu$ l reaction containing 15  $\mu$ l 10 $\times$  NEB buffer 4, 1.5  $\mu$ l of 10 mg ml<sup>-1</sup> BSA, 5  $\mu$ g pSicoR-mCherry and 50 U each of *Xho*I and *Hpa*I (NEB), and incubated overnight at 37  $^{\circ}$ C. Vector was then treated (without purification) with Antarctic Phosphatase (NEB) in a 150- $\mu$ l reaction containing 16.9  $\mu$ l Antarctic Phosphatase buffer (NEB) and 12.5 U enzyme for 4 h at 37  $^{\circ}$ C. Vector was then separated on an 0.8% agarose gel, cut out and extracted from the gel using the QIAquick gel purification kit (Qiagen). Vector was then phenol-chloroform-extracted, ethanol-precipitated and resuspended in 20  $\mu$ l water. A second library type was cloned using primers for mir30-context shRNAs as previously described<sup>4,5</sup>, using the above parameters except that cloning was performed by digestion with *Xho*I and *Eco*RI (NEB).

Ligations were performed in a 20- $\mu$ l reaction containing 500 ng vector, 30 ng insert, 2  $\mu$ l 10 $\times$  ligase buffer (NEB) and 2,000 U T4 DNA ligase (NEB), and were incubated at 16  $^{\circ}$ C for 16 h.

To preserve the diversity of the library, colonies were scraped from twelve 15-cm plates directly after transformation of the ligation mixture (six 100- $\mu$ l transformations of max-efficiency DH5a cells (Invitrogen)). Plates contained about 50,000 colonies each. The collected cell pellet was used for maxiprep (Qiagen) directly, and enough DNA was recovered from the plates for many viral preparations.

**Isolation of CD45, LAIR1 or CD3-targeting shRNAs.** To select shRNAs that specifically target the human *CD45* receptor mRNA, *LAIR1* (*CD305*), or *CD3*, we prepared concentrated virus from the CD-antigen library and control (no insert) pSicoR-mCherry vectors as described previously<sup>16</sup>. We infected  $2 \times 10^6$  human Raji B cells, U937 cells or Jurkat cells (all from American Type Culture Collection) with virus at a multiplicity of infection of 0.1 to ensure single integration sites of the viruses. After 7 d of culture,  $\sim 5 \times 10^7$  cells were incubated on ice for 15 min with 500  $\mu$ l of PBS containing 20% normal mouse serum, 5% BSA and 10% fetal calf serum to block nonspecific interactions. Subsequently, phycoerythrin-conjugated antibody to CD45 (anti-CD45; BD Biosciences), allophycocyanin-conjugated anti-LAIR1 (R&D) or allophycocyanin-conjugated anti-CD3 $\epsilon$  (BD Biosciences) was added and allowed to interact with the cells for 30 min, followed by two washes with Hanks' balanced salt solution supplemented with 2% fetal calf serum. mCherry-positive (as a marker for infected cells) and CD45- or LAIR1-reduced cells (to select potential active anti-*CD45* or anti-*LAIR1* shRNAs) were sorted by using a MoFlow cell sorter (Dako Cytomation). After a week of culture the same procedure was repeated, and the cells were cultured for another 7 d. Note that for the binning plus deep sequencing experiment, cells were infected and sorted after 7 d. For further details of binning and sorting, see **Supplementary Figure 1**. All of the sequences for the CD antigen shRNA library as well as the effective shRNAs targeting *CD45* and *LAIR1* are included as **Supplementary Data 1–3** online.

To retest anti-*CD45* shRNAs after the two-sort experiment, genomic DNA was isolated from the sorted cell fractions, and lentiviral shRNA integrations were amplified by PCR. They were then cloned into the pCR2.1-TOPO vector (Invitrogen) and sequenced. Identified *CD45*-targeting shRNAs were subcloned into pSicoR-mCherry, packaged into lentivirus and individually validated for their efficiency in *CD45* knockdown, essentially as described above. Analysis of cells after viral infection and surface antigen staining was performed on an LSR II flow cytometer (BD Biosciences) and the FACS data were analyzed by FlowJo software (Tree Star). Percentage knockdown was calculated by ((geometric mean of uninfected cells – geometric mean of infected cells)/geometric mean of uninfected cells)  $\times 100$ .

**Amplification of shRNA pools from genomic DNA.** shRNAs were amplified from genomic DNA in a 50- $\mu$ l PCR reaction consisting of 30  $\mu$ l water, 10  $\mu$ l 5 $\times$  Phusion GC buffer, 5  $\mu$ l 5  $\mu$ M primer mix, 1  $\mu$ l 10 mM dNTPs, 1.5  $\mu$ l DMSO, 750 ng genomic DNA and 1 U (0.5  $\mu$ l) Phusion polymerase. Cycling parameters were 98  $^{\circ}$ C for 30 s; then 25 cycles of 98  $^{\circ}$ C for 30 s, 56  $^{\circ}$ C for 15 s, 72  $^{\circ}$ C for 15 s; then 72  $^{\circ}$ C for 10 min. In some cases many PCR reactions were pooled on a Minelute column (Qiagen) before electrophoresis on 20% PAGE with 0.5 $\times$  TBE running buffer, electroelution and concentration on a second column.

**Deep sequencing and quantitation of read accuracy.** Genomic DNA was prepared as before, and shRNAs were amplified as before except that sequences compatible with annealing to the Illumina flow cell were added. Sequencing was performed according to manufacturer's protocols (Illumina).

For assessing linearity of sequence counting over a dilution series, 31 68-mer oligonucleotides consisting of a unique 28-nt tag flanked by 21-nt and 19-nt common Illumina primer binding

sites were individually amplified by PCR and purified from an acrylamide gel. The dsDNA products were quantified using a BioAnalyzer (Agilent) and pooled using a dilution strategy designed to give a broad range of expected sequence tag concentrations. This

pool was subjected to deep sequencing, and reads were aligned with the tag library, allowing up to three mismatches against the tag.

16. Ventura, A. *et al. Proc. Natl. Acad. Sci. USA* **101**, 10380–10385 (2004).

