Supporting Online Material

Materials and Methods

Cell Lines and Virus Production

HMECs from a reduction mammoplasty were purchased from Lonza, immortalized with human telomerase, and maintained in MEGM (Lonza). HCC1143, HCC1954, HCC1937, T47D, RWPE, DU145, PC3, and LNCaP cells were purchased from the ATCC and maintained as directed. MSCV-PM-shRNA retroviruses were produced as previously described (*20*). Viral supernatants were stored at -80^{o}C prior to use.

Vectors and Cloning

Retroviral shRNA libraries used for the primary screens and individual shRNAs were cloned into MSCV-PM as previously described (*20*). The shRNA sublibrary for the validation screen was synthesized using parallel microarray synthesis (Agilent) followed by PCR-amplification of chip material using the forward primer siRNAF: CTAATTGATCTTCTCGAGAAGGTATATTGCTGTTGACAGTGAGCG and the reverse primer GFPr: TTTAAAGAATATACGCGTCCGCGTCGATCCTAGG. shRNAs were cloned into the XhoI/MluI sites of MSCV-PM-Mir30-PheS. The 3' mir30 sequence from MSCV-PM was added to this library using the FseI/EcoRI sites. Individual shRNAs from the validation screen for follow up experiments were synthesized as 100 bp oligos (IDT) and cloned into the XhoI/EcoRI sites of MSCV-PM-PheS. The sequences of individual shRNAs can be found in Table S6. All cloning was performed with enzymes from New England Biolabs, Qiaquick Gel Extraction kits from Qiagen, and DH5α bacteria from Invitrogen.

Pilot shRNA proliferation screen

A previously described (*20*) focused set of the second generation Elledge-Hannon human shRNA library consisting of 8203 shRNAs targeting 2924 kinases, phosphatase catalytic and regulatory subunits, ubiquitin-proteasome pathway genes, and genes implicated in cancer was used for a pilot screen. HMECs were transduced with the MSCV-PM-shRNA focused retroviral library in triplicate using 8μg/ml polybrene (Sigma) with an average representation of ~1000 cells per shRNA at an MOI of 2 in one pool. Viruses were removed following an overnight incubation. Initial reference samples of at least 1000 cells per shRNA were collected 72 h post-infection. The remaining cells were puromycin-selected (2 μg/ml) and propagated with a representation of ≥ 1000 cells per shRNA maintained at each passage. End samples were collected after 8 population doublings (PDs). Genomic DNA was harvested from initial and end samples. Half-hairpin barcodes were PCR-amplified, labeled, and hybridized to microarrays as described below. Screen data were analyzed using the method of significance analysis of microarrays (SAM) (*27*) to identify shRNAs that were consistently enriched across triplicates with a false-discovery rate of 20% and an average microarray signal greater than or equal to two-fold above the mean. Enriched shRNAs were defined by an average log2 ratio cutoff of two-fold.

Genome-wide shRNA proliferation screens

Genome-wide shRNA proliferation screens were performed as above for the pilot screen using the entire second generation Elledge-Hannon human shRNA library (*19*) which was previously cloned into MSCV-PM (*20*) and consists of a total of 74,905 shRNAs targeting 19,011 genes in six independent pools of approximately 12,800 shRNAs per pool. HMECs were screened for STOP genes, while HMEC, HCC1937, RWPE, DU145, PC3, LNCAP, HCC1954, HCC1143, and T47D cells were screened for essential GO genes. We harvested initial cell samples shortly after library infection and end samples after allowing the cells to undergo 8 population doublings (PD) while maintaining a representation of ≥ 1000 copies of each shRNA during passaging. Following harvesting of genomic DNA from the initial and end samples, half-hairpin barcodes were PCR amplified, labeled, and competitively hybridized to microarrays as described below to determine the abundance of each shRNA before and after the 8PDs. Screen data were analyzed as above to identify shRNAs that were consistently enriched across triplicates with a false-discovery rate of 5% and an average microarray signal greater than two-fold above the mean. Enriched shRNAs targeting STOP genes were defined by an average log2 ratio cutoff of two-fold (pools 1-3) or 1.8-fold (pools 4-6). Essential shRNAs targeting GO genes were defined by significant depletion in 5 of the 9 screens with average log2 ratio cutoffs of 1.5-fold (HCC1937, RWPE, DU145, PC3, LNCAP screens) or two-fold (HMEC, HCC1954, HCC1143, T47D screens) depletion with a false discovery rate of 5%.

Validation screen using a sublibrary

A sublibrary was designed against 1550 high confidence genes from the genome-wide and pilot proliferation screens containing approximately 12 new shRNAs per gene. Also included in this library were all shRNAs which scored in the primary screen against these 1550 genes and 100 negative control shRNAs targeting FF and EGFP. A total of 21,768 shRNAs were synthesized by parallel microarray synthesis on a custom microarray (Agilent). Following shRNA design and synthesis, we determined that this library targets a total of 1796 genes in the human genome (hg19). More genes are targeted than the library was originally designed against because some shRNAs target multiple genes. The screen was performed as described above for the pilot and genome-wide proliferation screens, however, the pools were deconvolved using next-generation Illumina sequencing. Sequencing reads were annotated to the shRNA library design file, and counts for each shRNA in a given sample were normalized to the total reads per sample. The average log2 ratio of each shRNA's abundance in the end to the initial samples was calculated across triplicates and normalized to the mean of 50 negative control shRNAs targeting FF. Screen data were analyzed with SAM to identify shRNAs that were consistently enriched across triplicates with a FDR = 5%. Enriched shRNAs were defined by an average log2 ratio cutoff of two-fold greater than the mean of the 50 negative control shRNAs targeting FF.

Genomic DNA preparation

Genomic DNA was harvested from initial and end samples by incubating in 10 mM Tris-HCl pH 8.0, 10 mM EDTA, 0.5% SDS, and 0.2 mg/ml proteinase K overnight at 55^{o}C, followed by the addition of 0.2 M NaCl. Genomic preps were phenol-chloroform extracted using Phase-Lock tubes (Fisher), treated with 25 μg/ml RNAse A at 37^{o}C for 3 h, phenol-chloroform extracted again, ethanol precipitated, and resuspended in 10 mM Tris-HCl pH 8.5.

Half-hairpin barcode PCR, probe labeling, and microarray hybridization

Half-hairpin barcodes from the primary screens were PCR-amplified from genomic DNA using the forward primer JH353F: TAGTGAAGCCACAGATGTA and the reverse primer BC1R: CCTCCCCTACCCGGTAGA, and 2 μg of the PCR product was Cy3- and Cy5-labeled as previously described (*20*). Custom microarrays of half-hairpin probes were synthesized by Nimblegen at a density of 13,000 probes per subarray for each genomic or pilot screening pool, and labeled PCR product was hybridized as previously described (*20*).

Illumina Sequencing

Half-hairpin barcodes from the validation screen were PCR-amplified from 30 μg genomic DNA in a reaction containing 200 μM dNTPs, 500 nM each of JH353F and BC1R primers, 1x PRIMESTAR GC buffer, and 50 units of PRIMESTAR HS (Takara). PCR was performed with the following program: 98°C 4 min, 29 cycles of 98°C 10 sec, 53°C 5 sec, 72°C 25 sec. PCR products for each replicate were pooled, isopropanol precipitated, resuspended, and gel-purified using QiaQuick Gel Extraction columns (Qiagen). Illumina sequencing adapters were added by a second round of PCR using 500 ng DNA template, and the forward primer p7+Loop: CAAGCAGAAGACGGCATACGATAGTGAAGCCACAGATGTA and the reverse primer p5+mir3: AATGATACGGCGACCACCGACTAAAGTAGCCCCTTGAATTC. Each reaction contained 200 μM dNTPs, 2 μM of each primer, 1x Accuprime buffer, 5% DMSO, and 5 units of Accuprime pfx (Invitrogen). PCR was performed with the following program: 95°C 2 min, 5 cycles of 95°C 15 sec, 50°C 30 sec, 68°C 30 sec, 27 cycles of 95°C 15 sec, 56°C 30 sec, 68°C 30 sec, and a final extension at 68°C for 2 min. PCR products for each replicate were pooled and gel-purified using QiaQuick Gel Extraction columns (Qiagen). Samples were sequenced with the primer mir30-EcoRI: TAGCCCCTTGAATTCCGAGGCAGTAGGCA.

Cancer Cell Line Analysis to Determine Homozgyous Deletion Frequencies

Copy number profiles of 611 cancer cell lines described previously (*13*) were examined to determine homozygous and hemizygous deletion frequencies per gene. Copy numbers were normalized to a median of 2 for each cell line. Deletions were defined as normalized copy-numbers of 0-0.5 (homozygous deletion) and 0.5-1.25 (hemizygous loss), corresponding to copy-number modes observed across all cell lines (Fig. S6). The hemizygous and homozygous deletion frequencies for each gene were calculated as the fraction of cell lines exhibiting hemizygous or homozygous deletions at that gene; these were summed to determine the total deletion frequency. The frequency of homozygous deletions for genes located in recurrent focal deletion peaks was determined as the average homozygous deletion frequency across all genes within the peaks.

Comparison to Cancer Sequencing Data

Cancer sequencing information used for comparison was published previously and can be found online at COSMIC (*16*). 526 tumors in the COSMIC database that were previously analyzed using whole genome sequencing were examined for homozygously inactivated genes where a nonsense or frameshift mutation was observed and the second allele was lost by deletion or reduction to homozygosity in the same tumor. Furthermore, we examined data from these 526 tumors to compile a set of genes containing a nonsense or frameshift mutation, independent of zygosity. We mapped these gene sets to deleted regions derived from a cross-tumor analysis of 3,131 cancer samples and to STOP gene sets from our primary and validation screens. For each comparison, we computed the statistical significance of the observed enrichment/depletion using Fisher's Exact Test.

Comparison to Cancer Deletion Regions

Cancer deletion peak information used for comparison was published previously (*13*) and can be found online at www.broadinstitute.org/tumorscape. We mapped all genes scoring as significantly enriched in our primary and validation shRNA screens to a set of significantly deleted regions derived from a cross-tumor analysis of 3,131 cancer samples. Genes from our primary and validation screen candidate STOP and GO lists (annotated with hg19 gene symbols) that were not mapped to the human genome version hg18 were excluded from the comparisons. For each comparison, we first computed the statistical significance of the observed enrichment/depletion using Fisher's Exact Test. To confirm these results, we next evaluated the fraction of screen hits residing at least partially within a deletion peak to the expected value as estimated through 1,000 random permutations of all deletion peaks across the genome. These permutations were performed by first circularizing the genome, removing the genes in existing deletion regions to prevent re-sampling, and shifting each deletion peak by a random offset; when a peak crossed a chromosome, it was split into two peaks. The size of each permuted peak was adjusted so that the number of genes contained in the entire permuted dataset was equal to the number of genes contained in the original deletion set. The p-value for enrichment of STOP genes/depletion of GO genes was then calculated as (1+ # (permutations exceeding observed enrichment/depletion))/(# permutations).

MCA Proliferation Assays

MCAs were performed as described (*21*), in which equal numbers of non-color HMECs expressing candidate shRNAs or FF and GFP-positive HMECs expressing FF were mixed and plated in triplicate, grown for 5 days, and assessed by FACS analysis. Assays were normalized to the percent non-color FF HMECs competed against GFP-positive FF HMECs.

Supplemental Text

Supplemental Results

Cancer genomes show significant rewiring due to genomic instability and mutagenesis that drives in tumorigenesis. Although copy number alteration studies have discovered candidate oncogenes in recurrent amplifications, putative tumor suppressor gene identification has proven more elusive, often because deleted regions encompass so many genes. The two-hit hypothesis suggests that two mutations in the same gene are required for tumorigenesis, indicating a recessive disease. However, there are now many examples of haploinsufficient tumor suppressors such as *CHK1, ATR, PTEN, CDKN1B and DICER* in which loss of only a single allele contributes to tumorigenesis (*9-11*)**.** Current models, however, do not explain why large regions of recurrent deletions containing many genes are found in most human cancers (*12, 13*). Importantly, the vast majority of these deletions are hemizygous, as copy number analysis of over 700 cancer cell lines recently showed (*12*). Whether multiple genes within a hemizygous deletion region contribute to the tumorigenic phenotype or are simply a byproduct of genomic instability during malignant evolution that targets a single, recessive or haploinsufficient tumor suppressor gene remains to be determined.

Examination of recurrent focal deletion peaks across cancers. A thorough copy number analysis of over 3000 tumors revealed 82 regions of recurrent focal deletion with an average of 6 per tumor, and some cancers have an average of 10 or more of these focal deletions per tumor (Fig. 1A). A key question arising from these studies is if the genes lost in these recurrent deletions are driving the cancer phenotype and, if so, how.

Many altered processes can promote tumorigenesis including those that delay differentiation, induce angiogenesis, increase invasion, bypass senescence, promote anchorage independent survival, and increase proliferation. Of these categories, proliferation is likely to contain the most genes as proliferation is integrated into all developmental decisions. Since many pathways can alter proliferation and subtle changes in proliferation rates can have profound effects on tumor cell fitness and lead to clonal selection, we sought to investigate whether recurrent deletion regions found in cancers were affecting regulators of cellular proliferation. Our previous studies have shown that it is much easier to either stimulate or inhibit the proliferation of normal cells than it is alter proliferation of a cancer cell using in vitro tissue culture shRNA proliferation screens. This would be expected if tumor cells have already undergone genetic changes that optimize proliferation and resistance to killing. The interpretation of this observation is that many of the same changes that promote the proliferation of normal cells have already occurred in most cancer cells. Thus, identifying these proliferation-relevant genes in normal cells could allow us to uncover which alterations in cancer cells are driving proliferation. This is, in part, based on the assumption that genes that act to restrain proliferation in tissue culture will behave the same way in tumors *in vivo*.

As discussed in the text, we define proliferation regulators as falling into two broad categories, STOP genes that restrain proliferation and GO genes that promote proliferation. There are 2 classes of GO genes that can be distinguished by their phenotypes when their levels are altered.

Class 1 GO genes are required for optimal proliferation but only display a phenotype when levels are reduced. This class includes essential genes required for cell viability and includes many factors involved in basic cellular functions such as replication, transcription, translation and the cell cycle. Conversely, Class 2 GO genes, whose overproduction increases proliferation and survival, are often implicated in cancer and encode such notable proteins as cyclins D1, D2, and D3 and a variety of growth factors and their receptors (*28*, *29*). Depending on the specific tumor, i.e. if the gene is amplified or overexpressed, reduction of Class 2 GO gene activity can reduce proliferation.

Genome-wide shRNA screen to identify regulators of cell proliferation. As many tumor suppressors and oncogenes control cell proliferation and survival in cell culture, proliferation is a good model for certain aspects of tumorigenesis. For this reason, we were interested in whether recurrent deletions in cancer may be driven, in part, by the loss of proliferation regulators. To identify candidate STOP genes that constrain normal cell proliferation and survival, we performed an shRNA proliferation screen as described in the text in HMECs derived from a reduction mammoplasty (Figure S2A). We harvested initial cell samples shortly after library infection and end samples after allowing the cells to undergo 8 population doublings (PD) while maintaining a representation of $\geq$ 1000 copies of each shRNA during passaging. Following PCR-amplification of half-hairpin barcodes from genomic DNA of initial and end samples, amplicons were labeled and competitively hybridized to a microarray to determine the abundance of each shRNA before and after the 8 PDs. By comparing the log2 ratio of each shRNA's abundance in end vs. initial samples, enriched shRNAs were identified that increase normal cell proliferation (Figure 2A, red). Screen data were analyzed as previously described (*20*) using significance analysis of microarrays (SAM) to identify shRNAs that were consistently enriched $\geq$ 1.8-fold across triplicates with a false-discovery rate of 5%, which represents a minimum 15% increase in proliferation and/or survival per generation. Approximately 15% of the STOP genes identified (537) were enriched with multiple shRNAs.

To validate growth regulators, we tested individual shRNAs using a multi-color competition assay (MCA) (*21*), in which equal numbers of GFP negative control cells expressing an shRNA targeting firefly luciferase (FF) and non-color cells expressing a candidate shRNA were mixed, grown for 5 days, and assessed by FACS analysis (Fig. S2B). The percent of cells expressing each candidate shRNA was normalized to FF and plotted for 77 independent shRNAs enriched in the proliferation screen and compared to that of positive control shRNAs targeting p21 and Rb (Fig. S2C). Even with this very short assay, we observed a validation rate of approximately 51%, demonstrating that our screen identified genes that constrain proliferation. This is likely to be an underestimate as weaker hairpins whose phenotypes depend on integration into optimal expression environments are likely to only score in a longer assay which gives the subset of highly penetrant shRNAs time to emerge and take over the culture (see below).

Generation of a sublibrary to validate high-confidence STOP genes. We designed, synthesized, and cloned a new retroviral shRNA sublibrary against a higher-confidence gene list of 1555 genes as described in the text containing 21,768 shRNAs with approximately 12 additional shRNAs per gene (Fig. 2C). Enriched shRNAs from the primary screens that target these genes and negative control shRNAs were also included in the sublibrary. We performed a secondary validation screen in triplicate using this shRNA sublibrary similar to that shown Fig.

S2A and deconvolved the starting and end samples using high-density Illumina sequencing. STOP genes validating with multiple shRNAs are more likely to be true on-target regulators of cell proliferation, and we identified many STOP genes validating with 8 or more shRNAs. Most of these shRNAs (7 targeting p53 and 8 targeting Rb) increased proliferation ≥ 4-fold, compared to only 8% of the negative control FF shRNAs. Together, these data indicate that the validation screen can distinguish between authentic regulators of cell proliferation and false-positive genes that scored in the primary screen due to off-target effects or other sources of variability.

Many known TSG pathways enriched in the validation screen (Fig. 2F). For example, shRNAs targeting TGF-beta pathway genes, whose activation results in increased expression of the CDK inhibitor p21 and cell cycle arrest, were enriched in the screen. Other regulators of the G1/S transition validated in the validation screen as well, including several other CDK inhibitors (p27, p21, p57), GSK-3β, which targets cyclin D1 for degradation, and inhibitors of E2F transcription factors (Rb and p107). Depletion of several regulators of the p53 pathway (p300, p53, Gadd45b), apoptosis (Apaf1, Caspase 3 and Caspase 6), and the spindle checkpoint (securin) also increased cell proliferation with multiple shRNAs in the validation screen.

The STOP gene set is highly enriched for genes frequently deleted in tumors. A cursory examination of our validation screen hits revealed numerous tumor suppressor genes whose depletion with ≥ 4 shRNAs increased cell proliferation ≥ 4-fold as described above. These established tumor suppressors are mutated or deleted in a variety of cancers and include *EP300* (colorectal, breast, pancreatic and other cancers), *FBXW7* (T-ALL, colorectal, and endometrial cancers), *NF2* (neurofibromatosis, meningioma, and renal cancer), *BRCA2* (breast and ovarian cancers), *RB1* (retinoblastoma, sarcoma, breast and small cell lung cancers), *SMAD4* (pancreatic and colorectal cancers), and *TP53* (breast, colorectal, lung and other cancers) (Fig. S4B). The presence of these known tumor suppressor genes in our STOP gene sets suggests our cell proliferation screens are likely to have identified novel tumor suppressors and that cell proliferation in this HMEC system is relevant to tumorigenesis.

The fact that more STOP genes exist in regions of recurrent focal deletion than expected as described in the text indicates that the hemizygous recurrent focal deletions found in cancers contain more proliferation suppressors than one would expect randomly and suggests that single copy loss of these regions may increase proliferation. Parallel analyses using the frame-shift and nonsense mutation gene set or a missense mutation gene set did not reveal enrichment for such loss-of-function mutants in regions of recurring deletion, further suggesting that many of these recurring deletions are unlikely to be acting through classical two-hit mechanisms of tumorigensis.

Optimization of STOP and GO gene densities in recurrent cancer deletion regions. Sixty-six of the 82 cancer deletion peak regions (80.5%) contain at least 1 STOP gene (Table S9). The 16 deletion peak regions without a STOP gene are gene-poor deletions, often containing one large, non-essential gene and containing a total of only 47 genes with an average of only 2.9 genes per deletion, compared to an average of 23.6 genes per deletion across the entire deletion set. Analysis of the regions containing STOP genes provided several findings of interest. Multiple STOP gene candidates are present in each deletion, with STOP genes comprising 22.6% of the genes in these regions. Even for mid-sized deletions containing a bona fide tumor suppressor gene like *RB1*, *PTEN*, *TP53* and *SMAD4*, multiple STOP genes are present in each deletion. In the *RB1* focal deletion region, 5 of 9 genes are STOP genes. For these 4 gene regions containing a known tumor suppressor, 32.4% of the genes are STOP genes, suggesting that these deletions might have greater selective advantage than point mutation loss of the tumor suppressor alone. Also, if the TSG is truly recessive, the presence of an adjacent, haploinsufficient STOP gene would help clonal selection if both genes were deleted. Furthermore, some deletions have very high percentages of STOP genes. For example, STOP genes comprise 32.4% of the 76 genes in the chr6:76686994-105211031 deletion, 31.4% of the 86 genes in the chr8:43190171-72909145 deletion, 29% of the 86 genes in the chr6:101022997-121353987 deletion, and 27% of the 140 genes in the chr5:85995953-133474474 deletion. These observations suggest that cancer cells may select for hemizygous deletion of more densely clustered STOP genes in a single deletion event because the decreased gene dosage of multiple tumor suppressors provides a competitive advantage.

To test the hypothesis that deletions might avoid essential Class 1 GO genes, we assembled an *in silico* list of high probability essential Class 1 GO genes involved in critical cellular processes as annotated by KEGG pathway analysis that include basal transcription and RNA polymerase, the spliceosome, the ribosome, DNA replication, fatty acid biosynthesis, amino-acyl tRNA synthesis, and mRNA transport, encompass 473 genes that we had mapped to chromosomal locations. This gene set demonstrated a significant depletion ($p = 0.012$) from recurrent deletion regions with 28.3% fewer genes present in deletion regions than expected as described in the text (Figure 4A).

We confirmed this using an experimentally derived set of Class 1 GO genes whose depletion limit proliferation and survival. It should be noted that we did not attempt to identify Class 2 GO genes, as that would require a cDNA overproduction screen. Since both normal and cancer cells are dependent on these essential GO genes, we analyzed data from our HMEC proliferation screen as well proliferation screens on 1 normal prostate epithelial (RWPE), 4 breast cancer (HCC1143, HCC1937, HCC1954, and T47D), and 3 prostate cancer cell lines (DU145, PC3, LNCAP) for shRNAs that reduced cell proliferation and viability $\geq$ 1.5 fold in 5 of the 9 cell lines examined (Table S11). The STOP and GO screens were performed with the same shRNA library, and occasionally there is a conflict in which one shRNA for a gene enriches while another shRNA is depleted, resulting in a few genes that are part of both the STOP and GO lists. In these cases, we removed the conflicted gene from both lists. None of the individual shRNAs appeared on both lists. Even with the conflicting genes removed, the STOP gene permutation analysis remains statistically significant (Fig. S5B, C). The Class 1 GO gene set is enriched for functions in cellular machinery as described in the text. When we examined the location of GO genes within recurrent deletions, we found that more than half (58.5%) of deletion regions contained zero GO genes. Furthermore, only 5.2% of the genes in the peaks containing STOP genes were GO genes (98 of 1896 genes). In contrast to STOP genes, permutation analysis confirmed that recurrent deletion regions are found in regions with low densities of GO genes. These data suggest that haploinsufficiency of both STOP and GO genes within deletions in sporadic tumors may drive tumorigenesis.

Supplemental Discussion

Among the many drivers of tumorigenesis, recurrent deletions are the least well understood. Copy number and deep sequencing analyses have revealed that deletions, both chromosome arm and focal, are a much more frequent cause of gene inactivation than point mutations, with deletions covering an average of 17% of the genome per tumor (*13*). These studies, as well as our current analysis, show that deletions can affect thousands of genes per tumor, but the vast majority of these large deletions are hemizygous (*12*). We find that homozygous deletions occur 0.78% of the time for genes within recurrent focal deletion regions. Most studies so far have focused on identifying recessive tumor suppressor genes in which both copies are inactivated, either by deletion, silencing, mutation or a combination of these. If the two-hit model of tumorigenesis were to operate for all deletions, these deletions should each cover a potent tumor suppressor. However, our analysis of tumor sequencing data suggests that only 22% (18 of 82) of recurrent deletion regions could be explained by the presence of a known or putative recessive tumor suppressor gene. While the loss of a recessive tumor suppressor gene can account for a small subset of deletions, it cannot explain all of these events. Furthermore, it is even possible that many of these deletions are merely neutral passengers with no selective value for the tumor. To investigate these questions we exploited advances in RNA interference (RNAi) technology (*20, 30, 31*) to functionally annotate genes with respect to the cancer relevant phenotypes of proliferation and survival, properties shared with many *bona fide* tumor suppressors and ones we thought could provide an incremental selective advantage to cancer cells following single copy loss of many genes. A genetic screen based on this fundamental property yielded a set of candidate STOP genes that restrain proliferation and candidate essential GO genes that promote proliferation that showed unusual patterns of distribution in regions of recurrent hemizygous deletion in tumors.

Our model suggests that cancers might evolve in part by the selection of hemizygous somatic deletions encompassing high densities of STOP genes and low densities of GO genes. This strategy promotes proliferation and survival due to the cumulative reduction in TSG dosage while avoiding deleterious effects due to reduced dosage of genes that promote proliferation. The net proliferation and survival fitness of a given deletion is the integrated sum of the haploinsufficient phenotypes of the STOP and GO genes present in that deletion. Importantly, our analysis of GO gene depletion from recurrent deletion regions suggests that as much as 28% of genes in humans could display haploinsufficiency, some of which may be due to monoallelic expression imbalances. Because a fraction of the predominantly hemizygously deleted genes are occasionally homogyzously deleted, this could account for some of the depletion of GO genes observed. Furthermore, the excess STOP genes present in the deletion region over the expected will account for a small reduction, approximately 2.8%, in the number of GO genes expected since they are non-overlapping gene sets. Therefore, we predict that the percentage of genes acting in a haploinsufficient manner is likely to be less than 22-28%, and we conservatively suggest that 15-25% may behave in that fashion. The only published systematic analysis of haploinsufficiency of genes was done for growth phenotypes in yeast (*32*)**.** In that study, 3% of yeast genes were found to show haploinsufficiency for colony size. However, many yeast genes may have nothing to do with growth. If one looks at only essential genes, 10% show haploinsufficiency.

One concern was the possibility that many recurrent deletions are present simply because they are located adjacent to fragile sites (FRA). We examined the 39 known fragile sites previously described (*12*) and found that only 10 of these overlapped with the 82 recurrent deletions. Importantly, 8 of these 10 were in regions of very small deletions encompassing a total of 33 genes, of which 8 were STOP genes. In the case of the remaining two large deletions overlapping with FRA sites, both of them have known or putative tumor suppressors in them already (i.e. *APC* and *ATM*) arguing that they are likely to have selective value to the tumor and are not merely non-functional neutral deletions. Thus, fragile sites generating deletions with no selective value are likely to play a minor role in this analysis and are unlikely to affect the overall statistical enrichments we have observed.

Not all of the recurrent deletions are due to loss of the genes involved because a small number actually have a deletion that leads to an oncogenic fusion. For example in prostate cancer, TMRSS2-ETS2 fusions could explain some examples of recurrent deletion number 78 in our dataset. A total of 68 cancer samples have focal deletions covering the TMPRSS2-ERG peak region. These include only 15 prostate cancers with the consistent boundaries indicative of TMPRSS2-ERG fusion. We would like to make two additional points in this regard. The recurrent deletion region is an average of many deletions whose boundaries are not rigid and usually go several genes to either side of what the eventual boundary is ultimately declared to be, and it does appear that this fusion probably influenced the choice of the boundaries for deletion 78. Secondly, however, even in the case where a productive and positively selected transcription fusion oncogene is generated, it does not mean that the genes lost in the intervening region do not also contribute in a positive way to tumorigenesis. In this particular case, 6 of the intervening 20 genes are STOP genes and their deletion could still contribute to proliferation even in the presence of a TMRSS2-ETS2 fusion.

Our study provides evidence that large hemizygous deletions targeting multiple, adjacent genes that restrain proliferation are preferentially selected for during tumorigenesis. This suggests that these recurrent deletions may exhibit properties of a contiguous gene syndrome. Partial gene dosage due to deletion of multiple adjacent genes in a single deletion region has been shown as the cause of several classical contiguous gene syndromes such as 22q11.2 deletion syndrome, Angelman syndrome, and Prader-Willi syndrome, which produce complex phenotypes. The severity and distinguishing combination of abnormalities found with these patients is thought to result from the combined haploinsufficiencies of multiple genes, much as we hypothesize occurs in cancer.

The Cancer Gene Island Model provides a theoretical resolution to the conflict between the two-hit hypothesis of recessive tumor suppressors for sporadic tumors and the clonal expansion theory of tumor progression in the case of deletions. If a TSG is truly recessive, loss of one copy of the TSG would not result in a phenotype, and hence no clonal expansion would occur in that cell. Without clonal expansion, the likelihood that a mutation in the second allele would occur in the same cell is greatly diminished. However, if the deletion or mutation displays a proliferative phenotype on its own, clonal expansion of that event can occur and contribute to tumorigenesis alone. A subsequent second hit may or may not occur to further promote proliferation. For cases where clear inactivation of a TSG occurs in cancer by point mutation, e.g. *APC* in colon cancer, or gene silencing of both alleles, this would predict that the TSG is likely to exhibit some level of haploinsufficiency. Alternatively, if the potent recessive TSG is tightly linked to a haploinsufficient STOP gene, a deletion encompassing both genes might provide the clonal expansion that would facilitate mutation of the second allele of the recessive TSG.

Our study demonstrates that genes important for cancer development can be identified through the intersection of genome-wide loss-of-function screens and large-scale somatic copy number variation studies. The degree of enrichment of STOP genes in deletions that we report is likely to be an underestimate because the STOP gene list is likely to be noisy, and thus contains genes that do not contribute to proliferation due to RNAi off target effects. False negatives in shRNA screens are also a problem. This noise increases the expected overlap and decreases the enrichment observed. Therefore, we anticipate that better validated on-target RNAi reagents and more genetic screens in additional normal cell lines will refine our STOP and GO gene sets and significantly improve the statistical power of future analyses, just as the secondary validation screen provided a greater enrichment for STOP genes and the *in silico* use of KEGG gene categories of essential genes provided a greater depletion of essential GO genes. The statistics would also be strengthened if we knew which genes are haploinsufficient as they might be even more enriched than recessive STOP genes in general. In addition, performing this screen in multiple normal cell lines might also give a better list of candidate STOP genes because one does not expect HMECs to be a perfect model for all cancers. While there is no guarantee that a STOP gene in culture behaves as a STOP gene in vivo, many known TSGs are STOP genes. In addition, the enrichment of STOP genes in deletions also suggests that many of these genes do act the same in tumor cells and in vitro.

An unresolved issue is the degree to which gene silencing acts to inactivate genes in tumors. This impinges on our inability to identify more putative tumor suppressors across from recurrent deletions. We examined 526 tumors that had undergone whole genome sequencing and found only 62 examples of homozygous loss-of-function mutations (nonsense or frame-shift mutations). These genes harbor either two different inactivating mutations or one mutant allele without the corresponding wild-type allele due to either gene conversion or deletion of the other allele. If these sequenced cancers had 6 deletions per tumor, as we have seen in copy number analysis of > 3000 cancers, we would have assayed 3156 deletions in these 526 tumors. However, we only identified loss-of-function mutations in 10 genes across from these ~3156 deletions, with a average of 1.8 mutations per gene, excluding *TP53*. If the two-hit hypothesis were at play to explain the recurrent deletions, we might have expected many more examples of homozygous inactivation. Even if gene silencing were 10 times more frequent than inactivating point mutations, we would only observe 180 examples of a gene having one copy silenced and the other deleted out of ~3156 opportunities, less than 6% of the cases. Thus, while gene silencing events through methylation might potentially explain some advantages of deletions, it is unlikely to significantly affect the analysis presented here with respect to haploinsufficiency, although future studies will be required to rigorously establish the degree to which gene silencing acts to mimic homozygous inactivation of TSGs. However, even if loss-of-function mutations or silencing events across from recurrent deletions are identified, the deletion may still have a selective advantage on its own. Since growth regulators populate these regions and many are likely to be haploinsufficient, one would expect these regulators to contribute to proliferation of tumors when inactivated either hemizygously or homozygously as might be the case across from a deletion.

Future tests of the Cancer Gene Island hypothesis will come from several directions. One will be the reconstruction of recurrent deletions in normal cells and cancer cells to examine haploinsufficiency of deletions with respect to proliferation. An important step will also be the full-scale analysis of many tumors for point mutations, deletions, methylation and transcriptional profiles so that the precise genetic makeup of a tumor is known at the systems level. This will allow complete analysis of the number of genes that are inactivated in these regions of recurrent deletions. In addition, further rigorous analyses of copy number variation across different tumor types is likely to significantly redefine the boundaries of recurrent deletions and are likely to refine the hemizygous deletion landscape, providing further tests of this hypothesis in the future.

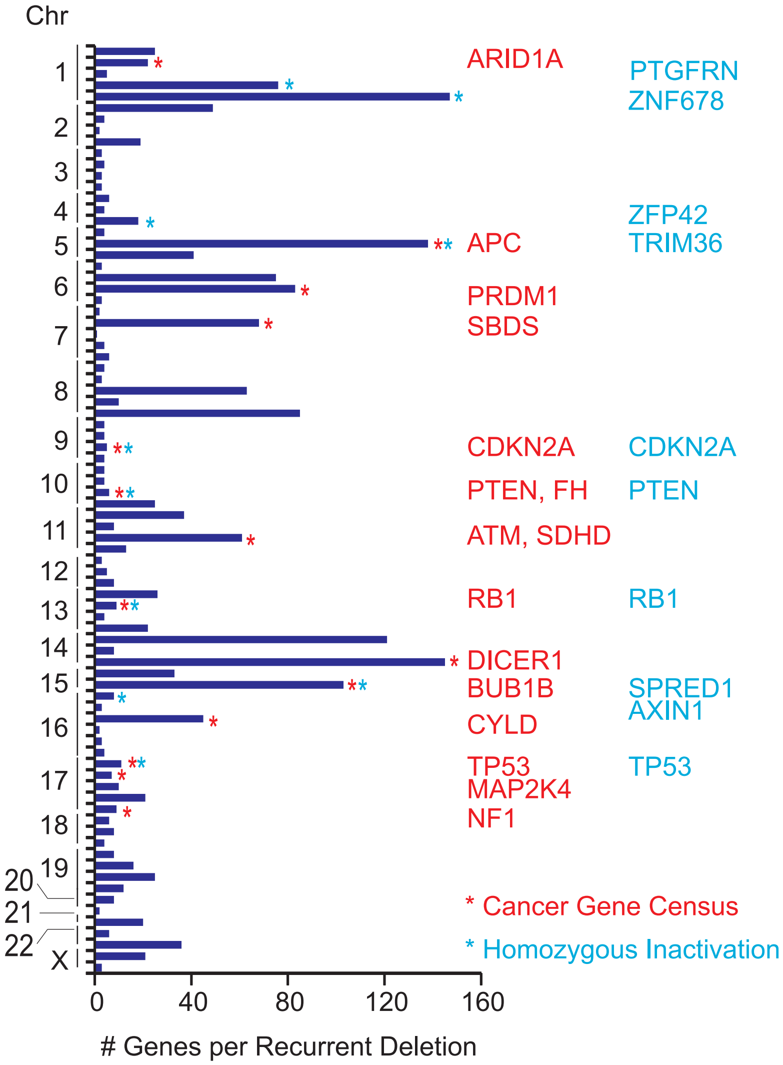


Fig. S1. Recurrent hemizygous deletions encompass many genes

A) The number of genes per recurrent deletion was plotted by chromosome location. Red and blue asterisks denote the presence of a known TSG from Cancer Gene Census or a homozygously inactivated gene from COSMIC whole genome sequencing, respectively. Each known or putative TSG is listed in red or blue, respectively. (Chr: chromosome)

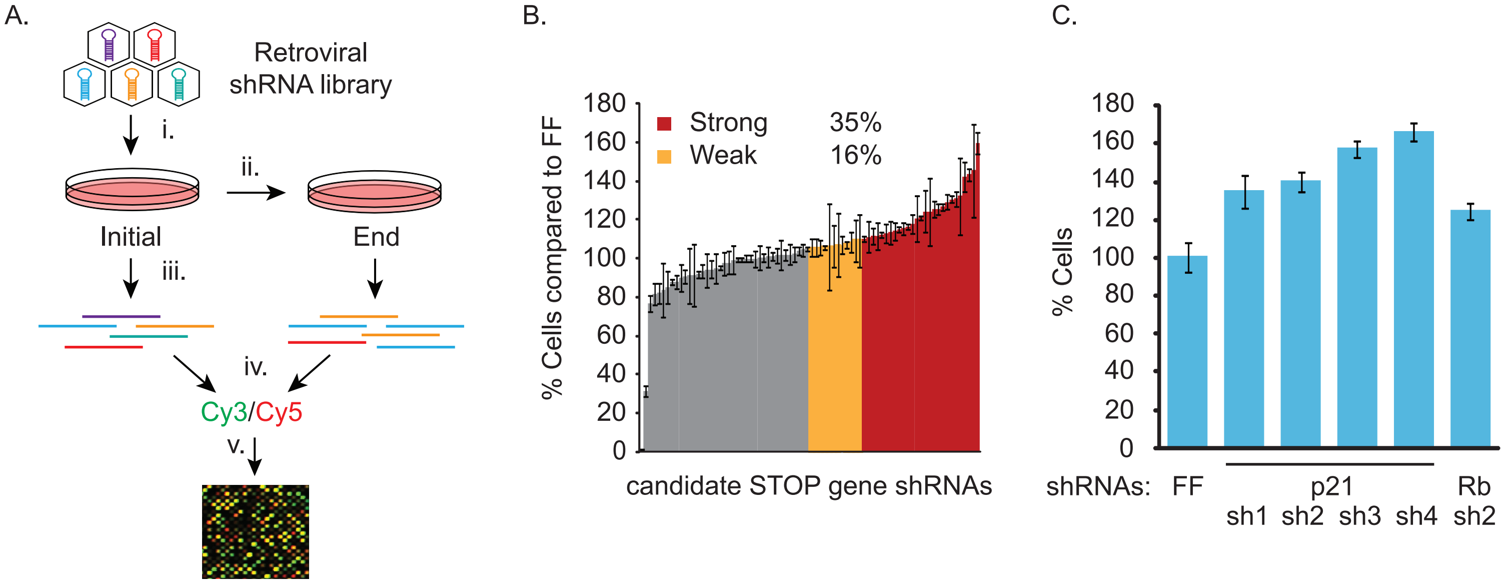
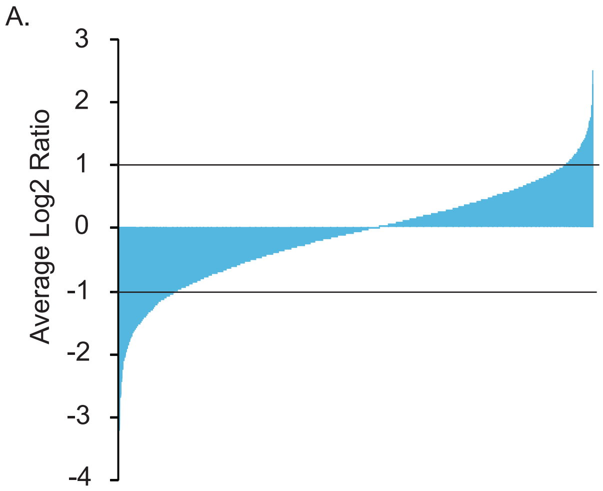


Fig. S2. A genome-wide proliferation screen identifies candidate STOP genes

A) Schematic of the primary shRNA proliferation screen. HMECs were transduced with virus particles expressing a genome-wide shRNA library (i) and allowed to grow under normal growth conditions for 8 population doublings (PD) (ii). Half-hairpin barcodes were PCR-amplified from initial (day 3, PD=0) and end (PD=8) genomic DNA samples, labeled with either Cy5 (initial) or Cy3 (end) dyes (iv), and competitively hybridized to a microarray (v). Screens were performed in triplicate.

B) Validation of primary screen STOP gene shRNAs using a multi-color competition assay (MCA). 77 independent shRNAs were examined for their ability to increase HMEC proliferation in a competition assay compared to a control shRNA targeting firefly luciferase (FF). Non-color HMECs expressing candidate STOP gene shRNAs or FF were plated in triplicate along with an equal number of GFP-positive HMECs expressing FF in the same well. The number of STOP gene shRNA-expressing cells after 5 days was measured by flow cytometry, averaged across triplicates, normalized to those expressing FF, and compared to positive control shRNAs targeting p21 and Rb (Fig. S2C). shRNAs were classified as either strong (≥ 10% increase, red), weak (5-10% increase, orange), or not validated (<5% increase, grey) to determine validation rates for the screen.

C) To test whether MCAs could be used to measure differences in proliferation, shRNAs targeting the known proliferation suppressors p21 and Rb were examined for their ability to increase HMEC proliferation in a competition assay compared to a control shRNA targeting FF. Non-color HMECs expressing p21 shRNA 1-4, Rb sh2, or FF were plated in triplicate in 24-well plates along with an equal number of GFP-positive HMECs expressing FF in the same well. The percent of non-color and GFP-positive HMECs in each MCA was measured after 5 days by flow cytometry and averaged across triplicates. The percent of p21 or Rb shRNA expressing cells were normalized to those expressing FF. p21 shRNAs 1-4 and Rb sh2 all demonstrated ≥ 10% increase in proliferation compared to FF.



B.

shRNAs	Genes	Enriched shRNAs	Enriched Genes	Genes Enriched with Multiple shRNAs
8203	2924	355 (4.3%	305 (10.4%)	42 (1.4%)

Fig. S3. A pilot shRNA screen identified regulators of cell proliferation.

A) Average log2 ratios of end vs. initial samples for > 8200 shRNAs in the pilot proliferation screen. Average log2 ratios across triplicates were ranked low to high and plotted. Enriched shRNAs are those with average log2 ratios ≥ 1 while lethal shRNAs are those with ratios ≤ 1 (black lines).

B) The number and percentage of genes and shRNAs that were enriched in the pilot proliferation screen with a single or multiple shRNAs are shown.

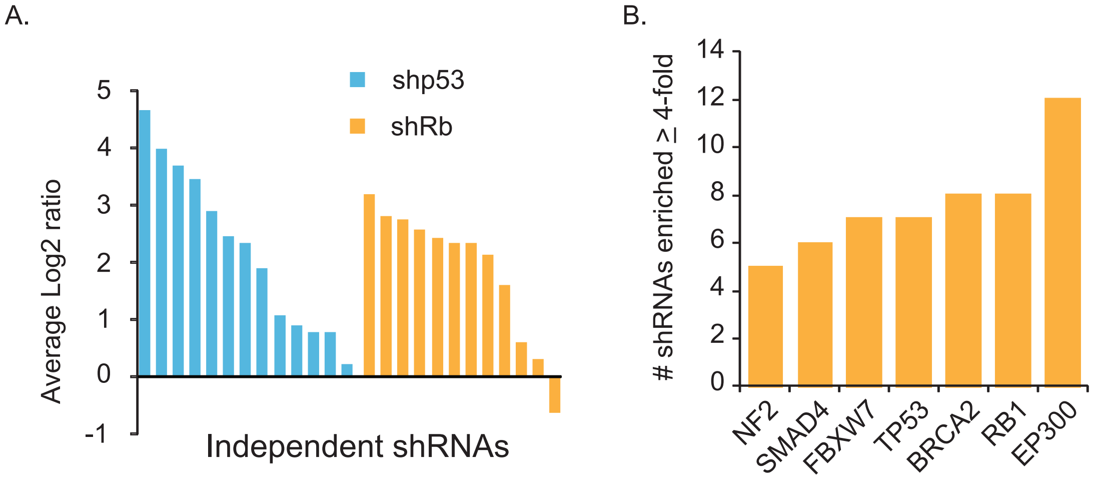


Fig. S4. A validation screen identifies known TSGs

A) Multiple shRNAs targeting p53 and Rb increase proliferation in the validation screen. Average log2 ratios for all independent shRNAs targeting p53 (blue) or Rb (orange) in the validation screen were ranked high to low and plotted. Nine out of 13 shRNAs targeting p53 and 9 out of 12 shRNAs targeting Rb were enriched greater than 2-fold compared to FF controls.

B) Multiple shRNAs targeting known tumor suppressors increase proliferation in the validation screen. The number of shRNAs that increased proliferation ≥ 4-fold compared to FF controls is shown for known tumor suppressor genes in various cancers.

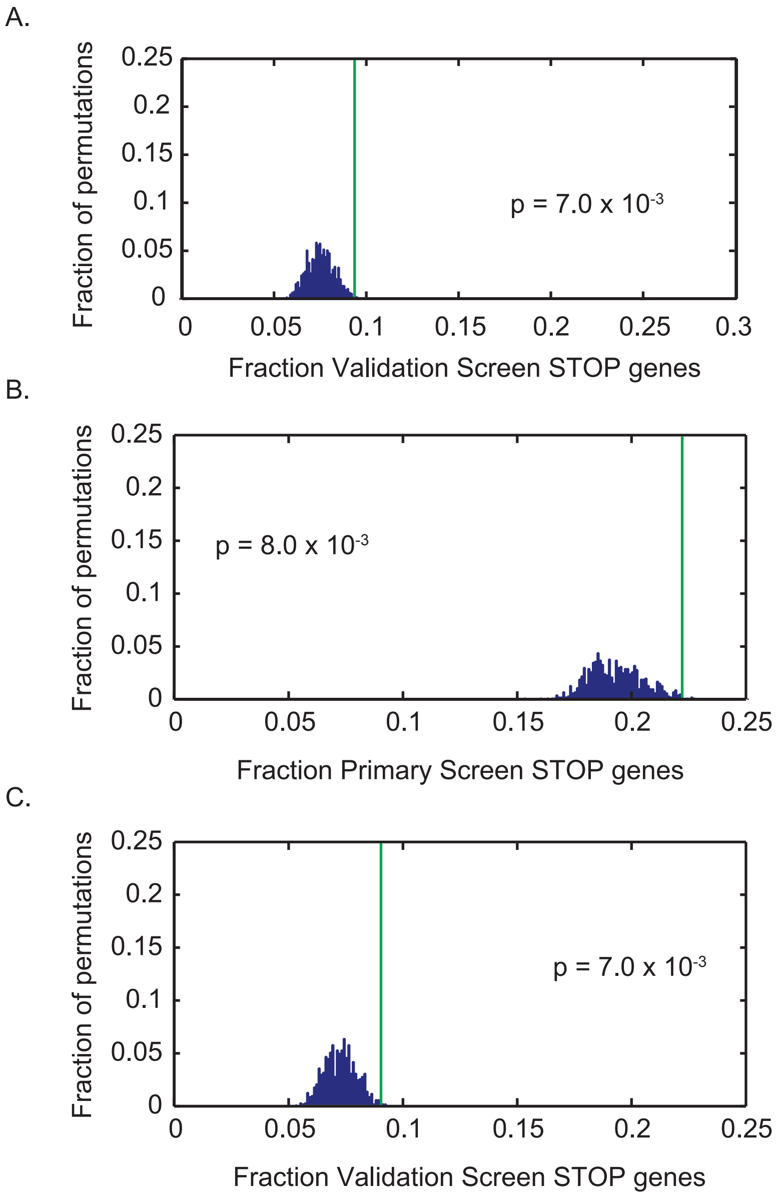


Fig. S5. Candidate STOP gene sets are significantly enriched within recurrent cancer deletion peak regions.

A) Validation STOP genes are significantly clustered within recurrent cancer deletion peak regions. The validation STOP gene set was mapped to genomic location, and cancer deletions were overlaid. The percentage of validation STOP genes in cancer deletions (green line) was compared to the distribution observed after 1000 permutations of the deletion peaks across the genome.

B, C) STOP gene sets without conflicting candidate STOP and GO genes significantly cluster in cancer deletion peak regions. Genes which overlapped in the STOP and GO gene sets due to one shRNA enriching and a different shRNA dropping out of the primary screen (B) or validation screen (C) were removed from both gene lists. These non-overlapping candidate STOP gene sets were mapped to genomic location and cancer deletion peak regions were overlaid. The percentage of candidate STOP genes found in cancer deletion peaks (green line) was compared to the distribution observed after 1000 permutations of the deletion peak regions across the genome.

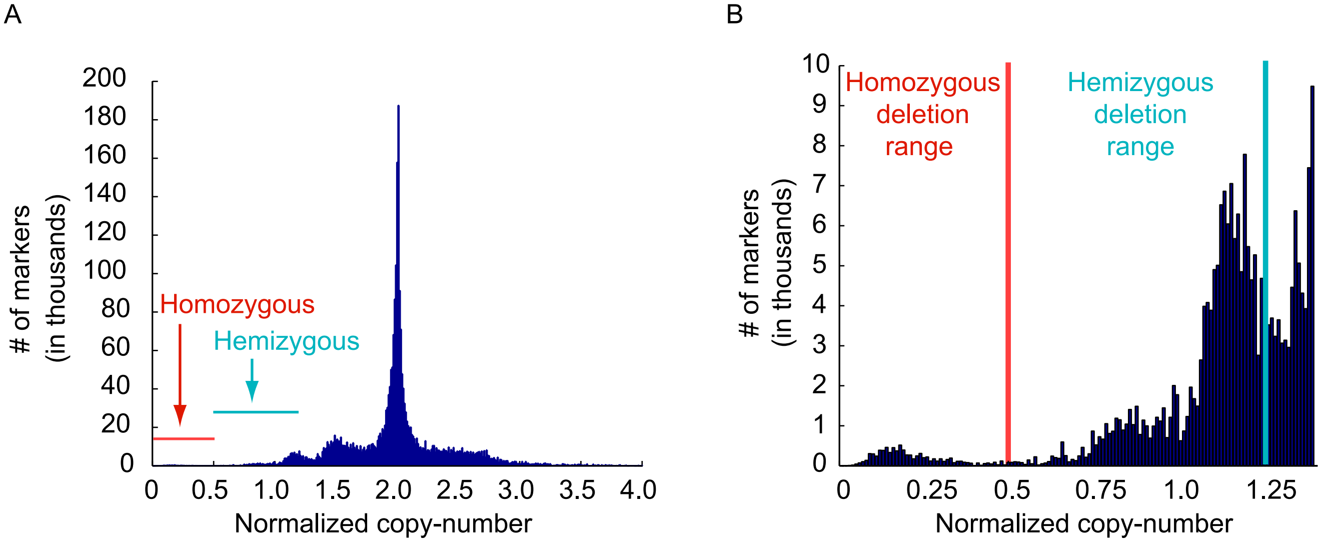


Fig. S6. Histograms of normalized copy numbers in the 82 peak regions of deletion, across 611 cell lines.

(A) The total number of SNP markers with normalized copy-numbers ranging from 0 to 4 is shown. For each cell line, the median copy number is normalized to two. The largest numbers of SNP markers are observed to have copy numbers near this median level. Other modes in this histogram represent additional copy number levels in subsets of the 611 cell lines. We set thresholds for detecting homozygous and hemizygous deletions at normalized copy numbers of less than 0.5 and 0.5-1.25 copies, respectively, as indicated.

(B) The region including normalized copy numbers less than 1 is magnified. The mode at less than 0.2 copies represents homozygous deletions. The deviation from zero corresponds to background signal intensity. We set the threshold for detecting homozygous deletions at a local minimum separating this mode and the higher modes that represent hemizygous loss.

Supporting Tables

Table S1. Recurrent deletion peak definitions

Table S2. Known recessive tumor suppressor genes according to the Cancer Gene Census

Table S3. Homozygously inactivated genes observed by whole genome sequencing in COSMIC

Table S4. Log2 ratios for shRNAs targeting STOP genes that significantly increased cell viability in the primary genome-wide screen

Table S5. Log2 ratios for shRNAs targeting STOP genes that significantly increased cell viability in the pilot screen

Table S6. Sequences of individual shRNA used in this study

Table S7. Log2 ratios for shRNAs targeting STOP genes that signficantly increased cell viability in the validation screen

Table S8. List of genes containing loss-of-function frameshift or nonsense mutations in 526 tumors previously analyzed by whole genome sequencing

Table S9. Number and percentage of STOP and GO genes found in each deletion peak

Table S10. Basal transcription factors, RNA polymerase, spliceosome, ribosome, DNA replication, fatty acid biosynthesis, amino-acyl tRNA synthesis, and mRNA export KEGG pathway essential GO genes

Table S11. Log2 ratios for shRNAs targeting GO genes that significantly reduced cell viability in normal and cancer cell lines

Table S12. Location of fragile sites within recurrent deletion peak regions