Cell

Resource

Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome

Teresa Davoli,[1,2,5] Andrew Wei Xu,[2,4,5] Kristen E. Mengwasser,[1,2] Laura M. Sack,[1,2] John C. Yoon,[2,3] Peter J. Park,[2,4] and Stephen J. Elledge[1,2,*]

[1]Howard Hughes Medical Institute, Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA
[3]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[4]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[5]These authors contributed equally to this work
*Correspondence: selledge@genetics.med.harvard.edu
http://dx.doi.org/10.1016/j.cell.2013.10.011

SUMMARY

Aneuploidy has been recognized as a hallmark of cancer for more than 100 years, yet no general theory to explain the recurring patterns of aneuploidy in cancer has emerged. Here, we develop *Tu*mor *S*uppressor and *On*cogene (TUSON) Explorer, a computational method that analyzes the patterns of mutational signatures in tumors and predicts the likelihood that any individual gene functions as a tumor suppressor (TSG) or oncogene (OG). By analyzing >8,200 tumor-normal pairs, we provide statistical evidence suggesting that many more genes possess cancer driver properties than anticipated, forming a continuum of oncogenic potential. Integrating our driver predictions with information on somatic copy number alterations, we find that the distribution and potency of TSGs (STOP genes), OGs, and essential genes (GO genes) on chromosomes can predict the complex patterns of aneuploidy and copy number variation characteristic of cancer genomes. We propose that the cancer genome is shaped through a process of cumulative haploinsufficiency and triplosensitivity.

INTRODUCTION

A key goal of cancer research is to identify genes whose mutation promotes the oncogenic state. Research over the last 40 years has identified numerous potent drivers of the cancer phenotype (Meyerson et al., 2010; Stratton et al., 2009; Vogelstein et al., 2013). Perhaps the most striking characteristics of cancer genomes are their frequent somatic copy number alterations (SCNAs) and extensive aneuploidies. Deletions and amplifications of whole chromosomes, chromosome arms, or focal regions are rampant in cancer, as are other rearrangements such as translocations and chromothripsis. Understanding how these events drive tumorigenesis is a major unmet need in cancer research.

Though ostensibly random, these alterations follow a nonrandom pattern that suggests that they are under selection and are likely to be cancer drivers rather than passengers. If so, we should be able to explain how they drive tumorigenesis. A recent clue as to how this might work came from the integration of a genome-wide RNAi proliferation screen with focal SCNA information (Solimini et al., 2012). The screen identified STOP and GO genes that are negative and positive regulators of cell proliferation, respectively. Hemizygous recurring focal deletions were enriched for STOP genes and depleted of GO genes, suggesting that the deletions maximize their protumorigenic phenotype through cumulative haploinsufficiency of STOP and GO genes. Haploinsufficiency describes a genetic relationship in a diploid organism in which loss of one copy of a gene causes a phenotype. The converse is triplosensitivity, in which an additional copy of a gene produces a phenotype. However, the distributions of STOP and GO genes were not able to predict aneuploidy or chromosome arm SCNA frequencies, perhaps because they represent only one aspect of tumorigenesis (proliferation) or are too diluted by nonhaploinsufficient genes. We hypothesized that the drivers of sporadic tumorigenesis might provide a more representative and potent set of STOP and GO genes with which to explore this phenomenon. Furthermore, this gene set may possess a higher frequency of haploinsufficiency.

In this study, we developed methods to identify tumor suppressor genes (TGSs) and oncogenes (OGs) from tumor DNA sequences. We implicate many new drivers in cancer causation and find many more cancer drivers than expected that exist in a continuum of decreasing phenotypic potential. Furthermore, we found that the distribution and potency of TSGs, OGs, and essential genes on chromosomes can explain copy number alterations of whole chromosomes and chromosome arms during cancer evolution through a process of cumulative haploinsufficiency and triplosensitivity.

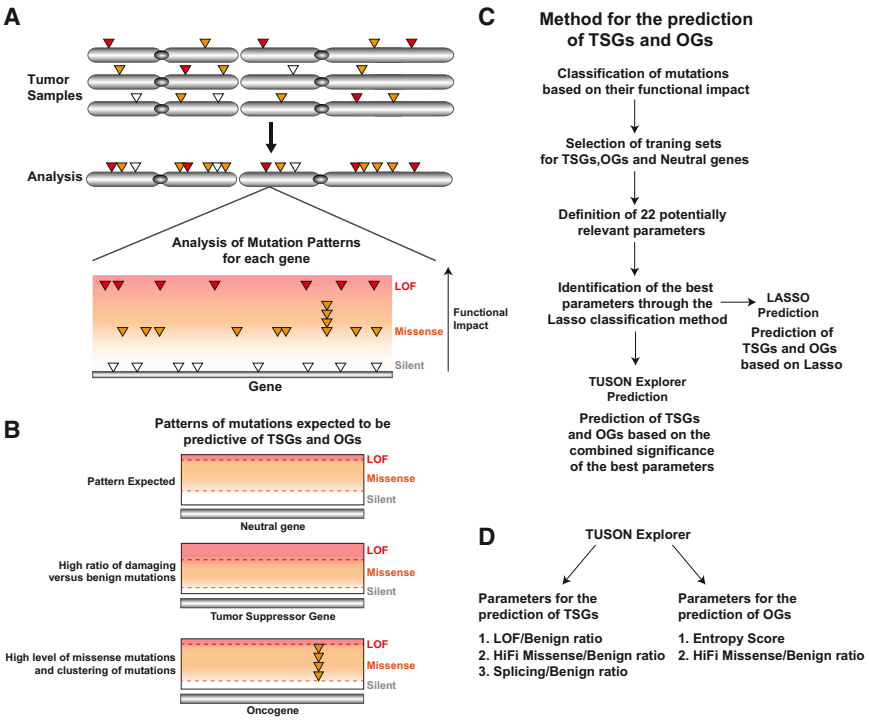

Figure 1. Prediction of TSGs and OGs Based on Their Mutational Profile

(A) Schematic representation of our method for the detection of cancer driver genes based on the assessment of the overall mutational profile of each gene. The somatic mutations in each gene from all tumor samples are combined and classified based on their predicted functional impact. The main classes of mutations (silent, missense, and LOF) are depicted.

(B) Schematic depicting the most important features of the distributions of mutation types expected for a typical TSG, OG, and neutral gene. Compared to "neutral" genes, TSGs are expected to display a higher number of inactivating mutations relative to their background mutation rate (benign mutations), and OGs are expected to display a higher number of activating missense mutations and a characteristic pattern of recurrent missense mutations in specific residues.

(C) A flowchart delineating the main steps in our method for identifying TSGs and OGs, from the classification of the mutations based on their functional impact to the identification of the best parameters through Lasso and their use for the prediction of TSGs and OGs by TUSON Explorer (or the Lasso method).

(D) Schematic related to (C) depicting the parameters selected by Lasso employed by TUSON Explorer for the prediction of TSGs and OGs (HiFI, high functional impact). For TSGs, the parameters are the LOF/Benign ratio, the HiFI/Benign ratio, and the Splicing/Benign ratio, whereas for OGs the parameters are the Entropy score and the HiFI/Benign ratio.

Also see Figure S1 and Table S1.

RESULTS

Cancer driver genes have been described as mountains and hills (Wood et al., 2007). Mountains are driver genes that are very frequently mutated in cancer, whereas hills represent less frequently mutated driver genes. It has become clear from recent international sequencing efforts that most potent drivers (mountains) have been discovered. A key issue is how to determine the identity of the significant but less frequently mutated drivers, the hills. A recent analysis searching for very high confidence cancer drivers in a database of ~400,000 mutations estimated that there were 71 TSGs and 54 OGs (Vogelstein et al., 2013). It is likely that there also exist additional functionally significant cancer drivers with weaker phenotypes and lower probabilities that are selected less frequently. A central question is how to identify these genes. In principle, with more samples analyzed, greater statistical significance can be placed on the outliers, allowing discovery of lower penetrance drivers. However, it is likely that there is more information present in the current data that may allow these lower frequency events to be detected.

To approach this question, we sought to devise a method to predict TSGs and OGs in cancer based on the properties of gene mutation signatures of these two distinct classes of driver genes. We hypothesized that the proportion of the different types of mutations with different functional impact would be informative in predicting these two types of drivers (Figure 1A). Each gene has a background mutation rate that is dependent on transcription, replication timing, and possibly other unknown parameters, and this rate can be estimated by the number of mutations that are unlikely to affect its function (such as silent or functionally benign mutations), whose observed frequency is not dependent on selective pressure during cancer evolution. The proportion of functionally relevant mutations of particular classes compared to this background mutation rate will be dependent on the degree of selection and will predict the likelihood that a gene will act as a cancer driver. TSGs and OGs can be distinguished among the cancer driver genes based on the characteristic pattern of the different types of mutations (i.e., loss of function [LOF], missense, silent) that are typically observed for those two classes of drivers relative to neutral genes, as illustrated in Figure 1B.

Identification of Parameters Predicting TSGs and OGs

We set out to determine the most reliable parameters for the prediction of TSGs and OGs in an unbiased way (Figure 1C). We used sequence data from >8,200 tumors from the COSMIC (Forbes et al., 2010) and TCGA (http://cancergenome.nih.gov/) databases and a recently published database (Alexandrov et al., 2013) comprising >1,000,000 mutations (Figure S1 and Table S1 available online). We defined a list of 22 parameters primarily based on the different classes of mutations and used the classification method Lasso and three training sets of known TSGs and OGs (from the Cancer Gene Census, Futreal et al., 2004) (Table S2A) and neutral genes to identify those parameters that best predict the two classes of driver genes (see Experimental Procedures). We employed PolyPhen2 to predict the

functional impact of missense mutations in order to classify them into those with potentially high (HiFI) or low (LoFI) functional impact (Adzhubei et al., 2010). LoFI mutations are typically conservative amino acid changes or changes in poorly conserved residues. We defined the combination of silent and LoFI missense as "Benign" mutations to provide a larger, more reliable value for estimating background mutation rates. We also defined the LOF mutations as the combination of nonsense and frameshift mutations. As a majority of known OGs show an atypical distribution of recurrent mutations in one or a few key residues, we utilized entropy, a well-defined concept in physics and information theory (Shannon and Weaver, 1949), to measure the degree of reoccurring mutations within a gene. The Entropy score represents the weighted sum of the probabilities, across a gene, that a site is mutated (see Experimental Procedures). The best parameters found by Lasso for the prediction of TSGs and OGs are described below and are visualized in Figures 1D and 2.

Tumor Suppressors versus Neutral Genes

The most predictive parameters for TSGs are: (1) the ratio of LOF mutations to Benign ($p = 2.51 \times 10^{-28}$, Wilcoxon, one-tailed test); (2) the ratio of Splicing to Benign mutations ($p = 4.6 \times 10^{-13}$); (3) the ratio of HiFI missense to Benign mutations ($p = 3.2 \times 10^{-14}$); and (4) high-level deletion frequency ($p = 1.46 \times 10^{-8}$). A 20-fold cross-validation shows a high prediction accuracy of 93.2% on these training sets (Figure 2A and Table S2B).

Oncogenes versus Neutral Genes

The most predictive parameters for OGs are: (1) the entropy for missense mutations ($p = 2.2 \times 10^{-14}$); (2) the ratio of HiFI missense mutations to Benign mutations ($p = 1.2 \times 10^{-9}$); and (3) high-level amplification frequency ($p = 1.4 \times 10^{-6}$). The 20-fold cross-validation accuracy is 85.2% (Figure 2B and Table S2B).

Tumor Suppressor Genes versus Oncogenes

One important aim of our prediction method is the discrimination between TSGs and OGs. The most predictive parameters between these two sets are: (1) the ratio of LOF to Benign mutations ($p = 2.5 \times 10^{-16}$); (2) high-level amplification frequency ($p = 1.3 \times 10^{-9}$); (3) high-level deletion frequency ($p = 7.6 \times 10^{-6}$); and (4) the ratio of Splicing to Benign mutations ($p = 9.9 \times 10^{-7}$). The 20-fold cross-validation accuracy is 91.9%. Overall, Lasso identified parameters that make intuitive sense for these classes of genes and clearly delineated TSGs and OGs from each other and from neutral genes. In sum, we identified independent parameters that strongly predict and distinguish between TSGs and OGs (Figures 2C and 2D and Table S2B).

Identifying OGs and TSGs

Having identified the most predictive parameters, we developed a method we call *Tu*mor *S*uppressor and *On*cogene Explorer (TUSON Explorer) that combined selected parameters to derive an overall significance and ranking for each gene as a potential TSG or OG (Figure 1D). First, we derived a p value for each gene for the ratios of LOF/Benign, Splicing/Benign, HiFI/Benign, and Missense Entropy based on the comparison to the neutral gene set (see Experimental Procedures). For the LOF/Benign parameter, we applied a correction to normalize for the nonuniform codon usage among genes for the occurrence of nonsense mutations (see Experimental Procedures). Finally, we used an extension of Liptak's method to provide a combined p value for the selected parameters for each gene. For TSGs, the combined p values (and q values) were derived from individual values from the LOF/Benign, Splicing/Benign, and HiFI/Benign parameters. For OGs, the combined values were derived from the Missense Entropy and the HiFI/Benign parameters (Figure 1D). The LOF/Benign parameter for discrimination between TSGs and OGs was subsequently utilized to define a final list of OGs and TSGs (see Experimental Procedures). TUSON Explorer does not take into account SCNA information, and this allows us to perform a rigorous analysis of our cancer driver genes for their abilities to predict the frequency of deletion and amplification (see below).

As a second strategy to predict the probability of a given gene being a TSG or OG, we employed the Lasso model, which also takes into account SCNAs (see Experimental Procedures). The ranked lists of predicted TSGs and OGs by both Lasso and TUSON Explorer are contained in Tables S3A and S3B. This list provides a facile look-up table that can be easily sorted for different parameters for all those who are interested in the mutational behavior of a given gene in this data set.

Both ranking strategies performed similarly and eliminated the problems of inappropriately including giant genes and genes in highly mutable regions (Dees et al., 2012), without the need to consider expression level or replication timing (Lawrence et al., 2013). Importantly, both of our strategies distinguish between TSGs and OGs, which are predicted to have functionally opposite roles in the control of cell growth and have different implications for potential cancer therapeutics.

Estimates of the Numbers of TSGs and OGs

A ranking of this nature consists of truly significant genes mixed with false-positive genes that obtain low p values by chance under the null hypothesis. Thus, we sought to get an estimate of the minimum number of TSGs and OGs by analyzing the distribution of the combined p values for each class of cancer driver genes. To achieve this, we utilized a histogram-based method (Mosig et al., 2001) to estimate the number of rejected hypotheses from the distributions of the combined p values calculated for each gene. With our data set, this method estimated ~320 TSGs and ~250 OGs (Extended Experimental Procedures). This long list of TSGs and OGs suggests that there are many more drivers than anticipated and that they exist in a continuum of decreasing potency (Discussion and Figure 7A). For the analyses described below, we considered the top 300 TSGs (q value < 0.18) and 250 OGs (q value < 0.22) as our working lists. Given the fact that the deviation of the mutation signatures from the normal pattern is a function of the degree of selection and the frequency of mutation, increasing the number of tumor samples will detect even more cancer drivers of progressively weaker selective pressure. To determine the potential number of TSGs upon additional sequencing, we applied TUSON and estimated the number of TSGs (using Mosig's method) on random subsets of the data set with increasing numbers of samples and observed that the number of predicted TSGs continues to increase with additional samples. We observed that the rate of increase in the predicted number of TSGs decreases slightly at

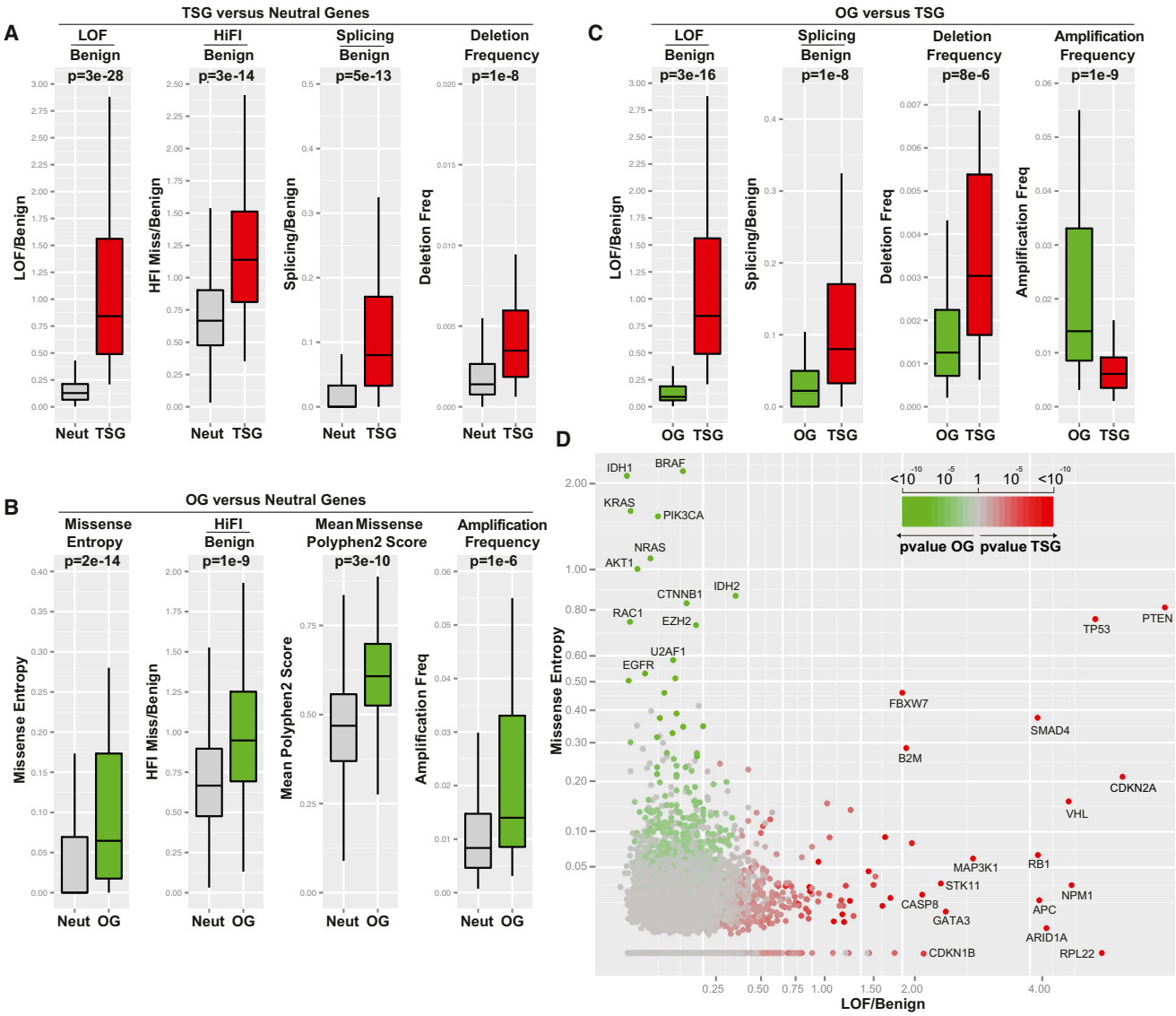

Figure 2. Best Parameters Selected by Lasso for the Prediction of TSGs and OGs

(A–C) Box plot representations of the distribution of the values for the indicated parameters in the neutral genes (gray), TSG (red), and OG training set (green). The median, first quartile, third quartile, and outliers in the distribution are shown. The p value for the difference between the two indicated distributions is shown as derived by the Wilcoxon test.

(A) Box plots showing the distribution of LOF/Benign, HiFI missense/Benign, and Splicing/Benign ratios and the high-level frequency of focal deletion among the neutral gene set and the TSG set.

(B) Box plots showing the distribution of Missense Entropy, HiFI missense/Benign, mean of PolyPhen2 score, and the high-level frequency of focal amplification among the neutral gene set and the OG set.

(C) Box plots showing the distribution of LOF/Benign and Splicing/Benign ratios, high-level frequency of deletion and amplification among the TSG and OG sets.

(D) Plot of the LOF/Benign ratio and Missense Entropy for each gene, the best parameters for discriminating between TSGs and OGs. Specific genes with high levels of LOF/Benign or Missense Entropy are shown along with their p values for being a TSG or an OG (TUSON Explorer).

Also see Figure S2 and Tables S2 and S3.

the highest number of samples examined (Figure S2), indicating a possible plateau at very large number of samples.

PAN-Cancer Mutational Analysis

Gene Ontology (GO) term and pathway analysis of our list of potential TSGs showed enrichment for functions that are highly relevant to tumorigenesis, including cell-cycle control, embryonic development, promotion of differentiation, apoptosis, and blood vessel development (Table S3C and Figure 3). In addition, there was a strong enrichment for transcriptional regulation ($q = 6.19 \times 10^{-11}$) and chromatin modification ($q = 5.7 \times 10^{-12}$). Furthermore, we noticed an enrichment for genes involved in

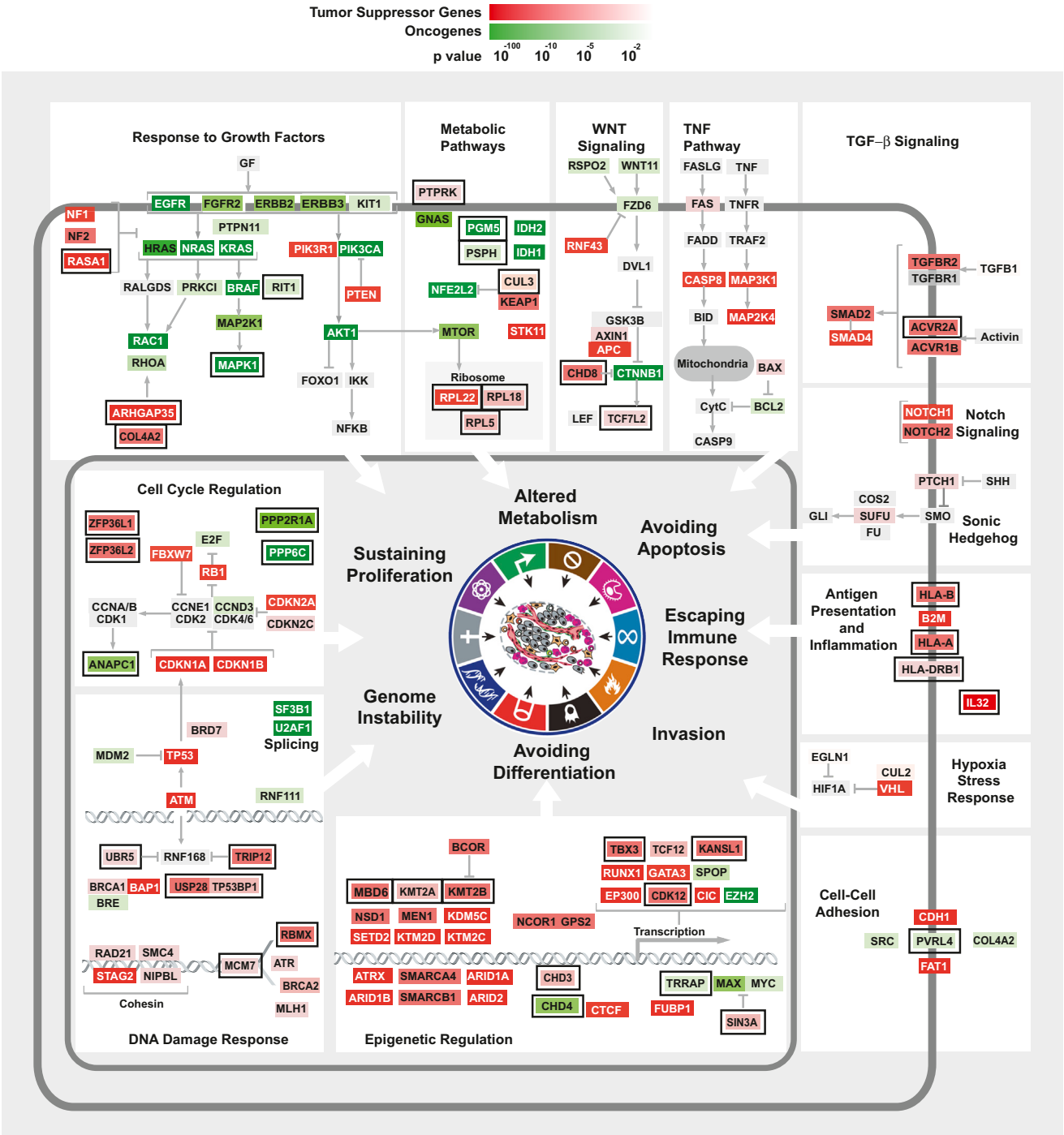

Figure 3. Representation of Predicted TSGs and OGs within Their Cellular Pathways

Placement of predicted cancer drivers within specific cellular pathways. TSGs and OGs were predicted by TUSON Explorer. The predicted TSGs and OGs belonging to many known cellular pathways or complexes are shown, along with how they generally correspond to the hallmarks of cancer. TSGs are shown in red, whereas OGs are shown in green; color intensity is proportional to the combined p value as indicated. For some pathways, additional genes absent from the predicted TSGs and OGs were added and marked in gray for clarity of the pathway representation. Although several genes are known to affect multiple pathways and hallmarks, only one function is presented for the sake of limiting the complexity of the diagram. An external black box outside of the colored gene box highlights genes previously less well characterized for their roles in tumorigenesis.

See also Figure S3 and Table S3.

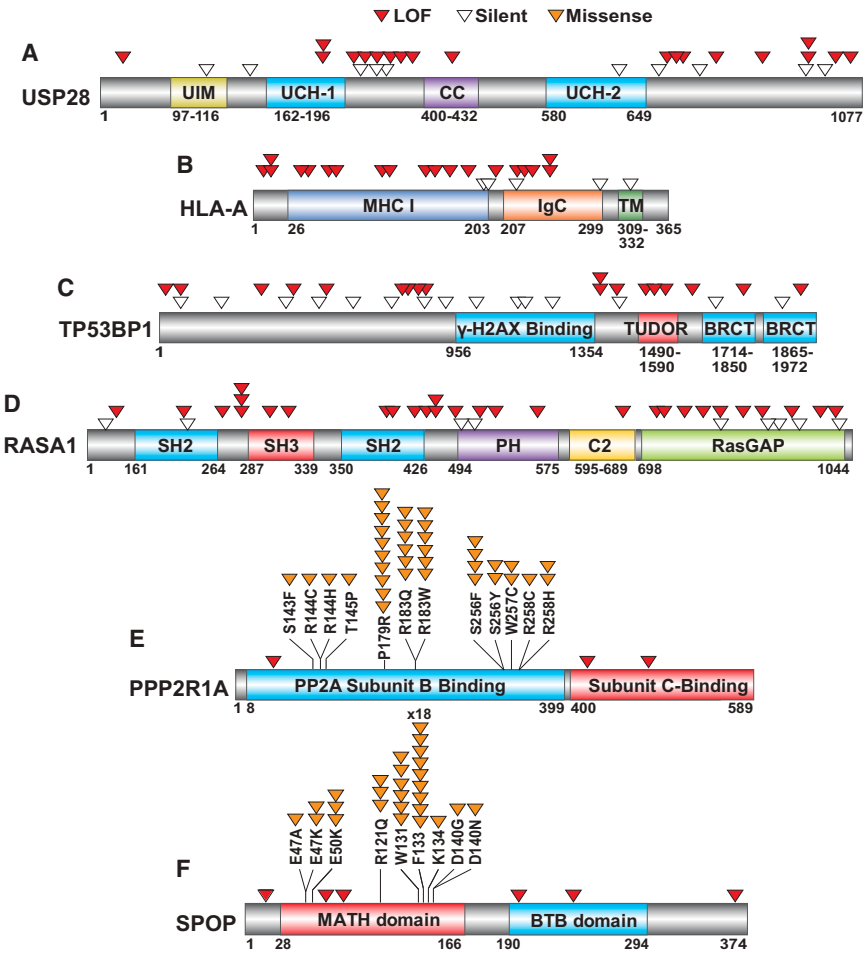

Figure 4. Representation of Mutation Patterns in Representative Predicted TSGs and OGs

(A–F) The mutational patterns of selected TSGs and OGs are depicted. For TSGs (A–D), the locations of LOF (red) and silent (white) mutations within the coding regions are shown. For OGs (E and F), the location of recurrent missense mutations (orange) and of LOF mutations (red) within the coding regions are shown. *USP28*, *TP53BP1*, and *RASA1* are previously less well-characterized candidate TSGs in the TUSON PAN-Cancer analysis. *SPOP* and *PPP2R1A* are previously less well-characterized candidate OGs.
See also Table S3.

New Potential Cancer Drivers

New components of pathways previously linked to tumorigenesis have also been detected (Figure 4). For example, the DNA damage response pathway is central to the maintenance of genomic stability, and both members of a key DDR complex, the TP53BP1/USP28 complex (Zhang et al., 2006), which are substrates of the ATM kinase (Matsuoka et al., 2007), were identified within the top 110 TSGs ($q < 0.15$, Figures 3 and 4A–4C). Two components that regulate ATM-dependent chromatin remodeling, *UBR5* and *TRIP12* (Gudjonsson et al., 2012), are also high on the TSG list ($q < 0.25$). *RBMX*, which controls *ATR* and *BRCA2* expression (Adamson et al., 2012), ranked 76th on the list ($q = 1.1 \times 10^{-4}$). There are several candidate OGs with enzymatic functions that could serve as drug targets (Figure 4E and Tables S3B and S7B), including three phosphatases (*PPP6C*, *PTPN11*, *PTPRF*) and regulators such as *PPP2R1A*, as well as several kinases (*MAPK1*, *MAPK8*, *BRSK1* among others). There are many other new potential TSGs and OGs on these lists that cannot be discussed here due to limitations of space, but several of these are presented in Figure 3.

Consistent with the enrichment of cell-cycle and apoptosis GO terms, integration of the PAN-Cancer analysis with functional gene sets revealed that essential genes are significantly depleted for deleterious mutations (see below). An exception to that finding was the presence of *RPL22*, *RPL5*, and *RPL18* large ribosomal subunit genes in the top 210 TSGs ($q < 0.07$; Figure 3). Interestingly, heterozygous mutations in ribosomal genes promote tumorigenesis in zebrafish (Lai et al., 2009). Furthermore, familial mutations in ribosomal proteins have been associated with Diamond-Blackfan anemia, which is associated with an increased risk of leukemia (Willig et al., 2000).

Analysis of Individual Tumor Types

Identification of cancer drivers using the PAN-Cancer analysis favors discovery of genes whose functions contribute to many

the immune system ($q = 5.8 \times 10^{-3}$), particularly in antigen processing and presentation represented by the MHC class I system. Two major HLA genes (*HLA-A* and *HLA-B*) were in the top 90 candidate TSGs ($q < 0.0002$), and the β2 microglobulin (*B2M*) gene, which is an obligatory complex component of both HLA proteins, ranked 43rd ($q = 9.2 \times 10^{-9}$) on our TSG list, underscoring that escaping from immunosurveillance is a significant selective force in tumorigenesis (Hanahan and Weinberg, 2011) (Figures 3 and 4B). Furthermore, *IL32*, which stimulates the immune responses of NK cells and CD8+ T cells that monitor MHC status (Conti et al., 2007), is also in the top 50 TSGs. Unexpectedly, negative regulation of cell adhesion ($q = 4.32 \times 10^{-4}$) was enriched, indicating that increase of cell adhesion may confer a selective advantage to tumor cells. Traditionally it has been thought that reducing adhesion promotes tumorigenesis; however, recent findings suggest a potentially different role for cell-to-cell-adhesion. First, it has been shown that circulating tumor cells exist in clusters in the blood (Hou et al., 2011). Second, *PVRL4*, which ranked well in our Lasso OG list, was shown to promote transformation through cell adhesion, as do several other oncogenes like *MYC*, *KRAS*, *PI3K*, and loss of *PTEN* (Pavlova et al., 2013). Thus, promotion of adhesion may be a driving force in tumorigenesis.

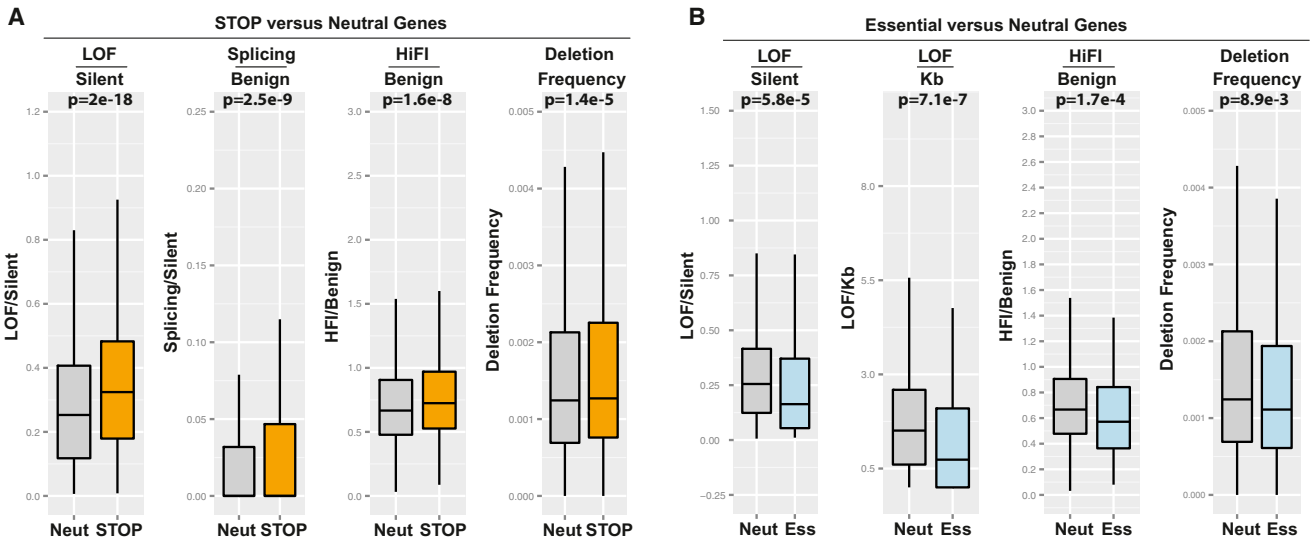

Figure 5. Behavior of Functional Gene Sets Relative to TSG Parameters

(A and B) Box plot representation of the distribution of the values for the indicated parameters in the neutral genes compared with the essential genes and STOP genes. The median, first quartile, third quartile, and outliers in the distribution are shown. The p value for the difference between the two indicated distributions is shown as derived by the Wilcoxon test.

(A) Box plots showing the distribution (orange) of LOF/Silent, Splicing/Benign, and HiFI missense/Benign ratios (gray) and the high-level frequency of focal deletion among the neutral gene and STOP gene sets.

(B) Box plots showing the distribution of LOF/Silent, LOF/Kb, HiFI missense/Benign ratios and the high frequency of focal deletion among the neutral gene set (gray) and the essential gene set (light blue).

See also Tables S3A, S5A, and S5B.

different types of cancer. Certain cancer drivers may miss the cutoff for significance in the PAN-Cancer analysis because they are primarily involved in controlling tissue-specific differentiation networks or because they are rate limiting for a particular function in only certain tissues. Thus, we anticipate that new drivers can be discovered through analysis of mutation signatures in individual tumor types despite their lower numbers. We performed the same analysis as above for each of 20 tumor types (Tables S1, S4A, and S4B). This analysis found many TSGs that are specific for one tissue type such as *CDH1* and *GATA3* in breast adenocarcinoma, *VHL* and *PBMR1* in kidney clear renal cell carcinoma, and *ID3* and *NPM1* in hematological malignances (Table S4A). Genes whose FDRs for the different subtypes were below 0.25 were all already relatively highly ranked in the PAN-Cancer analysis. This indicates that the majority of tissue-specific drivers were detected in the PAN-Cancer analysis.

We wanted to determine how many new TSGs might be expected from the analysis of a new cancer subtype. For this, we calculated the number of TSGs in the whole data set lacking an individual tumor type (Tables S4C) and compared this list to the TSGs in that tumor type, which averaged 14 genes. We found that the average new cancer type added about five TSGs to the PAN-Cancer list. Thus, on average, ~70% of the TSGs detected in a single tumor type were already detected in the PAN-Cancer analysis performed after excluding the mutations in that type of tumor. This suggests that most cancer genes selected during tumor evolution act in cellular pathways whose role in tumorigenesis is widespread among different tumor types.

Analysis of TSGs and OGs

Behavior of Functional Gene Sets

The PAN-Cancer mutation data set allows us to interrogate the behavior of functional gene sets derived through experimental approaches. We previously showed that STOP genes are overrepresented in regions of deletion (Solimini et al., 2012). Examination of their abundance in the set of candidate TSGs showed that STOP genes are significantly enriched in the TSG set ($p = 0.0031$, Fisher's exact test) comprising ~10% of the top 300 TSGs (68% more than expected). The STOP gene set showed a 50% higher ratio LOF/Silent than the average for the neutral gene set ($p = 2.0 \times 10^{-18}$; Figure 5A). Furthermore, the STOP genes showed a significant increase in the Splicing/Benign and HiFI/Benign ratios, two of the most potent parameters for the prediction of TSGs (Figure 5A). This analysis further underscores the fundamental connection between cell proliferation and cancer.

We next investigated a high-confidence set of 145 genes predicted to be essential at the cellular level based on their housekeeping cellular functions and their high evolutionary conservation (Table S5A and Experimental Procedures). This set was depleted from regions of recurring deletions (Beroukhim et al., 2010) by 43% ($p = 0.0198$, Fisher's exact test), and a larger set of 332 essential genes was depleted by 25% ($p = 0.014$). Examination of the LOF/Silent ratio showed that, for the set of 145 genes, the frequency of LOF /silent was 27% lower than the rate for the neutral gene set ($p = 5.8 \times 10^{-5}$; Figures 5B and S5B). Additionally, the LOF/kb and HiFI/Benign ratios were also significantly decreased in the essential gene set. Given that the vast majority of the mutations and deletions in question

are heterozygous, the reduced LOF mutation and deletion frequency of the essential genes as a group argues that between 25% and 45% are haploinsufficient. Interestingly, our TSGs were enriched in recurring focal deletions (68%, p = 0.000281) and were depleted from recurring amplifications (28%, p = 0.015), whereas the OGs were enriched in amplifications (25%, p = 0.046) and depleted from focal deletions (23%), indicating that amplifications are also likely to be Cancer Gene Islands.

General Properties of Cancer Drivers

High Interactivity

To search for unique properties of TSGs and OGs, we examined the degree to which these drivers participate in protein complexes using the CORUM database of experimentally validated human protein complexes (Ruepp et al., 2010). We found that both TSGs and OGs were significantly more likely to be in protein complexes than a typical protein. The 13.4% of all proteins found in CORUM are in a complex. However, 36.7% of the predicted TSGs were in complexes ($p = 3.1 \times 10^{-24}$), and 26.4% of the predicted OGs were in complexes ($p = 3.5\ 10^{-8}$; Figure S3A).

High Betweenness

A second property of complexes is the degree to which they are connected to other proteins and complexes. We explored this by assessing a property called ''betweenness,'' which is proportional to the number of times the protein is part of the shortest paths between all pairs of proteins in a network. High betweenness indicates a greater connectivity. The TSG and OG candidate gene lists were mapped onto the most current BioGRID human protein-protein interaction network (Stark et al., 2006). Both the predicted TSGs and OGs show a high degree of betweenness (TSG $p = 6.16 \times 10^{-32}$, OG $p = 1.68 \times 10^{-6}$; Figure S3B), indicating that they are optimally positioned to impact information flow through networks.

Greater Length

Proteins with greater interactivity often have more domains. Thus, we examined gene length. Cancer drivers are significantly longer than the average gene (1,700 nt), with the mean for TSGs at 3,234 nt ($p = 2 \times 10^{-21}$) and OGs at 2,107 nt ($p = 9.7 \times 10^{-6}$). Importantly, this observation is also characteristic of the genes in our training sets (TSGs, 4,133 nt, $p = 6.7 \times 10^{-10}$; OGs, 2,260 nt, p = 0.0016).

An Unusually High Concentration of TSGs on the X Chromosome

While examining the distribution of TSGs across chromosomes, we found that the X is unusually enriched for TSGs (p = 0.0042, exact binomial test) relative to autosomes. Examining the top 300 TSGs, we find that, although only 3.9% of all genes are on the X, it contains 7.3% of all predicted TSGs (86% more than expected) and was the only chromosome with a significant enrichment of TSGs (Table S5C). Given the fact that the X is functionally haploid in both males and females, this observation has certain implications for evolutionary selection of cancer drivers during tumorigenesis and haploinsufficiency of TSGs (see Discussion).

Interestingly, in the top 400 TSGs, we found two potential TSGs on the Y, *ZFY* and *UTY* ($q < 0.22$). Both have homologs on the X that escape X inactivation, each of which also displays tumor suppressor properties: *ZFX* (p = 0.019) and *UTX*/*KDM6A* ($p = 3.3 \times 10^{-46}$). This could explain the observation that frequent Y nullisomy is observed in prostate, renal cell, head and neck, Barret's esophageal adenocarcinoma, bladder, pancreatic adenocarcinoma, and other cancers at frequencies of 30%–80% (Bianchi, 2009; Kowalski et al., 2007).

Furthermore, we analyzed the silent mutation rates along entire chromosomes and found an enhanced mutation rate on the X chromosome relative to autosomes in males (30% increase, $p = 1.1 \times 10^{-9}$). This increase is even greater in females (77.5%, $p = 1.6 \times 10^{-11}$) (Table S5D). Possible explanations for this phenomenon are detailed in the discussion.

Distribution and Potency of Cancer Drivers on Chromosomes Predict Arm and Chromosome SCNA Frequencies

In addition to focal SCNAs, a less frequent but significant chromosomal alteration is whole-arm loss or gain. We hypothesized that the distribution and potency of TSGs and OGs on chromosomes might explain the average frequency of chromosomal whole-arm SCNAs seen in cancer. To this end, we generated a *ch*romosome *arm* score, Charm, that provides an assessment of each arm based on the density of TSGs and OGs and their potency (weights of TSGs and OGs are based on their rank on their respective lists and serve as a metric for their potency). The Charm score represents a measure of the amount of positive or negative growth and survival potential that wild-type OGs or TSGs might normally impart to a given arm and therefore how SCNAs might impact cancer evolution by altering this balance during tumorigenesis. Importantly, for Charm calculations, we employed the parameters from TUSON Explorer, which does not include copy number information. To lessen the diluting impact of false positives for this analysis, we applied stringency cutoffs of a q value of 0.25 for TSGs and 0.35 for OGs and a minimum of 10 missense mutations for OGs and 8 LOF mutations for TSGs to get a stringent list of 264 TSGs and 219 OGs (see Experimental Procedures). The analysis of the $Charm^{TSG}$ score versus frequency of chromosomal arm deletion revealed a strong positive correlation ($r = 0.578$, $p = 5.8 \times 10^{-5}$, Pearson correlation; Figure 6A and Table S6A). Interestingly, the $Charm^{TSG}$ score also showed a strong negative correlation with arm amplification frequency, and thus a high $Charm^{TSG}$ score indicates a significantly reduced tendency for a chromosome arm to be amplified ($r = -0.59$, $p = 2.8 \times 10^{-5}$; Figure 6B). Simple TSG densities without weighting by rank also showed correlations with arm deletions (Figure S4A), but these correlations are improved by Charm. In contrast to $Charm^{TSG}$, the $Charm^{OG}$ score showed a negative correlation with arm deletion frequency ($r = 0.52$, $p = 3.2 \times 10^{-4}$; Figure 6C). Moreover, the density of OGs positively correlated with arm amplification frequency ($r = 0.45$, $p = 1.8 \times 10^{-3}$, Figure 6D) but was not improved by the Charm score (data not shown).

We reasoned that, like GO genes in focal deletions, the chromosome arms most frequently deleted in cancer would be depleted of genes that promote the fitness of cancer cells. Using our in silico list of essential genes, we estimated their fitness potency by estimating their avoidance of damaging mutations using the (LOF + HiFI)/Benign ratios. By determining a $Charm^{Ess}$ score for each arm, we found a negative correlation between

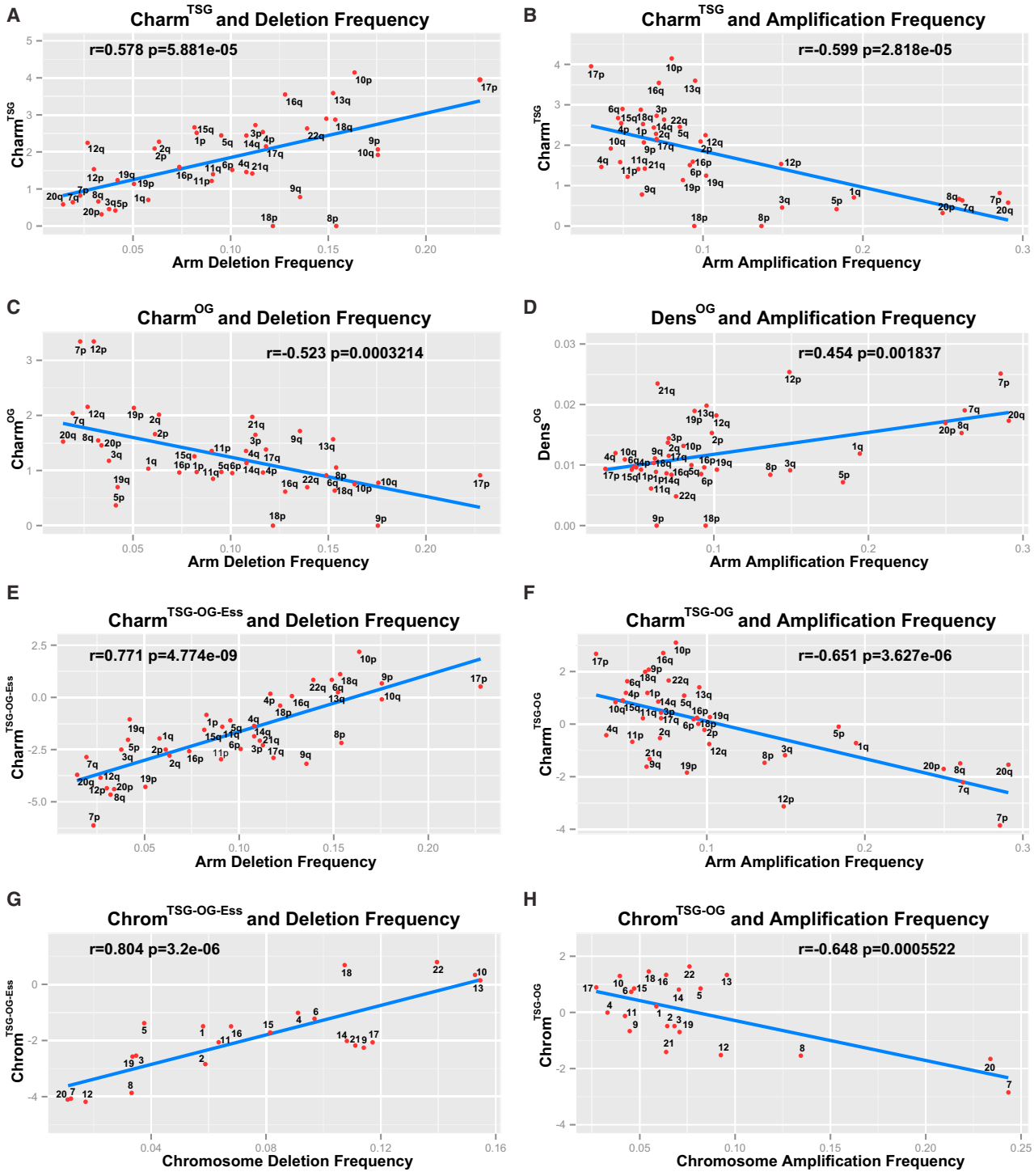

Figure 6. Charm Score, Chrom Score, and Copy Number Alterations: Correlation Analysis

(A–F) The Pearson's correlation analysis of the Charm scores or Density score for the indicated gene sets (A–D) and/or combinations of these sets (E and F) relative to the frequency of arm-level deletion or amplification. Ess, essential genes.

(G and H) The correlations of the Chrom scores ($Chrom^{TSG-OG-Ess}$ and $Chrom^{TSG-OG}$) relative to the chromosome-level deletion or amplification frequency. The Charm scores refer to a weighted density of TSGs, OGs, or essential genes present on each chromosome arm, where each TSG or OG is weighted based on its rank position within the list of predicted TSGs and OGs ranked by TUSON Explorer and each essential gene is weighted based on its (LOF + 1/2 × HiFI)/Benign ratio. The Chrom score is the equivalent of the Charm score for whole chromosomes.

See also Figures 4 and 5 and Table S6.

$Charm^{Ess}$ scores and the frequency of arm-level deletions ($r = 0.34$, $p = 1.6 \times 10^{-2}$; Figure S4D). No correlation was found between $Charm^{Ess}$ and amplification frequency, as expected.

Because the $Charm^{TSG}$, $Charm^{OG}$, and $Charm^{Ess}$ scores correlate with arm-level deletion, we combined them by giving a positive weight to the $Charm^{TSG}$ score and a negative weight to the $Charm^{OG}$ and $Charm^{Ess}$ scores to derive a cumulative $Charm^{TSG\text{-}OG\text{-}Ess}$ score. The $Charm^{TSG\text{-}OG\text{-}Ess}$ score gave an even stronger positive correlation with arm deletion frequency ($r = 0.77$, $p = 4.7 \times 10^{-9}$; Figure 6E and Table S6A). For amplification, we used the $Charm^{TSG\text{-}OG}$ score and found a strong negative correlation with amplification frequency ($r = 0.65$, $p = 3.6 \times 10^{-6}$; Figure 6F). We also combined amplification and deletion frequencies into a single score for copy number variation on each arm and compared that to the $Charm^{TSG\text{-}OG}$ score. This also gave a strong significant correlation ($r = 0.74$, $p = 2.7 \times 10^{-8}$; Figure S5A).

We extended our analysis of cancer driver scores and SCNAs to whole-chromosome aneuploidy using its Charm equivalent score that we call Chrom (Figures 6G, 6H, S4E–S4H, S5B, S5E, and S5F). $Chrom^{TSG}$ significantly correlated with chromosome deletion frequency ($r = 0.66$, $p = 3.7 \times 10^{-4}$; Figure S4E) and anticorrelated with amplification frequency ($r = 0.54$, $p = 4.0 \times 10^{-3}$; Figure S4F). Impressively, when we combined all three classes—TSGs, OGs, and essential genes—the $Chrom^{TSG\text{-}OG\text{-}Ess}$ was strongly predictive of the frequency of chromosome loss ($r = 0.80$, $p = 3.2 \times 10^{-6}$; Figure 6G), and $Chrom^{TSG\text{-}OG}$ was predictive of chromosome gains ($r = 0.64$, $p = 5.5 \times 10^{-4}$; Figure 6H). Very similar results were obtained using just the TUSON ranking without stringency cutoffs (Figures S5C–S5F and Table S6B).

Together, these data strongly argue that a selective force in generating chromosomal arm and whole-chromosome SCNAs derives from the integration of the relative densities and potencies of positively and negatively acting cancer drivers on a particular chromosome. Thus, the SCNAs in cancer genomes may be selected during tumor evolution through cumulative haploinsufficiency for deletions (as previously proposed for STOP genes in focal deletions [Solimini et al., 2012]) and through cumulative triplosensitivity for amplifications (see Discussion).

DISCUSSION

In this study we analyzed the mutational data from >8,200 sporadic cancers to predict cancer driver genes. We determined the most predictive parameters for identifying TSGs and OGs and used them to develop an algorithm called TUSON Explorer to predict the probability that an individual gene functions as a TSG or an OG in cancer. This unbiased approach demonstrated that the probability of being a cancer driver can be assessed by the significance of the distortion of its mutational pattern from the pattern expected for a "neutral" gene. Combining data from our analyses of drivers and copy number changes, the average tumor in our data set has a mean number of ~1 OG mutation, ~3 TSG mutations (LOF and damaging missense), ~3 chromosomal arm gains, ~5 chromosomal arm losses, ~2 whole-chromosome gain, ~2 whole-chromosome losses, ~12 focal deletions, and ~11 focal amplifications (Zack et al., 2013). Thus, SCNAs comprise a very large proportion of cancer-driving events.

A Continuum of Cancer Driver Genes

A central conclusion from this study is that there are likely to be many more cancer drivers than anticipated. Our estimate of the number of TSGs based either on the combined significance of the different parameters or on the single best parameter for the prediction of TSGs, i.e., the LOF/Benign ratio, predicted ~320 TSGs with the current database from 8,200 tumors. Likewise, we also predict more OGs than anticipated. The view of the cancer landscape emerging from our analysis does not contain a clear cutoff for predicting cancer drivers. Instead, there exists a continuum of decreasing probability of a given gene being a driver (either TSG or OG). This probability is revealed by the degree of selection that the gene experiences during tumor evolution, which should be proportional to the phenotypic effect caused by its loss or gain. This continuum of decreasing potency of potential cancer drivers is likely to correspond to a continuum of increasing numbers of genes with decreasing phenotypic severity, as illustrated schematically in Figure 7A. In addition, we hypothesize that events that simultaneously affect multiple weak drivers can cumulatively have an effect equal to a single potent driver. Our modeling of the progressively higher number of driver genes identified as increasing numbers of tumors are analyzed suggests that this number will continue to climb as more sequence information becomes available but may be beginning to plateau. However, the newly identified drivers are likely to display progressively less potency with lower therapeutic significance. This is analogous to GWAS studies for which increasing sample sizes allow the identification of progressively weaker acting variants.

Our analysis provides a probability of each gene being a cancer driver, and as such, there will be false positives regardless of the threshold of minimum probability that we employ. Identifying bona fide drivers from the regions with significant p values but higher FDR values, i.e., weaker phenotypic signatures, can be aided by considering other information such as their involvement in SCNAs, biochemical connections to known OGs and TSGs, and functional information gleaned from the literature. These heuristic methods can be used to increase confidence and rescue genes onto the likely cancer driver list (Tables S7A and S7B).

PAN-Cancer and Tissue-Specific Analysis

Analysis of individual tumor types identified distinct sets of drivers in each tumor type, but the majority of these were also identified in the PAN-Cancer analysis as lower confidence candidates (Tables S4A–S4C). Thus, although there is clearly tissue specificity, there is still significant overlap among different tumor types and a PAN-Cancer analysis samples a sufficient number of similar tumors to detect most of the largely tissue-specific or tissue-biased cancer drivers. Our analysis suggests that significantly deeper sequencing of individual tumor types is unlikely to uncover many new potent drivers beyond what we have already identified and further sequencing is likely to suffer from diminishing returns. This view is consistent with a recent review that argues that nearly all potent

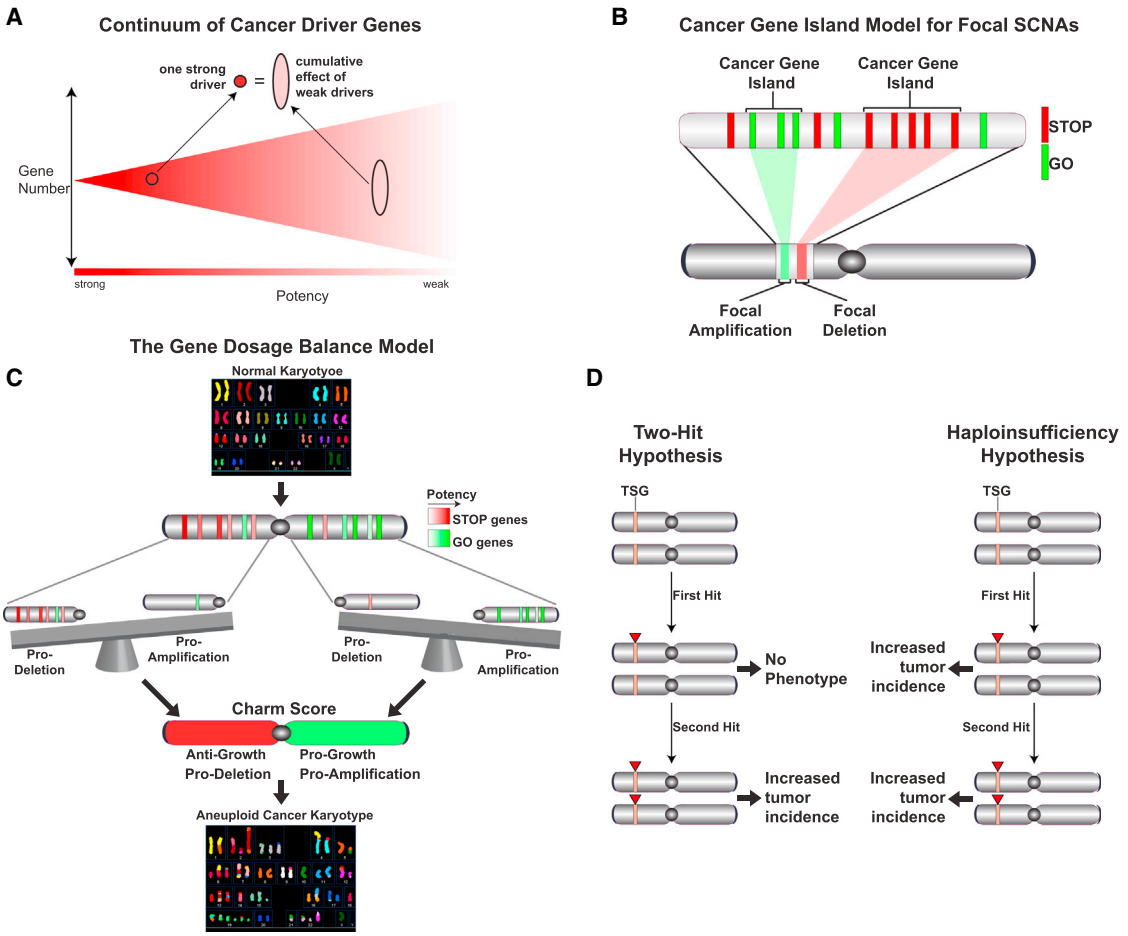

Figure 7. Cumulative Haploinsufficiency and Triplosensitivity Shape the Cancer Genome

Illustrative schematics of different concepts highlighted in the Discussion.

(A) The phenotypic continuum of cancer drivers.

(B) The cancer gene island model for focal SCNAs.

(C) The cumulative gene dosage balance model for predicting the patterns of aneuploidy. (The panel depicts the concept for arm-level SCNAs.)

(D) Comparison of the predictions of Knudson's Two-Hit Hypothesis for TSGs compared to the Haploinsufficiency Hypothesis presented in this study.

drivers have been identified (Vogelstein et al., 2013). Sequencing of more rare and relatively unexplored cancer types may identify a few novel potent drivers that are specific to those tumor types, but the vast majority of potent drivers will already have been seen in other cancers. The major effects of continued sequencing will likely be to solidify the continuum by bringing much weaker drivers into the realm of statistical significance.

Properties of New Potential Cancer Driver Genes

Analysis of the lists enriched for cancer drivers revealed several general properties that distinguish them from nondriver genes. The lists of both TSGs and OGs are strongly enriched both for residence in protein complexes and for a property known as betweenness, which is a measure of the degree to which a set of genes is enriched for hubs within an interaction network. Thus, the driver genes are much more highly connected than the average protein in the human gene network and are longer. Highly connected nodes are better positioned to control the flow of information, and their removal or hyperactivation will have the highest impact across a network due to their centrality.

Unexpected Properties of the X Chromosome

Given the potentially deleterious effects of mutating TSGs, we anticipated that TSGs would be depleted from the X chromosome by natural selection, as the X is haploid in males and is functionally haploid in females due to dosage compensation. However, our analysis revealed just the opposite—namely, that the X has 86% more TSGs than expected. Oncogenes, on the other hand, are not overrepresented on the X. The likely explanation is that a deleterious mutation in a TSG on the X is more penetrant because there is not a WT copy to compensate for its loss. This further suggests that natural selection has not completely depleted TSGs from the X, possibly because cancer is largely a postreproductive disease.

We found a higher mutation rate for the X than for autosomes, and this is further exaggerated in females. In females, the additional increase in X mutability is likely due to the presence of the inactive X, which has very little transcription and, hence, less transcription-coupled repair and is enriched in late-replicating heterochromatin, which tends to be more mutagenic (Stamatoyannopoulos et al., 2009). The mechanism underlying these differences and their biological significance remains to be determined. However, these differences might indicate that the mutation rates of whole chromosomes are set by evolution and that the higher mutability of the X is advantageous over evolutionary time if it also occurs in the germline.

Haploinsufficiency and Cancer

The clonal expansion theory of tumorigenesis argues that, in order for an individual mutation to be selected, it must cause an expansion of the clone derived from that mutant cell by increasing its relative proliferation and survival (Vogelstein and Kinzler, 1993). This is intuitive for OGs, as they are dominant, but it is less so for TSGs. For a hemizygous mutation in a TSG to be selected in cancer, we have to assume that either the mutation is dominant negative or the TSG is haploinsufficient. Our current analysis of the degree to which essential genes are absent from hemizygous recurring focal deletions, coupled with the reduced frequency with which essential genes experience LOF mutations in tumors, conservatively suggests ~30% haploinsufficiency overall among human genes (Experimental Procedures). A recent analysis of haploinsufficiency by the mouse knockout consortium (White et al., 2013) found that 42% of genes examined produced a phenotype when heterozygous, similar to our estimates. Evidence suggesting that our sporadic TSG list is largely haploinsufficient comes from a comparison of the enrichment in focal deletions of STOP genes versus our sporadic TSGs. STOP genes, which are TSG-like, are enriched by 20% (Solimini et al., 2012). If we assume that only 30% of this gene set is haploinsufficient and that all of the selective enrichment comes from haploinsufficient genes, then a list of purely haploinsufficient STOP genes would be expected to be enriched by 67%. Perhaps coincidentally, our list of TSGs is enriched 68% in recurring focal deletions, suggesting that a significant proportion, and possibly all, of sporadic TSGs are haploinsufficient.

We propose that two classes of TSGs might exist: those that are haploinsufficient and contribute to sporadic cancer and those that are *haplosufficient* and do not significantly contribute to sporadic cancer through mutation. Circumstances under which organisms inherit only one functional copy of those haplosufficient TSGs might result in cancer because loss of the second allele would produce a selectable phenotype. This situation occurs with familial TSGs and the classic Two-Hit model of tumorigenesis. Our hypothesis is consistent with the fact that, out of a list of 73 familial TSGs culled from the literature, only 32% of them had a combined q value <0.25 in the PAN-Cancer analysis (Table S3D). Another circumstance with only one functional allele per cell occurs on the X, where we see a ~86% higher density of TSGs than on the autosomes. If the predicted rate of ~30% haploinsufficiency is correct, then one might expect a ~200% increase over autosomes, but negative selective pressure on the X could have reduced that number. Thus, it is possible that there are actually similar densities of TSGs on the X and autosomes (haploinsufficient sporadic TSGs and haplosufficient potential TSGs), but those on the X realize their tumorigenic potential at a higher rate than do those on the autosomes.

The PAN-Cancer Mutational Analysis Predicts Aneuploidy in Cancer

Aneuploidy is a hallmark of cancer and can have both advantageous and deleterious consequences for cells (Tang and Amon, 2013; Luo et al., 2009), but there is no general theory that explains how patterns of aneuploidy emerge. Knowing the identity and potential potency of cancer drivers has allowed us to uncover a driving force behind selection of arm- and chromosome-level SCNAs. Our analysis using Charm and Chrom as an integrated assessment of the density and potency of the different classes of cancer driver genes on chromosomes displayed a robust ability to predict the patterns of whole-arm amplifications and deletions and aneuploidy (Figures 6, S4, and S5). The fact that the Charm score improves the correlations with SCNAs compared to the simple gene density of the different classes of genes indicates that the ranking of driver genes by TUSON Explorer is likely to represent an accurate estimate of the potency of their phenotypic effect in cancer and further supports the continuum theory.

$Dens^{OG}$ and $Charm^{OG}$ do not predict arm amplification as well as $Charm^{TSG}$. This reduced predictive potential is likely to be because the OGs were selected on the basis of the ability to be activated by mutation and because simply increasing the dosage by 50% might not strongly impact the networks they control. $Charm^{OG}$, however, does show a strong negative correlation with arm deletion frequency, indicating that, normally, the WT OGs are acting to promote proliferation and survival and the cumulative reduction of their levels by 50% is deleterious. In this respect, the OGs are behaving like the essential genes, and the inclusion of a high-confidence list of 332 essential genes together with OGs and TSGs further improves the predictive ability for arm deletions (Figure 6E). As expected, the essential genes have no predictive power for amplifications.

$Charm^{TSG}$ strongly predicts whole-arm deletions. Unexpectedly, it also strongly predicts arm amplification, providing a strong negative correlation. This suggests that increasing the gene dosage of a group of TSGs can have deleterious effects on tumorigenesis through the process of cumulative triplosensitivity. If TSGs are truly haploinsufficient, their WT protein levels may be only marginally sufficient to execute their roles. If so, TSGs may well be more sensitive to increased gene dosage to further enhance their pathways than typical genes. In other words, haploinsufficient genes may be more likely to display triplosensitivity. This property of sporadic TSGs being both haploinsufficient and triplosensitive, therefore, may make their cumulative Charm score an even better parameter to explain SCNAs of chromosome arms and aneuploidy in general. Developing a combined $Charm^{TSG\text{-}OG\text{-}Ess}$ and $Chrom^{TSG\text{-}OG\text{-}Ess}$ score can now predict ~80% of the frequency of arm and chromosome loss and ~65% of the amplifications observed across all cancers.

Although the correlation between Charm/Chrom scores and SCNAs is striking, there are several areas for improvement. The first area concerns our lack of knowledge of the full complement of essential genes and which of these are haploinsufficient. Second, only a subset of OGs will be dosage sensitive, and this knowledge would improve the correlation. In addition, there are two classes of OGs. Class I contains classical oncogenes such as KRAS that are activated by mutation but whose WT copies are not necessarily oncogenic after overexpression and will not be predictive of amplification. Class II contains genes such as cyclin D that can be activated by overexpression but are difficult to activate by missense mutations and thus lack a mutational signature. Class II OGs cannot be identified with confidence through mutational signatures yet are likely to display triplosensitivity and would positively correlate with amplification. Third, some TSGs can be difficult to distinguish from OGs. These are TGSs that have low haploinsufficiency but can produce a selectable phenotype by generation of dominant-negative alleles. Such genes will lack a strong LOF signature but will show a significant number of deleterious missense mutations, which are likely to predominantly occur in one or a few crucial residues, thus conferring a significant Entropy score. In addition, early SCNA events might influence subsequent events, as is the case when specific aneuploidy co-occurs (Ozery-Flato et al., 2011), which would confound our analysis to some degree. Finally, refining these lists of cancer drivers will only improve their predictive power. The current programs for prediction have their strengths and weaknesses and are likely to be further improved in the future. More precise knowledge of these essential and cancer driver genes should significantly improve SCNA predictability and our understanding of the cancer genome. Finally, the SCNA frequencies might vary according to tumor type; thus, comparison of data sets within one tumor type might provide more predictive power. In addition, we do not know the background frequency of SCNAs upon which selection acts, so the observed SCNA frequency cannot be normalized like mutation rates can, and therefore, the observed SCNA frequency detected might reflect both frequency of the event and its selective power, which could confound the correlation.

Models of Cancer Evolution

Our work suggests a very important role for cumulative haploinsufficiency and triplosensitivity operating during cancer evolution to drive tumorigenesis. In each genomic region, there are STOP (TSG) and GO (OG and essential) genes that will exert a negative or positive phenotypic effect on tumorigenesis. Both for focal deletions as illustrated by the Cancer Gene Island Model (Figure 7B) and for chromosomes and chromosomal arm SCNAs as indicated by the Charm and Chrom analysis (Figure 7C), the integrated cumulative balance of these positive and negative tumorigenic effects of individual genes affected in each SCNA event provides the selective potency to that event and can predict its frequency across cancers.

For the past 40 years, the tumor suppressor field has been guided by Knudson's classical Two-Hit Hypothesis of tumorigenesis for familial cancers. Though there are certainly bona fide examples of the Two-Hit Model in sporadic cancer, this model conflicts with the theory of clonal evolution of sporadic cancer in the assumption that the first hit is fully recessive and a second hit is required to contribute to tumorigenesis. While it is difficult to measure the frequency with which the Two-Hit Hypothesis operates in cancers because the role and extent of methylation inactivation is not yet known in each tumor, analysis of LOF mutational events suggests that it may be a relatively infrequent event except in the case of a few genes such as *TP53* (p53) and *CDKN2A* (p16), both of which are inducible responders to oncogenic stress, which can increase during tumorigenesis. In the cases of sporadic cancer, wherein the two-hit hypothesis does operate, it is still possible and even probable that the genes involved are haploinsufficient to begin with. Our results have led us to propose that the vast majority, if not all, of sporadic TSGs are likely to be haploinsufficient and that, therefore, sporadic TSGs are most likely to operate through the Haploinsufficiency Model shown in Figure 7D. It is important to note that these hypotheses are not mutually exclusive, as loss of the second allele of a haploinsufficient TSG, the second hit, will undoubtedly provide a stronger selective pressure than the first hit. However, a tumor has multiple paths through which to evolve, and it may not require loss of that second allele as it obtains growth-promoting power through the accumulation of other events.

In 1914, Theodor Boveri proposed that specific "chromosome constitutions can be produced such that the cells that harbor it are driven to unrestrained proliferation" (Boveri, 1929). Recurring patterns of aneuploidy exist in tumors, but whether they exist because of the frequency of occurrence of each individual SCNA event or because they are selected due to a tumorigenic phenotypic effect was not known. Here, we propose that the cumulative phenotypic effects of gene dosage alterations of STOP and GO genes provide the selective pressure that is responsible for the recurrent patterns of copy number variation observed in cancer. Our findings support the hypothesis put forward by Boveri exactly one century ago that aneuploidy is not only a hallmark of cancer, but is also a driving force during the evolution of human cancer.

EXPERIMENTAL PROCEDURES

Somatic Mutation Data Set

The data set of somatic mutations included data from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) research network and from the Catalogue of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) and the data set published by Alexandrov et al. (2013). The data set contained ~1,200,000 mutations from 8,207 tumor samples from >20 tumor types (Table S1) and will be available at http://elledgelab.med.harvard.edu/. All data related to SCNAs were derived from the TCGA Genome Data analysis Center at the Broad Institute (Zack et al., 2013).

TUSON Explorer Predictions

The PolyPhen2 algorithm (Adzhubei et al., 2010) was used to predict the functional impact of each missense mutation and to classify them as high functional impact (HiFI) or low functional impact (LoFI). We defined the four following classes of mutations: (1) Benign mutations: Silent + LoFI Missense; (2) Loss of Function mutations (LOF): Nonsense and Frameshift mutations; (3) Splicing mutations: mutations affecting splicing sites; and (4) HiFI missense mutations. An additional parameter considered was the Entropy score, which measures the degree of randomness of the distribution of missense mutations.

Among 22 potential parameters, we selected the most predictive ones by using the Lasso prediction model and three training sets of known TSGs and OGs (from the Cancer Gene Census, Futreal et al., 2004) and putative neutral genes. LOF/Benign, Splicing/Benign, and HiFI/Benign ratios were selected by Lasso for the prediction of TSGs, and the HiFI/Benign ratio and the Entropy score were selected for the prediction of OGs. TUSON predictions are based on the calculation of a combined p value (and q value) of the selected parameters by using an extended version of the Liptak method (Tables S3A and S3B). Based on the combined p values derived with the TUSON method, we estimated the total number of predicted TSGs and OGs by using a histogram-based method (Mosig et al., 2001).

Charm and Chrom Score and Correlation with Frequency of SCNAs

For each arm and chromosome, respectively, the Charm and Chrom scores for a certain gene set (TSGs, OGs, or essential genes) represent the density of the genes contained in that set weighted by their predicted potency. The potency of each gene corresponds to its rank position within its gene set list ranked by the TUSON p value or by the (LOF + 1/2 × HiFi)/Benign ratio for the essential genes. For the cumulative $Charm^{TSG\text{-}OG\text{-}Ess}$ and $Chrom^{TSG\text{-}OG\text{-}Ess}$ score, the scores of OGs and essential genes were subtracted from the scores relative to the TSGs. The correlation analysis was performed using one-sided Pearson's correlation test between the Charm or Chrom score and the frequency of deletion and amplification of each arm or chromosome across all tumors (Table S6).

Analysis of Functional Gene Sets

The STOP gene list was derived from an analysis performed using RNAi gene enrichment ranking (RIGER) algorithm (Cheung et al., 2011) on a previously described functional shRNA-based proliferation screen (Solimini et al., 2012; Table S5A). An in silico list of 332 essential genes was derived by considering the intersection between the lists of genes predicted to be housekeeping genes and highly conserved genes (Marcotte et al., 2012; Table S5A). We used the Fisher's exact test to examine the significance of the association between the presence of a gene in recurrent SCNAs (Beroukhim et al., 2010) and its presence among a certain gene set.

For additional information, see the Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.10.011.

ACKNOWLEDGMENTS

We thank Simon Forbes and Michael Stratton (Wellcome Trust Sanger Institute, UK) for the data from the Cosmic dataset (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/), Eric Wooten for help on data extraction, and Chad Creighton and Kim Rathmell for allowing access to their unpublished data on KICH SCNAs. We also thank Andrew Futreal, Semin Lee, David Livingston, David Page, Mary-Claire King, Jim Lupski, Rameen Beroukhim, Matt Meyerson, Atanas Kamburov, and Paul Edwards for helpful advice and Bert Vogelstein, Judith Glaven, and members of the Elledge lab for helpful comments on the manuscript. This work was funded by a DOD Breast Cancer Innovator Award and NIH grant to S.J.E., U54LM008748 to P.J.P., and K08DK081612 to J.C.Y. S.J.E. is an investigator with the Howard Hughes Medical Institute. We apologize to our colleagues whose papers we could not cite due to space limitations.

Received: August 20, 2013
Revised: September 26, 2013
Accepted: October 8, 2013
Published: October 31, 2013

REFERENCES

Adamson, B., Smogorzewska, A., Sigoillot, F.D., King, R.W., and Elledge, S.J. (2012). A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. Nat. Cell Biol. *14*, 318–328.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Bianchi, N.O. (2009). Y chromosome structural and functional changes in human malignant diseases. Mut. Res. *682*, 21–27.

Boveri, T. (1929). The Origin of Malignant Tumors (Baltimore, MD: Williams and Wilkins).

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Conti, P., Youinou, P., and Theoharides, T.C. (2007). Modulation of autoimmunity by the latest interleukins (with special emphasis on IL-32). Autoimmun. Rev. *6*, 131–137.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., et al. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. *38*(Database issue), D652–D657.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Gudjonsson, T., Altmeyer, M., Savic, V., Toledo, L., Dinant, C., Grøfte, M., Bartkova, J., Poulsen, M., Oka, Y., Bekker-Jensen, S., et al. (2012). TRIP12 and UBR5 suppress spreading of chromatin ubiquitylation at damaged chromosomes. Cell *150*, 697–709.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hou, J.M., Krebs, M., Ward, T., Sloane, R., Priest, L., Hughes, A., Clack, G., Ranson, M., Blackhall, F., and Dive, C. (2011). Circulating tumor cells as a window on metastasis biology in lung cancer. Am. J. Pathol. *178*, 989–996.

Kowalski, J., Morsberger, L.A., Blackford, A., Hawkins, A., Yeo, C.J., Hruban, R.H., and Griffin, C.A. (2007). Chromosomal abnormalities of adenocarcinoma of the pancreas: identifying early and late changes. Cancer Genet. Cytogenet. *178*, 26–35.

Lai, K., Amsterdam, A., Farrington, S., Bronson, R.T., Hopkins, N., and Lees, J.A. (2009). Many ribosomal protein mutations are associated with growth impairment and tumor predisposition in zebrafish. Dev. Dyn. *238*, 76–85.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Luo, J., Solimini, N.L., and Elledge, S.J. (2009). Principles of cancer therapy: oncogene and non-oncogene addiction. Cell *136*, 823–837.

Matsuoka, S., Ballif, B.A., Smogorzewska, A., McDonald, E.R., 3rd, Hurov, K.E., Luo, J., Bakalarski, C.E., Zhao, Z., Solimini, N., Lerenthal, Y., et al. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science *316*, 1160–1166.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nat. Rev. Genet. *11*, 685–696.

Mosig, M.O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann, A. (2001). A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics *157*, 1683–1698.

Ozery-Flato, M., Linhart, C., Trakhtenbrot, L., Israeli, S., and Shamir, R. (2011). Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. Genome Biol. *12*, R61.

Pavlova, N.N., Pallasch, C., Elia, A.E., Braun, C.J., Westbrook, T.F., Hemann, M., and Elledge, S.J. (2013). A role for PVRL4-driven cell-cell interactions in tumorigenesis. Elife *2*, e00358.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. *38*(Database issue), D497–D501.

Shannon, C.E., and Weaver, W. (1949). The Mathematical Theory of Communication (Urbana, IL: University of Illinois Press).

Solimini, N.L., Xu, Q., Mermel, C.H., Liang, A.C., Schlabach, M.R., Luo, J., Burrows, A.E., Anselmo, A.N., Bredemeyer, A.L., Li, M.Z., et al. (2012). Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science *337*, 104–109.

Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nat. Genet. *41*, 393–395.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*(Database issue), D535–D539.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Tang, Y.C., and Amon, A. (2013). Gene copy-number alterations: a cost-benefit analysis. Cell *152*, 394–405.

Vogelstein, B., and Kinzler, K.W. (1993). The multistep nature of cancer. Trends Genet. *9*, 138–141.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

White, J.K., Gerdin, A.K., Karp, N.A., Ryder, E., Buljan, M., Bussell, J.N., Salisbury, J., Clare, S., Ingham, N.J., Podrini, C., et al.; Sanger Institute Mouse Genetics Project. (2013). Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell *154*, 452–464.

Willig, T.N., Gazda, H., and Sieff, C.A. (2000). Diamond-Blackfan anemia. Curr. Opin. Hematol. *7*, 85–94.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108–1113.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.-Z., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

Zhang, D., Zaugg, K., Mak, T.W., and Elledge, S.J. (2006). A role for the deubiquitinating enzyme USP28 in control of the DNA-damage response. Cell *126*, 529–542.