Supplementary Information

Supplementary Figure Legends

Supp. Fig. 1 related to Fig. 1 Distribution of the number of mutations per samples across different tumor types

A boxplot-based representation of the distribution of the number of somatic mutations in the different tumor types, ordered according to their median is shown. The median, first quartile, third quartile and outliers in the distribution are shown.

Supp. Fig. 2 related to Fig. 2 Variation of the number of predicted TSGs with increasing number of samples

Subsets of the mutation dataset containing decreasing numbers of mutations are analyzed with the TUSON Explorer method and the number of predicted (true positive) TSGs is estimated through the method proposed by Mosig et al. (see Experimenthal Procedures, Mosig et al., 2001). Additionally, 812 additional samples were added to the current dataset from the published database recently described by Alexandrov et a., 2013 (Alexandrov et al., 2013) to generate the data point with the highest number of samples

Supp. Fig. 3. Related to Fig. 3 CORUM and Betweenness analysis on predicted OGs and TSGs

A) Analysis of the involvement of predicted TSGs and OGs in protein complexes. The CORUM dataset of human protein complexes was used to determine the fraction of TSGs and OGs that belong to known protein complexes. Corresponding p-values are shown as determined by the Wilcoxon test.

B) Betweenness analysis on the predicted TSGs and OGs. Enrichment of high betweenness of TSGs and OGs were computed on the interaction network from BioGRID using the Kruskal–Wallis one-way analysis of variance (p-values are shown).

Supp. Fig. 4 related to Fig. 6 Correlation analysis of density, Charm and Chrom for TSGs, OGs and essential genes and arm- and chromosome-level deletion and

amplification frequency

A correlation analysis (Pearson's correlation) of the arm-level or chromosome-level frequency of deletion and amplification with the indicated parameter among the relative density (Dens), Charm score, Chrom score for TSGs, OGs or Essential genes and different combination of these scores as indicated. The r-value and p-value for each correlation is shown. The lists of TSGs and OGs are the same used for Fig. 6.

Supp. Fig. 5 related to Fig. 6 Correlation analysis of density, Charm and Chrom for TSGs, OGs and essential genes and arm- and chromosome-level deletion and amplification frequency

A correlation analysis (Pearson's correlation) of the arm-level or chromosome-level frequency of deletion and amplification with the indicated parameter among the Charm score, Chrom score for TSGs, OGs or Essential genes in different combination of these scores as indicated. The r-value and p-value for each correlation is shown. The correlations shown in panels A-B are based on the lists of TSGs and OGs used for Fig. 6, while the correlations shown in panels C-F, marked with an asterisk, are based on the top 300 TSGs and top 250 OGs ranked by TUSON combined q-value, without taking into account additional parameters or cutoffs.

Supplementary Table Legends

Supp. Table 1 related to Fig. 1: List of tumor types, N of samples and mutations in each.

Supp. Table 1 contains the list of the tumor types present in the dataset their relative number of samples and total number of mutations. In addition, this table shows the tumor types for which an individual analysis is shown in Supp. Table 4.

Supp. Table 2 related to Fig. 2: Lists of TSG and OG training sets and the Neutral genes set and Behavior of parameters in the binary classifications

- Supp. Table 2a: **Lists of TSG and OG training sets and the Neutral genes set**
  Supp. Table 2a contains the TSG and OG training sets as well as the Neutral gene set, derived as described in the methods.
- Supp. Table 2b: **Behavior of parameters in the binary classifications**
  Supp. Table 2b shows the 22 parameters for the prediction of TSGs and OGs, together with their median in the indicated gene set: Neutral genes, TSG training set and OG training set.

Supp. Table 3 related to Fig. 3: Ranking of TSGs and OGs by TUSON Explorer and Lasso and related analyses

- Supp. Table 3a: **Ranking of TSGs by TUSON Explorer and Lasso**
  This table shows all the parameters used for the prediction of TSGs with relative p-values and q-values (see methods), as indicated. The Total N of Silent, Missense, LOF and Splicing mutations, the coding sequence length (CDS), the ratios of the different classes of mutations and the corresponding p-values and q-values are shown as specified. In addition, the p-value and q-value for deletion and amplification are shown. Columns X and Y contain the combined p-value and q-value from TUSON Explorer method and column Z contains the probability of gene of being TSG as predicted by the Lasso method.
- Supp. Table 3b: **Ranking of OGs by TUSON Explorer and Lasso**
  This table shows all the parameters used for the prediction of OGs with relative p-values and q-values (see methods), as indicated. The Total N of Silent, Missense, Entropy Score, the CDS and the ratios of different classes of mutations and their p-values and q-values are shown as specified. In addition, the p-value and q-value for deletion and amplification are shown. Columns X and Y contain the combined p-value and q-value from TUSON Explorer method and column Z contains the probability of gene of being OG as predicted by the Lasso method.
- Supp. Table 3c: **GO-term enrichment analysis for TSGs predicted by TUSON**
  GO-term enrichment analysis using DAVID of the top 300 candidate TSGs predicted by the PAN-Cancer analysis (http://david.abcc.ncifcrf.gov/, Huang et al., 2009).

- Supp. Table 3d: **Analysis and ranking of familial TSGs in sporadic cancer**
  A list of familial TSGs with the corresponding p-values and q-values derived from the TUSON Explorer analysis on sporadic tumors.

Supp. Table 4 related to Fig. 3: TUSON Explorer prediction of TSGs and OGs on single tumor types and related analysis

- Supp. Table 4a: **TUSON Explorer prediction of TSGs on single tumor types**
  Supp. Table 4a contain the TUSON Explorer p-value and q-value for each of the indicated tumor type for the prediction of TSGs.
- Supp. Table 4b: **TUSON Explorer prediction of OGs on single tumor types**
  Supp. Table 4b contain the TUSON Explorer p-value and q-value for each of the indicated tumor type for the prediction of OGs
- Supp. Table 4c: **Analysis of TSGs of single tumor types compared to the TSGs in the PAN-Cancer analysis**
  For each of indicated tumor types Supp. Table 4c contains the number of TSGs predicted by the PAN-Cancer analysis on the mutation dataset after removing the corresponding type of tumor (column B), the number of TSGs predicted by the analysis of the indicated specific tumor type (column C). Column E and F report the number and % of TSGs among the tissue specific ones (column C) not found in the PAN-cancer TSGs (column B).

Supp. Table 5 related to Fig. 5: Analysis of functional gene sets, analysis of TSGs on the chromosomes and analysis of mutations in females versus males

- Supp. Table 5a: **Analysis of functional gene sets (STOP and Essential genes)**
  Supp. Table 5a contains the lists of STOP genes and the two lists of *in silico* derived Essential genes (145 and 332 genes) as described in the methods. In addition, this table reports the list of genes localized in regions of recurrent focal deletions or amplifications (Beroukhim et al., 2010).
- Supp. Table 5b: **Relative density of TSG on the chromosomes.**
  Supp. Table 5b shows the relative density of predicted TSGs (first 300 genes from

the TUSON Explorer prediction, combined p-value) on the different chromosomes, including its deviation from the expected density from the overall gene distribution on the chromosomes and the significance (p-value) of this deviation (one tailed binomial test).

- Supp. Table 5c: **Mutation density on X and autosomes in males and females**
  Supp. Table 5c shows the analysis performed on males and females tumor samples relative to the occurrence of Silent and LOF mutations on the chromosome X. The total N of samples, the density of silent mutations on the X (normalized for the number of copies of X) and the number of LOF and Silent mutations on the TSGs on the X chromosome in males and females are shown.

Supp. Table 6 related to Fig. 6: Density, Charm, Chrom scores for TSGs, OGs and Essential genes and frequency of deletion and amplification across cancers

Supp. Table 6 shows the values of the various parameters used to establish the correlation between SCNAs data and the gene density, the Charm scores and Chrom scores for TSGs, OGs and the Essential genes. For each arm and chromosome and each class of genes (TSGs, OGs and Essential genes) we show the density, Charm and Chrom score. Additionally, the $Charm^{TSG\text{-}OG}$ and the $Charm^{TSG\text{-}OG\text{-}Ess}$ are shown (see experimental procedure for the method of calculation). Lastly, the average frequency of arm- and chromosome-level deletion and amplification is shown. Supp. Table 6a contains the values used for the tighter list of TSGs and OGs using stringency cutoffs (see methods); 6b contains the values used for the top 300 TSGs and top 250 OGs after ranking according to the TUSON q-value, without stringency cutoffs.

Supp. Table 7 related to Fig. 3: Manually curated list of genes predicted by TUSON Explorer to have TSG-like or OG-like signatures.

- Supp. Table 7a: **Manually curated list of genes predicted by TUSON Explorer to have TSG-like signatures**
  Supp. Table 7a lists the gene symbol, description, key parameters, and heuristically-derived confidence level for the top 1000 genes ranked by TUSON

to have a pattern of mutation similar to TSGs. Q-values for the parameters utilized by TUSON are included, along with the combined p-values and q-values (TUSON Explorer), and the probability of each gene being a TSG as predicted by Lasso. A manual confidence column indicates our overall level of confidence of each gene being a true TSG (see supplementary methods). We report GO terms and KEGG pathways associated with each gene. For a subset of genes, we summarize evidence from the literature implicating the gene as a cancer driver or suggesting a putative functional role in cancer. Relevant publications are referenced by their PMID. We categorize the literature status of these genes as Known, Putative, Unclear or Context-dependent, or Novel (see supplementary methods).

- Supp. Table 7b: **Manually curated list of genes predicted by TUSON Explorer to have OG-like signatures**

  Supp. Table 7b lists the gene symbol, description, key parameters, and heuristically-derived confidence level for the top 1000 genes ranked by TUSON to have a pattern of mutation similar to OGs. Q-values for the parameters utilized by TUSON are included, along with the combined p-values and q-values (TUSON Explorer), and the probability of each gene being an OG as predicted by Lasso. A manual confidence column indicates our overall level of confidence of each gene being a true OG (see supplementary methods). We report GO terms and KEGG pathways associated with each gene. For a subset of genes, we summarize evidence from the literature implicating the gene as a cancer driver or suggesting a putative functional role in cancer. Relevant publications are referenced by their PMID. We categorize the literature status of these genes as Known, Putative, Unclear or Context-dependent, or Novel (see supplementary methods).

Supplemental Experimental Procedures

Somatic mutation dataset and SCNA data.

The dataset of somatic mutations included data from all-exome and all-genome sequencing published by Alexandrov et al. (Alexandrov et al., Nature 2013) (3653 tumors), from the Catalogue Of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) (738 tumors) and data from the Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) research network (3816 tumors). Tumor samples present in more than one dataset were excluded and the final dataset contained 1,195,228 mutations from 8207 tumor samples from more than 20 tumor types (Supp. Table 1). Hypermutated tumors with more than 2000 mutations in coding sequences where excluded from the analysis. This dataset was used for the mutational analysis and TUSON Explorer predictions and will be available at http://elledgelab.med.harvard.edu/. All data related to SCNAs were derived from Zach et al., 2013.

Definition of classes of mutations

We utilized the PolyPhen2 Hum-Var prediction model (Adzhubei et al., 2010) to weight the functional impact of each missense mutation and to classify them as high functional impact (HiFI) or low functional impact (LoFI) based on their probability of being damaging predicted by the PolyPhen2 HumVar algorithm. The HiFI and LoFI mutations were defined as mutations with a probability of being damaging higher than 0.447 (corresponding to Possibly or Probably damaging in Hum-Var prediction) or lower than 0.25, respectively. Based on this predicted functional impact, we defined these four classes of mutations, which were used to define the parameters in Lasso (see below):

- Benign mutations: Silent + LoFI Missense
- Loss of Function mutations (LOF): Nonsense and Frameshift mutations
- Splicing mutations: mutations affecting splicing sites (more than 95% of splicing mutations are in the first two positions at donor or acceptor sites)
- HiFI missense mutations (damaging missense mutations)

Entropy of missense mutations

Mutations often occur recurrently on a few residues in oncogenes, suggesting that tumor evolution selection acts to favor of them. We measure the strength of the selection force by the entropy score, which measures the degree of randomness in where the mutations occur. Advantages of using entropy include 1) it can sensitively detect mild selection forces distributed over many residues; 2) the score is largely independent of gene size or the total number of mutations.

Given a protein, we represent a somatic mutation by $M_i = (p_i, A_i)$, where p_i represents its location on the protein, and A_i represents the new amino acid into which it has mutated. We scan the protein and obtain the number of occurrence, n_i, for each mutation type M_i, the total number of different mutation types k, and the total number of all mutations n. The observed frequency for each mutation type M_i is computed as $f_i = n_i/n$. The entropy of the observed data is calculated as

$$S = \sum_{i=1}^{k} -f_i \cdot \ln(f_i),$$

where ln is the natural logarithm. The maximum entropy, where each mutation type has the same probability $p_i = 1/k$, is

$$S_0 = \sum_{i=1}^{k} -p_i \cdot \ln(p_i) = \ln(k).$$

We use the difference between the two scores, $\Delta S = S_0 - S$, to measure the selection force over the whole protein.

Under a null random model, the probability of observing k mutation types with occurrences $n_1, n_2, \ldots, n_k$ out of a total n mutations is

$$P = \binom{n}{n_1, n_2, \ldots\ldots, n_k} \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i}.$$

Using the Sterling approximation, we have

$$\ln(P) \approx n \cdot \Delta S.$$

The p-values were computed based on the above formula. We normalized *n* for all proteins to the average number of mutations in well-known oncogenes to avoid artifactual effects due to large gene sizes. For the entropy score calculation, hypermutated samples deriving from melanoma patients were excluded.

Definition of primary parameters for the prediction of OG and TSG

Based on the classification of mutations described above, the entropy score calculation and other gene-specific features we selected a set of 22 primary parameters associated with each gene for the prediction of TSG and OG.

1. Silent mutations/kb
2. Total N Missense mutations
3. Total N LOF mutations
4. Total N of Splicing mutations
5. Missense mutations/kb
6. LOF mutations/kb
7. Entropy score for missense mutations
8. LOF/Silent ratio
9. Splicing/Silent ratio
10. Missense/Silent ratio
11. HiFI missense/LoFI missense ratio
12. LOF/Benign ratio
13. Splicing/Benign ratio
14. Missense/Benign ratio
15. HiFI missense/Benign ratio
16. Average PolyPhen2 score for missense mutations
17. LOF/Total mutations
18. Missense/Total mutations
19. Splicing/Total mutations
20. LOF/Missense mutations
21. High-level Deletion frequency
22. High-level Amplification frequency

For the parameters 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, and 20 that involve ratios, a pseudo-count estimated from the median of each parameter calculated on all genes was added. Let us consider a gene G_i with the number of mutations M_i^A M_i^B in class A and B, along with their medians m^A and m^B. The parameter p_i^{AB} involving the ratio of mutations in class A and B was calculated as follows:

$$p_i^{AB} = \frac{M_i^A + \frac{1}{2} m^A}{M_i^B + \frac{1}{2} m^B}$$

Derivation of TSG, OG and Neutral Gene (NG) training sets

To derive the TSG and OG training sets, we started from the entire list of 468 genes reported to be mutated (somatic or germline) in cancer from the Cancer Gene Census (Futreal et al., 2004) (http://cancer.sanger.ac.uk/cancergenome/projects/census/). We excluded genes reported in translocations. We considered genes whose mutations were reported as Dominant for OG and reported as Recessive for TSG. Furthermore, we selected the most high confidence TSG and OG as those genes that have been implicated in tumorigenesis by strong experimental evidence in model organisms. The final lists of TSG and OG contained 50 genes each and are reported in Supplemental Table 2**.**

To derive the list of Neutral genes (NG), we considered the entire gene list (18700 protein coding genes) and we excluded all the genes satisfying any of the following criteria: a) belonging to the Cancer Gene Census list of mutated or rearranged genes; b) having been previously implicated in any Entrez databases at NCBI as oncogene or tumor suppressor gene (http://cbio.mskcc.org/CancerGenes/DescribeMethods.action); and c) belonging to a list of housekeeping genes or genes highly conserved across evolution (see Derivation of Functional gene sets). The final list of NG (10,900 genes) is contained in Supplemental Table 2.

Predictions based on classification models

Classification models were used to train and predict the probabilities of being a tumor suppressor gene or an oncogene for each gene. We used a statistical method called Lasso (least absolute shrinkage and selection operator) to narrow down our original list of 22

parameters. This method minimizes the residual sum of squares as in standard linear regression but with the added constraint on the sum of the absolute values of the coefficients (this tends to shrink some coefficients to zero and allows removal of those variables to simplify the model). We curated the final lists of parameters based on Lasso selection results, with manual removal of some obviously biased parameters such as the total number of mutations.

We first performed three binary classifications, between 1) TSG and OG, 2) TSG and NG and 3) OG and NG. Each classification estimated probabilities of each gene being in one of the classes. The final probabilities of being a TSG or OG were obtained by averaging the results of three classifiers. We tested both the Lasso classifier and the well-known Supporting Vector Machine (SVM) (Cortes and Vapnik, 1995). To compensate for Lasso's inflexibility of taking account of non-linear effects, we transformed each parameter x into 6 new parameters: x, x^2, x^3, x^4, x^5 and $\ln(x)$. The lambda value was determined by using a 20-fold cross validation approach. SVM is known for its efficient computation and for taking account of potentially non-linear effects of parameters. To determine the optimal weights for parameters, we performed a computationally-heavy grid search. We used the Gaussian Radial Basis function as the kernel, and used the R package e1071 for computation. Comparing the cross-validation results from Lasso and SVM, we found that Lasso always displayed higher accuracy; thus the final predictions were performed by Lasso only. To avoid issues associated with imbalanced classifications when the 10,900 NG set was used, we generated 20,000 random small NG sets of size 150 each. Each NG was covered 275 times among these random sets. With each random set, we performed a Lasso classification against either TSG or OG. The final results were averaged over these 20,000 results.

Tumor Suppressor and Oncogene Explorer (TUSON Explorer)

p-value and q-value calculation First we derived a p-value for each of the following parameters selected through the Lasso method for the prediction of TSGs or OGs. The parameters selected through Lasso for the prediction of TSGs were the following 3 parameters:

1. LOF/Benign ratio (Lasso coefficient β = -7.96)
2. HiFi Missense/Benign ratio (Lasso coefficient β =-0.61)
3. Splicing/Benign ratio (Lasso coefficient β = 0.89)

The parameters selected through Lasso for the prediction of OGs were the following 2 parameters:

1. Entropy Score (Lasso coefficient β = -1.21)
2. HiFi Missense/Benign ratio (Lasso coefficient β = 1.06)

Additionally, LOF/Benign was selected by Lasso as the best parameters for discriminating between TSGs and OGs (Lasso coefficient β =-0.79).

For the Entropy score, the p-values were calculated as described above. For the LOF/Benign, Splicing/Benign, and HiFI missense/Benign ratios, the p-values were computed by using the exact binomial test as follows. Let us consider a gene G_i among all genes G and define S_i,H_i and B_i to be its number of Splicing, HiFI missense and Benign mutations, respectively. Let N be the set of NG. We define E_S and E_H as

$$E_S = \sum_{j|G_j \in N} S_j \Big/ \sum_{j|G_j \in N} B_j$$

$$E_H = \sum_{j|G_j \in N} H_j \Big/ \sum_{j|G_j \in N} B_j$$

We compute the p-value for Splicing/Benign p_s and HiFI missense/Benign p_H for gene i as follows:

$$p_{Si} = \sum_{j=S_i}^{S_i+B_i} \binom{S_i + B_i}{j} \cdot \left(\frac{E_S}{1+E_S}\right)^j \cdot \left(\frac{1}{1+E_S}\right)^{S_i+B_i-j}$$

$$p_{Hi} = \sum_{j=H_i}^{H_i+B_i} \binom{H_i + B_i}{j} \cdot \left(\frac{E_H}{1+E_H}\right)^j \cdot \left(\frac{1}{1+E_H}\right)^{H_i+B_i-j}$$

For the LOF/Benign parameter, we applied an additional normalization in relationship to the non-homogeneous codon usage among genes. Since we are interested in the frequency of LOF mutations per gene, that number can deviate from the average simply on the basis of codon usage. The vast majority of nonsense mutations are single base changes and only 18 of 61 codons can mutate to a stop codon by a single nucleotide change. Therefore we normalized the number of nonsense mutations to the number of potential nonsense codons**.** The normalization factor for a gene is given by the ratio of the density of possible nonsense codons in the gene divided by the average density among all genes in the genome, as described below.

Let us define N_i, F_i to be the number of Nonsense and Frameshift mutations for gene G_i. We also define t_i as the number of possible nonstop codons (codons among the 18 codons that can be mutated to a nonsense by single nucleotide substitution) and n_i the total number of codons of gene G_i. The average density of possible nonstop codons among all genes is

$$d = \frac{\Sigma_j c_j}{\Sigma_j n_j}.$$

The relative density of possible nonsense codons for gene G_i is

$$d_i = \frac{t_i}{n_i} \cdot \frac{1}{d}.$$

We compute expected ratio of Nonsense/Benign E_{Ni} for the gene G_i as

$$E_{Ni} = \frac{d_i . \sum_{j|G_j \in N} N_j + \sum_{j|G_j \in N} F_j}{\sum_{j|G_j \in N} B_j}.$$

Using the exact binomial test, the p-value $p_{Li}$for LOF/Benign mutations for gene G_i is computed as follows:

$$p_{Li} = \sum_{j=N_i+F_i}^{N_i+F_i+B_i} \binom{N_i + F_i + B_i}{j} \cdot \left(\frac{E_{Ni}}{1+E_{Ni}}\right)^j \cdot \left(\frac{1}{1+E_{Ni}}\right)^{N_i+F_i+B_i-j}.$$

Furthermore, an additional normalization was applied to genes with a high CDS length. We first noticed that the distribution of CDS length has a long right tail with the skewness coefficient of 9. Likewise, a similar degree of skewness is present in the distribution of silent mutations, which mainly depends on gene length. We thus normalized the number of each class of mutations for all the genes whose total number of silent mutations is higher than 30 (2 standard deviation away from the mean). To illustrate this, if a gene G_i had a number of silent mutations S_i higher than 30, the numbers of its mutations M_i in each class of mutations (nonsense, frameshift, silent and HiFI) are normalized to $normM_i = M_i \cdot S_i/30$. The normalized number $normM_i$ of the mutations in each class was then used to compute the p-values as described above.

Calculation of a combined p-value

To derive a combined p-value, we combined the p-values deriving from the 3 or 2 parameters selected by Lasso for the prediction of TSGs and OGs, respectively. For this purpose, we used an extension of the Liptak's method (Liptak, 1958) with correlation correction because 1) correlations were observed between parameters and 2) different parameters have different effect sizes and need to have different weights. A grid search method was used to find the optimal weights which results in the best rankings for the TSG or OG genes in the training sets. For each gene, p-values from different parameters were transformed into Z-scores, $\boldsymbol{Z} = (Z_1, Z_2)$. The combined Z-score is computed as $Z_c = \frac{\boldsymbol{w} \cdot \boldsymbol{Z}}{\sqrt{\boldsymbol{w} \cdot \boldsymbol{\rho} \cdot \boldsymbol{w}'}}$, where $\boldsymbol{w}$ is the weight and $\boldsymbol{\rho}$ is the correlation matrix. The combined Z-score is then transformed back into the combined p-value. For TSG predictions, we combined p-values from LOF/Benign, Splicing/Benign and HiFI/Benign parameters. For OG predictions, we combine p-values from Missense entropy and HiFI/Benign parameters. The q-value was computed by using the Benjamini & Hochberg method (Benjamini and Hochberg, 1995).

The LOF/Benign ratio was predicted by Lasso as the best parameter discriminating between TSGs and OGs. For the final prediction of the OGs, in order to discriminate between OGs and TSGs, the genes having a q-value for LOF/Benign ratio lower than 0.3 were excluded from the final list. Conversely, for the TSGs, the genes having a q-value of

1 for the LOF/Benign parameter were excluded from the final list.

Estimate of the number of cancer genes

We implemented a histogram-based method proposed by Mosig et al. (Mosig et al., 2001) to estimate the numbers of TSGs and OGs, given the combined p-values. The method assumes a uniform distribution for the p-values of null hypotheses and estimates its number. The number of rejected hypotheses then corresponds to the number of TSG genes or OG genes. In addition, we validated the Mosig analysis by applying a simple FDR-based method to the ranking made by a single parameter using the q-value. This entails choosing a group of ranked genes starting from the top-ranked gene and multiplying that number by (1-q) for the last gene in the list to reveal the number of statistically significant gene signatures on the list. This analysis confirmed the conclusions from the Mosig method.

In addition to estimating the N of TSGs and OGs predicted by TUSON Explorer on the dataset used for the mutation analysis, we also assessed how the number of predictive TSGs by the analysis described above changes by analyzing datasets composed of progressively increasing numbers of samples and total N of mutations (Supp. Fig. 2). Random subsets (10 for each data point shown in the figure) of the mutation dataset were analyzed by TUSON Explorer method and the N of predicted TSGs was assessed. Additionally, 812 more samples were added to the current dataset from the published database recently described by Alexandrov et a., 2013 (Alexandrov et al., 2013) to generate the data point shown in the graph corresponding to the dataset with the highest number of samples (Supp. Fig. 2).

Estimating a list of p-values from the corresponding list of q-values

We designed a novel statistical method to convert a list of q-values into a list of p-values. This is needed as the significance of CNVs in GISTIC2 results are in the form of q-values, while p-values are needed to obtain the combined p-values (Mermel et al., 2011). We basically reversed the Benjamini & Hochberg method of calculating of q-values. If a CNV segment has q-value q and rank i among a total of n segments, the p-value is $p = q \cdot i/n$. When two segments with ranks $i - 1$ and i have the same q-value q, the p-

value for the $(i-1)$th segment is $p = q \cdot (i-0.5)/n$. The p-values generated in such a way are unbiased estimates from the corresponding list of q-values. These p-values and q-values for deletion and amplification are in Supp. Table 3a, b.

CORUM analysis

The CORUM database contains an experimentally-validated dataset of human protein complexes and was employed to determine the involvement of TSG and OG in protein complexes (Ruepp et al., 2010). Two-tailed binomial test was used to test the significance of the enrichment in genes involved in protein complexes.

Betweenness centrality

Betweenness centrality measures the importance or centrality of each gene in the network. For a network with n genes, define $\sigma_{i,j}$ as the total number of shortest paths between gene i and gene j, and $\sigma_{i,j}^{k}$ as the total number of $\sigma_{i,j}$ that passing through gene k. The betweenness centrality for gene k is defined as:

$$bw(k) = \frac{2}{(n-1)\cdot(n-2)} \cdot \sum_{\mathrm{i}\neq\mathrm{j}\neq\mathrm{k}} \frac{\sigma_{i,j}^{k}}{\sigma_{i,j}}.$$

We downloaded the human gene interaction network from BioGRID, which contains 15,843 genes and 217,215 interactions. We used the Brandes' algorithm (Brandes, 2001) to compute the betweenness centrality values for the full network. Kruskal–Wallis one-way analysis of variance was used to access whether the predicted TSGs and OGs have significantly larger betweenness centrality values compared to genes in the whole network.

Charm and Chrom score and correlation analysis

The Charm and Chrom scores represent a score assigned, respectively, to each chromosome arm or whole chromosome, which depend on the relative density and

potency of TSGs and OGs (and occasionally Essential genes) contained in it. We used density instead of the total number of TSGs, OGs and essential genes per arm because it is likely there is a penalty due to proteotoxic stress that occurs upon changing the dosage of many genes at once and this should be proportional to the number of genes in the region of interest. Thus, to approximate this penalty we normalized to the total number of genes per arm or chromosome. To determine the Charm and Chrom scores, we first selected the TSGs and OGS based on the following parameters derived from TUSON Explorer, a method based only on the mutation profile of each gene, independent of any SCNA information. The correlation analyses using the top 300 TSGs and top 250 OGs after ranking based on the TUSON p-values are shown in Supp. Fig. 5C-F. In Fig. 6, Supp. Fig. 4 and Supp. Fig. 5A-B we show the correlation analyses based on a more stringent list of TSGs and OGs selected as described below. We selected the TSGs having a combined q-value<0.25, a q-value for the LOF/Benign ratio <0.25 and a minimum of 8 LOF mutations (264 TSGs in total). For the OGs we selected the genes having a combined q-value<0.35, a q-value for the Entropy score <0.35 and a minimum of 10 Missense mutations (219 OGs in total). The correlation analysis for these lists is shown in Fig. 6, Supp. Fig. 4-5 and Supp. Table 6. In all cases the potency of each TSG and OG was estimated by its rank position on the list ranked by the combined q-value for TSGs and OGs respectively determined by TUSON Explorer.

Given our ranked list of T predicted TSGs and O predicted OGs among all genes G, each TSG or OG is assigned a weight w equal to $T\text{-}r$ *or* $O\text{-}r$, where r is the rank position of that gene in its respective list. Let us now consider an arm i and define T_i as the TSGs contained in that arm and O_i the OGs contained in it. In addition, N is the total number of genes contained in that arm. We determine the Charm scores for that arm as follows:

$$Charm_i^{TSG} = \sum_{j|G_j \in T_i} w_j / N_i; \; Charm_i^{OG} = \sum_{j|G_j \in O_i} w_j / N_i$$

$_i$

In addition to the TSGs and OGs, we also included our *in silico* list of Essential genes (332 genes, Supp. Table 5a). The importance of Essential genes was estimated based on

the (LOF + 1/2·HiFI)/Benign ratio, with genes with lower ratios having a higher weight (all the mutations in the dataset including hypermutated samples were used). Similarly to the TSGs and OGs, we determined a Charm score for the Essential genes $Charm_i^{Ess}$.

In order to combine the cumulative effects of TSGs and OGs on each arm, we derived a combined score. The $Charm_i^{Ess}$ and the $Charm_i^{OG}$ were given negative weights as their effects on tumorigenesis go in opposite directions of TSGs. Additionally, to derive this combined score, a normalization factor was used to equalize the total effects of the different gene lists. Given the number of all genes N and the $Charm_i^{TSG}$, $Charm_i^{OG}$ and $Charm_i{}^{Ess}$ scores, the combined score for arm *i* is as follows:

$$Charm_i^{TSG-OG-Ess} = Charm_i^{TSG} - \frac{\sum_i Charm^{TSG}}{\sum_i Charm^{OG}} \cdot Charm_i^{OG} - \frac{\sum_i Charm^{TSG}}{\sum_i Charm^{Ess}} \cdot Charm_i^{Ess}$$

The $Charm_i^{TSG-OG}$ was calculated in a similar way but omitting the $Charm_i^{Ess}$ term.

The Chrom score was calculated in a similar way to the Charm score, considering each entire chromosome instead of each arm individually.

For the correlation test, we used the one-sided Pearson's correlation. We determined the correlation between the frequency of arm-level deletion or amplification with the Charm scores and the frequency of chromosome-level deletion or amplification with the Chrom scores. In addition to the Charm and Chrom scores we also determined the correlation using the relative density of TSGs, OGs and essential genes, in the absence of a rank-based weighting.

Data for SCNAs was derived from the following tumor types (Zach et al., 2013). The tumor types are BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, STAD, SKCM, THCA, UCEC and KICH. The frequency of the different types of SCNAs was determined as described below. For each available tumor type, we calculated the frequencies of samples having deletions, high level deletions, amplifications and high level amplifications using the thresholds -0.415, -2, 0.32 and 0.807 respectively for the log copy ratios. The overall frequency of arm level SCNAs is the unweighted average

among all tumor types.

Functional gene sets (STOP and Essential) and analysis of their density in focal deletions

To define the list of STOP genes, we employed the data from functional genome-wide shRNA based proliferation screening (Solimini et al., 2012), specifically from the secondary validation screen. We performed an analysis using RNAi gene enrichment ranking (RIGER) algorithm (Cheung et al., 2011), using Kolmogorov-Smirnov statistics. The genes with a p-value <0.005 were considered. The final list of STOP genes is in Supplemental Table 8.

To define a list of genes predicted to be Essential, from the KEGG database we determine a list of genes predicted to be essential as being part of the following crucial biological processes including DNA replication, RNA transcription, mRNA transport, tRNA synthesis, RNA splicing, glycolysis, oxidative phosphorylation, fatty acid biosynthesis, fatty acid metabolism, purine metabolism, pyrimidine metabolism and amino acid metabolism (650 genes, KEGG list). Furthermore, we considered a list containing housekeeping genes and highly conserved genes (Marcotte et al., 2012). The housekeeping genes list contains 1722 genes expressed in more than 90% of analyzed tissues in a human expression database (Su et al., 2004). The list of highly conserved genes contains 1617 identified in 8 different species (*A.thaliana, B. taurus, C. elegans, C. familiaris, M. mulatta, M. musculus, R. norvegicus, and S. cerevisiae*), as determined by Paranoid (O'Brien and Fraser, 2005, Marcotte et al., 2012). The final list of Essential genes used in the functional gene set analysis contains 150 genes and was determined as the genes belonging to all three lists: KEGG, the housekeeping genes list and the highly conserved genes list (Supplemental Table 8). For the correlation analysis shown in Supplemental Fig. 4 a larger list of 332 genes was used, represented by genes belonging to both the housekeeping and the ortholog genes lists.

We determined whether there was a significant enrichment or depletion of STOP and Essential genes, respectively, within recurrent focal deletions. To this aim, we

considered the 82 regions of recurrent focal deletions defined in Beroukhim et al., 2010 (Beroukhim et al., 2010). We considered the genes found in recurrent deletions and how many among those belonged to the STOP or Essential genes. We used the Fisher's exact test to examine the significance of the association (contingency) between the presence of a gene in recurrent deletions and its presence among the Essential genes or the STOP genes.

Analysis of mutations in tumors from males and females

We used the data deriving from TCGA only because of the availability of gender information, after removal of the hypermutated tumors (see Somatic mutation dataset and SCNA data). We extracted the male-female information on a total of 627 female samples and 990 male samples of the following tumor types that do not have a strong gender-specificity: glioblastoma multiforme, head and neck adenocarcinoma, kidney clear cell carcinoma, lung adenocarcinoma and lung squamous cell carcinoma. To estimate the mutation background on the X chromosome and autosomes, we determined the average density of silent mutation per Mb of total coding region on X chromosome and on autosomes in each tumor (Supplemental Table 10). We employed the Wilcoxon test to test the significance of the difference between the mutation density on autosomes and the mutation density on the X chromosome in males and females. Additionally, we also derived the number of LOF and Silent mutations occurring on the TSGs present on the X in males and females (Supplemental Table 10).

Analysis of individual tumor types

We performed the analysis of 20 individual tumor types indicated in Supplemental Table 1. Hematological and lymphoid malignances were analyzed together and indicated as ALL.Hematological cancers. The analysis on single tumor types was performed using the TUSON Explorer method described in the previous sections for the PAN-Cancer analysis on the entire dataset. The q-values were calculated on the genes showing at least 3 mutations in the LOF and Silent mutation sets for TSGs and at least 5 missense mutations for OGs. The combined p-values and q-values for TSG and OG for the single tumor types are reported in Supplemental Table 7. In addition, for each of the 20 tumor types

indicated in Supp. Table 4c, we applied our method to the entire dataset excluding each individual tumor type, one at a time and compared the predicted TSGs found in this analysis (q-value<0.25) with the TSGs found in the analysis of the individual tumor type (q-value<0.25, Supp. Table 4c).

Heuristic approaches to define a manually curated list of TSGs and OGs

We collected additional information about TSGs and OGs, including SCNAs, gene function, and supporting literature, to assign an overall confidence level for each gene (increasing levels of confidence from 1 to 4). All genes were assigned a default confidence level of "1." For TSGs, confidence level was increased to "2" if more than 8 LOF or more than 15 overall mutations were analyzed. Confidence level was further increased to "3" if the gene deletion frequency was in the top 10% for all genes, if the gene was ranked 250th or better by Lasso, or if we found significant literature evidence supporting the gene's role as a TSG. Confidence level was increased to "4" if the gene met two or more of these criteria. For OGs, confidence level was increased to "2" if more than 12 missense mutations were analyzed. Confidence level was further increased to "3" if the gene amplification frequency was in the top 10% for all genes, if the gene was ranked 250th or better by Lasso, or if we found significant literature evidence supporting the gene's role as an OG. Confidence level was increased to "4" if the gene met two or more of these criteria.

For a subset of TSGs and OGs ranked highly by TUSON, we summarized any current literature evidence supporting each gene's role as a cancer driver. Based on these literature searches, we categorized the literature status of each gene as: Known, Putative, Unclear/context-dependent, or Novel. We classified genes as novel if there are no papers reporting evidence of alteration or a functional role in cancer.

For some genes, the existing evidence supports roles as both a TSG and OG, sometimes dependent upon tissue-specific or context-related factors; we classified these genes as unclear or context-dependent. We do not assert that these literature summaries are comprehensive, but rather that they highlight selected findings which inform our overall confidence for some of the genes.

For TSGs we assigned 83 genes a confidence level of "4", 125 a level of "3", 322 a level of "2" and 470 a confidence level of "1". Among the confidence level "4" genes, 4 were characterized as novel based on an absence of literature evidence suggesting a connection to cancer. Among the confidence level "3" genes, an additional 52 were characterized as novel.

For OGs we assigned 55 genes a confidence level of "4", 147 a level of "3", 699 a level of "2" and 99 a confidence level of "1". Among the 66 confidence level "4" and "3" genes for which literature searches were performed, 15 novel genes were identified based on absence of literature evidence suggesting a connection to cancer.

Note on RPL22 as a potential cancer driver

The most frequently mutated ribosomal gene in our study, RPL22, which ranked 22 on our TSG list, had a very unusual pattern of LOF mutations, primarily occurring as a frameshift in a stretch of As in tumor types associated with microsatellite-instability (for example uterine corpus endometrioid carcinomas, (Esteller et al., 1998; Goodfellow et al., 2003). While it could be non-phenotypic, its presence in ~10% of specific tumor types, such as uterine corpus endometrioid carcinoma, suggests that it is under positive selection, as recently suggested (Cancer Genome Atlas Research et al., 2013)

Estimates of Haploinsufficiency

We have performed three sets of analyses to detect happloinsufficiency based on the pattern of enrichment or depletion in focal deletions. The previous analysis of recurring focal deletions found an enrichment of ~20% of STOP genes and 22% depletion of GO genes in these regions suggesting a minimum of ~20% haploinsufficiency. As we do not have a definitive list of essential genes we used different functional and bioinformatics methods to obtain lists of potential GO genes. A similar analysis with a different set of 476 potential essential GO genes for human cells derived from the analysis of three independent shRNA based screens (Cheung et al.,

2011; Marcotte et al., 2012; Solimini et al., 2012) showed a 30% depletion from deletion regions (not shown). A third set of set of 332 GO genes predicted to be essential based on their conservation among different species and their expression in >90% of all human tissues showed ~27% fewer LOF/Silent mutations and a significant exclusion from focal deletions (45% more than expected), confirming the 30% estimate of haploisufficiency. Like the Cancer Gene Island hypothesis for focal deletions (Solimini et al., 2012), these observations argue that a substantial proportion of genes are haploinsufficient, 30-45%, and that deletions gain their selective advantage by the sum of the haploinsufficiency effects of the genes within that deletion. Additional evidence for haploinsufficiency comes from the studies of Xue et al, where they examined the genes on 8p22 that are frequently hemizygously deleted in liver cancer and found that multiple genes in that region acted cooperatively to restrain tumorigenesis (Xue et al., 2012). Together, our conservative estimate is 30% haploinsufficiency, between 20 and 40%.

Supplemental references

Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. *57*,289-300.

Bernal, M., Garcia-Alcalde, F., Concha, A., Cano, C., Blanco, A., Garrido, F., and Ruiz-Cabello, F. (2012). Genome-wide differential genetic profiling characterizes colorectal cancers with genetic instability and specific routes to HLA class I loss and immune escape. Cancer immunology, immunotherapy : CII *61*, 803-816.

Biankin, A.V., Waddell, N., Kassahn, K.S., Gingras, M.C., Muthuswamy, L.B., Johns, A.L., Miller, D.K., Wilson, P.J., Patch, A.M., Wu, J., *et al*. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature *491*, 399-405.

Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C., *et al*. (2010). Signatures of mutation and selection in the cancer genome. Nature *463*, 893-898.

Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology *25*, 163-177.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research *69*, 6660-6667.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., *et al*. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America *108*, 12372-12377.

Cortes, C., Vapnik, V (1995). Support-vector networks. Machine learning *20*, 273-297.

Del Campo, A.B., Kyte, J.A., Carretero, J., Zinchencko, S., Mendez, R., Gonzalez-Aseguinolaza, G., Ruiz-Cabello, F., Aamdal, S., Gaudernack, G., Garrido, F., *et al*. (2013). Immune escape of cancer cells with beta2-microglobulin loss over the course of metastatic melanoma. International journal of cancer Journal international du cancer.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., *et al*. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*, 1069-1075.

Esteller, M., Levine, R., Baylin, S.B., Ellenson, L.H., and Herman, J.G. (1998). MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. Oncogene *17*, 2413-2417.

Goodfellow, P.J., Buttin, B.M., Herzog, T.J., Rader, J.S., Gibb, R.K., Swisher, E., Look, K., Walls, K.C., Fan, M.Y., and Mutch, D.G. (2003). Prevalence of defective DNA mismatch repair and MSH6 mutation in an unselected series of endometrial cancers. Proceedings of the National Academy of Sciences of the United States of America *100*,

5908-5913.

Huang, D.W/, Sherman, B.T., Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*. *4*, 44-57.

Liptak, T. 1958. On the combination of independent tests. Magyar Tud. Akad. Mat. Kutato Int. Kozl. 3: 171–197.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L., *et al*. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discovery *2*, 172-189.

Mermel, Craig H., et al. "GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers." *Genome Biol* 12.4 (2011): R41.

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.

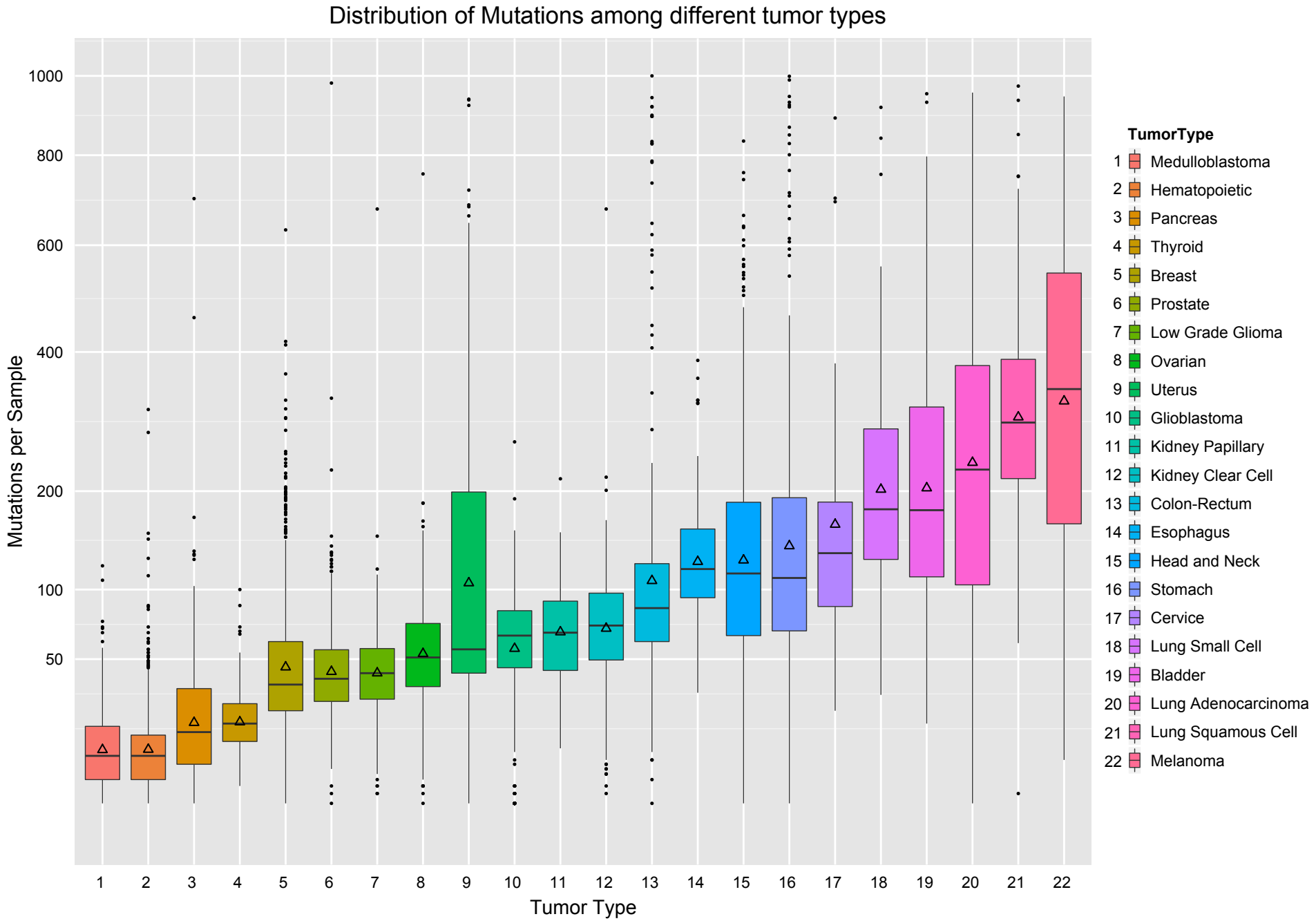
O'Brien, S.J., and Fraser, C.M. (2005). Genomes and evolution: the power of comparative genomics. Current opinion in genetics & development *15*, 569-571.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B *58*, 267-288.

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

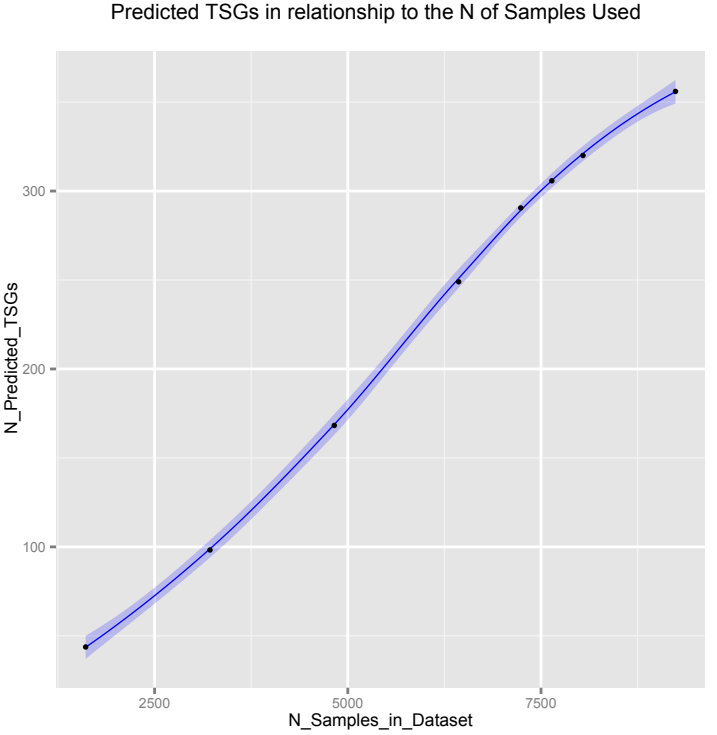
Supplemental Figure 1

Supp.Fig.1

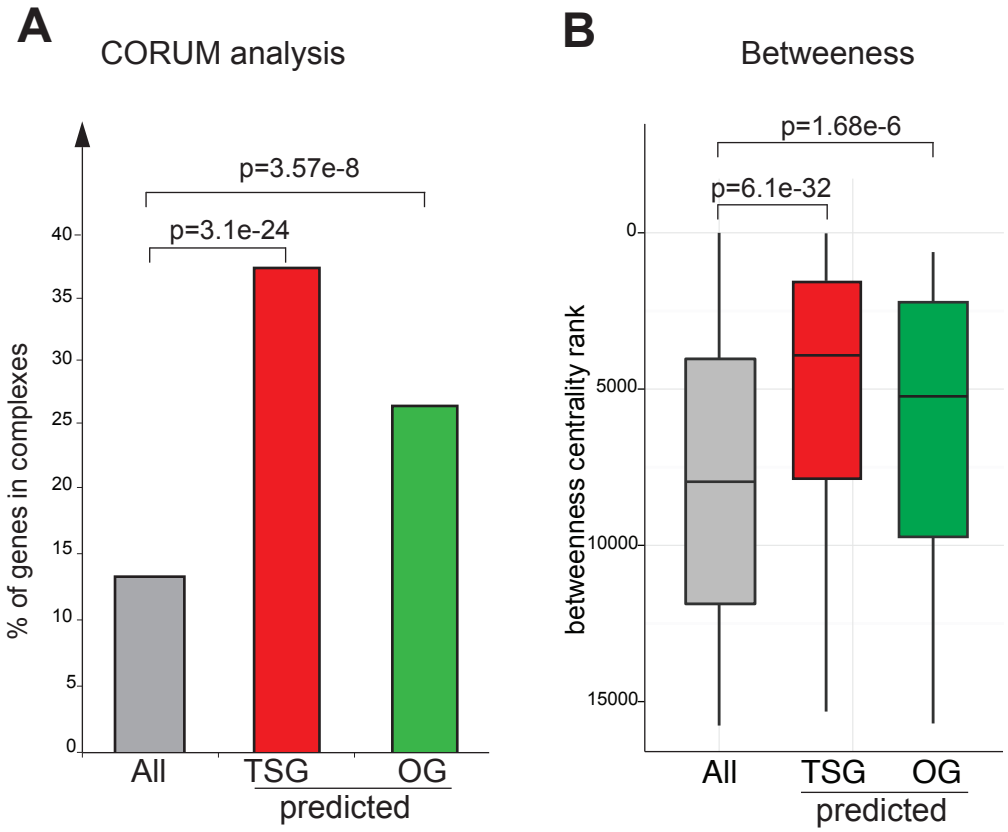


Supplemental Figure 2

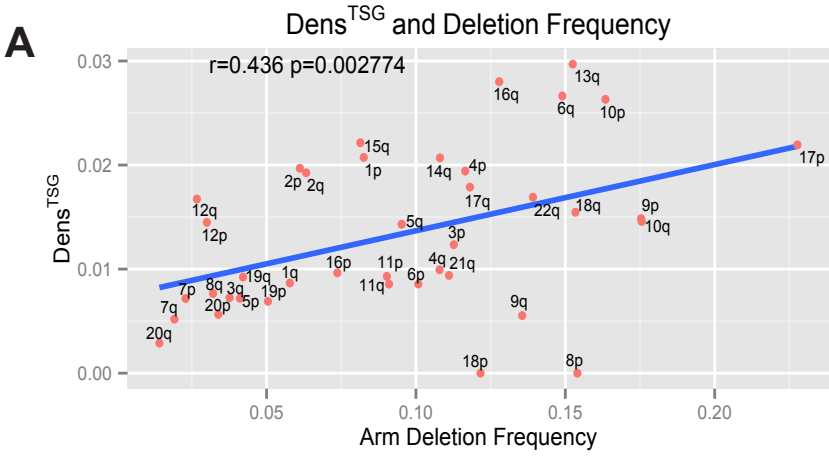
Supp. Fig. 2

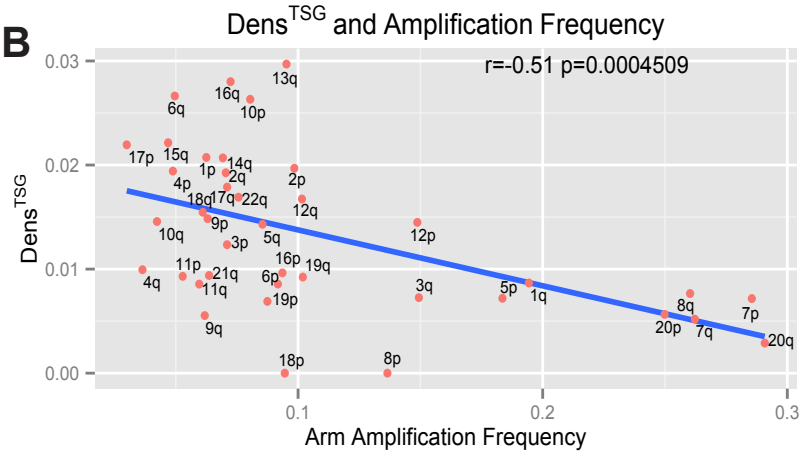


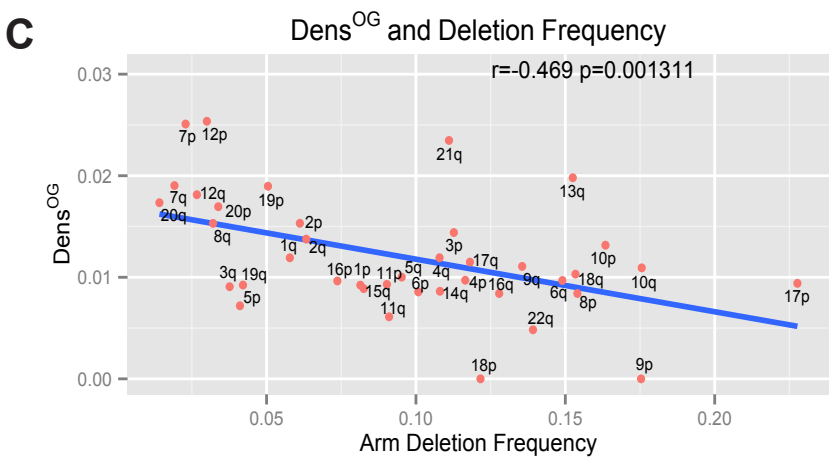
Sup. Fig. 3

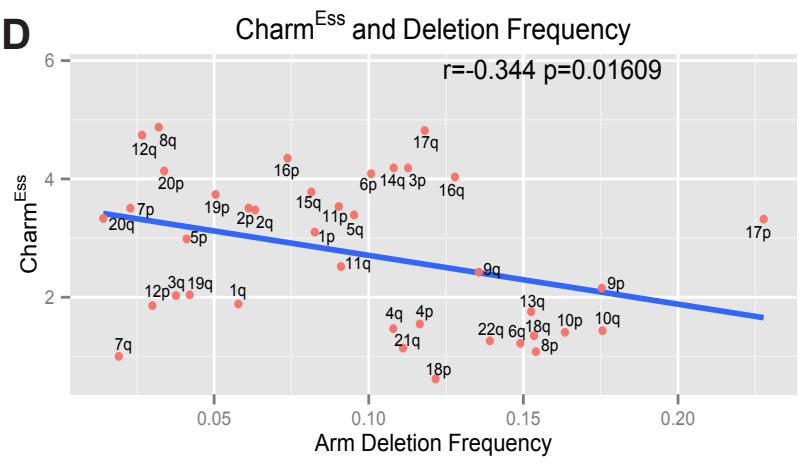

A
CORUM analysis
p=3.57e-8
p=3.1e-24
% of genes in complexes
40
35
30
25
20
15
10
5
0
All
TSG
OG
predicted
B
Betweeness
p=1.68e-6
p=6.1e-32
betweenness centrality rank
0
5000
10000
15000
All
TSG
OG
predicted

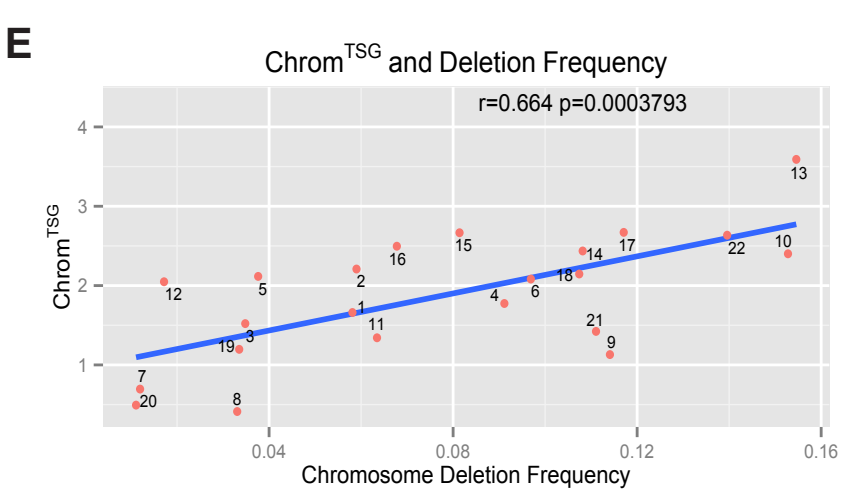
Supplemental Figure 4

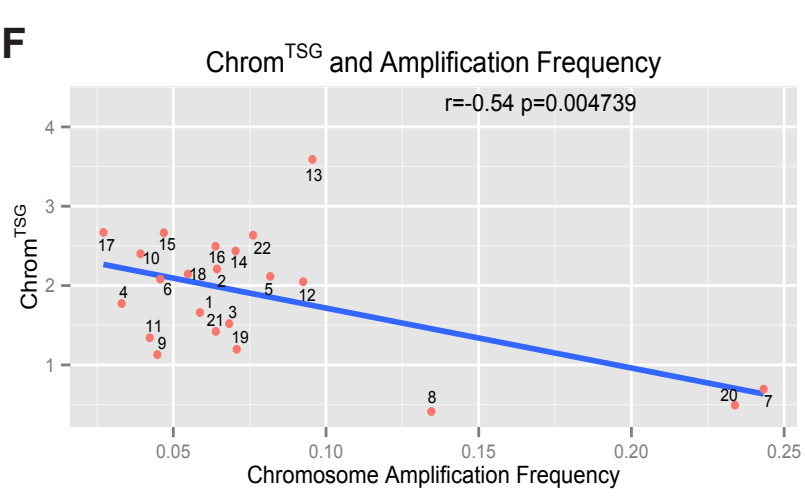


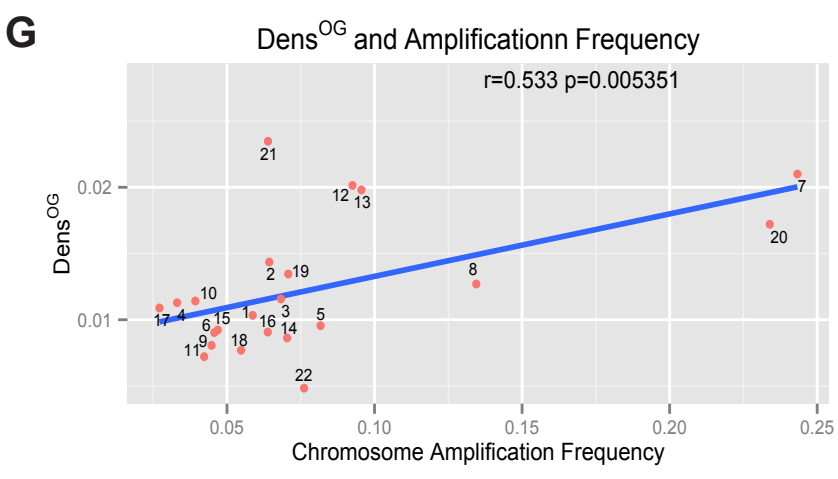


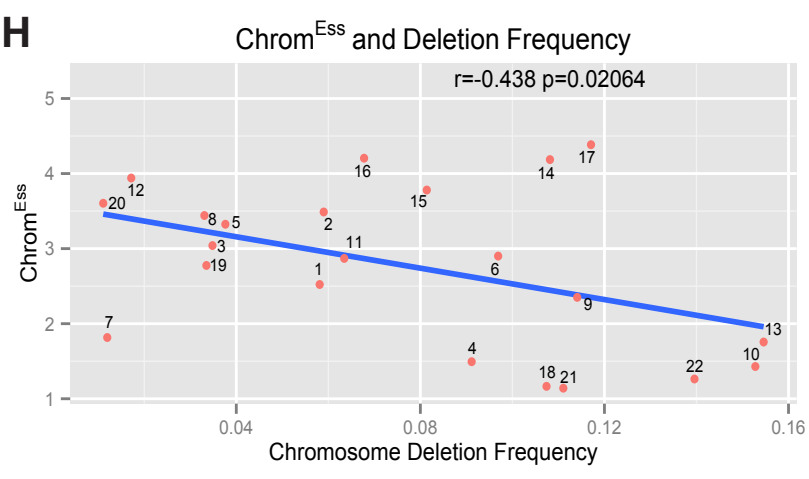












Supplemental Figure 5