STAR* METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Trey Ideker (tideker@ucsd.edu)

METHOD DETAILS

Data filtering

For somatic mutation-based analyses, we filtered out all silent and extra-exonic somatic mutations, with the exception of mutations on splice sites. More specifically, for TCGA data, all mutations with ‘Variant_Classification’ of 'Silent', ‘3'UTR’, 'Intron', ‘5'UTR’, 'RNA', ‘3'Flank’, and ‘5'Flank’ were removed. For MSKCC data, all mutations with a ‘Consequence’ of 'synonymous_variant', '3_prime_UTR_variant', 'intron_variant', and '5_prime_UTR_variant' were removed. As the MSKCC contains multiple samples for some patients, we filtered the data so that each patient is only represented once in the set. First, we prioritized samples for which 410 genes were analyzed on the most recent MSK-IMPACT platform over samples for which only 341 genes were analyzed on the older MSK-IMPACT platform, if a patient had samples analyzed by both platforms. Any other patient duplicates were removed by simply choosing the sample with lowest index. For combined analyses of somatic mutations and CNVs, only high-level CNV events were utilized (GISTIC scores of −2 and +2).

Data inclusion

For somatic mutation-based analyses, we included all TCGA and MSKCC cohorts of tumor types with ≥30 genes within the 341 MSK-IMPACT cancer gene panel with >20 mutations within the respective cohort (**Table S1, S2, S3**). For MSKCC data, we first made cohorts comparable to TCGA, by subsetting based on (detailed) tumor type. For the included MSKCC cohorts, this

mapping was performed as follows. BLCA: 'DetailedTumorType' = 'Bladder Urothelial Carcinoma'; BRCA: 'DetailedTumorType' = 'Breast Invasive Ductal Carcinoma', 'Breast Invasive Lobular Carcinoma', 'Breast Mixed Ductal and Lobular Carcinoma', 'Breast Invasive Carcinoma, NOS', 'Breast Invasive Cancer, NOS', 'Breast Carcinoma', or 'Breast Invasive Mixed Mucinous Carcinoma'; LUAD: 'DetailedTumorType' = 'Lung Adenocarcinoma'; SCKM: 'DetailedTumorType' = 'Cutaneous Melanoma'; COADREAD: 'GeneralTumorType' = 'Colorectal Cancer' (**Table S4**). For combined analyses of somatic mutations and CNVs, only (TCGA) samples for which somatic mutation and GISTIC calls were both available were included in the analysis. For the consensus molecular subtype (CMS) stratified DISCOVER analysis of the TCGA colorectal cancer cohort, 498 tumors with available CMS classification were used (rather than the full set of 559 tumors).

Tumor mutation and alteration load calculations

For tumor mutation load calculations, the binary (filtered) patient-by-gene matrix was summed over all genes sequenced for the respective cohort (whole exome for TCGA cohorts, 341 genes for the MSKCC cohorts). We chose this measure for mutation load (rather than summing the number of individual variants in the MAF files for each patient) for compatibility with mutual exclusivity testing. Namely, mutual exclusivity detection methods take binary patient-by-gene matrices as input, and the tumor mutation load is then given by the number of genes with at least one mutation. For alteration load calculations (combined analyses of somatic mutations and CNVs), a binary patient-by-gene ‘alteration matrix’ was created with 0's and 1’s representing unaltered (no mutation or high-level CNV) and altered (mutation or high-level CNV) cases, respectively. This binary matrix was then summed over all genes.

Mutation/Alteration Load Association (MLA/ALA) calculation

To obtain a standardized measure for the association of a gene's mutation profile to mutation load (Mutation Load Association, MLA), or alteration load (Alteration Load Association, ALA) we used logistic regression, as implemented by the Python package statsmodels (statsmodels.discrete.discrete_model.Logit). For each tumor type, the mutation/alteration profile of each gene (separately) was regressed on the tumor mutation load (+ intercept). The coefficient fitted for the tumor mutation/alteration load corresponds to the log of the odds ratio. To compare these coefficients between genes with different mutation/alteration frequencies, they need to be standardized. This standardized association was calculated by dividing the fitted coefficient by the standard error. Relatively low values imply tendencies of genes to be mutated in tumors with low mutation/alteration load, high values imply enrichment of mutations in tumors with high mutation/alteration load.

QUANTIFICATION AND STATISTICAL ANALYSIS

Cancer gene enrichment calculation

To investigate whether genes with extreme (low or high) MLAs tend to be cancer genes, we calculated the MLA of all human genes in each cohort. We defined 'cancer genes' as the genes within the panel of 341 established onco- and tumor suppressor genes sequenced for all patients in the MSKCC cohorts (**Table S1**). Next, for each cohort, we used Fisher's exact test to assess whether these cancer genes were significantly enriched in the 25 genes with lowest or highest MLA.

Mutual exclusivity analyses

For mutual exclusivity analyses, data were arranged in a binary patient-by-gene format. Pairwise gene-gene mutual exclusivities were tested using DISCOVER (Canisius et al., 2016), MEMo (Ciriello et al., 2012), WExT (Leiserson et al., 2016), Fisher's exact test, and MEGSA

(Hua et al., 2016). MEMo was adapted for pairwise mutual exclusivity testing by comparing the number of tumors mutated in at least one of the two genes to its expectation by chance based on 10,000 marginal-preserving permutations of the mutation matrix. For WExT, we used the highly accurate saddlepoint approximation, since the heavy computational requirements of the recursive formula approach made its use unfeasible. The mutation probability matrix used by WExT was estimated using 10,000 degree-preserving permutations. For the consensus molecular subtype (CMS) stratified DISCOVER analysis of the TCGA colorectal cancer cohort, the estimation of background mutation probabilities was performed separately for each subtype. In this way, subtype-specific differences in gene mutation frequencies are considered when calculating level of mutual exclusivity expected by random chance. Then, these probability matrices were merged and used for mutual exclusivity testing. For mutual exclusivity testing, significance was defined as a P-value < 0.05.

DATA AND SOFTWARE AVAILABILITY

All data were obtained from publicly available sources. Somatic mutation data were obtained from The Cancer Genome Atlas (TCGA) Research Network (http://cancergenome.nih.gov/) and the first 10,000 patients of the Memorial Sloan Kettering MSK-IMPACT Cancer Center (MSKCC). For TCGA data, mutation calls of TCGA's final project, the PanCanAtlas, were downloaded from Synapse (syn7824274, wiki https://www.synapse.org/#!Synapse:syn7214402/wiki/405297) on September 18th, 2017. These mutation calls were generated in a standard fashion across all samples, resulting in a uniform dataset. For MSKCC data, mutation calls were downloaded from the cBioportal for cancer genomics (http://cbioportal.org/msk-impact) on August 8th, 2017. TCGA copy number variation (CNV) data and GISTIC tumor-by-gene calls were accessed from the Broad Firehose Analysis Pipeline in January 2019. Colorectal cancer Consensus Molecular Subtypes (CMS) were obtained from the original paper (Guinney et al. Nature Med., 2015).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Data analysis was done using Python 2.7	Python	
Seaborn	Python library	https://seaborn.pydata.org/
Pandas	Python library	https://pandas.pydata.org/
Statsmodels	Python library	https://pypi.org/project/statsmodels/
Scipy	Python library	https://www.scipy.org/install.html
Matplotlib	Python library	https://pypi.org/project/matplotlib/
Rpy2	Python library	https://pypi.org/project/rpy2/
DISCOVER	(Canisius et al., 2016)	http://ccb.nki.nl/software/discover/#installation
MEMo	(Ciriello et al., 2012)	https://omictools.com/memo-2-tool
WExT	(Leiserson et al., 2016)	http://compbio.cs.brown.edu/projects/wext/
MEGSA	(Hua et al., 2016)	http://dceg.cancer.gov/tools/analysis/megsa/