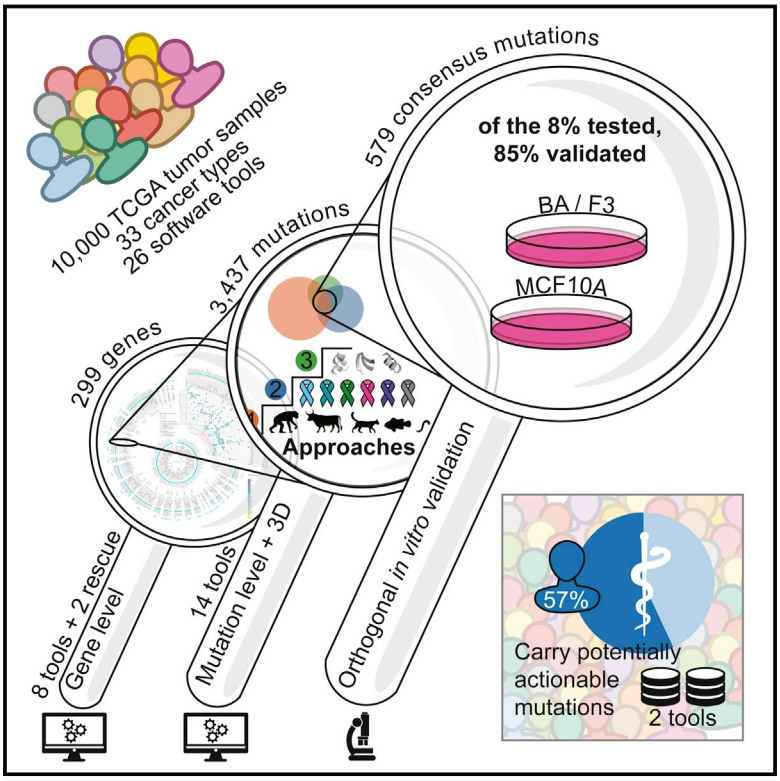
Article

Cell

Comprehensive Characterization of Cancer Driver Genes and Mutations

Graphical Abstract



Highlights

- PanSoftware applied to PanCancer data identified 299 cancer driver genes
- Driver genes and mutations are shared across anatomical origins and cell types
- *In silico* discovery of ~3,400 driver mutations coupled with experimental validation
- 57% of tumors harbor potentially actionable oncogenic events

Authors

Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, ..., Gordon B. Mills, Rachel Karchin, Li Ding

Correspondence

karchin@jhu.edu (R.K.),
lding@wustl.edu (L.D.)

In Brief

A comprehensive analysis of oncogenic driver genes and mutations in >9,000 tumors across 33 cancer types highlights the prevalence of clinically actionable cancer driver events in TCGA tumor samples.

Bailey et al., 2018, Cell *173*, 371–385
April 5, 2018 Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2018.02.060

Cell

Comprehensive Characterization of Cancer Driver Genes and Mutations

Matthew H. Bailey,[1,2,31] Collin Tokheim,[3,4,31] Eduard Porta-Pardo,[5,6,31] Sohini Sengupta,[1,2] Denis Bertrand,[7] Amila Weerasinghe,[1,2] Antonio Colaprico,[8,9,10] Michael C. Wendl,[2,11,12] Jaegil Kim,[13] Brendan Reardon,[13,14] Patrick Kwok-Shing Ng,[15] Kang Jin Jeong,[16] Song Cao,[1,2] Zixing Wang,[17] Jianjiong Gao,[18] Qingsong Gao,[1,2] Fang Wang,[17] Eric Minwei Liu,[19] Loris Mularoni,[20] Carlota Rubio-Perez,[20] Niranjan Nagarajan,[7]

(Author list continued on next page)

[1]Division of Oncology, Department of Medicine, Washington University in St. Louis, St. Louis, MO 63110, USA
[2]McDonnell Genome Institute, Washington University, St. Louis, MO 63108, USA
[3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[4]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA
[5]Barcelona Supercomputing Centre (BSC), Barcelona, Spain
[6]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA
[7]Computational and Systems Biology, Genome Institute of Singapore, Singapore, 138672
[8]Interuniversity Institute of Bioinformatics in Brussels (IB)[2], 1050 Brussels, Belgium
[9]Machine Learning Group (MLG), Département d'Informatique, Université Libre de Bruxelles (ULB), Boulevard du Triomphe, CP212, 1050 Bruxelles, Belgium

(Affiliations continued on next page)

SUMMARY

Identifying molecular cancer drivers is critical for precision oncology. Multiple advanced algorithms to identify drivers now exist, but systematic attempts to combine and optimize them on large datasets are few. We report a PanCancer and PanSoftware analysis spanning 9,423 tumor exomes (comprising all 33 of The Cancer Genome Atlas projects) and using 26 computational tools to catalog driver genes and mutations. We identify 299 driver genes with implications regarding their anatomical sites and cancer/cell types. Sequence- and structure-based analyses identified >3,400 putative missense driver mutations supported by multiple lines of evidence. Experimental validation confirmed 60%–85% of predicted mutations as likely drivers. We found that >300 MSI tumors are associated with high PD-1/PD-L1, and 57% of tumors analyzed harbor putative clinically actionable events. Our study represents the most comprehensive discovery of cancer genes and mutations to date and will serve as a blueprint for future biological and clinical endeavors.

INTRODUCTION

Over the past decade, the Cancer Genome Atlas (TCGA) has coordinated a monumental enterprise of data generation and genomic investigation across 33 cancer types. Numerous notable findings have emerged from this project (https://cancergenome.nih.gov/publications). The individual TCGA projects motivated the development of many bioinformatic algorithms oriented toward discovery, characterization, and prioritization of cellular processes driving cancer based on pathways (Creixell et al., 2015), genes (Ding et al., 2014), or individual variations (Gonzalez-Perez et al., 2013) (Key Resources Table; STAR Methods). Despite this remarkable progress, algorithms do not entirely agree on certain candidate cancer driver genes and mutations, necessitating expert curation to filter likely false positive findings. Previous PanCancer analyses (Tamborero et al., 2013b) have been limited to fewer cancer types and have largely avoided nominating rare driver mutations.

TCGA is now concluding the most sweeping cross-cancer analysis yet undertaken, namely the "PanCancer Atlas project." This project includes the uniform analysis of all TCGA exome data by the Multi-Center Mutation-Calling in Multiple Cancers (MC3) network, yielding unbiased interpretation of the entire 10,437 tumor samples dataset (Ellrott et al. 2018). Here, we describe our analysis of the MC3 somatic mutation set using 26 diverse bioinformatics tools (Figure S1A). Merging results from these tools and manual curation ultimately identified 299 cancer genes. In parallel with functional validation in cell lines, eight other tools and one novel aggregating algorithm characterized mutations having the strongest phenotypic consequences. Four additional tools leveraged protein structural data to elucidate clusters of mutations in three-dimensional space. Finally, the five remaining tools expounded on copy number, RNA abundance, and clinical association using networks, machine learning, and database mining algorithms to further corroborate mutation level findings. The systematic and deep nature of these findings will serve cancer research far into the future.

Isidro Cortés-Ciriano,[21,22,23] Daniel Cui Zhou,[1,2] Wen-Wei Liang,[1,2] Julian M. Hess,[13] Venkata D. Yellapantula,[1,2] David Tamborero,[20] Abel Gonzalez-Perez,[20] Chayaporn Suphavilai,[7] Jia Yu Ko,[7] Ekta Khurana,[19] Peter J. Park,[21,22] Eliezer M. Van Allen,[13,14] Han Liang,[16,17] The MC3 Working Group, The Cancer Genome Atlas Research Network, Michael S. Lawrence,[13,24] Adam Godzik,[6] Nuria Lopez-Bigas,[20,25] Josh Stuart,[26] David Wheeler,[27] Gad Getz,[13] Ken Chen,[17] Alexander J. Lazar,[28] Gordon B. Mills,[16] Rachel Karchin,[3,4,29,32,*] and Li Ding[1,2,12,30,32,*]

[10]Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL 33136, USA
[11]Department of Mathematics, Washington University in St. Louis, St. Louis, MO 63130, USA
[12]Department of Genetics, Washington University in St. Louis, St. Louis, MO 63110, USA
[13]The Broad Institute, Cambridge, MA 02142, USA
[14]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[15]Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[16]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[17]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[18]Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[19]Meyer Cancer Center and Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA
[20]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
[21]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[22]Ludwig Center at Harvard, Boston, MA 02115, USA
[23]Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
[24]Department of Pathology, Massachusetts General Hospital Cancer Center, 55 Fruit Street, Boston, MA 02114, USA
[25]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[26]University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA
[27]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA
[28]Departments of Pathology, Genomic Medicine, & Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[29]Department of Oncology, Johns Hopkins University, Baltimore, MD 21287, USA
[30]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63110, USA
[31]These authors contributed equally
[32]Lead Contact
*Correspondence: karchin@jhu.edu (R.K.), lding@wustl.edu (L.D.)
https://doi.org/10.1016/j.cell.2018.02.060

RESULTS

Mutational Dataset and Driver Gene Identification Power

Mutation calls were produced by the Multi-Center Mutation Calling in Multiple Cancers (MC3) working group that harmonized the results of seven algorithms (Ellrott et al. 2018) (STAR Methods). To reduce the false-positive rate for driver gene discovery, we implemented three strategies to optimize driver detection and data quality (Figure S1B; STAR Methods). Briefly, we excluded 344 hypermutator samples because of artifactual sensitivity to high background mutation rates (Figure 1A). All mutations that passed the MC3 filter criteria were included. In addition, a less stringent filter was applied to samples from ovarian serous cystadenocarcinoma (OV) and acute myeloid leukemia (LAML) projects, as exome data for these two cancer types have distinct characteristics not amenable to our standard filtering. Finally, samples marked with inconsistent pathology were excluded. Our driver detection dataset ultimately consisted of 9,079 samples having 1,457,702 total mutations (Figure S1B), where the number of mutations per sample was widely distributed across cancer types, as previously noted (Figures 1B and 1C) (Kandoth et al., 2013; Lawrence et al., 2013; Tamborero et al., 2013b).

For individual cancer types, analyses were sufficiently powered to detect genes mutated at a median of 6.1% above background mutation rates (Figure 1D). Power largely correlated with cohort size, with lower values observed for lymphoid neoplasm diffuse large B cell lymphoma (DLBC) (25.5%, n = 37), cholangiocarcinoma (CHOL) (20.5%, n = 34), and uterine carcinosarcoma (UCS) (14.9%, n = 55) and the highest statistical power for breast invasive carcinoma (BRCA) (2.3%, n = 779), brain lower grade glioma (LGG) (2.8%, n = 510), and thyroid carcinoma (THCA) (2.3%, n = 491). We saw modest increase in statistical power for 12 individual cancer types previously analyzed by the TCGA PanCancer effort (Kandoth et al., 2013), but the addition of 21 individual cancer types to our current PanCancer analysis increased power to <1% prevalence (Figure S1C).

Landscape of Cancer Driver Genes

The final consensus list consists of 299 unique genes: 258 genes obtained from a systematic approach and 41 additional genes recovered after manual curation of previous TCGA marker papers with the majority (26 out of 41, 63%) supported by additional -omics network tools not used in original significantly mutated gene (SMG) detection studies (Figures 1A and S2; Table S1; STAR Methods). Here, we focus on the 258 genes set, but acknowledge the limitations of a systematic approach by including the 41 manually rescued genes in our final list.

The list recovers most of the previously described driver genes for the majority of cancer types. In fact, in 20 out the 31 cancer

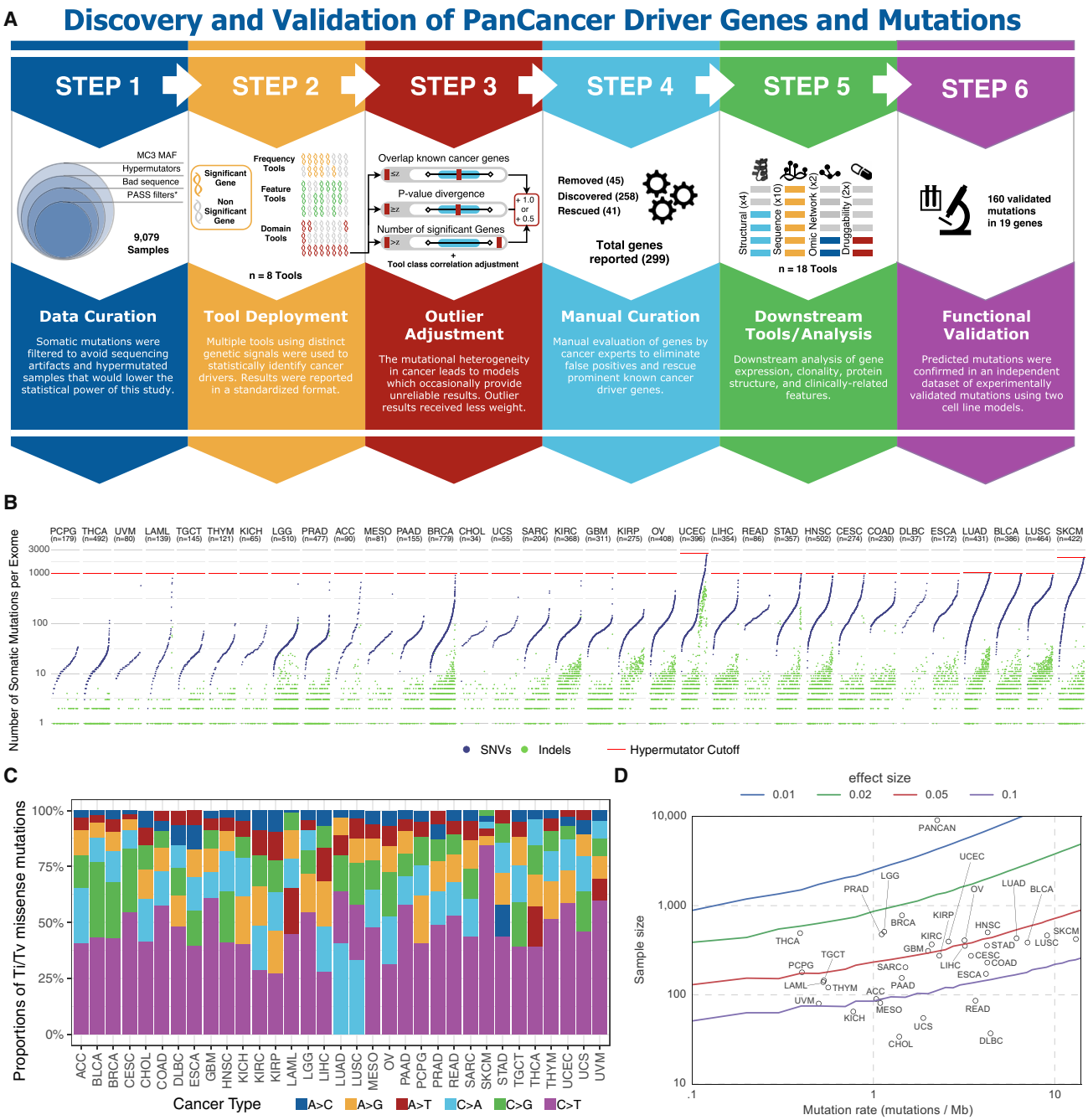


Figure 1. Cancer Driver Gene Discovery Strategy, Power, and Mutations

(A) We identified six main steps to identify and discover driver genes in cancer: data curation, tool development, outlier adjustment, manual curation, downstream tool analysis, and functional validation.

(B) Somatic mutations per sample are plotted for each sample and cancer type. Mutations are separated into SNVs (blue) and indels (green). The selected hypermutator cutoff for each cancer is shown in red.

(C) Transition and transversion proportions are shown for six nucleotide changes. The stacked proportion bar chart is sorted by increasing transition/transversion fraction.

(D) Statistical power for detection of cancer driver genes at defined fractions of tumor samples above the background mutation rate (effect size with 90% power) is depicted. Circles indicate each of 33 cancer types placed according to the study sample size and median background mutation rate.

See also Figures S1 and S2 and Table S6.

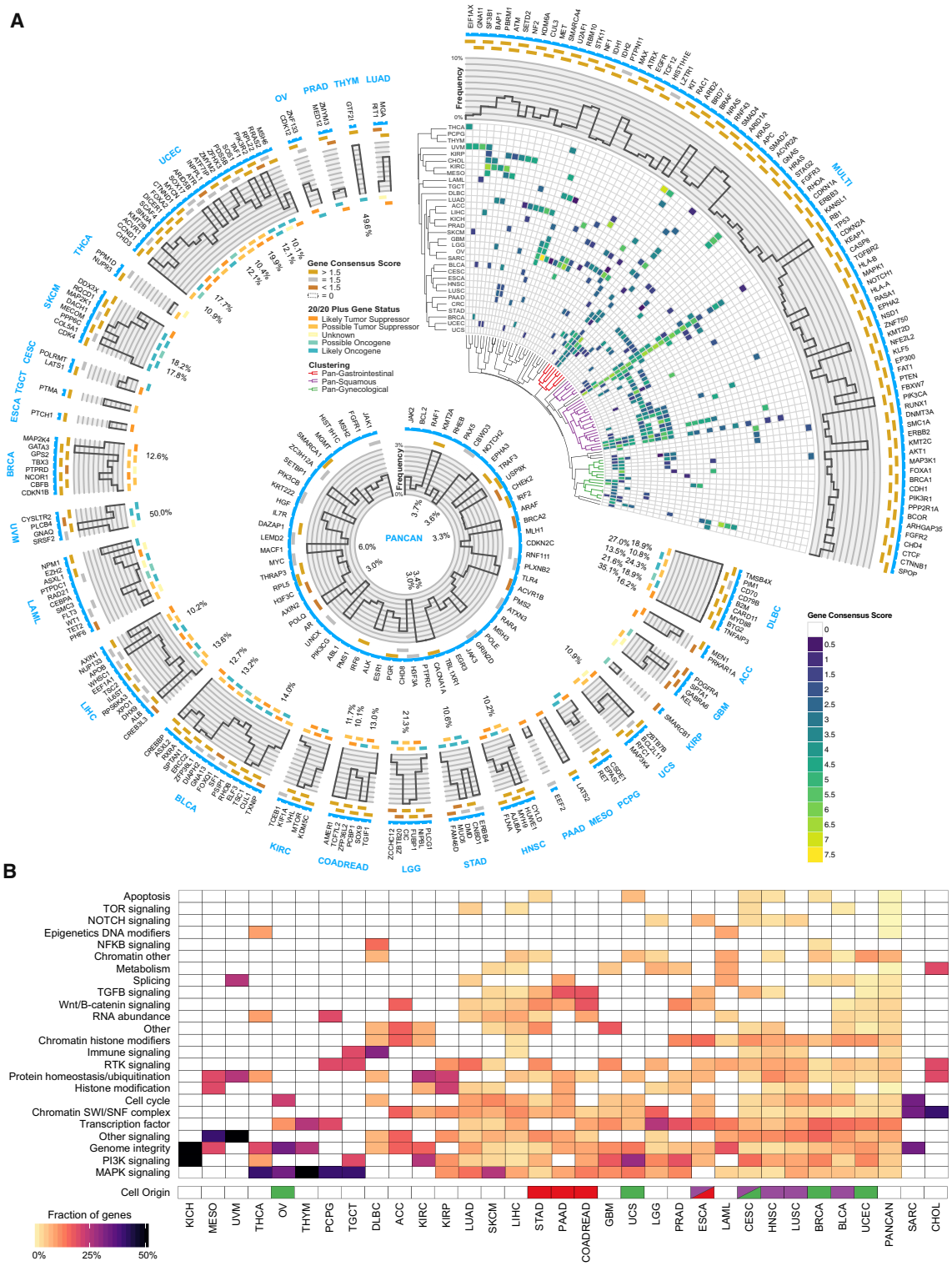


Figure 2. Cancer Driver Gene Discovery Workflow

(A) Circos (Krzywinski et al., 2009) plot displays 299 cancer genes. Each sector indicates a unique cancer type (text in blue) with predicted drivers unique to that cancer type listed (gene name in black). Only tissues having at least one unique driver gene are shown. The top right sector shows all genes found significant in multiple cancer types. Next, a categorical score of gold, silver, or bronze is assigned to each gene based on the highest consensus score. If a gene was not scored and required rescue, then the field is empty. The next ring illustrates the mutation frequency of a gene. For the top-right wedge the PanCancer frequency is used,

(legend continued on next page)

types included in our study that had either been previously published or for which we had an internal list of known cancer driver genes, the recovery rate is 80% or higher (Figures S2D and S2E). The most significant outliers are stomach adenocarcinoma (STAD) and the previous PanCancer study, for which we only recovered around 70% of the previously described genes (Figure S2D). The consensus list also includes 59 novel genes that had not been described previously and other known drivers not previously associated with a given tissue (Table S1; STAR Methods). Predictions of known cancer driver genes in new cancer types include *ATRX* in adrenocortical carcinoma (ACC), *KMT2C*, *CTNNB1*, and *PTEN* in bladder urothelial carcinoma (BLCA), and *ARID1A* and *KRAS* in BRCA. Entirely novel predictions include *GNA13* in BLCA (a homolog of the known drivers *GNAQ* and *GNA11*), *RRAS2* in uterine corpus endometrial carcinoma (UCEC) (with shared homology in *KRAS* and *HRAS*), and *KIF1A* in head and neck squamous cell carcinoma (HNSC) (a kinesin of the same family of the cancer driver *KIF5B*).

The number of detected cancer driver genes varies among cancer types, with kidney chromophobe (KICH) having the fewest (2 genes) and UCEC having the most (55 genes). Furthermore, the ratio of predicted tumor suppressor genes to oncogenes widely varies by tissue (Figure S4B). We observed a significant positive correlation (Pearson's R = 0.66, p value = 4.1×10^{-5}) between average mutation burden in a cancer type and the number of identified consensus genes (Figure S3B). Study-based calculations for powered effect size in each cancer type did not entirely explain this phenomenon (Pearson's R = −0.31, p value = 0.09) (Figure S3C). Regarding the associations of driver genes with different cancer types, many genes (142 out of 258) are associated with a single cancer, whereas 87 genes have driver roles in two or more cancer types, with an additional 29 genes uniquely identified using PanCancer approaches on all samples combined. As expected, *TP53* is the most extreme case (27 cancer types), followed by *PIK3CA*, *KRAS*, *PTEN*, and *ARID1A*, each of which is associated with 15 or more cancer types (Figures 2A and S4A).

We clustered cancer types according to the consensus scores of their associated genes. Remarkably, some cancer types are grouped by tissue of origin, such as LGG and glioblastoma multiforme (GBM), and others by cell of origin. The most significant of the cell origin clusters spans all squamous cancer types (BLCA, cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], esophageal carcinoma [ESCA], HNSC, and lung squamous cell carcinoma [LUSC]) (permutation test, adjusted $p < 0.01$) and includes several transcription factors (*ZNF750*, *NFE2L2*, or *KLF5*), chromatin and histone modifiers (*KMT2D*, *EP300*, or *NSD1*), and various PI3K pathway genes (*PIK3CA*, *PTEN*, or *MAPK1*). We found two additional significant clusters (permutation test, adjusted $p < 0.05$) that group gynecological (UCS, CESC, UCEC, OV, and BRCA), as well as gastrointestinal cancers (COADREAD, pancreatic adenocarcinoma [PAAD], ESCA, and STAD) (Figures 2A and S4A; STAR Methods).

Finally, we classified the consensus driver genes by cancer-related biological processes and associated pathways (Figure 2B; Table S2). For most genes, the categories (excluding "other" and "other signaling") clearly reflect known processes involved in carcinogenesis, namely "transcription factor" (39 genes), "RTK signaling" (16) and "RNA abundance" (15), "protein homeostasis/ubiquitination" (15), "chromatin histone modifiers" (15), "genome integrity" (14), "chromatin other" (14), and "immune signaling" (10). The last group is of particular interest, given the connection between driver genes and immune response (Thorsson et al., 2018). In terms of cancer types, most have at least one cancer driver that belongs to either genome integrity (28 out of 33 cancer types) or the MAPK or PI3K signaling pathways (24 and 22 cancer types, respectively). Notably, squamous cancer types have higher proportions of chromatin histone modification genes, as well as receptor-tyrosine kinase and immune signaling.

Approaches to Driver Mutation Discovery

Not all mutations in a cancer driver gene have equal impact (Torkamani and Schork, 2008), with consequences frequently depending on position within the protein and amino acid change (Carter et al., 2009). We explored this issue across the entire PanCancer dataset, classifying 751,876 unique missense mutations by examining the 299 identified cancer driver genes, according to their predicted oncogenic effect. We combined the output of three different categories of tools into consensuses approaches (STAR Methods): (1) tools distinguishing benign versus pathogenic mutations using sequence (CTAT population); (2) tools distinguishing driver versus passenger mutations using sequence (CTAT cancer); and (3) tools discovering statistically significant three-dimensional clusters of missense mutations (structure based). These tool groups identified 10,098 (1.3% of total missense mutations), 4,595 (0.6%), and 1,469 (0.2%) unique amino acid substitutions, respectively (Figure 3A). Differences in the number of predicted driver mutations for each approach are likely due to tool design and requirements, i.e., dependence of structural clustering tools on available three-dimensional protein structures (either experimental or homology-based) yields fewer predicted driver mutations.

When benchmarked against OncoKB (Chakravarty et al., 2017), a manually curated dataset of cancer mutations annotated according to likely oncogenic effect, cancer-focused

whereas cancer-type-specific frequencies are used in the remaining sectors. Where frequencies exceed the y axis limit of 10%, the innermost label indicates the frequency (not shown are *PIK3CA* = 11.8% and *TP53* = 37.5%). The final ring uses a five-point scale from orange to teal to represent each gene from likely tumor suppressor to likely oncogene, respectively, according to the 20/20+ algorithm. Finally, in the top-right slice, we show hierarchical clustering of the gene consensus scores for genes that were found in more than one cancer type (note: CRC refers to the COADREAD cancer type). Additionally, significant gene clusters (permutation test) identified pan-gastrointestinal (red), pan-squamous (purple), and pan-gynecological tissues (green). The middle ring illustrates all genes that were found only using PanCancer results or were otherwise rescued.

(B) Heatmap showing clustering of different cancer types by pathway/biological process affected by associated consensus driver genes. Cell of origin for pan-gynecological, pan-gastrointestinal, and pan-squamous are colored as indicated above.

See also Figures S2, S3, and S4 and Tables S1, S2, S3, S4, and S7.

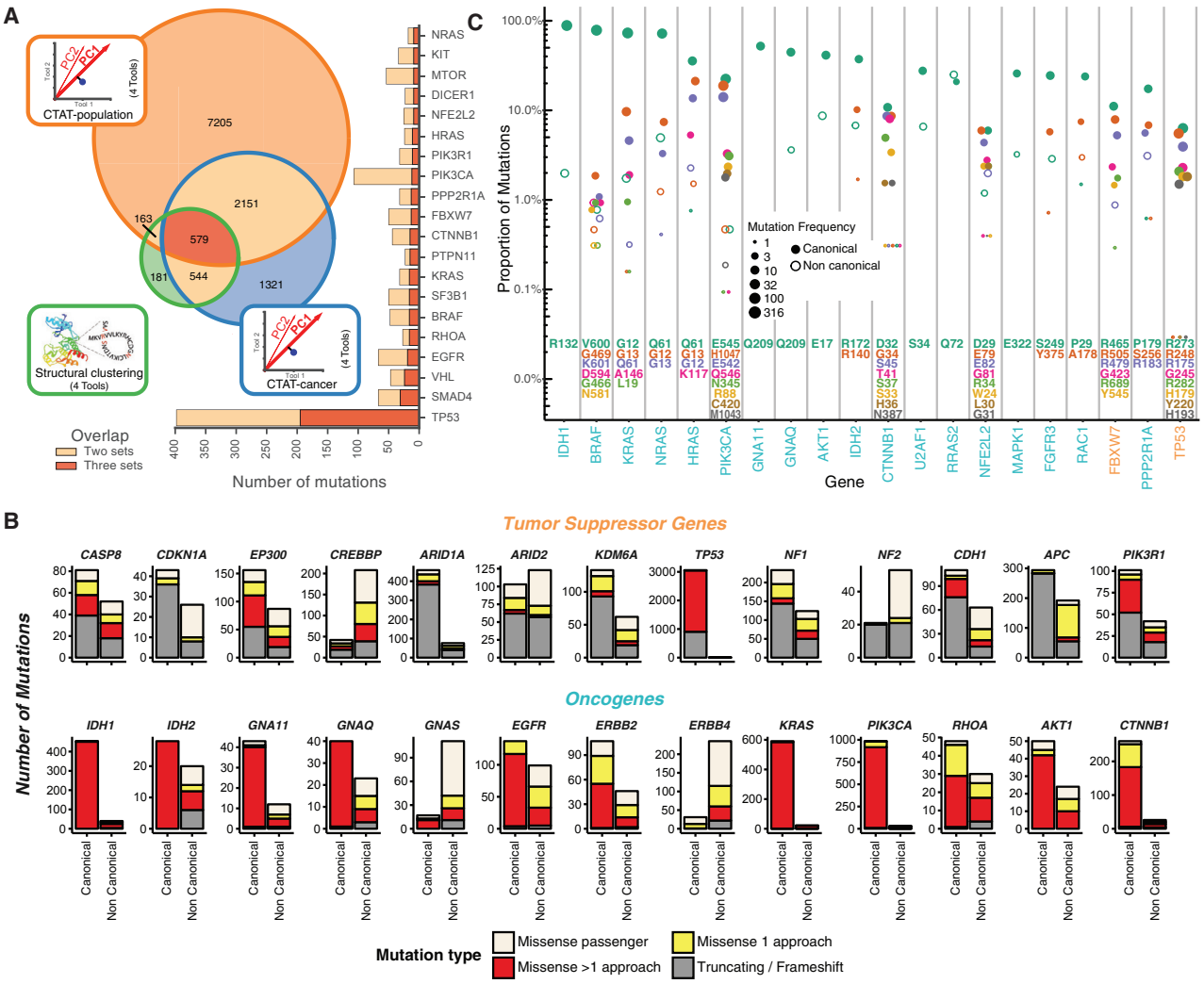


Figure 3. Driver Mutation Discovery Approaches, Overview, Overlap, and Contrasts

(A) Venn diagram indicates the total number of mutations overlapping among three consensus approaches: CTAT population, CTAT cancer, and structural clustering. Adjacent bar chart indicates the top 20 genes sorted by three-set intersecting mutation counts.

(B) Driver gene discovery identified gene-tissue pairs (canonical genes) in tumor suppressors and oncogenes. However, some gene-tissue pairs were not identified in driver discovery (non-canonical). Mutation frequency from canonical and non-canonical cancer genes are displayed and divided among four mutation classes: truncation/frameshift mutations (gray); missense mutations uniquely identified by only one approach (yellow, see A); missense mutations identified by multiple approaches (red, see A); and missense passenger mutations not identified by any approach (off-white).

(C) Mutation percentage out of all missense and truncating/frameshift mutations within a gene is shown on the y axis (log scale). Point size is log scaled and represents amino acid position frequency. The top 23 genes ordered by increasing mutational diversity (normalized entropy) and only the 9 most frequently mutated amino acid positions for each gene are shown.

See also Figure S5 and Table S4.

algorithms had superior predictive value than algorithms distinguishing benign and pathogenic mutations (Figure S5). The CTAT cancer score outperformed all individual sequence-based approaches.

Overall, 9,919 predicted cancer driver mutations in our cohort (3,437 unique mutations) were identified by ≥ 2 approaches from CTAT population, CTAT cancer, or structural clustering. These mutations affect 5,782 tumor samples. These missense driver mutations represent a greater fraction of the total mutations in oncogenes than in tumor suppressors (Figure 3B). In this latter group, most mutations seem to be truncations or frameshifts, consistent with previous observations (Vogelstein and Kinzler, 2004). Nevertheless, some tumor suppressor genes also have high numbers of missense driver mutations, such as *EP300*, *CREBBP*, *CASP8*, *PIK3R1*, and *TP53* (Figure 3B). An interesting example is *CDH1*, which is primarily affected by truncating or frameshift mutations in BRCA (75 out of 85 mutations), but mostly targeted by missense driver mutations in STAD (21 out of 25 mutations). This suggests differing roles for *CDH1* in these two cancer types.

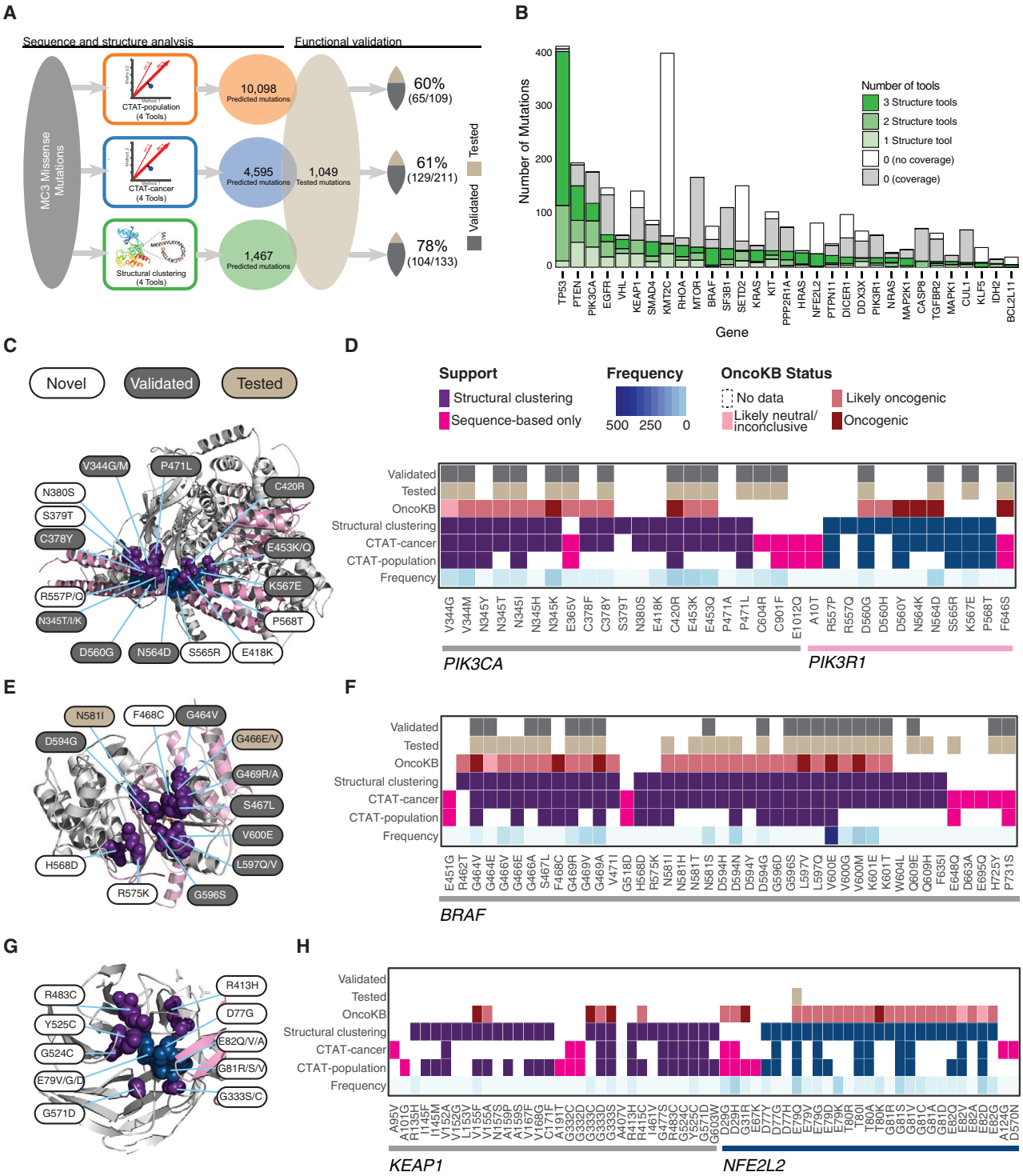


Figure 4. Driver Mutation Discovery and Validation

(A) Steps taken to assess consensus among mutation-level predictions using sequence-based and structural clustering tools and comparing them to an orthogonal set of functionally validated mutations. From left to right: gray box represents missense mutations that were processed by 12 tools from three categories (population-based, cancer-focused, and structural clustering tools) and combined into three consensus approaches (CTAT population, CTAT cancer, and structural clustering). The total number and percentage of functionally validated/tested mutations are also shown.

(B) Number of mutations (y axis) found by structural tools for each gene (x axis) are shaded according to support by structural tools (green). Those mutations without support are distinguished by two categories, with (gray) and without (white) available protein structure.

(legend continued on next page)

We were intrigued by missense driver mutations detected in cancer types where the gene was not predicted to be a driver. This subset is particularly important for genotype-driven clinical trials (Gagan and Van Allen, 2015). Overall, there are 1,719 tissue-unmatched likely driver mutations (19% of the total) in 1,431 patients (16%) and 502 patients whose only predicted missense driver mutations affect genes not yet known to play a role in that cancer type. For example, we identified 28 patients with predicted *EGFR* driver mutations in cancer types where *EGFR* is not yet identified as a common driver gene, such as HNSC, STAD, LUSC, UCEC, ESCA, and liver hepatocellular carcinoma (LIHC). In extreme cases, such as *ERBB4* or *GNAS*, these mutations actually represent the majority of predicted driver missense mutations in the gene (Figure 3B). Additionally, we found that 2% (10/457) of *IDH1* missense events that occur at position R132 are found in cancers not typically known to carry such mutations, i.e., BLCA (n = 2), BRCA (2), COADREAD (2), lung adenocarcinoma (LUAD) (2), pheochromocytoma and paraganglioma (PCPG) (1), and thymoma (THYM) (1) (Figure 3C). Furthermore, we observed that $RRAS2^{Q72}$, a predicted oncogene in UCEC (n = 5 samples) with strong homology to $KRAS^{Q61}$ and $HRAS^{Q61}$, was exceptionally mutated in cancer types where it was not previously recognized: UCS (n = 1), LUSC (1), LUAD (1), prostate adenocarcinoma (PRAD) (1), HNSC (1), and TCGT (1). Any analysis focusing only on driver genes and mutations known in that cancer type would very likely miss presumed driver mutations for those patients.

Functionally Validated Mutations Confirm Structure-Based Analysis

We used an independent dataset of 1,049 experimentally tested somatic mutations to validate our driver mutation prediction (Ng et al., 2018). Briefly, mutations were introduced in two cancer cell lines, Ba/F3 and MCF10A, and were evaluated for oncogenicity based on survival and growth (STAR Methods). In total, 160 mutations from 19 genes were validated in this dataset. The percentage of functionally validated mutations increased from 60% predicted with CTAT population, to 61% for those found by CTAT cancer, and 78% for structure-based analysis (Figure 4A). Among the 579 mutations predicted by all three approaches (Table S4), 39 of the 46 tested (85%) were validated. Further, the sensitivity and specificity of identifying driver mutations annotated by OncoKB suggests performance is generalizable to larger gene sets (Figure S5E). These results support the value of the prediction algorithms used in our study and the advantage of combining multiple tools. Also, we would like to note that this approach only addresses true positive findings and represents a floor estimate for computational predictions.

Structural-based mutations clustered on 66 proteins, including one cluster on KLF5, a gene not previously identified in PanCancer studies and ranked among the top 30 clusters by PanCancer mutation frequency (Figure 4B). We sought to further examine predictions of the three approaches in various well-established cancer driver genes, such as *PIK3CA*/*PIK3R1*, *BRAF*, and *KEAP1*/*NFE2L2* (Figures 4C–4H). The interface between PIK3CA and PIK3R1 contains a cluster of mutations found by at least two of the approaches and includes both validated mutations and some not tested. D560G, N564D, and K567E are validated mutations that closely cluster to non-tested mutations R577P/Q, S565R, and P568T in *PIK3R1*. Similarly, *PIK3CA* contains validated mutations C378Y, V344G/M, N345T/I/K, P471L, C420R, and E418K clustering with non-tested mutations S379T, N380S, and E418K. These non-tested mutations are excellent candidates for further experimental validation due to both their close proximity to known validated driver mutations and their support from sequence-based approaches (Figures 4C and 4D). *BRAF* also contains clusters similar to this *PIK3CA*/*PIK3R1* cluster, with a mixture of validated and novel mutations (Figures 4E and 4F).

Additionally, there are many genes that contain mutations found by all three approaches, but that were not tested experimentally, including *KEAP1*, *NFE2L2*, *RHOA*, *MTOR*, *MAP2K1*, and *VHL*. Nevertheless, many of these driver mutations have orthogonal evidence from OncoKB. For example, G333D/S mutations in *KEAP1* have an OncoKB status of likely oncogenic and oncogenic, respectively (Figures 4G and 4H). Also, *NFE2L2* mutations cluster closely with *KEAP1* mutations along the protein-protein interface (D77, E82, G81, and E79). While they were not experimentally validated, all have an OncoKB status of either likely oncogenic or oncogenic. Other *KEAP1* mutations in the same cluster found by all three approaches are R483C, Y525C, G524C, G571D, and R413H. However, none of these mutations were tested in our dataset, nor have evidence from OncoKB. Given their proximity to the validated *KEAP1* sites and the bioinformatic evidence that we found, these mutations are ideal candidates for follow-up validation experiments.

Overall, this analysis demonstrates the complementarity of sequence-based and structure-based approaches. For example, E365V, C604R, and C901F in *PIK3CA*, F646S in *PIK3R1*, and H725Y and P731S in *BRAF* were found only by the former and were experimentally validated (Figures 4D and 4F). Conversely, R462T in *BRAF* was only found by the latter and is annotated as likely oncogenic in OncoKB (Figures 4F and 4H).

Hypermutated Phenotypes and Immune Infiltrates

Environmental and biological factors, such as tobacco exposure, UV, and microsatellite instability (MSI), contribute to the tumorigenic hypermutator phenotype (Roberts and Gordenin,

(C–H) Heatmaps (D, F, and H) coupled with protein structures (C, E, and G) are shown in panels for proteins PIK3CA/PIK3R1 (PDB: 4OVU), BRAF (4MBJ), and KEAP1/NFE2L2 (3ZGC), respectively, and display whether a particular mutation was detected by sequence-based (CTAT population or CTAT cancer) or structure-based approaches (at least two structural tools). Purple/teal colors distinguish proteins (PIK3CA/PIK3R1 and KEAP1/NFE2L2 pairs) for mutations found by structure-based approaches, and pink boxes indicate mutations found only by sequence-based approach. Additionally, for each mutation, frequency (blue gradient), OncoKB status (red gradient), testing status (tan), and validation status (gray) are provided. All mutations found by structure-based approaches in each of the three genes are shown with a few additional mutations that are only found by sequence-based approaches. Key mutations are highlighted from heatmaps and labeled with white, gray, and tan labels referring to novel, validated, and tested (not validated) mutations, respectively.
See also Table S4.

2014). Because many hypermutated samples were excluded in the driver-discovery dataset, we performed additional analyses to explore genes associated with this phenotype. Using mutation signature analysis, we found that 90% (309/344) of the samples that we labeled as hypermutated have MSI, UV, POLE, APOBEC, or smoking as their primary signature (Figure 5A). MSI and POLE are particularly prevalent, accounting for 56% of the hypermutated samples. As expected, many cancer genes involved in MSI and mismatch repair (MMR), i.e., *POLE*, *MLH1*, *MSH3*, and *MSH2* (Alexandrov et al., 2013; Kim et al., 2013), are frequently mutated in these samples (Table S5; STAR Methods).

We expanded our analysis on mutation signatures by estimating MSI status using MSIsensor (Niu et al., 2014) across all samples (n = 9,423). 338 tumors have a score 4 (indicative of an MSI-high phenotype). MSIsensor scores were correlated with validated gel assays in a subset of hypermuated samples (n = 180, multiple regression model, p value $< 2 \times 10^{-16}$, $r^2 = 0.504$; STAR Methods). We identified canonical MSI cancer types (UCEC, colon adenocarcinoma [COAD], and STAD) as having the highest average MSI scores across all samples (Figure 5B). We also observed 73 tumors with high MSI scores from non-canonical cancers, i.e., 2% of OV (n = 7), and 2% of CESC (n = 5). We observed that OV tumors have a higher mean MSIsensor score when compared to other tissues, which is consistent with previous findings (Cortes-Ciriano et al., 2017). 4 of 5 CESC MSI samples harbored mutations in genes known to be involved in MSI, including 1 sample with 2,644 somatic mutations that carried frameshift deletions in both *MLH3* and *MSH3*.

MSI cases show improved response to immune checkpoint therapy, independent of histology (Brahmer et al., 2012; Gryfe et al., 2000; Le et al., 2015). Thus, we tested whether the samples with high MSIsensor scores exhibited similar patterns of immune infiltration between environmental and biological mechanisms. Using RNA-seq abundance data, we calculated PD-L1, PD-L2, PD-1, CD8A, and CD8B expression in MSI-high and microsatellite stable (MSS) samples to identify via association those samples that would likely benefit from immunotherapy (Figure 5C; STAR Methods). We observed a significant difference between immune infiltrates when comparing samples with high MSIsensor scores ($\geq$4) to others with low MSIsensor scores (<4) from COADREAD, STAD, and UCEC (Figures 5C), in agreement with previous findings about these cancer types. We then tested whether the other three most prevalent signatures in hypermutators, i.e., smoking, UV, and APOBEC, have similar patterns of immune infiltrate expression. However, only suggestive evidence (t test, p value < 0.05) was found for PD-1 overexpression in hypermutated bladder cancer (BLCA) samples with the APOBEC signature (Figure 5D). Together, these findings corroborate the known relationship between total mutational burden and expression of immune modulators, but suggest that MSI may be particularly immunogenic. Additionally, an examination of BRCA samples revealed that 11 of 12 hypermutated samples harbor at least one mutation in MSI associated genes (1 with hypermethylated *MLH1*) and had increased expression in PD-L1, PD-L2, and CD8A when compared to non-hypermutated cases (t test p values < 0.01, < 0.01, and < 0.05, respectively; Figure S6A). Similar findings in CESC and LUSC illustrate potential driver mechanisms in a subset of cases often overlooked in driver gene discovery analysis (Figures S6B and S6C).

Therapeutic Implications of Molecular Events

We used two different databases to assess therapeutic implications of molecular events in our dataset: Precision Heuristics for Interpreting the Alteration Landscape (PHIAL) (Van Allen et al., 2014) and Database of Evidence for Precision Oncology (DEPO; http://depo-dinglab.ddns.net). Both databases cast therapeutic projections based on FDA-approved therapies, clinical trials, published clinical evidence, and, in the case of PHIAL, the TARGET database. PHIAL works at the gene level, whereas DEPO focuses on specific mutations (STAR Methods). We emphasize that, while the implications and results of this section have been curated based on the literature, many of these results are still undergoing rigorous scientific/clinical testing. However, eligibility for clinical trials based on demonstration a particular driver mutation still falls within the rubric of a clinically actionable mutation.

We observed that both the fraction of samples and proportion of alteration types varied across tissue types. By PHIAL heuristics, 52% of all samples contained at least one putatively actionable alteration (Figure 6A), while 65% of samples had at least one putatively actionable or biologically relevant alteration from TARGET. Using DEPO, we found that 30% of samples in our dataset had at least one clinically actionable mutation (Figure 6B).

Using PHIAL, the most common putatively actionable alterations across the entire dataset were *CDKN2A* deletions (13%), *PIK3CA* mutations (12%), *MYC* amplifications (8%), *BRAF* mutations and amplifications (8%), and *KRAS* mutations (7%). *CDKN2A* loss may predict sensitivity to CDK4/6 inhibitors and affects over 40% of GBM, mesothelioma (MESO), and ESCA patients. *PIK3CA* mutations, which may predict sensitivity to PIK3CA inhibitors, affected 45% of patients with UCEC; *MYC* amplifications, prognostic in glioma and pancreatic cancer, were also present in 33% of OV samples. *BRAF* mutant samples made up over half of THCA and skin cutaneous melanoma (SKCM) patients, suggesting sensitivity to RAF inhibitors. Finally, we also found high fractions of patients with pancreatic, colon, rectum, and lung adenocarcinomas with *KRAS* mutations (between 70% and 30% in all cases). While these mutations are currently of limited utility in untreated pancreatic and lung adenocarcinomas, they predict resistance to anti-EGFR therapies in colorectal adenocarcinoma.

Similar to PHIAL, *PIK3CA*, *BRAF*, and *KRAS* contributed to the most number of samples with potentially actionable alterations from DEPO. SKCM, uveal melanoma (UVM), LGG, PAAD, COAD, and THCA have higher prevalence of clinically actionable alterations. When looking at the most common clinically actionable alterations by cancer type (Figure S7D), some of the same genes as PHIAL are key avenues for potential targeting, such as *BRAF* (V600E) for SKCM. Some key differences occur for uveal melanoma (UVM), in which *GNAQ* (Q209P) and *GNA11* (Q209P/L) mutations are present in 34% and 43% of cases, respectively. These mutations may be sensitive to MEK inhibitors in SKCM undergoing clinical trials. Additionally, MEK inhibitors are being deployed for UVM to target the *GNAQ/GNA11* mutations, but may require additional agents to show clinical benefit

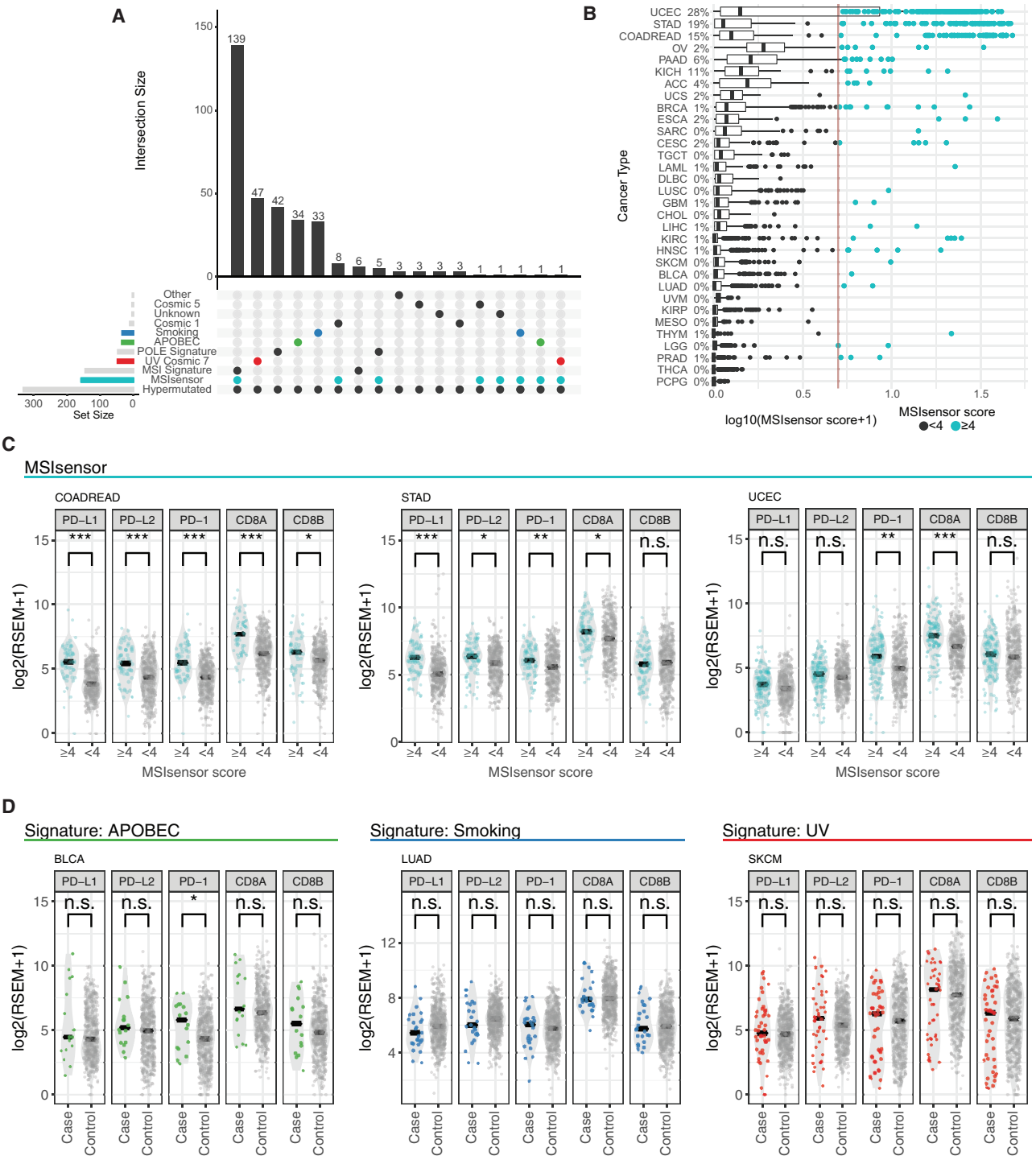


Figure 5. Hypermutators Exhibit Multiple Signatures, Microsatellite Instability, and Immune Infiltration Expression

(A) UpSetR (Conway et al., 2017) plot highlights the intersection of multiple signatures and phenotypes with hypermutated samples.

(B) MSI scores segregated by cancer types. MSI score threshold is displayed with a vertical line. The percentage of samples with high MSI is displayed to the right of each cancer type. Boxplots indicate median MSI score with 25th and 75th percentile hinges and whiskers that extend to 1.5*IQR.

(C and D) RNA-seq abundance of different immune biomarkers across signatures and MSI phenotypes defined by MSIsensor (C) and mutation signatures (D). Stars indicate significance levels using a two-sided t test to calculate p values (* < 0.05, ** < 0.01, and *** < 0.001).

See also Figure S6 and Table S5.

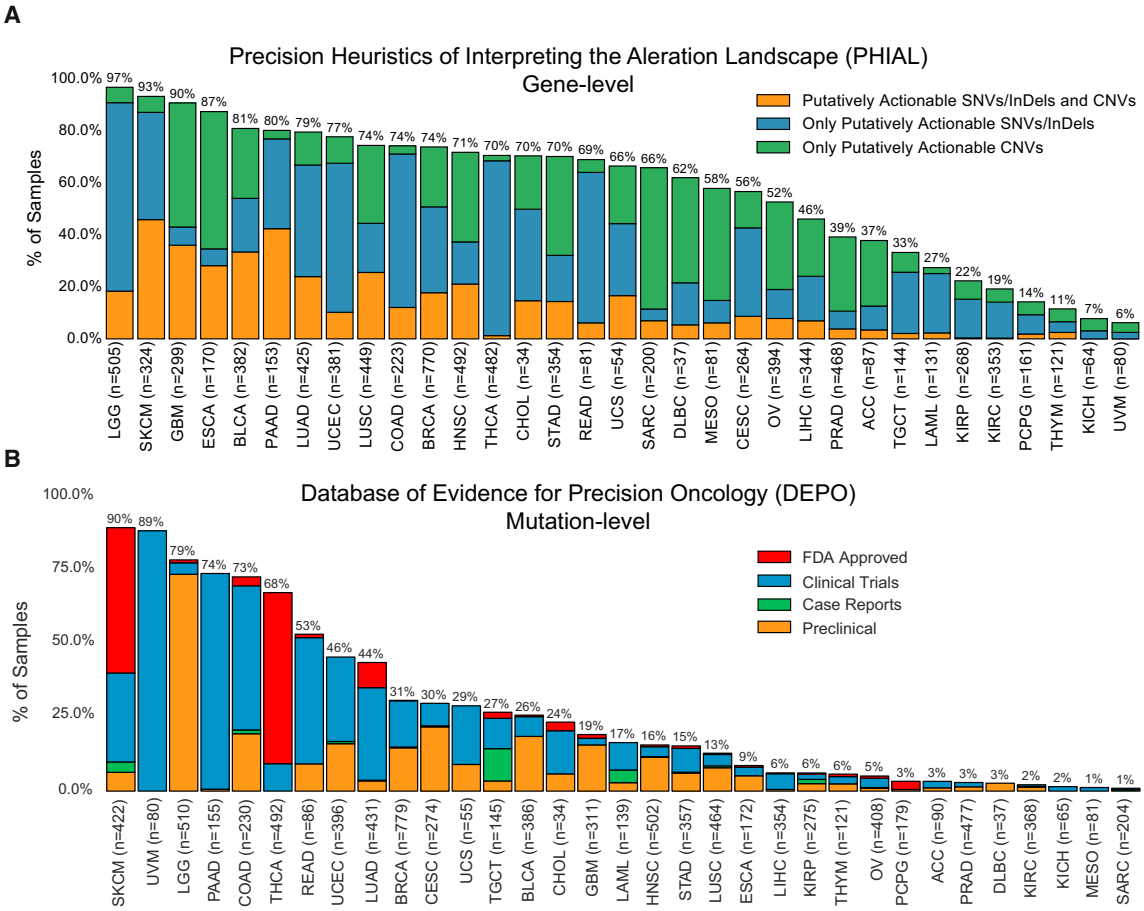


Figure 6. Putative Actionability across TCGA Studies

(A) Percentage of samples (y axis) with at least 1 putatively actionable SNV/indel/CNV (orange), SNV/indel (blue), and CNV only (green) for each cancer type (x axis) from the TARGET database. Sample size is also given for each cancer type in x axis labels. Only 8,775 samples are represented because of limitations of copy number data.

(B) Percentage of samples (y axis) with a druggable mutation (missense, indel, frameshift, and nonsense) from DEPO in each cancer type (x axis) at various stages of approval: FDA approved (red), Clinical Trials (blue), Case Reports (green), and Preclinical (orange). 9,079 samples are represented.

See also Figure S7.

(Carvajal et al., 2014). For THCA, in addition to *BRAF*, *NRAS* mutations (Q61R/K) are present in 8% of samples and could be sensitive to MEK inhibitors via repurposing; some *NRAS* mutations are sensitive in SKCM to MEK inhibition in clinical trials, particularly when combined with CDK4 inhibition (Adjei et al., 2008; Ascierto et al., 2013; Dummer et al., 2017; Iams et al., 2017). *PIK3CA* mutations (H1047R/E545K/E542K) are also prevalent in BRCA, CESC, and COAD at 24%, 20%, and 16%, respectively, in addition to UCEC, and each of these cancer types could also benefit from PI3K inhibition. Due to clinical realities and context specific pathogenesis, these percentages likely represent a ceiling of current molecular intervention potential.

DISCUSSION

We performed a PanCancer and PanSoftware analysis of one of the largest available cancer genomics datasets, identifying 299 cancer driver genes. The gene list is limited by focus on point mutations and small indels without consideration of copy-number variations (Zack et al., 2013), genomic fusions (Yoshihara et al., 2014), or methylation events (De Carvalho et al., 2012). Nevertheless, it represents the most comprehensive effort thus far to identify cancer driver genes and will serve as an important research asset.

Many important issues in the field remain unresolved, for example the similarity of driver gene sets across cancer types (Hoadley et al., 2014), mutation order and timing (founder vs. progression mutations) (Ding et al., 2012; McGranahan et al., 2015), interactions among mutations (Raimondi et al., 2016), the consequences of different mutations affecting the same gene (Torkamani and Schork, 2008), reliable tools for distinguishing driver mutations from passengers (Greenman et al., 2007), relationships between mutational signatures and driver genes (Alexandrov et al., 2013), differences between mutation burden and neoantigen load (Rizvi et al., 2015), and the implications for therapeutics (Van Allen et al., 2014). Using the

consensus genes and the functional mutations found in this study, we provided partial answers to these important questions. For example, we identified a series of clusters grouping various cancer types according to their cellular origin, highlighting the importance of the pan-squamous, pan-gynecological, and pan-gastrointestinal studies of the PanCancer Atlas.

Another important result is the dataset of 3,442 predicted driver mutations from both sequence-based and three-dimensional structure-based approaches. Because not all mutations in driver genes are actually drivers themselves, identifying the true-driver mutation subset remains a key challenge. We also used an external, independent experimental dataset to successfully validate predictions from three different approaches that predict cancer driver mutations. Our results suggest that cancer-specific sequence-based approaches outperform those aimed at detecting pathogenic variants in general. Structure-based approaches are more specific than sequence-based approaches at predicting driver mutations, but with reduced sensitivity. While functional validation confirmed true positive predictions, it gives no information regarding false negatives. Thus, what is reported here represents a lower bound. Our assay was unable to capture other factors relevant to positive selection, such as tumor microenvironment, metastasis, interactions with treatment, or the immune system. While caution must be taken when extrapolating, these observations are consistent with other functional studies on individual proteins or a subset of the proteome that have shown that mutations affecting the same three-dimensional functional regions are likely to have similar phenotypes (Brenan et al., 2016). However, we also found several instances in which sequence-based approaches captured driver mutations overlooked by structure-based approaches. Considering both approaches as complementary can improve prediction sensitivity.

We estimate that approximately half of the 10,000 TCGA samples studied here harbor a clinically relevant mutation, by predicting either sensitivity or resistance to certain treatments or clinical trial eligibility. For instance, the finding of *GNAQ* or *GNA11* mutation in uveal melanoma does not have a standard of care treatment, but a canonical activating mutation in one of these genes does allow consideration of a suite of rationally designed clinical trials (such as MEK ± PI3K inhibitors and other approaches). Under these broader considerations, we estimate that 57% (SD = 26.7%) of the TCGA cases harbor at least one potentially clinically actionable target.

The findings reported here and by the larger TCGA enterprise represent early steps toward a new era in cancer research and ultimately in cancer treatment. Studies will move beyond focusing on individual genes toward systematically integrating the myriad aspects of the cancer genome, including the interrelationships among its somatic and germline variations (Carter et al., 2017) and the tumor microenvironment and the immune system (Thorsson et al., 2018). Although this study represents the largest cancer gene and mutation study to date, we are mindful that the corpus of cancer driver genes and mutations may still be incomplete. However, it is likely that the community is nearing the beginning of the end of this phase of research, as larger cohorts continue to be examined with longer-range and longer-read sequencing technologies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data Preparation
  - Driver Discovery Approach
  - Standardized Result Reporting
  - Creation of A High Confidence Gene Set
  - Gene Discovery Weighting Strategy
  - Driver Mutation Discovery
  - Experimental Validation Data
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical Power Analysis of Driver Gene Identification
  - Anatomical Clustering of Cancer Driver Genes
  - Likely False Positive Gene Filter
  - Ctat Score
  - Normalized Entropy Score
  - Hypermutators and Immune Infiltrates
  - Druggability and Clinical Association
- DATA AND SOFTWARE AVAILABILITY
  - Algorithms used to create the consensus list
  - Population-based sequence algorithms
  - Cancer-focused algorithms
  - Structure-based algorithms
  - Additional algorithms

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.02.060.

ACKNOWLEDGMENTS

We thank patients who contributed to this study and the NCI Office of Cancer Genomics and acknowledge NIH grants from the NHGRI (U54 HG003273, U54 HG003067, and U54 HG003079) and grants from the NCI (U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, P30 CA016672, BP 2016-00296, and U24 CA211006).

AUTHOR CONTRIBUTIONS

L.D. and R.K. provided scientific direction and guided the data analysis. E.P.-P., M.H.B., S.S., and C.T. drafted the manuscript, and L.D., M.C.W, R.K., and A.J.L. revised the manuscript. M.H.B., C.T., E.P.-P., S.S., A.W., B.R., S.C., and A.C. generated the figures. P.K.-S.N., K.J.J, Z.W., and F.W performed experimental work, and G.M. provided the functional validation for somatic mutations. V.D.V., A.L., K.C., A.G., J.S., N.L.-B., A.G.-P., W.-W.L., D.W., E.V.A., G.G., M.L., E.K., M.C.W., and H.L. contributed additional scientific input and manuscript editing. B.R., S.S., and A.L. provided translational medicine insights and the figures, and L.D., M.H.B., S.C., W.-W. L., J.K., P.J.P, and I.C.-C. contributed signatures analysis of hypermutators and microsatellite unstable tumors. S.S. and Z.W. compiled mutation validation figures and furnished additional writing. A.W., D.B., S.C., and A.C. performed RNA-seq, copy number, and gene expression impact analyses,

and K.J.Y., C.S., J.H., D.C., N.N., C.R.-P., D.T., L.M., E.M.L., Q.G., J.J.G., A.W., D.B., M.H.B., E.P.-P., and C.T. were responsible for computations, including execution of all driver discovery tools. C.T., M.H.B., E.P.-P., and M.C.W. developed algorithmic and statistical procedures to aggregate results.

DECLARATION OF INTERESTS

Michael Seiler, Peter G. Smith, Ping Zhu, Silvia Buonamici, and Lihua Yu are employees of H3 Biomedicine, Inc. Parts of this work are the subject of a patent application: WO2017040526 titled "Splice variants associated with neomorphic sf3b1 mutants." Shouyoung Peng, Anant A. Agrawal, James Palacino, and Teng Teng are employees of H3 Biomedicine, Inc. Andrew D. Cherniack, Ashton C. Berger, and Galen F. Gao receive research support from Bayer Pharmaceuticals. Gordon B. Mills serves on the External Scientific Review Board of Astrazeneca. Anil Sood is on the Scientific Advisory Board for Kiyatec and is a shareholder in BioPath. Jonathan S. Serody receives funding from Merck, Inc. Kyle R. Covington is an employee of Castle Biosciences, Inc. Preethi H. Gunaratne is founder, CSO, and shareholder of NextmiRNA Therapeutics. Christina Yau is a part-time employee/consultant at NantOmics. Franz X. Schaub is an employee and shareholder of SEngine Precision Medicine, Inc. Carla Grandori is an employee, founder, and shareholder of SEngine Precision Medicine, Inc. Robert N. Eisenman is a member of the Scientific Advisory Boards and shareholder of Shenogen Pharma and Kronos Bio. Daniel J. Weisenberger is a consultant for Zymo Research Corporation. Joshua M. Stuart is the founder of Five3 Genomics and shareholder of NantOmics. Marc T. Goodman receives research support from Merck, Inc. Andrew J. Gentles is a consultant for Cibermed. Charles M. Perou is an equity stock holder, consultant, and Board of Directors member of BioClassifier and GeneCentric Diagnostics and is also listed as an inventor on patent applications on the Breast PAM50 and Lung Cancer Subtyping assays. Matthew Meyerson receives research support from Bayer Pharmaceuticals; is an equity holder in, consultant for, and Scientific Advisory Board chair for OrigiMed; and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to LabCorp. Eduard Porta-Pardo is an inventor of a patent for domainXplorer. Han Liang is a shareholder and scientific advisor of Precision Scientific and Eagle Nebula. Da Yang is an inventor on a pending patent application describing the use of antisense oligonucleotides against specific lncRNA sequence as diagnostic and therapeutic tools. Yonghong Xiao was an employee and shareholder of TESARO, Inc. Bin Feng is an employee and shareholder of TESARO, Inc. Carter Van Waes received research funding for the study of IAP inhibitor ASTX660 through a Cooperative Agreement between NIDCD, NIH, and Astex Pharmaceuticals. Raunaq Malhotra is an employee and shareholder of Seven Bridges, Inc. Peter W. Laird serves on the Scientific Advisory Board for AnchorDx. Joel Tepper is a consultant at EMD Serono. Kenneth Wang serves on the Advisory Board for Boston Scientific, Microtech, and Olympus. Andrea Califano is a founder, shareholder, and advisory board member of DarwinHealth, Inc. and a shareholder and advisory board member of Tempus, Inc. Toni K. Choueiri serves as needed on advisory boards for Bristol-Myers Squibb, Merck, and Roche. Lawrence Kwong receives research support from Array BioPharma. Sharon E. Plon is a member of the Scientific Advisory Board for Baylor Genetics Laboratory. Beth Y. Karlan serves on the Advisory Board of Invitae.

Received: August 26, 2017
Revised: November 22, 2017
Accepted: February 23, 2018
Published: April 5, 2018

REFERENCES

Adjei, A.A., Cohen, R.B., Franklin, W., Morris, C., Wilson, D., Molina, J.R., Hanson, L.J., Gore, L., Chow, L., Leong, S., et al. (2008). Phase I pharmacokinetic and pharmacodynamic study of the oral, small-molecule mitogen-activated protein kinase kinase 1/2 inhibitor AZD6244 (ARRY-142886) in patients with advanced cancers. J. Clin. Oncol. *26*, 2139–2146.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *Chapter 7*, Unit 7.20. 41.

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.

Ascierto, P.A., Schadendorf, D., Berking, C., Agarwala, S.S., van Herpen, C.M., Queirolo, P., Blank, C.U., Hauschild, A., Beck, J.T., St-Pierre, A., et al. (2013). MEK162 for patients with advanced melanoma harbouring NRAS or Val600 BRAF mutations: a non-randomised, open-label phase 2 study. Lancet Oncol. *14*, 249–256.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.-P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–689.

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., Huntsman, D.G., Caldas, C., Aparicio, S.A., and Shah, S.P. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. *13*, R124.

Bertrand, D., Chng, K.R., Sherbaf, F.G., Kiesel, A., Chia, B.K., Sia, Y.Y., Huang, S.K., Hoon, D.S., Liu, E.T., Hillmer, A., and Nagarajan, N. (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res. *43*, e44–e44.

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.; Australian Pancreatic Cancer Genome Initiative (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature *491*, 399–405.

Brahmer, J.R., Tykodi, S.S., Chow, L.Q., Hwu, W.-J., Topalian, S.L., Hwu, P., Drake, C.G., Camacho, L.H., Kauh, J., Odunsi, K., et al. (2012). Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. N. Engl. J. Med. *366*, 2455–2465.

Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N.S., Zhu, C., Bagul, M., Goetz, E.M., et al. (2016). Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. Cell Rep. *17*, 1171–1183.

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 Res. *69*, 6660–6667.

Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics *14* (*Suppl 3*), S3.

Carter, H., Marty, R., Hofree, M., Gross, A.M., Jensen, J., Fisch, K.M., Wu, X., DeBoever, C., Van Nostrand, E.L., Song, Y., et al. (2017). Interaction landscape of inherited polymorphisms with somatic events in cancer. Cancer Discov. *7*, 410–423.

Carvajal, R.D., Sosman, J.A., Quevedo, J.F., Milhem, M.M., Joshua, A.M., Kudchadkar, R.R., Linette, G.P., Gajewski, T.F., Lutzky, J., Lawson, D.H., et al. (2014). Effect of selumetinib vs chemotherapy on progression-free survival in uveal melanoma: a randomized clinical trial. JAMA *311*, 2397–2405.

Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: a precision oncology knowledge base. JCO Precis. Oncol. Published online May 16, 2017. https://doi.org/10.1200/PO.17.00011.

Chen, T., Wang, Z., Zhou, W., Chong, Z., Meric-Bernstam, F., Mills, G.B., and Chen, K. (2016). Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types. BMC Genomics *17* (*Suppl 2*), 394.

Consortium, G.; GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940.

Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M., and Park, P.J. (2017). A molecular portrait of microsatellite instability across multiple cancers. Nat. Commun. *8*, 15180.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., et al.; Mutation Consequences and Pathway Analysis Working Group of the International Cancer Genome Consortium (2015). Pathway and network analysis of cancer genomes. Nat. Methods *12*, 615–621.

De Carvalho, D.D., Sharma, S., You, J.S., Su, S.-F., Taberlay, P.C., Kelly, T.K., Yang, X., Liang, G., and Jones, P.A. (2012). DNA methylation screening identifies driver epigenetic events of cancer cell survival. Cancer Cell *21*, 655–667.

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.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Ding, L., Wendl, M.C., McMichael, J.F., and Raphael, B.J. (2014). Expanding the computational toolbox for mining cancer genomes. Nat. Rev. Genet. *15*, 556–570.

Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P.D., Cooper, D.N., Ryan, M., and Karchin, R. (2013). CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics *29*, 647–648.

Douville, C., Masica, D.L., Stenson, P.D., Cooper, D.N., Gygax, D.M., Kim, R., Ryan, M., and Karchin, R. (2016). Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum. Mutat. *37*, 28–35.

Dummer, R., Schadendorf, D., Ascierto, P.A., Arance, A., Dutriaux, C., Di Giacomo, A.M., Rutkowski, P., Del Vecchio, M., Gutzmer, R., Mandala, M., et al. (2017). Binimetinib versus dacarbazine in patients with advanced NRAS-mutant melanoma (NEMO): a multicentre, open-label, randomised, phase 3 trial. Lancet Oncol. *18*, 435–445.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., McLellan, M., Sofia, H.J., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. *6* https://doi.org/10.1016/j.cels.2018.03.002.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44* (D1), D279–D285.

Frampton, G.M., Fichtenholtz, A., Otto, G.A., Wang, K., Downing, S.R., He, J., Schnall-Levin, M., White, J., Sanford, E.M., An, P., et al. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat. Biotechnol. *31*, 1023–1031.

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.

Gagan, J., and Van Allen, E.M. (2015). Next-generation sequencing to guide cancer therapy. Genome Med. *7*, 80.

Gao, J., Chang, M.T., Johnsen, H.C., Gao, S.P., Sylvester, B.E., Sumer, S.O., Zhang, H., Solit, D.B., Taylor, B.S., Schultz, N., and Sander, C. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. *9*, 4.

Gonzalez-Perez, A., Deu-Pons, J., and Lopez-Bigas, N. (2012). Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. *4*, 89.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Gryfe, R., Kim, H., Hsieh, E.T., Aronson, M.D., Holowaty, E.J., Bull, S.B., Redston, M., and Gallinger, S. (2000). Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N. Engl. J. Med. *342*, 69–77.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al.; Cancer Genome Atlas Research Network (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Iams, W.T., Sosman, J.A., and Chandra, S. (2017). Novel targeted therapies for metastatic melanoma. Cancer J. *23*, 54–58.

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. *48*, 214–220.

Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper, D.N., Bernstein, J.A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat. Genet. *48*, 1581–1586.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kim, T.-M., Laird, P.W., and Park, P.J. (2013). The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell *155*, 858–868.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

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.

Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. N. Engl. J. Med. *372*, 2509–2520.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to drive high quality survival outcome analytics. Cell *173*, this issue, 400–416.

Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G.B., and Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS ONE *8*, e77945.

McGranahan, N., Favero, F., de Bruin, E.C., Birkbak, N.J., Szallasi, Z., and Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci. Transl. Med. *7*, 283ra254.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. *17*, 128.

Ng, P.C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. Genome Res. *12*, 436–446.

Ng, P.K.-S., Li, J., Jeong, K.J., Shao, S., Chen, H., Tsang, Y.H., Sengupta, S., Wang, Z., Bhavana, V.H., Tran, R., et al. (2018). Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell *33*, 450–462.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

Niknafs, N., Kim, D., Kim, R., Diekhans, M., Ryan, M., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum. Genet. *132*, 1235–1243.

Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.-W., Zhang, Q., McLellan, M.D., et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet. *48*, 827–837.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M.D., Wendl, M.C., and Ding, L. (2014). MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics *30*, 1015–1016.

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G.R., Creixell, P., Karchin, R., Vazquez, M., Fink, J.L., Kassahn, K.S., Pearson, J.V., et al.; International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group (2013). Computational approaches to identify functional genetic variants in cancer genomes. Nat. Methods *10*, 723–729.

Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. *39*, D465–D474.

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A pan-cancer catalogue of cancer driver protein interaction interfaces. PLoS Comput. Biol. *11*, e1004518.

Porta-Pardo, E., and Godzik, A. (2014). e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics *30*, 3109–3114.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics *21*, 3435–3438.

Pritchard, C.C., Salipante, S.J., Koehler, K., Smith, C., Scroggins, S., Wood, B., Wu, D., Lee, M.K., Dintzis, S., Adey, A., et al. (2014). Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J. Mol. Diagn. *16*, 56–67.

Raimondi, F., Singh, G., Betts, M.J., Apic, G., Vukotic, R., Andreone, P., Stein, L., and Russell, R.B. (2016). Insights into cancer severity from biomolecular interaction mechanisms. Sci. Rep. *6*, 34490.

Reimand, J., and Bader, G.D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol. *9*, 637.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118.

Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., et al. (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science *348*, 124–128.

Roberts, S.A., and Gordenin, D.A. (2014). Hypermutation in human cancer genomes: footprints and mechanisms. Nat. Rev. Cancer *14*, 786–800.

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L.B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat. Genet. *47*, 505–511.

Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al.; Oslo Breast Cancer Consortium (OSBREAC) (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400–404.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013a). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238–2244.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L., and Lopez-Bigas, N. (2013b). Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci. Rep. *3*, 2650.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Yang, T.-H.O., Porta-Pardo, E., Gao, G., Plaisier, C.L., Eddy, J.A., Plaisier, C.L., et al. (2018). The Immune Landscape of Cancer. Immunity *48*. https://doi.org/10.1016/j.immuni.2018.03.023.

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D.M., Kim, R., Ryan, M., Masica, D.L., and Karchin, R. (2016a). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. *76*, 3719–3731.

Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016b). Evaluating the evaluation of cancer driver genes. Proc. Natl. Acad. Sci. U S A *113*, 14330–14335.

Torkamani, A., and Schork, N.J. (2008). Prediction of cancer driver mutations in protein kinases. Cancer Res. *68*, 1675–1682.

Van Allen, E.M., Wagle, N., Stojanov, P., Perrin, D.L., Cibulskis, K., Marlow, S., Jane-Valbuena, J., Friedrich, D.C., Kryukov, G., Carter, S.L., et al. (2014). Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat. Med. *20*, 682–688.

Vogelstein, B., and Kinzler, K.W. (2004). Cancer genes and the pathways they control. Nat. Med. *10*, 789–799.

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

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Watson, I.R., Takahashi, K., Futreal, P.A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. Nat. Rev. Genet. *14*, 703–718.

Wong, W.C., Kim, D., Carter, H., Diekhans, M., Ryan, M.C., and Karchin, R. (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics *27*, 2147–2148.

Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol. *11*, R53.

Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R. (2014). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene *34*, 4845–4854.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Public MC3 MAF	Ellrott et al., 2018	https://gdc.cancer.gov/about-data/publications/mc3-2017
Clinical Data	Liu et al., 2018	https://gdc.cancer.gov/about-data/publications/pancanatlas
Target Drug Database - Phial	Van Allen et al., 2014	https://github.com/vanallenlab/2017-tcga-mc3_phial
DEPO	S.S., L.D., S.Q. Sun, R.J. Mashl, A.D. Scott, W. Wang, P. Batra, L.-B. Wang, and M.A. Wyczalkowski, unpublished data	http://depo-dinglab.ddns.net
OncoKB	Chakravarty et al., 2017	http://oncokb.org
Mutation Validation	Ng et al., 2018	N/A
Software and Algorithms		
20/20+	Tokheim et al., 2016b	https://github.com/KarchinLab/2020plus
MutSig2CV	Lawrence et al., 2014	http://archive.broadinstitute.org/cancer/cga/mutsig_run
MuSiC2	Dees et al., 2012	https://github.com/ding-lab/MuSiC2
OncodriveCLUST	Tamborero et al., 2013a	http://bg.upf.edu/group/projects/oncodrive-clust.php
OncodriveFML	Mularoni et al., 2016	http://bbglab.irbbarcelona.org/oncodrivefml/home
ActiveDriver	Reimand and Bader, 2013	http://individual.utoronto.ca/reimand/ActiveDriver/
CompositeDriver	This paper	https://github.com/khuranalab/CompositeDriver
HotMAPS	Tokheim et al., 2016a	https://github.com/KarchinLab/HotMAPS
CHASM	Carter et al., 2009	http://www.cravat.us/CRAVAT/
VEST	Carter et al., 2013	http://www.cravat.us/CRAVAT/
e-Driver	Porta-Pardo and Godzik, 2014	https://github.com/eduardporta/e-Driver
CanDrA	Mao et al., 2013	http://bioinformatics.mdanderson.org/main/CanDrA
HotSpot3D	Niu et al., 2016	https://github.com/ding-lab/hotspot3d
3DHotSpots.org	Gao et al., 2017	http://3dhotspots.org/3d/
e-Driver3D	Porta-Pardo et al., 2015	https://github.com/eduardporta/e-Driver
DriverNET	Bashashati et al., 2012	http://www.shahlab.ca
OncoIMPACT	Bertrand et al., 2015	https://github.com/CSB5/OncoIMPACT
MutationAssessor	Reva et al., 2011	http://mutationassessor.org/r3/
SIFT	Ng and Henikoff, 2002	http://sift.jcvi.org
PolyPhen2	Adzhubei et al., 2013	http://genetics.bwh.harvard.edu/pph2/
fathmm	Shihab et al., 2013	http://fathmm.biocompute.org.uk
transFIC	Gonzalez-Perez et al., 2012	http://bbglab.irbbarcelona.org/transfic/home
CTAT-score	This Paper	https://gdc.cancer.gov
MSIsensor	Niu et al., 2014	https://github.com/ding-lab/msisensor

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, Li Ding (lding@wustl.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The Cancer Genome Atlas (TCGA) collected both tumor and non-tumor biospecimens from 10,224 human samples with informed consent under authorization of local institutional review boards (https://cancergenome.nih.gov/abouttcga/policies/informedconsent). Here we used variants recently uniformly re-annotated that are publically available in mutation annotation file (MAF) format at the GDC (https://gdc.cancer.gov/about-data/publications/mc3-2017).

METHOD DETAILS

Data Preparation

A publicly available MAF file (syn7824274, https://gdc.cancer.gov/about-data/publications/mc3-2017) was recently compiled by the MC3 Working Group and is annotated with filter flags to highlight potential artifacts or discrepancies. This dataset represents the most uniform attempt to systematically provide mutation calls for TCGA tumors. The MC3 effort provided consensus calls from 7 software packages (Ellrott et al., 2018). Flagged artifacts include: non-exonic regions, whole-genome amplified (WGA) samples, exclusion lists, blood/tumor derived pairs, strand-bias, contamination estimations, oxo-guanine artifacts, low normal read depth, polymorphisms common in EXAC (Lek et al., 2016), mutations present in a panel of normal samples, non preferred tumor normal pairs, and mutations outside the regions of interest for any caller. If a mutation was not assigned any flag and was called by 2 or more variant calling software packages, it received a 'PASS' identifier. We restricted our analysis to PASS calls with the exception of samples from OV and LAML, which were some of the earliest sequenced by TCGA. Preparations for these samples utilized whole genome amplified (WGA) DNA, an important factor in that the WGA process can induce artifactual mutations. Of the 412 OV and 141 LAML samples present in our data 347 (84%) and 141 (100%), respectively, had variants derived from WGA DNA. In order to maintain sample sizes and uniformity in mutation calling, we did not filter mutations containing only 'wga' filter tags from these two cancer types. We recognize multiple limitations of this mutation call set, including the lack of structural variants and copy number alterations, as well as variability in sequencing depth and tumor purity. The above limitations may lead to variability in mutation detection; however, the MC3 dataset reflects the state-of-the-art in consensus mutation detection.

We also excluded highly mutated samples. These hypermutators were defined as samples with a mutation count exceeding Tukey's outlier condition, i.e., greater than 1.5 times the interquartile range above the third quartile in their respective cancer types (3Q + 1.5*IQR). Designation as a hypermutator also required the number of mutations in a sample to exceed 1,000, a heuristic that limited the number of discarded samples in low mutation rate cancer types (Figure S1). LUAD, SKCM, and UCEC had hypermutator thresholds greater than 1,000 mutations (1,047, 2,122, and 2,545, respectively) (Figure 1B). We also excluded samples that were flagged by the analysis-working group based on pathology, but allowed "RNA degradation" samples to remain, as this factor is not particularly relevant for most driver prediction tools based on mutations (https://www.dropbox.com/sh/wglgggbgketh982/AABJEqQ2QdCEruy9c6UXBdjba?dl=0) (Table S6). The final driver-discovery dataset consisted of 9,079 samples having a total of 791,637 missense mutations, 323,884 silent mutations, 96,196 3′ UTR mutations, 57,900 nonsense mutations, 42,251 intronic mutations, 42,251 Frameshift deletions, 34,266 5′ UTR, 21,804 splice site mutations, 19,856 RNA mutations, 11,305 frameshift insertions, 7,622 3′ flanking mutations, 6,419 5′ flanking mutations, 6,144 in-frame deletions, 1,362 translation start site mutations, 964 nonstop mutations, and 632 in-frame insertions.

Driver Discovery Approach

Using multiple tools can overcome numerous technical issues that confound individual statistical analyses to find driver genes, such as heterogeneous mutation rate across the genome (Lawrence et al., 2013), inflated significance for long genes (Watson et al., 2013), and false positive calls in cancers with high mutation rates (Tokheim et al., 2016b). We used 26 computational tools, spanning 10 different institutions, to identify mutation-based driver genes and driver mutations (Figure S1A). We divided the analysis into two phases: (I) driver gene-discovery and (II) gene and in-silico mutation validation (Figure 1C; STAR Methods). In the first phase, we applied 8 different tools comprising algorithms based on mutation frequency (MuSiC2 [Dees et al., 2012] and MutSig2CV [Lawrence et al., 2014]), features (20/20+ [Tokheim et al., 2016b], CompositeDriver [https://github.com/khuranalab/CompositeDriver] and OncodriveFML [Mularoni et al., 2016]), clustering (OncodriveCLUST [Tamborero et al., 2013a]), and externally defined regions (e-Driver [Porta-Pardo and Godzik, 2014] and ActiveDriver [Reimand and Bader, 2013]). The second phase used an additional 16 tools to further characterize the consensus genes from phase one. The collection was comprised of 8 mutation-level algorithms (SIFT [Ng and Henikoff, 2002], PolyPhen2 [Adzhubei et al., 2013], MutationAssessor [Reva et al., 2011], transFIC [Gonzalez-Perez et al., 2012], fathmm [Shihab et al., 2013], CHASM [Carter et al., 2009], CanDrA [Mao et al., 2013] and VEST [Carter et al., 2013]), 4 structure-based (HotSpot3D [Niu et al., 2016], HotMAPS [Tokheim et al., 2016a], 3DHotSpots.org [Gao et al., 2017] and e-Driver3D [Porta-Pardo et al., 2015]), 2 network and –omic integration tools (OncoIMPACT [Bertrand et al., 2015], DriverNet [Bashashati et al., 2012]), and 2 algorithms to identify clinically-actionable events (PHIAL [Van Allen et al., 2014] and DEPO [S.Q. Sun, R.J. Mashl, S. Sengupta, A.D. Scott, W. Wang, P. Batra, L.-B. Wang, M.A. Wyczalkowski, L. Ding, unpublished data]). Each tool reported gene or mutation level scores and/or p values along with a brief description of recommended cutoff thresholds or filters. Finally, the CTAT algorithm was applied separately to population based and cancer based tools. This accounts for the remaining 2 tools (this manuscript) for a total of 26 tools (https://gdc.cancer.gov/about-data/publications#/?groups=PanCanAtlas).

Tools integrating –omics data analyzed a smaller subset of TCGA, since we had to remove 75 samples that had problems regarding RNA-degradation. This issue did not affect the algorithms based only on somatic mutation data, so these 75 samples were included in their analyses (Table S6).

Standardized Result Reporting

Despite the variety in available data within the TCGA cohort, each of the 26 tools supplied tissue and PanCancer level predictions and results. We defined a standardized file format to facilitate multi-tool comparison, so each tool supplied information on genes, transcripts, missense mutations, scores, p values, q-values and additional information needed for tool specific requirements.

Creation of A High Confidence Gene Set

We identified a preliminary total of 2,101 potential drivers by taking the union of genes predicted by the eight driver-gene discovery tools. As illustrated in Figure S2A, the increased number of false positive genes is likely due to any individual tool's capability to maintain sound statistical properties that handle a complex set of factors such as tumor heterogeneity, increased mutation rates, and variable sample sizes. We refined this list by calculating, for each gene predicted in each cancer type, a consensus score that compensated for outlier results and correlation among tools (Figure S2 and Table S1; https://gdc.cancer.gov/about-data/publications#/?groups=PanCanAtlas). The consensus score was defined as a weighted sum of the number of tools that predicted the gene to be a driver in each cancer type (see Gene Discovery Weighting Strategy). We required a minimum of two tools to agree, where both could not be outliers (score $\geq$ 1.5). Although it is difficult to distinguish the overall performance improvement on a small number of held out CGC genes (Figure S3A), the weighting strategy did have higher specificity (p = 4.3e-8, McNemar test), which is preferable given concerns of false positives. Regardless, the consensus score performance on identifying CGC genes (Figure S3A) support previous reports that merging the results from different algorithms improve cancer driver discovery (Tamborero et al., 2013b).

To maximize the coverage of our analysis and ensure the accuracy of our final list, we reviewed previous findings in 31 individual cancer types and PanCancer-12 from TCGA. For cancer types not yet having a TCGA publication, we consulted with the relevant analysis working groups (LIHC, testicular germ cell tumors [TGCT], UVM, sarcoma [SARC], PAAD, and THYM). We included in our final consensus list all those genes that were previously described as drivers by experts in the cancer-specific analysis of TCGA datasets and were also identified by at least one of the eight algorithms, even if they did not meet our consensus score threshold ($\geq$1.5) (Figure 2A). This resulted in an additional 54 gene-cancer pairs, such as *ATR*, *CHEK2*, *IDH2,* and *ERCC2* in the PanCancer dataset and *FOXA1* in BLCA, *HRAS* in SKCM, and *MET* in LUAD (Figures S2B–S2F). The majority of this effort resulted in linking cancer genes identified by our strategy to additional cancer types based on previous literature (32/54).

The process of identifying genes in previous TCGA publications consisted in the following steps:

1. We manually reviewed all the official marker papers for each cancer type of The Cancer Genome Atlas. When no official paper was yet available, we contacted the lead analyst of the cancer type to access the official list of cancer driver genes.
2. We listed all the genes that were identified in the main text of one of the main figures of the corresponding paper as significantly more mutated than expected by chance.
3. Once we had the genes from each cancer type, we checked whether these genes had also been identified in our analyses by, at least, one algorithm. Note that both the mutation calls and the samples from the original TCGA paper and our analysis of each cancer type differ to some extent, so it is possible that genes which were previously identified by MutSigCV or MuSiC are not found by these algorithms in our analysis.
4. If a gene had been identified in the dedicated cancer type, deemed important enough to be highlighted in the main text/figure of the paper, and was also identified by at least one of our 8 gene-level discovery tools, we rescued it for our final list (Table S1).

To limit false positives in the expanded list, we applied linear discriminant analysis (Figure S2C) (see Likely False Positive Gene Filter). We identified and removed 45 genes from the consensus we detected as likely false positives. These included *CACNA1E* in PanCancer, *COL11A1* in LUAD, *DST* in GBM, and *TTN* in SKCM. The consensus list from the above systematic approach consisted of 258 unique genes (Table S1). The average number of non-silent mutations per sample in our consensus gene list varied substantially by cancer type ranging from < 1 in 12 cancer types (ACC, CHOL, KICH, kidney renal papillary cell carcinoma [KIRP], LAML, MESO, PCPG, PRAD, SARC, TGCT, THCA, and THYM) to 7.3 in UCEC. A median of 85% of tumors harbored non-silent mutations in consensus genes across cancer types (Figure S3F).

Given the limitations of a systematic approach, we additionally manually rescued 41 genes (Table S1). In the rescue attempt, we started with a list of genes identified from previous TCGA marker papers but not found from our systematic approach. We rescued genes with supportive evidence from the following sources: hypermutator phenotype related genes (since we excluded hypermutated samples in our systematic discovery; 6 genes), established cancer genes from LAML because of low quality variant calling originating from liquid tumor contamination of the normal samples (6 genes), genes supported by omic network tools (DriverNet and OncoIMPACT; 25 genes), and a gene supported by all three approaches from the driver mutation discovery (1 gene). Addition of genes to the final list was subjected to expert manual curation (3 genes).

The final consensus gene list consisted of 299 unique genes across 33 cancer types and the PanCancer dataset (Figure 2A; Table S1). The list captures most previously described driver genes for the majority of cancer types. We overlapped the cancer driver genes obtained from the consensus approach without manual curation with those from 5 independent studies in 4 cancer types (BRCA, PRAD, PAAD, and LIHC) of which one is whole-genome sequencing. The consensus approach always had a greater inter-study overlap, with an average increase of 26% over only using a single tool, either MuSiC2 or MutSig2CV (Barbieri et al., 2012; Biankin et al., 2012; Nik-Zainal et al., 2016; Schulze et al., 2015; Stephens et al., 2012) (Table S3). Among the 299 genes we identified 59 novel genes that were not previously identified in 6 previous PanCancer publications (Frampton et al., 2013; Kandoth et al., 2013; Lawrence et al., 2014; Pritchard et al., 2014; Tamborero et al., 2013b; Vogelstein et al., 2013) or the cancer gene census list (http://cancer.sanger.ac.uk/census/) (Futreal et al., 2004) (Table S1).

Gene Discovery Weighting Strategy

Tools predicting cancer genes were weighted according to their performance in each cancer type, receiving half the weight if a result was deemed an outlier, thereby obligating additional tool agreement (Figure S2A). Specifically, we examined quality metrics across tools and within the same tool, which allowed us to identify outlier results. We marked outliers based on the quasi-majority of three criteria: low concordance with known cancer genes, high divergence of p value distribution from theoretical expectation, and abnormally high number of significant genes. The first criterion evaluated the fraction overlap of significant genes with a previously manually curated set of driver genes from (Vogelstein et al., 2013) compared with the median across all tools. The second criterion examined whether the divergence of observed p values from those theoretically expected by the Mean Log Fold Change (MLFC) (Tokheim et al., 2016b) was greater than the median of all tools, which may indicate a tool's statistical assumptions may not be well satisfied. The third criterion examined whether a tool's prediction for particular cancer types appeared as an outlier in terms of the number of significant genes compared against all of the results for that tool (Tukey's outlier criterion: number significant > 3Q + 1.5*IQR). We calculated a gene consensus score by summing the tools that declared the gene as being significant, with a weight of 1 for non-outlier results and 0.5 for outlier results.

We also provided a score that is more stringent, which could be used by others to create a somewhat smaller set of confident driver genes (Table S1). Here, due to similarities in algorithmic decisions, we adjusted these consensus gene scores to compensate for correlation between tools of the same class (i.e., frequency, feature, and domain based tools). The contribution of a tool whose inference is uncorrelated with other tools is recorded by simple addition of its score to the running total. However, some tools show correlation at sufficient levels that their contributions should properly be considered in aggregate. For example, MuSiC2 and MutSig2CV are highly correlated, as are CompositeDriver and OncodriveFML (Figure S2G). For such tool pairs, we actually add the union of their scores, S1 U S2, to the running total in the form of

$$S_1 \cup S_2 = S_1 + S_2 - S_1 \cap S_2 = S_1 + S_2 - \frac{\varrho}{2}(S_1 + S_2) = \left(1 - \frac{\varrho}{2}\right)(S_1 + S_2) \quad \text{(Eq. 1)}$$

where ρ is the Pearson's coefficient between these two tools. We applied this procedure for pairs of tools whose variances exceeded 10%, i.e., for correlations greater than 0.32. Small changes of this threshold did not have any meaningful effect.

Driver Mutation Discovery

To maximize the coverage of our analysis we used 12 tools that look for three distinct hallmarks of "driverness." We utilized four tools that distinguish pathogenic mutations from benign polymorphisms on a population level (SIFT [Ng and Henikoff, 2002], PolyPhen2 [Adzhubei et al., 2013], VEST (version 3 scores) [Carter et al., 2013] and MutationAssessor [Reva et al., 2011]), four tools specifically designed to distinguish between driver and passenger somatic mutations (CHASM [Wong et al., 2011], CanDrA [Carter et al., 2013], fathmm [Shihab et al., 2013] and transFIC [Gonzalez-Perez et al., 2012]) and four tools that leverage information from protein structures (HotSpot3D [Niu et al., 2016], HotMAPS [Tokheim et al., 2016a], 3DHotSpot.org [Gao et al., 2017] and e-Driver3D [Porta-Pardo et al., 2015]). In order to combine the predictions from the sequence-based approaches we used principal component analysis to develop a Combined Tool Adjusted Total (CTAT) scores for both, population-based and cancer-specific scores (STAR Methods). Principal component analysis has been previously shown successful in a similar task of prioritizing germline mutations (Ionita-Laza et al., 2016). We also combined the results from three-dimensional tools by adding the number of tools that predicted a specific position as belonging to a cancer-mutation cluster. Finally, to limit the number of false positives, we focused our analysis on the genes of our consensus driver list.

To define the CTAT score thresholds, we used the maximum balanced accuracy when predicting OncoKB mutations "oncogenic" or "likely oncogenic" (Figures S5C and S5D). This yielded a threshold of 1.2 for CTAT-population and 2.4 for CTAT-cancer. For the structural algorithms, we report a mutation as likely driver if at least 2 algorithms identify it within a cluster. Finally, we evaluated the performance of each CTAT score using mutations from OncoKB labeled as "likely oncogenic" or "oncogenic" as true-positives.

Experimental Validation Data

For experimental validation to assess tool performance, we utilized experimental data provided by Gordon Mills at MD Anderson Cancer Center (Ng et al., 2018). 1049 mutations were tested in 2 growth-factor dependent cell models, Ba/F3 and MCF10A. Both models depend on specific growth factors for survival, with which they cease proliferating. It is hypothesized that a mutation is a driver if it confers survival advantage to cells even in the absence of these growth factors. Mutations were introduced in the cells and the dependent growth factors were withdrawn; subsequently, cell viability was measured. Every experiment had 2 negative controls, 3 positive controls, and a corresponding wild-type (WT) of the mutation tested. In general, we considered a mutation to be 'validated' if the cell viabilities of the mutations were higher than those of the wild-type.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical Power Analysis of Driver Gene Identification

We performed the statistical power analysis of driver gene identification at various prevalences (effect size = 0.1, 0.05, 0.02, and 0.01, fraction of samples above background) with 90% power, based on a previously established approach of elevated mutation rate (Lawrence et al., 2014). We used a binomial model implementation (https://github.com/KarchinLab/cancerSeqStudy),

previously described (Tokheim et al., 2016b). Default parameters were used. We placed each cancer type or PanCancer analysis according to the median mutation rate (per mega base) and number of samples (n shown in Figure 1C). Mutation rate per mega base was calculated through using sequencing coverage of samples obtained from the MuSiC2 analysis.

Anatomical Clustering of Cancer Driver Genes

We performed hierarchical clustering of the gene consensus scores for the 87 genes that were found in more than one cancer type (Figure S3E), thereby clustering both genes and cancer types (n = 32 cancer types, COAD and READ merged by maximum consensus gene score). The correlation distance metric and average linkage were used to avoid clustering purely based on the total number of consensus genes for a cancer type. Clusters of genes were defined based on cutting the dendrogram at a depth chosen by manual inspection. Each gene cluster was tested for enrichment in three groups of cancer types using a permutation test: Pan-squamous (BLCA, CESC, LUSC, HNSC, and ESCA), Pan-gynecological (BRCA, UCEC, UCS, CESC, and OV), and Pan-gastrointestinal (STAD, COADREAD, ESCA, and PAAD). This involved, for each cluster and group of cancer types, an initial calculation of the total gene consensus score from the observed data. Labels for the cancer types were then permuted 10,000 times and the total gene consensus score was subsequently recalculated based on the permuted cancer type labels. Lastly, P values were calculated as the fraction of permuted iterations that met or exceeded the observed total gene consensus score. P values were then multiple test corrected across all genes using the Benjamini-Hochberg FDR method.

Likely False Positive Gene Filter

We attempted to harness the collective ability of the analysis tools in order to remove remaining genes that were likely false positives using Fisher's linear discriminant analysis (LDA). This is a PanCancer filter in the sense that we selected features by manually examining 4 attributes for each of the tools. Specifically, for each gene, we compiled average P value over all cancers and the Pearson correlation coefficient, regression slope, and y-intercept of a least-squares fit between the cancer background mutation rates and tool P values. We then looked for the largest difference of means in units of standard deviations for these 4 attributes between a set of true positive list in the form of the 127 genes from Kandoth et al., 2013 versus an internally-curated list of 488 false positives (Table S7). We ultimately chose 4 features: the correlation coefficient from MuSiC2, the average P values from OncodriveFML and 20/20+, and the y-intercept from 20/20+. To harness these features collectively, we then solved the LDA linear algebra problem using decomposition, where the coefficient matrix is comprised of the within-groups variances, the vector of unknowns contains the feature weights, and the right hand side is the vector of the difference of means of the features. We then chose a conservative cut-point such the true positives were unlikely to be caught in the filter, reflecting 90% sensitivity for keeping associations found in Cancer Gene Census genes. Using the 4 LDA weights and the cut-point, we then ran the candidate gene list through the filter, removing all genes that failed the cut-point. However, we omitted from this filtering any gene already established as being a cancer gene and any "out-of-context" gene, meaning ones that showed obvious specificities to a single cancer.

Ctat Score

We developed the **C**ombined **T**ool **A**djusted **T**otal (CTAT) score to distinguish missense mutations that are cancer drivers from passenger mutations. The CTAT score combines multiple individual tools that prioritize missense mutations. To normalize each score, we calculated the z-score by subtracting the mean score and then dividing by the standard deviation. We then performed principal component analysis (PCA) using ScikitLearn v0.18.0 and used the score along the first principal component as our CTAT score, representing the scalar projection onto the first eigenvector. Only missense mutations that had no missing values for each of the combined tools were used in generating the principal component analysis. We performed this procedure on two distinct categories of tools, "population-based" tools that distinguish damaging/pathogenic germline missense variants from common polymorphisms (SIFT, PolyPhen2, VEST, and MutationAssessor), and "cancer-focused" tools designed to distinguish somatic missense mutations that are drivers from passengers (CHASM, CanDrA, fathmm, and transFIC). To score the remaining missense mutations that did have a missing score, we imputed missing scores of the individual tool with the mean for the method. Imputation was only performed for the cancer-focused tools as the population-based tools had too many missing values.

Normalized Entropy Score

We calculated a score to characterize consensus genes on their diversity of amino acid positions that contain either missense, frameshift, or truncating mutations. Because genes may be of different length and have different background mutation rates, we used a normalized entropy score (E) (Tokheim et al., 2016b):

$$E = \frac{-\sum_{i=1}^{n} p(i) log_2(p(i))}{log_2(n)}, \quad \text{(Eq. 2)}$$

where, for each gene, n is the total number of mutated positions and p(i) represents the fraction of mutations for the i-th mutated position. The normalized entropy score takes values between 0 and 1, with values closer to one indicating an even spread of mutations across all mutated positions.

Hypermutators and Immune Infiltrates

Hypermutator samples were defined above as those tumors with mutation counts greater than 1.5 times the interquartile range above the third quartile in their respective cancer types (3Q + 1.5*IQR). Additionally, mutations in a sample needed to exceed 1000, a heuristic that limited the number of discarded samples in low mutation rate cancer types (Figure S1). Three cancer types, LUAD, SKCM, and UCEC, had hypermutator thresholds greater than 1000 mutations (1047, 2122, and 2545 respectively) (Figure 1B).

18 global mutational signatures were originally calculated for each of the hypermutator samples according to Alexandrov et al., 2013 with a minimum cosine similarity ranging from 0.57 to 0.99. These signatures were then aggregated into the 9 representative signatures presented: POLE was comprised of "POLE" and "MSI - COSMIC14 (POLE+MSI)"; MSI combined "MSI - COSMIC15," "MSI - COSMIC20 (POLD+MSI)," "MSI - COSMIC21," "MSI - COSMIC26," and "MSI - COSMIC6"; COSMIC signature 5 combined "COSMIC5," and "ERCC2 - COSMIC5," unknown is comprised of "Unknown" (many of which were attributable to noise from WGA and 3 hypermutated samples were not performed in this analysis); UV, smoking, APOBEC, COSMIC1, and COSMIC5 signatures did not require aggregation; and other was comprised of "COSMIC17," "COSMIC22 - aristolochic acid signature" and "COSMIC3 – BRCA" (Figure 5A). A primary signature for each sample was calculated by identifying as the max score from each signature.

MSIsensor (Niu et al., 2014) was applied to all 9,423 samples in our dataset. We used the authors' recommended cut-off of greater than or equal to 4 in order to indicate MSI-High status. Scores below 4 cannot reliably distinguish been MSI-Low and MSS. More information on this tool is found in DATA AND SOFTWARE. 357 scores were generated from BAM files other than those used for variant calling by the MC3 Working group. Of the 357 samples, 29 had MSIscores greater than or equal to 4. 16 of these 29 samples (55%) had at least one frameshift/nonsense, missense mutatiation in gene involved in MSI or MMR phenotype (*POLE, MLH1, MLH3, MGMT, MSH6, MSH3, MSH2, PMS1,* or *PMS2*) or had high *MLH1* methylation. Results from 180 gel-assays were provided by The Broad Institute to assess MSIsensor scores. Using a multiple regression model, quantitative MSI scores correlated with qualitative results from the gel-assay (MSI-H, MSI-l, and MSS, p value $< 2x10^{-16}$, $r^2 = 0.504$); thus, justifying the use of MSIsensor.

PD-L1, PD-L2, PD-1, CD8A, and CD8B RPPA expression data were collected from FIREHOSE (January 28, 2016). By cancer type, samples were stratified by MSIsensor score status (Figure 5C), hypermutatator and mutation signatures status (Figure 5D), and hypermutator status alone (Figures S7A–S7C). Significance was calculated using two-sided t test statistics.

Druggability and Clinical Association

PHIAL is a heuristic clinical interpretation algorithm and database of tumor alterations relevant to genomics-driven therapy (TARGET) and was created in 2014 to identify putatively actionable or biologically relevant alterations in patient tumor sequence data. Although it was developed to study patients individually, PHIAL was applied to all 8775 samples that had both SNV/indel and thresholded copy number data available across TCGA MC3 and all 33 individual TCGA studies. PHIAL (1.2.0) using TARGET 1.4.2 and Cosmic v79 was applied to all 8775 samples that had both SNV/indel and thresholded copy number data available across TCGA MC3 and all 33 individual TCGA studies. TARGET contains 50 alteration-therapeutic assertions based on FDA-approved therapies, clinical trials, or published clinical evidence of genetic alteration-therapeutic action relationships which was leveraged by PHIAL to bin variants as *putatively actionable*, if both the gene and alteration type match an assertion, or *biologically relevant*, if only the gene matches.

DEPO version 1.0 (S.Q. Sun, R.J. Mashl, S. Sengupta, A.D. Scott, W. Wang, P. Batra, L.-B. Wang, M.A. Wyczalkowski, L. Ding, unpublished data; http://depo-dinglab.ddns.net) is a manually curated database of single nucleotide polymorphisms or SNPs (missense, frameshift, and nonsense mutations), in-frame insertions and deletions (indels), copy number variations (CNVs), and expression changes that are paired with drug responses. For present purposes, we focused strictly on SNPs and indels. For each variant-drug pair, there is an associated tumor type, an effect (sensitive or resistant), and a level of evidence describing the quality of data supporting the pair at various stages of approval: FDA-approved, clinical trials, case reports, and preclinical. We queried our samples for presence of druggable alterations from DEPO regardless of cancer type. The cancer type that had the highest level of evidence for a drug-variant pair was considered the "on-label" cancer type and all other cancer types were deemed to be "off-label" (Figure S7D). Cancer types containing an off-label variant were still considered to be 'druggable' via repurposing.

DATA AND SOFTWARE AVAILABILITY

Algorithms used to create the consensus list

20/20+

20/20+ is a Random Forest machine learning algorithm for predicting oncogenes and tumor suppressor genes from somatic mutations. 20/20+ uses features capturing mutational clustering, evolutionary conservation, predicted functional impact of variants, mutation consequence types, gene interaction network connectivity, and other relevant covariates. 20/20+ version 1.1.0 was run using default parameters, as described previously (Tokheim et al., 2016b), except where the number of simulations was increased to 100,000. We applied gene hold-out cross-validation to perform predictions without over-fitting. Additionally, for cancer type specific predictions, we held out all mutations from the corresponding cancer type in our training set. P value QQ-plots suggest well-calibrated predictions that are not inflated for false positives and results show substantial overlap with the cancer gene census (Futreal et al., 2004) and curated driver genes (Vogelstein et al., 2013). Genes were deemed significant if either the oncogene, tumor suppressor gene, or driver score had a q-value of less than or equal to 0.05. 20/20+ was also used to categorize the consensus genes as

either a oncogene, tumor suppressor gene, or unknown. A "likely" oncogene or tumor suppressor gene was determined using q-value threshold of 0.05, while "possible" status was assigned to the remaining genes with a p value less than or equal to 0.05.

MutSig2CV

MutSig2CV (Lawrence et al., 2014) analyzes somatic point mutations discovered in DNA sequencing, identifying genes mutated more often than expected by chance given inferred background mutation processes. Genes were deemed significant at a q-value threshold of 0.1. MutSig2CV consists of three independent statistical tests, described briefly below:

Abundance (CV)

The most important step for inferring genes' mutational significance is to properly classify whether the gene is highly mutated relative to some background mutation rate (BMR), which varies on a macroscopic level across patients and genes and on a microscopic level across sequence contexts. MutSig accounts for all three of these aspects, renormalizing BMR on a per-gene, -patient, and -context level.

Clustering (CL)

Genes often harbor mutational hotspots, specific sites that are frequently mutated. While abundance calculations bin mutations on the gene level, clustering bins mutations on the local site level, which allows MutSig to differentiate between genes with uniformly distributed mutations and genes with localized hotspots, assigning higher significance to the latter.

Conservation (FN)

MutSig uses evolutionary conservation as a proxy for determining the functional significance of a mutated site. It assumes that genetic sites highly conserved across vertebrates have greater functional significance than weakly conserved sites. MutSig assigns a higher significance to genes that experience frequent mutations in highly conserved sites.

MuSiC2

MuSiC2 (Dees et al., 2012) version 0.2 is a frequency based tool used to identify significantly mutated genes (https://github.com/ding-lab/MuSiC2). Significance is determined by comparing a calculated background mutation frequency to a convolution for specific transition, transversion, and CpG variants. Default parameters were used for initial SMG identification. A recent update to MuSiC2 provides a long gene filter, which seeks to remove false positives by virtue of finding genes whose elevated mutation tallies are due primarily to their larger size rather than their mutational significance. Briefly, it systematically tightens the p value threshold for longer genes (> 5000nt) based on a table test of uncoupling gene status (significant versus not significant) from gene size (long gene versus typical-size gene).

OncodriveCLUST

OncodriveCLUST (Tamborero et al., 2013a) identifies genes with non-silent mutations that cluster together in protein sequence more than expected based on a background distribution of synonymous mutations. OncodriveCLUST was run through a local installation of IntOGen pipeline (available at https://bitbucket.org/intogen/intogen-pipeline). Different minimum mutation thresholds were set manually, according to the mutation burden of the different cancer types: 3 (in ACC, CHOL, DLBC, ESCA, GBM, KICH, kidney renal clear cell carcinoma (KIRC), KIRP, LGG, MESO, PAAD, PCPG, PRAD, READ, SARC and THYM), 5 (in BRCA, CESC, COAD, LAML, LIHC, OV, TGCT, THCA, UCS, UVM and the PANCANCER run), 7 (in HNSC, SKCM and STAD), 10 (in BLCA) and 12 (in LUAD, LUSC and UCEC). Next, we applied a custom expression filter in each cancer type by filtering out genes whose median expression level was lower than 6 log2 RSEM in that particular cancer type. Genes were found significant at a q-value threshold of 0.05.

OncodriveFML

OncodriveFML (Mularoni et al., 2016) identifies genes that have greater accumulation of mutations that have higher predicted function impact (functional impact bias). The predicted impacts of mutations were scored using CADD (Kircher et al., 2014). The mean CADD score for mutations was compared to permuted mutations within the same gene to calculate an empirical p value. The results have been calculated considering all the observed mutations in CDS regions. CDS regions were extracted from Gencode release 19 (https://www.gencodegenes.org/releases/19.html). The annotations include all CDS where both the "gene_type" and the "transcript_type" were tagged as "protein_coding." The analysis was performed using OncodriveFML version 1.0.2-alpha with the coding indels option specified. The configuration file contained the default parameters with the following exceptions (https://bitbucket.org/bbglab/oncodrivefml/downloads/PanCanAtlas.conf). Genes were deemed significant at a q-value of 0.25.

ActiveDriver

ActiveDriver detects genes that are enriched in somatic mutations located in post-translationally modified sites, such as phosphorylation, acetylation, or ubiquitination sites. It identifies driver genes using a logistic regression that takes into account, among other factors, the position of the PTM sites and the distribution of the mutations (Reimand and Bader, 2013). ActiveDriver (v0.010, default parameter) was run using the database ActiveDriver_HG38. Due to high mean log fold change (MLFC) values, genes were deemed significant at a q-value of 0.0001.

e-Driver

This algorithm identifies protein regions that are enriched in somatic missense mutations using a binomial test and assuming mutations are distributed randomly across the protein. The protein regions can be linear (Porta-Pardo and Godzik, 2014) or three-dimensional (Porta-Pardo et al., 2015). The current analysis uses PFAM domains (Finn et al., 2016) and disordered regions predicted by Foldindex (Prilusky et al., 2005) for the linear analysis. We used the regions described in https://github.com/eduardporta/e-Driver/.

CompositeDriver

We have developed CompositeDriver v0.1 (https://github.com/khuranalab/CompositeDriver), a novel computational method considering both mutation recurrence and functional impact of mutations to identify signals of positive selection. For all mutations within a gene's protein coding region, a composite score was calculated through summation of mutation recurrence multiplied by the functional impact score (Jagadeesh et al., 2016). For each gene, a p value was computed by testing whether the observed composite score is significantly higher than the null distribution. To build the null distribution from the background, the same numbers of mutated positions were repeatedly drawn (default is 10^5 times) from other protein coding regions of similar replication timing and similar mutation context (Alexandrov et al., 2013). The Benjamini-Hochberg method for multiple hypothesis correction and q value cut-off of 0.05 was used.

Population-based sequence algorithms

VEST

VEST (Variant Effect Scoring Tool) is a machine learning method that predicts the functional significance of missense mutations observed through genome sequencing, allowing mutations to be prioritized in subsequent functional studies based on the probability that they impair protein activity (Carter et al., 2013; Douville et al., 2016). VEST version 3.0 scores were retrieved from the CRAVAT web server (v4.3) (Douville et al., 2013).

MutationAssessor

MutationAssessor (Reva et al., 2011) uses residue conservation across species to identify the impact of non-synonymous mutations. Scores were obtained using the precompiled database ljb26_all from ANNOVAR v20150322 (Wang et al., 2010).

PolyPhen2

Polymorphism Phenotyping v2 (PolyPhen2) (Adzhubei et al., 2013) is a machine learning approach that computes the functional impact of missense mutations. The method uses sequence-based and structure-based features to train a naive Bayes classifier. Scores were obtained using the precompiled database ljb26_all from ANNOVAR (Wang et al., 2010).

SIFT

Sorting Intolerant from Tolerant (SIFT) SIFT (Ng and Henikoff, 2002) predicts the functional impact of missense mutations using sequence homology. Scores were obtained using the precompiled database ljb26_all from ANNOVAR v20150322 (Wang et al., 2010).

Cancer-focused algorithms

CHASM

CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations) is a machine learning method that predicts the functional significance of somatic missense mutations observed in the genomes of cancer cells, allowing mutations to be prioritized in subsequent functional studies, based on the probability that they give the cells a selective survival advantage (Carter et al., 2009). CHASM scores (precompute version 3.0) were retrieved from the CRAVAT web server (v4.3) (Douville et al., 2013).

CanDrA

CanDrA (Mao et al., 2013) is a machine learning program that predicts cancer-type specific driver missense mutations based on 96 structural, evolutionary and gene features computed by over 10 other functional prediction algorithms such as CHASM, SIFT, and MutationAssessor. CanDrA used COSMIC, TCGA, and CCLE data for training and is heavily optimized to perform cancer-type specific driver mutation analysis (Chen et al., 2016). If a mutation appeared more than once, the maximum CanDrA score was taken. In this work, the CanDra "plus" version was run under default parameters using the "general" cancer type database.

fathmm

Functional Analysis Through Hidden Markov Models (fathmm) (Shihab et al., 2013) uses Hidden Markov modeling to represent the protein domain shared across human proteins and to estimate the functional impact of mutations. Using cancer-associated polymorphisms from CanProVar and putative neutral polymorphisms from UniProt, fathmm prioritizes mutations that are associated with cancer versus those that simply impact the function of a protein. Scores were obtained using the precompiled database FATHMM cancer v2.3 (http://fathmm.biocompute.org.uk/database/fathmm.v2.3.SQL.gz).

transFIC

Transformed Functional Impact score for Cancer (transFIC) (Gonzalez-Perez et al., 2012) assesses the functional impact of tumor non-synonymous SNVs by accounting for baseline tolerance of functional variants in relation to genes. This is performed by grouping genes by ontologies and assessing the tolerance of gene sets using functional scores provided by SIFT, PolyPhen2, and MutationAssessor. By transforming scores based specific ontologies in cancer datasets, modified transFIC scores outperformed original scores generated by other cancer specific tools. transFIC (v1.0, default parameters) was run using the gosmf database and applied to MutationAssessor predictions.

Structure-based algorithms

HotMAPS

Hotspot Missense mutation Areas in Protein Structures (HotMAPS) (Tokheim et al., 2016a) detects somatic mutation hotspot regions in 3D protein structures residing within a single protein chain or spanning protein chains (https://github.com/KarchinLab/HotMAPS;

v1.1.3). Protein structures were obtained from the Protein Data Bank (PDB) and homology models from the ModPipe human 2013 dataset (ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/), built with Modeler 9.11 (Pieper et al., 2011). Missense mutations were mapped to each protein structure or homology model using the MySQL database of Mutation position imaging toolbox (MuPIT) (Niknafs et al., 2013). The preferred biological assembly from MuPIT was used when multiple biological assemblies were available for a protein structure. HotMAPS calculates a p value for missense mutated residues containing a higher than expected density of missense mutations. Multiple hypothesis testing correction was performed using the Benjamini-Hochberg approach, and the significance threshold was set at a q-value of 0.01.

HotSpot3D

HotSpot3D (Niu et al., 2016) is a suite of algorithms (https://github.com/ding-lab/hotspot3d) that identifies spatial mutation clusters on 3D protein structures. For this manuscript, we used version 1.4.1. A pairwise distance measure is calculated for nearest-atoms/average-amino-acid on protein structure. Networks are then built by properly linking pairwise distances to corresponding mutations. Initialized by the distance matrix of the edges, clusters are constructed using the Floyd–Warshall shortest-paths algorithm to obtain the geodesics. We weighted this algorithm to bias centroid sections toward frequently mutated missense mutations. Finally, a closeness-centrality measure, or the sum of centralities over each mutation in a cluster, was used to describe features in the genes we identified here. For this study we used the following cutoffs: For intra-molecular clusters: 1) no linear amino-acid chain distance cutoff was enforced, 2) pairwise distances were calculated using the average amino-acid structure difference, 3) only mutation pairs with protein specific p values less than 0.05, and 4) the maximum network radius was 10 Angstroms. For inter-molecular clusters: 1) no linear amino-acid chain distance cutoff was enforced, 2) pairwise distances were calculated using the average amino-acid structure difference, 3) only mutation pairs with protein specific p values less than 0.05, and 4) the maximum network radius was 20 Angstroms.

3DHotSpots.org

The algorithm behind 3DHotspots.org identifies statistically significant clusters of missense cancer mutations in 3D structures (Gao et al., 2017). Missense mutations were mapped to 3D protein structures using G2S web services (https://g2s.genomenexus.org/) (March 2017). Only alignments with a sequence identity of 90% or above were included. The contact map of each structure chain was then calculated. Two residues with any pair of atoms within 5Å were considered in contact. A 3D cluster is defined by a central residue and at least one contact neighbor residue. A 3D cluster is identified as significantly mutated if its residues were more frequently mutated than expected by chance, as determined by a permutation-based test. Details of the methodology and the tool are available at https://github.com/knowledgesystems/mutationhotspots. Version 1.0.1 with default parameters was used in this analysis.

e-Driver3D

This algorithm identifies protein regions that are enriched in somatic missense mutations using a binomial test and assuming mutations are distributed randomly across the protein. The three-dimensional analysis is based on a library of protein interaction interfaces extracted from the Protein Data Bank30. The interaction interfaces are defined for each pair of protein chains in each PDB coordinates file as all the residues of a chain with a carbon atom within 5 Å of a carbon atom of the other chain. We used the interfaces described in https://github.com/eduardporta/e-Driver/interfaces_human_genome.txt.

Additional algorithms

DriverNET

DriverNet (Bashashati et al., 2012) is a package to predict functional important driver genes in cancer by integrating genome data (non-synonymous SNVs, indels, and copy number alteration) and transcriptome data (gene expression data). The different data types are integrated using an influence graph (Wu et al., 2010). We ran DriverNet (v1.6.0, numberOfRandomTests = 500, weight = FALSE, purturbGraph = FALSE, purturbData = TRUE) and genes with q-value of 0.05 were deemed significant.

OncoIMPACT

OncoIMPACT (Bertrand et al., 2015) is a model-driven approach to integrate omics profiles (genomics and transcriptomics) and provides patient-specific cancer driver gene predictions. It uses a gene interaction network to associate mutations (non-synonymous SNVs, indels and copy number alterations) with transcriptomic changes (Wu et al., 2010). We measured the transcriptomic change of each patient as the log2 fold change of the patient gene expression value with the cancer type median gene expression value. OncoIMPACT (v0.9.4) was run using default parameters. The top 50 predicted genes were used for the consensus gene list building.

MSIsensor

Written in C++, MSIsensor (version 0.2) is an algorithm that distinguishes microsatellite instable (MSI) tumors from microsatellite stable (MSS) samples based on tumor/normal sequence data (Niu et al., 2014). Homopolymer regions of 5 or more nucleotides in length are aggregated separately in tumor/normal pairs and compared using a χ^2 statistic. MSI-high was calculated as an MSI score $\geq$ 4. Parameters for running MSIsensor "msi" command are as follows: –l (minimal homopolymer size) = 1 and –q (minimal microsatellite size) = 1. These settings are not minimal number of repeats, but rather the minimal number of nucleotides to consider within the repeat.

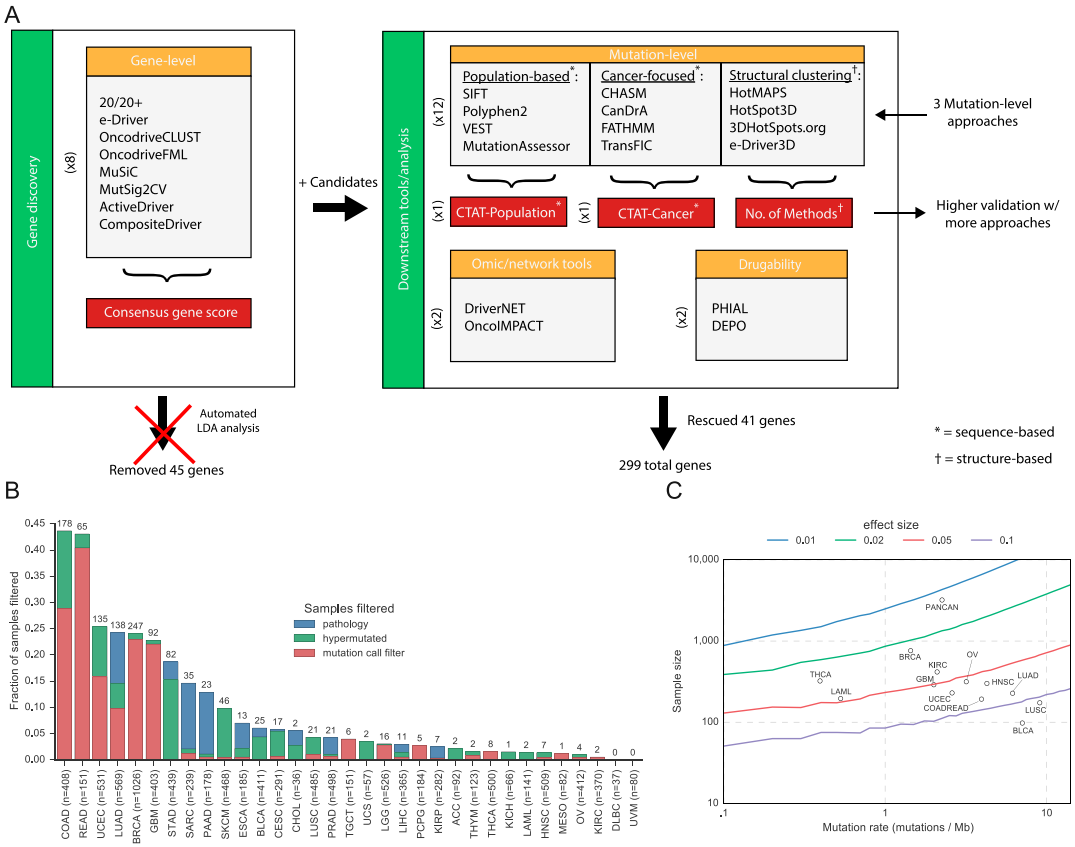
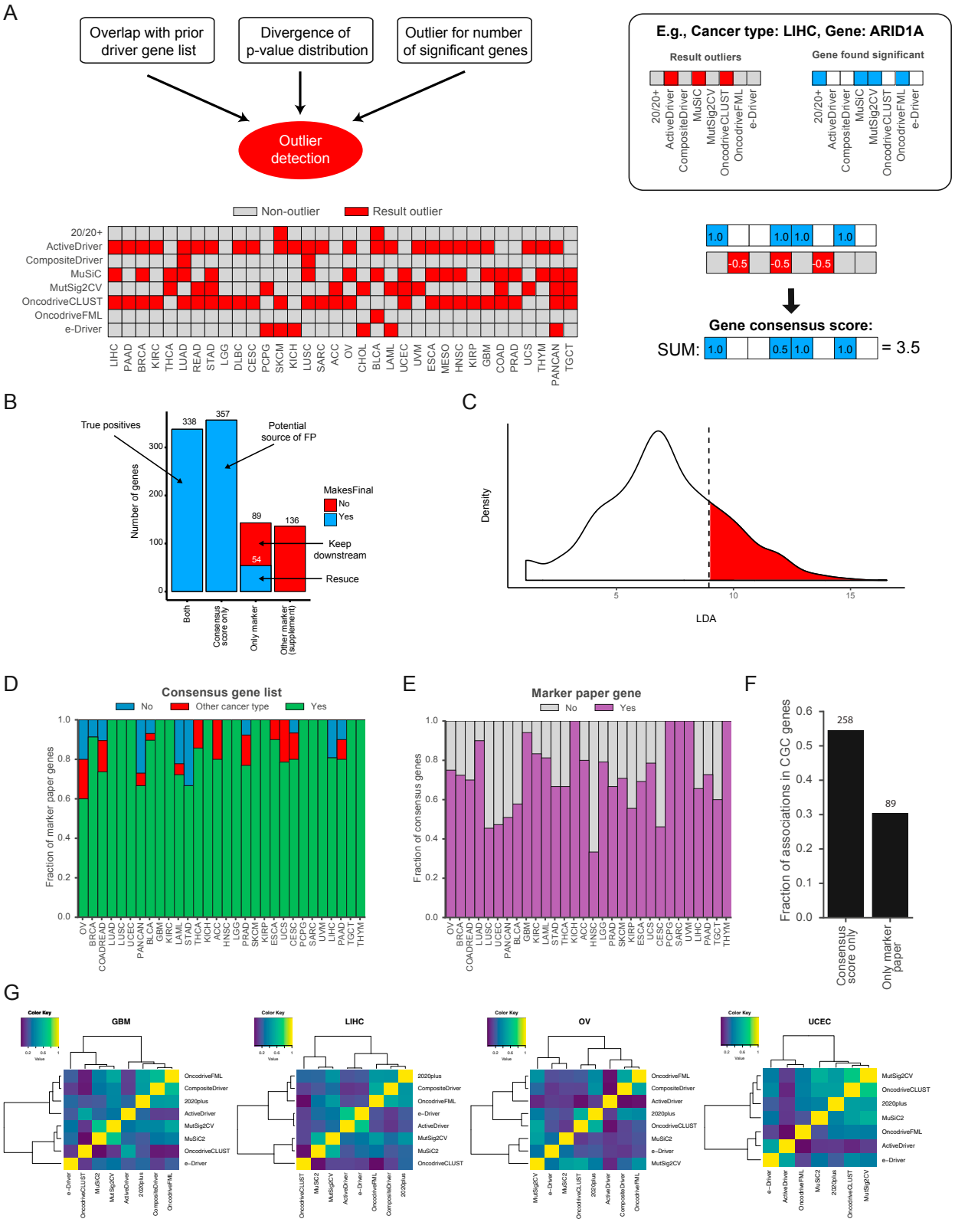


Figure S1. MAF Filtering and Power Comparison, Related to Figure 1

(A) Overall schema showing how we used the different algorithms and the input from the literature to identify our cancer driver gene consensus list and the driver mutations.

(B) Fraction of samples filtered through three quality assurance filters: a mutation call filter, hypermutated samples, and samples excluded by pathology review. Numbers above bars indicate the number of samples completely dropped. N refers to the total samples before filtering.

(C) Statistical power analysis for detection of driver genes at defined fraction of tumor samples above the background mutation rate (effect size). Circles indicate each of 12 cancer types or all cancer types together ("PANCAN") from the original TCGA analysis of 12 cancer types (PanCan-12) placed according to the study sample size and median background mutation rate across samples.



(legend on next page)

Figure S2. Consensus Gene Scores and SMG Filtering, Related to Figures 1 and 2

(A) Left, outlier detection was performed on a per analysis and method basis. Outliers were marked (red) based on the quasi-majority of three criteria: (1) low concordance with known cancer genes from Vogelstein et al. (lower than median); (2) high divergence of p value distribution from theoretical expectation (higher than median); and (3) abnormally high number of significant genes (> 1.5x the interquartile range above the third quartile). The first two criteria were assessed based on the other tools within a single analysis, while the third criterion was assessed based on the same tool's results over all the individual cancer types (excluding the PanCancer analysis). Right, example calculation of the gene consensus score for *ARID1A* in the cancer type LIHC. A result from an outlier is down weighted, receiving a weight of 0.5 instead of 1.0. The gene consensus score is the sum of weights for tools finding that gene as significant.

(B) Overlap of consensus gene list with prior TCGA marker papers.

(C) Likely false positives were detected with a high Linear Discriminant Analysis (LDA) score threshold representing 90% sensitivity for keeping associations found in Cancer Gene Census genes. LDA was trained to distinguish common false positives in exome sequencing from previous TCGA PanCancer marker papers. The LDA threshold was only applied to the potential source of false positive genes.

(D) Fraction of marker paper genes highlighted in the main text that were also found in our consensus gene list.

(E) Fraction of our consensus gene list found in previous TCGA marker papers.

(F) Fraction of associations found in the Cancer Gene Census (CGC) that were either found only in the consensus gene list or TCGA marker paper.

(G) Four heatmaps indicate the relationship between algorithms used in driver gene discovery for 4 cancer types GBM, LIHC, ovarian serous cystadenocarcinoma (OV), UCEC (left to right). Pairwise Pearson 2-tailed correlation coefficients were calculated from driver prediction p values generated by each tool and in each cancer type. Strength of the correlation coefficient (R) is displayed in colors ranging from yellow (strong) to blue (weak).

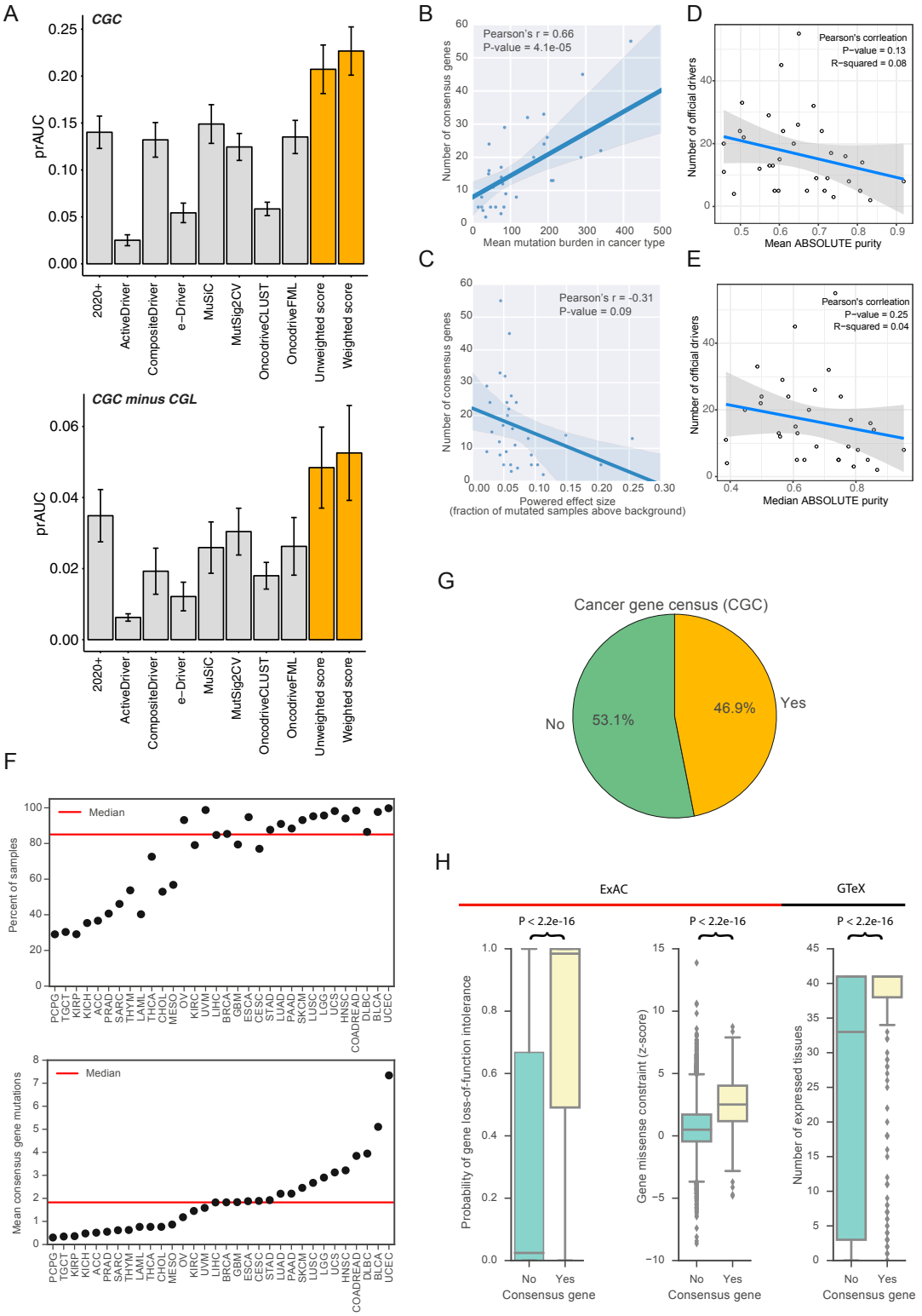


Figure S3. Characteristics of Consensus Genes, Related to Figure 2

(A) Predictive power of each individual driver gene detection method (in gray) and of the weighted and weighted scores (in orange). The predictive power was measured as prAUC, using all the genes in the Cancer Gene Census and a set that additionally excludes Cancer Genome Landscape genes used in outlier detection. Error bars, calculated by bootstrapping, indicate one standard deviation.

(legend continued on next page)

(B) The number of consensus genes in each cancer type positively correlated with the average mutation burden. Shaded area indicates 95% bootstrapped confidence interval.
(C) Given the variability in powered effect size (fraction of mutated samples above background with 90% power) in this study, there is a negative but not significant correlation with the number of consensus genes in each cancer type. COAD and rectum adenocarcinoma (READ) were excluded because analysis was performed separately, but the final consensus genes were merged.
(D) Pearson correlation between the number driver genes identified and median purity was calculated and plotted.
(E) Pearson correlation between the number driver genes identified and mean purity was calculated and plotted. Summary statistics for p value and r-squared value are reported in the top right corner of panels (D) and (E).
(F) Percent of samples containing a non-silent mutation stratified by cancer type. The red line indicates the median across cancer types (left) and average number of non-silent mutations in consensus genes per sample (right).
(G) Pie chart showing percent of consensus genes which are found in the Cancer Gene Census with annotations for small somatic mutations (missense, splice site, indel, and nonsense).
(H) Consensus genes showed a higher probability for loss-of-function intolerance and missense mutation constraint of germline mutations based on ExAC (Lek et al., 2016) and were expressed (RPKM > 1) in a wider number of tissues from GTeX (version 6) (Consortium, 2015). Given the high correlation of gene expression in the 11 brain regions assessed from GTEx, we took the median of multiple brain tissues, as done in Lek et al. (2016).
Boxplots indicate median MSI score with 25th and 75th percentile hinges and whiskers that extend to 1.5*IQR.

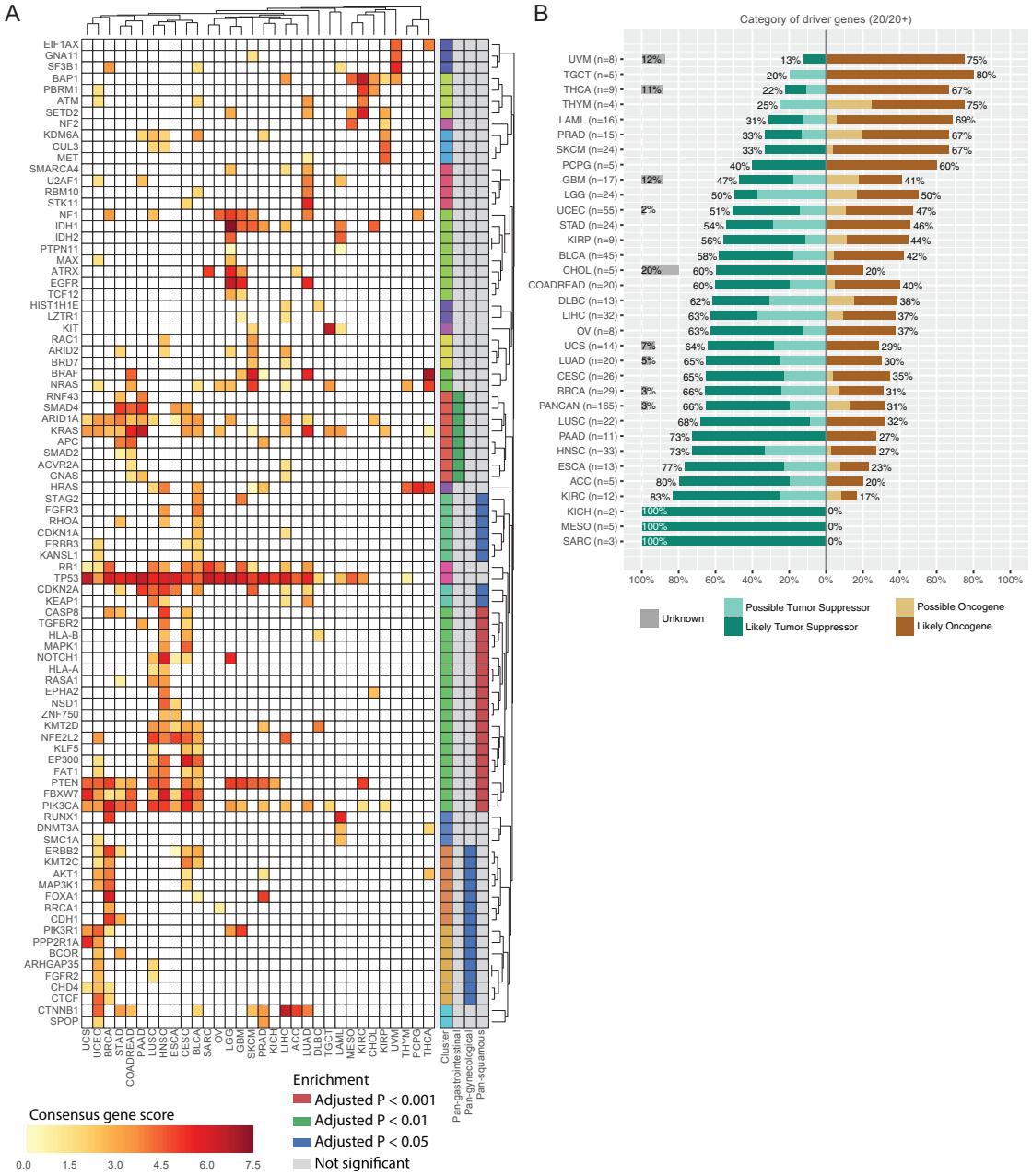


Figure S4. Molecular Properties of Cancer Driver Genes, Related to Figure 2

(A) Hierarchical clustering of the gene consensus scores for genes that were found in more than one cancer type. The correlation distance metric and average linkage was used. Each gene cluster was tested for enrichment in three groups of cancer types, in order: Pan-squamous (BLCA, CESC, LUSC, HNSC, and ESCA), Pan-gynecological (UCEC, UCS, CESC, OV, and BRCA), and Pan-gastrointestinal (STAD, COADREAD, ESCA, and PAAD). Significant gene clusters are based on a permutation test assessing the total gene consensus score (10,000 iterations) and are progressively colored gray (not significant), blue (Adjusted $p < 0.05$), green (Adjusted $p < 0.01$), and red (Adjusted $p < 0.001$). P values were multiple test corrected across all genes using the Benjamini-Hochberg FDR method. Gene clusters are shown as distinct colors in the first column of the row annotation bar. Clusters of genes were defined based on cutting the dendrogram at a depth chosen by manual inspection.

(B) Percentage of consensus genes predicted as either oncogene (brown), tumor suppressor gene (green), or unknown (gray) by the 20/20+ algorithm, an improved version of the 20/20 rule (Vogelstein et al., 2013). The 20/20+ algorithm uses a supervised-learning approach (random forests) and bases predictions on the mutational patterns observed within a gene. "Likely" and "Possible" statuses were determined at a threshold of 0.05 for q-value (Benjamini-Hochberg method) and p value, respectively. Consensus genes were designated as "Unknown" if they did not meet these thresholds. N represents the number of significant genes in each cancer type.

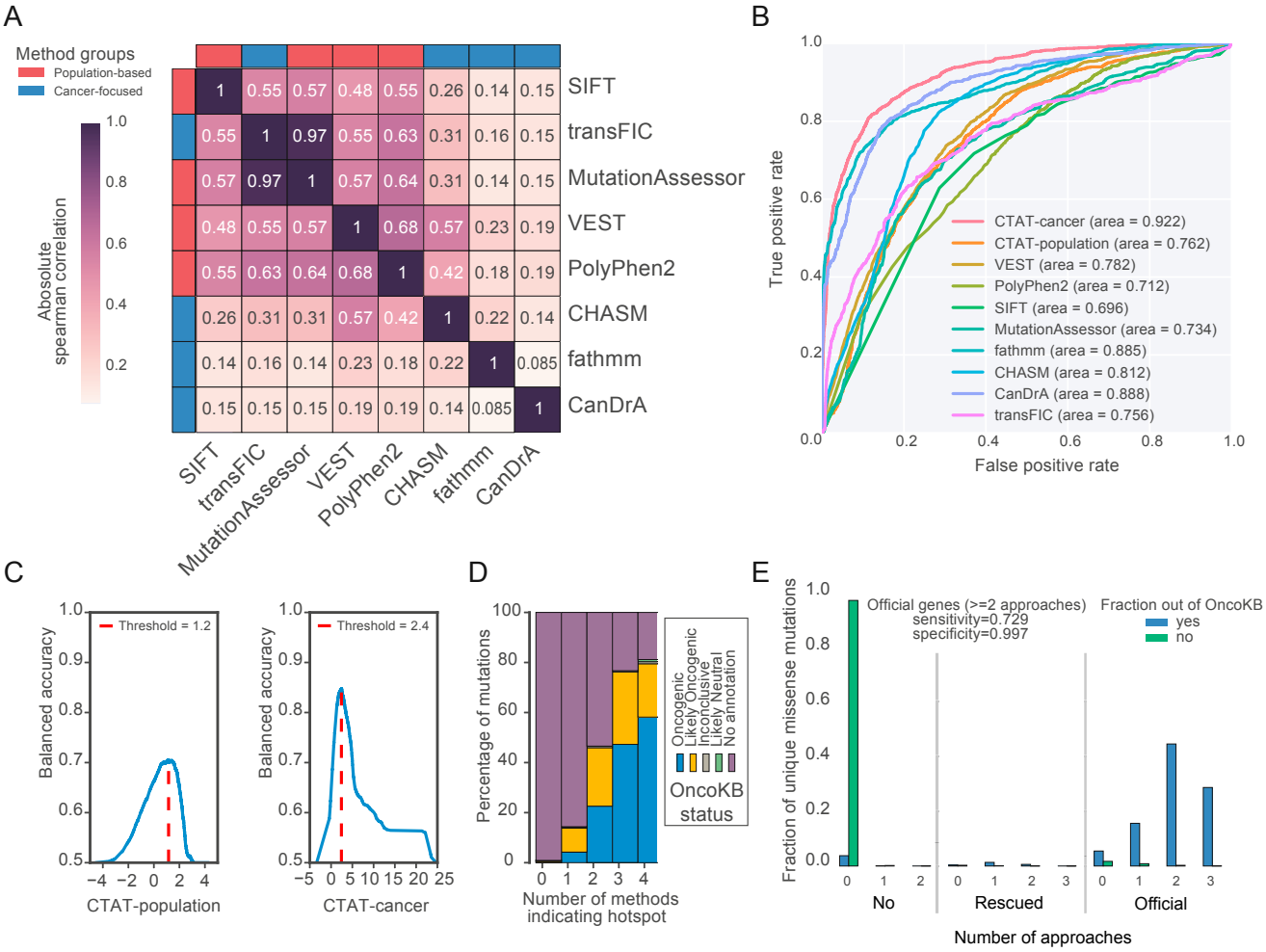


Figure S5. Characteristics and Implementation of Driver Mutation Analysis, Related to Figure 3

(A–E) Eight sequence-based tools scored missense mutations to prioritize likely driver mutations over passenger mutations. (A) The absolute Spearman correlation between different sequence-based tools is shown, where tools are arranged in order by hierarchical clustering using a Euclidean distance metric. Tools that distinguish pathogenic missense mutations from neutral polymorphisms are labeled "population-based" (red), while tools focused on distinguishing passenger somatic missense mutations from cancer drivers are colored blue. A consensus score (named (C)ombined (T)ool (A)djusted (T)otal—CTAT) for the "population-based" tools and "cancer-focused" tools was developed. (B) Receiver operator curves (ROC) compared CTAT-population and CTAT-cancer scores to 8 sequence-based tools. We used OncoKB annotation of "Oncogenic" and "Likely Oncogenic" versus all other missense mutations in consensus genes as a benchmark. Area under the curve (AUC) calculations are presented for each of the individual 8 sequence-based tools and two sequence-based consensus approaches. (C) We determined the optimal score threshold based on balanced accuracy (red dashed line) for CTAT-population (left) and CTAT-cancer (right). Missense mutation hotspots were also detected based on four structural tools that utilize three-dimensional protein structures. (D) The percentage of missense mutations labeled as "Oncogenic" or "Likely Oncogenic" in OncoKB steadily increased with greater number of structural tools, indicating an amino acid residue was a hotspot. (E) Fraction of unique missense mutations in this study either in or not in the OncoKB, which is stratified by the number of mutation-level approaches in agreement (Population-based, Cancer-focused, and Structural clustering). The gray line separates where mutations were found in our consensus gene list (not found, manually rescued, or official).

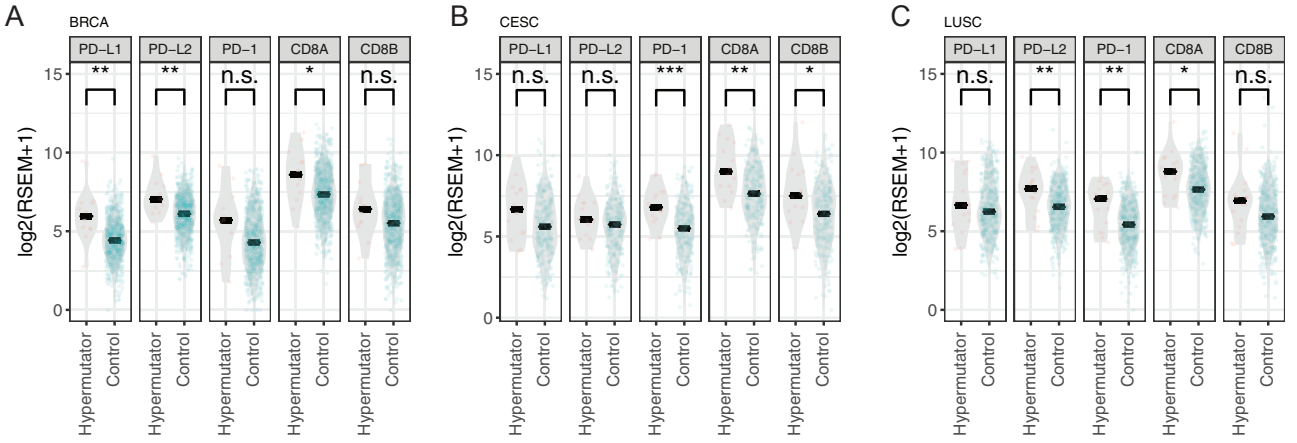


Figure S6. Relationship between Hypermutated Samples and Immune System Markers, Related to Figure 5

(A–C) RNA-Seq abundance of different immune biomarkers for MSI phenotypes defined by MSIsensor. Stars indicate significance levels from a two-sided t test to calculate p values (* < 0.05, ** < 0.01, *** < 0.001) for BRCA (A), CESC (B) and LUSC (C).

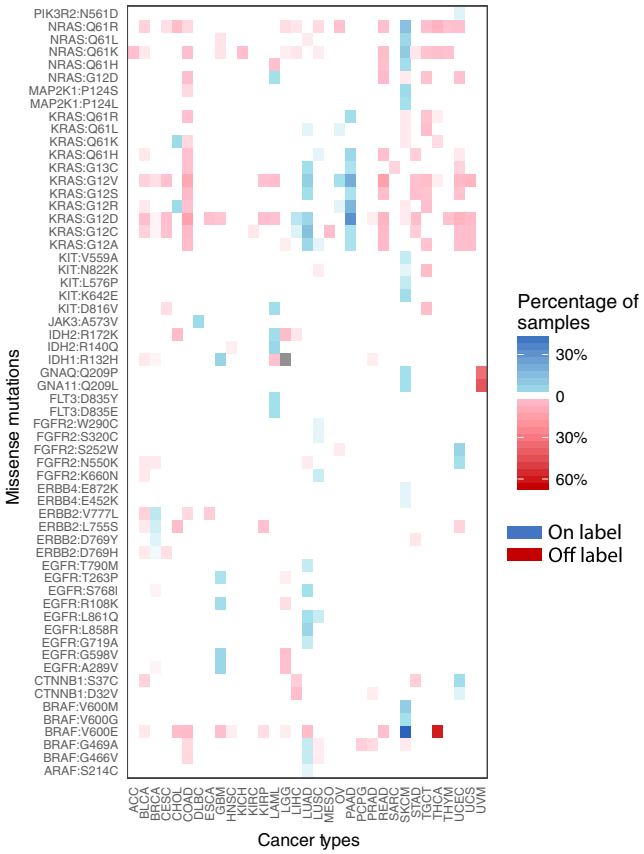


Figure S7. On-Label/Off-Label Calculations for Druggable Mutations in Cancer, Related to Figure 6

Missense mutations from consensus gene calling were annotated using the DEPO database. Here the proportion of samples in a cancer type (x axis) with on-label (blue) or off-label (red) therapeutic options are provided for specific missense mutations (y axis). Briefly, on-label refers to mutation specific treatments that have been clinically tested for a given cancer type. Off-label designations refer to potential drug therapies not heavily tested for said cancer types. Only druggable mutations present in the largest number of tumor samples across the TCGA cohort are displayed.