**Supplementary Note 1: List size for driver gene prediction**

Driver prediction methods are commonly used to order genes and facilitate the selection of candidate genes in a cancer type for additional analyses and experiments. We evaluated how many driver genes would be expected in a cancer type, given the overlap in mutations from tumors of a cancer type, and assuming between 1 and 5 driver events per patient. This number can then be estimated by solving a Set Cover problem on the bipartite graph defined by the set *G* of mutated genes, the set *P* of patients and edge representing genes mutated in a patient. For a given number *x* of driver events in a patient, we estimated the minimal number of driver genes in the cohort by computing the minimal number of nodes of *G* required to cover each node of *P* by *x* edges. This analysis suggested that the median number of driver genes in a cohort is expected to be between 10 and 50 (**Supplementary Figure 4**), motivating the choice of our thresholds.

**Supplementary Note 2: Machine learning approaches for consensus generation**

We explored the use of machine learning approaches, such as support vector machines (SVM) and random forests (RF) to see if they could improve over the simple rank aggregation idea used in ConsensusDriver. For each tumor type, we trained SVM/RF classifiers on the remaining 14 data sets with predictions from all 18 methods as features (see **Supplementary Methods**). Genes were ranked based on confidence scores from the classifiers to compute combined scores for the top 10 and top 50 predictions for each method (**Supplementary Figure 14**). We noted that our best SVM or RF predictors did not improve over the best individual method or ConsensusDriver. Further experimentation/optimization and more complete gold-standards may be needed to successfully use these approaches for consensus predictions. On the other hand, an alternate rank aggregation approach (Robust Rank Aggregation) readily recapitulated ConsensusDriver’s performance (**Supplementary Figure 14**), though it suffers from the drawback of being much more computationally expensive.

**Supplementary Methods**

**Other consensus approaches**

For each cancer type, Random Forest and SVM classifiers were given two inputs: a matrix $S$ containing training samples consisting of gene prediction scores for each method on each cancer types and a vector $C$ of classifications that indicates for each gene entry if it is a driver or a passenger based on gold-standard datasets. Genes were labeled as drivers if they were annotated as a driver for the associated tumor-type in the NCG4 high confidence [1] or CGC [2] driver lists. Genes were labeled as passengers if they were not in the CGC or the NCG4 (high confidence or putative) driver lists. For the gene prediction scores in $S$ we used $-log10(p-value)$ for methods providing FDR values, the score reported by the methods for FI and FIC methods as well as OncoIMPACT$,$ and 0 or 1 for passengers and drivers respectively, if the method did not return a score or FDR value (NetBox and HotNet2). The scores for each method were then scaled across all cancer types to a range of 0-1 (except for OncoIMPACT which was scaled separately for each cancer type). On average there were 51 driver and 308 passenger genes per cancer type. To account for the unbalanced dataset, weights proportional to driver/passenger frequencies were introduced into the regularization parameter of the SVM model. Hyper-parameters of the SVM and the random forest were learned using a stratified 5-fold cross validation based on a grid search. For the SVM model, multiple kernel types were analyzed (linear, polynomial of varying degrees and radial basis function). Results based on polynomial kernel (degree 2) are reported here as they provided the best results.

We evaluated the performance of the machine learning classifiers using a leave-one-out validation approach: we trained the model on 14 cancer types and evaluated it on the remaining one. SVM and Random Forest classifiers were developed using the Python scikit-learn package [3].

The robust rank aggregation approach was evaluated by providing ranked lists from each method as input to the R package RobustRankAggreg [4].

**Processing of expression data**

For the OV, GBM and BRCA datasets, array-based normalized genes expression values were used, while normalized RNA-seq read count values were used for all other cancer types (normalized with DESeq2 [5]). For each cancer type, a set of representative normals was selected by hierarchical clustering (Euclidean distance) of gene expression profiles and manual identification of the cluster with expression profile that was most divergent from tumor profiles (see Supplementary File 2 for list of selected normals and tumors that were excluded for clustering with them). Differential expression analysis (tumor *vs* normal, as required by OncoIMPACT [6] and DawnRank [7]) was performed using t-test (for microarray data) or DESeq2 (for RNA-seq data) and genes selected based on FDR cutoff of 0.05 (with average read count > 200 to exclude lowly expressed genes).

**Parameters used for different driver prediction methods**

A majority of the methods could be run directly using either GDAC data or variant annotation from annovar [8] (v20150322, UCSC, REFSEQ and ENSEMBL gene annotations), and did not require any data massaging and/or parameter choices. DawnRank[[1]](#footnote-1) [7] (v1.2, mu=20, maxit=100, epsilon=1e-04), DriverNet [9] (v1.6.0, numberOfRandomTests=500, weight=FALSE, purturbGraph=FALSE, purturbData=TRUE, prediction with FDR $\leq $ 0.05), fathmm [10] (v2.2, -w Cancer, variant annotated as CANCER); MutSigCV [11] (v1.4, FDR $\leq $ 0.05), Oncodrive-CIS [12] (v1.1.0, prediction with FDR $\leq $ 0.05), OncodriveCLUST [13] (v0.5.0, -m 3, prediction with FDR $\leq $ 0.05), OncoIMPACT [6] (v0.9.3, prediction from stringent mode) and S2N [14] (viaCNAmet1.2 [15], favorSynergetic=TRUE, strictChecks=FALSE, perms=1000, na.limit=0.1, gainData=TRUE, strictLim=0.05) were run using their default parameters.

Predictions for FI methods were obtained using annovar [16] (v20150322, UCSC genes, with LJB26 database). To identify predicted functional mutations the following thresholds were used: SIFT [17] score < 0.05), Polyphen2 [18] Hvar variant with score < 0.446, MutationAssessor [19] variant annotated as H or M, MutationTaster [20] variant annotated as A or D predictions. CHASM [21] (via CRAVAT 3.2, gene with FDR < 0.2) was run using classifiers specific to the cancer type being analyzed. transFIC [22] (v1.0, using the gosmf database) was applied to MutationAssessor predictions as on the IntOGen web site (https://www.intogen.org/). CHASM [21] provides gene-specific predictions (file Gene\_Level\_Analysis.Result.tsv, where the gene score corresponds to the best score observed on a point mutation in that gene). For other FIC and FI methods the average score per gene (across patients, selecting the best score per patient) was used to order genes.

ActiveDriver [23] (v0.010, gene with FDR < 0.2) was run using the database ActiveDriver\_HG38 that contains phosphorylation, ubiquitination, acetylation, and methylation sites. OncodriveFM [24] (v0.6.0, -e median, gene with FDR $\leq $ 0.05) was run using predictions from SIFT, Polyphen2 (HVAR) and MutationAssessor. As recommended, the scores for synonymous variants were fixed to the worst score of each method (1, 0, and -2 respectively) and the scores for frameshift and nonsense variants were fixed to the best score for each method (0, 1 and 3.5 respectively). NetBox (v1.0, p\_value\_threshold=0.05, num\_global\_trials=0, num\_local\_trials=0, shortest\_path\_threshold=2) analysis was performed using the top 500 most frequently mutated genes in the NetBox network, similar to the original analysis [25]. HotNet2 [26] (v1.0.1, delta\_permutations=100, significance\_permutations=1000) was run using the iRefIndex network [27]. The filtering criteria used in the original manuscript are not implemented in HotNet2 [26] but similar critera were used here, filtering all genes with high mutation frequency (>0.02) that are not identified as significant by MutSigCV. NetBox and HotNet2 predictions were ranked according to mutation frequencies.

DriverDB predictions were obtained from <http://driverdb.tms.cmu.edu.tw/driverdbv2/cancer.php> and are based on the output of the following methods: ActiveDriver [23], Netbox [25], OncodriveFM [24], MutSigCV [11], Dendrix [28] , MDPFinder [29], Simon [30] and MEMo [31]. Genes predicted by 2 or more methods were selected and ranked using the order provided on the DriverDB website. MutSig predictions were obtained from Lawrence et al [8] supplementary materials.

**References**

[1] An O, Pendino V, D’Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database (Oxford) 2014;2014:bau015.

[2] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.

[3] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. … Mach Learn … 2012;12:2825–30.

[4] Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 2012;28:573–80. doi:10.1093/bioinformatics/btr709.

[5] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550. doi:10.1186/s13059-014-0550-8.

[6] Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BKH, Sia YY, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res 2015;43:e44.

[7] Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med 2014;6:56.

[8] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164–e164. doi:10.1093/nar/gkq603.

[9] Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol 2012;13:R124.

[10] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 2013;34:57–65.

[11] Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.

[12] Tamborero D, Lopez-Bigas N, Gonzalez-Perez A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. PLoS One 2013;8:e55489.

[13] Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 2013;29:2238–44.

[14] Hautaniemi S, Ringnér M, Kauraniemi P, Autio R, Edgren H, Yli-Harja O, et al. A strategy for identifying putative causes of gene expression variation in human cancers. J Franklin Inst 2004;341:77–88. doi:10.1016/j.jfranklin.2003.12.005.

[15] Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 2011;27:887–8.

[16] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164–e164. doi:10.1093/nar/gkq603.

[17] Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4. doi:10.1093/nar/gkg509.

[18] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.

[19] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118.

[20] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014;11:361–2. doi:10.1038/nmeth.2890.

[21] Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009;69:6660–7.

[22] Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med 2012;4:89. doi:10.1186/gm390.

[23] Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol 2013;9:637.

[24] Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res 2012;40:e169.

[25] Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One 2010;5:e8918.

[26] Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2014;47:106–14.

[27] Razick S, Magklaras G, Donaldson IM. iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics 2008;9:405. doi:10.1186/1471-2105-9-405.

[28] Leiserson MDM, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol 2013;9:e1003054.

[29] Zhao J, Zhang S, Wu L-Y, Zhang X-S. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 2012;28:2940–7. doi:10.1093/bioinformatics/bts564.

[30] Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics 2011;27:175–81. doi:10.1093/bioinformatics/btq630.

[31] Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res 2012;22:398–406. doi:10.1101/gr.125567.111.

1. We stopped DawnRank processing of the BRCA data set after 15 days. Average score and median score were computed over the remaining 14 cancer types for this method. [↑](#footnote-ref-1)