bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Systematic integration of protein affecting mutations, gene fusions, and copy number alterations into a comprehensive somatic mutational profile

Shawn S Striker [1], Sierra F Wilferd [1], Erika M Lewis [1], Samantha A O'Connor [1], Christopher L Plaisier [1]*

[1] School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ

*To whom correspondence should be addressed: Christopher Plaisier, E-mail: plaisier@asu.edu, Phone: (480) 965-6832, Address: P.O. Box 879709, Tempe, AZ 87287-9709

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

A gene can be mutated across a tumor cohort by protein affecting mutations (PAMs), gene fusions, or copy number alterations (CNAs). These mutations can have a similar phenotypic effect (i.e., allelic heterogeneity) and should be integrated into a unified gene mutation profile. We provide OncoMerge as a somatic mutation integration platform that tames allelic heterogeneity, discovers causal mutations, integrates binary PAM and fusion with quantitative CNA data types, and overcomes known obstacles in cancer genetics. OncoMerge was applied to the 9,584 patient tumors from 32 cancers profiled by the TCGA Pan-Cancer Atlas to validate the novel integration methods. Integration increased the number and frequency of somatically mutated genes and improved the prediction of the somatic mutation role as either activating or loss of function. Using OncoMerge integrated somatic mutations boosts the power to infer active gene regulatory networks that increase the connectedness of the networks and incorporate more somatic mutations and regulators associated with cancer biology. We extracted transcription factor (TF) regulatory networks and found that they were enriched with feedback and feed-forward loop network motifs. Subsequent, signed network motif analysis demonstrated that coherent switch-like feedback motifs and delay-inducing feed-forward loops were the only enriched configurations. This enrichment pattern suggests that evolution in general or in the tumor microenvironment is selecting for these coherent functional configurations. The OncoMerge integrated somatic mutations provide a more comprehensive platform for studies linking somatic mutations to downstream cancer phenotypes and will lead to novel biological insights in clinical samples.

Keywords: somatic mutations, allelic heterogeneity, cancer, oncogene, tumor suppressor gene

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The selective pressures driving the accrual of somatic mutations that affect cancer phenotypes are shared across cancers. This phenomenon leads to genes being somatically mutated across multiple cancers, e.g., oncogenes and tumor suppressors. The three main types of somatic mutations that modify the function of a gene or render it non-functional are: 1) protein affecting mutations (PAMs), 2) gene fusions, and 3) copy number alterations (CNAs). A PAM is a point mutation, short insertion, or short deletion inside a gene's coding region or splice sites [1]. Gene fusions occur when genomic rearrangements join two genes into a novel chimeric gene or place a promoter in front of a new gene, causing misexpression [2]. Finally, CNAs occur frequently in tumors where whole chromosomes, chromosomal arms, or localized genomic segments are duplicated or deleted [3,4]. Somatic mutation via PAM, gene fusion, or CNA can have similar effects on cancer phenotypes, i.e., allelic heterogeneity. This interchangeability and the erratic circumstances that produce somatic mutations lead to the mixture of mutation types observed in large cohorts of patient tumors [1].

Describing how somatic mutations in a gene impact cancer phenotypes requires integrating the information from all three mutation types. Currently, most studies linking somatic mutations to cancer phenotypes focus on one mutation type. This leads to missing associations for mutations primarily found in another type and reduced power to detect associations for mutations with high allelic heterogeneity that span the mutation types. Thus, a current obstacle facing those studying the downstream effects of somatic mutations is the lack of an established method for integrating PAMs, gene fusions, and CNAs into a comprehensive gene mutation profile. The lack of integration methods is due to several complicating factors. Firstly, the allelic heterogeneity observed in and between tumors means that different mutations in the same gene can be equivalently oncogenic. Second, it is challenging to discern driver (causal) from passenger (non-causal) somatic mutations. Third, an algorithm must be able to systematically integrate the binary PAM and gene fusion (mutated or not) with the quantitative copy number from CNAs. Lastly, some tumors have drastically higher somatic mutation rates than other tumors (e.g., microsatellite instability[5] and hypermutation[6]). These higher mutation rates confound any frequency-based integration approach and drive the discovery of spurious somatic mutations. We developed OncoMerge to fill the somatic mutation integration niche by providing an algorithm that systematically overcomes these obstacles to generate an integrated gene mutation profile. The input for the OncoMerge algorithm is the output from state-of-the-art methods for detecting PAMs (MC3[1] and MutSig2CV[7]), transcript fusions (PRADA[2,8]), and CNAs (GISTIC2.0[9]). The integrated mutation profiles will give more power to detect associations with cancer phenotypes by capturing all tumors with a somatically mutated gene leading to a more comprehensive understanding of how genetic alterations drive cancer phenotypes.

The tremendous amount of cancer genome sequencing data generated in the last ten years has enabled efforts to discover and catalog somatic mutations across many cancers [1,10]. Many algorithms have been developed to discern which somatic mutations are drivers and how the mutations affect genes [6,7,11–15]. The impact of somatic mutations can be classified as activating (Act) gene function (typically found in oncogenes), or loss of function (LoF) (typically found in tumor suppressor genes)[16]. It has also been demonstrated that the systematic integration of PAM and CNA somatic mutations for a gene improves the ability to determine Act or LoF status[16]. These foundational studies have created a platform to develop an algorithm that systematically integrates the three somatic mutation types.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The systematic integration of somatic mutations requires choosing a gene-level model that determines how the data for the three somatic mutation types will be integrated, which in OncoMerge is called the somatic mutation role. We determine the somatic mutation role by employing rules similar to those used in OncodriveROLE[16] (**Figure 1**). The possible somatic mutation roles in OncoMerge are PAM, Fusion, CNA amplification (CNAamp), CNA deletion (CNAdel), Act, or LoF. The PAM, Fusion, CNAamp, and CNAdel somatic mutation roles use the somatic mutation profile of the role in the integrated mutation matrix. The Act and LoF are integrated mutation roles that harness allelic heterogeneity. Allelic heterogeneity is especially prevalent in tumor suppressor genes, where mutations at many positions in a gene can impede its function to prevent cancer phenotypes[4]. Allelic heterogeneity is less prevalent for oncogenes where a small number of specific gain of function alleles are needed to drive cancer phenotypes[4]. Genes underlying CNAs can add another layer of information as tumor suppressors are often deleted, which has an equivalent oncogenic effect as missense or truncating PAMs. The LoF role is designated when PAMs, Fusions, and CNAdels are integrated. Oncogenes are often amplified as this typically leads to overexpression of the underlying genes, which has a similar positive effect on gene function as a gain of function PAM. The Act role is designated when PAMs, Fusions, and CNAamps are integrated. Systematic determination of the somatic gene role and application of the rules laid out above will be used to integrate the three mutation types into a comprehensive somatic mutation profile.

The algorithms developed to discern somatic mutation drivers for cancers provide a set of gold standard mutations with gene roles that can be used to assess the performance of the new OncoMerege algorithm. The gold standards are classified by whether the somatic mutation of a gene was cancer-specific or not. The TCGA consensus[6] and Cancer Gene Census (CGC) from COSMIC[15] were used to develop gold standards with cancer-specific somatically mutated gene roles. The TCGA consensus is a list of driver genes identified from the TCGA Pan-Cancer Atlas labeled with somatic mutation role (oncogene or tumor suppressor) and cancer type. The CGC from COSMIC is an expert-curated database of human cancer driver genes labeled with somatic mutation role (oncogene and tumor suppressor) and cancer type. The 20/20 rule[4], OncodriveROLE[16], and Tokheim ensemble[14] were used to develop gold standards with somatically mutated gene roles. The 20/20 rule defines oncogenes by requiring >20% of mutations in recurrent positions and tumor suppressors as >20% of recorded mutations are inactivating (missense or truncating)[4]. OncodriveROLE is a machine learning algorithm that classifies genes according to their role (Act or LoF) based on well-curated genomic features[16]. The Tokheim ensemble is an ensemble-based method that integrates MutSigCV, 20/20+, and TUSON methods for predicting gene roles (oncogene and tumor suppressor)[14]. Comparisons of somatic mutation role between OncoMerge and the gold standards were facilitated by converting oncogenes to Act and tumor suppressors to LoF. Finally, a combined gene role agnostic gold standard was developed based on a union of all somatic mutations from all five gold standards. These gold standards were used to assess the utility of filters and the quality of the OncoMerege integrated somatic mutation matrices through their ability to recall somatic mutations with the appropriate gene role.

A primary goal of OncoMerge is to construct a comprehensive somatic mutation profile that will increase the power to identify how mutations modulate cancer phenotypes. Previously, we have used the Systems Genetics Network AnaLysis (SYGNAL) pipeline[17] to build causal and mechanistic gene regulatory networks (GRNs) for 31 cancers from the TCGA Pan-Cancer Atlas[18]. Using SYGNAL, we link somatic mutations through the GRN to the hallmarks of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cancer[19–21], thereby linking somatic mutations to cancer phenotypes. These SYGNAL GRNs describe how somatic mutations influence transcription factor (TF) or miRNA expression, which modulates the expression of downstream genes. In SYGNAL, somatic mutations are used as input for the Network Edge Orienting (NEO) portion of the pipeline that infers causal flows of information (somatic mutation → TF or miRNA regulator → bicluster of co-regulated genes). Thus, we use the OncoMerge integrated somatic mutation matrices in SYGNAL GRN inference to demonstrate the increased power to identify how mutations modulate cancer phenotypes.

TFs are a significant factor in regulating gene expression in a cell, and interactions between TFs could be used to explain much of the overall transcriptional state of a cell. Neph et al., 2012 constructed a human TF gene regulatory network by integrating genome-wide digital genomic footprinting with DNA recognition motifs across 41 cell types[22]. The network architecture of three-node network motifs was investigated and shown to have a pattern similar to other biologically derived networks[23–25]. Because these TF regulatory networks were generated based on DNA binding alone, they are not an active representation of the effect on transcript levels but static DNA binding maps. On the other hand, SYGNAL GRNs are trained using coexpression as an integral element of network construction. Therefore, SYGNAL GRNs can be considered active because transcriptional effects support regulatory interactions. We compare and contrast the underlying architecture of active TF regulatory networks from SYGNAL relative to static TF regulatory networks from DNA binding maps.

As proof of principle, we apply OncoMerge to the multi-omic characterization of 32 cancers by the TCGA PanCancer Atlas to develop filters and demonstrate a meaningful benefit for downstream analyses. We demonstrate the power of using an integrated mutation matrix in downstream analysis by re-analyzing the causal relationships for pan-cancer SYGNAL networks[18]. We constructed transcription factor (TF) regulatory networks[22] and generated triad significance profiles (TSPs)[24] to investigate the underlying network architecture[23–25]. We provide the complete OncoMerge code, comprehensive mutation matrices for 32 TCGA cancers, regulatory networks for 31 cancers, and TF regulatory network architecture for 25 cancers. These studies demonstrate that OncoMege efficiently integrates PAMs, fusions, and CNAs into a comprehensive mutational profile that strengthens downstream analyses linking somatic mutations to cancer phenotypes.

Methods

Clinical and molecular data from TCGA

These studies used standardized, normalized, batch corrected, and platform-corrected multi-omics data generated by the Pan-Cancer Atlas consortium for 11,080 participant tumors[18]. Complete multi-omic profiles were available for 9,584 patient tumors. TCGA aliquot barcodes flagged as "do not use" or excluded by pathology review from the Pan-Cancer Atlas Consortium were removed from the study. The overall survival (OS, OS.time) data used were obtained from Liu et al. 2018[26].

- Somatic protein affecting mutations (PAMs) in TCGA – Somatic PAMs were identified by the Multi-Center Mutation Calling in Multiple Cancer (MC3) project[1] and were downloaded from the ISB Cancer Gateway in the Cloud (ISB-CGC; https://isb-cgc.appspot.com/). PAMs were required to have a FILTER value of either: PASS, wga, or native_wga_mix. In addition, all PAMs needed to be protein-coding by requiring that Variant_Classification had one of the following values: Frame_Shift_Del,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Missense_Mutation, Nonsense_Mutation, Nonstop_Mutation, Splice_Site, or Translation_Start_Site. Additionally, mutation calls were required to be made by two or more mutation callers (NCALLERS > 1). When both normal tissue and blood were available, the blood was used as the germline reference.

- Statistical significance of PAMs in TCGA – The likelihood that a gene is somatically mutated by chance alone was determined using MutSig2CV[11] and downloaded for each cancer from the Broad GDAC FIREHOSE (https://gdac.broadinstitute.org/). Genes with a MutSig2CV False Discovery Rate (FDR) corrected p-value (q-value) less than or equal to 0.1 were considered significantly mutated[11].
- Somatic transcript fusions in TCGA – The TumorFusions portal[2] provides a pan-cancer analysis of tumor transcript fusions in the TCGA using the PRADA algorithm[8].
- Somatic copy number alterations (CNAs) in TCGA – Genomic regions that were significantly amplified or deleted were identified using Genomic Identification of Significant Targets in Cancer (GISTIC2.0)[9] and downloaded for each cancer from the Broad GDAC FIREHOSE.

Somatic mutation data import and preprocessing

An essential first step in OncoMerge is loading up and binarizing the somatic mutation data. The somatic mutation data comprised of four primary matrices: 1) PAMs, 2) fusions, 3) CNA amplifications (CNAamps), and 4) CNA deletions (CNAdels) (**Figure 1**). In addition, two derivative matrices Act and LoF are created by merging the PAM with the CNAamps or CNAdels matrices, respectively (**Figure 1**). All files are formatted as comma-separated values (CSV) files with genes as rows and patients as columns unless otherwise noted.

- PAM matrix - The matrix values are [0 or 1]: zero indicates the gene is not mutated in a patient tumor, and one indicates the gene is mutated in a patient tumor.
- Fusion matrix - The matrix values are [0 or 1]: zero indicates no gene fusion in a patient tumor, and one indicates the gene fused to another genomic locus in a patient tumor.
- CNAamp and CNAdel matrices – The all_thresholded_by_genes.csv GISTIC output file is used to populate the CNAamp and CNAdel matrices. The all_thresholeded_by_genes matrix values range from -2 and have no positive bound, and the values indicate the copy number relative to the background. A cutoff of greater than or equal to 2 was used to identify deep amplifications and less than or equal to -2 for deep deletions. Only deep amplifications or deletions were included in these studies due to heterogeneity of cell types and tumor biopsy purity. Oncomerge allows this threshold to be modified through a command line parameter ('-gt' or '--gistic-threshold').
    - CNAamp matrix – The matrix values are [0 or 1]: zero indicates a gene is not amplified in a patient tumor, and one indicates the gene is amplified in a patient tumor.
    - CNAdel matrix – The matrix values are [0 or 1]: zero indicates a gene is not deleted in a patient tumor, and one indicates a gene is deleted in a patient tumor.
- Act matrix – The Act matrix is the bitwise OR combination of the PAM, Fusion, and CNAamp matrices. The Act matrix has genes as rows and patients as columns. The matrix values are [0 or 1]: zero indicates the gene is not mutated or amplified in a patient tumor, and one indicates the gene is either mutated, fused, amplified, or some combination in a patient tumor.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- LoF matrix – The LoF matrix is the bitwise OR combination of the PAM, Fusion, and CNAdel matrices. The LoF matrix has genes as rows and patients as columns. The matrix values are [0 or 1]: zero indicates the gene is not mutated or deleted in a patient tumor, and one indicates the gene is either mutated, fused, deleted, or some combination in a patient tumor.

Seeding OncoMerge with putative somatic mutations

OncoMerge focuses on likely causal somatic mutations by considering only somatic mutations that were statistically shown to be mutated more often than expected by chance alone. These statistically significant mutations were used as seeds for OncoMerge integration. Somatic PAMs used as seeds were identified with MutSig2CV q-values less than or equal to 0.1[7] and a mutation frequency greater than 5%. Gene fusions used as seeds were identified as significant in PRADA[8] and a mutation frequency greater than 5%. CNAamps or CNAdels used as seeds were identified as significantly amplified or deleted from the amplified genes (amp_genes) or deleted genes (del_genes) GISTIC output files with residual q-values less than or equal to 0.05. CNAs from sex chromosomes (X and Y) were excluded. Genes from sex chromosomes can enter OncoMerge as seeds from PAMs or fusions. These seed genes become the starting point of the OncoMerge integration. Subsequent steps determine if Act or LoF merged mutation profiles or their component PAM, Fusion, CNAamp, or CNAdel mutation roles are the most appropriate integration model for a gene.

Merging somatic mutations in OncoMerge

The mutation role for each seed gene is assigned based on the following criteria (Supp. Fig 1):

- If Act frequency (PAM+Fusion+CNAamp) > PAM+Fusion frequency and the Act frequency ≥ 5% then the mutation role is set to Act.
- Else LoF frequency (PAM+Fusion+CNAdel) > PAM+Fusion frequency and the LoF frequency ≥ 5% then the mutation role is set to LoF.
- Else if the gene mutation role is not set to Act or LoF:
    - If the gene is a PAM seed gene (MutSig2CV q-value ≤ 0.1 and frequency ≥ 5%) and has a frequency greater than Fusion, CNAamp, and CNAdel, then the mutation role is set to PAM.
    - Else if the gene is a Fusion seed gene (TumorFusion.org frequency ≥ 5%) and has a frequency greater than PAM, CNAamp, and CNAdel, then the mutation role is set to Fusion.
    - Else if the gene CNAamp frequency ≥ 5% and has a frequency greater than PAM, Fusion, and CNAdel, then the mutation role is set to CNAamp.
    - Else if the gene CNAdel frequency ≥ 5% and has a frequency greater than PAM, Fusion, and CNAamp, then the mutation role is set to CNAdel.

Permuted q-value (PQ) filter

For putative Act and LoF mutations, a permuted q-value is computed by randomizing the order of rows in the PAM, Fusion, and CNA mutation matrices' and then calculating the randomized frequency distribution for Acts and LoFs. The observed frequency for an Act or Lof mutation is then compared to the randomized frequency distribution to compute the permuted p-value. Permuted p-values are corrected into q-values using the multiple-test Benjamini-Hochberg FDR-based correction method. Only Acts or LoFs that had a permuted q-value ≤ 0.1 were retained. Any Act or LoF with a permuted q-value > 0.1 was set to the mutation role of either PAM,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fusion, CNAamp, or CNAdel based on which mutation role had the highest frequency. The permuted q-value cutoff can be set through a command line parameter ('-pq', --perm_qv').

Minimum final frequency (MFF) filter

A low-pass genomic filter was applied to each CNA locus if the CNA locus had ≥ 10 underlying genes. The number of genes underlying a CNA locus can be set through a command line parameter ('-mlg', --min_loci_genes'). The filter keeps only the gene(s) with the maximum mutation frequency, and all genes with the maximum mutation frequency are kept for ties.

Microsatellite hypermutation censoring (MHC) filter

The TCGA tumors used in this study have been characterized for both MSI[5] and hypermutation[6] (**Supplementary Table 1**). The tumors with MSI or hypermutation are loaded as a blocklist of patient IDs through a command line parameter ('-bl' or '--blocklist'). All tumors in the blocklist are excluded from consideration by the PQ and MFF filters while determining the genes to include in the final somatic mutation matrix. The mutation status for blocklist tumors are included in the final integrated mutation matrix.

OncoMerge outputs

OncoMerge provides four output files that provide valuable information about the integration process and the final integrated mutation matrix that can be used in downstream studies. Here is a brief description of each file and its contents:

- oncoMerge_mergedMuts.csv – The integrated mutation matrix is comprised of genes (rows) by patient tumors (columns) of mutation status after integration by OncoMerge. The matrix values are [0 or 1]: zero indicates that the gene is not mutated in a patient tumor, and one indicates that the gene was mutated in a patient tumor.
- oncoMerge_CNA_loci.csv – A list of the genes mapping to each CNAamp or CNAdel locus included in the OncoMerge integrated mutation matrix.
- oncoMerge_ActLofPermPV.csv – List of all significant Act and LoF genes, their OncoMerge mutation role, frequency, empirical p-value, and empirical q-value. This output is before the application of the low-pass frequency filter.
- oncoMerge_summaryMatrix.csv – Matrix of genes (rows) by all information gathered by OncoMerge.

To aid in comparisons between runs, we provide the save permutation option ('-sp' or '--save_permutation') to output permutation results so that the same permuted distribution can be used with different parameters in separate runs. We also provide the load permutation option ('-lp' or '--load_permutation') to load up the permuted distribution from a previous run. The permuted distributions are saved in the following files if requested:

- oncomerge_ampPerm.npy, oncomerge_delPerm.npy – Snapshot of the non-deterministic permutation results from combining PAM, Fusion, and CNAamp or PAM, Fusion, and CNAdel frequencies, respectively.

Gold standard cancer-specific gene role validation datasets

Gold standard datasets are vital to validating the usefulness of each feature in OncoMerge. Two different sources of gold standard cancer-specific gene role (Act or LoF) datasets were used to validate the OncoMerge predicted tumor-specific gene roles:

- TCGA consensus: The TCGA consensus was constructed by Bailey et al., 2018 wherein they catalog a list of 299 unique oncogenesis associated genes[6]. In the TCGA

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

consensus 280 cancer-specific oncogene roles were identified, and 417 cancer-specific tumor suppressor roles were identified (**Supplementary Table 2**).

- Cancer Gene Census (CGC): The CGC was developed by Catalogue of Somatic Mutations in Cancer (COSMIC) as an expert-curated database of human cancer-driving genes[15]. CGC cancers were mapped to the TCGA cancers by manual curation (Supplementary Table 2). In the CGC 205 cancer-specific oncogene roles were identified, and 304 cancer-specific tumor suppressor roles were identified (**Supplementary Table 2**).

Gold standard gene role validation datasets

Three different sources of gold standard gene role (Act or LoF) datasets were used to validate the OncoMerge predicted gene roles:

- 20/20 rule: The 20/20 rule defines oncogenes (Act) by requiring >20% of mutations in recurrent positions, and tumor suppressors (LoF) as >20% of recorded mutations are inactivating (missense or truncating)[4]. With the 20/20 rule, 54 oncogene roles were identified, and 71 tumor suppressor roles were identified (**Supplementary Table 2**).
- OncodriveROLE: The OncodriveROLE is a machine learning algorithm that classifies genes according to their role based on well-curated genomic features[16]. With OncodriveROLE, 76 oncogene (Act) roles were identified, and 109 tumor suppressor (LoF) roles were identified (**Supplementary Table 2**).
- Tokheim Ensemble: Ensemble-based method from Tokheim et al., 2016[14], which integrates MutSigCV, 20/20+, and TUSON methods for predicting gene roles. With the Tokheim Ensemble, 78 oncogene (Act) roles were identified, and 212 tumor suppressor (LoF) roles were identified (**Supplementary Table 2**).

Computing overlap between OncoMerge and gold standards

A hypergeometric enrichment statistic was used to compute the significance of overlap observed between each gene role in OncoMerge versus the gold standards. When possible, the tumor specificity of the gene role was taken into consideration (TCGA consensus and CGC). Enrichment p-values less than the Bonferroni corrected alpha value of 0.002 were considered significant.

TCGA Pan-Cancer SYstems Genetics Network AnaLysis (SYGNAL)

The mRNA and miRNA expression data required to run SYGNAL were obtained from Thorsson et al., 2018[18]. The SYGNAL pipeline is composed of 4 steps and command-line parameters for all programs are described in detail in Plaisier et al., 2016[17]. Each cancer was run separately through the pipeline to reduce the confounding from tissue of origin differences. Highly expressed genes were discovered for each cancer by requiring that genes have greater than or equal to the median expression of all genes across all conditions in ≥ 50% of patients[18]. These gene sets were then used as input to SYGNAL to construct the gene regulatory networks (GRNs) for each cancer.

The underlying cMonkey2 biclustering results are identical to those from Thorsson et al., 2018[18] as they do not rely upon genetic information. Using Network Edge Orienting (NEO)[17,27] somatic mutations are integrated with bicluster and regulator expression in the next step. The systems genetics analysis with NEO was modified from Thorsson et al., 2018 in two ways: 1) we removed constraints to identify immune-related regulatory interactions, which substantially increased the size of the network by including additional patient survival-associated biclusters

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not associated with immune functions; and 2) the OncoMerge integrated mutation matrix was used and compared against the PAM only mutation matrix used previously in Thorsson et al., 2018[18].

TF regulatory network construction for PanCan-SYGNAL networks

A TF regulatory network was built for each cancer in three steps (**Figure 6A**). First, the TFs regulating survival-associated biclusters were extracted from each cancer's SYGNAL GRN. Second, a preliminary $TF_{regulator} \rightarrow TF_{target}$ regulatory network was constructed based on the presence of a binding site for a putative $TF_{regulator}$ in the promoter of a TF_{target} from the Transcription Factor Target Gene Database[17] (http://tfbsdb.systemsbiology.net). TF family expansion[17] was used to supplement TFs that did not have an experimentally determined DNA recognition motif in the database. The assumption was that the motifs within a TF family would not vary significantly. Therefore TF family members from the TFClass database[28] with a known DNA recognition motif can be used as a proxy for a TF with no known DNA recognition motif. Finally, the putative $TF_{regulator} \rightarrow TF_{target}$ regulatory network was filtered by requiring a significant Pearson correlation between the mRNA expression of the $TF_{regulator}$ and TF_{target} (Pearson's $|R| \geq 0.3$ and p-value ≤ 0.05). The sign of the correlation coefficient can be used to determine the role of a regulatory interaction: a positive correlation coefficient equates to the $TF_{regulator}$ being an activator, and a negative correlation coefficient equates to the $TF_{regulator}$ being a repressor. Networks with fewer than 50 interactions were not included in the analyses as they were not sufficiently powered to run the network motif analysis. The cancer regulatory networks for DLBC, KICH, KIRP, OV, TGCT, and THYM were excluded from further studies.

TF regulatory network motif analysis

Three-node network motifs were enumerated from the TF regulatory networks using mfinder[23] in the same manner as Neph et al., 2012[22] and used to compute triad significance profiles (TSPs)[24]. The parameters used with mfinder v1.20 were[22]: motif size set at 3 (-s 3), requested 250 random networks to be generated (-r 250), and the Z-score threshold was set at -2000 to ensure all motifs are reported (-z -2000). All Z-scores were extracted for each cancer and converted to triad significance profiles using the methods of Milo et al., 2004[24].

For consistency, the TF regulatory networks for the 41 different cell types from Neph et al., 2012[22] were downloaded from http://www.regulatorynetworks.org/ and analyzed using the same approach described above.

Signed network motif analysis incorporating TF regulator interaction roles

The enrichment of signed feed-forward loops (FFLs), regulated feedback, and regulating feedback network motifs was computed using FANMOD[25], which takes into consideration TF regulatory roles (activation and repression). The command line version of FANMOD from IndeCut[29] was used with default parameters, except for the inclusion of regulatory role (colored edges)[25] (fanmod 3 100000 1 <input_file> 1 0 1 2 0 1 0 1000 3 3 <output_file> 1 1). Z-scores for signed FFLs, regulated feedback, and regulating feedback network motifs were extracted for each cancer and converted to triad significance profiles using the methods of Milo et al., 2004[24]. The signed FFL network motifs are broken down into C1, C2, C3, C4, I1, I2, I3, and I4, as described previously[30].

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Establishing a baseline for the integration of somatic mutations

Somatic mutations play a significant role in cancer pathogenesis, and the main mutation types are PAMs, fusions, and CNAs (amplifications and deletions). Somatic mutation of the same gene with different mutation types can have similar downstream effects on cancer phenotypes. We have developed OncoMerge as a systematic method to integrate PAM, fusion, and CNA somatic mutations into a more comprehensive mutation matrix for subsequent analyses. OncoMerge systematically integrates somatic mutations and defines a role for each gene (**Figure 1**): PAM, fusion, CNA deletion (CNAdel), CNA amplification (CNAamp), Activating (Act), and Loss of Function (LoF). The role assigned to a gene describes the rubric used to integrate the data from the source data matrices.

A significant part of developing OncoMerge was constructing and optimizing the statistical filters that provide an essential quality control step to identify somatically mutated genes that are more likely to be functional in tumor biology. The selection and optimization of OncoMerge statistical filters were performed using the 9,584 patient tumors from 32 cancers profiled by the TCGA Pan-Cancer Atlas[1,6]. We used three metrics to assess the value of potential filters: 1) impact on the number of somatically mutated genes (**Figure 2A**); 2) impact on the distribution of the number of genes mapping to genomic loci (**Figure 2B**); and 3) significance of the overlap between somatically mutated genes from OncoMerge with gold standard datasets (including overlap with gene roles and tumor-specific gene roles; **Figure 2C**; **Supplementary Table 3**). These metrics ensure that the integrated somatic mutations are consistent with prior knowledge and that the size of CNA mutations does not overwhelm the integration algorithm.

Next, we determined the integration baseline by applying OncoMerge to the TCGA Pan-Cancer Atlas without filtering. Slightly less than one-third of the genome was considered somatically mutated in at least 5% or greater of tumors in at least one of the 32 cancers (30% or 6,028 genes, **Figure 2A**). We observed a highly significant overlap between OncoMerge somatically mutated genes and the combined gold standard (genes = 395, p-value = 1.1 x 10^{-44}, **Figure 2C**) when gene role was not considered. Significant overlaps existed between the LoF somatic mutations from three gold standards (TCGA consensus, CGC, and Vogelstein) with the somatic mutations with the LoF predicted role from OncoMerge (**Figure 2C**). None of the comparisons of Act somatic mutations were significantly overlapping (**Figure 2C**). Many of the 6,028 genes map to the same copy number alteration genomic locus (**Figure 2B**). These unfiltered results reveal two main integration biases. First, there is no overlap of Act somatic mutations with previously identified Act mutations. Second, the integration with CNAs is causing the inclusion of many passenger mutations mapping to the same genomic locus. OncoMerge applied to the TCGA Pan-Cancer Atlas without filtering provides a baseline to benchmark success. Addressing the integration biases we observed is the impetus we had for developing and optimizing filters for OncoMerge.

Developing an optimal filtering strategy for the integration of somatic mutations

A key consideration in developing OncoMerge was that integrating the somatic mutation types should highlight the functional somatic mutations over passenger mutations. Therefore, we created two filters designed to prioritize somatically mutated genes that are more likely to be functional. The first filter determined if the final mutation frequency after integrating PAM, fusion, and CNA somatic mutations is larger than expected by chance alone. A permutation-based approach empirically determined the background integrated mutation frequency distribution.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Then the observed frequencies are compared to the randomized background distribution to calculate permuted p-values, which are corrected using the Benjamini-Hochberg method to provide permuted q-values. A permuted q-value ≤ 0.1 denotes a significant final mutation frequency. The permuted q-value (PQ) filter reduced the number of somatically mutated genes to 5,630 (**Figure 2A**). This filtering improved LoF somatic mutations from three to four gold standards (TCGA consensus, CGC, Vogelstein, and OncodriveROLE) with the somatic mutations that had the LoF predicted role from OncoMerge. Still, the Act comparisons did not show significant enrichment (**Figure 2C**). The PQ filter had a minimal impact on the number of genes per locus (**Figure 2B**). This lack of significant overlap for Act somatic mutations demonstrates that further filtering is required.

The second filter deals with passenger gene somatic mutations. An average CNA encompasses 3.8 ± 7.9 Mb of genomic sequence[31], and genomic segments of this size typically include many genes. These large genomic regions make it difficult to determine which of the affected genes are the functional gene(s) underlying the CNA locus without integrating additional information. We assert that passenger genes underlying a CNA locus can be considered noise and can be identified by the lack of allelic heterogeneity. Thus, functional gene(s) can be identified through allelic heterogeneity that boosts the somatic mutation frequency for a gene above the background CNA frequency. We designed a low-pass filter that retains only the gene(s) with the maximum final frequency (MFF). The MFF filter is only applied if a locus has more than ten genes. Application of the MFF filter dramatically reduced the number of somatically mutated genes from 6,028 to 1,459 (**Figure 2A**) and the number of genes per locus (**Figure 2B**). We additionally observed a marked improvement in overlap with the gold standards. Significant enrichment was observed for four Act gold standards with somatic mutations that OncoMerge predicts to be Act, and all five of the LoF gold-standard versus OncoMerge predicted LoF comparisons (**Figure 2C**). The MFF filter directly addresses the issue of too many genes in a CNA locus. Removing more than three-quarters of the somatically mutated genes improves the overlaps with gold standards.

We then assessed the impact of applying both the PQ and MFF filters. Simultaneous application of both filters led to a slight reduction in the number of somatically mutated genes beyond the MFF filter (1,398 genes; **Figure 2A**), and the improvement in the number of genes per locus was retained (**Figure 2B**). There was also an improvement in the significant overlap with gold standards where all five LoF gold-standard versus OncoMerge predicted LoF and four Act gold-standard versus OncoMerge predicted Act were significant (**Figure 2C**). Importantly, none of the gold standard Act versus LoF or LoF versus Act comparisons were significant for any filter combination, demonstrating that the OncoMerge predicted roles are consistent with prior knowledge.

Reducing biases due to microsatellite instability and hypermutation

Microsatellite instability (MSI) and hypermutation phenotypes drastically increase the number of somatic mutations in a tumor. The PQ and MFF filters and OncoMerge's core algorithm rely upon somatic mutation frequency which is susceptible to confounding by MSI or hypermutation. Fortunately, all TCGA tumors used in this study are characterized for both MSI[5] and hypermutation[6] status (**Figure 3A**). We observed a highly significant positive correlation between MSI/hypermutation frequency and the total number of somatic mutations per cancer after integration by OncoMerge (R = 0.69 and p-value = 1.1×10^{-5}). This strong positive correlation demonstrates that MSI/hypermutation is likely inflating the number of somatic

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mutations discovered by OncoMerge. Therefore, we created the MSI and hypermutation censoring filter (MHC) to exclude these tumors while OncoMerge determines which genes to include in the final somatic mutation matrix. The mutation status for tumors with MSI and hypermutation are included for genes in the final integrated mutation matrix. Applying the MHC filter alongside the PQ and MFF filters reduced the overall number of somatically mutated genes (1,133 genes; **Figure 2A**) and had minimal impact on the number of genes per locus (**Figure 2B**; **Supplementary Table 4**). The combined PQ, MFF, and MHC filters decreased the correlation between the MSI/hypermutation frequency (R = 0.53 and p-value = 1.7 x 10^{-3}). All ten of the gold standard Act vs. Act and LoF vs. LoF comparisons were significant. These results established that the MHC filter is valuable for removing passenger mutations introduced by tumors with severely increased somatic mutation rates. The PQ, MFF, and MHC filters comprise the default and final OncoMerge filter set. The filters deal with known complications in cancer genetics and ensure that the mutation roles in the integrated matrix are correctly assigned.

Benefits of an integrated somatic mutation matrix

We evaluated the benefits of systematic somatic mutation integration by comparing OncoMerge integrated somatic mutation matrices to those from PAMs. The PAM somatic mutation matrices were used as a reference point because we have successfully used them as the sole source for somatic mutations in previous studies[17,18]. We assessed the benefits of integration by tabulating the number of somatic mutations and their roles (**Figure 3B**), the number of genes added by integration (**Figure 3C**), and the increase in somatic mutation frequency due to integration (**Figure 3E**). Impressively, Act and LoF mutations represented the bulk of the somatic mutations in 30 cancers (**Figure 3B**). The papillary thyroid carcinoma (THCA) and kidney chromophobe (KICH) were the only cancers that lacked Act or LoF mutations. Consistent with Agrawal et al. 2014[32], THCA had only three mutations with a frequency ≥ 5% BRAF, NRAS, and RET. On the other hand, KICH was under-sampled in the TCGA Pan-Cancer atlas (n = 65), and LoF and Act mutations would likely be discovered with the inclusion of more patient tumors.

We then investigated how many new genes the integration added for each cancer. Integration added at least one somatically mutated gene for each cancer (**Figure 3C**), and more than eighty somatically mutated genes for BLCA, LUAD, and UCEC (**Figure 3C**). The somatically mutated genes added by OncoMerge make the integrated somatic mutation matrices more comprehensive.

Next, we investigated the frequencies of the somatic mutations from the OncoMerge integrated mutation matrices. The genes with the highest frequency map to well-known oncogenes (e.g., BRAF) and tumor suppressors (e.g., APC and TP53; **Figure 3D**). The two tumor suppressor genes APC and TP53 were mutated in greater than eighty percent of the tumors for multiple cancers (**Figure 3D**). The APC gene was mutated in greater than eighty percent of tumors for colon adenocarcinoma (COAD) and rectal adenocarcinoma (READ). The TP53 gene was mutated in greater than eighty percent of tumors for esophageal carcinoma (ESCA), lung squamous carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), rectal carcinoma (READ), and uterine carcinosarcoma (UCS). These frequently mutated genes in the OncoMerge integrated mutation matrices are consistent with prior knowledge of somatic mutations for each cancer.

Finally, we calculated the frequency added through integration by subtracting the integrated mutation frequency from the PAM frequency. The most substantial increases in somatic

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mutation frequency were observed for TMPRSS2-ERG in prostate adenocarcinoma (PRAD) and CDKN2A in mesothelioma (MESO), glioblastoma (GBM), diffuse large B-cell lymphoma (DLBC), and esophageal carcinoma (ESCA; **Figure 3E**). Neither TMPRSS2-ERG nor CDKN2A would have been identified as somatically mutated without incorporating fusions and CNAs, respectively. These findings demonstrate that OncoMerge significantly improves the number and frequency of somatically mutated genes in most cancers. Also, these results show that the systematic integration of PAM, fusion, and CNA somatic mutations is crucial for obtaining a comprehensive mutation matrix for each cancer.

Pan-cancer somatic mutations capture many known tumor suppressors and oncogenes

Genes mutated in multiple cancers are of great interest as selective pressures have found a common solution in different contexts to influence cancer phenotypes. Therefore, we searched for genes somatically mutated in at least five cancers in the OncoMerge integrated mutation matrices. The resulting gene list could be broken down into two groups of somatic mutations: the LoF set (n = 28, **Figure 4A**) and the Act set (n = 18, **Figure 4B**). The FBXW7, KMT2C, and KMT2D somatic mutations were challenging to classify as LoF or Act. The genes FBXW7 and KMT2D were somatically mutated with PAMs in six and seven cancers, respectively (**Figure 4A**). The gene KMT2C (also known as MLL3) was primarily LoF and PAM but had the mutation role of Act for ovarian cancer (OV) (**Figure 4B**). Based on a literature search, all three genes have been classified as tumor suppressors[33–35]. Therefore, we grouped FBXW7, KMT2C, and KMT2D mutations with the LoF set.

The pan-cancer somatically mutated genes harbored many well-known tumor suppressors and oncogenes (**Figure 4C**). As expected, tumor suppressors[33] were significantly enriched in the LoF group (overlap = 20, p-value = 2.0×10^{-20}), and oncogenes[36] were significantly enriched in the Act group (overlap = 8, p-value = 9.2×10^{-9}). The top three most somatically mutated tumor suppressors were TP53, PTEN, and CDKN2A. These three tumor suppressors control important checkpoints in the cell cycle making them functionally interesting. The gene TP53 was somatically mutated in 24 cancers, primarily by PAMs, but four LoF were also observed for glioblastoma (GBM), liver hepatocellular carcinoma (LIHC), prostate adenocarcinoma (PRAD), and sarcoma (SARC). The top three most mutated oncogenes across cancers were PIK3CA, KRAS, and CCNE1. Two of these genes (PIK3CA and KRAS) become overactive kinases when mutated, and CCNE1 is a fundamental part of the cell cycle regulatory machinery. Both PIK3CA and KRAS have PAM and Act mutation roles across the different cancers, and only the NFE2L2 gene has a similar mixture of PAM and Act mutation roles. The remainder of the oncogenes are like CCNE1 in that the gene somatic mutation roles are all Act. These pan-cancer analyses further validate the systematic somatic mutation integration by OncoMerge through the unbiased recall of tumor suppressors and oncogenes.

Improving gene regulatory network inference

A major goal of developing OncoMerge was to construct an integrated somatic mutation profile that would increase the power to identify how mutations modulate cancer phenotypes. Previously we used PAMs from the cancers in the TCGA Pan-Cancer Atlas as input for SYGNAL to construct gene regulatory networks (GRNs)[18]. SYGNAL GRNs are composed of causal and mechanistic interactions linking somatic mutations to a TF or miRNA regulator to a co-regulated set of genes (bicluster). Somatic mutations in SYGNAL are used as input for the Network Edge Orienting (NEO) portion of the pipeline that infers causal flows of information

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(somatic mutation → TF or miRNA regulator → bicluster of co-regulated genes). Therefore, we recomputed NEO analyses using the OncoMerge integrated somatic mutation matrices for each cancer in the TCGA Pan-Cancer Atlas to demonstrate the increased power to detect causal flows of information. The resulting networks were filtered to include only biclusters with good quality co-expression that were significantly associated with patient survival. Regulatory interactions were required to be both causal (significant evidence of information flow between a mutation → regulator → bicluster) and mechanistic (enrichment of regulator binding sites in the promoter or 3' UTR of the bicluster genes). We compare SYGNAL GRNs inferred using OncoMerge integrated mutation matrices (**Supplementary Table 5**) with SYGNAL GRNs inferred using the legacy PAM-based mutation matrices from Thorsson et al., 2018.

The GRNs are comprised of nodes and edges. The degree of a node is the number of edges connecting it to other nodes. The average degree is a standard network metric computed as the average of all node degrees in the network. We found that the average degree was larger for 26 OncoMerge GRNs relative to legacy GRNs (**Figure 5A**). The exceptions were GBM (average degree was equal) and COAD and STAD (legacy had a larger average degree). COAD and STAD have many MSI and hypermutation tumors (**Figure 3A**), suggesting that the MHC filter removed spurious associations. Furthermore, the hypothesis that MSI and hypermutation inflated the average degree of GRNs is supported by the reduction in the number of COAD mutations in the OncoMerge GRN relative to the legacy GRN (**Figure 5B**). Thus, we have increased the average degree in the networks and addressed a systematic bias found in legacy networks.

Next, we considered the number of mutations in each GRN predicted to modulate the activity of regulators. The OncoMerge GRNs contained more somatic mutation nodes than the legacy GRNs for all cancers but COAD, likely due to MSI and hypermutation as described above (**Figure 5B**). Then, we assessed the recall of somatic mutations previously associated with each cancer from the DisGeNET database[37]. All but two OncoMerge GRNs recalled more previously associated somatic mutations than the legacy GRNs (**Figure 5C**). The exceptions were UVM with the same amount and COAD with fewer (**Figure 5C**). This demonstrates that OncoMerge integrated mutation matrices provide increased power for linking somatic mutation matrices into GRNs, and improve the capture of somatic mutations previously associated with each cancer.

Finally, we considered the number of predicted causal and mechanistic transcription factor (TF) regulators in each GRN. The OncoMerge GRNs contained more predicted TF regulators than legacy GRNs for all but GBM, which had one less TF (**Figure 5D**). We also assessed the recall of TFs previously associated with each cancer from the DisGeNET database[37,38]. Twenty-four of the OncoMerge GRNs recalled more previously associated TFs than legacy GRNs (**Figure 5E**). The GBM and KIRP GRNs had the same amount, and KICH and UVM had no recall of previously associated TFs in either GRN (**Figure 5E**). In summary, using OncoMerge integrated mutation matrices in GRN construction builds more extensive and biologically meaningful networks.

Comparing active and static TF regulatory network architectures

The interactions between TFs are important for generating the transcriptional state of a human cell. The underlying architecture of TF regulatory networks, comprised of TFs and their interactions, are typically explored by enumerating all three-node network motifs and computing

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

their enrichment or depletion into triad significance profiles (TSPs)[24]. Most studies of network motif enrichment have relied upon unsigned interactions[22,24,39–42], which ignore whether the interaction is activating or repressing. To facilitate comparisons, our first analysis of network architecture uses unsigned TSPs to compare static and active TF regulatory networks. Static TF regulatory networks were constructed using chromatin accessibility and DNA binding motifs for 41 cell types[22]. These TF regulatory networks are static because they do not incorporate gene expression data in their construction. Active TF regulatory networks are derived from the OncoMerge augmented SYGNAL pan-cancer GRNs, that were trained using patient tumor transcriptional data and therefore are comprised of active TF regulatory interactions. Using the following steps, we constructed TF regulatory networks for each cancer from the pan-cancer SYGNAL GRNs (**Figure 6A**). First, we extracted all the TF regulators from the pan-cancer GRNs. Interactions between TFs were inferred based on the presence of DNA binding motifs from the TF target gene database[17], and a significant correlation between the TF regulator and TF target in patient tumor expression (Pearson's $|R| \geq 0.3$ and p-value ≤ 0.05; **Figure 6A**; **Supplementary Table 6**). The enrichment (or depletion) of motifs in the network was computed using TSPs[24]. Triad significance profiles were calculated for twenty-five TF regulatory networks and summarized as the median TSP (**Figure 6A** & **B**). We excluded the cancer types DLBC, KICH, KIRP, OV, TGCT, and THYM because they had too few inferred regulatory interactions (< 50 interactions). Finally, we recomputed the TSPs for the static TF regulatory networks using a more recent version of the mfinder algorithm (**Figure 6B**).

The median TSPs of the active and static TF regulatory networks were highly correlated (R = 0.75, p-value = 3.0×10^{-3}; **Figure 6B**). Demonstrating that the architecture of the active network resembles the static network. However, the maximum enriched network motifs were different. The regulated and regulating feedback motifs (motifs 108 and 46) were the most highly enriched motifs from the static TF regulatory networks and were still enriched, although not as significant as in the active networks. In contrast, the feed-forward loop (FFL, motif 38) is the most highly enriched motif in the active TF regulatory networks. These two motifs are quite similar in structure and differ only by a single edge. Feedback motifs and FFLs can be further broken down into ten and eight signed network motifs that each have a unique functional output[30]. Thus exploring the enrichment of signed network motifs allows the discovery of what functions are being selected for by evolution in general and the microcosm of tumor biology.

Coherent feed-forward loops enriched in active TF regulatory networks

Incorporating the sign of the regulatory interactions (activating or repressing) splits the FFL motif into eight signed network motifs classified as coherent (C1, C2, C3, C4) and incoherent (I1, I2, I3, I4)[30]. Simulation studies have demonstrated that coherent FFLs lead to delays in target gene expression, and incoherent FFLs accelerate target gene expression[30]. FFLs were significantly enriched in active TF regulatory networks, which led us to question whether coherent, incoherent, or both FFLs were enriched. In active GRNs, the sign of the correlation between the TF regulator to TF target can be used to determine the sign of the interaction ($R > 0$ equates to activation, $R < 0$ equates to repression). The four coherent FFLs were enriched in the active TF regulatory networks (**Figure 6C**; **Supplementary Table 7**), and incoherent FFLs were severely under-enriched ($Z \ll 0$). In summary, coherent FFLs were enriched in our active TF regulatory networks, suggesting that transcriptional delay mechanisms must provide a valuable function for TF regulatory networks.

Coherent switch-like feedback motifs enriched in active TF regulatory networks

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The regulated and regulating mutual feedback motifs have a two-node feedback loop at their core. The double-positive and double-negative two-node mutual feedback loops act like switches[43]. We tested the twenty signed regulated and regulating mutual feedback network motif configurations for enrichment in TF regulatory networks. Three regulating and three regulated signed mutual feedback motifs (**Figure 6C**; **Supplementary Table 7**). These six enriched regulated and regulating mutual feedback motifs had a commonality in their configuration. Firstly, all the network motifs were coherent. Coherent regulated and regulating feedback loops have interaction signs between the feedback loop that are either double-positive or double-negative. And the regulated or regulating node interacts with the feedback loop nodes using the same sign for double-positive feedback loops and the opposite sign for double-negative feedback loops. Thus, there are three coherent configurations for both regulated and regulating mutual feedback motifs making six total, coinciding with the six enriched configurations (**Figure 6C**; **Supplementary Table 7**). The enriched motifs containing a double-positive feedback loop had the same interactions with the non-feedback loop node, both activating or repressing (**Figure 6C**). The enriched motif containing a double-negative feedback loop had opposing interactions with the non-feedback loop node, one activating and one repressing (**Figure 6C**). These enriched signed network motifs are the configurations that function as molecular switches[44]. Again, evolution has selected for coherent network motif configurations likely because of their function.

Discussion

We developed OncoMerge to integrate PAMs, fusions, and CNAs into a more accurate representation of the somatic mutation landscape of patient tumors. The OncoMerge integration algorithm and three filters (PQ, MFF, and MHC) effectively address the issues of allelic heterogeneity and the unification of binary and quantitative mutation data. These issues have forced most studies of somatic mutations to focus on one somatic mutation type and were the impetus for us to develop OncoMerge for the integration of the three most common somatic mutation types. We tested OncoMerge by integrating the somatic mutation data from 32 cancers from the TCGA Pan-Cancer Atlas. Comparison to gold standards confirmed that the genes and roles selected by OncoMerge were accurate. The integration of somatic mutation types had several quantifiable benefits for somatically mutated genes. First, most somatically mutated genes had an integrated role of Act or LoF, demonstrating that consolidation of allelic heterogeneity is vital to achieving a complete picture of somatic mutations for a patient cohort. Second, genes somatically mutated primarily by fusions and CNAs were added by the integration. Lastly, the frequency of many somatically mutated genes increased due to the integration of the three somatic mutation types. We used the integrated somatic mutations as input to SYGNAL to demonstrate improvements in power for systems genetics-based inference of GRNs. Using integrated somatic mutations increased the average connectedness of the GRNs by incorporating more somatic mutations and regulators previously linked to cancer biology. Next, we found that while the underlying architecture of active SYGNAL TF regulatory networks and static DNA binding TF regulatory networks were similar overall, the top most enriched network motifs were different. We discovered that switch-like feedback and delay-inducing feed-forward loop motifs were enriched in TF regulatory networks. We developed and tested a novel systematic integration tool and demonstrated that integrated somatic mutations improve our ability to link somatic mutations with cancer phenotypes.

The construction of active GRNs enabled the exploration of signed network motifs and led to the discovery that specific signed network motif configurations are being enriched. The SYGNAL

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

GRNs construction method identifies active gene regulatory interactions by discovering interactions that are supported by gene expression data from patient tumors [PMID = 27426982]. On the other hand, prior networks were static maps of DNA binding sites constructed using digital genomic footprinting and the similarity of the underlying sequence of the footprints for known DNA binding motifs[22]. The active networks use a correlation-based method to determine TF regulatory roles (activator or repressor) for the interactions, which is not possible using static binding maps. Analyzing signed network motifs provides a leap forward in understanding how the underlying architecture of GRNs functions in real-world biological systems. OncoMerge integrated somatic mutations offer a more solid platform to infer active GRNs that can be used to explore the functional architecture of TF regulatory networks.

We discovered that coherent regulated and regulating feedback and FFL network motifs were enriched in cancer TF regulatory networks. We cannot say whether this enrichment of network motifs will generalize to all active GRNs or if this is a cancer-specific phenomenon. In normal organismal development, feedback motifs have been previously shown to be important for cell fate decision-making[45,46]. On the other hand, in tumor cells and other cells in the tumor microenvironment, the enriched feedback motifs may be maintaining a cell fate, or the disease could be coopting the circuit to drive tumor biology. Likewise, coherent FFL network motifs have also been associated with enhanced drug resistance[47]. These coherent motifs are relevant for normal and diseased cell biology, and evolution has specifically selected these motif configurations because of their unique functional outputs.

We provide the Oncomerge software in several standard distribution formats to facilitate future studies that aim to integrate somatic mutations. The source code is available on GitHub (https://github.com/plaisier-lab/OncoMerge). Finally, a Docker image was created that can be run as a virtual machine with all dependencies pre-installed (https://hub.docker.com/r/cplaisier/oncomerge). Detailed documentation is provided, along with a tutorial that describes the use of OncoMerge. The goal of disseminating OncoMerge in these ways is to give end-users flexibility to choose what distribution method best fits their computational platform.

Additionally, we provide the OncoMerge integrated somatic mutation matrices for those planning studies that use somatic mutations from the TCGA Pan-Cancer Atlas (https://doi.org/10.6084/m9.figshare.20238867). These integrated somatic mutation matrices can be used for any downstream analyses incorporating somatic mutations and will provide the same power boost observed in our studies. In addition, we also offer the pan-cancer SYGNAL GRNs and TF regulatory networks as supplementary tables to expedite systems genetics studies of TCGA cancers. We hope these accessible results will facilitate studies linking somatic mutations to downstream cancer phenotypes and lead to novel biological insights in clinical samples.

Future improvements to the OncoMerge algorithm include a more quantitative integration approach for the somatic mutations, a replacement for or an improved maximum final frequency filter, aggregation across pathways, and a determination of whether other genomic features may be integrated (ecDNA[48] or epigenomics[49]). Additionally, in future single-cell studies with both transcriptome and genome information, it would be helpful to have an OncoMerge implementation that integrates PAM, fusion, and CNA for every single cell. We envision OncoMerge as a valuable tool in the somatic mutation characterization pipeline. We hope that it

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

will facilitate multi-omic studies and lead to novel discoveries that can be translated into clinical insights.

Acknowledgments

This work was supported by NIH-NINDS Award # 1R01NS123038-01, and 1R01NS119650-01. The authors also acknowledge Robert Schultz for assistance in preliminary studies, and the Cancer Genome Atlas Research Network for the TCGA Pan-Cancer Atlas multi-omic patient tumor profiles.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

2. Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res.* **46**, D1144–D1149 (2018).

3. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

4. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

5. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis. Oncol.* **2017**, (2017).

6. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

7. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

8. Torres-García, W. *et al.* PRADA: pipeline for RNA sequencing data analysis. *Bioinforma. Oxf. Engl.* **30**, 2224–2226 (2014).

9. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

10. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

11. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

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

13. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

15. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

16. Schroeder, M. P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinforma. Oxf. Engl.* **30**, i549-555 (2014).

17. Plaisier, C. L. *et al.* Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. *Cell Syst.* **3**, 172–186 (2016).

18. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

19. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

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

21. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).

22. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).

23. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).

24. Milo, R. *et al.* Superfamilies of evolved and designed networks. *Science* **303**, 1538–1542 (2004).

25. Wernicke, S. & Rasche, F. FANMOD: a tool for fast network motif detection. *Bioinforma. Oxf. Engl.* **22**, 1152–1153 (2006).

26. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

27. Plaisier, C. L. *et al.* A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet.* **5**, e1000642 (2009).

28. Wingender, E., Schoeps, T. & Dönitz, J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41**, D165-170 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

29. Ansariola, M., Megraw, M. & Koslicki, D. IndeCut evaluates performance of network motif discovery algorithms. *Bioinforma. Oxf. Engl.* **34**, 1514–1521 (2018).

30. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11980–11985 (2003).

31. Harbers, L. *et al.* Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. *Front. Oncol.* **11**, 700568 (2021).

32. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).

33. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* **44**, D1023-1031 (2016).

34. Hillman, R. T. *et al.* KMT2D/MLL2 inactivation is associated with recurrence in adult-type granulosa cell tumors of the ovary. *Nat. Commun.* **9**, 2496 (2018).

35. Gala, K. *et al.* KMT2C mediates the estrogen dependence of breast cancer through regulation of ERα enhancer function. *Oncogene* **37**, 4692–4710 (2018).

36. Liu, Y., Sun, J. & Zhao, M. ONGene: A literature-based database for human oncogenes. *J. Genet. Genomics Yi Chuan Xue Bao* **44**, 119–121 (2017).

37. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).

38. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

39. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).

40. Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).

41. Stergachis, A. B. *et al.* Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365–370 (2014).

42. Li, Y. *et al.* Construction and analysis of dynamic transcription factor regulatory networks in the progression of glioma. *Sci. Rep.* **5**, 15953 (2015).

43. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

44. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).

45. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).

46. McCauley, B. S., Weideman, E. P. & Hinman, V. F. A conserved gene regulatory network subcircuit drives different developmental fates in the vegetal pole of highly divergent echinoderm embryos. *Dev. Biol.* **340**, 200–208 (2010).

47. Charlebois, D. A., Balázsi, G. & Kærn, M. Coherent feedforward transcriptional regulatory motifs enhance drug resistance. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **89**, 052708 (2014).

48. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).

49. Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. & Ciriello, G. Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Rep.* **25**, 1066-1080.e8 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure legends

Figure 1. OncoMerge integrates PAMs, fusions, and CNAs into an integrated mutation matrix with the most suitable mutation type for each gene. The input data for OncoMerge includes the PAM, transcript fusion, and CNA matrices. OncoMerge then generates six matrices (PAM, Fusion, CNAamp, CNAdel, Act, and LoF) and uses the mutational frequency and statistical filters to determine each gene's most suitable mutation type.

Figure 2. OncoMerge inferred activating and loss of function mutations overlap significantly with prior knowledge from five independent gold standard datasets. **A.** Impact of filter sets on the number of somatically mutated genes inferred by OncoMerge in at least one cancer. **B.** Impact of filter sets on the distribution of genes per CNA locus using the same set of filtering conditions (y-axis is distributed on a log scale). The dashed line indicates the ten genes per loci cutoff that invokes the MFF filter. **C.** Enrichment of the gold standard (GS) activating (Act) or loss of function (LoF) somatic mutations with OncoMerge (OM) Act or LoF somatic mutations for each filtering condition: no filters (None); permuted q-value filter (PQ); maximum final frequency (MFF); combined PQ and MFF; and combined PQ, MFF, and microsatellite and hypermutation censoring filter (MHC). After Bonferroni multiple hypothesis correction, significant enrichments are highlighted in red (p-value $\leq 4.8 \times 10^{-4}$). The orange arrowheads indicate OM Act vs. GS Act, and the green arrowheads indicate OM LoF vs. GS LoF.

Figure 3. Summary of effect on number and frequency of somatic mutations after integrating mutation types. **A.** Frequency of hypermutation and microsatellite instability across cancers. **B**. Number and distribution of mutation types. **C**. Number of somatically mutated genes with a frequency ≥5% added after integration. **D**. Integrated somatic mutation frequencies. **E**. Increases in somatic mutation frequency relative to PAM frequency after integration.

Figure 4. Pan-cancer somatic mutations with a consistent functional impact across at least five cancers. **A**. Pan-cancer somatic mutations from the loss of functions group. **B**. Pan-cancer somatic mutations from the activating group. **C**. Prior knowledge of tumor suppressor or oncogene status for each somatically mutated gene (black square indicates known tumor suppressor or oncogene activity).

Figure 5. Demonstrating improvements in downstream SYGNAL analysis by comparing GRNs constructed with an OncoMerge integrated somatic mutation matrix versus a legacy network using only PAMs. **A.** Average degree of nodes in the PanCaner SYGNAL networks. OncoMerge = orange, legacy = yellow. **B.** Mutations per cancer network. OncoMerge = red, legacy = blue. **C.** Mutations that overlap with genes previously associated with a specific cancer in DisGeNET. OncoMerge = red, legacy = blue. **D.** TFs per cancer network. OncoMerge = green, legacy = purple. **E.** TFs that overlap with genes previously associated with a specific cancer in DisGeNET. OncoMerge = green, legacy = purple.

Figure 6. The architecture of functional disease-specific TF regulatory networks from human tumors. **A**. Active TF regulatory network construction pipeline: 1) TFs from all cancer regulatory networks were identified, 2) A putative map of TF regulatory network interactions was constructed, 3) TF → TF relationships were filtered using Pearson's correlations computed from patient tumor data, and 4) compute the triad significance profiles using mfinder. **B**. Comparison of active TF regulatory network based on SYGNAL GRNs (red) to the static TF regulatory network based on ENCODE DNA binding and accessibility (blue, Neph et al., 2012). **C**.

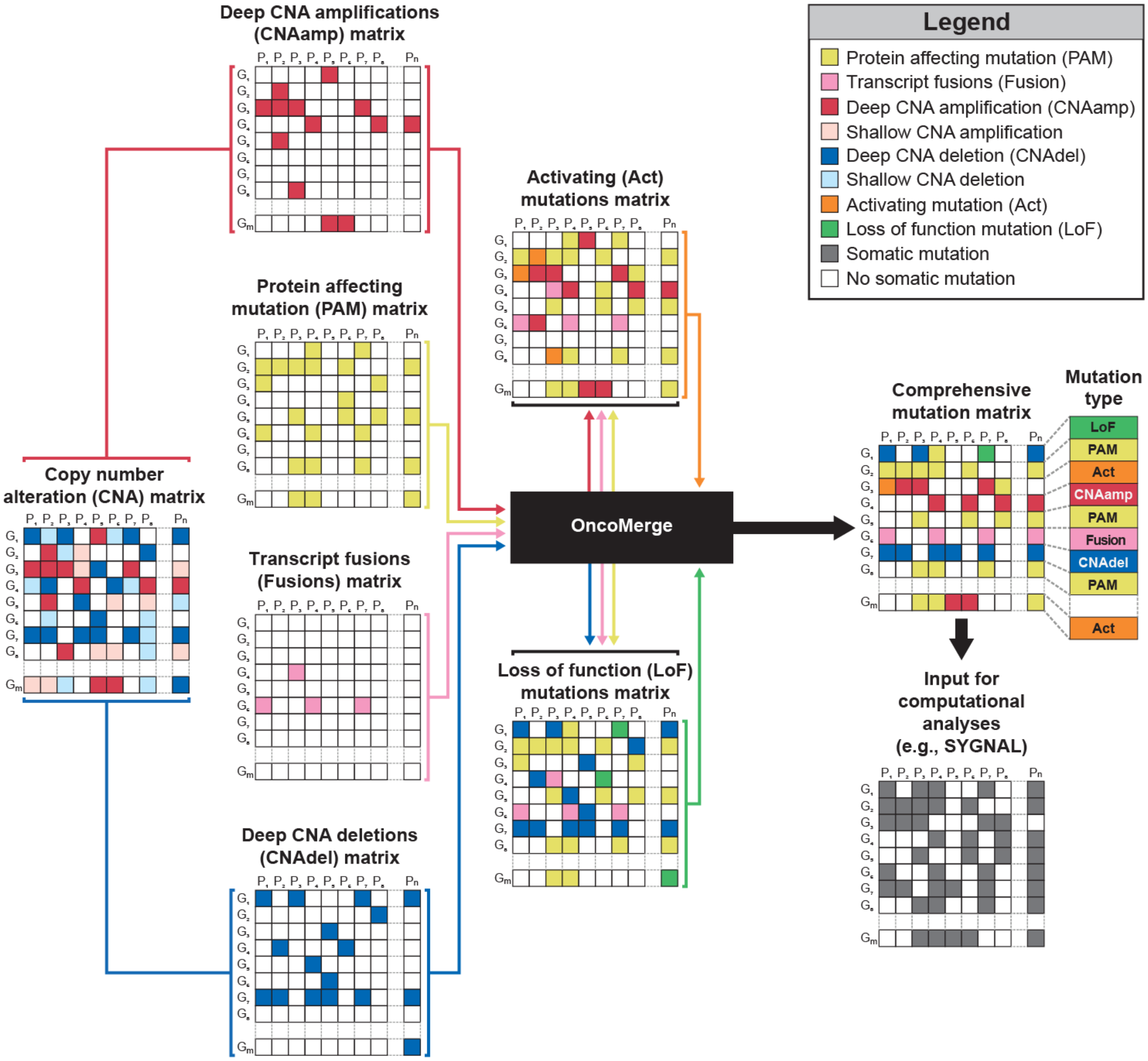
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FANMOD enrichment normalized Z-scores for the three most enriched motifs from the active TF regulatory network after incorporating TF regulatory interaction roles (activation or repression). The first row, titled Coherent motifs, is shaded when the motif configuration is coherent and white when it is incoherent. Normalized Z-scores are reported for each cancer, and diagonal dashed lines are inserted when no Z-score was returned. The network motif can be found at the bottom of each column, colored with regulatory roles (activation = green arrow, repression = red perpendicular line). C1, C2, C3, C4 = coherent FFLs. I1, I2, I3, I4 = incoherent FFLs.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

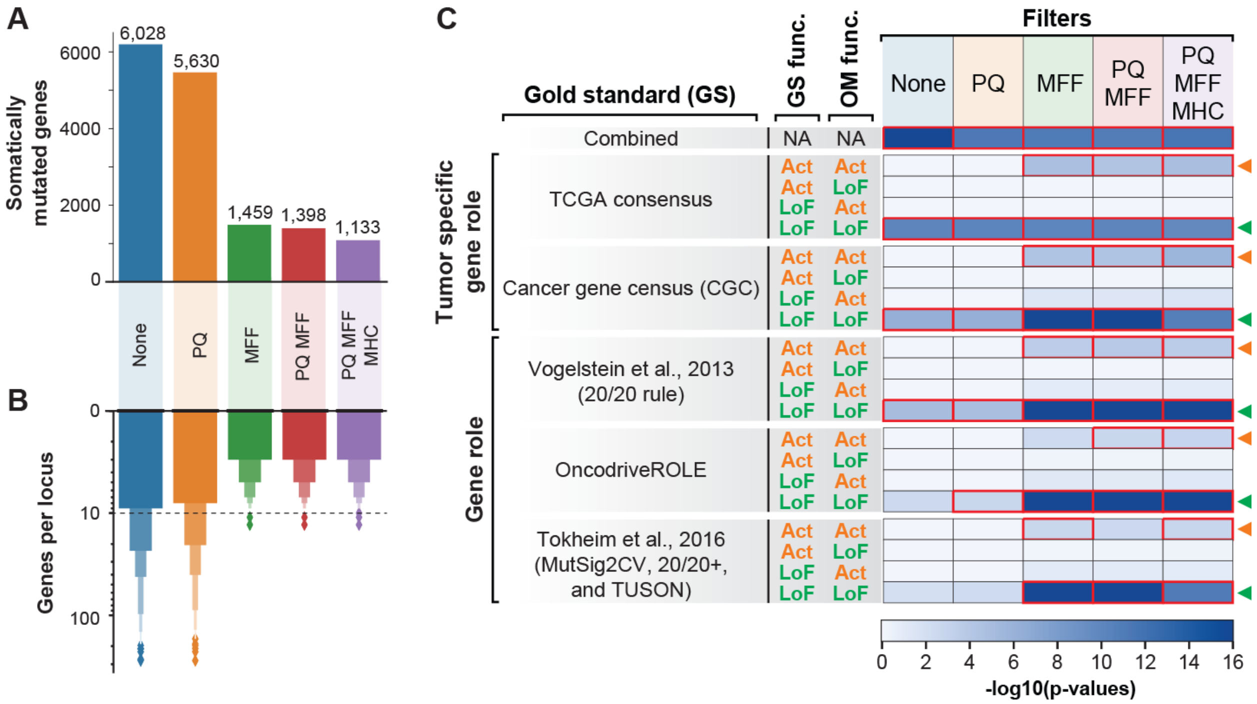
Figures

Figure 1.


Legend
Protein affecting mutation (PAM)
Transcript fusions (Fusion)
Deep CNA amplification (CNAamp)
Shallow CNA amplification
Deep CNA deletion (CNAdel)
Shallow CNA deletion
Activating mutation (Act)
Loss of function mutation (LoF)
Somatic mutation
No somatic mutation
Deep CNA amplifications (CNAamp) matrix
Protein affecting mutation (PAM) matrix
Transcript fusions (Fusions) matrix
Deep CNA deletions (CNAdel) matrix
Copy number alteration (CNA) matrix
Activating (Act) mutations matrix
Loss of function (LoF) mutations matrix
OncoMerge
Comprehensive mutation matrix
Mutation type
LoF
PAM
Act
CNAamp
PAM
Fusion
CNAdel
PAM
Act
Input for computational analyses (e.g., SYGNAL)
P1 P2 P3 P4 P5 P6 P7 P8 Pn
G1 G2 G3 G4 G5 G6 G7 G8 Gm

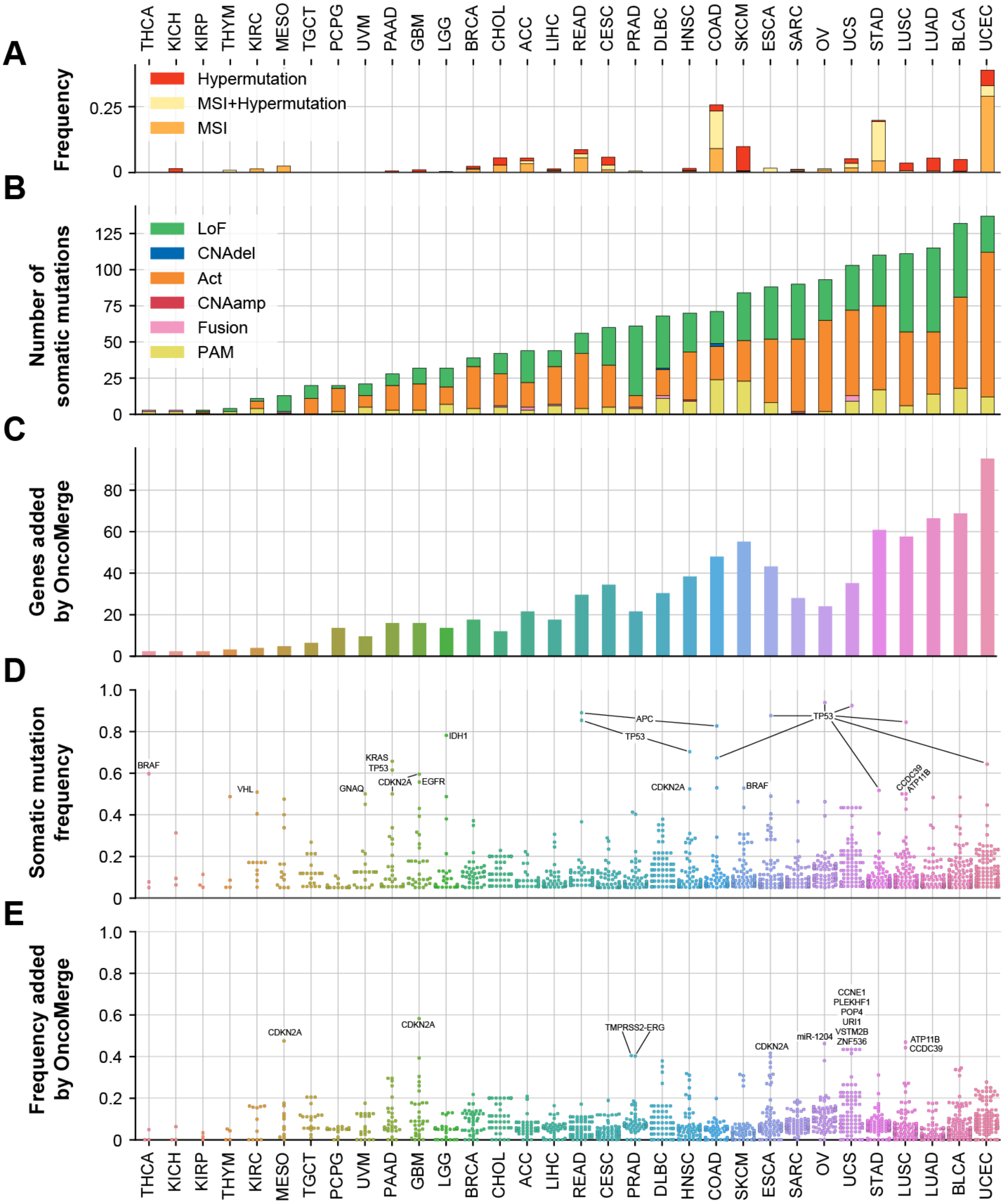
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2.



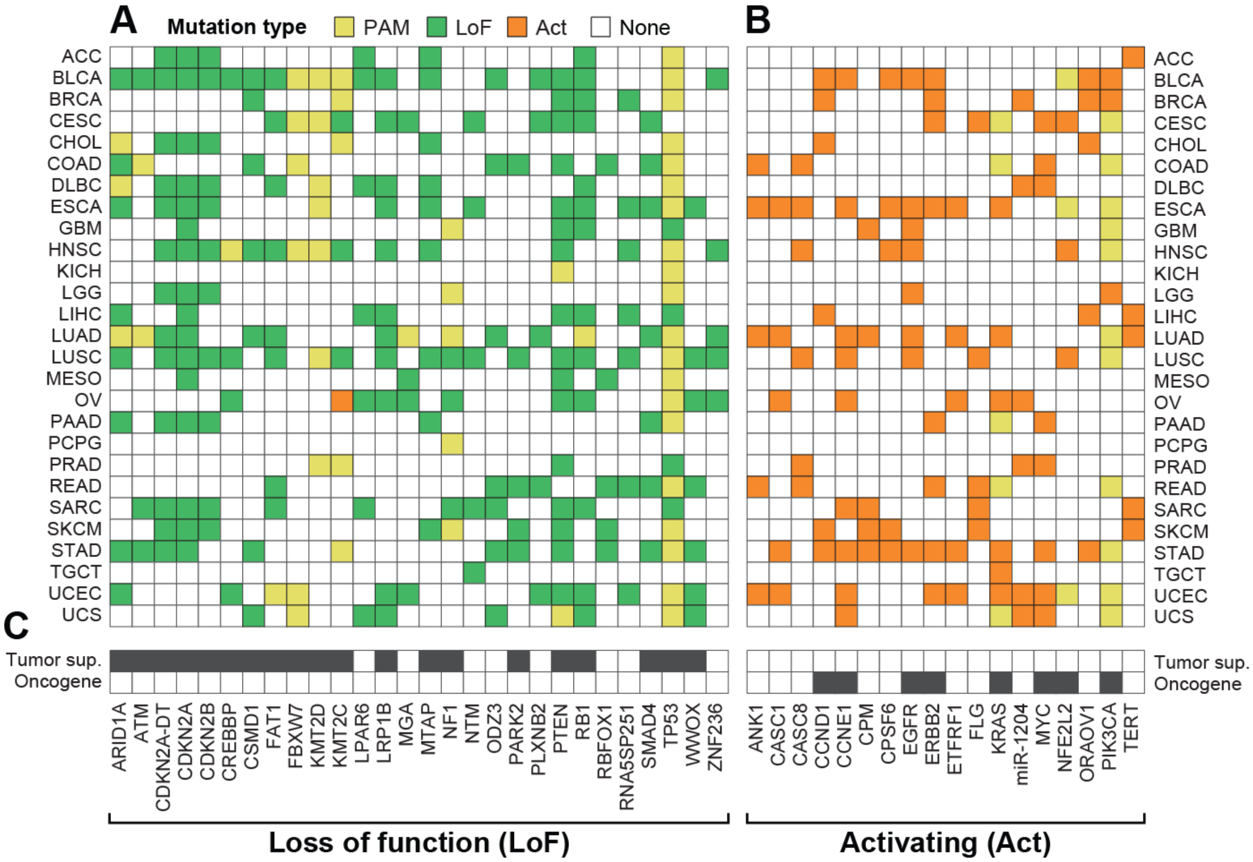
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3.



bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

917 **Figure 4.**

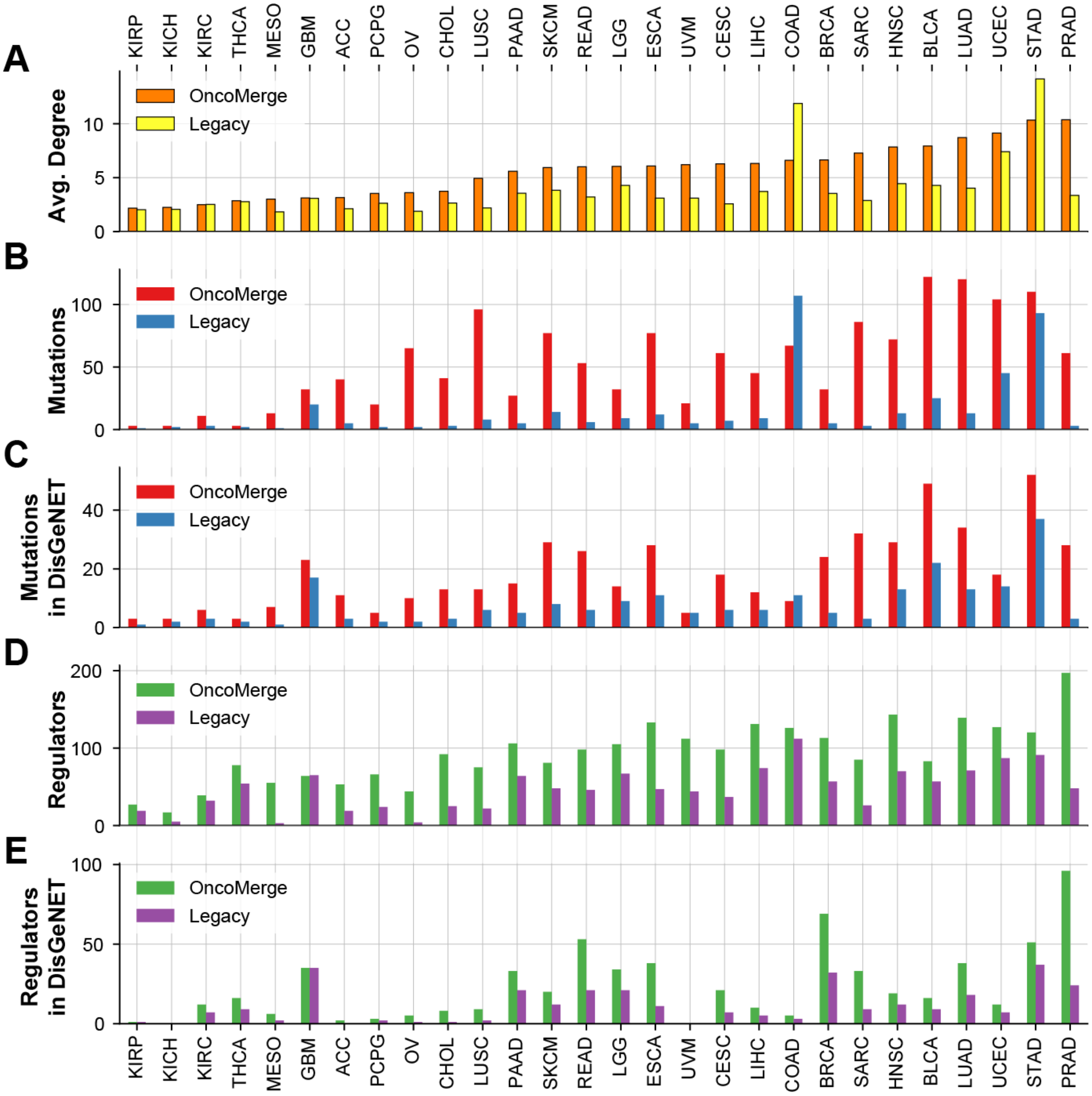


918

919

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

920 **Figure 5.**

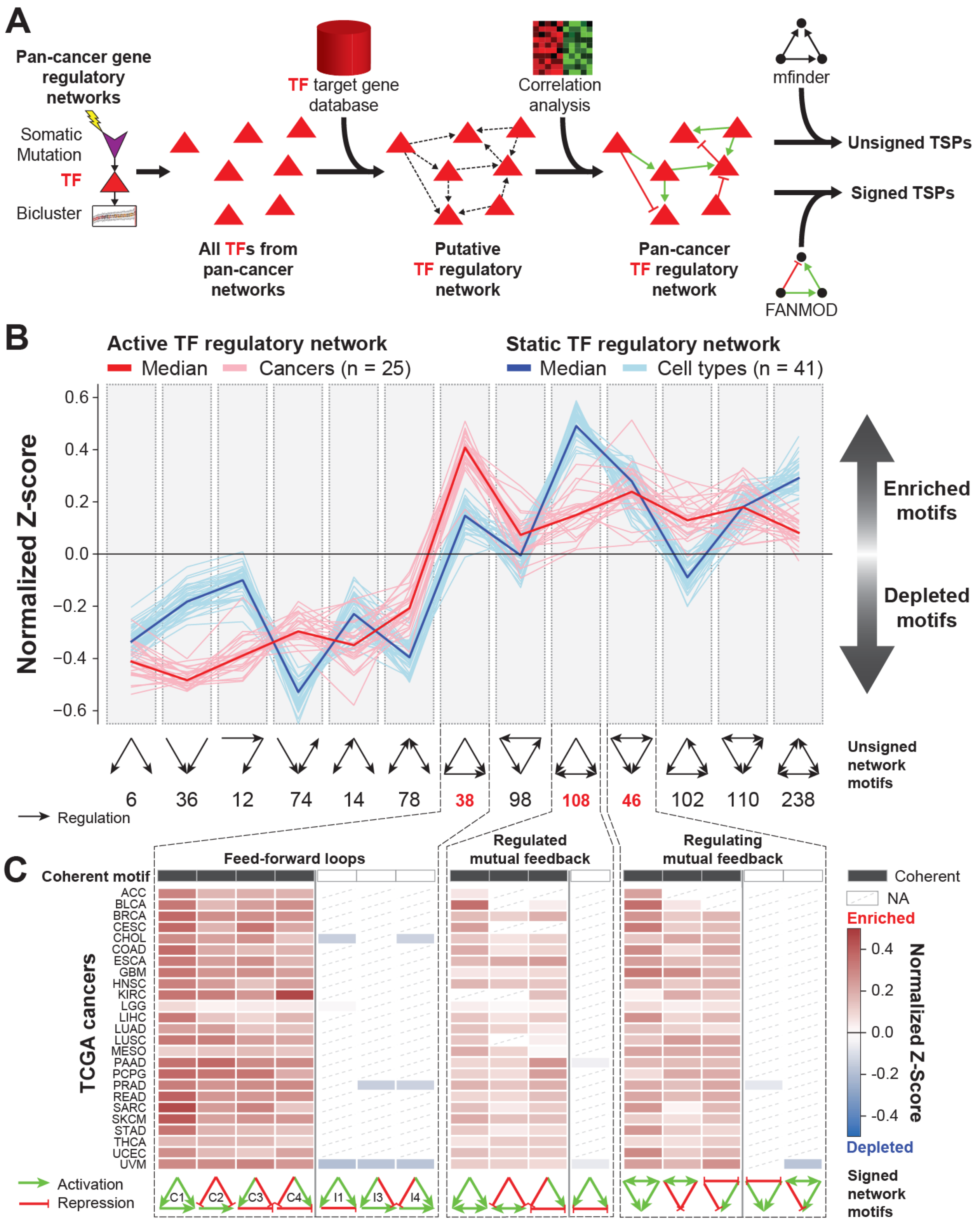


921

922

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

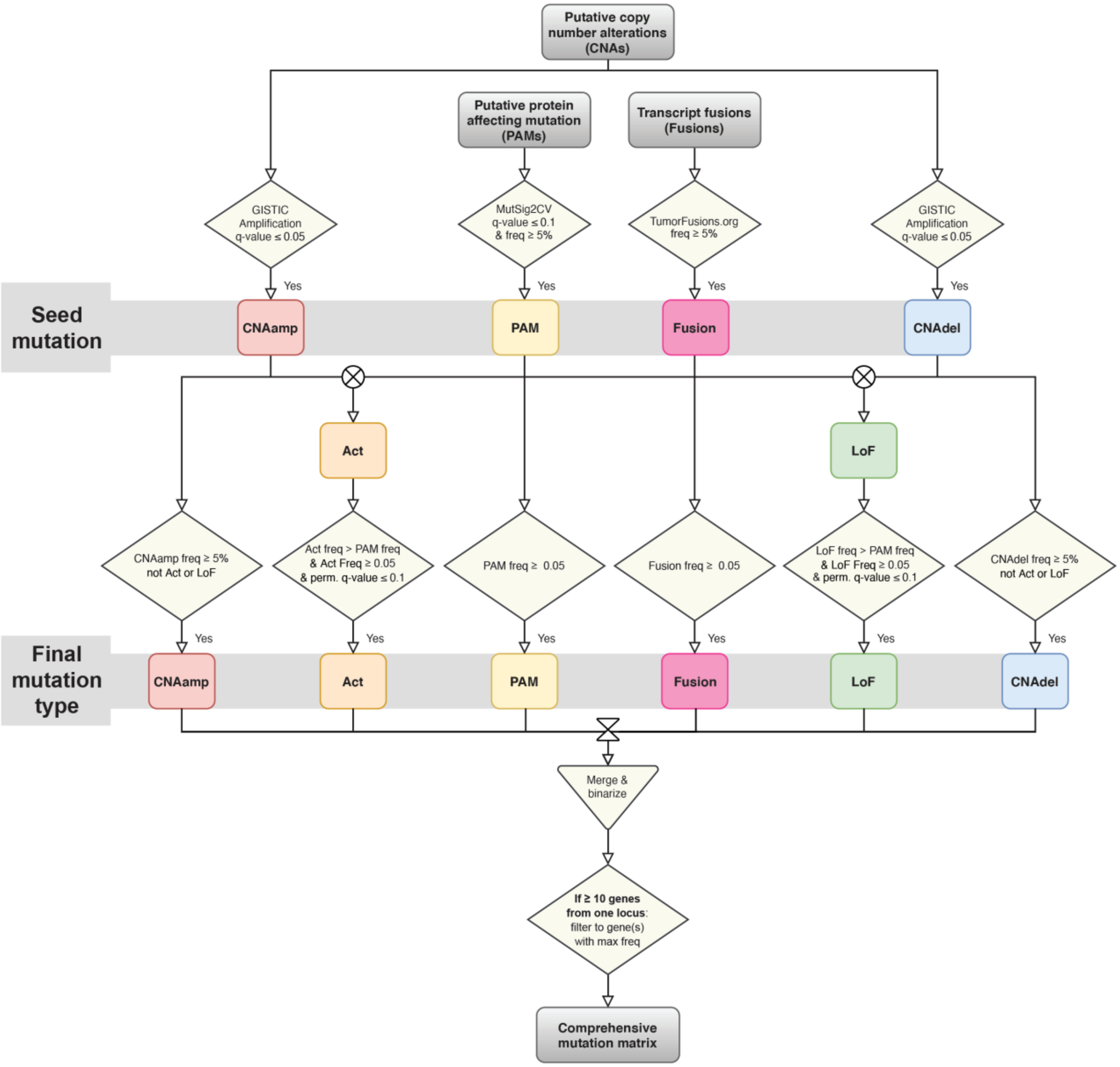
Figure 6.



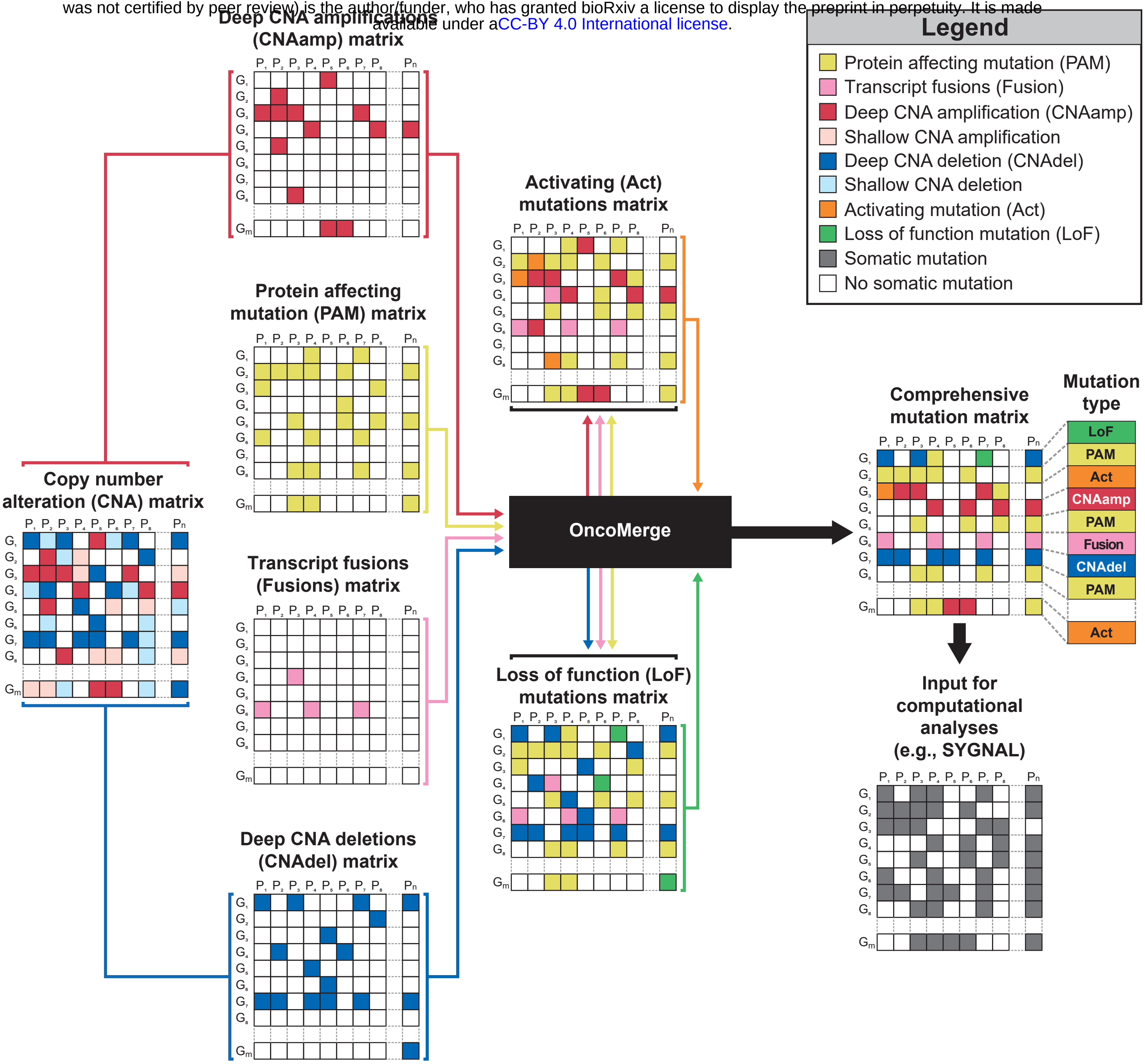
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary figures

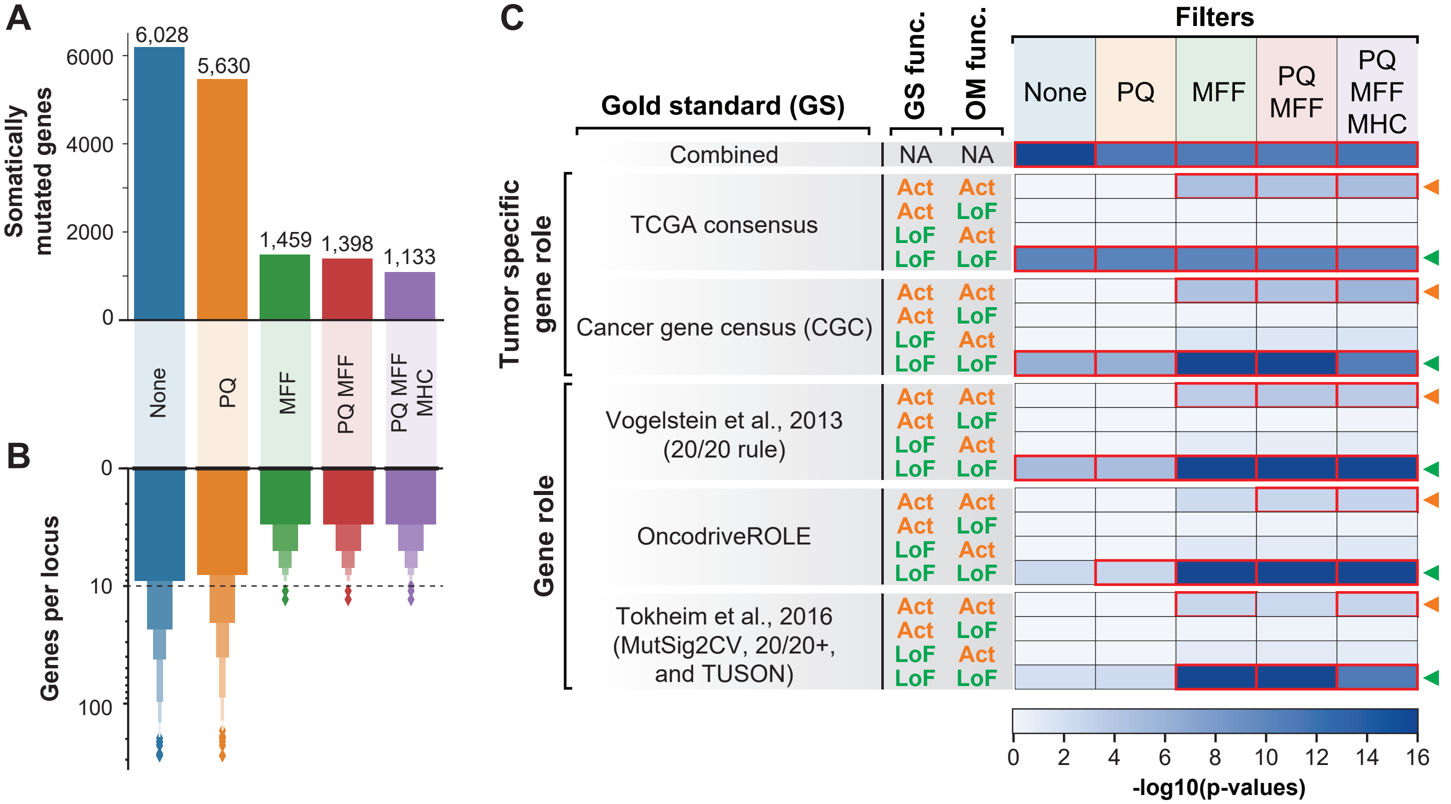
Supplementary Figure 1. OncoMerge flow-chart that describes how the putative protein affecting mutation (PAM), transcript fusions (Fusion), and putative copy number alteration (CNA) data are integrated and filtered to generate a integrated mutation matrix.

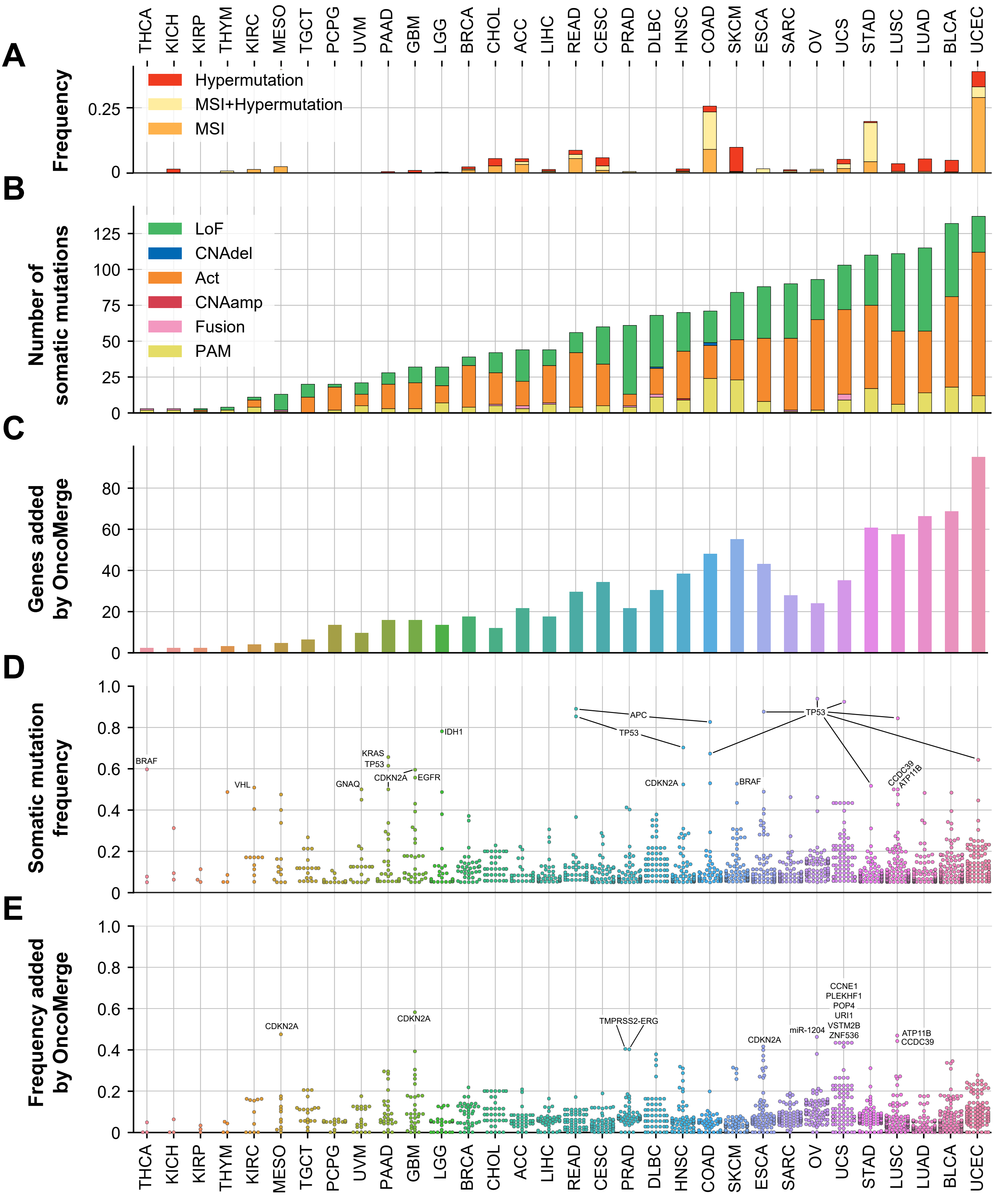

Putative copy number alterations (CNAs)
Putative protein affecting mutation (PAMs)
Transcript fusions (Fusions)
GISTIC Amplification q-value ≤ 0.05
MutSig2CV q-value ≤ 0.1 & freq ≥ 5%
TumorFusions.org freq ≥ 5%
GISTIC Amplification q-value ≤ 0.05
Yes
Yes
Yes
Yes
Seed mutation
CNAamp
PAM
Fusion
CNAdel
Act
LoF
CNAamp freq ≥ 5% not Act or LoF
Act freq > PAM freq & Act Freq ≥ 0.05 & perm. q-value ≤ 0.1
PAM freq ≥ 0.05
Fusion freq ≥ 0.05
LoF freq > PAM freq & LoF Freq ≥ 0.05 & perm. q-value ≤ 0.1
CNAdel freq ≥ 5% not Act or LoF
Yes
Yes
Yes
Yes
Yes
Yes
Final mutation type
CNAamp
Act
PAM
Fusion
LoF
CNAdel
Merge & binarize
If ≥ 10 genes from one locus: filter to gene(s) with max freq
Comprehensive mutation matrix

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

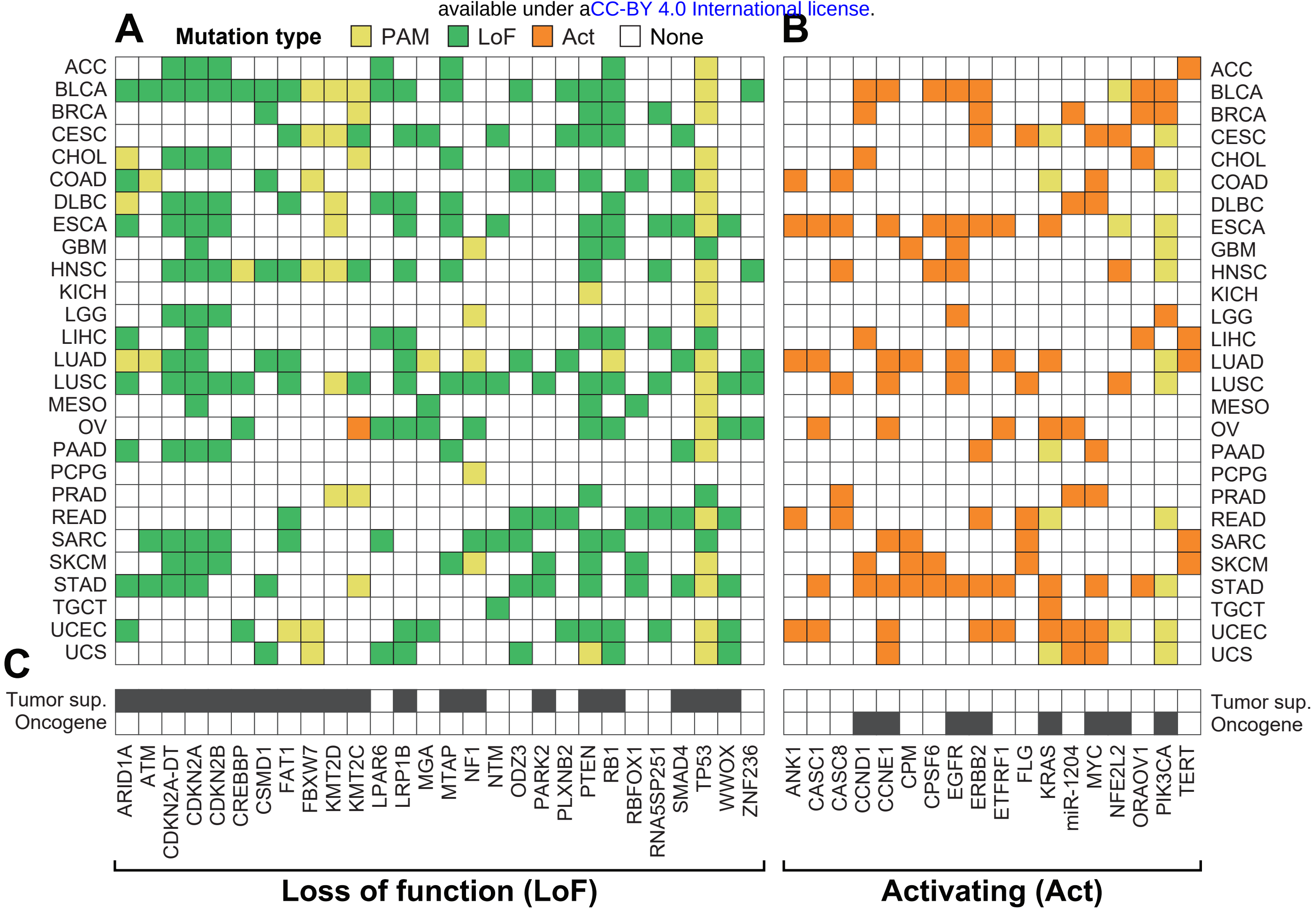

Legend
Protein affecting mutation (PAM)
Transcript fusions (Fusion)
Deep CNA amplification (CNAamp)
Shallow CNA amplification
Deep CNA deletion (CNAdel)
Shallow CNA deletion
Activating mutation (Act)
Loss of function mutation (LoF)
Somatic mutation
No somatic mutation
Deep CNA amplifications (CNAamp) matrix
Copy number alteration (CNA) matrix
Protein affecting mutation (PAM) matrix
Transcript fusions (Fusions) matrix
Deep CNA deletions (CNAdel) matrix
Activating (Act) mutations matrix
Loss of function (LoF) mutations matrix
OncoMerge
Comprehensive mutation matrix
Mutation type
LoF
PAM
Act
CNAamp
PAM
Fusion
CNAdel
PAM
Act
Input for computational analyses (e.g., SYGNAL)
P1 P2 P3 P4 P5 P6 P7 P8 Pn
G1 G2 G3 G4 G5 G6 G7 G8 Gm

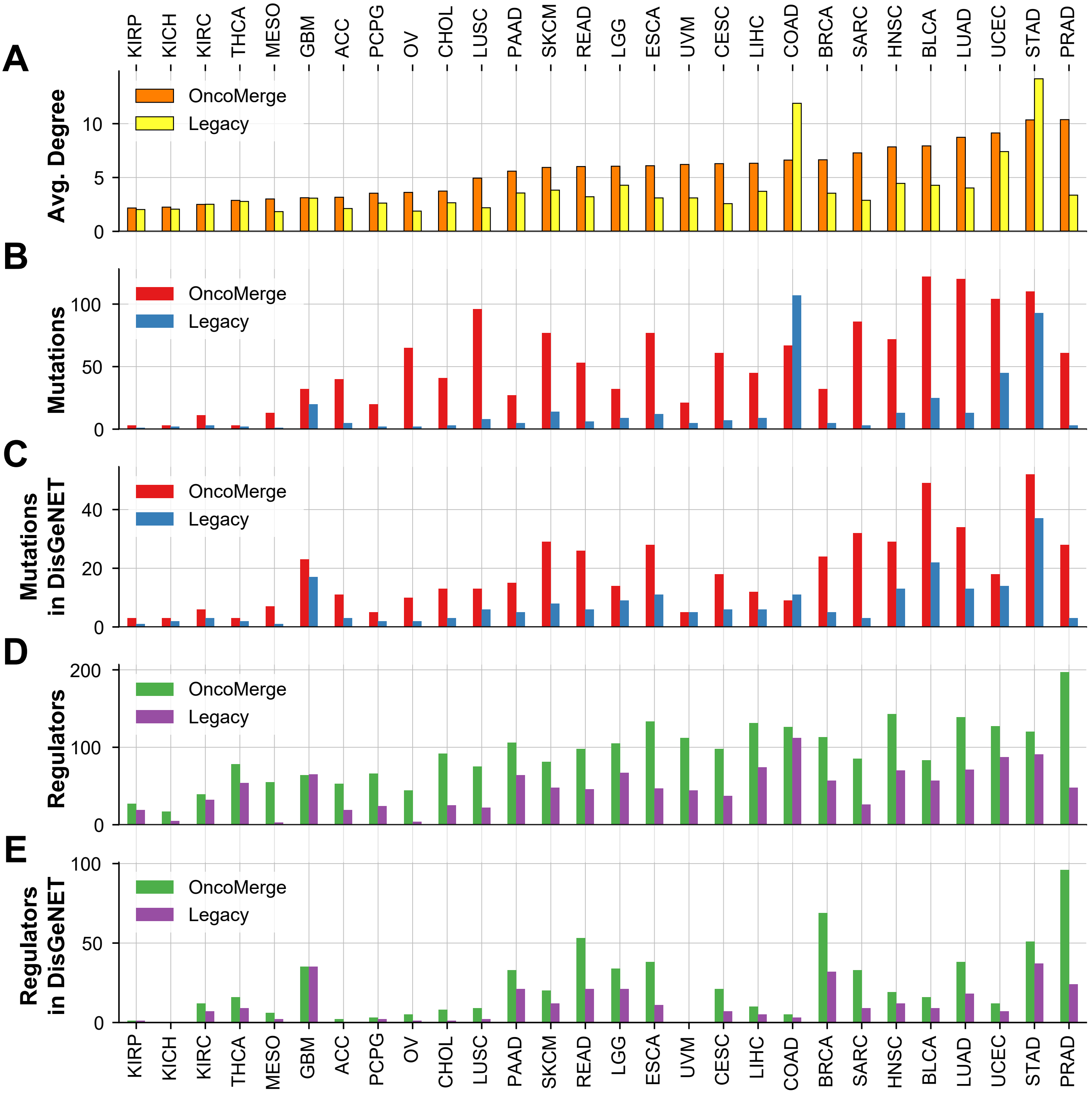
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



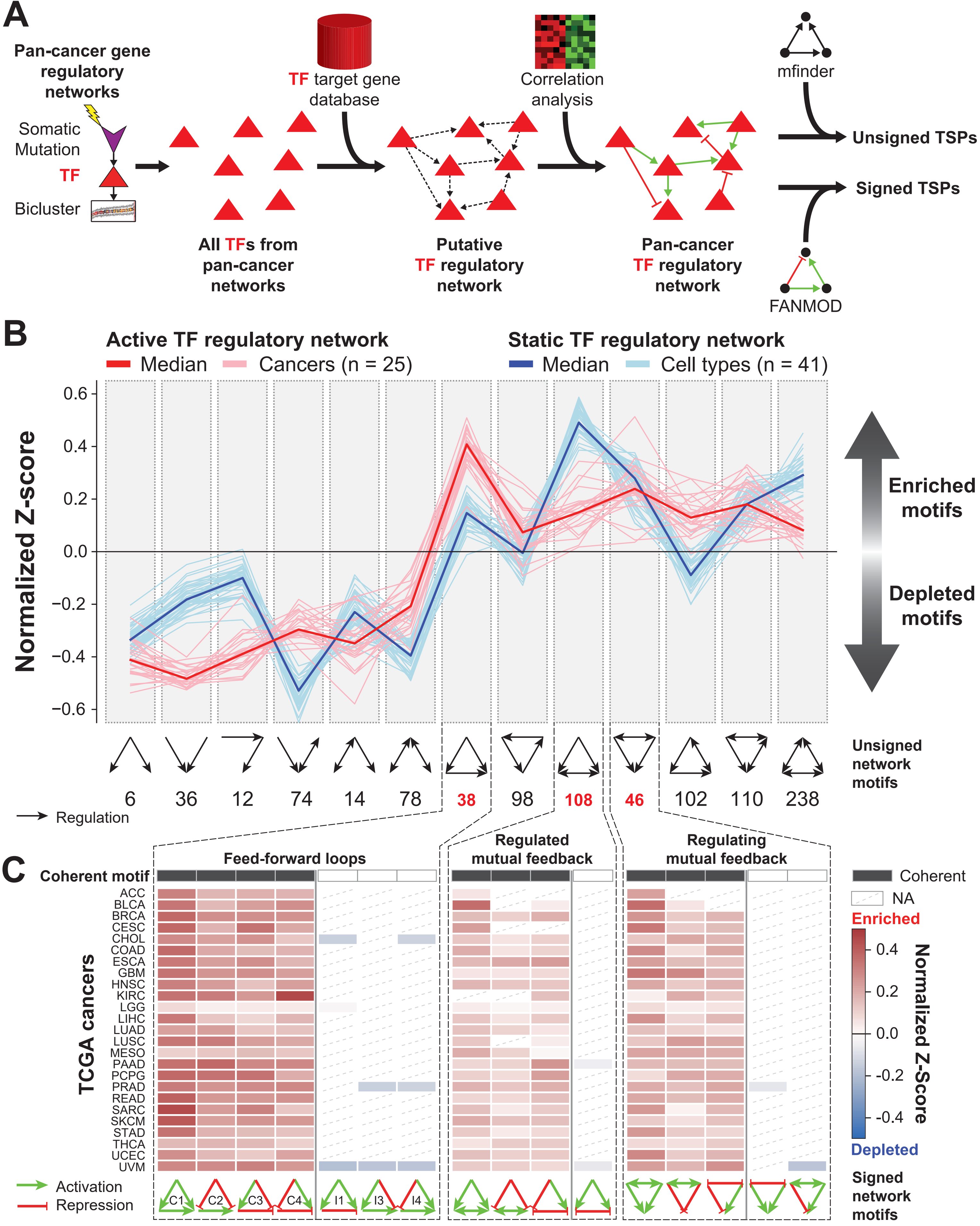

A
B
C
D
E
THCA
KICH
KIRP
THYM
KIRC
MESO
TGCT
PCPG
UVM
PAAD
GBM
LGG
BRCA
CHOL
ACC
LIHC
READ
CESC
PRAD
DLBC
HNSC
COAD
SKCM
ESCA
SARC
OV
UCS
STAD
LUSC
LUAD
BLCA
UCEC
Frequency
0.25
0
Hypermutation
MSI+Hypermutation
MSI
Number of somatic mutations
125
100
75
50
25
0
LoF
CNAdel
Act
CNAamp
Fusion
PAM
Genes added by OncoMerge
80
60
40
20
0
Somatic mutation frequency
1.0
0.8
0.6
0.4
0.2
0
BRAF
VHL
GNAQ
KRAS
TP53
CDKN2A
EGFR
IDH1
APC
TP53
CDKN2A
BRAF
TP53
CCDC39
ATP11B
Frequency added by OncoMerge
1.0
0.8
0.6
0.4
0.2
0
CDKN2A
CDKN2A
TMPRSS2-ERG
CDKN2A
miR-1204
CCNE1
PLEKHF1
POP4
URI1
VSTM2B
ZNF536
ATP11B
CCDC39

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




A
B
C
D
E
KIRP
KICH
KIRC
THCA
MESO
GBM
ACC
PCPG
OV
CHOL
LUSC
PAAD
SKCM
READ
LGG
ESCA
UVM
CESC
LIHC
COAD
BRCA
SARC
HNSC
BLCA
LUAD
UCEC
STAD
PRAD
Avg. Degree
0
5
10
OncoMerge
Legacy
Mutations
0
50
100
Mutations in DisGeNET
0
20
40
Regulators
0
100
200
Regulators in DisGeNET
0
50
100

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.22.501139; this version posted July 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


A
Pan-cancer gene regulatory networks
Somatic Mutation
TF
Bicluster
All TFs from pan-cancer networks
TF target gene database
Putative TF regulatory network
Correlation analysis
Pan-cancer TF regulatory network
mfinder
Unsigned TSPs
Signed TSPs
FANMOD
B
Active TF regulatory network
Median
Cancers (n = 25)
Static TF regulatory network
Median
Cell types (n = 41)
Normalized Z-score
Enriched motifs
Depleted motifs
Regulation
Unsigned network motifs
C
Coherent motif
Feed-forward loops
Regulated mutual feedback
Regulating mutual feedback
TCGA cancers
ACC
BLCA
BRCA
CESC
CHOL
COAD
ESCA
GBM
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
MESO
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
THCA
UCEC
UVM
C1
C2
C3
C4
I1
I3
I4
Activation
Repression
Coherent
NA
Enriched
Depleted
Normalized Z-Score
Signed network motifs