Gene-to-Protein-to-Disease: a segmentation-based approach to identify new cancer prognostic signatures of exome sequence variants

Ginny X.L. Li[1,2], Dan Munro[3], Damian Fermin[4], Christine Vogel[3], Hyungwon Choi[1,5,*]

1. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

2. Saw Swee Hock School of Public Health, National University of Singapore, Singapore

3. Center for Genomics and Systems Biology, Department of Biology, New York University, New York, USA

4. Department of Pediatric Nephrology, University of Michigan Medical School, Ann Arbor, Michigan, USA

5. Institute of Molecular and Cell Biology, Agency for Science, Technology, Research, Singapore

*Corresponding author: Hyungwon Choi (hwchoi@nus.edu.sg)

Abstract

Mutations in protein coding and non-coding regions can be early drivers of tumorigenesis and tumor progression. However, the mutations typically occur at variable positions across individuals, resulting in the data being too sparse to test meaningful associations between variants and phenotypes. To overcome this challenge, we devised a novel approach called Gene-to-Protein-to-Disease (GPD) which uses new sequence units to accumulate variant information through segmentation-based mapping. We found that the variant frequencies *per* sequence unit were highly reproducible between two large cancer cohorts. Survival analysis identified 1,640 sequence units in which somatic mutations had deleterious effects on survival. Significance association was not detected in 1,150 units defined on 1,035 genes by conventional gene-level analysis, providing new prognostic signatures. Importantly, GPD also identified co-dependent sequence variants: units with accumulating somatic mutations impact survival differently depending on the distribution of germline variants in surrounding regions, indicating that a person's genetic predisposition interacts with the effect of later mutations with respect to cancer survival.

Introduction

Accumulation of mutations in the genome can impair cellular function, lead to abnormal proliferation of cells, and eventually cause cancer [1]. Today's massively parallel sequencing technology can detect such variation in the whole genome or exome, and simultaneous analysis of paired tumor and normal samples enables reliable calling of somatic mutations and germline variants. Large-scale sequencing projects such as The Cancer Genome Atlas (TCGA) have successfully mapped the mutation landscape of >11,000 patients across 33 cancer types [2–5]. The logical next step is to test for the association between cancer genome variants and clinical outcomes such as prognosis and response to therapeutics [6–8]. For example, somatic driver mutations are responsible for tumor maintenance, progression, and metastasis [9], indicators of patient survival.

Indeed, mutations at specific loci can make a marked difference in disease prognosis. For example, tumors from papillary thyroid carcinoma patients with the missense mutation V600E in the *BRAF* gene are often associated with more rapid tumor growth and higher death risk than the patients without the mutation [10]. Further, missense mutations affect the transcriptional activity of the tumor suppressor TP53 in different ways, as has been shown in a study of ~1,800 primary breast cancer patients [11]. Mutations found within the DNA-binding motif, especially those at codon 179 and 248, are associated with worse patient survival than mutations at other positions. As such, systematic identification of prognostically informative variants across cancer types remains a major aim of cancer genomics.

For this reason, many algorithms for calling driver mutations have been developed. MuSic [12], OncodriveCLUST [13], and MutSigCV [14] identify the genes that harbor significantly more mutations than others based on cancer-specific background mutations; SIFT [15], FATHMM [16], PolyPhen2 [17], and CHASM [18] identify mutations by evaluating their functional impact; e-Driver [19] and ActiveDriver [9] focus on mutations residing at structurally important sites; and Network-Based Stratification [20], PARADIGM [21], TieDIE [22], and DriverNet [23] prioritize driver mutations using network- and pathway-based

approaches. Even more so, in a massive meta-analysis of the Pan-Cancer Atlas project, Bailey *et al.* [24] applied 26 different computational tools and identified 299 consensus driver genes and ~3,400 mutations across 33 cancer types -- one of the most comprehensive such studies known to-date.

Despite these developments, it remains challenging to test statistical association between mutations and meaningful clinical outcomes such as patient prognosis. One reason behind this challenge is the extreme heterogeneity of tumor genomes [25]. In contrast to studies of single-nucleotide polymorphisms in healthy individuals, somatic tumor mutations rarely reside at the same locus across many patients [26]. As a result, most tumor variants have very low frequency, e.g. on average <0.011% across TCGA Pan-Cancer Atlas patients. Further, the issue is aggravated by varying genome instability across patients, as is typical for cancer cells [27]. As a consequence, analysis of somatic mutations at the locus level inherently lacks the statistical power to detect prognostic markers [28].

A common remedy to this challenge is to count mutations *per* gene, or to classify the mutation status of a gene as 'mutated' or 'not mutated' in each patient [29]. A disadvantage of this approach is that it considers samples with different mutation profiles as carriers of the same mutation burden, regardless of the positions of mutations. A systems-oriented approach called, often referred to as network-based stratification methods, propagates the binary mutation status of genes onto a protein-protein interaction network and identifies interacting genes enriched with adjacent mutation events as etiologically important players [20, 30–32]. While this method more sensitively identifies mutation clusters than locus-based analysis, it ignores the positioning of mutations in the actual gene and subsequently fails to provide directly testable hypotheses on the gain or loss of protein function.

To counter these challenges and enable sensitive identification of prognostic exome variants while preserving information on the location of the mutations, we propose a new approach called Gene-to-Protein-to-Disease (GPD), which maps somatic mutations and germline variants onto sequence units derived for protein coding or non-coding regions. GPD splits proteins into functionally interpretable segments which comprise domains, linker regions,

and small sequence windows with post-translational modification sites. GPD then accumulates somatic mutations and germline variants within these different sequence units. It therefore transforms the original variant data into substantially less sparse count data and enables statistical analysis for disease prognosis. GPD outperforms conventional locus- and gene-level analyses with respect to resolution and sensitivity. The approach successfully recovers a handful of cancer driver mutations as prognostic markers, and also enables identification of interactions between the effects of somatic mutations and germline variants on survival outcomes.

Methods

Data Source. Exome sequencing data for somatic mutations in TCGA datasets were downloaded from the Genomic Data Commons (MC3 Public MAF) [3]. Germline variant data were downloaded from the ISB-CGC web interface, where the VCF file is deposited [4]. We further annotated the file with the snpEff tool (genome version GRch37.75) [33]. In addition, we gathered 348,658 protein modification sites (244,852 phosphorylation, 46,352 ubiquitination, 25,807 acetylation, 20,691 methylation, 8,388 SUMOylation, and 2,568 glycosylation sites) from PhosphoSitePlus [34], and 45,607 domains, families and repeats for 19,076 genes from Pfam [35].

GPD segments exons and adjacent genomic regions of each gene into three types of units (Figure 1A). First, Protein Information Units (PIU) are the genomic regions encoding protein domains (e.g., as defined by the Pfam database[35]) or +/-5 amino acid-long windows around protein modification sites (e.g., as reported in the PhosphoSitePlus database). Sequence regions between PIUs are defined as linker units (LU). The linker units not only include linker regions between domains, but also cover unannotated, repeat or disordered regions. The regions outside the protein-coding sequences, including untranslated regions, introns, and regulatory regions are collectively defined as non-protein coding units (NCU). Similar to linkers, non-coding units are gene-specific, i.e. these regions are assigned to the closest gene in the genome.

Survival outcome data were downloaded from TCGA Pan-Cancer Clinical Data Resource [5], which contains clinical information for 10,793 patients (367 non-primary skin cutaneous melanoma patients with metastatic tumors were excluded). The data resource provides four types of clinical endpoints, including overall and disease-specific survival, as well as disease-free and progression-free time interval. For the purpose of consistency, we used overall survival as the endpoint of interest in this study. Abbreviations for cancer types have been adapted from TCGA available at https://gdc.cancer.gov/resource-tcga-users/tcga-code-table/tcga-study-abbreviations.

Unsupervised analysis. GPD produces a feature matrix of dimension *n* by *p* in each cancer type, where *n* and *p* refer to the number of patients and the number of eligible sequence units, respectively. We applied principal component analysis to these data matrices and recorded principal component scores of each patient on all principal components. We then divided the subjects into groups of high and low scores (median as threshold) in the principal component most strongly associated with overall survival. In **Supplementary Figures 1, 2, 8** and **9**, we use the principal component for which the scores are the most significantly associated with overall survival in each cancer type.

Selection of prognostic sequence units (main effect model). The association between mutation count of each sequence unit and overall survival was evaluated using a Cox proportional hazard model. To focus on the association with long-term survival outcome, we excluded patients with less than 90 days survival time, which may confound the association analysis due to variable short-term response to treatment. For each cancer, sequence units with mutations in three or more patients were selected for model fitting. Age, gender and race were adjusted in all models (when available). The Cox model posits the hazard function as a product of the baseline risk function at time *t* and the relative risk based on covariates. For each sequence unit or gene, the hazard function for subject *i* is given by:

$$\lambda(t|X_i) = \lambda_0(t)exp(X_i \cdot \beta)$$

where $\lambda_0(t)$ denotes the baseline hazard function and $X_i = \{Count_i, Age_i, Gender_i, Race_i\}$ are the covariates. The models were fitted for somatic mutations and germline variants, separately. For comparison, we repeated the analysis but counted mutations *per* gene (gene-level analysis), instead of *per* sequence unit. For each model, the proportional hazard assumption was tested by evaluating the association between transformed survival time and the scaled Schoenfeld residuals. Genes or units in which the assumption was significantly violated (Chi-squared, *p*-value < 0.05) were removed from the list of candidate prognostic markers.

Identification of interactions between effects. The hazard models accounting for the interaction between somatic mutations and germline variants consider three main covariates: the somatic mutation count of a unit, the total count of germline variants in the same gene, and the interaction between the two. Similar to the main effect models above, models were adjusted for gender, age and ethnicity. The hazard function is identical to the above, but X_i denotes a vector of covariates as:

$$X_i = \{Count_{i_{somaticUnit}}, Count_{i_{germlineAllUnits}}, Count_{i_{somaticUnit}} * Count_{i_{germlineAllUnits}}, Age_i, Gender_i, Race_i\}.$$

For the test of interaction, the model is fitted for a selection of sequence units only: the unit must have positive somatic mutation counts in ≥ 10 subjects and positive germline variants in the gene containing the unit.

Cancer pathway enrichment analysis. For each cancer, we chose all genes containing one or more units eligible for the association analysis. We downloaded 209 cancer-related pathways using the Enrichr library [36]. We then used the hypergeometric test to compute the significance score for the enrichment of a cancer pathway in the genes containing survival-associated sequence units.

GPD software implementation. The GPD framework was developed as an R package. The package is publicly available at https://github.com/ginnyintifa/GPD, with user-defined GPD units and control parameters for mapping and analysis. A user manual is available at the GitHub site. The GPD implementation maps mutations to sequence units. The default genomic coordinates of protein domains and protein modification sites are provided in the package, but the user can also provide the information from other sources in the specified format. Users need to provide mutation data.

Results

GPD maps variants to sequence units in exons and adjacent genomic regions

Figure 1A illustrates GPD's sequence units by mapping hypothetical somatic mutation data of a patient to a gene: two mutations reside within a protein domain, one mutation near a protein modification site, two mutations in a linker unit, and three mutations in the non-coding unit. Using the protein domains from Pfam database and the protein modification sites from PhosphoSitePlus database as PIUs, GPD mapped somatic mutations to a total of 270,470, 17,252, and 17,266 PIUs, linker, and non-coding regions across 17,744 human genes, respectively.

Next, we counted for each sequence unit the coding and non-coding variants per tumor sample as derived from TCGA. For PIUs we considered only non-synonymous variants and found that different cancers showed unique somatic mutation frequencies. For example, uterine cancer (UCEC) had on average >1,000 mutations *per* patient; skin and colon cancers (SKCM and COAD) had on average >500 mutations *per* patient. By contrast, the exomes in many other cancers (e.g. KICH, UVM, TGCT, THCA and PCPG) had an average of fewer than 30 somatic mutations per patient.

The overall distribution of somatic mutations in the exomes of individuals was remarkably sparse and heterogeneous. Less than 8% of the somatic mutations appeared at the same locus in at least two patients (**Figure 1B**). The sparsity in the data was significantly reduced when accumulating mutations *per* gene or *per* sequence unit: while 99% of the genes showed mutations, we found 68%, 96%, and 98% of PIUs, linker units, and non-coding units, respectively, mutated in at least two patients. The variable mutation frequencies across different sequence units within the same genes suggest that GPD's approach provides a higher-resolution map than gene-level analyses, as it can focus on smaller sequence areas while still preserving coverage and statistical power to compare patient groups.

When examining the distribution of mutations across genes, we noticed that almost half the somatic mutations fell within PIUs (49%), one third fell within linker units (32%), and

19% were in non-coding units (**Figure 1C**). In the current definition of GPD's sequence units, 67% of the coding sequence is covered by PIUs and 33% is covered by linker units. As such, within the protein coding regions, both unit types have about equal overall chances to carry a mutation: the expected frequency of a mutation is 0.17 and 0.16 for PIUs and linker regions, respectively.

In contrast, the observed frequencies were different per unit type in germline variants: as many as 88% were found in the non-coding units and only 6% resided in PIUs and linker units. The results imply that genetic predisposition of a person appears to be primarily encoded in non-coding regions that may affect expression levels rather than structure and function of the resulting protein.

Figure 1D shows the PIUs with the largest number of somatic mutations in TCGA patients. These PIUs are located in well-known cancer genes. For example, the PIU for the DNA-binding domain of *TP53* gene showed consistent somatic mutations across several cancers. Other mutations surrounded the serine phosphorylation site S269 which is crucial for the conformational stability of the protein [37]. In another gene, *KRAS*, we find the highest frequency of somatic mutations in the PIU for the RAS domain in pancreas, lung, rectum, and colon cancer (PAAD, LUAD, READ, and COAD). The RAS domain also accumulates most mutations in the *NRAS* gene in 103 non-metastatic skin cancer (SKCM). Further, we find frequent mutations in thyroid, skin, and colon cancers (THCA, SKCM, and COAD) in a region in the *BRAF* gene which is densely populated with phosphorylation and ubiquitination sites (T599, K601, S602, S605, and K601). Similarly, the cancer gene *PTEN* accumulated mutations around acetylation sites (lysine) K125 and K128. **Supplementary Table 1** lists the sequence units with high mutation rates in specific cancers.

Next, to assess the reproducibility of the unit-level mapping, we validated the above results with somatic mutation data for nine cancers from an independent cohort (MSK-IMPACT) [38]. Due to the lack of coverage of mutations in non-coding regions in MSK-IMPACT, we mapped variants to PIUs and linker units only. We found that the somatic mutation frequencies for the sequence units are highly reproducible between the TCGA and MSK-IMPACT cohorts

(**Figure 1E**). Therefore, we suggest that somatic mutations have propensity to be enriched in specific GPD units within each gene.

Somatic mutations across sequence units reveal prognostic signatures

Next, we tested whether the total mutation counts from select sequence units enable discovery of prognostic signatures, i.e. sets of sequence units or genes harboring mutations that can prognosticate outcomes of specific cancers. While we did not expect a single mutation or sequence unit to predict the clinical outcome, we aimed at assembling a panel of genes accumulating somatic mutations with some of their sequence units that show prognostic potential. We first used principal component analysis to construct orthogonal axes of large variation across the mutation data mapped to sequence units. We then tested whether any principal component is associated with the overall patient survival within each cancer.

To do so, we used Cox proportional hazard models (see **Methods**). In each cancer, we stratified the patients into groups of high and low principal component scores, respectively, using the median as threshold, and tested the difference in the survival curves. Such analysis was not feasible at the locus level analysis due to extreme sparsity of data at nearly all loci.

In 24 out of 33 cancer types, we found one or more principal components being statistically significantly associated with overall survival (**Supplementary Figure 1**): high and low scoring groups for the most significantly associated principal component showed clear separation at p-value < 0.01 (12 cancers) and < 0.05 (24 cancers, Kaplan-Meier curves) for the log-rank test. This separation of the principal component scores derived from sequence unit data was better for the majority of cancer types than when using gene level data (**Supplementary Figures 2 and 3A**).

Similarly, mutation counts of many sequence units showed better prognostic potential than whole-gene mutation counts. In fact, patients with identical mutation counts at the gene level could have distinct survival outcomes depending on whether the mutation lied in a critical functional unit (e.g. protein domain) or not, especially in cancers with abundant somatic mutation events, such as bladder, colon, and lung cancers (BLCA, COAD, LUAD and LUSC).

For example, **Supplementary Figure 4** shows an example of patients grouped according to the binary mutation status of *TP53* gene, who do not show significant survival difference. However, when we compared survival curves between the group with mutation in the vicinity of ubiquitination site K132 and methylation sites R110 and the group without any mutation, the former group had a substantially worse survival curve. Similarly, ubiquitination site K860 in *EGFR* in lung cancer (LUAD) patients also has a similar prognostic potential whereas mutation status of the entire gene does not. MH1 and MH2 domains in *SMAD4* may also be prognostic for colon cancer (COAD) patients, and linker units in many genes appear to be more relevant than the whole gene to survival in many cancers.

Somatic mutation signatures associated with survival are highly cancer specific

To identify sequence units harboring somatic mutations with prognostic potential, we tested each unit for positive association with overall survival in Cox proportional hazard models. Across the 33 cancer types, we discovered 695, 571, and 374 unique PIUs, linker and non-coding regions, respectively that were associated with overall survival in at least one cancer type (*q*-value 0.1). The most such survival signatures lied in breast cancer (264 PIUs, 261 linker units and 133 non-coding units), followed by colon, liver, and lung cancer (COAD, LIHC, and LUSC and LUAD; **Figure 2A**).

GPD's unit-level analysis discovered survival associated units on genes which were undetected by the analysis at the gene level. Conventional gene-level analysis discovered 1,395 genes to be significantly associated with survival, while our analysis at the sequence unit level identified 1,526 genes bearing at least one associated sequence unit. However, only 449 genes are shared between the two analyses. The 1,150 survival associated sequence units in the remaining 1,035 genes would not have been identified without GPD analysis (**Supplementary Table 3**). Examples of these cases include SNF family in the *SLC6A5* gene in head and neck cancer (HNSC) patients, V-set domain in the *MCAM* gene in lung cancer (LUAD) patients, and tyrosine phosphorylation site Y355 in the *PIK3CA* gene in colon cancer (COAD) patients.

Notably, most significant associations with survival were specific to one cancer type: only twelve of the 695 significant PIUs were significantly associated with overall survival in two or more cancer types (**Supplementary Table 2**). The result indicates that the impact of somatic mutations on prognosis differs by cancer type and only few mutations are shared across cancers. The shared PIUs include the P53 domain in *TP53*, the C2 PI3K-type domain in *PIK3CA*, the Actin domain in *ACTRT3*, and the ERM family motif in *NF2,* linker regions in 18 genes, e.g. *ANK3*, *DOCK6*, *ASPA1*, and *INTS1*, and non-coding regions in 11 genes, e.g. *ABCA11P*, *CRQ*, *TPD52*, and *TRPS1.* Expectedly, these genes include many pan-cancer genes, i.e. known tumor suppressors or oncogenes implicated in multiple cancer types.

To compare the ability of sequence unit- and gene-level analysis to stratify different risk groups, we next clustered patients into good and poor survival outcome groups considering the total mutation counts of all selected sequence units in both analyses (see **Methods**). This analysis was applied to 25 cancer types, excluding cancers without statistically significant units (CHOL, DLBC, KICH, PAAD, TGCT, THCA, UCS and UVM). As expected, the disease prognosis was clearly separable between the two groups in each cancer when the data was analysed at the unit level (**Figure 2B** and **Supplementary Figure 5**; log-rank test p-value < 0.05). When we repeated the same analysis using gene-level mutation counts, survival curves were significantly separated in many cancer types, similar to unit-level analysis. In ten cancer types (LGG, SARC, CESC, ESCA, MESO, COAD, PRAD, BRCA, GBM, and THYM), however, the unit-level analysis led to better stratification of survival risk groups than the gene-level analysis (Kaplan-Meier curves in dashed line in **Figure 2B**, -log10 p-values compared in **Figure 2C**).

Survival associated GPD units provide cues to impaired protein functions

A major advantage of GPD lies in its ability to provide insights into the functional impact of mutations. For example, *SMAD4* is a well-known tumor suppressor that plays a critical role in the transforming growth factor beta (TGF-beta) signalling pathway. Sixty-one somatic mutations were reported on this gene in colon cancer (COAD, n=413), including 44 missense,

nine frameshift, six nonsense, one in-frame deletion and one intron mutations (**Figure 3**). About 69% (42 of 61) of these mutations were mapped onto the PIUs for the two domains of SMAD4 protein, namely MH1 and MH2. MH1 domain is a DNA binding domain and MH2 domain is mainly implicated in transcriptional activation and homo- and heteromeric complex formation among the SMAD proteins [39]. Sixteen out of 27 protein modification sites of SMAD4 resided on these two domains. Notably, 68% (30 of 44) of total missense mutations occurred in the MH2 domain, which represents only 38% of the protein sequence length -- illustrating the existence of mutational hotspots around the modification sites. We found such a hotspot in a PIU with 12 missense mutations around position R361. Further, deceased patients with shorter survival time bore more frequent mutations in the MH1 domain, MH2 domain and a window covering ubiquitination or SUMOylation site K113 than patients with longer follow-up or deceased patients with longer survival time (**Supplementary Figure 6**). This association was consistent with the results from the Cox proportional hazard model, which shows significant deleterious effect of the mutations in MH1 domain (q-value = 0.0005). The results also agreed with a previous study showing that missense mutations in MH1 can alter protein stability *in silico* [40].

Another example is *ERBB4*, a known oncogene which encodes a receptor tyrosine-protein kinase. This enzyme is responsible for signal reception from neuregulins and EGF family members on the cell surface [41]. The ErbB4 protein contains several domains which carry the signal transduction function such as the Furin-like and the kinase (Pkinase-tyrosine) domain. We found a cluster of tyrosine phosphorylation sites on the tail of the protein that are phosphorylated upon ligand binding. GPD mapped a total of 16 missense mutations and 1 deletion from the 767 ER positive breast cancer patients to the sequence units in this gene (**Figure 3**). Three missense mutations and the frameshift deletion lie within the Furin-like domain (p-value = 0.026, q-value = 0.199), together with five missense mutations in nearby linker units (p-value = 0.005, q-value = 0.051). These mutations associated with patients with shorter survival time, suggesting that the mutations alter ErbB4 function through damaged cysteine-rich regions in the Furin-like domain. Such interpretation is consistent with findings in

an independent ER-positive breast cancer cohort which demonstrated that the risk of death increases in patients with ERBB4 mutations [[42].

Survival-associated sequence units overlap with known driver mutations

We next benchmarked survival-associated sequence units against known driver mutations and their corresponding genes (**Supplementary Tables 4 & 5**). To do so, we used the 299 cancer driver genes and ~3,400 driver mutations identified *in silico* by Bailey *et al.*, who employed a multitude of software tools to obtain the consensus list with data from the TCGA Pan-Cancer Atlas cohort [24].

We identified multiple PIUs and one linker region within 13 known driver genes, including *ERBB2*, *EGFR*, *TP53*, and *HIST1H1E*. Further, using position information for the core set of 579 missense driver mutations we mapped the mutations to sequence units. Over 232 (40%) of the core mutations localized to 12 prognostic signatures which include Furin-like domain and Receptor L-domain in *EGFR*, Furin-like domain in *ERBB2*, phosphorylation site Y220 and the DNA binding domain in *TP53*, a glycosylation site at T217 in *CDH1*, C2 PI3K-type domain in *PIK3CA*, and Peptidase C14 domain in *CASP8*.

As a further validation of GPD's predictive abilities, we examined cancer pathways that were enriched in the genes containing significant units (p-value < 0.05) (see **Methods**). We identified 40 pathways enriched in these genes in at least four cancer types (**Supplementary Figure 7**). For example, integrin signalling and integrin cell surface interaction were most frequently enriched in multiple cancers (LUAD, LUSC, CESC, HNSC, SKCM, and COAD).

Germline variant mapping to sequence units unveils a different landscape of prognostic signatures

Next, we mapped >467 million germline variants to GPD units in 17,872 genes and performed the same association analysis as for the somatic mutations. While the absolute number of germline variants was substantially larger than that of somatic mutations, germline variants appeared mostly in non-coding regions and had comparatively few variants in protein-coding

regions (**Figure 1C**). On average, a patient in the Pan-Cancer Atlas cohort had ~43,000 germline variants in and near the exome.

Although the majority of cancers showed a similar distribution of germline variants, some cancers tended to have more germline variants *per* patient than others (**Supplementary Figure 8**). This bias suggests that the genetic predisposition is partially cancer-specific. For example, colon and rectum cancer patients had the largest number of variants (COAD, READ; median count ~80,000), whereas skin, paraganglia, uvea, kidney, and brain cancer patients had much fewer (SKCM, PCPG, UVM, KICH, and GBM; median count ~30,000). This difference may be due to technical reasons, such as possible batch effects in TCGA germline variant calls [43].

Almost all genes (99%) had at least one germline variant per patient (**Supplementary Figure 9**). Overall, 61%, 95%, and 99% of the PIUs, linker units, and non-coding units, respectively, had mapped variants in at least two patients. The list of frequently mapped GPD units is reported in **Supplementary Table 6**.

We next applied principal component analysis to the germline variant count data to test whether germline data alone can be associated with survival outcomes. Using the principal components of the unit- and gene-level data most significantly associated with survival, we stratified all patients in each cancer into two groups (median of principal component scores). The log-rank tests of the survival curves showed substantially larger p-values (i.e. less significant results) than those observed in the analysis of somatic mutation data (**Supplementary Figure 3B**). This result implies that the genetic predisposition of a patient does only rarely affect survival of the cancer.

These rare exceptions include some cancers, e.g. colon, bladder, kidney, prostate and liver cancers (COAD, BLCA, KIRC, PRAD and LIHC), in which the survival curves showed better separation between the two survival groups, suggesting that pathological germline variants may indeed possess prognostic potential. Overall, unlike in the somatic mutation analysis, the sequence unit-level analysis did not outperform the gene-level analysis. (**Supplementary Figures 3B, 10, and 11**).

We next set out to examine which sequence units harboring germline variants were associated with overall survival, following the same workflow as for the somatic mutations. We found that 929, 427, and 17,086 PIUs, linker, and non-coding units were significantly associated with overall survival in at least one tumor type. When we examined the regression coefficients, 880 PIUs (95%), 369 LUs (86%) and 15,913 NCUs (93%) showed deleterious effect with statistical significance (**Supplementary Table 7**). However, although they account for a small proportion, 49 PIUs (5%), 58 linker units (14%), and 1,173 non-coding units (7%) had statistically significant, negative coefficients (*q*-value 0.1), indicating that a small fraction of germline variants had positive effects on survival.

Stomach cancer (STAD) showed the largest number of significant associations for sequence units with germline variants, involving 216 PIUs (211 positive), 99 linker units (97 positive) and 10,086 non-coding units (**Figure 4A**). Only eleven of the 880 deleterious PIUs showed significant associations with overall survival in two or more cancer types, including RhoGEF domain in *ARHGEF37* gene, 7tm Family in olfactory receptor gene *OR4Q2*, TGF-beta domain in *BMP8B* gene, and Cadherin domain in *PCDHGB3* gene, to name a few. Meanwhile, linker units in six genes (*ANKDD1B*, *SLC26A8*, *SRRM2*, *TPRA1*, *ZNF202* and *ZNF398*) showed statistically significant associations for overall survival in two or more types of cancer, and non-coding units in 11,510 genes had possible deleterious effects in at least two cancer types. The non-coding unit in *FAM210A* had variants associated with poorer survival outcomes in as many as nine cancer types.

By contrast, gene-level analysis discovered significantly associated genes in 13 cancer types only. Stratifying patient survival outcomes by the germline variant counts on these genes shows less significant separation compared with clustering using unit-level analysis results (**Figure 4B** and **4C, Supplementary Figure 12**). Given that most germline variants reside in the non-protein coding units (NCU), it appears that pathological germline variants may be more enriched in the coding regions (**Figure 4A**). When we performed the gene-level analysis with germline variants strictly in coding regions, we indeed observed better separation of

survival curves in many cancers (data not shown), but not as clear as the separation we observed in the analysis with somatic mutations.

Interactions between somatic and germline variants in the association with overall survival

We have shown that GPD allows for detection of sequence units enriched with variants that are associated with overall survival. Next, we exploited GPD's capability to test for interactions between germline variants and somatic mutations within each gene in the association with overall survival. Such analysis is unique and of high clinical importance, as each subject's germline genome may modulate the effects of somatic mutations.

We surveyed all GPD units for which ≥10 patients had at least one somatic mutation in the GPD unit and a germline variant in the corresponding gene. We only considered the units for which the association between somatic mutation counts and overall survival was significant (p-value < 0.05). These requirements rendered 5,822, 8,453, and 4,850 PIUs, linker, and non-coding units in 8,466 genes eligible. For a given sequence unit, we tested for the main effects of somatic mutation counts and germline variant counts as well as the interaction between the two (see **Methods**). We did not adjust for multiple testing since the relatively small number of units qualified for the testing.

Figure 5A shows that uterine cancer (UCEC) had the largest number of interaction effects, followed by colon (COAD), lung (LUSC and LUAD), stomach (STAD) cancers. **Supplementary Table 8** provides the list of 594 total significant interactions across all the cancer types. Interestingly, interaction terms in colon, lung, and stomach cancers (COAD, LUSC, and STAD) had more positive coefficients than negative ones (68%, 70%, and 96% respectively), suggesting synergistic adverse effects of simultaneous occurrence of somatic mutations and germline variants on overall survival in those cancers.

Figure 5B shows examples demonstrating how the interaction between germline variants and somatic mutations modified the hazard ratio. For example, there were multiple occurrences of both germline variants and somatic mutations in the disordered arm u3 domain

of the *APC* gene. Sixty-two out of 378 colon cancer (COAD) patients had at least one somatic mutation in this sequence unit and 373 patients had germline variant(s). We found that the difference in the hazard ratios between a patient with somatic mutation and a patient without was much greater if the patients had more germline variants. Similar patterns were observed in the *KMT2C* and *TP53* genes in their respective cancer types. In comparison, as seen in the bottom panel of **Figure 5B**, having very few germline variants along with a somatic mutation decreased the hazard ratio and elevated the risk of death substantially.

In sum, these findings suggest that substantial germline variation in key sequence units predisposes patients to a higher risk of death. Somatic mutations coupled with more germline variants in the same gene tend to have greater negative effects on survival, while somatic mutations surrounded by few germline variations may have protective effects.

Discussion

We present a new approach, GPD, which summarizes sequence variation and mutations in small units of protein-coding and non-coding regions. We used protein domains and families and small windows surrounding protein modification sites as the sequence units for data summarization. GPD's software architecture enables the user to easily define their own GPD units, rendering the analysis framework highly flexible. In particular, linker regions are currently an amalgamation of unannotated areas on protein sequences, and this can be improved by expanding the functional annotation and 3D structure characterization of proteins.

GPD has two main advantages. First, mutation count data for GPD's sequence units immediately allow for downstream statistical analysis to test their association between with clinical outcomes. GPD mitigates the sparsity problem of locus-level exome sequencing data, addressing the inherent drawback that most exome variants do not occur at the same position with reasonably high frequency across different subjects. Second, GPD exploits the fact that variant data is not randomly distributed across a gene's sequence but accumulates in specific

regions. Therefore, it does not sacrifice the statistical power of conventional gene-level summaries, while providing higher resolution than previous approaches.

Further, supporting subsequent experimental analysis of the functional impact of any mutations, the sequence units point immediately to the relevant regions within a gene and quantify enrichment of mutations in those specific regions. This information can immediately streamline prognosis of mutations in specific patients.

Another major advance in GPD is the quantitation of the combinatorial impact of different types of sequence variation, leading to the identification of interactions between existing genetic predisposition of a patient (germline variants) and tumorigenic events (somatic mutations) with respect to disease prognosis. To the best of our knowledge, this interaction has not been systematically addressed through statistical association analysis of large-scale data sets. We found many genes in which a patient had few germline variants and somatic mutations in the same gene were associated with better overall survival. By contrast, other genes had many pre-existing, germline variations, we found that additional somatic mutations almost exclusively had deleterious effects on overall survival. This finding implies that genetic predisposition (germline variants) affects the impact of subsequent mutation events on a gene and their indirect impact on patient survival -- highlighting a new avenue for individualized treatment based on the patient's genotype.

To date, only few other studies have explored the interaction between germline and somatic variation, but in limited fashion. For example, Carter *et al*. [44] analysed ~6,000 tumors and identified 412 genetic interactions between germline polymorphisms and major somatic events. However, the analysis was limited to 138 known oncogenes and tumor suppressor genes and focused on the effect of inherited polymorphisms in germline on the likelihood of somatic mutations, rather than discovering the prognostic value of the interaction between the two classes of sequence variation.

GPD's results are validated through their overlap with known cancer genes. Although GPD analysis reported sequence units that did not overlap with the list of cancer driver genes, forty percent of the most reliable driver mutations (232 of 579) are within 12 of GPD's survival-

associated sequence units, suggesting half the driver mutations carry predictive information on disease prognosis. Furthermore, most sequence units with somatic mutations were highly patient-specific: they occurred in only few samples across cancers. However, panels of exome variants with prognostic potential will likely be more diverse than the panel of driver mutations, especially given different treatment course among the patients. Future analysis for prognostic panel selection should be thoroughly validated in a large number of independent cohorts.

Acknowledgments

This work was supported in part by grants from Singapore Ministry of Education (MOE2016-T2-1-001 and MOE2018-T2-2-058 to H.C.) and the National Institutes of Health (NIH 5R35GM127089-02 to C.V.). The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflict of Interest

None of the authors have any competing interests.

Figures

Figure 1. GPD mapping scheme and mapping results of somatic mutations across 33 cancers. (**A**) GPD's protein information units (PIUs) consist of domains and windows surrounding protein modification sites (+/- 5 amino acids around a PTM site) in proteins (yellow); Regions of the protein sequence that are not covered by a domain and a PTM site-containing window are called linker units (LUs) (blue). All sequence regions that do not encode the final protein product are named non-protein coding units (NCUs). Mutations (red triangles) are mapped to the GPD units defined above. GPD's unit-level data are per unit counts of mutations. Locus-level and gene-level counts are also illustrated for comparison. (**B**) Histograms of the mutation-carrying loci, genes, PIUs, LUs and NCUs in 10,793 subjects (x-axis) across all TCGA cancers. For example, a majority of mutations occurring at a specific locus have frequency of one or two individuals. By contrast, genes with any somatic mutations have a median count of 64. (**C**) Percentage of mutations mapped to one of the three types of units in a gene. (**D**) Percentage of patients carrying mutations in the 50 most frequently mapped PIUs across 33 cancers. Stars indicate the PIUs mentioned in the main text. (**E**) Mutation frequency (logarithm base 2) of each PIU and LU among patients in TCGA Pan-Cancer cohort and MSK-IMPACT cohort.

Figure 2. Association analysis between GPD somatic unit-level data and overall survival. (**A**) The total number and the proportion of significantly associated units (PIU, LU and NCU) for each cancer. (**B**) In each cancer, patients with and without mutations in any of the significant units are separated into two groups (red and blue solid line, see **Methods**). Kaplan-Meier curves are drawn for the two groups in five cancers with the most significant difference in survival curves between the two risk groups. For comparison, patients with and without mutation in any of the significant genes are separated into two groups (red and blue dashed line). (**C**) Minus the logarithm of *p*-value (base 10) of the log rank test for the difference in survival curve between the two groups across 20 cancers. Diagonal line indicates equal

significance value for both analyses. X-axis and y-axis show the significance score for GPD's unit-level analysis and gene-level analysis, respectively.

Figure 3. Detailed mutational profiles of *SMAD4* in colon cancer (COAD) patients and *ERBB4* in ER positive breast cancer (BRCA) patients. Different types of mutations are shown with dots in different colors. Frequency of mutation at individual loci is indicated with the height of the vertical line under the dots. PIUs are shown along the bottom as colored triangles (PTM sites) and orange bars (domains). Somatic mutation count of MH1 in *SMAD4* and LU in *ERBB4* were statistically significantly associated with overall survival.

Figure 4. Association between GPD unit-level germline variant counts and overall survival. (**A**) The total number and proportion of significantly associated units (PIU, LU and NCU) for each cancer. (**B**) Kaplan-Meier plots for five cancers with the most significant difference in survival curves between the two risk groups. Similar to somatic mutation data, patients with and without mutations in any of the significant units are separated into two groups (red and blue solid line), and likewise for mutations in significant genes (red and blue dashed line). (**C**) Minus the logarithm of *p*-value (base 10) of the log rank test for the difference in survival curve between the two groups across 20 cancers. Diagonal line indicates equal significance value for both analyses. X-axis and y-axis show the significance score for GPD's unit-level analysis and gene-level analysis, respectively.

Figure 5. Testing the interaction effect of somatic and germline variants in cancer prognosis. (**A**) The number of statistically significant interaction effects in each cancer, among the eligible units (see **Methods**). (**B**) Comparison of the hazard ratios between patient groups with different somatic and germline variant levels. Patients are divided into four groups based on whether they have zero or at least one somatic mutation in the unit and whether they have fewer or more germline variants than the 75% quantile across all patients.

References

1. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015;349:1483–9. doi:10.1126/science.aab4082.

2. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018;173:305–20.e10.

3. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. 2018;6:271–81.e7.

4. Huang K-L, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell. 2018;173:355–70.e14.

5. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173:400–16.e11.

6. Dellinger TH, Smith DD, Ouyang C, Warden CD, Williams JC, Han ES. L1CAM is an independent predictor of poor survival in endometrial cancer - An analysis of The Cancer Genome Atlas (TCGA). Gynecol Oncol. 2016;141:336–40.

7. Castro E, Goh C, Olmos D, Saunders E, Leongamornlert D, Tymrakiewicz M, et al. Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. J Clin Oncol. 2013;31:1748–57.

8. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21:938–45.

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

10. Xing M. BRAF mutation in thyroid cancer. Endocrine Related Cancer. 2005;12:245–62.

11. Olivier M. The Clinical Value of Somatic TP53 Gene Mutations in 1,794 Patients with Breast Cancer. Clinical Cancer Research. 2006;12:1157–67. doi:10.1158/1078-0432.ccr-05-1029.

12. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.

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

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

15. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.

16. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. Predicting the functional

consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29:1504–10.

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

18. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27:2147–8.

19. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics. 2014;30:3109–14.

20. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10:1108–15.

21. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.

22. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics. 2013;29:2757–64.

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

24. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;174:1034–5.

25. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell. 2015;27:15–26.

26. Pickrell J, Berisa T, Segurel L, Tung JY, Hinds D. Detection and interpretation of shared genetic influences on 40 human traits. doi:10.1101/019885.

27. Jeggo PA, Pearl LH, Carr AM. DNA repair, genome stability and cancer: a historical perspective. Nature Reviews Cancer. 2016;16:35–42. doi:10.1038/nrc.2015.4.

28. Kiiski JI, Fagerholm R, Tervasmäki A, Pelttari LM, Khan S, Jamshidi M, et al. FANCM c.5101C>T mutation associates with breast cancer survival and treatment outcome. Int J Cancer. 2016;139:2760–70.

29. Candido-dos-Reis FJ, Song H, Goode EL, Cunningham JM, Fridley BL, Larson MC, et al. Germline mutation in BRCA1 or BRCA2 and ten-year survival for women diagnosed with epithelial ovarian cancer. Clin Cancer Res. 2015;21:652–7.

30. Hristov BH, Singh M. Network-Based Coverage of Mutational Profiles Reveals Cancer Genes. Cell Systems. 2017;5:221–9.e4. doi:10.1016/j.cels.2017.09.003.

31. Le Morvan M, Zinovyev A, Vert J-P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. PLoS Comput Biol. 2017;13:e1005573.

32. Zhong X, Yang H, Zhao S, Shyr Y, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC Genomics. 2015;16 Suppl 7:S7.

33. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6:80–92.

34. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015;43 Database issue:D512–20.

35. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32.

36. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:W90–7.

37. Fraser JA, Vojtesek B, Hupp TR. A novel p53 phosphorylation site within the MDM2 ubiquitination signal: I. phosphorylation at SER269 in vivo is linked to inactivation of p53 function. J Biol Chem. 2010;285:37762–72.

38. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23:703–13.

39. Miyaki M, Kuroki T. Role of Smad4 (DPC4) inactivation in human cancer. Biochemical and Biophysical Research Communications. 2003;306:799–804. doi:10.1016/s0006-291x(03)01066-0.

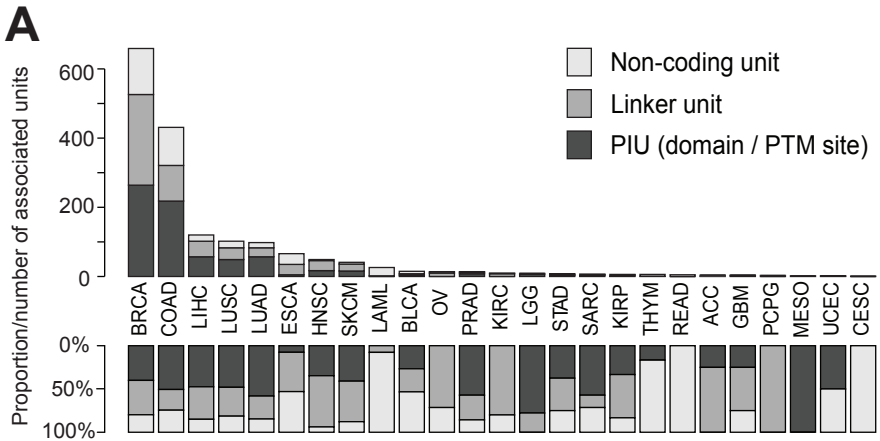
40. Fleming NI, Jorissen RN, Mouradov D, Christie M, Sakthianandeswaren A, Palmieri M, et al. SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. Cancer Res. 2013;73:725–35.

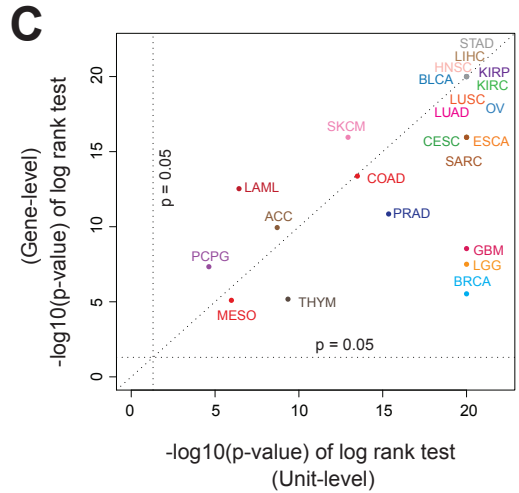
41. Veikkolainen V, Vaparanta K, Halkilahti K, Iljin K, Sundvall M, Elenius K. Function of ERBB4 is determined by alternative splicing. Cell Cycle. 2011;10:2647–57.

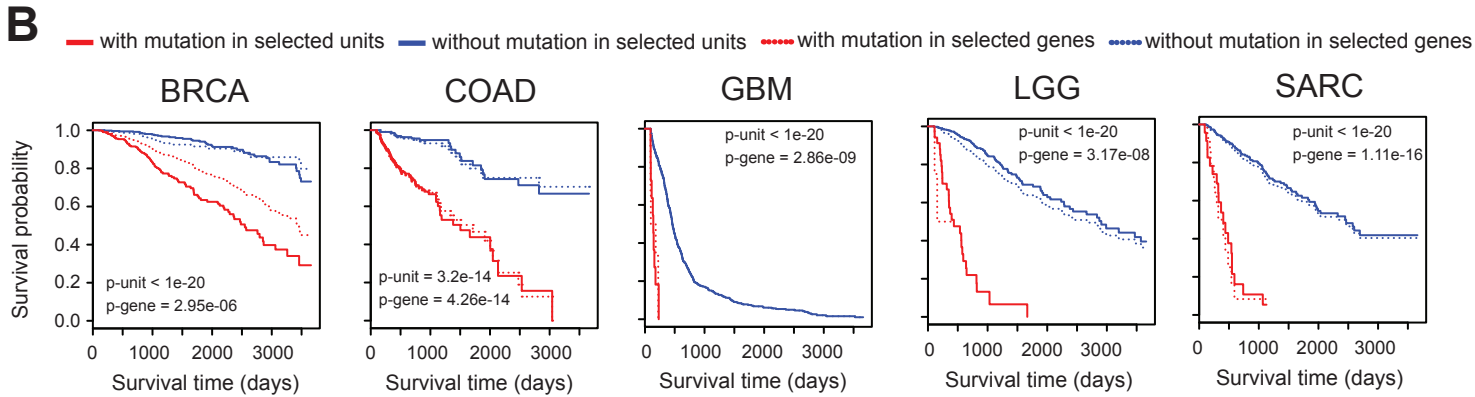
42. Griffith OL, Spies NC, Anurag M, Griffith M, Luo J, Tu D, et al. The prognostic effects of somatic mutations in ER-positive breast cancer. Nat Commun. 2018;9:3476.

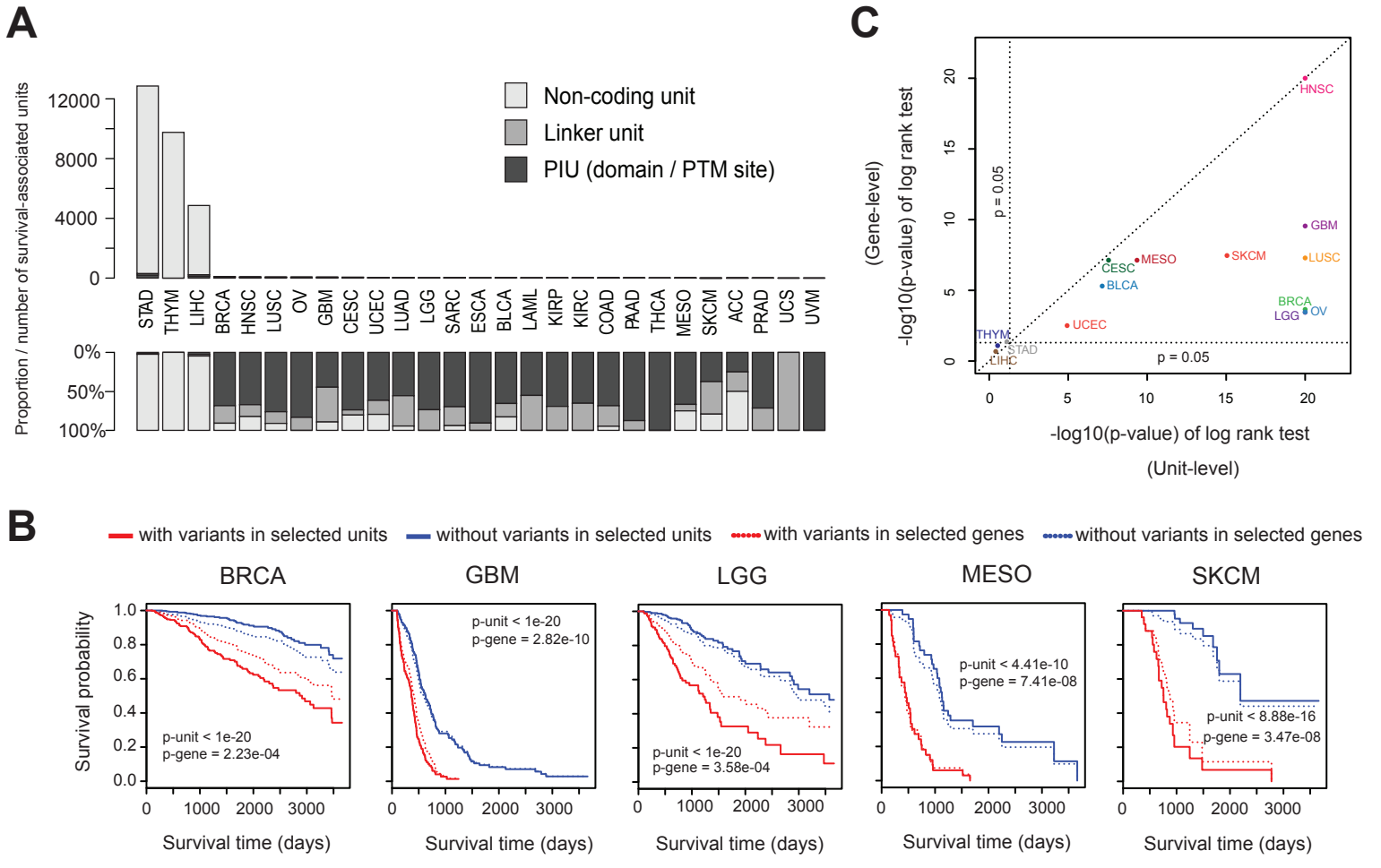
43. Buckley AR, Standish KA, Bhutani K, Ideker T, Lasken RS, Carter H, et al. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. BMC Genomics. 2017;18:458.

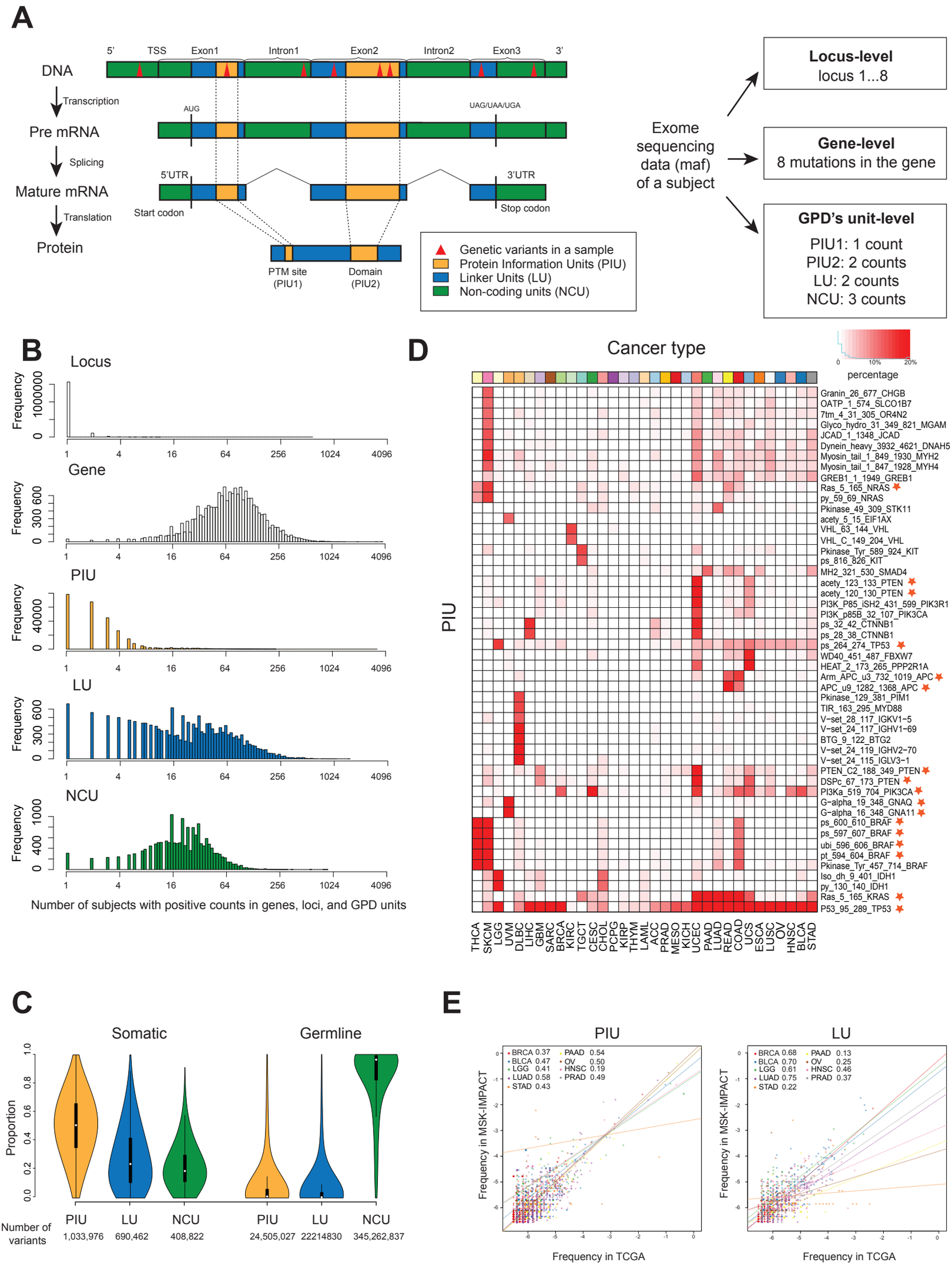
44. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, et al. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. Cancer Discov. 2017;7:410–23.

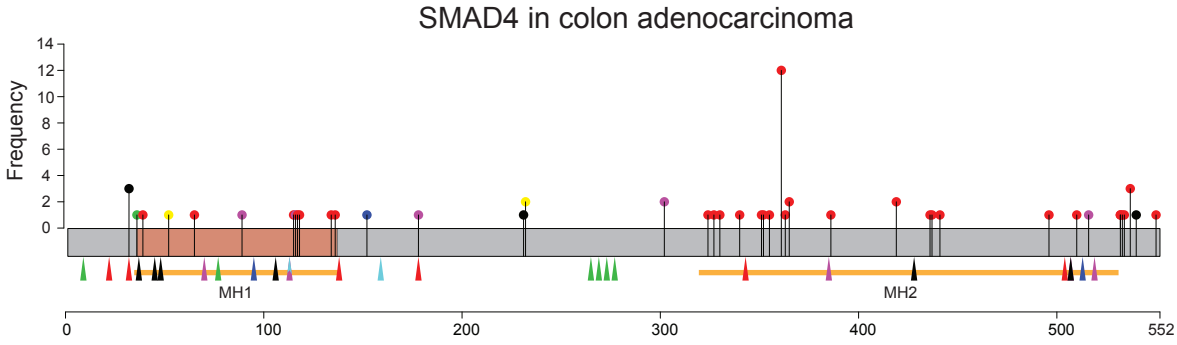

A
Non-coding unit
Linker unit
PIU (domain / PTM site)
Proportion/number of associated units
600
400
200
0
0%
50%
100%
BRCA
COAD
LIHC
LUSC
LUAD
ESCA
HNSC
SKCM
LAML
BLCA
OV
PRAD
KIRC
LGG
STAD
SARC
KIRP
THYM
READ
ACC
GBM
PCPG
MESO
UCEC
CESC

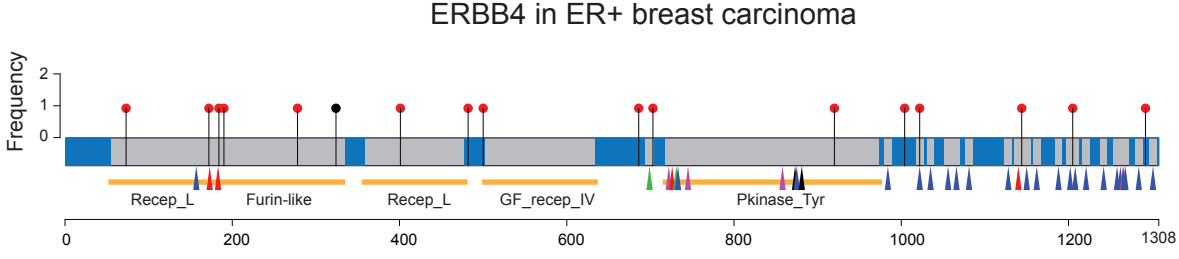

C
(Gene-level)
-log10(p-value) of log rank test
-log10(p-value) of log rank test
(Unit-level)
p = 0.05
p = 0.05
STAD
LIHC
HNSC
BLCA
KIRP
KIRC
LUSC
OV
LUAD
SKCM
CESC
ESCA
SARC
COAD
LAML
ACC
PRAD
PCPG
GBM
LGG
BRCA
THYM
MESO
0
5
10
15
20


B
with mutation in selected units
without mutation in selected units
with mutation in selected genes
without mutation in selected genes
BRCA
COAD
GBM
LGG
SARC
Survival probability
1.0
0.8
0.6
0.4
0.2
0.0
p-unit < 1e-20
p-gene = 2.95e-06
p-unit = 3.2e-14
p-gene = 4.26e-14
p-unit < 1e-20
p-gene = 2.86e-09
p-unit < 1e-20
p-gene = 3.17e-08
p-unit < 1e-20
p-gene = 1.11e-16
0
1000
2000
3000
Survival time (days)


A
Proportion / number of survival-associated units
12000
8000
4000
0
Non-coding unit
Linker unit
PIU (domain / PTM site)
STAD THYM LIHC BRCA HNSC LUSC OV GBM CESC UCEC LUAD LGG SARC ESCA BLCA LAML KIRP KIRC COAD PAAD THCA MESO SKCM ACC PRAD UCS UVM
0%
50%
100%
C
(Gene-level) -log10(p-value) of log rank test
-log10(p-value) of log rank test (Unit-level)
p = 0.05
p = 0.05
HNSC
GBM
SKCM
LUSC
MESO
CESC
BLCA
BRCA
LGG
OV
UCEC
THYM
STAD
LIHC
B
with variants in selected units
without variants in selected units
with variants in selected genes
without variants in selected genes
BRCA
GBM
LGG
MESO
SKCM
Survival probability
p-unit < 1e-20
p-gene = 2.23e-04
p-unit < 1e-20
p-gene = 2.82e-10
p-unit < 1e-20
p-gene = 3.58e-04
p-unit < 4.41e-10
p-gene = 7.41e-08
p-unit < 8.88e-16
p-gene = 3.47e-08
Survival time (days)


A
5'
TSS
Exon1
Intron1
Exon2
Intron2
Exon3
3'
DNA
Transcription
Pre mRNA
AUG
UAG/UAA/UGA
Splicing
Mature mRNA
5'UTR
3'UTR
Start codon
Stop codon
Translation
Protein
PTM site (PIU1)
Domain (PIU2)
Genetic variants in a sample
Protein Information Units (PIU)
Linker Units (LU)
Non-coding units (NCU)
Exome sequencing data (maf) of a subject
Locus-level
locus 1...8
Gene-level
8 mutations in the gene
GPD's unit-level
PIU1: 1 count
PIU2: 2 counts
LU: 2 counts
NCU: 3 counts
B
Locus
Gene
PIU
LU
NCU
Frequency
Number of subjects with positive counts in genes, loci, and GPD units
D
Cancer type
percentage
PIU
THCA SKCM LGG UVM DLBC LIHC GBM SARC BRCA KIRC TGCT CESC CHOL PCPG KIRP THYM LAML ACC PRAD MESO KICH UCEC PAAD LUAD READ COAD UCS ESCA LUSC OV HNSC BLCA STAD
C
Somatic
Germline
Proportion
PIU LU NCU PIU LU NCU
Number of variants
1,033,976 690,462 408,822 24,505,027 22214830 345,262,837
E
PIU
LU
Frequency in MSK-IMPACT
Frequency in TCGA

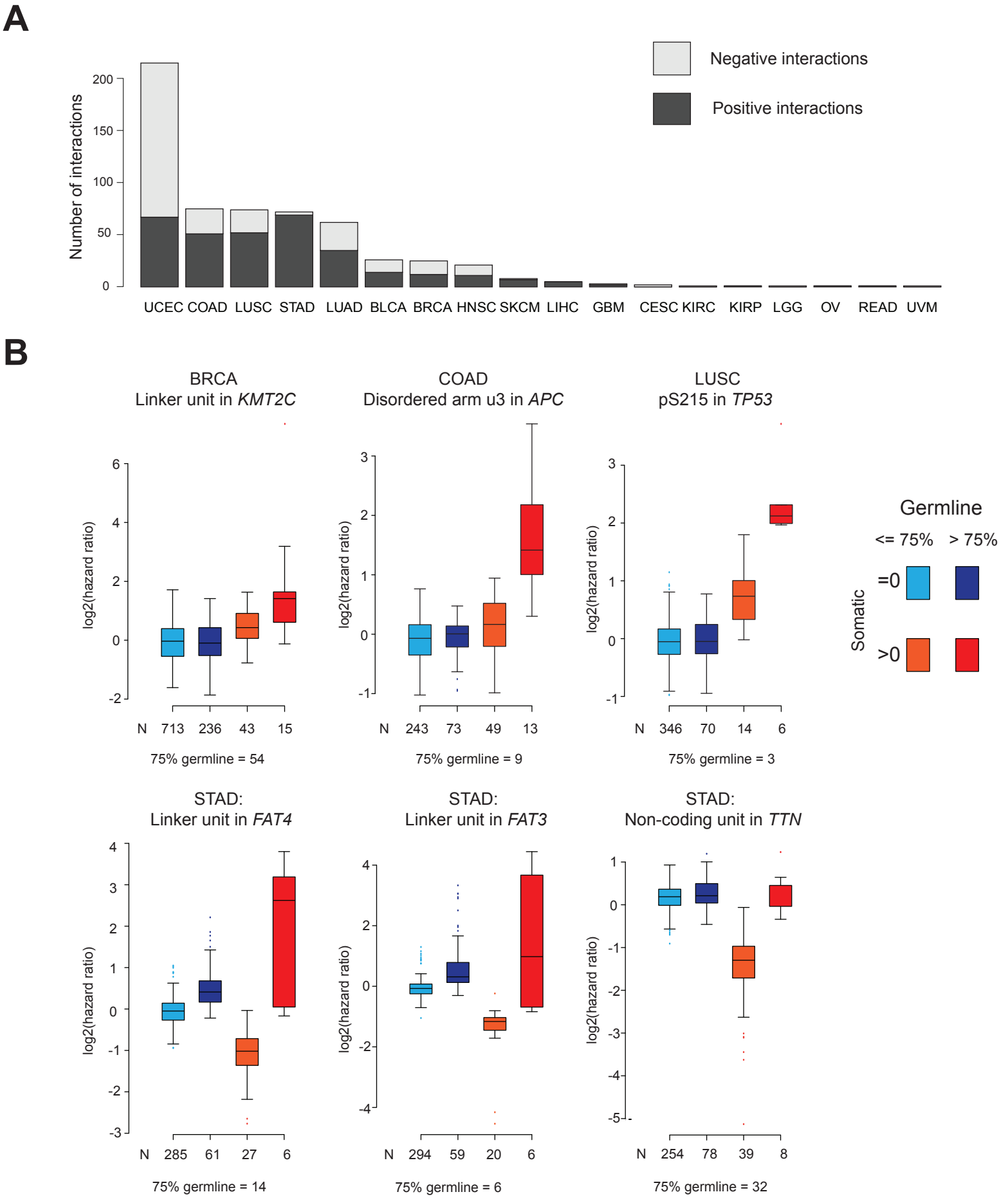




- Missense Mutation
- Frame Shift Deletion
- Frame Shift Insertion
- Nonsense Mutation
- In Frame Deletion
- Intron

- Phosphorylation (serine)
- Phosphorylation (threonine)
- Phosphorylation (tyrosine)
- Ubiqutination (lysine)
- Acetylation (lysine)
- SUMOylation (lysine)

- Protein domains
- Significant PIU
- Significant linker units


A
Negative interactions
Positive interactions
Number of interactions
200
150
100
50
0
UCEC COAD LUSC STAD LUAD BLCA BRCA HNSC SKCM LIHC GBM CESC KIRC KIRP LGG OV READ UVM
B
BRCA
Linker unit in KMT2C
COAD
Disordered arm u3 in APC
LUSC
pS215 in TP53
log2(hazard ratio)
N 713 236 43 15
75% germline = 54
N 243 73 49 13
75% germline = 9
N 346 70 14 6
75% germline = 3
Germline
<= 75% > 75%
Somatic
=0
>0
STAD:
Linker unit in FAT4
STAD:
Linker unit in FAT3
STAD:
Non-coding unit in TTN
N 285 61 27 6
75% germline = 14
N 294 59 20 6
75% germline = 6
N 254 78 39 8
75% germline = 32