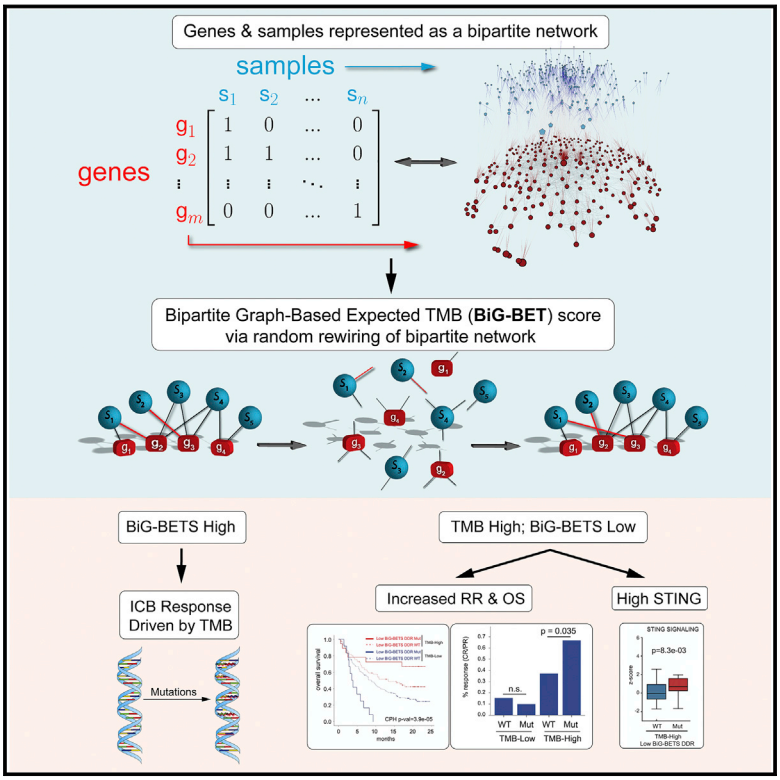
Cell Reports
Medicine

Report

A bipartite graph-based expected networks approach identifies DDR genes not associated with TMB yet predictive of immune checkpoint blockade response

Graphical abstract



Highlights

- Univariate tests are inappropriate to associate gene mutation with elevated TMB
- The TMB paradox clarifies why most genes associate with high TMB
- Denoting tumors and genes as a bipartite network resolves the TMB paradox
- TMB high tumors with low BiG-BETS DDR gene mutation respond well to ICB

Authors

William H. Weir, Peter J. Mucha, William Y. Kim

Correspondence

peter.j.mucha@dartmouth.edu (P.J.M.), wykim@med.unc.edu (W.Y.K.)

In brief

Weir et al. demonstrate that a networks-based approach resolves the TMB paradox, accurately defines genes associated with elevated TMB, and delineates a cohort of patients (TMB high, low BiG-BETS DDR mutant) with high predictive power for response and prolonged overall survival with ICB.

Weir et al., 2022, Cell Reports Medicine *3*, 100602
May 17, 2022
https://doi.org/10.1016/j.xcrm.2022.100602

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Report

A bipartite graph-based expected networks approach identifies DDR genes not associated with TMB yet predictive of immune checkpoint blockade response

William H. Weir,[1] Peter J. Mucha,[1,2,3,*] and William Y. Kim[4,5,6,7,8,*]

[1]Curriculum in Bioinformatics & Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[3]Department of Mathematics, Dartmouth College, Hanover, NH, USA
[4]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[5]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[6]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7]Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[8]Lead contact
*Correspondence: peter.j.mucha@dartmouth.edu (P.J.M.), wykim@med.unc.edu (W.Y.K.)
https://doi.org/10.1016/j.xcrm.2022.100602

SUMMARY

Immune checkpoint blockade (ICB) has had remarkable success for treatment of solid tumors. However, as only a subset of patients exhibit responses, there is a continued need for biomarker development. Numerous reports have shown a link between tumor mutational burden (TMB) and ICB response, while others have identified a link between ICB response and mutation in DNA damage repair (DDR) genes. However, it remains unclear to what extent mutations in DDR genes hold predictive value above and beyond their association with TMB. Herein, we present a networks-based test and bipartite graph-based expected TMB score (BiG-BETS) with higher specificity for discriminating DDR genes and pathways that are associated with elevated TMB. Moreover, we find that mutations in certain DDR genes that are not associated with elevated TMB (low BiG-BETS) are nevertheless predictive of ICB benefit in high TMB patients, demonstrating that their inactivation contributes to ICB response in a TMB-independent manner.

INTRODUCTION

Immune checkpoint blockade (ICB) has achieved remarkable success in many solid tumors. Nonetheless, only a minority of patients respond and ICB is associated with significant financial toxicity;[1] therefore, the ability to better predict ICB response has the potential to impact both patient survival and quality of life. Several genomic markers have demonstrated consistent predictive power for ICB response, including tumor intrinsic properties such as tumor mutational burden (TMB) as well as RNA expression signatures (i.e., a T cell inflamed gene expression profile).[2] While high levels of TMB in particular have been consistently associated with ICB response,[3] even patients with high TMB levels have response rates below 40%. Therefore, despite these important observations, there remains significant patient heterogeneity in ICB response that is not explained by existing biomarkers.

TMB represents the balance between a tumor's exposure to a mutagenic process (i.e., UV radiation, carcinogen, etc.) and the integrity of the cellular DNA damage repair (DDR) pathways. Consistent with this notion, an elevated TMB is frequently seen in tumors associated with carcinogens (i.e., melanoma and UV radiation and lung cancer and cigarette smoke)[3,4] and has also been associated with mutations in some DDR genes.[5–9] In contrast, in a comprehensive study, the Cancer Genome Atlas (TCGA) DDR working group assessed whether mutations in DDR genes were associated with elevated TMB. They found only two DDR genes that when mutated were significantly associated with a higher TMB than other genes in the cohort.[10] We sought to resolve this discrepancy as well as to assess whether mutations in DDR genes have predictive power for ICB response independent of TMB.

In this study, we address the dilemma as to whether DDR gene mutations are a cause or consequence of elevated TMB. We show that, using traditional univariate tests, the vast majority of genes when mutated are associated with elevated TMB, illustrating that these approaches are ill suited to define relationships between gene mutation and TMB. This is explained by the fact that the readout of interest (TMB) is confounded by the variable it is being associated to (mutations in a gene). We illustrate how the "TMB paradox," refashioned from the well-established friendship paradox in network science, accounts for why the vast majority of genes are associated with an elevated TMB when using univariate testing. Furthermore, we show that

representing tumors and their mutated genes as a bipartite network allows for development of a bipartite graph-based expected TMB score (BiG-BETS) that more accurately defines DDR genes associated with high or low TMB, termed high and low BiG-BETS DDR genes, respectively. Finally, while we note that having a mutation in a high BiG-BETS DDR gene did not add predictive power to ICB benefit for patients with TMB high tumors, low BiG-BETS DDR gene mutation in TMB high tumors enriched for patients with elevated stimulator of interferon genes (STING) pathway activity significantly increased ICB response and overall survival benefit to ICB.

RESULTS

The majority of genes, when mutated, are associated with elevated TMB by univariate test

While it is logical that ineffective DNA damage or repair mechanisms would result in increased mutational load, all human genomic studies to date describing an association between the presence of DDR mutations and elevated TMB have been correlative. Indeed, an equally plausible explanation is that mutation in a DDR gene (or any gene) is simply more likely in TMB high tumors because more genes are mutated. We therefore sought to understand whether tumors with DDR gene inactivation associate with elevated TMB as a cause or consequence of DDR gene mutation.

We noted that the majority of studies linking DDR gene inactivation to elevated TMB assessed statistical significance using a univariate test (t test or Mann-Whitney U) comparing mutant tumors with non-mutant tumors.[7–9,11,12] This approach is intrinsically biased because the readout of interest (TMB) is confounded with the variable it is being associated to (mutations in a gene). By using a univariate test, tumors with a higher TMB will have a higher likelihood of having mutations in the genes being tested (i.e., DDR pathways). Moreover, aggregating mutations across genes into a pathway further increases the effect of this bias.

To demonstrate this bias, we applied the classic univariate approach (Mann-Whitney U [MWU] test) to the pan-cancer TCGA dataset, assessing whether mutation of a gene correlates with elevated TMB. Strikingly, we found that 98% (17,860/18,151) of genes when mutated were associated with elevated TMB relative to their respective non-mutated out-groups, even when corrected for multiple comparisons (Figure 1A), and that DDR genes as a group did not have a significantly higher proportion of genes that when mutated were associated with elevated TMB (Figure 1B). Moreover, when we applied the classic univariate approach (MWU test) to query whether inactivation of any of the DDR pathways (nucleotide excision repair [NER], base excision repair [BER], non-homologous DNA end joining [NHEJ], mismatch repair [MMR], Fanconi anemia [FA], homologous recombination [HR], and damage sensor [DS]) as defined by the TCGA DDR group[10] were associated with increased TMB, we found that inactivation of each DDR pathway was highly significantly associated with elevated TMB (Figure S1A), in keeping with previously published work.[7–9,12,13] Furthermore, examination of the distribution of the mean TMB values of tumors with a mutation in each gene found that the average mean TMB linked to DDR mutations was only slightly higher (24.1 versus 23.26) than the average mean TMB for non-DDR genes (Figure 1C). The fact that this univariate approach rejects the null hypothesis for the vast majority of genes and that the mean TMB for DDR gene mutations is not different than that of mutations in non-DDR genes suggests that the application of univariate statistical testing is invalid.

The friendship paradox accounts for why the majority of genes associate with elevated TMB by univariate testing

It is counterintuitive that the majority of genes (when mutated) have a higher mean TMB (~24) than the mean TMB of all samples (~5; Figure 1D, vertical dotted line). We have labeled this phenomenon the "TMB paradox," because it reflects the well-established "friendship paradox" found in network analysis. The friendship paradox holds that, in a social network, most people have fewer friends than their friends do.[14] In other words, for most nodes in a network, their neighbors or "friends" will on average have a higher degree (number of connections) than the node itself. This arises because higher degree nodes count toward the degree in multiple neighboring nodes and thus are oversampled (see STAR Methods: Proof of friendship paradox). Analogously, highly mutated samples contribute toward the average TMB for many of the genes in the dataset, resulting in an outsized effect (Figure S1B). Since the univariate t test and Mann-Whitney U tests are testing for differences in central tendencies (i.e., the mean or the median, respectively), the TMB paradox explains why the majority of genes have a highly significant association with elevated TMBs. The TMB paradox is therefore a manifestation of the oversampling bias introduced by highly connected nodes (i.e., high TMB tumors), and this oversampling bias is what makes univariate tests (t test and MWU) inappropriate to identify which genes are associated with higher levels of TMB. Therefore, the majority of the 98% of genes associated with elevated TMB by the univariate approach (Figure 1A) are likely an artifactual result of the TMB paradox rather than the underlying biology.

Representing tumors and their mutated genes as a bipartite network facilitates development of a BiG-BETS to define DDR genes associated with high TMB

We hypothesized that a networks-based approach and a more appropriate, joint statistical test for whether mutation in a given gene is specifically associated with higher levels of TMB would better discriminate which DDR genes are truly associated with an elevated TMB. To this end, we accounted for the TMB paradox by converting the pan-cancer TCGA tumors and their respective mutated genes (moderate + high consequence; see STAR Methods) into a bipartite network, where genes and tumors represent the two classes of nodes and the edges (connecting two nodes) indicate the mutated genes within a tumor (Figure 1E). By re-casting our data into a bipartite network, we see in Figure 1E that the TMB for a sample is roughly equivalent to the number of edges it has (i.e., its degree) and that the average TMB associated with mutation in a given gene is the average degree of its neighbors (Figures 1E and S1B).

We then leveraged this bipartite network representation to derive a null model of TMB distribution for each DDR gene or pathway. Specifically, random sampling (permutations) of the

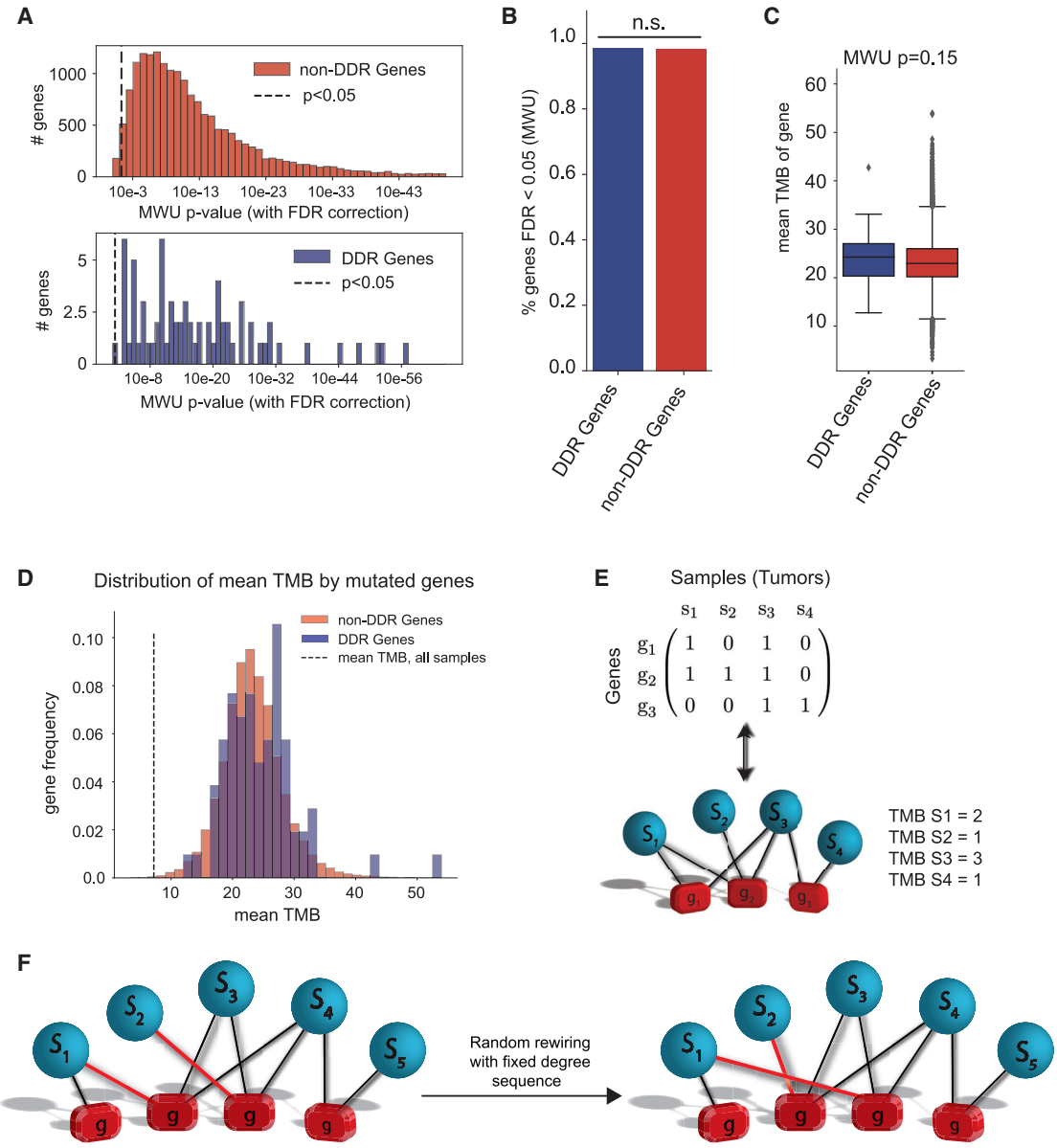


Figure 1. Univariate testing inappropriately associates most genes with elevated TMB, and re-casting samples and mutations as a bipartite network overcomes this limitation

(A) Distribution of Mann-Whitney U test p values (with multiple test correction) on reversed log scale across all genes in pan-TCGA dataset (red) versus DDR genes only (blue). For each gene, the MWU test compares distribution of TMB values for samples with a mutation in the gene versus all samples in the cohort. Right of dashed black line represents $p < 0.05$. FDR, false discovery rate.

(B) Percentage of genes in which mutations are significantly associated with elevated TMB by the MWU test (with FDR correction) broken down by DDR genes (blue) and non-DDR genes (red). n.s., not significant.

(C and D) Distribution of mean TMB values for mutated sample set for all genes (red) versus DDR genes (blue) in TCGA (compared using MWU test in [C]). Vertical dotted black line in (D) denotes the overall mean TMB for the cohort of samples. Diamonds in (C) represent outliers (data >1.5 * IQR).

(E) Schematic representation of converting the mutational data in a matrix to a bipartite network.

(F) Schematic representation of BiG-BETS network rewiring process to sample from the bipartite configuration model. Random pairs of edges are selected to be exchanged to generate new samples.

bipartite network was performed by stochastically rewiring the network while maintaining the degree distribution (number of edges of each node) of the original dataset (this null model is known as the configuration model; see STAR Methods).[15,16] Generation of a null model through permutation allowed us to compare the actual mean TMB for tumors with mutation in a given gene or pathway against the expected distribution under random sampling (Figure 1F).

Application of this BiG-BETS to the TCGA pan-cancer dataset found that BiG-BETS returned a more uniform distribution of p values (Figure S2A), and only a subset of DDR genes when mutated have a significant association with elevated TMB (Figure 2A), hereafter referred to as "high BiG-BETS DDR gene" (Figure 2B). It was reassuring to see that MMR genes, such as MSH3, MLH1, and MSH2, had some of the highest BiG-BETSs (Figure 2B). We noted that several DDR genes, such as ATR or CHEK1, which had previously been associated with elevated TMB by others,[9] were no longer found to have significant associations with elevated TMB by BiG-BETS (Z scores of -1.90 and -0.058, respectively; Figure S2B). A comparison of the number of genes significantly associated with elevated TMB by univariate (MWU) and our BiG-BETS permutation test demonstrates that BiG-BETS has drastically fewer genes that are significantly associated with elevated TMB (Figure 2A).

We next used BiG-BETS to compute DDR pathway level Z scores (Figure 2C). Unlike the univariate MWU approach, which found that all DDR pathways were significantly associated with elevated TMB (Figure S1A), BiG-BETS found that only mutations in the MMR and NER pathways were significantly associated with higher levels of TMB (Z score = 2.23 and Z score = 3.04, respectively; Figure 2C).

Interestingly, we found that, across all genes, the genes with the lowest BiG-BET Z scores were highly enriched for a number of biological processes (Figure 2D), including the negative regulation of cell proliferation and chromatin remodeling. This finding is congruent with the fact that cancer types with low overall levels of mutational burden are often driven by epigenetic changes and disruption in chromatin architecture.[17] Reassuringly, we see independent validation using a real-world clinically sequenced dataset (Samstein et al.[18]; see STAR Methods). The BiG-BET permutation-test-derived Z scores obtained using our approach correlated well between the pan-TCGA and Samstein datasets (using the 468 genes in common between the two datasets; Figure S2C). Finally, while the above analysis was performed using non-synonymous mutations defined to be of "moderate + high consequence" (see STAR Methods and Figure S3A), we found a high level of correlation between these BiG-BETSs and those derived from a more restrictive "high consequence" set of mutations (Figures S3A and S3B). There was also excellent agreement across the DDR genes and pathways (Figure S3B). Interestingly, a set of genes had BiG-BETSs that were noticeably lower when using the moderate + high consequence categorization (Figure S3B). We observed that the vast majority of them were genes with hotspot mutations (defined using a curated list from Miao et al.[19]). These hotspot mutations (i.e., *BRAF* ^{V600E}, *IDH1* ^{R132H}, etc.) are included in the moderate + high consequence set but excluded from the high consequence set. Further validating our method, past work has shown that some of these hotspot mutations, such as *EGFR*, *BRAF* ^{V600E}, and *IDH1* ^{R132H}, are associated with a lower TMB.[20–22]

Mutations in low BiG-BETS DDR genes add predictive power for ICB response in TMB high tumors

Given TMB's correlation with ICB response, it is not surprising that inactivation of some DDR genes or pathways has been associated with clinical benefit from ICB. For example, cancers with loss of the MMR machinery have increased response to pembrolizumab,[23–25] and mutations in specific genes, such as POLE[26] or BRCA2,[27] have also been shown to correlate with ICB response. However, whether DDR gene inactivation increases predictive power to ICB response over TMB alone is unclear, with one retrospective study suggesting increased predictive power of DDR mutation on top of TMB,[6] while another larger study from a large prospective phase II trial (IMvigor210) found that DDR mutations did not have predictive power for ICB response beyond its relationship to TMB.[7]

To assess whether inactivation of a specific DDR gene adds predictive power to ICB response over TMB alone, we first analyzed two large, annotated genomic datasets of ICB-treated patients, IMvigor210 (phase II trial of atezolizumab in urothelial cancer) and Samstein (real-world, pan-tumor, Memorial Sloan Kettering Cancer Center [MSKCC]).[7,18] There was reasonable overlap in DDR genes between the IMvigor210 and Samstein datasets. IMvigor210 and Samstein contained 19 and 31 DDR genes, respectively, with 17 DDR genes contained in both datasets (Figure S3C). We first validated that TMB correlated with overall survival (OS) (IMvigor210 and Samstein datasets; Figures S4A and S4B) and response (IMvigor210; Figure S4B). Given the current dogma that DDR mutations enhance ICB response through their potential to increase TMB,[3,5] we tested the hypothesis that mutations in high BiG-BETS DDR genes (i.e., those associated with high TMB) would correlate with increased ICB benefit merely because of their association with high TMB. In keeping with this notion, mutation in a high BiG-BETS DDR gene did not add any predictive power for response or OS in TMB high tumors (Figures S4C and S4D; IMvigor210 Cox proportional hazard [CPH] p = 0.34 and Samstein CPH p = 0.11).

In contrast, there was a strong interaction between mutation in a low BiG-BETS DDR gene and both response and OS with TMB as a covariate. Specifically, we found that, in patients with TMB high tumors, a mutation in a low BIG-BETS DDR gene was associated with the longer OS (Figures 3A and 3B) and increased response rate to ICB (Figure 3C). This effect on OS is seen within both the IMvigor210 and Samstein datasets (median survival 20.0 versus 10.7 months in IMvigor210 [Figure 3A] and 14.5 versus 10.0 months in Samstein [Figure 3B] for TMB high tumors with a low BiG-BETS DDR mutation versus wild type [WT]) and is reinforced in the clinical response data for IMvigor210, as 67% of TMB high tumors with a low BiG-BETS DDR mutation had ICB response versus a 37% response rate for TMB high tumors that were low BiG-BETS DDR WT (Figure 3C; p = 0.035). Similar results were seen when the analysis was done with only the subset of DDR genes that overlapped between the Samstein and IMvigor210 datasets (n = 17; Figures S3C and S5A).

Given the importance of this finding, we sought to further validate our BiG-BETS method. We therefore generated a metadataset (hereafter called "Weir_metadataset") of 424 patients from available cohorts on cBioportal, including melanoma,[27–29] non-small-cell lung cancer (NSCLC),[30,31] and a recently published real-world cohort of bladder cancer patients from our own institution.[32] Using the Weir_metadataset, we observed similar findings as we saw in the Samstein and IMvigor210 datasets. Specifically, patients with TMB high tumors and mutation in a

CellPress
OPEN ACCESS

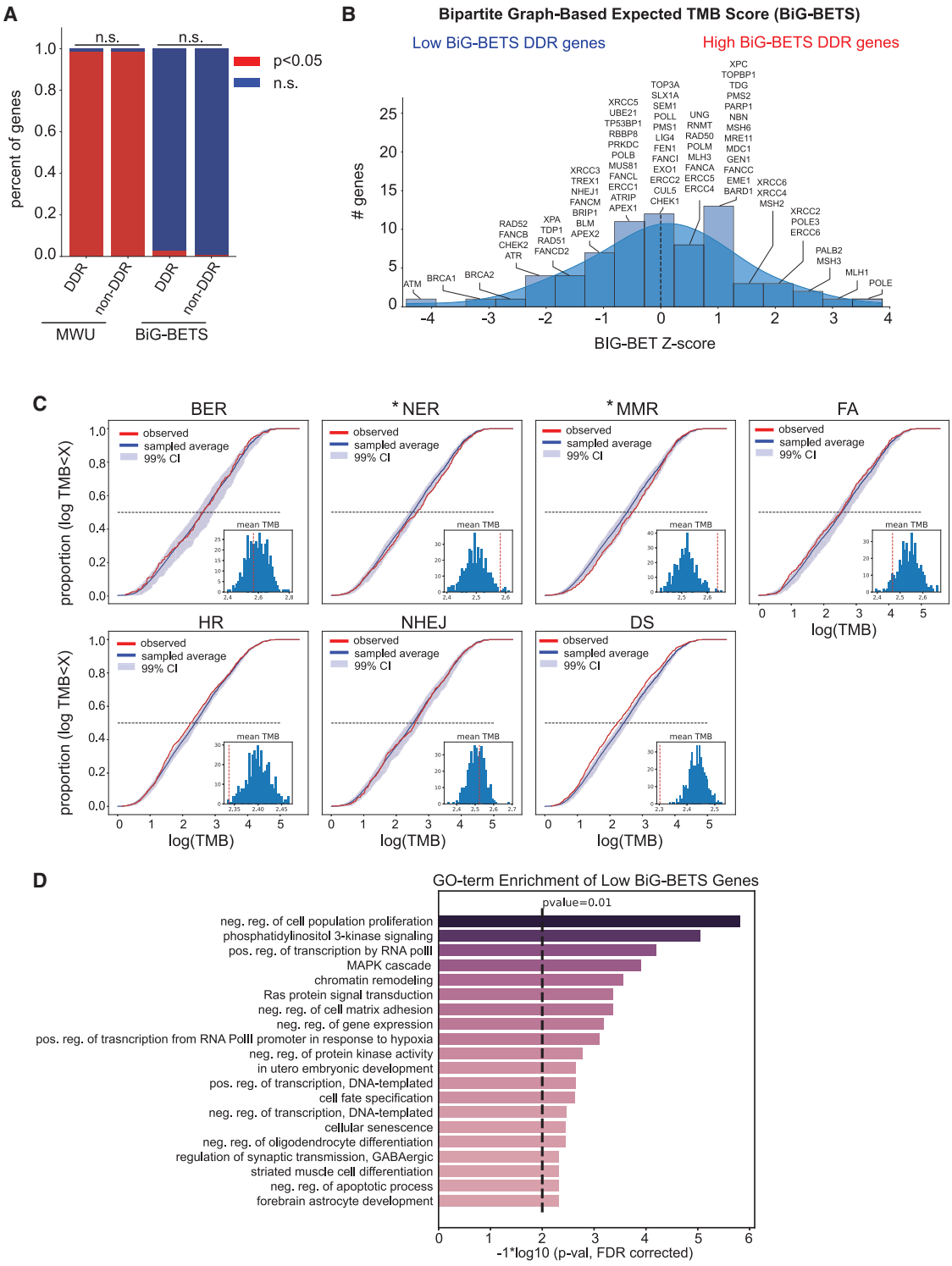


Figure 2. A networks-based model and permutation test (BiG-BETS) are superior to univariate model

(A) Percentage of genes that are significant (with multiple test correction) using MWU test (left two bars) and significant by the networks-based test (right two bars). Differences between DDR and non-DDR genes computed were assessed with chi-squared test with p value shown above the corresponding bars; n.s., not significant.

(B) Distribution of BiG-BETS *Z* scores for the DDR genes. Low versus high *Z* score DDR genes are defined using *Z* score < 0 and *Z* score > 0 (dashed vertical line), respectively, with individual genes in each bin listed about the plot.

(legend continued on next page)

low BiG-BETS DDR gene had prolonged OS (median OS 19.3 versus 14.5 months; Figure 3D) and increased response (response rate [RR] 48% versus 29%; Figure 3E) relative to those with TMB high tumors WT for low BiG-BETS DDR genes. Therefore, mutation in a low BiG-BETS DDR gene has additional predictive power for ICB benefit over TMB alone. This observation seems to be specific to DDR genes, as when we performed a parallel analysis using chromatin remodeling genes from the Gene Ontology Resource (GO term, chromatin remodeling; GO: 0006338) we found no association between having a mutation in a low BiG-BETS, chromatin remodeling gene and improved clinical outcome from ICB in high TMB tumors (Figure S5B).

Mutation in low BiG-BETS DDR genes in TMB high tumors is not merely prognostic and is predictive across individual tumor types

To ensure that mutation in a low BiG-BETS DDR gene in a TMB high tumor is not merely prognostic, we analyzed the subset of TCGA tumors that overlap with the ICB-treated tumor types from the Samstein dataset (see STAR Methods). Since the vast majority of the TCGA tumors were collected and sequenced prior to the widespread use of ICB, they should serve as an ICB naive cohort. To this end, we looked at OS by TMB and BiG-BETS DDR mutation status in the TCGA Samstein overlap cohort (Figure 3F). In this non-ICB-treated cohort, there were no differences in survival in TMB high tumors based on low BiG-BETS DDR gene mutation status (hazard ratio = 0.73 [0.46–1.15]; CPH p = 0.177). Therefore, the BiG-BETS score is not merely prognostic.

We wanted to understand whether the clinical benefit from ICB seen in TMB high, low BiG-BETS DDR mutant tumors is present within individual tumor types. Of the individual tumor types within the Samstein dataset, most tumor types did not have enough TMB high, low BiG-BETS DDR mutant samples to see the conditional effect. Only NSCLC and bladder cancer had more than n = 5 samples with high TMB, low BiG-BETS DDR mutations. We therefore combined patients from the Weir_meta-dataset with the Samstein and the IMvigor210 cohorts (hereafter called Weir_combined). Analysis of the Weir_combined dataset showed that NSCLC patients with TMB high, low BiG-BETS DDR mutations had a significantly prolonged OS (CPH p = 0.001), while melanoma and bladder cancer, while not significant, showed a similar trend (Figures 3G and 3I). We also assessed our method on a kidney cancer dataset[33] as well given the frequent treatment of renal cell carcinoma (RCC) patients with ICB. Interestingly, RCC patients did not appear to show any difference in OS or response on the basis of either TMB or low BiG-BETS DDR mutation status (Figure S5C). We hypothesize that this might reflect that ICB response in RCC is because neoantigens are driven by frameshift mutations rather than single-nucleotide variants (SNVs)[34] or that other tumor-associated antigens (like cancer testis antigens or endogenous retroviruses) mediate the responses seen in this unique tumor type. Therefore, we believe that mutations in low BiG-BETS DDR genes in TMB high tumors correlate with enhanced clinical benefit to ICB across individual tumor types.

Tumors with mutation in low BiG-BETS DDR genes have enhanced STING activity

We hypothesized that alterations in these low BiG-BETS DDR genes may mediate an anti-tumor immune response through enhanced innate immunity or antigen presentation or possibly through production of neo-antigens in a way that is not captured by TMB alone.[3,5] Indeed, we saw that, in TMB high patients from the IMvigor210 dataset, low BiG-BETS DDR mutant tumors had elevated gene signatures scores of both STING as well as its key downstream transcriptional mediator, interferon regulatory factor 3 (IRF3) (Figure 4A), but interestingly did not show significant changes in other gene signatures associated with ICB response (i.e., CD8 T cell, EMT-Stroma, transforming growth factor β [TGF-β], or interferon gamma [IFNG] signatures; Figures 4B–4D). We noted that TMB high, low BiG-BETS DDR mutant tumors also had significantly elevated STING gene signature in the TCGA (with cancer types matched to Samstein; Figures S6A–S6D), though EMT_stroma and fibroblast TGF-beta response signatures (FTBRS) were also significantly different in this case. In aggregate, these data demonstrate that TMB high patients with a mutation in a low BiG-BETS DDR gene have increased response (67%) and prolonged OS when treated with ICB, potentially due to enhanced baseline STING activity.

DISCUSSION

In summary, we identified a sampling bias in currently used methods to test for an association between mutation in a specific gene or pathway and elevated TMB. Application of our method, BiG-BETS, accurately defines DDR genes associated with high and low TMB (high BiG-BETS DDR and low BiG-BETS DDR genes, respectively) and identifies that only inactivation of the MMR and NER DDR pathways are truly associated with elevated TMB. Moreover, we demonstrate that mutation in high BiG-BETS DDR genes do not hold predictive value for ICB response in TMB high tumors, because their predictive value is driven by their association with elevated TMB. In contrast, and of clinical importance, is that in TMB high tumors, mutation in a low BiG-BETS DDR gene is significantly associated with elevated STING pathway activity, increased ICB response (up to 67%), and prolonged OS benefit from ICB treatment.

Recent work by Hsiehchen and colleagues also examined correlation between DDR pathways and ICB benefit across a large

(C) Application of bipartite configuration test to the DDR pathways in the TCGA data. Each subplot shows the observed cumulative distribution of TMB for samples with a mutation in the genes of the specified pathway by the red solid line. The blue line shows the average cumulative distribution across 400 sampled networks, with the light blue band showing the 99% confidence interval (CI). Horizontal line at y = 0.5 denotes the median TMB for the distributions. Inset figures show a histogram of the means of the sampled distributions of TMB for samples with a mutation in the corresponding DDR pathway. The vertical red dashed line within the inset depicts the observed mean TMB in the actual dataset. *Z* scores were constructed by comparing the observed mean TMB with the sampled means.

(D) Significant gene ontology (GO) terms identified in the 50 lowest BiG-BET genes from the pan-TCGA dataset. p values corrected using Benjamini-Hochberg.

CellPress
OPEN ACCESS

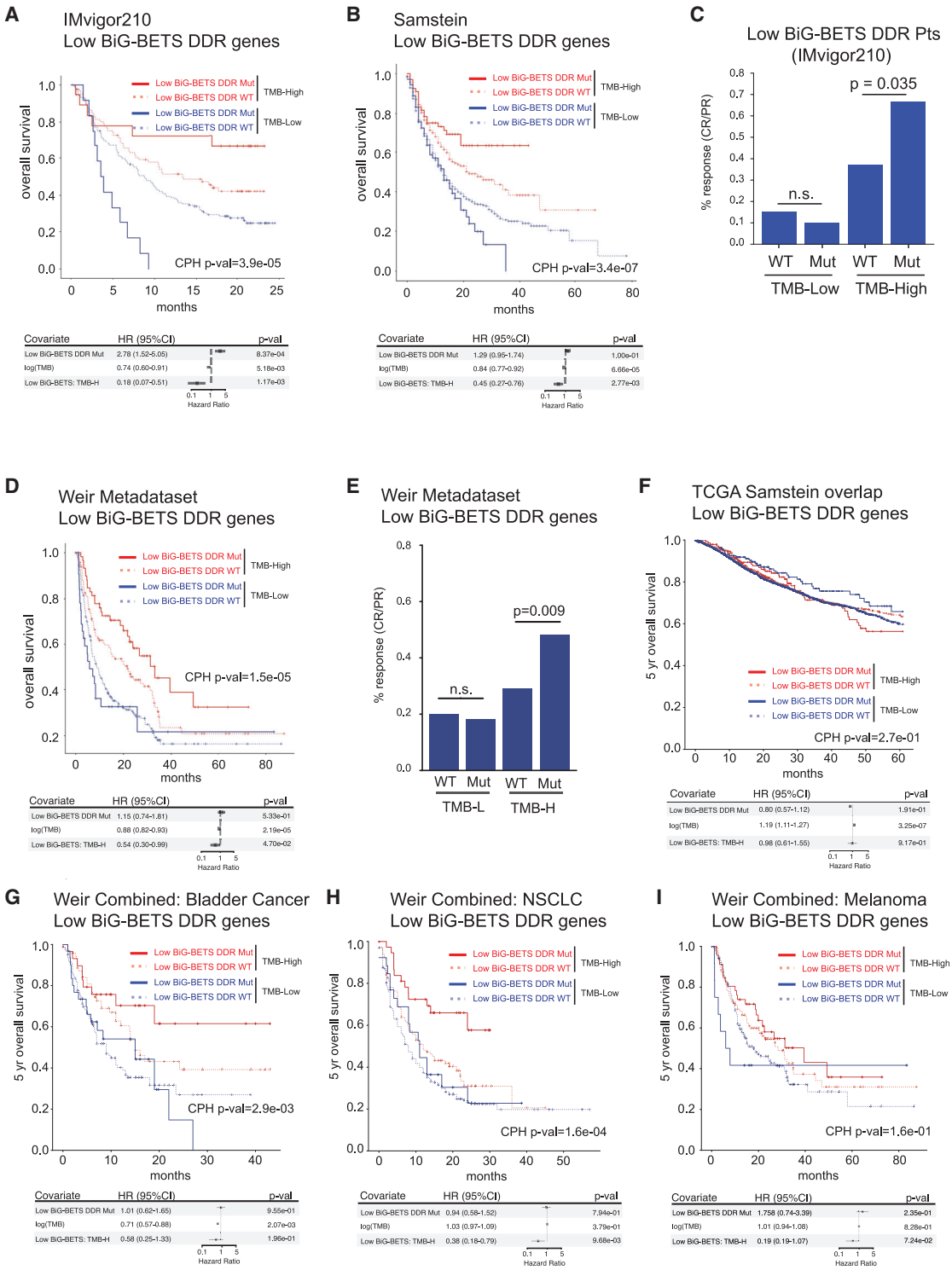


Figure 3. TMB high tumors with mutation in a low BiG-BETS DDR gene have improved survival and response

(A) Kaplan-Meier curves depicting overall survival (OS) in the IMvigor210 cohort broken down by TMB high (red lines) versus TMB low (blue lines) and into samples with a mutation in low BiG-BETS DDR genes (bold lines) and low BiG-BETS DDR WT tumors (dotted lines). Significance for survival curves determined by log likelihood ratio test of Cox-proportional harzards model. Table underneath shows forest plot of coefficients for CPH model jointly testing TMB (as continuous variable), mutation in low BiG-BETS DDR genes, and an interaction term between the two variables (denoted by low BiG-BETS; TMB-H). Patient counts for each category in TMB-H_MUT, TMB-H_WT, TMB-L_MUT, and TMB-L_WT were 19, 84, 12, and 159, respectively.

(legend continued on next page)

dataset of tumors with genomic annotation and found that patients whose tumors harbored mutations in the NER and HR pathways were associated with higher RR and OS.[35] Our work is distinct, as our primary objective was to faithfully define DDR genes that associate with low and high TMB status. We have found that patients with TMB high tumors and mutation in low BiG-BETS DDR gene have the best OS and RR. There are a number of important differences between our study and Hsiehchen and colleagues, including the number of DDR genes used (Hsiehchen n = 40; our study n = 72) as well as the classification of DDR genes into specific DDR pathways. Finally, while the authors suggest that the NER and HR mutations predict OS independent of TMB, we note that all of the NER genes from their paper have high BiG-BETS scores in our analysis. Therefore, by our analysis, mutations of genes in the NER pathway are associated with elevation in TMB.

Our BiG-BETS method and clinical observations have potentially important implications for patient care. First, as mentioned above, patients with TMB high, low BiG-BETS DDR mutant tumors have significantly increased ICB response (67%) and prolonged OS and should therefore be assessed in future clinical trials. Moreover, this tandem predictive biomarker can likely be improved upon with integration of other immunogenomic features, such as the integration of a T cell inflamed gene expression profile, which has shown predictive value for ICB response in pan tumor analysis[2] and is not correlated with TMB.[36] Second, a number of the low BiG-BETS DDR genes are kinases, multiple of which have small-molecule inhibitors in late-stage clinical trials (i.e., ATM, CHEK1, and WEE1). One would predict that treatment of a TMB high tumor that is *low* BiG-BETS DDR WT with one of these kinase inhibitors (in an attempt to replicate a low BiG-BETS DDR mutation) and ICB might mimic our genetic findings. In contrast, our method predicts that inhibition of *high* BiG-BETS DDR genes that are kinases (i.e., ATR) in combination with ICB would not necessarily benefit patients with TMB high tumors.

There is much clinical interest in examining the efficacy of ICB in patients with DDR mutations. Moreover, there are numerous clinical trials underway combining DDR inhibitors with ICB. These trials often use broad panels of DDR genes as inclusion criteria or focus on a specific DDR pathway (i.e., HR). Our data suggest that restricting biomarker selection to a specific DDR pathway is not wise, as our low BiG-BETS DDR genes are relatively evenly spread across all of the DDR pathways (Figure S6E). While a unifying explanation for how our low BiG-BETS DDR genes, which are spread across all DDR pathways, are associated with enhanced STING activity is challenging to envision, we see and validate this finding in both the IMvigor210 and TCGA datasets. Finally, our work is cautionary, as it suggests that, without selection of both (1) the subset of DDR genes with predictive power for ICB response (low BiG-BETS DDR genes) as well as (2) patients with TMB high tumors, trials integrating DDR mutations and inhibitors with ICB may show a lack of clinical benefit.

Limitations of the study

Limitations of the study include that it is based upon retrospective cohorts of patients, which can introduce a number of biases. It is therefore important that prospective validation of mutant low BiG-BETS DDR genes in TMB high tumors be carried out.

SYNOPSIS

Herein we develop a test, the bipartite graph-based expected TMB score (BiG-BETS), that resolves the TMB paradox, accurately defines genes associated with elevated TMB, and remarkably delineates a cohort of patients (TMB high, low BiG-BETS DDR mutant) with high predictive power for ICB response and prolonged overall survival.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Description of the datasets
  - DNA damage repair pathways
  - Chromatin remodeling pathway
  - Definition of high and moderate + high consequence mutations
  - Bipartite-graph based-expected TMB score (BiG-BETS)

(B) Kaplan-Meier (KM) curves depicting OS in the Samstein et al. cohort broken down along the same lines as (A) with corresponding coefficients in CPH model below. Patient counts for each category in TMB-H_MUT, TMB-H_WT, TMB-L_MUT, and TMB-L_WT were 67, 307, 72, and 861, respectively.
(C) Percentage of patients with response (complete or partial response) to ICB in the IMvigor210 dataset. Patients are stratified by TMB (TMB high versus TMB low) and into samples with a low BiG-BETS DDR gene mutation or not (WT). The number of patients who were responders (complete response [CR] or partial response [PR]) in each category from left to right is 12, 1, 28, and 10, respectively. Significant differences between groups were tested using chi-squared.
(D) KM curves depicting OS in the Weir metadataset (see STAR Methods for full description) broken down along the same lines as (A) and (C) with corresponding coefficients in CPH model below. Patient counts for each category in TMB-H_MUT, TMB-H_WT, TMB-L_MUT, and TMB-L_WT were 60, 117, 33, and 201, respectively.
(E) Response rates by low BiG-BETS DDR mutations in the Weir metadataset.
(F) Kaplan-Meier (KM) curves depicting OS in the TCGA samples (using tumor types overlapping with Samstein et al.) cohort broken down along the same lines as (A) with corresponding coefficients in CPH model below.
(G–I) KM curves depicting OS in a combined dataset that includes IMVigor210, Samstein et al., and Weir metadataset split out by tumor type, including (G) bladder cancer, (H) non-small cell lung cancer, and (I) melanoma. Each plot is broken down along the same lines as (A) and (B) with corresponding coefficients in CPH model below.

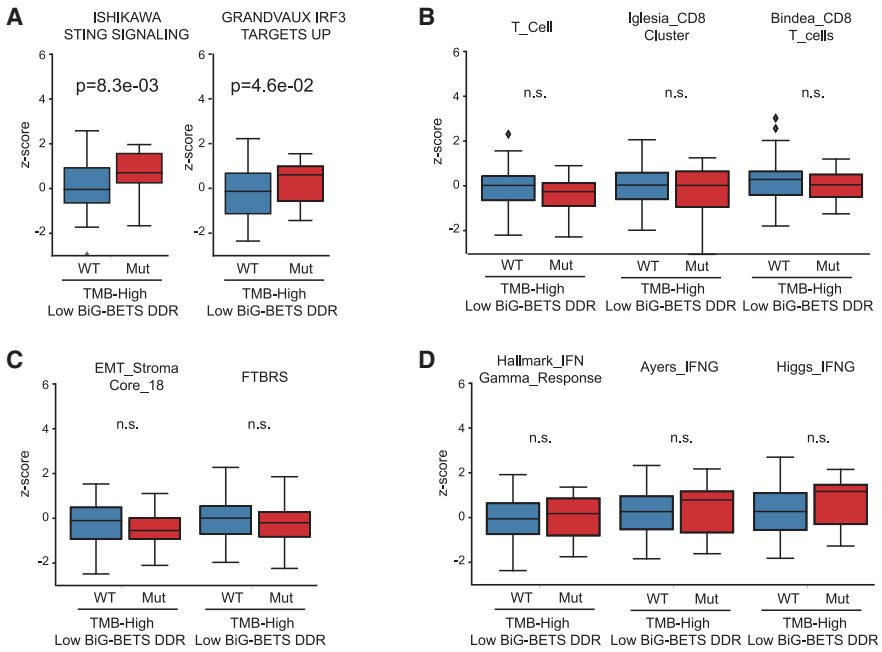


Figure 4. Mutation of low BiG-BETS DDR genes in TMB high tumors is associated with elevated STING and IRF3 gene signatures

A–D) Boxplots of indicated gene signatures in IMVigor210 patients stratified by TMB (TMB high versus TMB low) and low BIG-BETS DDR gene mutation or not (WT). For each signature, each sample is assigned a *Z* score based on the average expression level of all genes in the signature compared with the average across all samples (see STAR Methods). Significance was calculated using the Mann-Whitney U test with diamonds representing outliers (data>1.5*IQR).

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Survival and response analyses
  - Gene expression signatures analysis
  - Proof of the friendship paradox for bipartite network

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100602.

ACKNOWLEDGMENTS

We thank Ben Vincent for his insights and acknowledge the members of the Kim and Mucha Labs for useful discussions. This work was supported by the University Cancer Research Fund (UCRF) (W.Y.K.), the James S. McDonnell Foundation 21st Century Science Initiative (Complex Systems Scholar Award grant no. 220020315; W.H.W. and P.J.M.), and by grant R01DK111930 from the National Institute of Diabetes and Digestive and Kidney Diseases. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding organizations.

AUTHOR CONTRIBUTIONS

Conceptualization, W.H.W. and W.Y.K.; methodology, W.H.W. and P.J.M.; software, W.H.W.; validation, W.H.W. and W.Y.K.; formal analysis, W.H.W. and P.J.M.; investigation, W.H.W.; resources, none; data curation, W.H.W.; writing – original draft, W.H.W., P.J.M., and W.Y.K.; writing – reviewing and editing, W.H.W., P.J.M., and W.Y.K.; visualization, W.H.W.; supervision, P.J.M. and W.Y.K.; project administration, W.H.W.; funding acquisition, P.J.M. and W.Y.K.

DECLARATION OF INTERESTS

W.H.W., P.J.M., and W.Y.K. have filed a provisional patent on the use of networks-based approaches to identify predictors of immunotherapy response.

Received: July 14, 2021
Revised: January 7, 2022
Accepted: March 20, 2022
Published: April 14, 2022

REFERENCES

1. Tran, G., and Zafar, S.Y. (2018). Financial toxicity and implications for cancer care in the era of molecular and immune therapies. Ann. Transl. Med. *6*, 166.
2. Cristescu, R., Mogg, R., Ayers, M., Albright, A., Murphy, E., Yearley, J., Sher, X., Liu, X.Q., Lu, H., Nebozhyn, M., Zhang, C., et al. (2018). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science *362*, eaar3593.
3. Keenan, T.E., Burke, K.P., and Allen, E.M.V. (2019). Genomic correlates of response to immune checkpoint blockade. Nat. Med. *25*, 389–402.
4. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., Boyault, S., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.
5. Mouw, K.W., Goldberg, M.S., Konstantinopoulos, P.A., and D'Andrea, A.D. (2017). DNA damage and repair biomarkers of immunotherapy response. Cancer Discov. *7*, 675–693, Available from: http://cancerdiscovery.aacrjournals.org/content/7/7/675.article-info.
6. Teo, M.Y., Seier, K., Ostrovnaya, I., Regazzi, A.M., Kania, B.E., Moran, M.M., Cipolla, C.K., Bluth, M.J., Chaim, J., Al-Ahmadie, H., Snyder, A., et al. (2018). Alterations in DNA damage response and repair genes as

potential marker of clinical benefit from PD-1/PD-L1 blockade in advanced urothelial cancers. J. Clin. Oncol., Available from: http://ascopubs.org/doi/abs/10.1200/JCO.2017.75.7740#affiliationsContainer.

7. Mariathasan, S., Turley, S.J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel, E.E., III, Koeppen, H., Astarita, J.L., Cubas, R., Jhunjhunwala, S., et al. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature *554*, 544–548.

8. Wang, J., wang, Z., Zhao, J., Wang, G., Zhang, F., Zhang, Z., Zhang, Y., Dong, H., Zhao, X., Duan, J., Bai, H., et al. (2018). Co-mutations in DNA damage response pathways serve as potential biomarkers for immune checkpoint blockade. Cancer Res. https://doi.org/10.1158/0008-5472.CAN-18-1814.

9. Parikh, A.R., He, Y., Hong, T.S., Corcoran, R.B., Clark, J.W., Ryan, D.P., Zou, L., Ting, D.T., Catenacci, D.V., Chao, J., Fakih, M., et al. (2019). Analysis of DNA damage response gene alterations and tumor mutational burden across 17,486 tubular gastrointestinal carcinomas: implications for therapy. Oncologist *24*, 1340–1347.

10. Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., Liu, Y., et al. (2018). Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas. Cell Rep. *23*, 239–254.e6.

11. Chae, Y.K., Anker, J.F., Carneiro, B.A., Chandra, S., Kaplan, J., Kalyan, A., Santa-Maria, C.A., Platanias, L.C., and Giles, F.J. (2016). Genomic landscape of DNA repair genes in cancer. Oncotarget *7*, 23312–23321.

12. Chalmers, Z.R., Connelly, C.F., Fabrizio, D., Gay, L., Ali, S.M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., et al. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med. *9*, 480, Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0424-2.

13. Chae, Y.K., Anker, J.F., Bais, P., Namburi, S., Giles, F.J., and Chuang, J.H. (2017). Mutations in DNA repair genes are associated with increased neoantigen load and activated T cell infiltration in lung adenocarcinoma. Oncotarget *9*, 7949–7960.

14. Feld, S.L. (1991). Why your friends have more friends than you do. American Journal of Sociology *96*, 1464–1477, Available from: http://www.jstor.org/stable/2781907

15. Fosdick, B.K., Larremore, D.B., Nishimura, J., and Ugander, J. (2018). Configuring random graph models with fixed degree sequences. Siam Rev. *60*, 315–355.

16. Saracco, F., Clemente, R.D., Gabrielli, A., and Squartini, T. (2015). Randomizing bipartite networks: the case of the world trade web. Sci. Rep. *5*, 10595.

17. Gonzalez-Perez, A., Jene-Sanz, A., and Lopez-Bigas, N. (2013). The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. Genome Biol. *14*, r106.

18. Samstein, R.M., Lee, C.-H., Shoushtari, A.N., Hellmann, M.D., Shen, R., Janjigian, Y.Y., Barron, D.A., Zehir, A., Jordan, E.J., Omuro, A., Kaley, T.J., et al. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat. Genet. *51*, 202–206.

19. Miao, D., Margolis, C.A., Vokes, N.I., Liu, D., Taylor-Weiner, A., Wankowicz, S.M., Adeegbe, D., Keliher, D., Schilling, B., Tracy, A., Manos, M., et al. (2018). Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. Nat. Genet. *50*, 1271–1281.

20. Battaglin, F., Xiu, J., Baca, Y., Shields, A.F., Goldberg, R.M., Seeber, A., Habib, D., Soni, S., Puccini, A., Tokunaga, R., Arai, H., et al. (2020). Comprehensive molecular profiling of IDH1/2 mutant biliary cancers (BC). J. Clin. Oncol. *38*, 479.

21. Mar, V.J., Wong, S.Q., Li, J., Scolyer, R.A., McLean, C., Papenfuss, A.T., Tothill, R.W., Kakavand, H., Mann, G.J., Thompson, J.F., Behren, A., et al. (2013). BRAF/NRAS wild-type melanomas have a high mutation load correlating with histologic and molecular signatures of UV damage. Clin. Cancer Res. *19*, 4589–4598.

22. Negrao, M.V., Skoulidis, F., Montesion, M., Schulze, K., Bara, I., Shen, V., Xu, H., Hu, S., Sui, D., Elamin, Y.Y., Le, X., et al. (2021). Oncogene-specific differences in tumor mutational burden, PD-L1 expression, and outcomes from immunotherapy in non-small cell lung cancer. J. Immunother. Cancer *9*, e002891.

23. Lemery, S., Keegan, P., and Pazdur, R. (2017). First FDA approval agnostic of cancer site — when a biomarker defines the indication. N. Engl. J. Med. *377*, 1409–1412.

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

25. Le, D.T., Durham, J.N., Smith, K.N., Wang, H., Bartlett, B.R., Aulakh, L.K., Lu, S., Kemberling, H., Wilt, C., Luber, B.S., Wong, F., et al. (2017). Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. Science, eaan6733, Available from: http://science.sciencemag.org/content/early/2017/06/07/science.aan6733?utm_campaign=fr_sci_2017-06-08&et_rid=35386688&et_cid=1373712.

26. Mehnert, J.M., Panda, A., Zhong, H., Hirshfield, K., Damare, S., Lane, K., Sokol, L., Stein, M.N., Rodriguez-Rodriquez, L., Kaufman, H.L., Ali, S., et al. (2016). Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer. J. Clin. Invest. *126*, 2334–2340.

27. Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., et al. (2016). Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell *165*, 35–44.

28. Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S., Hollmann, T.J., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. N. Engl. J. Med. *371*, 2189–2199.

29. Allen, E.M.V., Miao, D., Schilling, B., Shukla, S.A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M.H.G., Goldinger, S.M., Utikal, J., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science *350*, 207–211.

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

31. Hellmann, M.D., Nathanson, T., Rizvi, H., Creelan, B.C., Sanchez-Vega, F., Ahuja, A., Ni, A., Novik, J.B., Mangarin, L.M.B., Abu-Akeel, M., Liu, C., et al. (2018). Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer. Cancer Cell *33*, 843–852.e4.

32. Rose, T.L., Weir, W.H., Mayhew, G.M., Shibata, Y., Eulitt, P., Uronis, J.M., Zhou, M., Nielsen, M., Smith, A.B., Woods, M., Hayward, M.C., et al. (2021). Fibroblast growth factor receptor 3 alterations and response to immune checkpoint inhibition in metastatic urothelial cancer: a real world experience. Br. J. Cancer, 1–10.

33. Braun, D.A., Hou, Y., Bakouny, Z., Ficial, M., Angelo, M.S., Forman, J., Ross-Macdonald, P., Berger, A.C., Jegede, O.A., Elagina, L., Steinharter, J., et al. (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. Nat. Med. *26*, 909–918.

34. Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Bakir, M.A., Chambers, T., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. *18*, 1009–1021.

35. Hsiehchen, D., Hsieh, A., Samstein, R.M., Lu, T., Beg, M.S., Gerber, D.E., Wang, T., Morris, L.G.T., and Zhu, H. (2020). DNA repair gene mutations as predictors of immune checkpoint inhibitor response beyond tumor mutation burden. Cell Rep. Med. *1*, 100034.

36. Spranger, S., Luke, J.J., Bao, R., Zha, Y., Hernandez, K.M., Li, Y., Gajewski, A.P., Andrade, J., and Gajewski, T.F. (2016). Density of immunogenic antigens does not explain the presence or absence of the T-cell-inflamed tumor microenvironment in melanoma. Proc. Natl. Acad. Sci., 201609376.

37. Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., Sofia, H.J., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *6*, 271–281.e7.

38. Miao, D., Margolis, C.A., Gao, W., Voss, M.H., Li, W., Martini, D.J., Norton, C., Bossé, D., Wankowicz, S.M., Cullen, D., Horak, C., et al. (2018). Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. Science, eaan5951, Available from: http://science.sciencemag.org/content/early/2018/01/03/science.aan5951?utm_campaign=fr_sci_2018-01-04&et_rid=35386688&et_cid=1771523.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA MC3 Mutations Calls	TCGA	https://gdc.cancer.gov/about-data/publications/pancanatlas
TCGA Expression Data	TCGA	https://gdc.cancer.gov/about-data/publications/pancanatlas
TCGA TMB Annotation	TCGA	https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin
IMVigor210	Mariathansan et al.	http://research-pub.gene.com/IMvigor210CoreBiologies/
Samstein et al.	Samstein et al.	https://www.cbioportal.org/study/summary?id=tmb_mskcc_2018
Weir Metadataset	This paper	Table S3
DDR Classification	Knijnenburg et al.	https://www.cell.com/cms/10.1016/j.celrep.2018.03.076/attachment/9ca123d4-fe1c-4849-b461-c1e549c6b57d/mmc2.xlsx
Software and algorithms		
BiG-BETS source code	this paper	https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/GBFHPB

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. William Y. Kim (wykim@med.unc.edu).

Materials

This study did not generate new unique reagents.

Data and code availability

- Standardized Datasets: This work did not generate any novel standardized datasets.
- Custom Computer Code: All original code has been released at both https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/GBFHPB as well on GitHub (https://github.com/wweir827/BIGBETS) and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

No experimental models were used in this work.

METHOD DETAILS

Description of the datasets

The cancer genome atlas (TCGA) pan-cancer MC3 dataset

The primary dataset we used for developing our method was the TCGA-pancan unified ensemble MC3 variant call set (downloadable at https://gdc.cancer.gov/about-data/publications/pancanatlas). See https://www.synapse.org/#!Synapse:syn7214402/wiki/405297 for further description.). This dataset used Whole Exome Sequence (WES) tumor samples from all TCGA centers and variants

were re-called in a uniform pipeline. This dataset includes 3.6 million, small variants from 10,295 tumor samples. This dataset is described further in Ellrott et al.[37] Filtering this list of variants on Moderate + High consequence (defined above) resulted in 1,000,011 variants in 19,255 genes in 10,164 different samples, while keeping High consequence variants resulted in 208,682 remaining variants in 18,284 different genes from 9,530 different samples. Figure S3B demonstrates excellent correlation in the BiG-BET score between the high impact and the high + moderate impact datasets, especially with regards to the DDR genes and pathways.

TMB values for TCGA were obtained from (https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin) combining both Silent and Non-Silent scores for each sample.

Clinical data used for survival analysis of the TCGA were obtained from the TCGA clinical data resource as detailed in Liu et al. 2018. We looked at the effect of low and high BiG-BET DDR mutations on overall 5-year survival as detailed above. We filtered the cohort to the cancer types that best reflected the composition of Samstein et al., keeping the following TCGA types: LUAD, LUSC, BRCA, SKCM, COAD, ESCA, KIRC, BLCA, and HNSC. This resulted in 4,287 samples. To calculate gene signatures profiles, the TCGA-PanCan expression data was obtained from https://gdc.cancer.gov/about-data/publications/pancanatlas and processed as described above in Gene expression signatures analysis.

IMvigor210

The IMvigor210 trial is a Phase II single arm study examining the response of patients with locally advanced or metastatic urothelial bladder cancer to atezoliziumab (anti PD-L1). A full description of the characteristics of the patient cohort can be found in.[7] We have used the publicly available dataset released by Mariathansan et al. which can be accessed at http://research-pub.gene.com/IMvigor210CoreBiologies/. The cohort consists of 260 patients with 1249 short variants across 160 different genes. Because less detailed annotations were available, we did not filter any of the mutations from this cohort.

Samstein et al. Cohort

To validate our clinical findings, we used a large, multi-trial cohort consisting of 1661 patients treated with different Immune Checkpoint Blockade (ICB) therapies and with targeted clinical sequencing, first compiled and analyzed in.[18] Sequencing was performed using the Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) panel. We downloaded the data from cbioportal using the following link: https://www.cbioportal.org/study/summary?id=tmb_mskcc_2018. We filtered down to the 1307 patients that received anti PD-1/PD-L1 therapies and kept the variants with one of the following high impact consequences: missense mutation, nonsense mutation, frameshift deletion, frameshift insertion, translation start site, or nonstop mutation. This resulted in a total of 19,057 variants in 468 different genes.

Weir meta-dataset

As an additional validation set, we also compiled a *meta*-dataset from several different studies available on cBioportal as well as a recently published study from our group. This *meta*-dataset included melanoma,[27–29] non-small cell lung cancer[30,31] and clear cell renal carcinoma.[38] The datasets from cbioportal can be accessed using the following link: https://www.cbioportal.org/study/summary?id=ccrcc_dfci_2019,skcm_mskcc_2014,skcm_dfci_2015,mel_ucla_2016,nsclc_mskcc_2018,nsclc_mskcc_2015. Additionally, we included a real-world metastatic urothelial carcinoma cohort from UNC.[32] The Weir metadataset included 407 total samples with 16,250 variants in 599 genes. As most of the datasets on cBioportal did not include TMB values, we used the total number of mutations for each sample as a surrogate. For our analyses, we considered a tumor to be TMB-H if it was in the top 50% of tumors by total mutation count. The compiled dataset is included in Table S3.

DNA damage repair pathways

We have relied on the core DNA Damage Repair pathways defined by Knijnenburg et al.[10] to conduct all of our pathway level analysis. The pathways are defined as follows:

BER	NER	MMR	FA	HR		NHEJ	DS
PARP1	CUL5	EXO1	FANCA	MRE11	EME1	LIG4	ATM
POLB	ERCC1	MLH1	FANCB	NBN	GEN1	NHEJ1	ATR
APEX1	ERCC2	MLH3	FANCC	RAD50	MUS81	POLL	ATRIP
APEX2	ERCC4	MSH2	FANCD2	TP53BP1	PALB2	POLM	CHEK1
FEN1	ERCC5	MSH3	FANCI	XRCC2	RAD51	PRKDC	CHEK2
TDG	ERCC6	MSH6	FANCL	XRCC3	RAD52	XRCC4	MDC1
TDP1	POLE	PMS1	FANCM	BARD1	RBBP8	XRCC5	RNMT
UNG	POLE3	PMS2	UBE2T	BLM	SHFM1	XRCC6	TOPBP1
	XPA			BRCA1	SLX1A		TREX1
	XPC			BRCA2	TOP3A		
				BRIP1			

Chromatin remodeling pathway

We also looked at genes associated with chromatin remodeling as annotated by the Gene Ontology (GO) project (http://geneontology.org/). We selected all genes associated with the GO:0006338 – 'chromatin_remodeling' or any of its associated sub-term, resulting in a set of 267 genes given in Table S4.

Definition of high and moderate + high consequence mutations

We started with the MC3 mutational dataset provided by the TCGA (https://gdc.cancer.gov/about-data/publications/pancanatlas). The TCGA MC3 dataset was then filtered to include only ''High Consequence'' non-synonymous mutations, which were defined as being categorized as a high consequence mutation by the Sequence Ontology and summarized by Ensembl here (https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html) ['stop lost', 'stop gained', 'transcript ablation', 'start lost', 'frameshift variant', 'splice_site', 'translation_start_site'] and were also categorized as having a PolyPhen score of 'probably damaging', 'possibly damaging', or 'unknown'.

We defined non-synonymous mutations of ''Moderate + High Consequence'' as moderate ['inframe insertion', inframe deletion', missense variant', and 'protein altering variant'] or high ['stop lost', 'stop gained', 'transcript ablation', 'start lost', 'frameshift variant', 'splice_site, 'translation_start_site'] consequence mutations by the Sequence Ontology and were also categorized as having a PolyPhen score of 'probably damaging', 'possibly damaging', or 'unknown' (Figure S3A).

Bipartite-graph based-expected TMB score (BiG-BETS)

We converted the Pan-Cancer TCGA tumors and their respective mutated genes into a bipartite network, where genes and tumors represent the two classes of nodes and the edges (connecting two nodes) indicate the mutated genes within a given tumor (Figure 1E). By re-casting our data into a bipartite network, we see in Figure 1E that the TMB for a sample is proportional to the number of edges it has (i.e. its degree) and that the average TMB associated with mutation in a given gene is essentially average degree of its neighbors (Figure S1B).

The bipartite network representation was used to derive a null model of TMB distribution for each gene. Specifically, random sampling (permutations) of the bipartite network was performed by stochastically rewiring the network while maintaining the degree distribution (number of edges of each node) of the original dataset (this null model is known as the configuration model, which for a bipartite network is further constrained to maintain the bipartite nature of the network).[15,16] Generation of a null model through permutation allowed us to compare the actual mean TMB for tumors with mutation in a given gene or pathway against the expected distribution under random sampling (Figure 1F).

The BiG-BET score consists of comparing the observed mean TMB for each gene or pathway's mutated sample set against the expected distribution under random sampling of bipartite networks that match the degree distribution of the original dataset. The null model for networks in which all networks with a given degree sequence are uniformly likely is known as the configuration model,[24] which has also been extended to bipartite graphs.[15] The bipartite configuration model can be envisioned by cutting across the edges in the original network and reconnecting the ''stubs'' at random with each possible set of pairings respecting the bipartite structure of the original network and being equally likely under the model (visualized in Figure 1F). We note that we have applied a rewiring procedure as described in[15] to sample this bipartite configuration model rather than the more direct ''stub matching'' approach to ensure that we sample the appropriate, more restricted model without self-loops and multi-edges.[3–5] We iterate the following steps:

1. Select two edges at random in the network: $(g_i,s_x) \epsilon \varepsilon$ and $(g_j,s_y) \epsilon \varepsilon$
2. Confirm that each edge is connected to a distinct pair of nodes: $g_i \neq g_j$ and $s_x \neq s_y$. If the two edges involve either the same sample node, or the same gene node, repeat 1.
3. Swap which gene is connected to which sample from these two edges. Add (g_j,s_y) and (g_i,s_x) to the set of edges while removing original edges.
4. Repeat 1-3.

Steps 1-3 constitute a single rewire of the network. To obtain a BiG-BET score for each individual gene, we rewire the network many times in sequence, keeping track of the edge changes so that the network becomes unrecognizable from the original network and the previous sample. This process is a Markov chain, generating a random network at each step conditioned only on its immediate predecessor that is independent of earlier networks. If run long enough, the process will generate all possible networks from the model with uniform probability. Prior to drawing samples from the Markov chain, we conduct $2m$''burn-in'' rewires, where m is the number of edges in the network, to give the process freedom to sample high likelihood regions under the model independent of the initialization at our observed data. We perform at least m rewires between samples to ensure that most edges in the network will have the opportunity to rewire.

We define the BiG-BET score as follows; Let the gene g_i have degree k_i. Let $\partial g_i^{obs} = \{s | (s, g_i) \in \varepsilon^{obs}\}$ denote the set of samples connected to in the bipartite representation of the original data, G_{obs} (that is the neighbors of g_i). Let be the $T_{obs} = \mathbb{E}_{s \in \partial g_i^{obs}}(TMB_s)$, average TMB for all samples connected to g_i in the observed dataset. We derive a Z score for a significant association between TMB and g_i as follows:

1. We sample R independent realizations of the bipartite network with fixed degrees sequences, according to the process described above.
2. For each sampled network, G_r we compute $T_r = \mathbb{E}_{s \in \partial g_i^r}(TMB_s)$, the average TMB for all neighbors of g_i in each sampled bipartite network G_r.
3. We compute the Z score for the observed graph using:

$$z_i = \frac{T_{obs} - \mathbb{E}_r T_r}{\widehat{\sigma}_r(T_r)}$$

where $\widehat{\sigma}_r(T_r)$ is the empirical standard deviation for the distribution of sampled $\{T_r\}$. All results in the paper have been derived using R = 400 samples from the bipartite configuration model. All BiG-BET scores for genes in the TCGA cohort are listed in Table S1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Survival and response analyses

All survival analyses were conducted using the Cox-proportional hazard model with a log likelihood ratio test (LLR) to test for the overall significance of the model and a t-test to assess the significance of individual variables in the model. For the depiction of Kaplan-Meier curves, TMB is treated as a binary variable with a threshold of TMB>10 defining the high TMB group. However, in the joint models depicted by the forest plots, TMB is treated as a continuous variable. Comparison of response across groups is conducted using a Chi-squared test. Samples were divided into groups based on the presence of a mutation within a High DDR Z score gene or Low Z score DDR gene, shown in Figure 2B.

The Low BIG-BET Z score DDR genes included:

APEX1, APEX2, ATM, ATR, ATRIP, BLM, BRCA1, BRCA2, BRIP1, CHEK2, ERCC1, ERCC2, EXO1, FANCB, FANCD2, FANCL, FANCM, MUS81, NHEJ1, POLB, PRKDC, RAD51, RAD52, RBBP8, TDP1, TP53BP1, TREX1, UBE2T, XPA, XRCC3, XRCC5.

The High BIG-BET Z score DDR genes were:

BARD1, CHEK1, CUL5, EME1, ERCC4, ERCC5, ERCC6, FANCA, FANCC, FANCI, FEN1, GEN1, LIG4, MDC1, MLH1, MLH3, MRE11, MSH2, MSH3, MSH6, NBN, PALB2, PARP1, PMS1, PMS2, POLE, POLE3, POLL, POLM, RAD50, RNMT, SEM1, SLX1A, TDG, TOP3A, TOPBP1, UNG, XPC, XRCC2, XRCC4, XRCC6.

Tumors with mutations in both a High and a Low DDR gene were considered in the High Z score category and excluded from the Low Z score category.

Gene expression signatures analysis

To calculate gene signatures profiles for each dataset, RNAseq expression data was obtained and filtered to the corresponding samples with mutational data. We log (1 + x) transformed the data and used a robust scaling (median centered and scaled by inter-quartile range) to normalize across samples. For each signature, we calculate the average expression of all genes within the signature and then assign each sample a Z score of the basis of its expression relative to the entire cohort. Signatures used in analysis are given in Table S2.

Proof of the friendship paradox for bipartite network

We can represent a network of N nodes and m edges with an $N \times N$ adjacency matrix,A, where the entries of A are defined as follows

$$\begin{cases} 1 & \text{if } (i,j) \in \varepsilon \\ 0 & \text{otherwise} \end{cases},$$

where we use ε to denote the set of edges present in the graph, indexed by the pair of nodes connected by each edge. For bipartite networks, we denote N_1 to be the number of nodes of class 1 and likewise N_2 the number in class 2, with $N_1 + N_2 = N$. In a bipartite network, each edge (i,j) connects a node from class 1 with one from class 2. The degree k_i of node i is given by the number of edges connected to that node: $k_i = \sum_j A_{ij}$. We let the degree distribution p_k give the fraction of nodes with degree k, representing the probability that a randomly chosen node will have that degree. We denote the class specific degree distributions as p_k^1 and p_k^2 to represent the fraction of nodes within each class with a given degree. The overall degree distribution,p_k, and the class specific degree distributions are related by

$$p_k = \frac{p_k^1 N_1 + p_k^2 N_2}{N}$$

In our gene-sample network, we are interested in the average degree across all samples with a mutation in a given gene. We show that this value, the average neighbor-of-a-gene degree, is typically greater than or equal to the average degree of the sample nodes in the network, following a proof similar to that for unipartite networks in.[35]

We begin by computing the probability that, after following a randomly chosen edge in our bipartite network, we arrive at a node of a given class with degree k. Without loss of generality we assume class 1 is our class of interest (the sample nodes). There are m

edges connected to nodes of class 1, so the probability of ending at a particular node with degree k is k/m. Since there are such nodes with degree k, the probability of following an edge to a class 1 node of degree k is

$$\frac{k}{m}N_1 p_k^1 = \frac{k}{\langle k\rangle_1}p_k^1,$$

where $m/N_1 = \langle k\rangle_1$ gives the average degree for nodes of class 1. That is, the average neighbor degree distribution is weighted by a factor of. We are more likely to choose a higher degree vertex by virtue of the simple fact that it has more edges connected to it. We can then compute the average neighbor degree by

$$\sum_k k\frac{k}{\langle k\rangle_1}p_k^1 = \sum_k \frac{k^2}{\langle k\rangle_1}p_k^1 = \frac{\langle k^2\rangle_1}{\langle k\rangle_1}.$$

We can compute the difference between the average neighbor degree and the average degree, restricted to nodes in class 1:

$$\frac{\langle k^2\rangle_1}{\langle k\rangle_1} - \langle k\rangle_1 = \frac{1}{\langle k\rangle_1}\left(\langle k^2\rangle_1 - \langle k\rangle_1^2\right) = \frac{1}{\langle k\rangle_1}Var_1(k),$$

where $Var_1(k)$ is the variance of the degree distribution restricted to nodes of class 1. This is strictly non-negative and is zero only in the case where all nodes of class 1 have the same degree. That is, except for the case where all nodes in the class have the same degree, the average neighbor degree is greater than the average degree. Furthermore, we see that this difference is proportional to the variance of the degree distribution of class 1, meaning that heavier-tailed class-restricted degree distributions give even bigger differences, and thus might be even more likely to be misanalysed by a univariate statistic that inadvertently mixes the roles of these two averages.