Beyond synthetic lethality: multiple gene interaction types play a key functional role in cancer

Assaf Magen, Avinash Das, Joo Sang Lee, Mahfuza Sharmin, Alexander Lugo, Silvio Gutkind, Eytan Ruppin*, Sridhar Hannenhalli*

Center for Bioinformatics and Computational Biology

University of Maryland, College Park, Maryland, USA

*Corresponding authors

Sridhar Hannenhalli & Eytan Ruppin

3104G Biomolecular Sciences Building

University of Maryland, College Park, MD

sridhar@umiacs.umd.edu

The phenotypic effect of perturbing a gene's activity depends on the activity state of other genes, reflecting the fundamental notion that genotype to phenotype linkage is mediated by a network of functionally interacting genes. The vast majority of contemporary investigations have focused on just one type of genetic interactions (GI) – synthetic lethality (SL). However, there may be additional types of GIs whose systematic identification may markedly enrich the molecular and functional characterization of cancer. Here, based on a novel data-driven approach, we identify ~72K GIs of 11 new types, shared across cancers. These GIs are highly predictive of patient survival, stratify breast cancer tumors into refined subtypes, and explain differences in patients' response to drugs and cancer driver genes' tissue-specificity. These results markedly expand the scope of cancer GIs and lay a strong conceptual and computational basis for future studies of additional types of GIs and for their translational applications.

Introduction

Many cellular functions are mediated by functionally interacting networks of genes. Genetic interactions (GIs), whereby the phenotypic effects of a gene's activity are modified by the activity of another gene, are thus a key to understanding complex diseases, including cancer, which involves an interplay among a myriad of genes[1–6]. GIs are of particular interest in cancer because, on one hand the dependence of one gene's phenotypic effect on the activity state of another gene provides opportunities for selective killing of cancer cells with minimal side effects[7], and on the other hand, it can render a drug targeting a particular gene ineffective because of unfavorable state of another gene, and it can result in emergence of resistance[8,9].

Three types of GIs have been studied so far, showing significant roles in cancer progression and patient survival, and suggesting novel therapeutic avenues. First and foremost are Synthetic Lethal (SL) interactions, describing the relationship between two genes whose individual inactivation results in a viable phenotype while their combined inactivation is lethal to the cell[1,9–11]. The vast majority of GI studies up until now have focused on the SL, which have long been considered as important for developing selective anticancer drugs[9]. Such drugs inhibit the SL partner of a gene that is inactivated by genomic alterations in the particular cancer[1,6,12], thus primarily killing the cancer cells with few cytotoxic effects on healthy cells. Another related class of GIs are Synthetic Dosage Lethal (SDL) interactions, where the

49 underactivity of one gene together with the over-activity of another gene is lethal but not each
50 event by itself[10,13,14]. SDLs are promising for targeting tumors with over-activated oncogenes,
51 many of which are difficult to target directly, by targeting their SDL partners[15–17]. A third class of
52 GIs are Synthetic Rescues (SR), where a change in the activity of one gene is lethal to the cell
53 but an alteration in its SR partner 'rescues' cell viability. SRs may play a key role in tumor
54 relapse and emergence of resistance to therapy[18–20]. Indeed, previous investigations have
55 shown that the overall numbers of functionally active SLs and SDLs in a given tumor sample is
56 highly predictive of patient survival[13]. SLs, and even SDLs and SRs whose investigation in
57 cancer has just began, represent just the 'tip of the GI iceberg', though, as there are many
58 additional types of GI that can be conceived at a conceptual level, but have remained
59 unchartered, and whose systematic exploration may have important functional ramifications for
60 cancer therapy.

61 Here we present a novel data-driven computational pipeline, called ''EnGIne"
62 (Encyclopedia of clinically significant GIs in cancer) for the identification of multiple types of GIs.
63 The application of EnGine to analyze 5288 TCGA samples (Methods) belonging to 18 different
64 cancer types reveals 71,946 significant GIs of 12 distinct types, having either positive or
65 negative associations with patient survival. A GI between two genes is considered functionally
66 active in a given tumor if the activity states of the two genes correspond to the definition of that
67 GI. We show that a tumor-specific 'fitness' score based on the active GIs in the tumor is highly
68 predictive of patient survival, comparing favorably with the gene-level approach, and that it can
69 be utilized for predicting a tumor's drug response. Furthermore, the GI activity profile of a tumor
70 can be used for stratifying breast cancer into functionally and clinically relevant subtypes. In
71 sum, these results provide the first comprehensive characterization of clinically relevant pan-
72 cancer GIs of 12 distinct basic types, substantively expanding the current knowledge of GIs in
73 cancer, and laying a strong conceptual and computational foundation for future studies of
74 additional GI types.

75

76 **Results**

77 **1. Overview of the Encyclopedia of Genomic Interactions (EnGIne) Pipeline**

78 The overall EnGine pipeline is summarized in Fig. 1 and the technical details are provided in the
79 Methods section. Here we provide an overview of the approach. Given a large set of tumor

transcriptomes (Fig. 1A), we first partition the expression level of each gene into low, medium, and high activity state. Thus, for a pair of genes, there are 9 combinations, or bins, of possible activity states for the two genes (Fig. 1B). For a given ordered pair of genes, each tumor sample maps to exactly one of the 9 bins. Our goal is to identify GI pairs of the form $(x, y, b, \pm\alpha)$ such that for the specific gene pair (x, y), the tumors in which the joint activity of (x,y) maps to bin b have a significant fitness advantage (+) or disadvantage (-) with effect size α, relative to all other tumors where the activity of (x,y) maps to a different bin. The effect size α is estimated by measuring the difference in the survival curves between those patients where the activity of (x, y) is in bin b in their tumors and those where it is not, as depicted in Fig. 1C. We term a specific GI pair $(x, y, b, \pm\alpha)$ as *functionally active* in a given tumor if the activity states of (x,y) fall in bin b in that tumor. Notably, we expect that the patients having a large number of functionally active interactions with negative effects on the tumor and a small number of such positive effect interactions are likely to have an overall higher survival and vice versa.

Our goal is to detect the genome-wide GIs in each of the 9 bins (Fig. 1B) searching over a space of ~163M gene pairs corresponding to ~18,000 genes. We analyzed 5288 TCGA samples for 18 cancer types (see Methods for initial filtering of TCGA samples). First, as an initial screening, we performed a Log-Rank test (depicted in Fig. 1C) for each gene pair in each of the 9 bins and retained the 223,946 genes pairs that exhibit a significant association with survival in any of the 9 bins (Methods). Second, if a potential GI in bin *b* has a differential effect on tumor fitness, we expect the number of tumors that map to bin *b* to be relatively enriched (for '+' interaction positively affecting tumor survival), or depleted (for '-' interaction negatively affecting tumor survival). We thus applied an additional filter (Fig. 1D) to retain only GIs with a molecular depletion/enrichment signal (Methods), yielding 179,444 gene pairs overall. Third, for each retained gene pair, in each of the 9 bins, we implemented a Cox proportional hazards model, specifically controlling for age, cancer-type, gender, and race, to assess whether being in a particular bin is associated with patient survival, either positively or negatively (Fig. 1E); while the second step aims to identify interactions that are selected for or against at the tumor level, the third step aims to identify interactions that also potentially affect survival – only interactions that have a consistent effect (positive on tumor growth and negative on survival, or *vice versa*) are retained. Further refining the initial filtering based on the log rank test, the Cox proportional hazards model enables us to specifically control for cancer type, race, gender, age and individual gene effects in the interactions identification process (Methods). Finally, to

identify the most significant GIs in each bin independently, we applied an empirical False Discovery Rate (FDR) correction based on the significance of the Cox interaction term of the 179,444 gene pairs (identified in the previous step) relative to those obtained for randomly shuffled gene pairing as the NULL control. At FDR <= 0.01%, this resulted in 71,946 predicted GIs across the 9 bins, of the form $(x, y, b, \pm\alpha)$, which form the final set of TCGA inferred GIs (Fig. 1E). Considering the symmetry among bins (bin 2 ~ bin 4 corresponding to low-medium expression interaction; bin 3 ~ bin 7 corresponding to low-high expression interaction; bin 6 ~ bin 8 corresponding to medium-high expression interaction), there are 6 unique types of interaction bins, and considering the two directions of the effect size yields a total of 12 basic types of GIs.

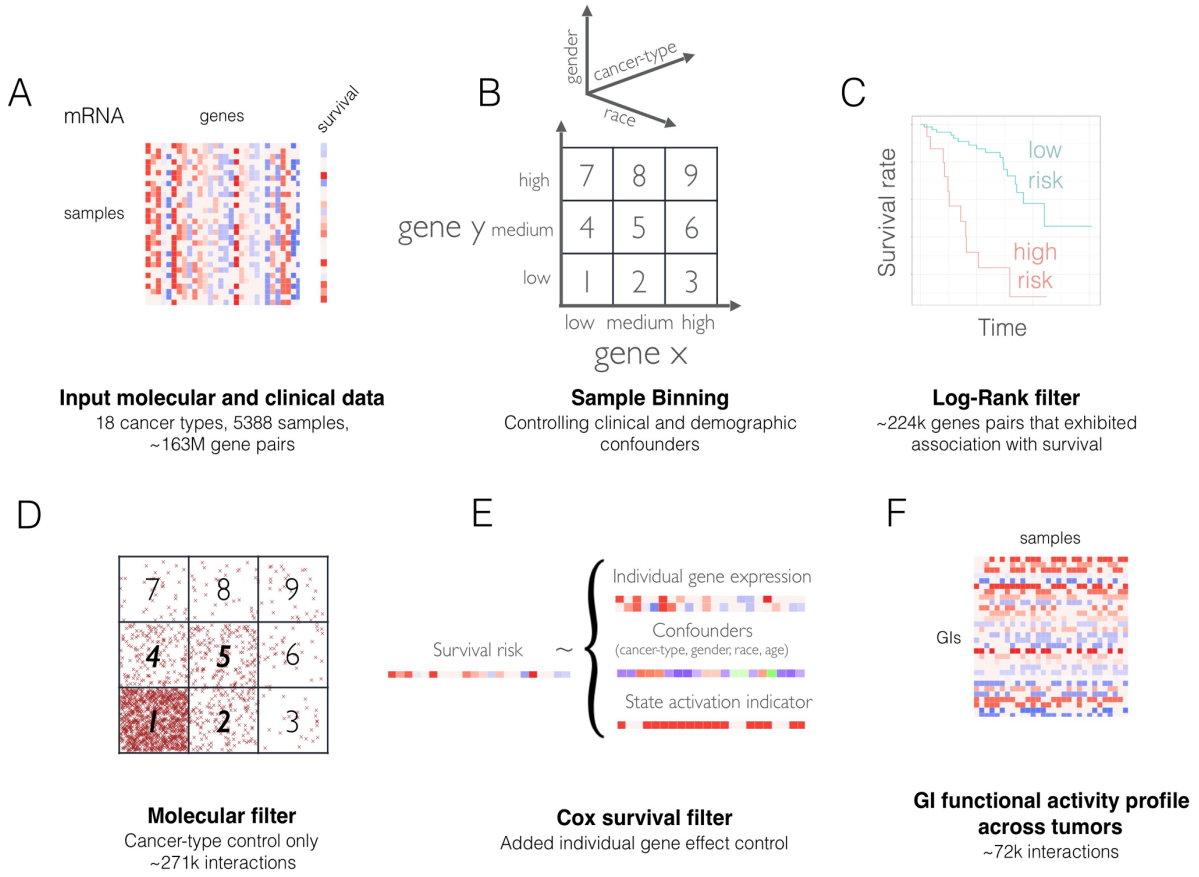


Figure 1. Overview of the EnGIne pipeline. Given a large set of tumor transcriptomes (A), we first partition the expression level of each gene into low, medium, and high activity state, resulting in 9 joint activity state bins for any two genes (B). Each gene pair and bin *b* induces a bipartition of the set of tumor samples based on whether the activity levels of the gene pair examined is in bin *b* in the specific tumor. The first step of EnGIne is a preparatory one, aiming to screen for the gene pairs that show distinct survival trends in the two sets of tumors in any of

129 the bins, based on log-rank test (C). Next, for a gene pair and a bin identified in the previous
130 step, we test whether the putative gene interaction in bin *b* has a differential effect on tumor
131 fitness, by testing for depletion/enrichment of samples in the bin *b* relative to expectation based
132 on individual genes (D). Finally, for each retained gene pair, in each of the 9 bins separately, we
133 fit a Cox proportional hazards model to assess whether being in a particular bin is associated
134 with a distinct (positive or negative) pattern of patient survival, followed by correction for multiple
135 hypotheses testing (E). The output of this computational pipeline is (a) a list of GIs of each of
136 the 12 types studied, and (b) GI profile in each of the individual tumor samples, defined as
137 activity state of each GI in the tumor sample (F).

138

139

140 **EnGIne TCGA analysis reveals survival-associated GIs involving known cancer drivers**

141 Applying EnGIne to 5288 tumor samples spanning 18 types of cancers results in the selection of
142 71,946 GIs of 12 different GI types (~0.02% of all the possible candidate gene pairs and GI
143 types tested). Considering the expectation that neighboring genes in the protein interaction
144 network (PIN) are more likely to be involved in a GI[21], we aimed to additionally construct a PIN-
145 supported GI network. Knowing that the a great many of the interactions may not correspond to
146 a direct PPI interaction[22], we filter the GIs that are separated by more than two edges in the
147 PIN, resulting in a PIN-supported GI network composed of 1704 gene pairs. Among the 1786
148 genes involved in these 1704 PIN GIs (Supplementary Table S1) we identified 133 known
149 cancer driver genes (Cosmic dataset[23]) that are associated with various cancer types
150 (enrichment P-value < 2.5E-22) and 50 breast cancer specific (Intogen dataset[24]) driver genes
151 (enrichment P-value < 7.7E-13).

152 Fig. 2A shows the distribution of the detected 1704 GIs across the 12 GI types. Notably,
153 SL interactions, whose functional activity has a negative effect on tumor fitness and are the
154 most experimentally studied type of GI, are surprisingly one of the least abundant types of GI,
155 and so are SDLs (1% of all GIs). Remarkably, the positive “anti-symmetric” type of SLs, which is
156 composed of interactions where the joint low activity of the two interacting genes is associated
157 with a higher tumor fitness, are 3 times more abundant than SLs. The Interaction between the
158 cancer genes GNAQ and JAK2 is one example of such a positive interaction in bin-1
159 (Supplementary Fig. S1). Our analysis reveals several other types of GIs to be much more
160 abundant than the SL and SDL-type GIs. In fact, in the 2 most abundant types of pan-cancer
161 GIs (Bin-2 and Bin-6), one of the genes has medium level of activity and only the extreme
162 activity of its partner gene reveals a phenotypic effect. For most GI bins we see a higher

163 proportion of GIs exerting a positive effect on tumor fitness, which corresponds to a negative
164 association with patient survival. Importantly, it is consistent with the hypothesis that the GIs
165 uncovered during the evolution of cancer are under positive selection.

166 We tested and validated EnGIne by capitalizing on a collection of 7297 interaction pairs
167 that were tested experimentally in SL *in vitro* screens, collected from the literature[25–30]. Each of
168 the three filtering steps of EnGIne (Fig. 1C-E), could discriminate the experimentally determined
169 positive SLs from the negative ones, with ROC-AUCs of 0.63 (P-value < 0.0005), 0.62 (P-value
170 < 0.001), and 0.59 (P-value < 0.012), respectively. The SL screening was done *in vitro*, a
171 platform known to exhibit major discrepancy with in-vivo data. Given that our goal is to identify
172 clinically-relevant effects and thus the EnGine predicted GIs are inferred directly from patient
173 data, this significant recapitulation of the SL signal in pan-caner heterogeneous tumor data,
174 while of modest accuracy, is encouraging and supports the contribution of each of the individual
175 steps in EnGIne. As shown in Supplementary Fig. S2 and detailed in Supplementary Results,
176 we found that known cancer drivers are enriched among GIs of various types, and interestingly,
177 relative to tumor suppressors, oncogenes show a greater enrichment and are especially
178 enriched among GIs in bin-3, having a positive effect on tumor fitness as expected.

179 We investigated the biological processes involved in each of the 12 broad types of GIs
180 shown in Fig. 2A. We found an enrichment (FDR <= 0.01) of several processes with well-
181 established links to cancer among the genes involved in these GIs (Methods). These include
182 cell cycle control, cell division, cell adhesion, extracellular matrix organization, angiogenesis,
183 hypoxia, the Wnt signaling pathway, and more (Supplementary File 2). An illustrative example is
184 a genetic interaction involving a breast tumor suppressor ERCC2, a transcription-coupled DNA
185 excision repair gene[31,32], and a breast cancer oncogene KLF4, a zinc finger transcription
186 factor[33]. Fig. 2B show survival patterns for different activity state combinations of these two
187 genes. It reveals two interesting trends: First, the over-activation of the oncogene and under-
188 activation of the tumor suppressor (bin 3) results in poorer survival than expected from the
189 individual gene effects (bins, 1, 2, 6, and 9). Second, and surprisingly, the survival curve reveals
190 a reversal of the effect of the tumor suppressor ERCC2 inactivity on survival when the
191 oncogene KLF4 has medium activity (bin 2), whose individual activity is actually associated with
192 better survival. This and several other examples of GIs involving a cancer driver gene
193 (Supplementary File 1) demonstrate that the context-specific effects of driver genes may show

194 very different trends than their previously established individual gene effects. Supplementary
195 Fig. S3 shows the extended GI-network involving the known driver genes.

196 Fig. 2C shows the predicted GIs involving the well-established oncogene Myc and tumor
197 suppressor BRCA1 obtained by target-specific FDR (Methods), highlighting their number and
198 variety. On one hand, this richness of GIs involving these driver genes might account for the
199 lack of efficacy in targeting cancer drivers in some cases, and on the other hand, our analysis
200 suggests combination therapies jointly targeting GI partners of these driver genes. Consistent
201 with its Myc's role as an oncogene, the GIs occurring when Myc has low activity mostly have
202 negative effect on tumor fitness. However, when Myc is activated, we find that the low
203 expression of PUF60, a repressor of Myc[34,35], is associated with higher tumor fitness (type 3+
204 interaction). Fig. 2D demonstrates that high expression of Myc does not significantly contribute
205 to poorer prognosis (P-value ~ 0.9), unless its suppressor PUF60 is lowly expressed (HR =
206 1.37, P-value < 5.0E-04), underscoring the importance of molecular context in developing anti-
207 Myc treatments. Although PUF60 is known to suppress Myc, we provide evidence for a
208 functional relationship that deviates from a simple epistatic interaction.

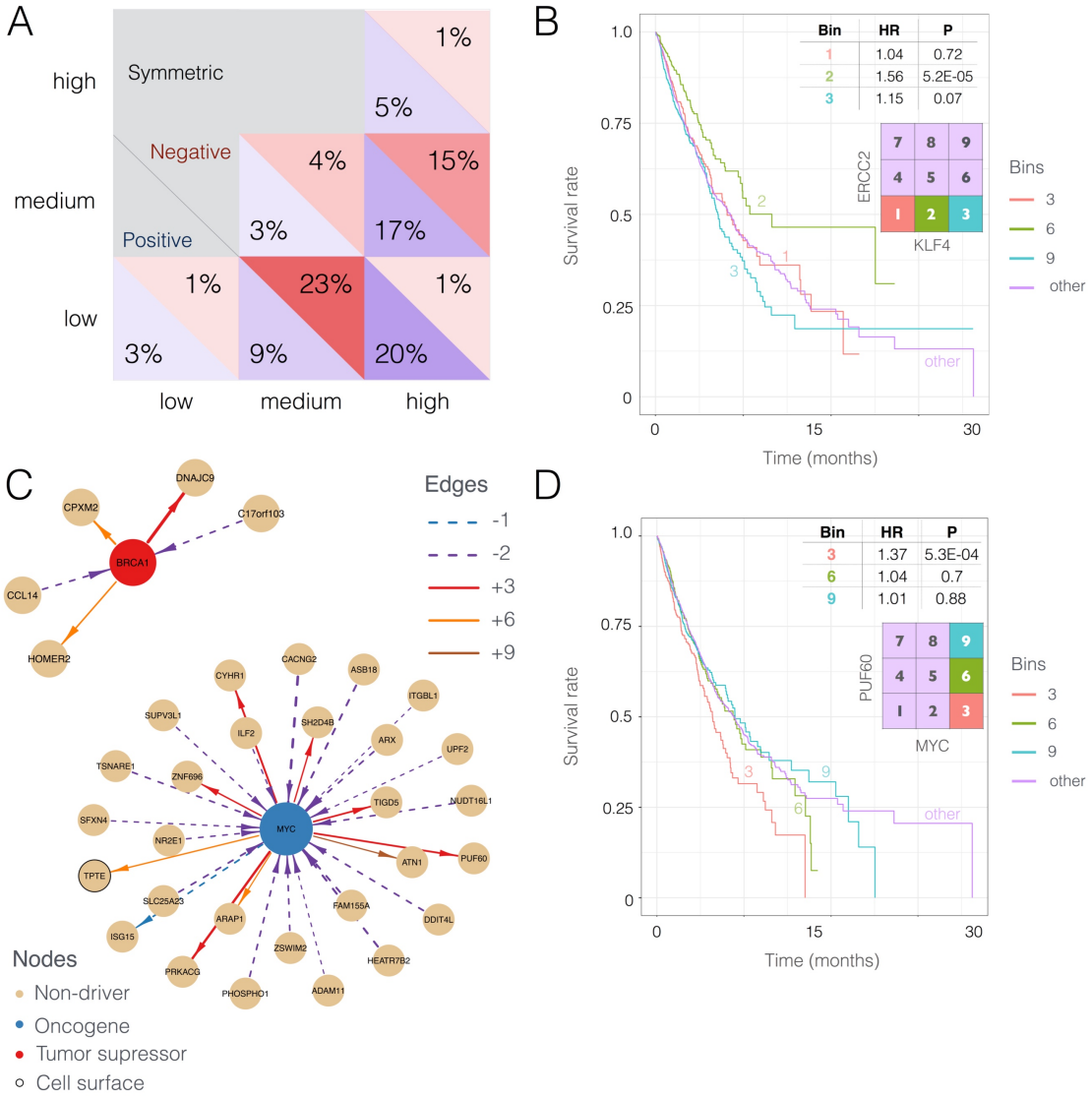


Figure 2 – Broad distribution and characteristics of the detected GIs and context-specific effect of cancer driver genes on survival. (A) Distribution of the 1704 significant PIN-supported GIs across 9 joint activity bins. The fractions of GIs in each bin are shown for GIs with positive (blue) and negative (red) effect on tumor fitness. Only the data in the lower half of the matrix are shown as the GIs are symmetric relative to the genes in a pair. (B) The Kaplan-Meier (KM) survival curve of GI involving ERCC2, a transcription-coupled DNA excision repair gene, known to be a breast cancer tumor suppressor, and KLF4, a zinc finger transcription factor known to be oncogenic in breast cancer, reveals increasingly poor survival by over-activation of the oncogene and under-activation of the tumor suppressor (bin 3). Strikingly, the effect of ERCC2 inactivity on survival is reversed when KLF4 has medium activity level (bin 2). (C) The predicted GIs involving the oncogene Myc and the tumor-suppressor BRCA1. (D) KM survival curve of GI involving Myc and its suppressor PUF60. High expression of Myc is associated with poor prognosis specifically at low activation of PUF60.

223 Finally, we quantified the ability of tumor-specific GIs to predict patient survival relative to
224 a comparable gene-wise approach, both in cross-validation setting in TCGA[15], as well as in an
225 independent breast cancer METABRIC dataset[36] (Fig. S4A, supplementary note 3 for additional
226 datasets discussion). As detailed in supplementary results, overall, the GI-based approach
227 achieves higher concordance than the gene-wise approach. An even bigger improvement is
228 observed in the independent dataset, testifying that the GI-based approach is generalizable,
229 while the individual-gene approach fails (concordance ≈ 0.51). Supplementary Fig. S4B depicts
230 the survival prediction accuracy of each GI type. As evident, interactions involving both genes in
231 their wild type mid-activity levels (i.e. bin 5) have negligible predictive power on survival,
232 testifying that more extreme levels of gene expression tend to be involved in functional GIs
233 affecting survival.

234

235 **Differential GI activation between drug-response groups**
236
237 To assess the extent to which tumor-specific GIs are associated with the tumor's response to a
238 drug we studied whether the detected GIs exhibit differential activity between patients who are
239 responsive to a drug and those who are not. For a given drug and cancer type having data on
240 responders and non-responders, we first identify drug-specific GIs based only on the samples
241 that do not have drug response information to avoid circularity (Methods); note that this network
242 is different from the generic GI networks presented in the previous sections, as focusing on
243 specific drug target genes requires a smaller multiple hypotheses burden, resulting in a greater
244 number of drug target-specific GIs than in the TCGA global pan-cancer network above. Also,
245 because the drugs considered are inhibitory, it suffices to focus on GI bins 1, 2, and 3, where
246 one of the genes (the drug's target) is presumed inactivated. For each identified GI, we test,
247 using Fisher exact test[37] whether the GI exhibits differential activation between the responders
248 and the non-responders (Methods). Note that, for positive GIs, we expect a lower activation
249 among responders and for negative GIs we expect a higher activation among responders.
250 However, owing to very small and unbalanced numbers of responders and non-responders
251 (5~35 samples per response group per drug), the Fisher test is underpowered and despite high
252 effect sizes, the p-values for many individual tests are not significant. Given that, we compared
253 the Fisher test p-values of the GIs with those obtained using randomly shuffled drug-response

254 labels, using paired Wilcoxon test[38] (Methods), performed separately for each drug-cancer type
255 pair and segregated by GI type.
256
257 We considered 12 drug-cancer type pairs that have tumor size reduction information following
258 treatment for at least 10 patients (at least 5 responders and 5 non-responders) in TCGA. Fig. 3A
259 shows the significance of differential activation of the drug-specific GIs, segregated by the 6 GI
260 types. Overall, in 18 of the 72 tests of drug-cancer type combinations, GIs of a particular type
261 exhibit differential activation between responders and non-responders, consistent with the
262 predicted effects of the GIs (5 fold enrichment for $P \leq 0.05$). Reassuringly, several drugs come
263 up as significant using SLs (GI type -1), consistent with previous reports showing the role of SLs
264 in mediating drug response[39]. Among the drugs, Gemcitabine, Lomustine and Paclitaxel exhibit
265 differential GI activation for most GI types (aggregate P-values = 4E-16 ~ 2E-11).
266
267 To further explore the most differentially activated GIs, we imposed an empirical FDR threshold
268 of 0.01 on the Fisher test P-value yielding a total of 521 GIs for the 12 drug-cancer type
269 combinations (Methods, Supplementary Table S2). Individual genes comprising the 521 GIs are
270 closer to each other's interacting gene in the PPI network relatively to shuffled pairs (Wilcoxon P
271 < 0.001, Methods) and have a significantly increased number of direct PPI interactions between
272 them (Fisher $P < 0.02$, Methods).
273
274 As an illustrative test case, we explored the GIs associated with the response to *Paclitaxel* in
275 Head and Neck Squamous Cell (HNSC) Carcinoma. We identified a GI involving the inactivation
276 of BCL-2 (known to suppress apoptosis, indirectly inhibited through phosphorylation[40]) and the
277 over-activation of ITPR1 (Inositol 1,4,5-trisphosphate receptor type 1, coding for IP3 Receptor)
278 negatively affecting tumor fitness. Interestingly, our analysis shows that this GI is activated
279 among the responders at a significantly higher ratio than among non-responders (odds-ratio ≈
280 11.1). Interaction between the IP3 receptor and BCL-2 is well characterized[41–43]. BCL-2 exerts
281 its oncogenic effect by inhibiting IP3-dependent channel opening and Ca^{2+} release from the ER
282 (by inhibiting mitochondrial Ca^{2+} overload), and thus preventing cancer cell apoptosis. Our
283 analysis strongly suggests that BCL-2 inhibition by *Paclitaxel* is especially effective when the
284 IP3 receptor is abundant and therefore increased Ca^{2+} release is possible. We find additional

interactions with literature evidence[44] among the *Paclitaxel* targets (Fig. 3B), however without complete experimental characterization, suggesting promising avenues for additional studies.

GIs explain cancer type-specific effects of cancer driver genes

Many of the known cancer driver genes affect tumor initiation and development in a tissue-specific manner. We assessed the extent to which the molecular context, characterized by the functionally active GIs, may explain the tissue-specific effects of cancer drivers. Toward this, we identified 15 oncogenes and 20 tumor suppressors whose effects are likely to be restricted to specific cancer types, based on preferentially high mutation rates in those cancer types, including breast, bladder, and gastric cancer (Methods, Supplementary Table S3). We then assessed the risk-score associated with each of these driver gene's GI partners in each of the TCGA samples (Methods). We hypothesized that the specific activity states of the GI partners of tumor-suppressors and oncogenes may explain their functional significance (as estimated by their effects on patient survival) in the specific tissue(s) where the driver's effect is known to manifest itself. Indeed, for 15 out of 35 (~43%) driver genes, we see a higher GI risk score (measuring the functional activity of the predicted GI partners) in the specific cancer samples belonging to the type they are known to be associated with in the literature, compared to their GI risk scores in the other cancer types (Wilcoxon rank-sum test, FDR < 0.1, Supplementary Table S3). For instance, HLF, a bZIP transcription factor, has been linked to lung and breast cancer based on its significantly greater missense mutation frequency in those cancer types[24]. We observed a significant difference (FDR < 1.07E-12) in GI activation between breast and lung cancer relative to the other tissues (Fig. 3C,D), where positive interactions are preferentially activated in these two tissues while negative interactions are preferentially activated in the background tissues, consistent with the increased tumor fitness in the foreground tissues. Among the GI partners, JAMIP1 is a member of the Integrated Breast Cancer Pathway and CLPTM1 is associated with Estrogen-Receptor Negative Breast Cancer. Overall, these results imply that cancer type-specific effects of many driver genes may be explained by their tissue-specific GI network activity (note that mutation data used here was not used in the inference of these GIs, so there is no circularity).

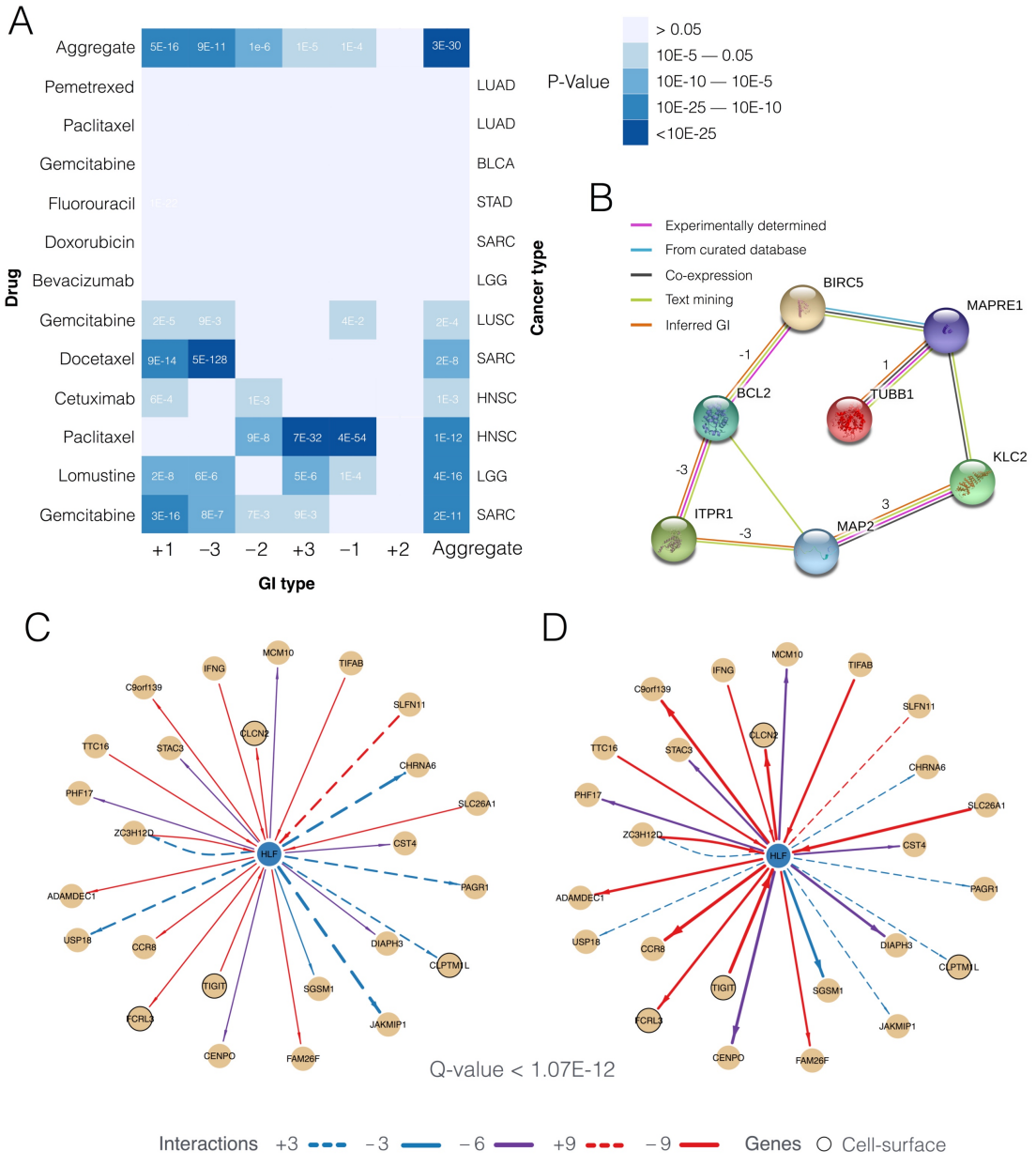


Figure 3 – (A-B) differential GI activation between drug response groups. (A) For each drug (left row labels) and each cancer type (right row label) combination, and for each GI type (columns), the heat plot shows the significance of differential activation of GIs in responders and non-responders consistent with expectation. The last column shows the significance when all GI types are aggregated. (B) The network shows the inferred functional interactions (based on STRING[44]) among the genes interacting with *Paclitaxel* targets, as well as inferred GI types. **(C-D) Differential GI activation between tissues in gene-specific GI-network.** For HLF-specific GI network the figure shows the activity states of GIs in breast and lung cancers (C - foreground tissues) and in other cancer types (D - background tissues). The edge weight (thickness) represents the fraction of samples in which the GI was functionally active. Several GIs can be

seen as differentially active in the two sets of cancer types. The figure also depicts cell surface proteins among the HLF's GI partners. The GI network-based sample-specific risk score is significantly higher (q-value < 1.07E-12) in breast and lung cancer relative to other cancer types, potentially mediated by a selective activation of positive GIs in the foreground tissues and negative GIs in the background tissues.

Stratifying breast cancer tumors based on their GI profiles

We set out to study whether functionally active GIs in a tumor can provide an interaction-based alternative to gene expression or sequence data conventionally used for tumor stratification into various sub-types. Given the cancer type specificity of many of the GIs studied above, we asked whether GI-based characterization of breast cancer is effective in providing a functionally relevant stratification of tumors to their subtypes. We used the raw GI profile of a tumor (which consists of 1704 binary values representing the functional activity of all PIN-supported GIs) to cluster the 1981 breast cancer samples in an independent METABRIC dataset [36] (recall that the GIs were identified in TCGA data) using a conventional Non-Negative Matrix Factorization (NMF) (Methods). Since NMF requires a predetermined number of clusters as input, we applied the method for a range of 2-15 clusters and assess the clustering quality using Dunn's index, repeating the clustering procedure used in the original work[36]. The Dunn's index achieved the maximum value for 10 clusters (Supplementary Fig. S7A) and upon closer inspection of the distributions of known breast cancer subtypes in these clusters (Supplementary Fig. S7B) we merged two of the clusters, thus yielding 9 clusters for further analyses.

As shown in Fig. 4A, the 9 clusters have distinct survival characteristics with an overall mean hazard ratio (HR) difference of 1.94 (P-value ~ 0). Supplementary Fig. S8A shows the survival characteristics obtained for the previously published clustering of the METABRIC samples[36]. As evident, both approaches obtain similar survival separation levels, but exhibit differences in their histopathological composition.

Currently Breast cancer has 5 well-established clinically distinct subtypes based on the tumors' histopathological attributes. Fig. 4B shows the fraction of these known subtypes among the 9 GI-based clusters. Several clusters are highly associated with specific subtypes such as Basal [triple-negative] (cluster 5), Luminal A (cluster 3,4) etc. Others show association with several subtypes, e.g., Luminal A and B are clustered together in cluster 8. Interestingly, the basal subtype, which are largely triple-negative and have poor prognosis, correspond to a distinct cluster in our analysis (cluster 5), consistent with their distinct clinical status. In the

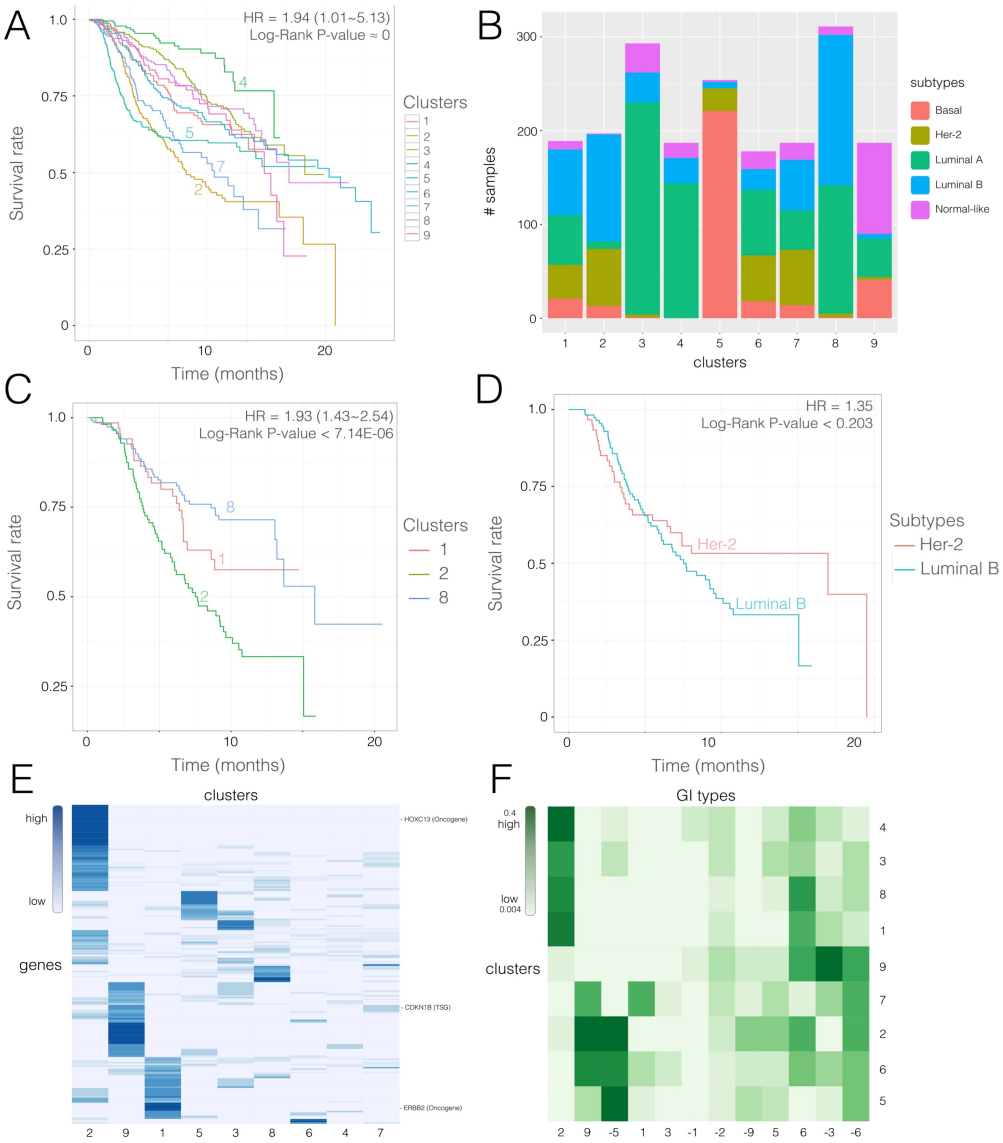
361 original METABRIC publication 50% of the samples were left unassigned to any of their 10
362 clusters in METABRIC while our GI-based clustering covers all samples. In order to provide an
363 equivalent comparison, we selected a matching number of samples (50%) such that we retain
364 the samples that are the most confidently associated with a cluster. Supplementary figure
365 S8B,C compare the GI-based clusters and the clusters in the original METABRIC publication in
366 terms of their histopathological subtype composition.

367 Fig. 4C,D illustrate the added survival predictive value of GI based characterization over
368 the classical histopathological ones. Our evaluation proceeds along two different perspectives:
369 (1) cases where known histopathological breast cancer subtypes are split across multiple GI-
370 based clusters (e.g. Luminal B across clusters 1, 2 and 8), and conversely (2), cases where one
371 GI-based cluster harbors multiple known histopathological subtypes (e.g. cluster 2 contains Her-
372 2 and Luminal B subtypes). In the former we find, e.g., that the 1989 Luminal B tumors that are
373 split across different GI-based clusters exhibit very distinct survival trends (Fig. 4C). Likewise, in
374 the latter case, we find that the survival trends of Luminal A and Luminal B histopathological
375 subtype samples that are assigned to the same GI-based cluster 8 do not show a significant
376 difference in their survival trends (Fig. 4D). We systematically identified 6 additional cases of the
377 above two kinds where (1) a known tumor histopathological subtype was split across multiple
378 GI-based clusters or (2) multiple known histopathological subtypes were assigned to the same
379 GI-based cluster (and each cluster has at least 30 samples). In each case of the first kind we
380 tested for differences in survival and in each of the second case we tested for lack thereof. As
381 shown in Supplementary Fig. S9, in 5 out of 6 cases we found that the GI-based clusters
382 provided a more accurate survival difference than the known histopathological subtypes. In
383 contrast, we identified 4 cases of GI-based clusters of the second kind and in none of them the
384 survival trends of the samples with different histopathology markers were significantly different
385 (Supplementary Fig. S10).

386 Additionally, we assessed whether the samples in GI-based clusters harbor distinct gene
387 mutations. Based on the TCGA mutation data, we identified 196 genes (Supplementary Table
388 S4) with significantly greater mutation frequency in one or more of the clusters, relative to their
389 overall mutation frequency in breast cancer (Methods). Fig. 4E shows the mutational frequency
390 profiles of these genes across the 9 clusters. Overall, the differentially mutated genes across
391 the GI based clusters include 10 cancer drivers: CDK12, CDKN1B, DNAJB1, ERBB2, EXT2,
392 FCGR2B, FNBP1, HOXC13, SEC24D, and PDGFRB. Cyclin dependent kinase Cdk12, involved

393 in DNA damage response, is known to affect BRAC1 transcription [45]. Likewise, cyclin CDKN1B
394 is known to be mutated in luminal breast cancer [46]. Our analysis suggests a more specific
395 enrichment of this mutation in subset of Luminal-A tumors (cluster 9 in Fig. 4B). The HER2
396 breast cancer subtype is conventionally characterized by overexpression of HER2 gene (also
397 known as ERBB2), but rare activating mutations of this gene have also been implicated in the
398 disease, independent of overexpression [47]. Our analysis reveals an enrichment of HER2
399 mutations in cluster 1, which includes only a small subset of annotated HER2 subtype tumors.
400 Most of the HER2 mutations (11 out of 13) are predicted to have a significant impact on the
401 protein structure (SIFT and PolyPhen measures, Methods). This may suggest a distinct
402 oncogenic mechanism underlying this subclass of tumors, which may express a HER2-like
403 subtype including a few samples where HER2 is over-expressed and many others where it is
404 not over-expressed, but mutated. As another example, the FCGR2B mutation modulates HER2
405 breast tumor's response to Trastuzumab [48]. Our analysis reveals a specific enrichment of
406 FCGR2B mutations in cluster 6, which includes several other breast cancer subtypes besides
407 HER2, pointing to the potential relevance of these mutations in tumors belonging to other
408 cancer subtypes as well (we emphasize that no mutation data was used at all in the GIs
409 inference via EnGIne).

410 We quantified the fraction of each of the 12 types of functionally active GIs among the
411 samples in each cluster (see Methods). Fig. 4F shows the active GI profiles of each cluster, and
412 reveals two broad subgroups, one including clusters 4, 3, 8, and 1 and another including
413 clusters 2, 5, 6, and 7. Interestingly, the two subgroups clearly segregate in terms of their
414 survival, testifying that the GI profiles capture a simplified yet robust characterization of the
415 clinical prognosis. Supplementary Fig. S11 compares the GI-profiles of clusters revealed in the
416 TCGA and the METABRIC breast cancer data and shows a high degree of consistency. A
417 global comparison of the GI profiles of the 9 clusters in the two datasets shows a Spearman
418 correlation of 0.67 (P-value < 2.4E-14) between the GI types composition of these clusters,
419 implying that GI-profiles are a robust characteristic of breast cancer tumors across different
420 tumor collections.



421

422 **Figure 4 – (A) Mean survival curves of the individuals in the 9 inferred GI-based breast**
423 **cancer subtypes**. **(B) Cluster subtype composition based on PAM50 breast cancer sub**
424 **typing**[49]. **(C-D) Survival trends of tumors of known histopathological cancer subtypes**
425 **within and across GI-based clusters.** (C) Luminal B samples that are split across GI-based
426 clusters 1, 2 and 8 show significant survival differences. (D) Her2 and Luminal-B type tumors
427 that are included within the GI-based cluster 2 exhibit similar survival trends. **(E) Mutational**
428 **profile of GI-based breast cancer subtypes**. Mutation profiles of 196 genes (rows) across the
429 9 GI-based clusters (columns). For each gene and each cluster, the figure depicts the fraction of
430 samples in the cluster in which the gene is mutated. **(F) GI types composition of the GI-based**
431 **breast cancer subtypes in the METABRIC dataset**. In clustering the samples based on GI
432 profile, each GI is probabilistically assigned to a single cluster, based on which, the composition
433 of GIs assigned to each cluster is obtained. The x-axes represent the 12 GI types (6 activity bins
434 and 2 directional effects on survival), and the y-axes represent the clusters. The colors
435 represent the fraction of cluster-assigned GIs of each GI type.

436

437 **Discussion**

438 Analyzing molecular and clinical data across thousands of tumors of dozens of types, here, for
439 the first time, we comprehensively map the landscape of 12 basic GI types in cancer. Our work
440 substantially extends previous investigations of gene interactions in cancer, which have been
441 almost entirely focused on synthetic lethality (SL: corresponding to positive effect on survival in
442 bin 1)[26,27,50], with a few studies of synthetic dosage lethality (SDL: corresponding to positive
443 effect on survival in bin 1)[11,12], to a total of 12 types of interactions. The identified GIs are
444 predictive of patient survival and drug response based on the tumor transcriptomes, and reveal
445 novel functionally and clinically relevant breast cancer subtypes. The set of functionally active
446 GIs thus provides a complimentary molecular characterization of tumors to those obtained by
447 histology and contemporary individual gene-centric transcriptomic and sequence based profiles.
448 A better concordance of the GI-based breast cancer subtypes with survival trends may be partly
449 due to the fact that GIs were inferred based on their impact on survival. However, interestingly,
450 the detected subtypes are marked by distinct mutational profiles, which were not utilized in
451 inferring the GIs. Overall, these results underscore the importance of molecular context
452 represented by functionally active GIs.

453 In this first conceptual foundation and application of EnGIne, we have analyzed the TCGA
454 compendium to identify GIs interactions shared across many cancer types. Identifying pairwise
455 genes interaction is only a first step toward capturing the true complexity of cellular networks.
456 Future work can go beyond the 9 basis GI types studied here to investigate more complex GI
457 types that involve different compositions of these basic types; for instance, a given interacting
458 gene pair can confer tumor fitness benefit in multiple co-activity bins and reduce tumor fitness in
459 others. Thus, while the results presented here go markedly beyond previous definitions and
460 analyses of GIs, they only begin to explore the full scope and clinical potential of GI-based
461 analyses of cancer, awaiting future investigations.

462

463

464

465

466 **Methods and Materials**

467 **Cancer datasets**

468 The Cancer Genome Atlas (TCGA)[15] consists of 8749 tumor samples across 28 cancer types, 3
469 races and 2 genders (data quality testing is described in supplementary note 1). We do gene
470 expression binning specifically for each combination of cancer type, race and gender.
471 Furthermore, we control for various clinical and demographic group specific effects. Because
472 using small groups of samples may result in insufficiently robust models, we filtered rare
473 combinations of clinical and demographic groups, resulting in 5288 mRNA samples derived
474 from patients spread across 18 cancer types, 3 races and 2 genders. The RSEM normalized
475 expression data was downloaded from the TCGA data portal and quantile normalization was
476 applied within each sample. METABRIC breast cancer dataset[36] consists of 1989 microarray
477 samples and was used for independent validation. Similarly, quantile normalization was applied
478 in each sample.

479 **Protein interaction Network (PIN)**

480 The PIN was obtained from a previously published resource called the HIPPIE which
481 aggregates physical protein interaction data from 10 source databases and 11 studies[21]. This
482 network consists of 15673 proteins and 203159 edges.

483 **Identification of gene interactions (GI) associated with cancer patient survival**

484 As shown in Fig. 1, we divide the range of each gene's expression across tumor samples into 3
485 equal-sized bins that correspond to the 3 activity states: low, medium and high expression.
486 Given a gene pair, each tumor sample is thus mapped into one of the 9 joint activity states of
487 the two genes. The choice of dividing a gene's activity into three classes, while somewhat
488 arbitrary, was made in consideration of interpretability of functional states and robustness of
489 inference. However, to account for differences in expression distributions across clinical and
490 demographic confounders, we apply subpopulation specific binning approach. We considered
491 the following categorical confounders: cancer type, race, and gender. We considered the
492 combination of confounder states for which there were at least 100 tumor samples
493 (Supplementary Table S5). Our binning is thus not confounded by various clinical and
494 demographic variables.

495 The GI pipeline consists of three steps that successively refine the predictions to arrive at high-
496 confidence set of predicted GIs. As such, the number of all pair-wise combinations of genes is
497 excessively large to apply a comprehensive Cox regression model.

498 **Step 1: Log-Rank.** In the first step, for each of the ~163 million gene pairs, say (X,Y), we
499 compute the Log-rank statistics [51] estimating the survival difference between the samples that
500 map to one of the 9 activity bins and the other 8 bins. We implemented the log-rank test in C++
501 for computational speed. To control for gene-wise effect, we compare the Log-rank statistics of
502 the gene pair (X,Y) (in a bin) with those for (X,U) and separately with those for (V,Y), where U
503 and V represent all other genes. We consider Log-rank(X,Y) to be significant if it among the top
504 0.1% relative to all (X,U) and top 0.1% among all (V,Y) gene pairs. We retain a gene pair if it is
505 deemed significant in any of the 9 bins. This procedure retained 210,612 gene pairs of the total
506 of ~163M.

507 **Step 2: Molecular enrichment and depletion.** For a gene pair, we expect that its positive
508 (respectively, negative) effect on survival should reflect in negative (respectively, positive)
509 selection on the cancer cell's viability. We therefore expect the number of samples mapped to
510 the corresponding activity bin relative to the gene pair to be depleted (respectively, enriched).
511 We only retained the gene pairs which the number of samples in a particular bin were lower
512 (bottom 45 percent) or respectively, greater (top 45 percent among all gene pairs) than
513 expectation, reducing the number of GIs to 271,096 across 179,444 gene pairs (recall that a
514 gene pair can participate in multiple GIs corresponding to multiple bins and effect on survival).
515 **Step 3: Cox proportional hazard test.** Cox proportional hazards model is the most widely
516 acceptable approach for modeling survival while accounting for censored data as well as
517 confounding factors. We modeled each of the retained gene pairs' effect on survival in each of
518 the 9 bins using the interaction status λ (active if the sample mapped to the bin and inactive
519 otherwise), along with the confounders. Specifically, we introduced the two individual genes σ_1,
520 σ_2 expression to model each gene's independent effect on survival, and additionally, clinical and
521 demographic confounders, namely, cancer type, race, gender, and age. The model is stratified
522 based on the discrete confounders, to account for differences in the baseline hazard (risk)
523 characteristics. We did not control for tumor stage and grade as these classifications reflect the
524 very same tumor characteristics our model aims to capture and such control would prevent us
525 from learning an important element of the disease. Control for genomic-stability and tumor purity
526 as a potential confounder is described in supplementary note 2.

527 $$risk \sim \sigma_1 + \sigma_2 + \lambda + age + strata(type, gender, race)$$

528 Cox modeling provided a P-value representing the significance of the effect of joint gene pair
529 activity on survival. To obtain a NULL distribution for the P-values, we repeated this process for
530 corresponding list of randomly shuffled pairs. We retained ~71K gene pairs in the top 1% of the
531 NULL P-values, corresponding to a nominal FDR of ~3.8%.

532 **Optional Step 4: Filtering by protein interactions.** To gain additional confidence in the
533 predicted GIs, given the greater tendency (and expectation) for neighbors in the Protein
534 Interaction Networks (PIN) to exhibit functional interactions (see Results section), we further
535 refined the GI set by retaining the pairs that are found within distance of 2 in the PIN. Overall we
536 obtain a set of 1704 GIs that exhibit molecular and clinical evidence in cancer as well as
537 evidence from the PIN network.

538

539 **GO enrichment of hub genes**

540 We conducted Gene Enrichment analysis on genes involved in GIs using GOrilla tool[52] taking all
541 17990 genes as background set. For enrichment analysis, we considered the genes into
542 multiple categories and the obvious categories can be the genes separated by 12 broad GI
543 types. However, the GI bins (Fig 1B) suggests that unlike in bin 1, 5, and 9, in bin 2, 3, and 9,
544 one of the 'x' gene and 'y' gene has higher or lower expression/activity than the other.
545 Therefore, considering bin number, patient survival, and gene activity states, we have total 18
546 (taking bin symmetricity into account) kinds of gene sets. For each gene sets, we conducted
547 Gene-Ontology enrichment analysis using Gorilla tool[52]. For the enrichment analysis, we have
548 used all 17990 genes as background set.

549

550 **Patient Survival Risk Prediction**

551 We applied a semi-supervised approach to assign a risk score to each patient according to the
552 functionally active interactions in the sample. Consider a GI involving genes X and Y conferring
553 a positive effect on the tumor fitness in a particular joint-activity bin B (Fig. 1). If in a sample
554 genes X and Y fall in bin B then then the GI is said to be 'functionally active' in the sample, and
555 a score of +1 is contributed to the overall tumor 'fitness'. Likewise if the GI has negative effect
556 on the tumor fitness then a score of -1 is contributed. For each sample, we compute the overall

557 score conferred by functionally active GIs (either in a bin-specific and effect-specific fashion, or
558 overall) in the sample. The higher the tumor 'fitness' score the lower the survival potential.
559 However, to make our approach comparable to gene-wise approaches[53], we assigned each
560 gene the sum of the contributions by all active GIs involving that gene. The estimated gene-wise
561 GI score is used as a predictor variable in a Cox model along with the confounding factors
562 discussed above to predict patient's survival.

563 This model is trained on the same data used for the GIs training and then validated on its
564 cross-validation counterpart, or in the case of independent validation, the model is trained on
565 the full TCGA dataset, and tested on an independent METABRIC breast cancer data with 1989
566 samples[36]. The prediction accuracy is estimated in terms of widely used C-index[54]. Several
567 previous publications have assessed survival risk prediction accuracy based on dichotomized
568 analysis where samples are separated into distinct low- and high-survival groups and their
569 survival curves the compared[51], which is prone to overestimating prediction accuracy. For
570 comparison, we have also performed accuracy estimation following the dichotomized
571 comparison of survival risks between the extreme cases of predicted survival risk groups, for
572 variable threshold to define the extreme (such as top versus bottom 10% or top versus bottom
573 20% and so on).

574 To compare the GI-based survival prediction with individual gene approach, we implemented an
575 analogous scheme for individual genes where the gene expression values were discretized into
576 3 expression levels (low, medium and high), and the discretized representation was used as a
577 predictor variable in a controlled Cox regression model to obtain the significance (P-value) of
578 each gene with respect to survival prediction. The most significant predictors (top 5%) were
579 chosen and precisely used as described for the GIs survival prediction procedure. An analogous
580 procedure was used to estimate the prediction accuracy based on both individual gene effects
581 and GIs.

582

583 **Characterization of differential GI activation between drug-response groups**

584 We inferred the GIs involving each drug's target gene based on the TCGA samples that do not
585 have that drug's response information to avoid circularity, and filtered based on FDR as for
586 global pan-cancer GI detection as above, but restricted to the target-specific GIs. For each of
587 the drug-specific GIs, we compare its activation frequency (whether the GI was functionally

588 activate or inactivate) among the responders (stable disease, partial response and complete
589 response categories) and non-responders group (clinical progressive disease categories), using
590 one-sided Fisher's exact test[37], where the alternative hypothesis is that negative (respectively,
591 positive) GIs are more frequently active among responders (respectively, non-responders).
592 However, given the extremely small and imbalanced sample sizes, and the conservative nature
593 of Fisher's exact test[55], we test whether the overall distribution of the obtained Fisher test p-
594 values are lower than those obtained using randomly shuffled drug-response labels, using one-
595 sided Wilcoxon test[38]. We thus obtain a p-value for each drug-cancer type pair, segregated by
596 GI type.

597

598 Then, for each gene pair in the inferred GIs list and, as a control, in a shuffled list of size 10x as
599 the original GIs list size, we compute the distance between the genes in the PPI network[21]. We
600 then use one-sided Wilcoxon test[38] to assess whether GI gene pairs are closer to each other
601 than random expectation. Alternatively, we also compare the number of directly GI gene pairs
602 having direct interaction using one-sided Fisher's exact test[37].
603

604 **Characterization of tissue-specific effect of cancer driver genes**

605 A recent study[56] of genes' somatic mutation profile across cancer types has identified and
606 characterized the tissue specificity of 459 candidate drivers. For each such candidate, we
607 matched the driver role annotation (oncogene or tumor-suppressor) obtained from the Census[23]
608 cancer genes dataset (genes with missing annotation were discarded) for 33 tumor suppressors
609 and 25 oncogenes. For each gene, we calculated the significant GI interactions involving this
610 gene, where the significance threshold is defined over the shuffled background interactions
611 involving this gene (90th percentile of the interaction confidence score). We further excluded
612 genes with 5 or fewer interactions with 300 or fewer samples where they are expressed,
613 resulting in 20 tumor suppressors and 15 oncogenes spanning 10 cancer types (Supplementary
614 Table S3). For each gene, the cancer types are partitioned into affected types (cancer types
615 affected by the driver) and the other unaffected cancer types. We hypothesized that higher-risk
616 will be associated with lower expression of a tumor-suppressor and higher expression of an
617 oncogene, much more in the affected cancer types relative to the unaffected cancer types.
618 Thus, we assess the associated risk-score of the samples found in the low (respectively, high)

619 expression bin for a tumor-suppressor (respectively, oncogene). Using a one-sided Wilcoxon
620 rank-sum test, we tested for higher risk-score in the samples in the affected cancer type in
621 comparison to those in unaffected types. After multiple hypotheses testing, 15 out of the 35
622 (~43%) driver genes were found to have significant tissue-specific GI-based risk score (FDR q-
623 value < 0.1, Supplementary Table S3).
624

625 **Breast cancer tumor stratification**

626 We represent a tumor sample as a vector indicating the functional activity status of each
627 predicted GI. This provides a survival-cognizant alternative to the widely-used gene expression
628 profile representation of a sample. We used this representation to stratify the METABRIC breast
629 cancer patients using Non-Negative Matrix Factorization (NMF)[57,58], which has suitable
630 statistical properties and has been shown to be effective in a variety of contexts[59]. Since NMF
631 requires a predetermined number of clusters, we performed the analysis for 2-15 clusters, and
632 assessed their fitness using Dunn's index[60], which quantifies compactness within and
633 separation across clusters. Similar analysis was performed on the TCGA breast cancer cohort.

634 The Hazard-ratio significance values were computed for each pair of clusters, while the P-
635 values were generated using multi-class log-rank test. For comparison purposes, to match our
636 estimated clusters' sizes to previously published METABRIC cluster sizes (~900 samples), we
637 constrained the number of samples in each cluster to the 100 samples that were found to be
638 most highly associated with the cluster.

639 Each cluster's GI profiles were constructed as follows. Our clustering approach – NMF,
640 assigned each GI to a single cluster. For each cluster, and for each of the 12 GI-types (6 bins in
641 Fig. 1 and the two directional effects on survival), we obtain the frequency of GIs of that type,
642 relative to all GI assigned to the clusters.

643 Mutation frequency analysis was performed on the TCGA clusters, where we first tested
644 whether the mutation frequency distribution of each gene differs significantly across clusters
645 using Chi-square test. The selected genes (Chi-Square test to ascertain the dependence of
646 mutation abundance on the interaction type) were then used to illustrate the mutational profiles
647 of the clusters.

648 The prediction values of a mutation's effect on the protein were obtained from the cBioPortal
649 repository. Out of the 196 differentially mutated genes found, 138 genes had matching extended
650 mutation information indicating their SIFT (sorts intolerant from tolerant amino acid
651 substitutions) and PolyPhene (polymorphism phenotyping) predictions. We calculated a gene-
652 wise ratio of mutations predicted to have a significant effect on the protein. 68 (49%) and 63
653 (45%) of which had at least 50% of the mutations belonging to the tested category in SIFT and
654 PolyPhene, respectively.

655 The breast cancer subtypes were derived using the widely accepted PAM50 algorithm[49]. The
656 METABRIC PAM50 subtypes were annotated in the original publication[36], while the TCGA
657 breast cancer subtypes were calculated using the original published class centroids[49].

658

References

1. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145,** 30–38 (2011).
2. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23,** 561–566 (2005).
3. Wong, S. L. *et al.* Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 15682–15687 (2004).
4. Zhong, W. & Sternberg, P. W. Genome-wide prediction of C. elegans genetic interactions. *Science* **311,** 1481–4 (2006).
5. Lu, X., Kensche, P. R., Huynen, M. a & Notebaart, R. a. Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. *Nat. Commun.* **4,** 2124 (2013).
6. Jerby-Arnon, L. *et al.* Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* **158,** 1199–1209 (2014).
7. Kaelin, W. G. The concept of synthetic lethality in the context of anticancer therapy. *Nat. Rev. Cancer* **5,** 689–98 (2005).
8. Fong, C. Y. *et al.* BET inhibitor resistance emerges from leukaemia stem cells. *Nature* **525,** 538–+ (2015).
9. Miyamoto, D. T. *et al.* RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **349,** 1351–6 (2015).
10. Stuhlmiller, T. J. *et al.* Inhibition of lapatinib-induced kinome reprogramming in ERBB2-positive breast cancer by targeting BET family bromodomains. *Cell Rep.* **11,** 390–404 (2015).
11. Sajesh, B. V., Guppy, B. J. & McManus, K. J. Synthetic genetic targeting of genome instability in cancer. *Cancers* **5,** 739–761 (2013).
12. Kroll, E. S., Hyland, K. M., Hieter, P. & Li, J. J. Establishing genetic interactions by a synthetic dosage lethality phenotype. *Genetics* **143,** 95–102 (1996).
13. Megchelenbrink, W., Katzir, R., Lu, X., Ruppin, E. & Notebaart, R. a. Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 12217–12222 (2015).
14. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* **43,** 656–62 (2011).
15. Chang, K. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
16. Rathert, P. *et al.* Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. *Nature* **525,** 543–547 (2015).
17. Luo, J., Solimini, N. L. & Elledge, S. J. Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. *Cell* **136,** 823–837 (2009).
18. Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W. & Friend, S. H. Integrating genetic approaches into the discovery of anticancer drugs. *Science (80-.).* **278,** 1064–1068 (1997).
19. Brough, R., Frankum, J. R., Costa-Cabral, S., Lord, C. J. & Ashworth, A. Searching for synthetic lethality in cancer. *Current Opinion in Genetics and Development* **21,** 34–41 (2011).
20. McLornan, D. P., List, A. & Mufti, G. J. Applying Synthetic Lethality for the Selective Targeting of Cancer. *N. Engl. J. Med.* **371,** 1725–1735 (2014).
21. Schaefer, M. H. *et al.* Hippie: Integrating protein interaction networks with experiment

707 based quality scores. *PLoS One* **7,** (2012).
708 22. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular
709 function. *Science (80-.).* **353,** aaf1420-aaf1420 (2016).
710 23. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–83
711 (2004).
712 24. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types.
713 *Nat. Methods* **10,** 1081–1082 (2013).
714 25. Lord, C. J., McDonald, S., Swift, S., Turner, N. C. & Ashworth, A. A high-throughput RNA
715 interference screen for DNA repair determinants of PARP inhibitor sensitivity. *DNA*
716 *Repair (Amst).* **7,** 2010–2019 (2010).
717 26. Martin, S. A. *et al.* Methotrexate induces oxidative DNA damage and is selectively lethal
718 to tumour cells with defects in the DNA mismatch repair gene MSH2. *EMBO Mol. Med.* **1,**
719 323–337 (2009).
720 27. Turner, N. C. *et al.* A synthetic lethal siRNA screen identifying genes mediating sensitivity
721 to a PARP inhibitor. *EMBO J.* **27,** 1368–77 (2008).
722 28. Luo, J. *et al.* A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal
723 Interactions with the Ras Oncogene. *Cell* **137,** 835–848 (2009).
724 29. Steckel, M. *et al.* Determination of synthetic lethal interactions in KRAS oncogene-
725 dependent cancer cells reveals novel therapeutic targeting strategies. *Cell Res.* **22,**
726 1227–45 (2012).
727 30. Bommi-Reddy, A. *et al.* Kinase requirements in human cells: III. Altered kinase
728 requirements in VHL-/- cancer cells detected in a pilot synthetic lethal screen. *Proc. Natl.*
729 *Acad. Sci. U. S. A.* **105,** 16484–16489 (2008).
730 31. Benhamou, S. & Sarasin, A. ERCC2/XPD gene polymorphisms and cancer risk.
731 *Mutagenesis* **17,** 463–469 (2002).
732 32. Bernard-Gallon, D. *et al.* DNA repair gene ERCC2 polymorphisms and associations with
733 breast and ovarian cancer risk. *Mol. Cancer* **7,** 36 (2008).
734 33. Akaogi, K. *et al.* KLF4 suppresses estrogen-dependent breast cancer growth by inhibiting
735 the transcriptional activity of ERalpha. *Oncogene* **28,** 2894–902 (2009).
736 34. Rahmutulla, B. *et al.* Alternative splicing of FBP-interacting repressor coordinates c-Myc,
737 P27Kip1/cyclinE and Ku86/XRCC5 expression as a molecular sensor for bleomycin-
738 induced DNA damage pathway. *Oncotarget* **5,** 2404–2417 (2013).
739 35. Matsushita, K. *et al.* Non-transmissible Sendai virus vector encoding c-myc suppressor
740 FBP-interacting repressor for cancer therapy. *World J. Gastroenterol.* **20,** 4316–4328
741 (2014).
742 36. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours
743 reveals novel subgroups. *Nature* **486,** 346–52 (2012).
744 37. Fisher, R. A. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of
745 P. *J. R. Stat. Soc.* **85,** 87 (1922).
746 38. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bull.* **1,** 80 (1945).
747 39. Jerby, L., Waldman, Y., Weinstock, A., Geiger, T. & Ruppin, E. Genome-wide detection of
748 synthetically-lethal genes uncovers a novel repository of selective cancer targets. 1–9
749 40. Ruvolo, P. P., Deng, X. & May, W. S. Phosphorylation of Bcl2 and regulation of
750 apoptosis. *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K* **15,** 515–522 (2001).
751 41. Oakes, S. A. *et al.* Proapoptotic BAX and BAK regulate the type 1 inositol trisphosphate
752 receptor and calcium leak from the endoplasmic reticulum. *Proc. Natl. Acad. Sci.* **102,**
753 105–110 (2005).
754 42. Chen, R. *et al.* Bcl-2 functionally interacts with inositol 1,4,5-trisphosphate receptors to

755 regulate calcium release from the ER in response to inositol 1,4,5-trisphosphate. *J. Cell*
756 *Biol.* **166,** 193–203 (2004).
757 43. Rong, Y.-P. *et al.* The BH4 domain of Bcl-2 inhibits ER calcium release and apoptosis by
758 binding the regulatory and coupling domain of the IP3 receptor. *Proc. Natl. Acad. Sci.*
759 **106,** 14397–14402 (2009).
760 44. Szklarczyk, D. *et al.* STRING v10: Protein-protein interaction networks, integrated over
761 the tree of life. *Nucleic Acids Res.* **43,** D447–D452 (2015).
762 45. Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation
763 of expression of DNA damage response genes. *Genes Dev.* **25,** 2158–72 (2011).
764 46. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast
765 cancer. *Nature* **486,** 400–4 (2012).
766 47. Sun, Z. *et al.* Analysis of different HER-2 mutations in breast cancer progression and
767 drug resistance. *J. Cell. Mol. Med.* **19,** 2691–2701 (2015).
768 48. Norton, N. *et al.* Association Studies of Fcγ Receptor Polymorphisms with Outcome in
769 HER2+ Breast Cancer Patients Treated with Trastuzumab in NCCTG (Alliance) Trial
770 N9831. *Cancer Immunol. Res.* **2,** 962–9 (2014).
771 49. Bernard, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic
772 subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
773 50. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic
774 strategy. *Nature* **434,** 917–21 (2005).
775 51. Harrington, D. P. & Fleming, T. R. A class of rank test procedures for censored survival
776 data. *Biometrika* **69,** 553–566 (1982).
777 52. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery
778 and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10,** 48
779 (2009).
780 53. Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across
781 tumor types. *Nat. Biotechnol.* **32,** 644–652 (2014).
782 54. Harrell, F. E., Lee, K. L. & Mark, D. B. in *Tutorials in Biostatistics, Statistical Methods in*
783 *Clinical Studies* **1,** 223–249 (2005).
784 55. Berkson, J. In dispraise of the exact test. Do the marginal totals of the 2X2 table contain
785 relevant information respecting the table proportions? *J. Stat. Plan. Inference* **2,** 27–42
786 (1978).
787 56. Rubio-Perez, C. *et al.* In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor
788 Types Reveals Targeting Opportunities. *Cancer Cell* **27,** 382–396 (2015).
789 57. Paatero, P. & Tapper, U. Positive Matrix Factorization - A Nonnegative Factor Model With
790 Optimal Utilization of Error-Estimates of Data Values. *Environmetrics* **5,** 111–126 (1994).
791 58. Lee, D. D. & Seung, H. S. Algorithms for Non-negative Matrix Factorization. in *NIPS* 556–
792 562 (2000). doi:10.1109/ICASSP.2006.1661352
793 59. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix
794 factorization. *Nature* **401,** 788–91 (1999).
795 60. Dunn, J. C. & Dunnf, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J.*
796 *Cybern.* **4,** 95–104 (1974).
797