HIT'nDRIVE: Patient-Specific Multi-Driver Gene Prioritization for Precision Oncology

Raunak Shrestha[1,2,*], Ermin Hodzic[3,*], Thomas Sauerwald[4], Phuong Dao[5], Kendric Wang[2], Jake Yeung[2], Shawn Anderson[2], Fabio Vandin[6], Gholamreza Haffari[7], Colin C. Collins[2,8], and S. Cenk Sahinalp[2,3,9,†]

[1]Bioinformatics Training Program, University of British Columbia, Vancouver, BC, Canada.

[2]Laboratory for Advanced Genome Analysis, Vancouver Prostate Centre, Vancouver, BC, Canada.

[3]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.

[4]Computer Laboratory, University of Cambridge, Cambridge, UK.

[5]National Center for Biotechnology Information, NLM, NIH, Bethesda, MD, USA.

[6]Department of Information Engineering, University of Padova, Padova, Italy.

[7]Faculty of Information Technology, Monash University, Melbourne, Australia.

[8]Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada.

[9]School of Informatics and Computing, Indiana University, Bloomington, IN, USA.

[*]Co-first authors

[†]Correspondence: cenk@sfu.ca

Supplemental Material

Contents

1 HIT'nDRIVE Framework . . . 5
2 Calculating Hitting Time on an Interaction Network . . . 6
3 Estimating Multi-Source Hitting Time via Single-Source Hitting Times . . . 7
4 Datasets and analysis . . . 11
5 Derivation of expression outlier genes . . . 14
6 Derivation of expression outlier gene weights . . . 15
7 HIT'nDRIVE sensitivity to least frequent driver gene . . . 15
8 Association of driver modules with patients' survival outcome . . . 16
9 Statistical significance of the overlap of driver genes with that of CGC database. . . . 17
10 HIT'nDRIVE: parameters . . . 18
11 HIT'nDRIVE: outlier stringency . . . 18
12 HIT'nDRIVE: network perturbation . . . 18
13 HIT'nDRIVE: underlying network . . . 18
14 HIT'nDRIVE: alteration types . . . 19
15 HIT'nDRIVE: random alterations and random outliers. . . . 19
16 Modified HIT'nDRIVE: when it is not required to prioritize one driver gene per patient. . . . 19
17 HIT'nDRIVE's ability to capture CGC genes . . . 20
18 Patient druggability . . . 20
19 Unsupervised classification of gene expression data . . . 21
20 Phenotype classification using CGC gene seeded modules . . . 22
21 Mutual exclusivity of driver modules . . . 23
22 HIT'nDRIVE using regulatory networks . . . 24
23 Correlation of predicted driver genes with alteration burden. . . . 25
24 Cancer-stage specific driver genes of colorectal cancer. . . . 26

List of Figures

Supplemental Fig. S1 HIT'nDRIVE identified driver genes with respect to varying parameter values in 100 select BRCA samples. 27

Supplemental Fig. S2 HIT'nDRIVE identified driver genes with respect to varying outlier stringency in 100 select BRCA samples. 28

Supplemental Fig. S3 HIT'nDRIVE identified driver genes with respect to network perturbation in 100 select BRCA samples. 29

Supplemental Fig. S4 HIT'nDRIVE identified driver genes with respect to underlying network used in 100 select BRCA samples. 30

Supplemental Fig. S5 HIT'nDRIVE identified driver genes with respect to different alteration types in 100 select BRCA samples. 31

Supplemental Fig. S6 HIT'nDRIVE identified driver genes using randomized input data in 100 select BRCA samples. 32

Supplemental Fig. S7 Modified HIT'nDRIVE not required to prioritize at least one driver gene per patient. 33

Supplemental Fig. S8 Genomic drivers of Glioblastoma 34

Supplemental Fig. S9 Genomic drivers of Glioblastoma 35

Supplemental Fig. S10 Genomic drivers of Ovarian Cancer 36

Supplemental Fig. S11 Genomic drivers of Ovarian Cancer 37

Supplemental Fig. S12 Genomic drivers of Prostate Cancer 38

Supplemental Fig. S13 Genomic drivers of Prostate Cancer 39

Supplemental Fig. S14 Genomic drivers of Breast Cancer 40

Supplemental Fig. S15 Genomic drivers of Breast Cancer 41

Supplemental Fig. S16 Driver Genes Distribution 42

Supplemental Fig. S17 Likelihood of HIT'nDRIVE to capture CGC Genes. 43

Supplemental Fig. S18 P-value Distribution of the likelihood of HIT'nDRIVE to pick CGC genes. 44

Supplemental Fig. S19 Distribution of patient druggability 45

Supplemental Fig. S20 Anti-cancer drugs targeting driver genes predicted by HIT'nDRIVE. 46

Supplemental Fig. S21 HIT'nDRIVE sensitivity to least frequent driver gene. 47

Supplemental Fig. S22 Schematic Diagram of HIT'nDRIVE-unsupervised approach to prioritize driver-modules 48

Supplemental Fig. S23 Phenotype Classification using CGC Genes Seeded Modules. 49

Supplemental Fig. S24 Comparison of phenotype classification accuracy achieved by HIT'nDRIVE-OptDis with that achieved by a set of differentially expressed genes. 50
Supplemental Fig. S25 Comparison of HIT'nDRIVE+OptDis based modules against randomly selected modules. 51
Supplemental Fig. S26 Mutual Exclusivity of Driver Modules 52
Supplemental Fig. S27 Module Expression Heatmap: TCGA-BRCA Dataset 53
Supplemental Fig. S28 Module Expression Heatmap: METABRIC-CAMBRIDGE Dataset 54
Supplemental Fig. S29 Module Expression Heatmap: METABRIC-VANCOUVER Dataset 55
Supplemental Fig. S30 Unsupervised Clustering of BRCA subtypes in TCGA-BRCA cohort. 56
Supplemental Fig. S31 Activity Score of BRCA subtype-specific modules containing *ESR1*. 57
Supplemental Fig. S32 Activity Score of BRCA subtype-specific modules containing *ERBB2*. 58
Supplemental Fig. S33 Heatmap of *NCOA3* driver module expression across different BRCA subtypes. . . . 59
Supplemental Fig. S34 BRCA subtype specific driver module (HER2-01) 60
Supplemental Fig. S35 BRCA subtype specific driver module (HER2-11) 61
Supplemental Fig. S36 BRCA subtype specific driver module (HER2-13) 62
Supplemental Fig. S37 BRCA subtype specific driver module (LUMA-09) 63
Supplemental Fig. S38 BRCA subtype specific driver module (LUMA-11) 64
Supplemental Fig. S39 BRCA subtype specific driver module (LUMA-29) 65
Supplemental Fig. S40 BRCA subtype specific driver module (LUMA-39) 66
Supplemental Fig. S41 BRCA subtype specific driver module (LUMB-23) 67
Supplemental Fig. S42 Metabolism of Estrogen 68
Supplemental Fig. S43 Drivers Modules of Ovarian Cancer 69
Supplemental Fig. S44 Drivers Modules of Prostate Cancer 70
Supplemental Fig. S45 EGFR-PI3K Signaling Pathway 71
Supplemental Fig. S46 Overview of Drug Response Analysis using HIT'nDRIVE + OptDis 72
Supplemental Fig. S47 Correlation between the number of driver genes predicted by HIT'nDRIVE with mutation rate and copy-number burden 73
Supplemental Fig. S48 HIT'nDRIVE predicted driver genes of Colorectal cancer (TCGA-COAD). 74

Supplemental Methods

1 HIT'nDRIVE Framework

HIT'nDRIVE naturally integrates genome and transcriptome data from a number of tumor samples for identifying and prioritizing sequence-wise altered genes as potential drivers. It "links" sequence-wise altered genes to genes with expression changes through a gene or protein interaction network. For that, it aims to find the *smallest* set of sequence-wise altered genes that can "explain" most of the observed gene expression alterations in the cohort. In other words, HIT'nDRIVE identifies the minimum number of potential driver genes which can "cause" a user-defined proportion of the downstream expression effects observed.

HIT'nDRIVE uses a particular "influence" value of a potential driver gene on other (possibly distant) genes based on the (gene or protein) interaction network in use. In order to capture the uncertainty of interactions of genes with their neighbours, it considers a random walk process which propagates the effect of sequence alteration in one gene to the remainder of the genes through the network. As a result, the influence is defined to be the inverse of hitting-time, which is the expected length (number of hops) of a random walk which starts at a given potential driver gene, and "hits" a given target gene the first time in a (protein or gene) interaction network. More specifically, for any two nodes $u, v \in V$ of an undirected, connected graph $G = (V, E)$, let the random variable $\tau_{u,v}$ denote the number of hops in a random walk starting from u and visiting v for the first time. Then the hitting-time $H_{u,v}$ is defined as $H_{u,v} = E[\tau_{u,v}]$ (Levin et al. 2008).

In order to capture synthetic lethality like scenarios, HIT'nDRIVE considers multiple sequence-wise altered genes as potential drivers. For that, we define the influence value (of a set of potential driver genes on a target) as the inverse of multi(source)-hitting time, i.e., the expectation of the smallest number of hops in one of the random walk processes, simultaneously starting at each one of the potential driver genes and ending at a given expression-wise altered gene for the first time. More specifically, let $U \subseteq V$ be a subset of nodes of G and $v \in (V - \{U\})$ be a single node. We thus define the multi(source)-hitting time $H_{U,v}$ as $H_{U,v} = E[\min_{u \in U} \tau_{u,v}]$.

HIT'nDRIVE formulates the process of potential driver gene(s) discovery in terms of the "random-walk facility location" (RWFL) problem, which, for a single patient can be described as follows.

Let $\mathscr{X}$ be a set of potential driver genes and $\mathscr{Y}$ be a set of expression altered (outlier) genes. Then, for a user defined k, HIT'nDRIVE can aim to return k potential driver genes as solution to the following optimization problem:

$$\arg\min_{X \subseteq \mathscr{X}, |X| = k} \max_{y \in \mathscr{Y}} H_{X,y}$$

where $H_{X,y}$ denotes the multi-hitting time from the gene set X to the gene y.

RWFL problem resembles the standard (minimax) "facility location" problem in which one seeks a set of nodes as facilities in a graph such that the maximum distance from any node in the graph to its closest facility is minimized. RWFL differs from standard facility location by its use of $H_{X,y}$ as a distance measure between a collection of nodes to any other node, which aims to capture the uncertainty in molecular interactions during the propagation of one or more signals, by random walks starting from one or more origins (reminiscent of the underlying Brownian motion). Since the standard facility location is an NP-hard problem, RWFL problem is NP hard as well. As shown in the next section, we overcome this difficulty by introducing a good estimate on the multi-hitting time that helps us to reduce RWFL problem to the weighted multi-set cover problem (WMSC), which we solve through an Integer Linear Programming (ILP) formulation. Although the use of set-cover for representing the most parsimonious solution in a bioinformatics context is not new (Hormozdiari et al. 2009), to the best of our knowledge, this is the first use of the multi-set cover formulation for maximum parsimony. In this formulation, we use a slightly different objective: given a user defined upper bound on the maximum multi-hitting time, we now aim to minimize the number of potential drivers that can "cover" (a user defined proportion of) the outlier genes. For more than one patient, we minimize the number of drivers that can "cover" (a user defined proportion of) patient-specific outliers such that each such outlier is covered by potential drivers that are aberrant in that patient.

2 Calculating Hitting Time on an Interaction Network

As mentioned before, HIT'nDRIVE estimates the multi-hitting time $H(U,v)$ between a set of nodes U and a single node v, as a function of independent hitting times $H(u,v)$ for all $u \in U$ - as will be shown later. To calculate exact values of $H(u,v)$ for all pairs of nodes in the network, we use the matrix inversion method as explained by Tetali *et al.* (Tetali 1999). Here we will briefly describe the method. For proofs please refer to Tetali (1999).

Let P denote the transition probability matrix (of size $n \times n$) of the interaction network with n nodes, and H its hitting-time matrix. $P_{i,j}$ represents the probability that random walk picks node u_j as its next step from node u_i, and $H_{i,j}$ represents the hitting-time $H(u_i,u_j)$. We assume that $P_{i,i} = 0$, for all $i \in \{1,...,n\}$, which forces the random walk to move from current node to one of its neighbours in every step. Lastly, let π be the stationary distribution of the network, where π_i represents the proportion of time that an infinite-length random walk in the network spends visiting node u_i.

Given P, H, and π, we define $(n-1) \times (n-1)$ matrices $\bar{P}$ and $\bar{H}$ as follows: $\bar{P}_{i,i} = \pi_i$, $\bar{P}_{i,j} = -\pi_i P_{i,j}$ and $\bar{H}_{i,i} = H_{i,n} + H_{n,i}$, $\bar{H}_{i,j} = H_{i,n} + H_{n,j} - H_{i,j}$, for all $i,j \in \{1,...,n-1\}$ such that $(i \neq j)$. We show how to calculate

hitting-times based on the following claim:

Theorem 1. *Given P, H, $\bar{P}$ and $\bar{H}$ as defined above, $\bar{P}\bar{H} = I_{n-1}$.*

For proofs please refer to Tetali (1999).

Note that $\bar{P}$ can be computed directly from the transition probability matrix P (following its definition) and we obtain $\bar{H}$ by inverting $\bar{P}$. Using the definition of $\bar{H}$ and proof of Theorem 1 (see Theorem 2.2 in Tetali (1999)) we obtain following formulae:

$$H_{i,n} = \sum_k N_k^{i,n}$$

$$H_{n,i} = \bar{H}_{i,i} - H_{i,n}$$

$$H_{i,j} = H_{i,n} + H_{n,j} - \bar{H}_{i,j}$$

Standard $O(n^3)$ matrix inversion method based on Gaussian elimination finishes this pre-processing step of calculating the exact hitting-times in a few hours for the interaction networks we analyzed.

3 Estimating Multi-Source Hitting Time via Single-Source Hitting Times

Given $U = \{u_1, u_2, \ldots, u_k\}$, we now show how to estimate $H_{U,v}$ by a function of independent pairwise hitting times $H_{u_i,v}$ for all $u_i \in U$. The estimate we use is

$$H_{U,v} \approx \frac{1}{\sum_{i=1}^{k} \frac{1}{H_{u_i,v}}}$$

Let the conductance of graph G be defined as $\Phi(G) = \min_{\emptyset \subsetneq S \subsetneq V} \frac{|E(S,V\setminus S)|}{\min\{\mathrm{vol}(S),\mathrm{vol}(V\setminus S)\}}$, where $vol(S)$ is the sum of degrees of the vertices of S. Many real-world networks including preferential attachment graphs are known to have large conductance (Mihail et al. 2006). For such graphs, our next theorem provides mathematical evidence for the accuracy of our estimate in (3).

Theorem 2. *Let $G = (V,E)$ be any graph with constant conductance $\Phi > 0$. Then there is an integer $C = C(\Phi) > 0$ such that, given an integer k, a set of nodes $U = \{u_1, u_2, \ldots, u_k\}$ and node $v \in V$ satisfying $\frac{1}{k \cdot \frac{\deg(v)}{2|E|}} \geq \log^{1.5} n$, the following inequality holds:*

$$H_{U,v} \leq C \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(v)}{2|E|}}.$$

In particular, for any pair of nodes u, v with $\deg(v) \leq \frac{2|E|}{\log^{1.5} n}$ we have $H_{u,v} = O(\frac{|E|}{\deg(v)})$.

For the proof of Theorem 2, it will be convenient to consider a lazy version of the random walk which stays at the current node in each step with probability $1/2$. Note that any hitting time (single-source or multi-source) of the lazy version of the random walk is always an upper bound on the corresponding hitting time of the standard random walk.

Lemma 3. *Let $G=(V,E)$ be a graph with constant conductance $\Phi>0$. For any pair of nodes $u,v\in V$ and number of steps t with $\omega(\log n)\le t\le \frac{2|E|}{\deg(v)}$, let $\mathscr{A}_{u,v,t}$ be the event that a random walk starting from u visits v within t steps. Then*

$$\Pr[\mathscr{A}_{u,v,t}]\ge \frac{\Phi^2}{280}\cdot t\cdot\frac{\deg(v)}{2|E|}.$$

Proof. We first record the following useful inequality (See Levin et al. (2008) for details). Let $P^s_{x,y}$ be the probability that a random walk starting at x visits node y in step s. Then,

$$\left|P^s_{x,y}-\frac{\deg(y)}{2|E|}\right|\le\sqrt{\frac{\pi(y)}{\pi(x)}}\cdot\lambda_{\max}^t,$$

where $\pi(w)=\frac{\deg(w)}{2|E|}$ for any $w\in V$, $\lambda_{\max}=\max\{\lambda_2,|\lambda_n|\}$ with $1=\lambda_1\ge\cdots\ge\lambda_n>-1$ being the eigenvalues of the transition matrix P. Since the random walk has loop probability $1/2$, $\lambda_n\ge 0$ and thus $\lambda_{\max}=\lambda_2$. Furthermore, by Cheeger's inequality, $\lambda_2\le 1-\frac{\Phi^2}{8}$. Hence

$$\left|P^s_{x,y}-\frac{\deg(y)}{2|E|}\right|\le\sqrt{\frac{\pi(y)}{\pi(x)}}\cdot\left(1-\frac{\Phi^2}{8}\right)^t,$$

which implies for every s with $t/2\le s\le t$, as $t=\omega(\log n)$,

$$\left|P^s_{u,v}-\frac{\deg(v)}{2|E|}\right|\le n^{-4}.$$

Let X be the random variable counting the number of visits to v within the time-interval $[t/2,t]$. Then, from the above,

$$\frac{t}{2}\cdot\frac{\deg(v)}{2|E|}\le \mathrm{E}[X]\le 2t\cdot\frac{\deg(v)}{2|E|}.$$

To apply the second moment method, we will now analyze the variance of X, denoted by $\mathrm{V}[X]$. Note that

$X = \sum_{s=t/2}^{t} X_s$, where $X_s = 1$ if the random walk visits u in step s and $X_s = 0$ otherwise. Then,

$$\begin{aligned}
\mathrm{V}[X] &\leq \sum_{s=t/2}^{t} \mathrm{E}[X_s] + 2 \sum_{t/2 \leq s < s' \leq t} \Pr[X_s = 1 \wedge X_{s'} = 1] - \Pr[X_s = 1] \cdot \Pr[X_{s'} = 1] \\
&= \sum_{s=t/2}^{t} \mathrm{E}[X_s] + 2 \sum_{t/2 \leq s < s' \leq t} \Pr[X_s = 1] \cdot (\Pr[X_{s'} = 1 \mid X_s = 1] - \Pr[X_{s'} = 1]) \\
&\leq \mathrm{E}[X] + 2 \sum_{t/2 \leq s < s' \leq t} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left(\left(\frac{\deg(v)}{2|E|} + (1 - \frac{\Phi^2}{8})^{s'-s} \right) - \left(\frac{\deg(v)}{2|E|} - n^{-4} \right) \right) \\
&\leq \mathrm{E}[X] + 2 \sum_{t/2 \leq s \leq t} \sum_{1 \leq i \leq t/2} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left((1 - \frac{\Phi^2}{8})^{i} + n^{-4} \right) \\
&\leq \mathrm{E}[X] + 2 \sum_{t/2 \leq s \leq t} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left(\frac{8}{\Phi^2} + t/2 \cdot n^{-4} \right) \\
&\leq \mathrm{E}[X] \cdot \left(2 + \frac{32}{\Phi^2} \right) + O(n^{-2}) \leq \frac{35}{\Phi^2} \cdot \mathrm{E}[X].
\end{aligned}$$

By the Paley-Zygmund inequality, for any $0 < \delta < 1$,

$$\Pr[X \geq \delta \cdot \mathrm{E}[X]] \geq (1-\delta)^2 \cdot \frac{\mathrm{E}[X]^2}{\mathrm{V}[X] + \mathrm{E}[X]^2} \geq (1-\delta)^2 \cdot \frac{1}{\frac{35}{\Phi^2} \cdot \frac{1}{\mathrm{E}[X]} + 1} \geq (1-\delta)^2 \cdot \frac{\Phi^2}{2 \cdot 35} \cdot \mathrm{E}[X],$$

where the last inequality follows from $\mathrm{E}[X] \leq 2$ which holds thanks to our upper bound on t. Choosing $\delta = \frac{1}{2}$ implies, as X is an integer-valued random variable,

$$\Pr[A_{u,v,t}] = \Pr[X \geq 1] \geq \Pr\left[X \geq \frac{1}{2} \cdot \mathrm{E}[X] \right] \geq \frac{\Phi^2}{8 \cdot 35} \cdot \mathrm{E}[X],$$

and due to the lower bound on $\mathrm{E}[X]$ derived earlier, the proof is finished. □

With the lemma at hand, we are now able to complete the proof of Theorem 2.

Proof. For any integer $\alpha \geq 1$, define $\tau = \tau(\alpha) := \alpha \cdot \frac{280}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(v)}{2|E|}}$. For any $1 \leq i \leq k$, let $\mathscr{E}_i$ be the event that the random walk starting from u_i does *not* visit v within τ steps. By partitioning the τ steps into consecutive sections of length $\log^{1.5} n$ and applying Lemma 3 to every section, we conclude that

$$\Pr[\mathscr{E}_i] \leq \left(1 - \frac{\Phi^2}{280} \cdot \log^{1.5} n \cdot \frac{\deg(v)}{2|E|} \right)^{\tau / \log^{1.5} n} \leq \exp\left(-\tau \cdot \frac{\Phi^2}{280} \cdot \frac{\deg(v)}{2|E|} \right).$$

As all k random walks are independent, it follows that

$$\Pr\left[\bigwedge_{i=1}^{k} \mathscr{E}_i\right] = \prod_{i=1}^{k} \Pr[\mathscr{E}_i] \leq \exp\left(-\tau \cdot \sum_{i=1}^{k} \frac{\Phi^2}{280} \cdot \frac{\deg(v)}{2|E|}\right) = \exp(-\alpha) \leq 2^{-\alpha}.$$

Hence the expected multi-source hitting time can be estimated as follows,

$$H_{\{u_1,\ldots,u_k\},v} \leq \frac{280}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(v)}{2|E|}} \cdot \sum_{\alpha=1}^{\infty} \alpha \cdot 2^{-\alpha} \leq \frac{560}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(v)}{2|E|}}$$

□

Note that the bound in Theorem 2 differs from our estimate in equation (3) in that $\frac{1}{H_{u_i,v}}$ is replaced by $\frac{\deg(v)}{2|E|}$. However, for graphs with constant conductance, we have $H_{u,v} \leq H_{\pi,v} + O(\log n)$, where $H_{\pi,v}$ is the hitting time for a random walk starting according to the stationary distribution π, given by $\pi(w) = \frac{\deg(w)}{2|E|}$ for every $w \in V$. Hence $\frac{2|E|}{\deg(v)} = H_{v,v} \leq H_{\pi,v} + O(\log n)$. Since $H_{\pi,v} = \sum_{u \in U} \pi(u) \cdot H_{u,v}$, it follows that, given any fixed node v, it holds for "most nodes" u that $H_{u,v}$ is not much smaller than $\frac{2|E|}{\deg(v)} - O(\log n)$.

Since Theorem 2 does not provide a strong mathematical bound, we attempted to measure the quality of the estimate by comparing it to exact multi-source hitting time values of randomly chosen set U. Obviously, computing the exact multi-hitting time over all possible sets of facilities in the network is computationally not feasible. Therefore, we performed 1000 iterations of the following experiment on the STRING v10 interaction network:

- Choose 10 nodes of the network uniformly at random, to be a set of facilities U.
- For each node $v \notin U$ estimate multi-source hitting time by performing 5000 random walk simulations from every $u \in U$ and in each simulation, measure the minimum time required for the first one of them to reach v. Take the average of the observations and call it the exact multi-hitting time $MHT_{U,v}$.
- Compute the relative error of the estimate $HT_{U,v}$ (estimated based on pair-wise hitting times) compared to $MHT_{U,v}$.
- Take the average relative error over all nodes $v \notin U$.

With 5000 random walk simulations, we hoped to get accurate enough estimate to be able to compute the relative error accurately without it taking too long. Over the 1000 iterations, we observed average error rate of 5.96%, with average error over non-facility nodes per iteration ranging from 3.03% to 8.06%. This suggests that our estimate

(which has the practical benefit of being linear and thus useful in a linear programming setting) is quite accurate in practice.

Furthermore, we would like to point out that the most extreme cases (of our approximation being inaccurate) are where one of the $u \in U$ is an immediate neighbour of v of degree one, so that $H_{u,v} = 1$ and the multi-hitting time of the entire set U should be 1. It is easy to see that if v has k such neighbours then the estimate will be approximately $\frac{1}{k}$, when it should be 1. We have analysed the candidate driver-outlier pairs in the bipartite graph that have hitting time 1 and obtained the following numbers: (a) TCGA-BRCA – 11 pairs with hitting time 1, where maximum number of candidate drivers connected to a single outlier with hitting time 1 is 1; (b) TCGA-GBM – 0 pairs with hitting time 1; (c) TCGA-OV – 12 pairs with hitting time 1, where maximum number of candidate drivers connected to a single outlier with hitting time 1 is 1; (d) TCGA-PRAD – 1 pair with hitting time 1. Since average non-zero hitting time in the STRING v10 hitting time matrix is 129322, and selected number of drivers is in order of tens (meaning that in most cases the estimate of $H_{U,v}$ will be inverse of sum of number 1 and inverses of much larger numbers), the estimate in these extreme cases will be close to 1 which represents the exact solution. Combined with the the results of the previously-explained randomized test that estimated average error to be 5.96%, it serves as further evidence that our approximation is quite accurate in practice.

4 Datasets and analysis

4.1 The Cancer Genome Atlas (TCGA)

We used publically available datasets of four major cancer-types glioblastoma multiforme (GBM) (The Cancer Genome Atlas Research Network 2008), Ovarian serous cystadenocarcinoma (OV) (The Cancer Genome Atlas Research Network 2011), breast adenocarcinoma (BRCA) (The Cancer Genome Atlas Research Network 2012), and prostate adenocarcinoma (PRAD) (The Cancer Genome Atlas Research Network 2015) from The Cancer Genome Atlas (TCGA) project. All data were obtained from TCGA data-portal in May 2014 which were mapped to GRCh37 genome build. Although TCGA has recently made available all data re-aligned to the newer GRCh38 genome build, to ensure compatibility, all TCGA data we have used in this study has been mapped to GRCh37.

4.1.1 Somatic mutations

Somatic mutation calls (level 2 data) from all available platforms/centers were merged. Only missense mutations, nonsense mutations and splice-site SNPs were marked as somatic-mutation alteration events.

4.1.2 Copy number aberrations

Copy number aberrations for GBM and OV, Agilent Human Genome CGH Microarray 244A (level 1) data files were used and for PRAD and BRCA, Affymetrix Genome-Wide Human SNP Array 6.0 (level 3) data files were used to generate the copy number profiles.

These Agilent FE format sample files were loaded into BioDiscovery Nexus Copy Number software v7.0, where quality was assessed and data was visualized and analyzed. All samples were mapped to the most recent genome build (hg 19, GRCh37) via Agilent probe identifiers and annotation (downloaded from Agilent's website) based on the 1M SurePrint G3 Human CGH Microarray 1x1M design platform. BioDiscovery's FASST2 segmentation algorithm, a Hidden Markov Model based approach, was used to make copy number calls. The FASST2 algorithm, unlike other common HMM methods for copy number estimation, does not aim to estimate the copy number state at each probe but uses many states to cover more possibilities, such as mosaic events. These state values are then used to make calls based on a log-ratio threshold. The significance threshold for segmentation was set at $= 5X10^{-6}$) also requiring a minimum of 3 probes per segment and a maximum probe spacing of 1000 between adjacent probes before breaking a segment. The log ratio thresholds for single copy gain and single copy loss were set at 0.2 and -0.23, respectively. The log ratio thresholds for two or more copy gain and homozygous loss were set at 1.14 and -1.1 respectively. Upon loading of raw data files, signal intensities are normalized via division by mean. All samples are corrected for GC wave content using a systematic correction algorithm. Only the high confidence copy number aberrations i.e. high copy number gain or homozygous deletions were marked as copy-number aberrant events. Finally, genes that harbour either a somatic-mutation aberrant event or a copy-number aberrant event were taken to be the final list of abberant genes at the genomic level.

4.1.3 Gene expression

We used microarray based gene-expression (Affymetrix HT Human Genome U133 Array Plate Set) (level-1) for GBM and OV data sets. Where as for BRCA and PRAD data sets, RNA-seq derived gene-expression were used (level-3). Gene expression profiles of normal and tumor phenotype were used as sample groups.

4.1.4 Gene fusions

Transcript fusions prediction calls for GBM, OV, BRCA and PRAD were obtained from TCGA Fusion gene Data Portal (http://www.tumorfusions.org) (Yoshihara et al. 2014). The fusion partner genes were tagged for gene-fusion alteration.

4.1.5 Colorectal cancer data

We obtained matched genomic (somatic mutation level-3, somatic copy-number aberration level-3) and transcriptome (gene-expression FPKM-UQ level-3) data for 429 samples from TCGA colorectal cancer (TCGA-COAD) project (data downloaded on 28th March 2017). This included 78 hypermutated cases and 351 non-hypermutated cases.

4.2 Genomics of drug sensitivity in cancer

Somatic mutation, copy-number alterations and gene-expression, and drug screening data of cancer cell lines were downloaded from Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al. 2016) website http://www.cancerrxgene.org/downloads. Data downloaded on August 2016.

4.3 Interaction networks

We used STRING version 10 (Szklarczyk et al. 2015) protein-interaction network which contains high confidence functional protein-protein interactions (PPI). Self-loops and interactions with missing HGNC symbols were discarded and interaction scores were divided by 1000 to obtain percentage-like reliability score. Only high confidence interactions with combined score of 0.9 or greater were selected. As a result we obtained a network of 10971 nodes with 214298 interactions.

In the case of prostate cancer, we integrated STRING-10 protein-protein interaction network with protein-DNA interaction network derived from Chip-seq experiments for transcription factors highly relevant to prostate cancer - *REST*, *FOXA1*, *AR*, *EZH2* (Sharma et al. 2013) and *ERG* (Rickman et al. 2012) resulting in a new combined network of 13517 nodes and 220190 interactions.

To simulate HIT'nDRIVE using different underlying network we used two additional interaction networks: Human Protein Reference Database - Protein-Protein Interaction Database (HPRD-PPI) network (version 9.0) (Keshava Prasad et al. 2009) and REACTOME pathway database (version 2015) (Fabregat et al. 2016).

4.4 Pathway enrichment analysis

The selected set of genes were tested for enrichment against gene sets of pathways present in Molecular Signature Database (MSigDB) v5.0 (Subramanian et al. 2005). A Fisher's exact test based gene set enrichment analysis was used for this purpose. A cut-off threshold of false discovery rate (FDR) ≤ 0.01 was used to obtain the significantly enriched pathways. An *R* implementation of GESD test is available at https://github.com/raunakms/GSEA-Fisher. Same procedure, as above, is used to assign biological functional to the gene-modules.

4.5 Validation dataset

For the validation of driver-modules we used the following gene-expression datasets: GBM: Murat-2008 (Murat et al. 2008), Sun-2006 (Sun et al. 2006); OV: Yoshihara-2009 (Yoshihara et al. 2009), Bowen-2009 (Bowen et al. 2009); PRAD: Taylor-2010 (Taylor et al. 2010), Grasso-2012 (Grasso et al. 2012), SMMU-PC (Second Military Medical University - prostate cancer patient cohort); BRCA:METABRIC (Curtis et al. 2012) and Richardson-2006 (Richardson et al. 2006).

5 Derivation of expression outlier genes

We used generalized extreme studentized deviate (GESD) test (Rosner 1983) to obtain the outlier genes. Unlike Grubbs test and the Tietjen-Moore test, GESD test only requires that an upper bound for the suspected number of outliers be specified. Given the upper bound, r, the GESD test essentially performs r separate tests: a test for one outlier, a test for two outliers, and so on up to r outliers.

An *R* implementation of GESD test is available at https://github.com/raunakms/GESD

Hypothesis: The GESD test is defined for the hypothesis:

- H_0: There are no outliers in the data set
- H_a: There are up to r outliers in the data set

Test statistic: Compute

$$R_i = \frac{max_i|x_i - \mu|}{\sigma}$$

with μ and σ denoting the sample mean and sample standard deviation, respectively. Remove the observation that maximizes $|x_i - \mu|$ and then recompute the above statistic with $n-1$ observations. Repeat this process until r observations have been removed. This results in the 'r' test statistics $R_1, R_2, ..., R_r$.

Critical region: Corresponding to the r test statistics, compute the following r critical values

$$\lambda_i = \frac{(n-i)t_{n-i-1,p}}{\sqrt{(n-i-1+t^2_{n-i-1,p})(n-i+1)}}$$

where $i = 1, 2, ..., r$, and t_p, v is the 100_p percentage point from the t distribution with v degrees of freedom.

$$p = 1 - \frac{\alpha}{2(n-i+1)}$$

here, α denotes the significance level.

The number of outliers is determined by finding the largest i such that $R_i > \lambda_i$

6 Derivation of expression outlier gene weights

Outlier-gene weights were calculated as follows: Let i denote genes, j denote patients and x_{ij} denote the gene-expression value of gene i in patient j. We then calculated the absolute value of z-score (z_{ij}).

$$z_{ij} = \frac{|x_{ij} - \mu_i|}{\sigma_i}$$

where, μ_i and σ_i respectively denotes mean and standard deviation of expression value of gene i. Next we performed Student's t-test in the gene-expression values of normal and tumor phenotypes. where, $\psi_i = -log(pvalue_{ttest})$. Finally, we calculate the outlier weight ω_{ij} as

$$\omega_{ij} = \frac{\psi_i z_{ij}}{\sum_i \psi_i z_{ij}}$$

7 HIT'nDRIVE sensitivity to least frequent driver gene

To demonstrate the sensitivity of HIT'nDRIVE to detect infrequent driver genes, we performed the following *in-silico* experiment. We selected 1000 TCGA-BRCA tumors. We label these 1000 TCGA-BRCA samples as the "Original" set of tumor samples. Our objective was to sub-sample TCGA-BRCA tumors with different sample-sizes such that the frequency distributions of mutations in the selected sub-samples are similar to that of original 1000 BRCA tumors. For this we first estimated the alteration-frequency distribution of the original 1000 tumor samples and calculated the mean (μ_{target}) and standard-deviation (σ_{target}) of the distribution. Our aim here is to find the sub-set of samples (with defined sample-size) such that the mean (μ_{obs}) and standard-deviation (σ_{obs}) of the sub-sampled population is very close to μ_{target} and σ_{target}) respectively. This can be represented as the following:

$$MINIMIZE \ (Score = |\mu_{obs} - \mu_{target}| + |\sigma_{obs} - \sigma_{target}|)$$

Here, we gave equal penalty to both the attributes mean and standard-deviation. We took a heuristic approach to solve the problem. We randomly sub-sampled, with a user defined sample-size, from the original set of tumor samples and calculated the above score. This step was repeated for 10,000 times (i.e. 10,000 different combination of samples of defined sample-size). Then the sub-sample set with least score was chosen for further HIT'nDRIVE analysis.

8 Association of driver modules with patients' survival outcome

To test for association of driver modules with patients' survival outcome, we developed a risk-score based on multi-gene (component genes of the module) expression. The risk-score (S) defined as a weighted sum of the normalized gene-expression values of the component genes in the module weighted by their estimated univariate Cox proportional-hazard regression coefficients (Beer et al. 2002) as given in the equation below.

$$S = \sum_{i}^{k} \beta_i x_{ij}$$

Here i and j represents a gene and a patient respectively, β_i is the coefficient of cox regression for gene i, x_{ij} is the normalized gene-expression of gene i in patient j, and k is the number of component genes in a gene-module. The normalized gene-expression values were fitted against overall survival time with living status as the censored event using univariate Cox proportional-hazard regression (Exact method).

Based on the risk-score values, patients were stratified into two groups: low-risk group (patients with $S < 33$ percentile of S), and high-risk group (patients with $S > 66$ percentile of S). Patients that fall in between (i.e. patients with $S >= 33$ percentile of S and $<= 66$ percentile of S) were discarded from the further analysis as these patients fall into intermediate-risk group and are bound to introduce noise while performing log-rank test.

Both Cox regression coefficients of each gene and risk-score cutoff values for each module were estimated from TCGA-BRCA cohort (training dataset), later these values were applied to METABRIC cohots (test dataset). To assess whether the risk-score assignment to high/low categories was valid, a log-rank test was performed for each module in both training and test datasets.

Finally, to identify the significant list of driver-modules that were robust enough to predict patients' survival, we calculated log-rank test pvalue, hazard-ratio (HR) (Wald test) and concordance-index (c-index) (Wald test).

9 Statistical significance of the overlap of driver genes with that of CGC database.

Suppose, for a cohort of cancer patients, we predict n_{total} number of driver genes using HIT'nDRIVE, out of which n_{cgc} number of driver genes are present in the Cancer Gene Census (CGC) database (of known cancer driver genes). Let, x be the total number of sequence altered genes (i.e. all potential driver genes) and let y of these x sequence altered genes be in CGC. This means that the probability that a randomly selected gene out of these sequence altered genes happens to be a CGC gene is $(\frac{y}{x})$.

The probability (p-value) that at least n_{cgc} out of n_{total} driver genes are identified in CGC is:

$$pvalue = \sum_{i=n_{\text{cgc}}}^{n_{\text{total}}} \binom{n_{\text{total}}}{i} \left(\frac{y}{x}\right)^{i} \left(1-\frac{y}{x}\right)^{n_{\text{total}}-i}$$

Next we consider driver genes in each patient. We also calculated the p-value for HIT'nDRIVE to pick at least p CGC drivers out of p' and pick at most q non-CGC drivers out of q' as follows

$$pvalue = \sum_{x=p'}^{x=p'+q'} \binom{p+q}{x} \left(\frac{p}{p+q}\right)^{x} \left(\frac{q}{p+q}\right)^{p'+q'-x}$$

Supplemental Results

10 HIT'nDRIVE: parameters

HIT'nDRIVE uses three user-specified input parameters:

1. α: fraction of outliers to be covered overall (across **all** patients)
2. β: fraction of outliers to be covered in **each** patient
3. γ: fractional lower bound on the sum of the incoming edge weights from driver genes selected by HIT'nDRIVE

HIT'nDRIVE is robust with respect to the changes in α and β but is somewhat sensitive to γ (Supplemental Fig. S1), as expected. However, as γ grows, the driver genes identified by HIT'nDRIVE do not change but simply grow in number by the addition of new driver genes, which indicates robustness of our method with respect to γ as well.

11 HIT'nDRIVE: outlier stringency

The higher the stringency we apply on the expression value change in a potential outlier, the fewer outliers we will identify, which in turn will result in fewer number of driver genes. However, the new set of driver genes obtained are, in general, a subset of the first set of driver genes, again indicating robustness (Supplemental Fig. S2).

12 HIT'nDRIVE: network perturbation

We used STRING v10 network for our analysis. The edges of the STRING v10 network was perturbed to different extent (between 1-10%) preserving the degree of the nodes in the network. HIT'nDRIVE analysis was performed using different perturbed networks. Proportion of common driver genes between the unperturbed network and each of the perturbed network were calculated (Supplemental Fig. S3). We observed that even though the edges of the network were perturbed, the list of driver genes did not change to a great extent (i.e. the overlap of driver genes was very high) as compared to the non-perturbed network even when the edges of the network were perturbed by up to 10%. This clearly demonstrates that HIT'nDRIVE is not biased towards network perturbations.

13 HIT'nDRIVE: underlying network

We evaluated the robustness of HIT'nDRIVE on three networks, namely STRING, HPRD and the REACTOME. Only 34% of the vertices in STRING, HPRD and the REACTOME are shared in all three networks; in terms of edges, an even smaller proportion of the edges. Not surprisingly, the more nodes the network has, the more driver

genes HIT'nDRIVE predicts. This is consistently observed across various parameter settings. What is noteworthy is that the percentage overlap between the driver genes predicted on the three networks is quite robust, i.e., the percentage of driver genes shared between all three networks is preserved across various parameter settings - e.g. this overlap is above 60% between the REACTOME and any of the other two networks, across various values of gamma - which is quite impressive. In fact the driver genes predicted on STRING are almost a superset of those predicted on REACTOME. See Supplemental Fig. S4.

14 HIT'nDRIVE: alteration types

We ran HIT'nDRIVE with SNVs, gene fusions and CNVs independently to evaluate how the resulting driver genes compare to those obtained by HIT'nDRIVE when applied to SNVs, gene fusions and CNVs simultaneously. The results again demonstrate that HIT'nDRIVE is reasonably robust with respect to the treatment of potential driver genes especially involving SNVs and CNAs, across various values of the gamma parameter (again, the choice of alpha and beta do not alter the results in a meaningful manner). For gene fusions, the fact that overlap is lower is not statistically meaningful as very few of the driver events are gene fusions. See Supplemental Fig. S5.

15 HIT'nDRIVE: random alterations and random outliers.

We compared the HIT'nDRIVE predictions of driver genes among observed mutations with those obtained through randomized mutations (Supplemental Fig. S6A) and random outliers (Supplemental Fig. S6B). There is a stark contrast between the two sets of driver gene predictions with respect to their overlap with the Cancer Gene Census (CGC) data set - conserved through different values of the γ parameter (the overlap is generally preserved across various settings of the remaining two parameters, namely α and β). Driver genes predicted in the non-randomized alteration (or non-randomized outliers) data not only (i) included a higher number of CGC genes (i.e. more number of true driver genes) as compared to that in driver genes predicted from randomized alterations (or randomized outliers) data, but also (ii) the number of CGC driver genes predicted through the use of non-randomized data increased quickly with increasing γ parameter, whereas it stays roughly the same when randomized data was used. Note that while performing randomization, the original gene labels (sequence-wise altered genes or expression-outlier genes) were randomly replaced by new ones while preserving their recurrence frequency distributions.

16 Modified HIT'nDRIVE: when it is not required to prioritize one driver gene per patient.

In HIT'nDRIVE, at least one gene is picked per patient (i.e. when the $\beta > 0$). This constraint is based on the implicit assumption that at least one causal mutation should be driving cancer (although there could be exceptions to this,

for example, the driver event could be something other than genomic alteration, and be in the form of methylation, aberrant expression of a regulatory RNA or a metabolite, they could all be incorporated in our framework, given matching data - which unfortunately is not available through TCGA). There are also important performance issues related to the value of beta: (1) Setting $\beta > 0$ significantly improves the robustness of our method with respect to the alpha parameter. In Supplemental Fig. S1, it can observed that the alpha parameter has minimal effect on the output of our method - provided beta is non-zero. If $\beta = 0$ (i.e. patients do not necessarily have one driver gene), our method is less robust, as can be seen in Supplemental Fig. S7B. In Supplemental Fig. S7C, especially for small values of alpha, the number of patients that do not have a driver gene increases as the value of gamma decreases. In the worst case, $\sim$40% of patients do not report a driver gene; this happens when $\alpha = 0.5$ and $\gamma = 0.02$. For guaranteeing robustness, the γ value should be set above 0.2 and the α value should be set above 0.7, which reduces to the fraction of patients with no driver genes to 5%. (2) Setting $\beta = 0$ significantly increases the running time of our method, from a couple of minutes to several days on very large datasets.

17 HIT'nDRIVE's ability to capture CGC genes

To check if HIT'nDRIVE is able to capture the true driver genes, we perform the following analysis. For the sake of this analysis, let us first assume that the cancer-type specific genes listed in CGC database are the true driver genes i.e. the ground truth. As described in the main manuscript text, we predicted potential driver genes in patients from four major cancer types using HIT'nDRIVE. For every patient analyzed, we compared the input (i.e. all sequence-wise altered gene) and the output (i.e. subset of the input sequence-wise altered genes that are predicted as potential driver genes) data for HIT'nDRIVE. We compared the amount of CGC true driver genes present in the input data versus amount of CGC true driver genes captured by HIT'nDRIVE.

The Supplemental Fig. S17 summarizes the results of this analysis. As can be seen, the likelihood of a sequence-wise altered CGC gene to be prioritized by HIT'nDRIVE is much higher than that of a non-CGC genes. Next, for each patient, we calculated the likelihood of HIT'nDRIVE to capture CGC genes (see supplemental methods for detailed p-value calculation). We found that majority of the samples analyzed have a very significant p-value (i.e. < 0.01) (Supplemental Fig. S18). This analysis demonstrates that HIT'nDRIVE is able to capture cancer driver genes, to a larger extent, in the patient samples analyzed.

18 Patient druggability

We checked for overlap of HIT'nDRIVE predicted driver genes with the druggability information from Rubio-Perez et al. (2015). In GBM, unlike other cancer types, more than 50% of patients could benefit from FDA approved

drugs. It was intriguing to note that a larger fraction of patients could actually benefit from drugs in clinical trials. The patient druggability data presented here is more or less similar to that presented in Rubio-Perez et al. (2015). However, slight differences were present due to different numbers of patient resulting in different number of driver genes in the two studies. Especially in the case of PRAD, the data presented in (Figure 2C) and Supplemental Fig. S16 had striking difference which was primarily due to discrepancies in the druggability databases. Majority of PRAD patients harboured TMPRSS2-ERG fusion which can be targeted using prap inhibitor (Chatterjee et al. 2013). This information is well covered in TARGET database but not present in data obtained from Rubio-Perez et al. (2015) resulting in far less number of patients that can benefit from targeted therapies (Supplemental Fig. S19).

We also assessed if the driver genes predicted by HIT'nDRIVE represents the targets of known anti-cancer drugs. For this we leveraged drug-target information for the drugs used in Genomics of Drug Sensitivity in Cancer project (Iorio et al. 2016). For every patient analyzed, we identified the drugs that could potentially be targeted against the (A) driver genes (predicted by HIT'nDRIVE) in each patient, (B) sequence-wise altered CGC genes present in each patient and (C) CGC genes prioritized as driver genes by HIT'nDRIVE in each patient.

We grouped the drugs into three tiers based on their level of clinical approval - Tier-I: clinically approved drugs, Tier-II: drugs currently in clinical trials and Tier-III: pre-clinical drugs We considered the potential driver genes, (A) either predicted by HIT'nDRIVE, (B) or the sequence-wise altered CGC genes, or (C) the intersection of (A) and (B) for each patient (Supplemental Fig. S20). Among these potential driver genes, we identified those which are targeted by clinically approved (Tier-I) drugs. About 75% of GBM patients and over 60% of OV and BRCA could be targeted by at least one such drug, i.e. these patients have at least one potential driver gene that is targeted by the drugs tested. In case of PRAD only about 20% of the patients have a potential driver gene targeted by a drug but this can be easily attributed to the fact that prostate cancer drugs primarily target androgen receptor *AR* which typically is not genomically altered but is rather alternatively spliced in primary prostate tumors. (We did not consider alternative splicing events in this study as potential drivers due to lack of RNA-Seq data.) We note that the proportion of patients with at least one drug targetable potential driver gene increases as we consider Tier-II and Tier-III drugs in addition to Tier-I drugs - demonstrating that driver genes (be it HIT'nDRIVE predicted or CGC driver genes) are indeed known targets for anti-cancer drugs.

19 Unsupervised classification of gene expression data

Our rationale for using driver-module identified by OptDis as a feature of classifying phenotypes (eg. tumor vs normal) is that the observable effects of true driver alterations on their immediate vicinity (in an interaction network) should be sufficient to discriminate normal samples from tumour samples. The genes in the immediate vicinity

of "top" potential driver genes are very limited in size compared to the entire gene set to be used by unsupervised clustering approach. To demonstrate this, we computed cross validation performance of whole set of differentially expressed genes and compared it to the accuracy of the modules on the TCGA-BRCA subtype expression datasets (Supplemental Fig. S24). There were 4657 genes on which the classifier was trained (using R caret package (Kuhn et al. 2016) LGOCV train method), achieving following accuracy: 93.41% in Basal, 88.16% in Her2, 66.08% in Luminal-A and 77.02% in Luminal-B. The fact that the classifier using only a few modules composed of genes (in the immediate vicinity of HIT'nDRIVE identified driver genes) performing better (Supplemental Fig. S24, Supplemental Fig. S25) than a classifier (albeit unsupervised) with access to the entire set of genes provides a strong evidence that HIT'nDRIVE identified genes are likely to be true driver genes.

The classification of breast subtypes (in TCGA samples) based on their gene-expression profiles using an unsupervised classification approach is shown in the Supplemental Fig. S30. Based on the dendrogram in the figure, BASAL subtype is very easy to classify as it forms a separate cluster. But in the case of rest of other subtypes, unsupervised classification did not reveal a distinguishable clusters hence difficult to classify. However, OptDis (supervised) classification outperformed unsupervised classification for every breast cancer subtypes. For these reasons, unsupervised classification is not a suitable approach for the problem. This clearly demonstrates the superiority of HIT'nDRIVE-OptDis classification approach using driver modules over naive unsupervised clustering approach.

20 Phenotype classification using CGC gene seeded modules

To evaluate the difference between HIT'nDRIVE predicted driver genes and a list of known driver genes, we performed the following experiments. First we compared the HIT'nDRIVE driver seeded module with CGC gene seeded module to classify tumor vs normal samples in TCGA-PRAD patient cohort. Note that among the four TCGA cancer cohorts we study in this paper, only the PRAD cohort includes non-trivial number of patients with no known driver genes (based on an unpublished study by PCAWG project) and thus provides a good testbed for novel driver gene identification by HIT'nDRIVE. As can be seen, HIT'nDRIVE identified driver seeded modules provide higher classification accuracy, potentially due to novel driver genes identified by HIT'nDRIVE.

The top HIT'nDRIVE modules associated with PRAD are seeded by (in the order of discriminative ability) *ERG*, *FOXA1*, *ERG/ACAN*, *PTEN* and *CDKN1B* (Supplemental Fig. S23A). All but *ACAN* are CGC genes associated with PRAD. HIT'nDRIVE successfully identified all these driver genes without the use of any information related to known PRAD driver genes from CGC. In addition, HIT'nDRIVE identified *ACAN*, a non-CGC gene as a potential driver gene of PRAD. In comparison, the modules identified for CGC PRAD driver genes were seeded by (again

in the order of discriminative ability) *ERG*, *FOXA1*, *NCOR2*, *BRAF*/*ERG* and *AR*/*KLK2* - missing *PTEN* due to potentially large overlap with other modules. Overall, the modules seeded by HIT'nDRIVE identified driver genes provide a higher accuracy in discriminating PRAD than CGC PRAD driver genes.

Next, we compared HIT'nDRIVE driver genes to CGC genes in breast cancer subtypes in TCGA-BRCA patient cohort. Note that breast cancer is possibly the best studied cancer type with respect to driver genes Thus it is not surprising that Basal, Her2 and Luminal-B subtypes show negligible differentiation between HIT'nDRIVE predictions and CGC based predictions (Supplemental Fig. S23B). This is due to big overlap between HIT'nDRIVE discovered modules and CGC modules (e.g. in BASAL, top 4 HIT'nDRIVE modules almost perfectly match the top 4 CGC modules - which, again, is not surprising since BRCA is a very well studied cancer with respect to driver genes). However, HIT'nDRIVE show some advantage in Luminal-A. HIT'nDRIVE outperformed the CGC genes from 43rd module onward. This may be due to HIT'nDRIVE predicted driver genes (seeds) such as *DMD*, *ROCK1*, *AGAP1*, *SHANK2* which are not part of CGC and these genes play important role in cancer.

21 Mutual exclusivity of driver modules

To investigate the association between the seed driver gene and component genes in the resulting driver module, for each cancer type, we first selected the top-50 scoring sub-networks and then combined the sub-networks seeded by the same driver gene. This resulted in 33, 36, 29 and 33 driver modules in GBM, OV, PRAD and BRCA cohorts respectively (Supplemental Table S10-13).

21.1 Glutathione S-Transferase (GST) module

We found Glutathione S-Transferase (GST) module as a top-scoring module in OV. The GST module consisted of 3 members of GST protein family - *GSTT1*, *GSTM5* and *GSTA3*; 2 members of Cytochrome P450 (CYP) protein family - *CYP2BC* and *CYP3A5*; 2 members of Alcohol Dehydrogenase (ADH) protein family - *ADH1B* and *ADH6*; 1 member of UDP Glucuronosyltransferase (UGT) family - *UGT2B17*; and Monoamine Oxidase B (*MAOB*). Developments of ovarian tumors are primarily regulated by female sex hormone - estrogen. Metabolism of estrogen may cause DNA damage by the formation of mutagenic DNA adducts and by generation of free radicals. Estradiol (estrogen) gets activated by CYP3A5 generating 4-hydroxyestradiol, gets oxidized to quinone intermediates leading to a carcinogenic pathway (Supplemental Fig. S42). GSTs and UGTs inactivate the estrogen and its intermediate metabolites, avoiding the formation of carcinogens, thus protecting cells against free radical damage and initiate tumor cell response against adjuvant cancer therapies including radiation and chemotherapy (Guillemette et al. 2004; Tew et al. 2011). Interestingly, we found mutual exclusivity of *GSTT1* and *UGT2B17* driver alterations in

OV patients (Supplemental Fig. S26A), both of which detoxify estrogen into inactive metabolites. Almost half of the patients in the OV cohort harboured *GSTT1* and/or *UGT2B17* alterations among which around 90% of the patients had homozygous deletions in either or both of the genes. Expression patterns of CYP and GST proteins have been known to influence the response to drug treatment and overall survival of the OV cancer patient (Ekhart et al. 2009). ADH proteins helps in ethanol metabolism to acetaldehyde, which inactivates GST proteins and thus inhibits anti-oxidative defence system and DNA-repair pathways.

21.2 Phosphoinositide-3-Kinase (PI3K) module

The second top-ranked module in BRCA was Phosphoinositide-3-Kinase (PI3K) module, which included four mutually exclusive driver genes: 3 members of PI3K protein family - *PIK3CA*, *PIK3R1*, *PIK3C2B*; and 1 member of Protein Tyrosine Phosphate (PTP) protein family - *PTPRM* (Supplemental Fig. S26C). Other genes in the module include *EGFR*, *PLCZ1*, *TNS1*, *PVRL3*, *SPRY2* and *FIGF* (VEGFD). PI3K pathway is frequently activated in many cancer types including BRCA as a result of genetic alteration targeting its key components (Fruman and Rommel 2014). Thus EGFR, PI3K, AKT and mTOR inhibitors are often used for different cancer types. HIT'nDRIVE analysis demonstrated the prevalence of genetic alterations in components of PI3K module: at least 34% of BRCA patients harboured alterations in this module. PI3K pathway is involved in regulation of diverse cellular processes (Weigelt and Downward 2012; Gordon and Banerji 2013), including cell proliferation, survival, and migration (Supplemental Fig. S45).

Although mutational heterogeneity between cancer patients adds noise to the data making cancer driver discovery more challenging, combination of different genetic alterations could lead to similar disease phenotype. Many recent cancer studies have shown that cancer driver genes often show a pattern of mutual exclusivity and are functionally related. The pairs of driver genes that show significant mutual exclusivity are also likely to demonstrate synthetic lethality interactions.

22 HIT'nDRIVE using regulatory networks

Our study focuses primarily on PPIs. A sub-network in this context is an amalgamation of chains of PPIs (Supplemental Fig. S43). In a limited way we have also attempted to enrich the PPI networks by incorporating (a) HiC data - with the premise that genes that are in close 3-D proximity in the nucleus encode proteins with a higher chance of interacting physically (allowing us to indirectly measure the likelihood of a PPI), and (b) ChIP-Seq data - from which we infer interactions involving transcription factor binding to the promoter region of associated genes. The HiC assisted enrichment of the network did not alter our results in a meaningful manner and thus the associated

results are excluded from the manuscript. The ChIP-Seq assisted enrichment, on on the other hand is meaningful and the associated results can be found below.

STRING v10 is a functional network and does contain regulatory interaction. However, the known regulatory interaction (i.e. directed interaction network in general) is very limited as compared to undirected protein-interaction network. This would limit the nodes in the influence matrix hence many critical driver genes of cancer and its interactors would not be left out. For this reason we decided to use undirected protein-interaction network. However, our method can use both directed (i.e. regulatory network) as well as undirected network.

In the case of prostate cancer, we integrated STRING-10 protein-interaction network with protein-DNA interaction network derived from Chip-seq experiments for transcription factors (i.e. TF-target network) highly relevant to prostate cancer - *REST*, *FOXA1*, *AR*, *EZH2* (Sharma et al. 2013) and *ERG* (Rickman et al. 2012) resulting in a new combined network of 13517 nodes and 220190 interactions.

Supplemental Fig. S44A shows a driver module with *PTEN* is a driver gene in TCGA-PRAD regulating its downstream genes. In the module, *EPHB4* (upregulation) is negatively regulated by *PTEN* (deletion/mutation, downregulation). Supplemental Fig. S44B shows a module containing TF-target interaction of *EZH2* and its target *MAPKAPK5*. Similarly, Supplemental Fig. S44C shows a driver module with *ERG* as a driver gene and its several targets as other component genes in the module.

23 Correlation of predicted driver genes with alteration burden.

To obtain the mutation rate, we calculated the somatic mutation frequency per Mb (considering mutations in protein-coding genes only). We obtained copy-number burden values (i.e. percentage of somatic copy-number genome changed) using BioDiscovery Nexus Copy Number software (Supplemental Table S20). Supplemental Fig. S47A summarizes the correlation between Mutation rate and copy-number burden. As reported in many recent studies, samples in OV, PRAD and BRCA had high copy-number burden. In case of GBM, majority of samples had more or less equal mutation and copy-number burden. A large number of COAD samples were hypermutated and few other samples had high copy-number burden.

Supplemental Fig. S47B shows the correlation of number of HIT'nDRIVE predicted driver genes with Mutation rate. Except for a number of hypermutated samples in COAD and few highly mutated samples in BRCA, the number of driver genes predicted by HIT'nDRIVE was not correlated with the somatic mutation rate of the respective sample. In case of COAD, a large number of driver genes were identified in hypermutated samples (30 driver genes per sample in average) as compared to non-hypermutated samples (10 driver genes per sample in average). Finally, Supplemental Fig. S47C shows the correlation of number of HIT'nDRIVE predicted driver genes with copy-number

burden. Here too we observed the number of HIT'nDRIVE predicted driver genes were largely independent of the somatic copy number burden in the genome. Therefore, except for the hypermutated cases, the number of HIT'nDRIVE predicted driver genes is independent of both mutation rate and copy-number burden.

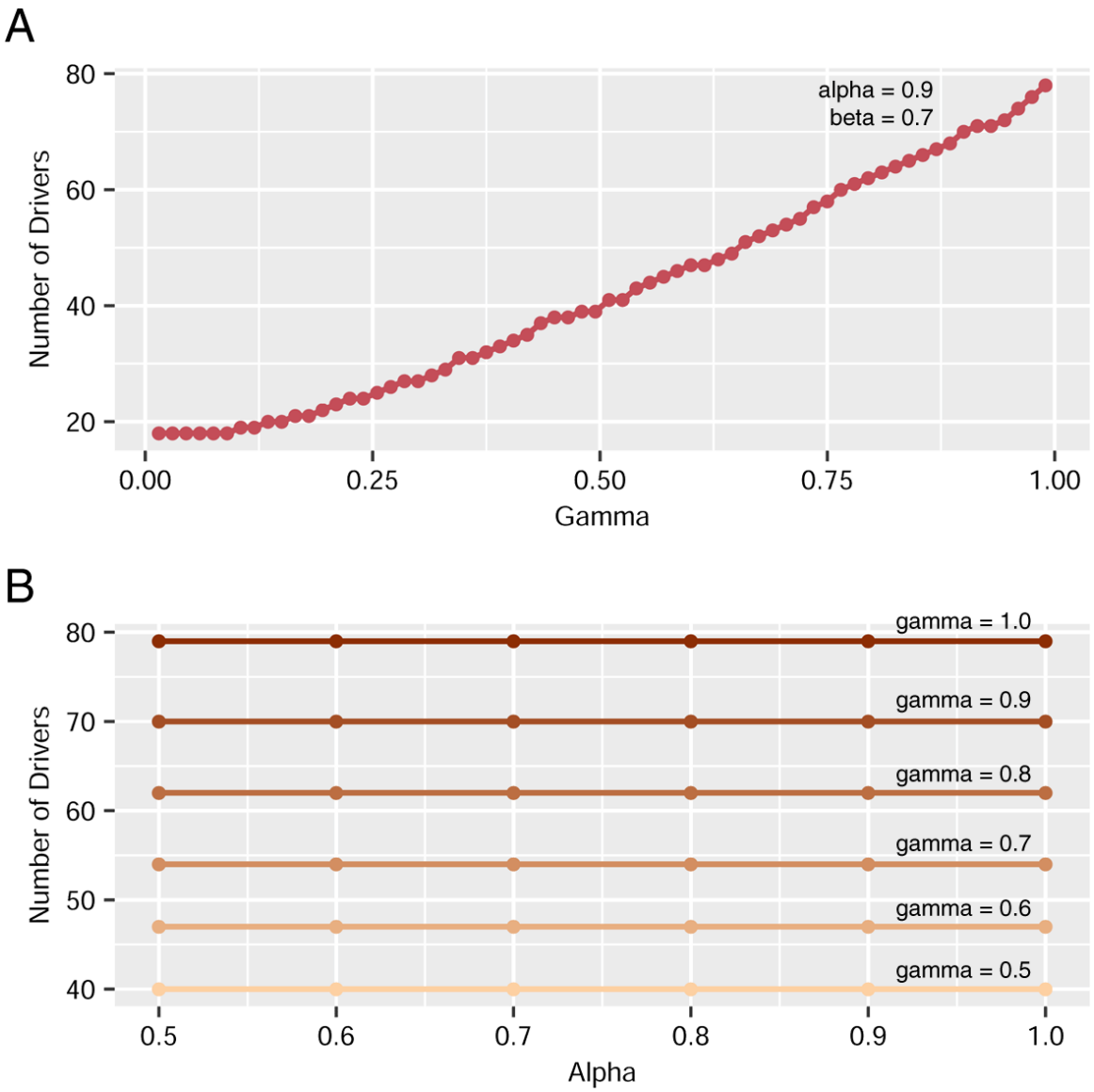
24 Cancer-stage specific driver genes of colorectal cancer.

We analyzed 429 cases of colorectal cancer (TCGA-COAD) with matched data for somatic mutation and/or copy-number aberration and RNA-seq gene-expression from TCGA. This included 78 hypermutated cases and 351 non-hypermutated cases.

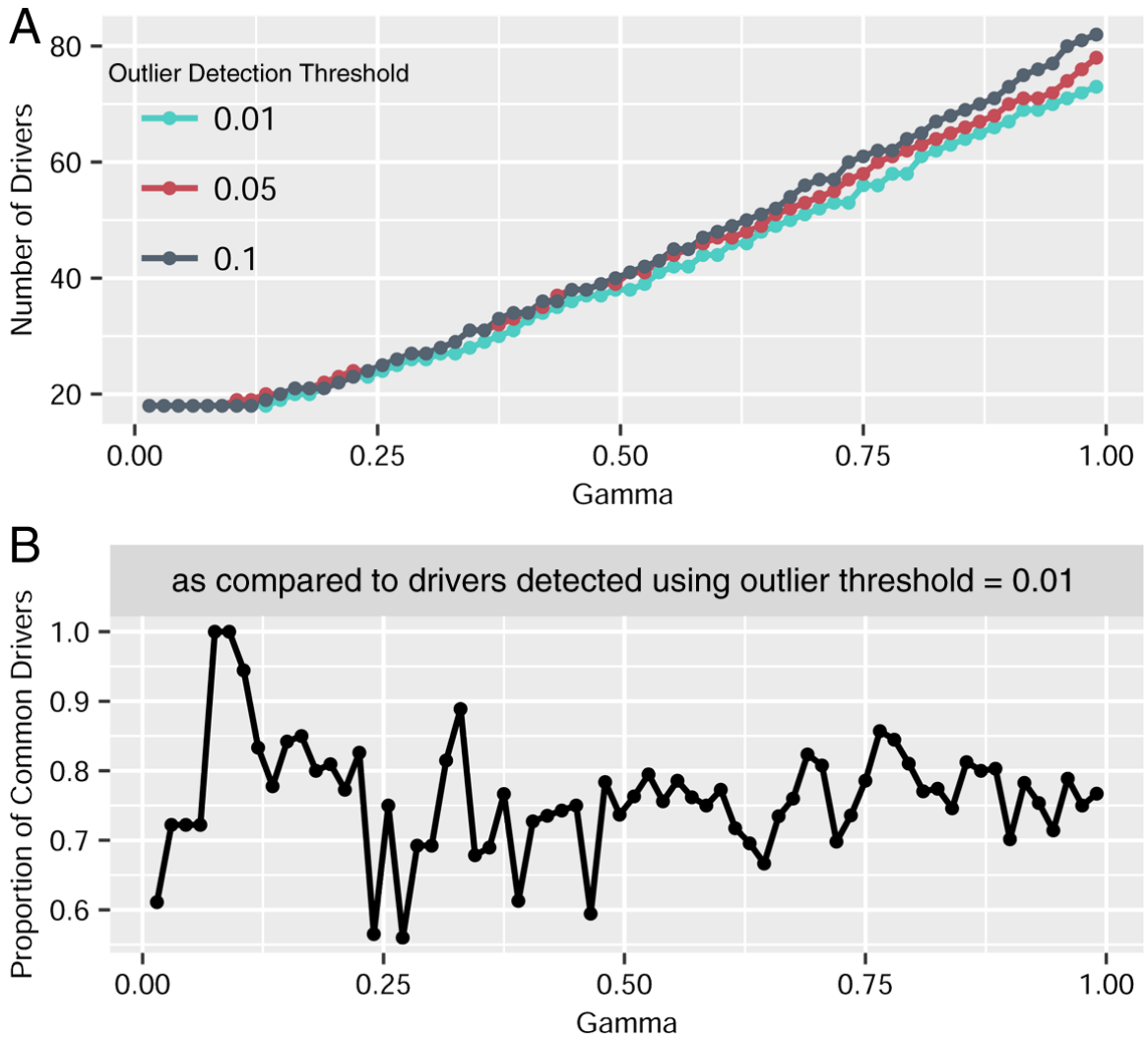
Using HIT'nDRIVE we identified driver genes for each patient sample (Supplemental Table S21) Altogether 310 unique driver genes were identified for all 429 samples analyzed. Strikingly, we found that the driver genes of hypermutated cases (140 driver genes) and non-hypermutated cases (193 driver genes) were markedly different with only 23 genes in common between the two groups. A large number of driver genes were identified in hypermutated samples (30 driver genes per sample in average) as compared to non-hypermutated samples (10 driver genes per sample in average) (Supplemental Fig. S48A).

We grouped the tumor samples based on the pathologic stages (T1, T2, T3 and T4). We focused on few known driver genes (*APC*, *TP53*, *KRAS*, *BRAF*, *SMAD4*, *MAP2K4*, *MAP3K4*, *PIK3CA*, *RNF43*) of colorectal cancer (Vogelstein et al. 2013; Dienstmann et al. 2017). Supplemental Fig. S48B summarizes the recurrent frequencies of above mentioned known driver genes of COAD broken down by different cancer stages. *APC* is known to initiate colorectal cancer and HIT'nDRIVE identified *APC* as the most recurrently altered driver gene in all four stages of tumor (indicating that it emerges in stage T1). Similarly, HIT'nDRIVE also predicted *TP53*, *MAP2K4* and *BRAF* as driver genes in all four stages of tumor, as expected. On the other hand, *KRAS*, *SMAD4* and *PIK3CA* are known driver genes of advanced stages of colorectal cancer. HIT'nDRIVE predicted *KRAS*, *SMAD4* and *PIK3CA* as driver genes only in stages T2, T3 and T4 and not in stage T1. Therefore this analysis demonstrates the capability of HIT'nDRIVE to predict stage specific driver genes of cancer.

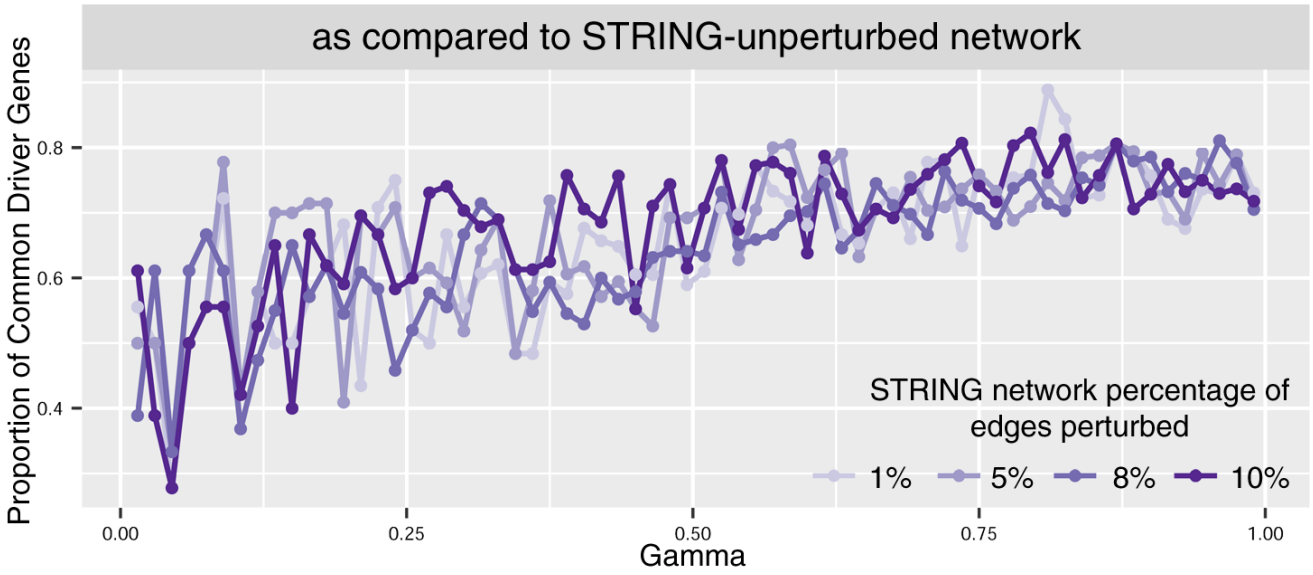
Supplemental Figures



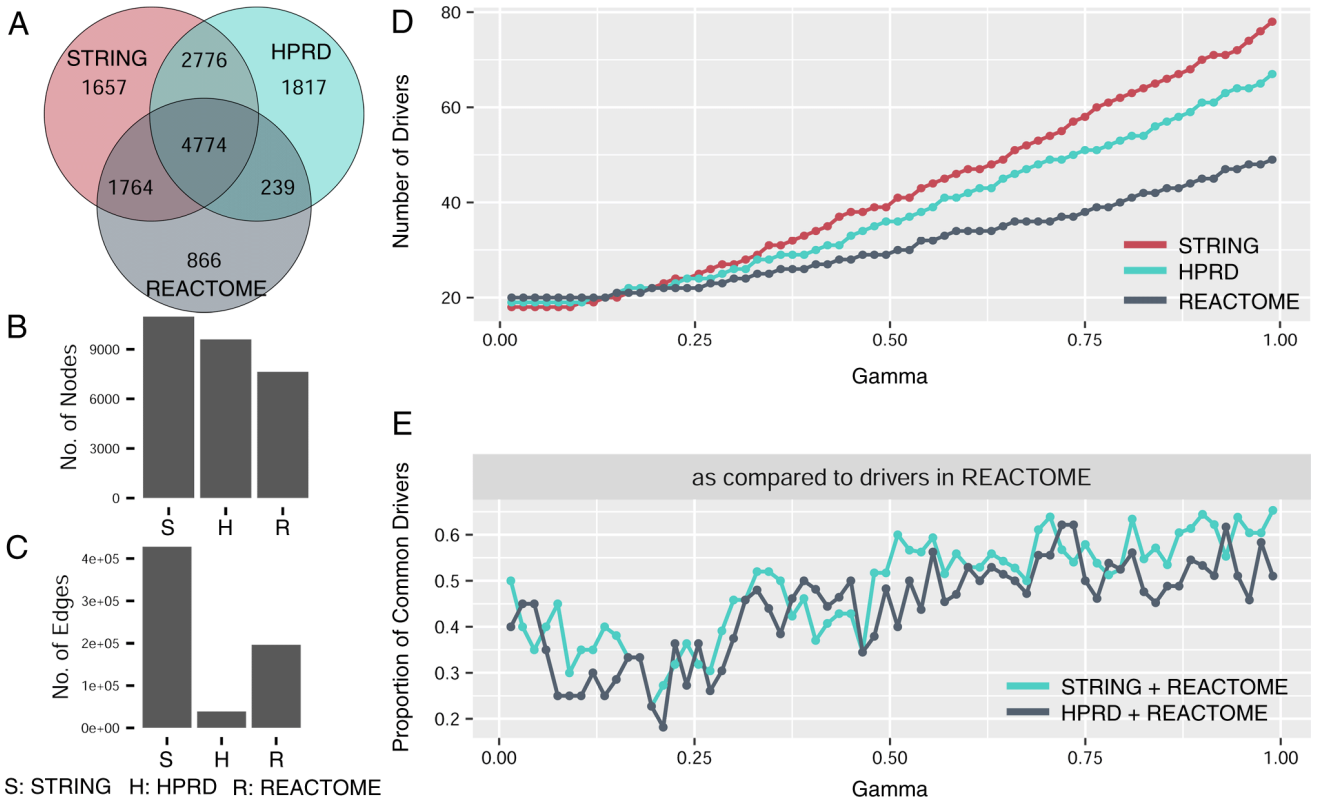
Supplemental Fig. S1. HIT'nDRIVE identified driver genes with respect to varying parameter values in 100 select BRCA samples. (See Supplemental Methods 7 for details) (A) The number of driver genes identified by HIT'nDRIVE with respect to varying values of γ, the fractional lower bound on the sum of the incoming edge weights from selected driver genes to each expression altered gene covered. As expected HIT'nDRIVE is sensitive to changes in γ. (B) The number of driver genes identified by HIT'nDRIVE with respect to varying values of α, the fraction of outlier genes to be covered across all patients. HIT'nDRIVE is highly robust with respect to the changes in α. (HIT'nDRIVE is also very robust to variation in β, which is the fraction of outliers to be covered for each patient).



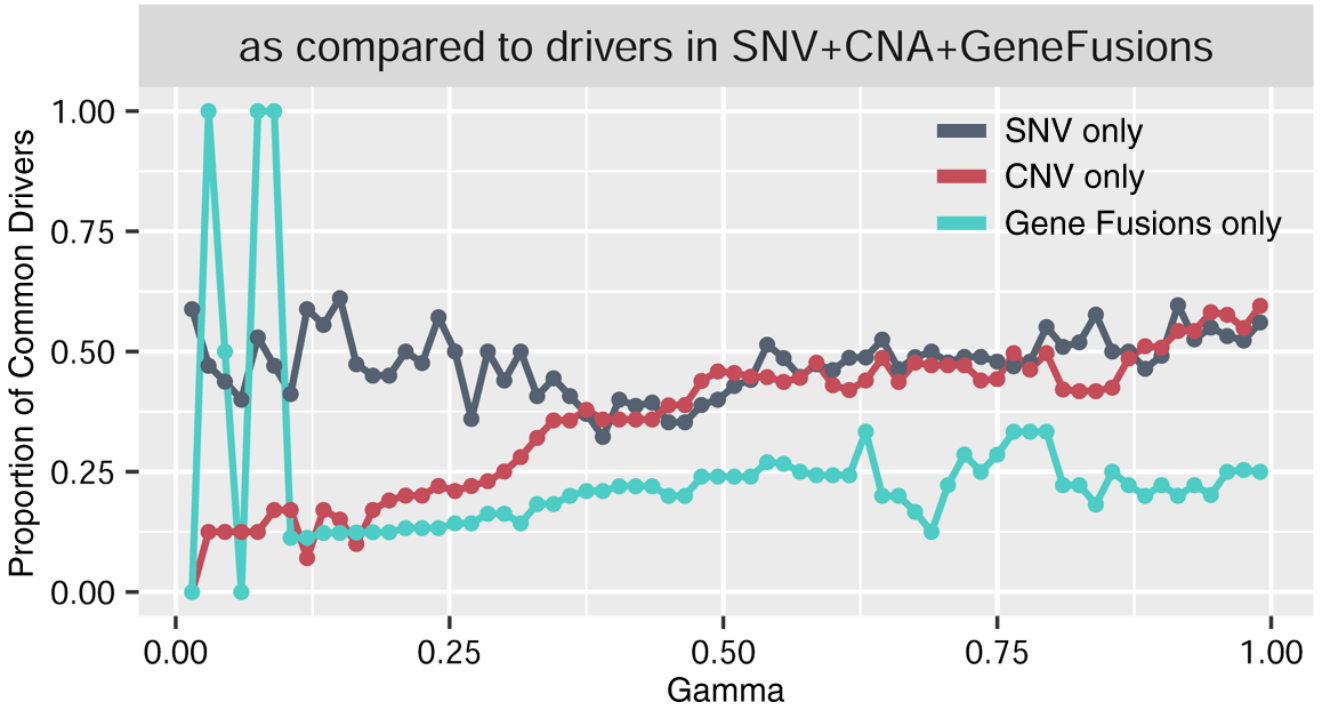
Supplemental Fig. S2. HIT'nDRIVE identified driver genes with respect to varying outlier stringency across 100 select BRCA samples. (see Supplemental Methods 7 for details.) (A) The number of driver genes identified by HIT'nDRIVE with respect to three outlier detection threshold values, across varying values of the γ parameter (see Supplemental Results 5 for details). An increase in the outlier detection threshold implies a slight decrease in the number of detected driver genes. (B) Proportion of HIT'nDRIVE detected driver genes obtained for outlier threshold of 0.01 which are also detected when the outlier threshold is 0.05 and 0.1. As can be seen, even though the number of driver genes increase with the value of γ, the proportion of driver genes jointly detected for three threshold values for outlier detection are very robust, roughly at 80%. More importantly, the increase in outlier detection threshold only decreases the number of driver genes and does not introduce new driver genes, implying robustness.



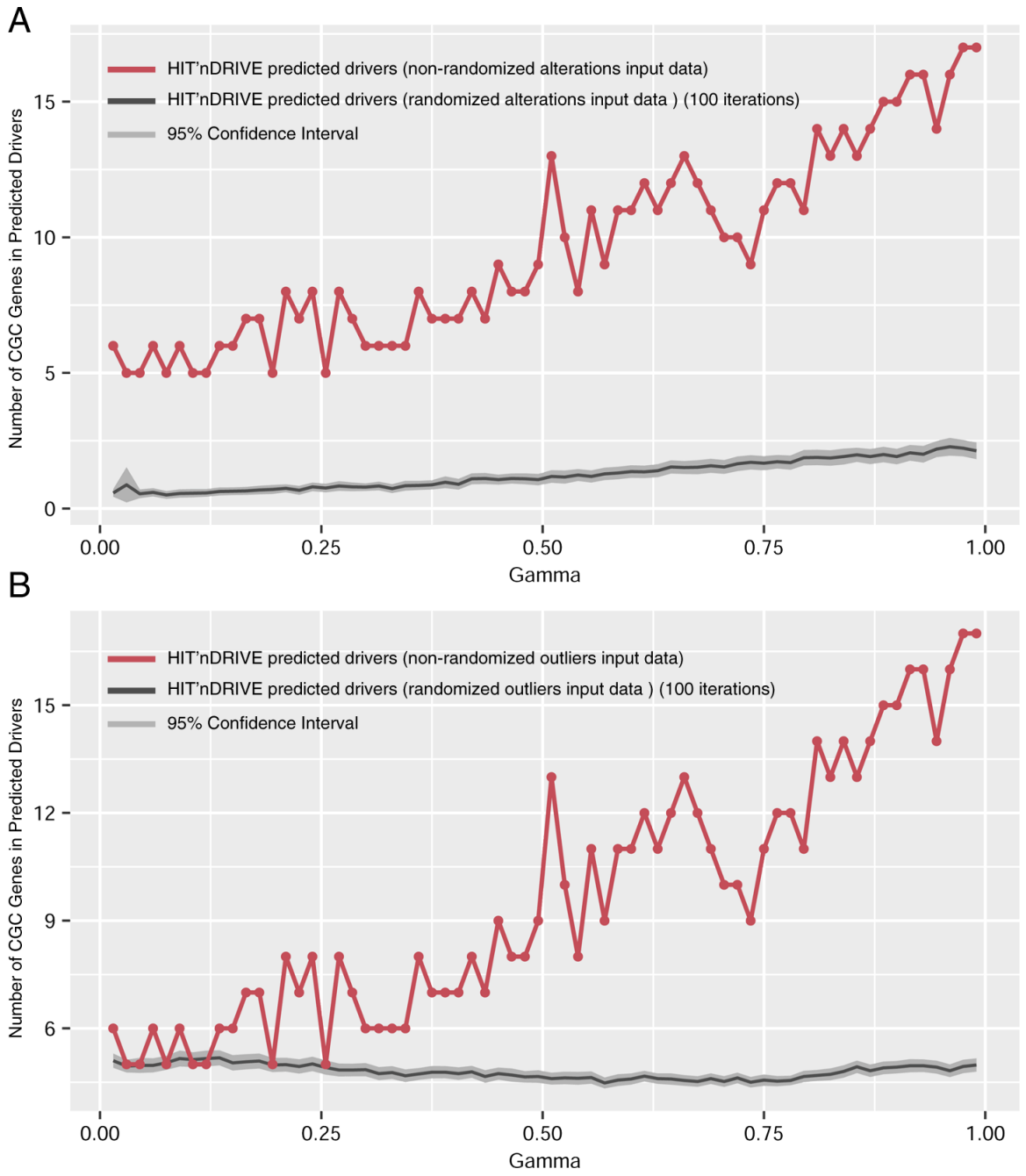
Supplemental Fig. S3. HIT'nDRIVE identified driver genes with respect to network perturbation in 100 select BRCA samples. The edges of the STRING ver-10 network was perturbed to different extent (between 1-10%) preserving the degree of the nodes in the network. HIT'nDRIVE simulation was performed using different perturbed networks. Proportion of common driver genes between the unperturbed network and each of the perturbed network were calculated.



Supplemental Fig. S4. HIT'nDRIVE identified driver genes with respect to underlying network used in 100 select BRCA samples. (A) Venn Diagram showing the overlap of nodes in the three different networks used - STRING v10 (only high-confident interactions), HPRD v9.0 and REACTOME v2015. (B) Comparison between the number of nodes in the network. (C) Comparison between the number of edges in the network. (D) Comparison between the number of driver genes detected using different networks. (E) Proportion of common driver genes between the networks (STRING-REACTOME and HPRD-REACTOME) as compared to driver genes detected using REACTOME network. A subset of 100 BRCA samples from TCGA were used for the simulation (See Supplemental Methods 7 for details).



Supplemental Fig. S5. HIT'nDRIVE identified driver genes with respect to different alteration types in 100 select BRCA samples. HIT'nDRIVE simulation was performed using different alteration types - SNV only, CNA only, Gene Fusions only and combination of SNV + CNA + Gene Fusions. Intersection of driver genes resulting when using individual alteration types alone and the driver genes resulting when using combination of SNV + CNA + Gene Fusions were calculated. Proportion of the intersection of driver genes as compared to the driver genes detected using combination of SNV + CNA + Gene Fusions were plotted. A subset of 100 BRCA samples from TCGA were used for the simulation (See Supplemental Methods 7 for details).



Supplemental Fig. S6. HIT'nDRIVE identified driver genes using randomized input data in 100 select BRCA samples. Driver genes predicted by HIT'nDRIVE in non-randomized data compared with the driver genes predicted using randomized (i.e. by gene label swapping for 100 iterations). (A) Randomized altered genes and (B) Randomized outlier genes. The Cancer Gene Census (CGC) genes present in the predicted driver geneset is displayed in the plot. A subset of 100 BRCA samples from TCGA were used for the simulation (See Supplemental Methods 7 for details).

A

Modified ILP formulation

$\min_{x_1,..,x_{|\mathscr{X}|}} \sum_i x_i$

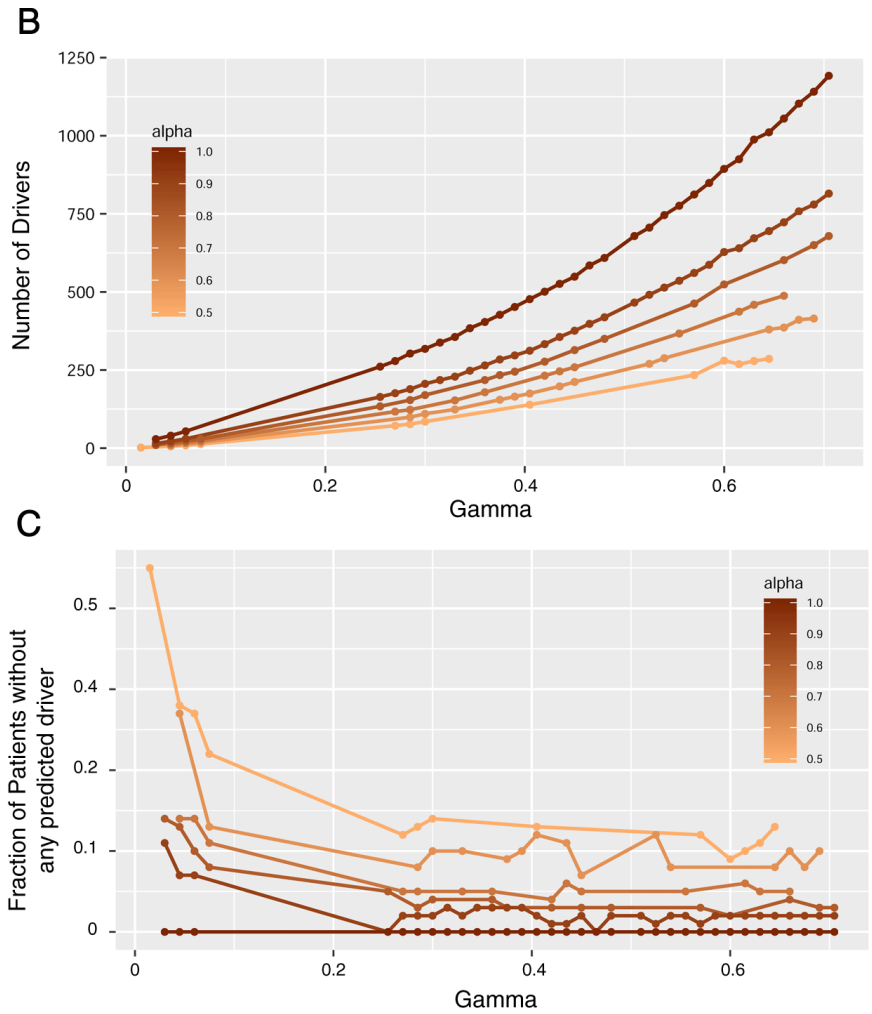
s.t.

$\forall i,j : x_i = e_{ij}$

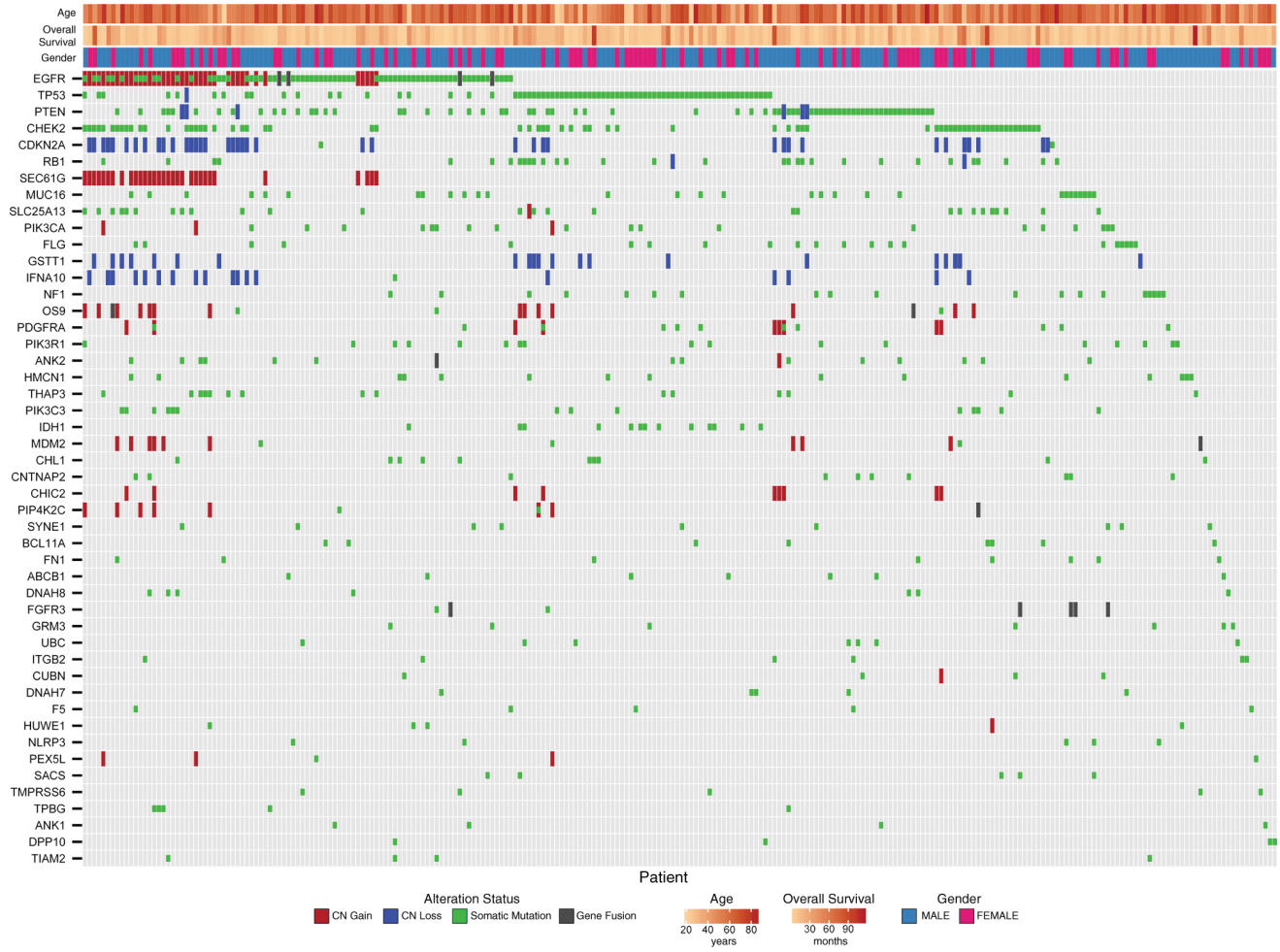
$\forall j : \sum_i e_{ij} w_{ij} \geq y_j \gamma \sum_i w_{ij}$

$\sum_j y_j \geq \alpha |\mathscr{Y}|$

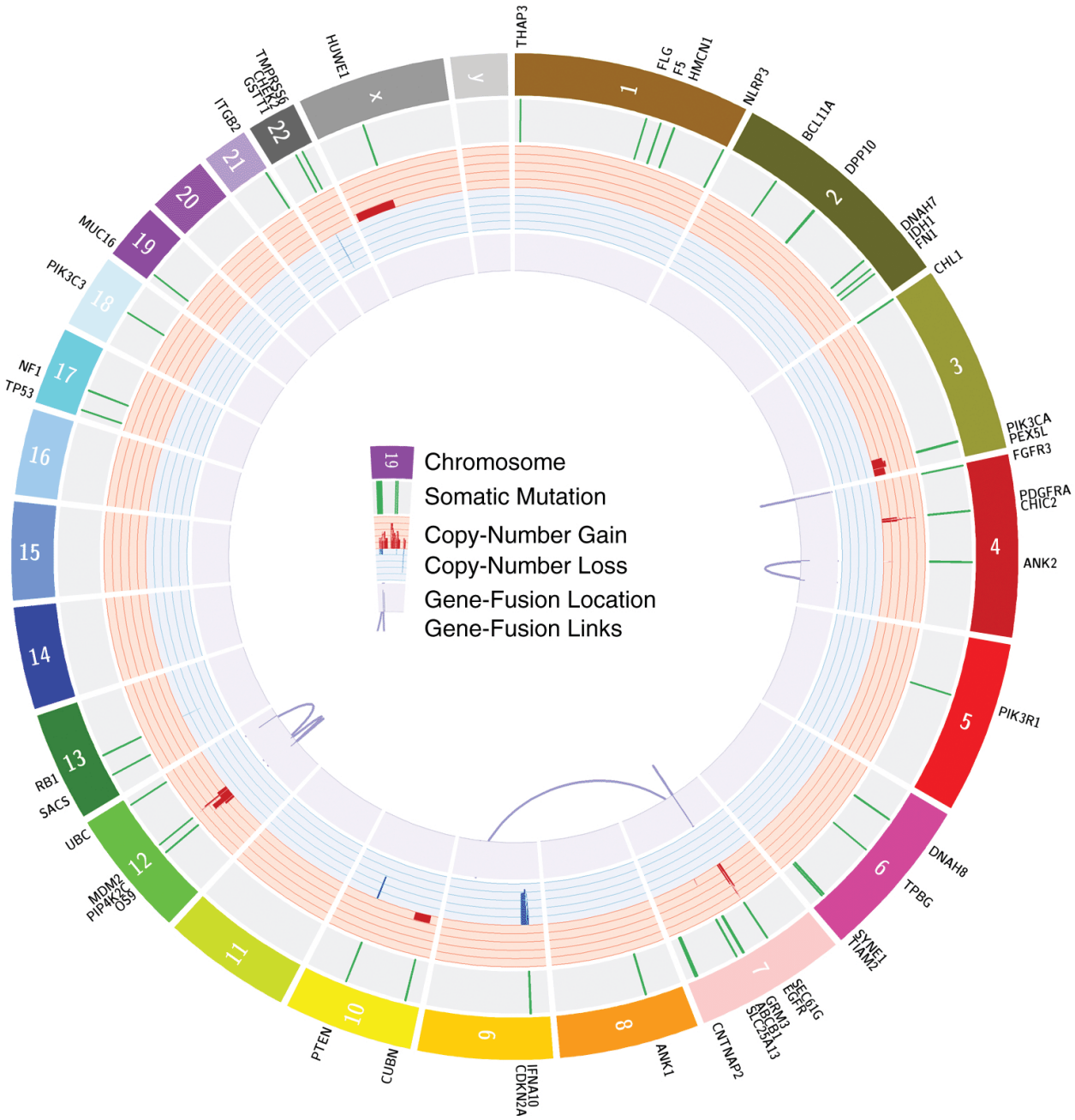
$x_i, e_{ij}, y_j \in \{0,1\}$



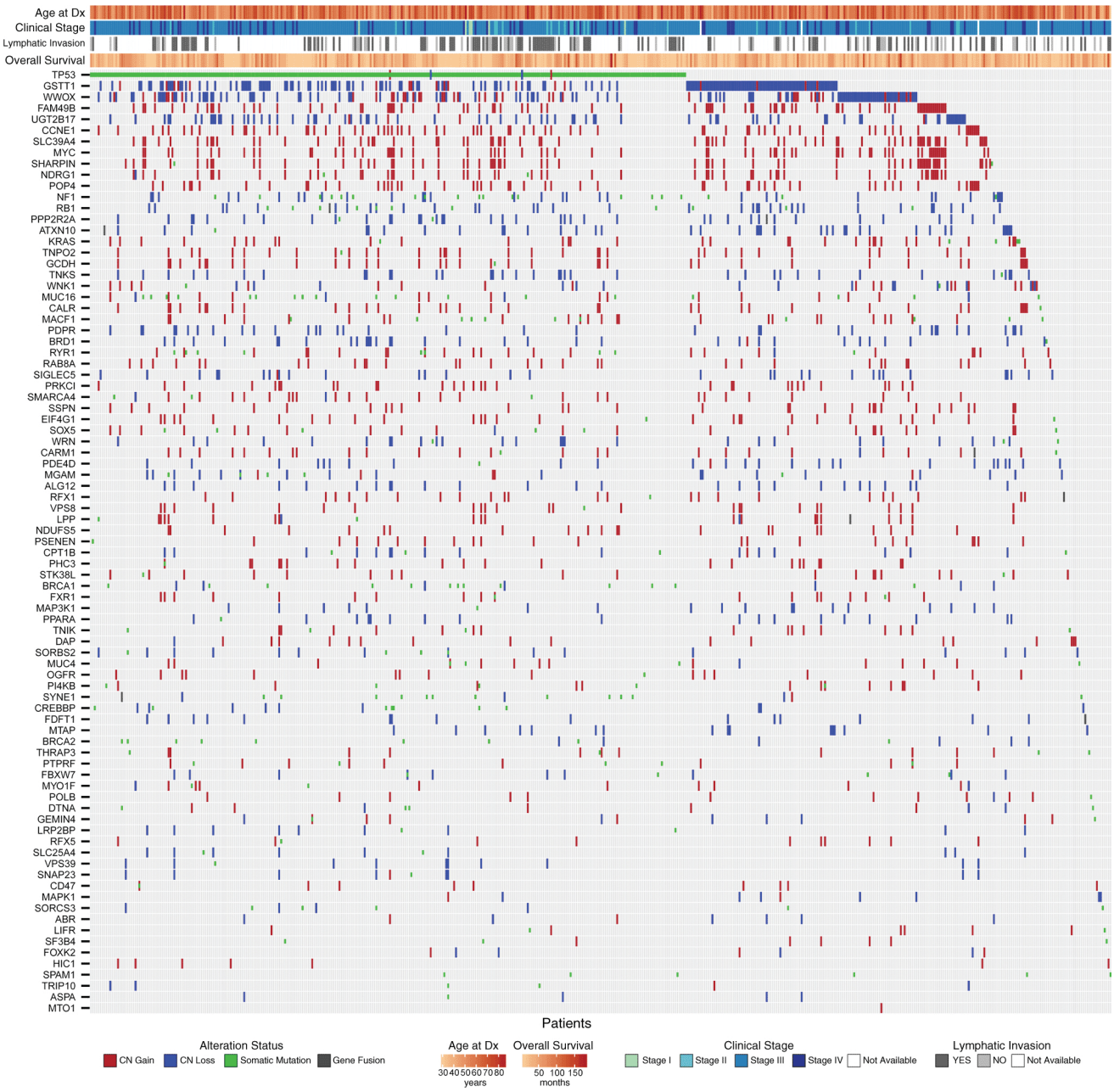
Supplemental Fig. S7. Modified HIT'nDRIVE not required to prioritize at least one driver gene per patient. (A) Modified ILP formulation where we removed the constraint that ensured at least one driver gene is prioritized per patient. (B) HIT'nDRIVE simulation with different values of gamma (γ) parameter with the modified ILP formulation as given in A. Each line represents different values of alpha (α) parameter, which controls the fraction of total outliers to be covered. (C) We calculated the fraction of patients with no driver genes prioritized, for the same set of driver genes prioritized in B.



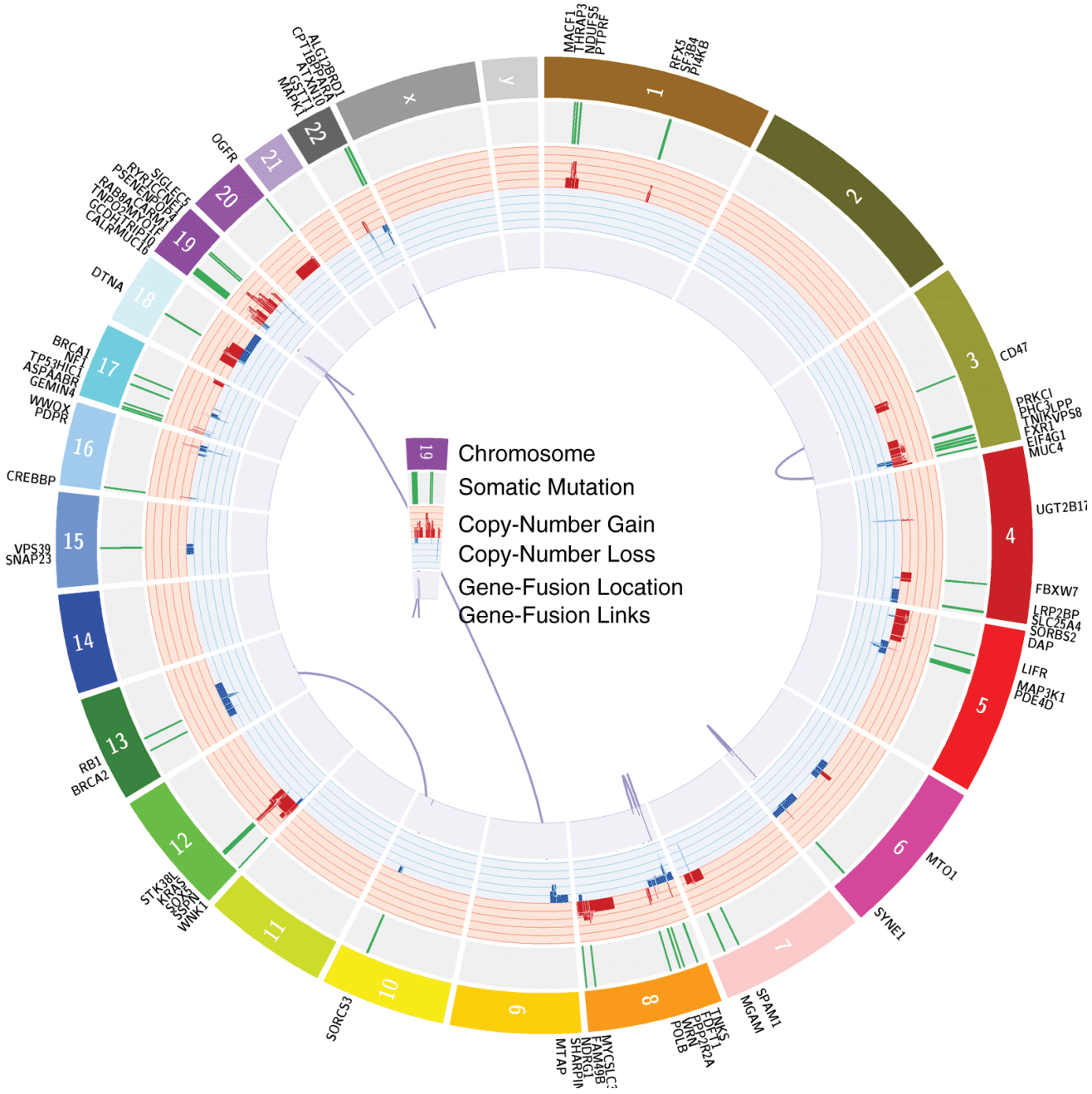
Supplemental Fig. S8. Genomic drivers of Glioblastoma. The spectrum of driver alterations (somatic mutations, CN amplifications, homozygous CN deletions and gene fusions) in GBM patients prioritized by HIT'nDRIVE. Alteration frequency of the genes is shown on the right and frequency of driver genes predicted per patient is shown on the top panel.



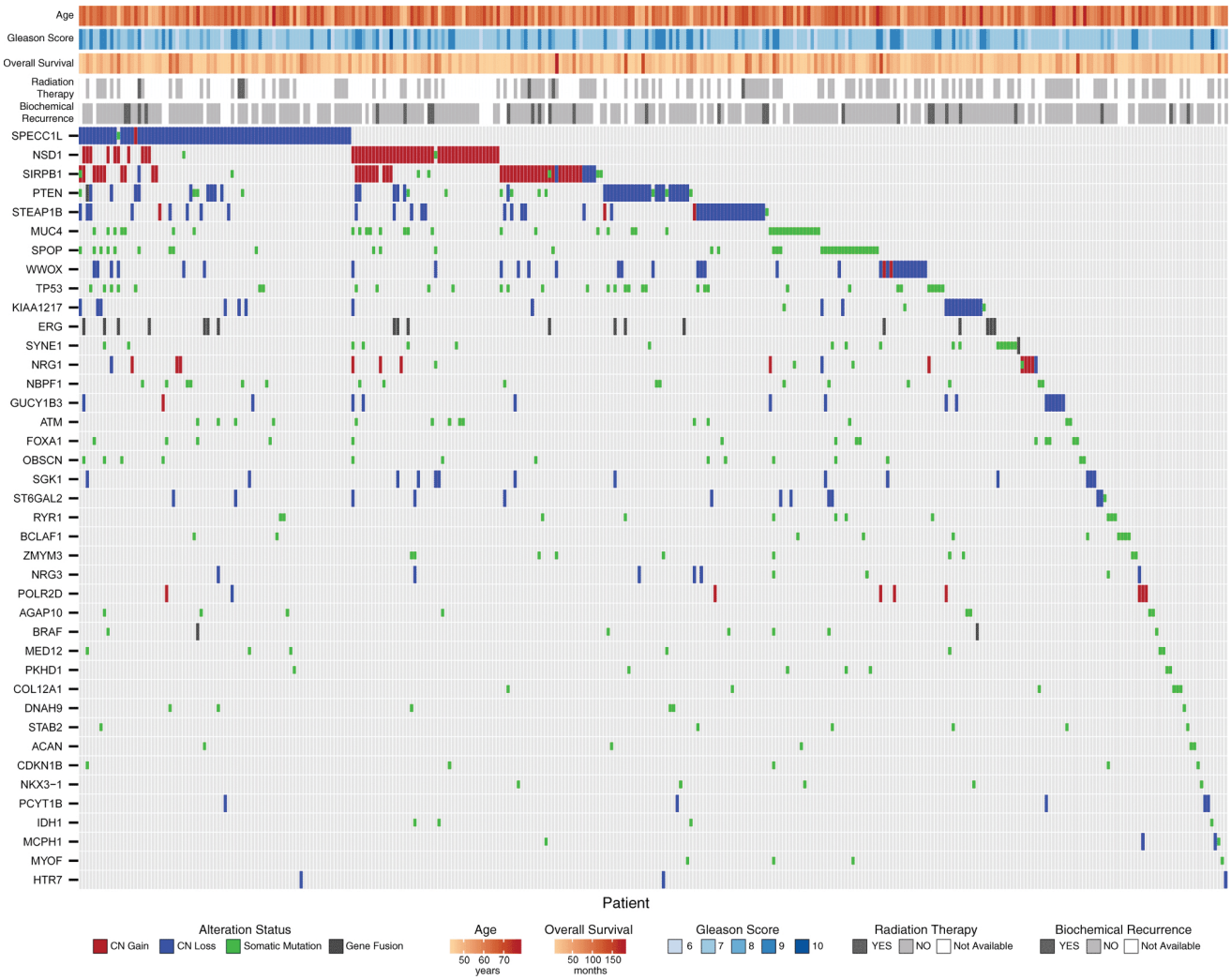
Supplemental Fig. S9. Genomic drivers of Glioblastoma. Circos plot showing the genomic position of the driver genes prioritized by HIT'nDRIVE in GBM.



Supplemental Fig. S10. Genomic drivers of Ovarian Cancer. The spectrum of driver alterations (somatic mutations, CN amplifications, homozygous CN deletions and gene fusions) in OV patients prioritized by HIT'nDRIVE. Alteration frequency of the genes is shown on the right and frequency of driver genes predicted per patient is shown on the top panel.

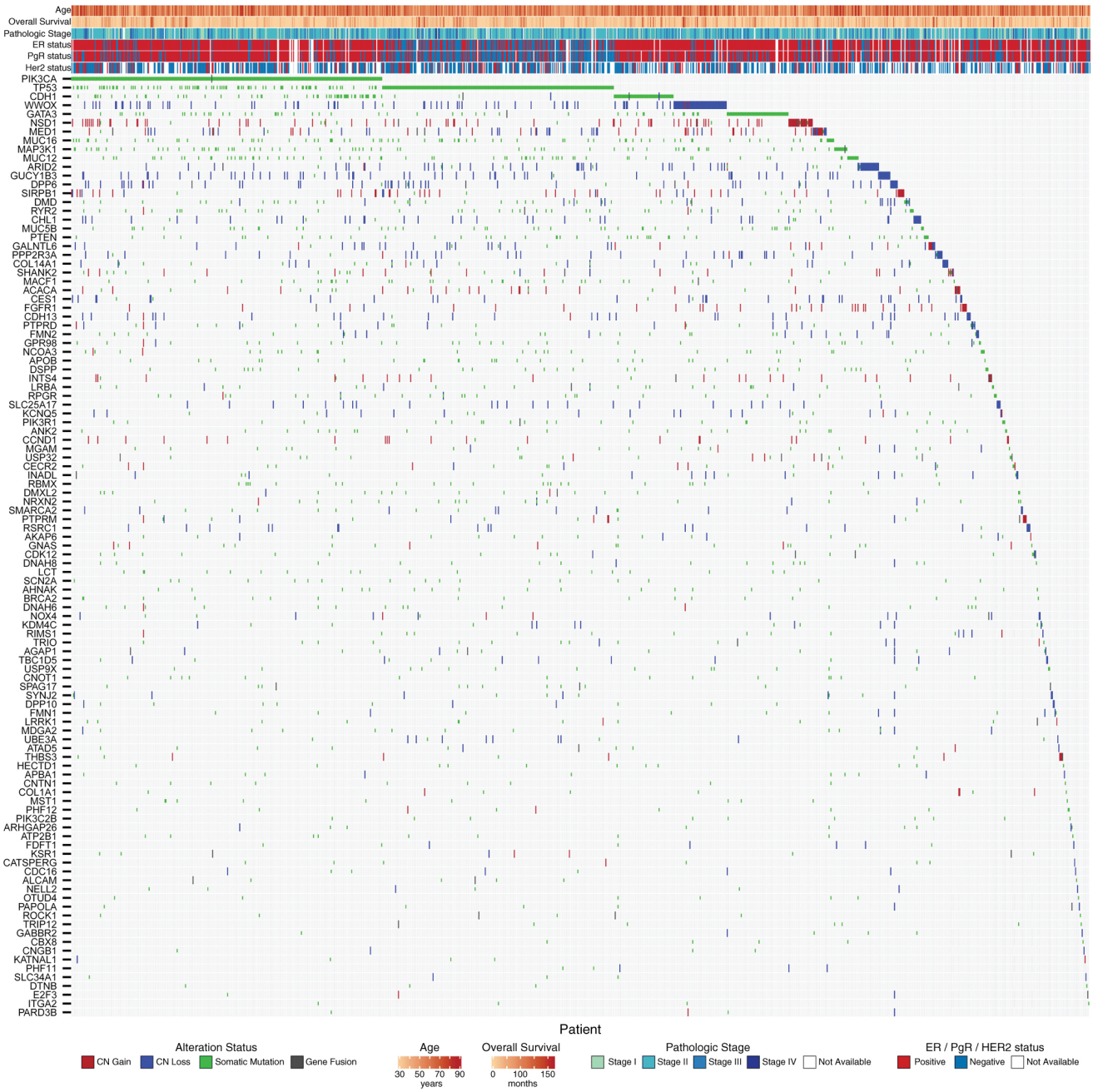


Supplemental Fig. S11. Genomic drivers of Ovarian Cancer. Circos plot showing the genomic position of the driver genes prioritized by HIT'nDRIVE in OV.

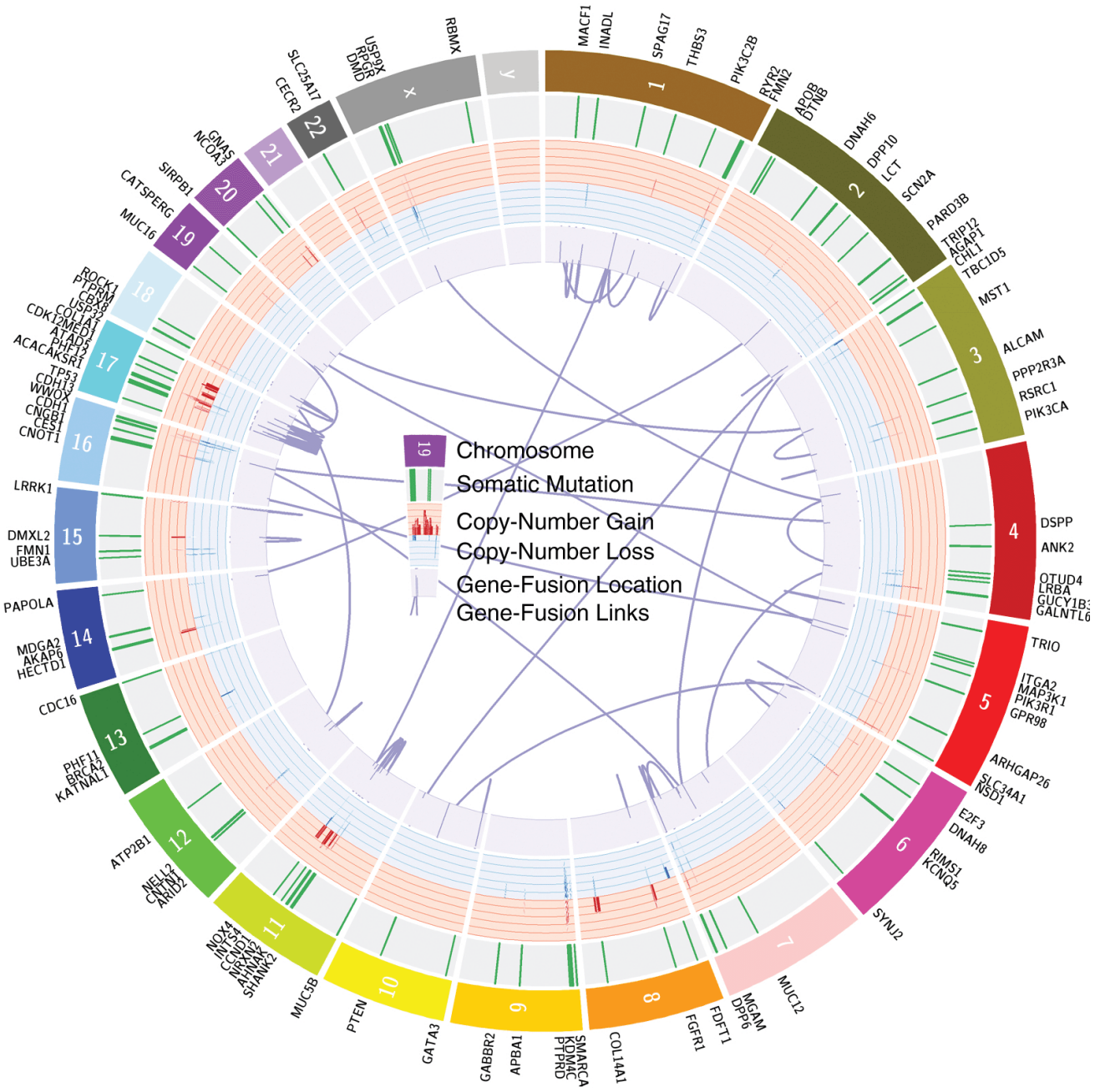


Supplemental Fig. S12. Genomic drivers of Prostate Cancer. The spectrum of driver alterations (somatic mutations, CN amplifications, homozygous CN deletions and gene fusions) in PRAD patients prioritized by HIT'nDRIVE. Alteration frequency of the genes is shown on the right and frequency of driver genes predicted per patient is shown on the top panel.

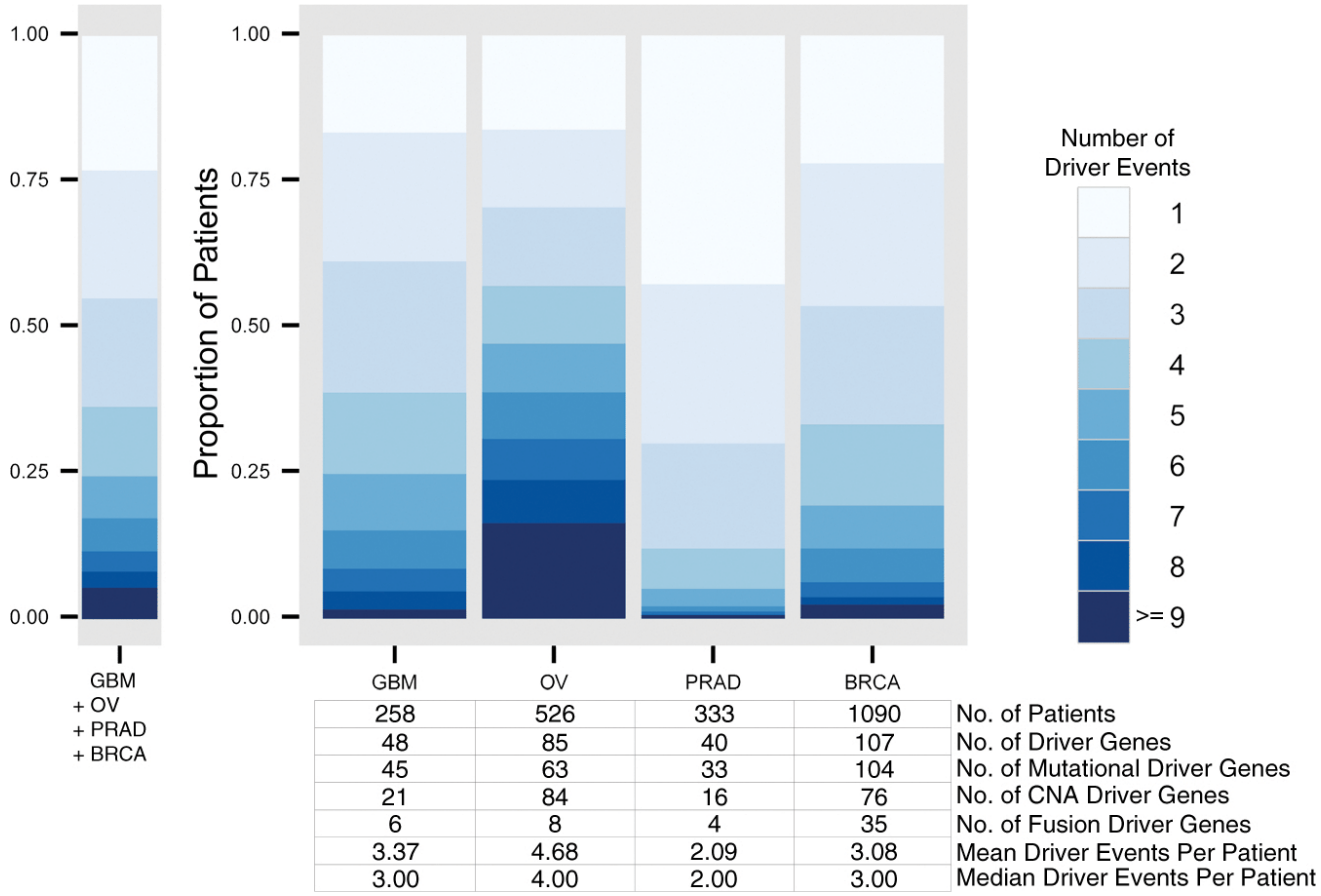
Supplemental Fig. S13. Genomic drivers of Prostate Cancer. Circos plot showing the genomic position of the driver genes prioritized by HIT'nDRIVE in PRAD.



Supplemental Fig. S14. Genomic drivers of Breast Cancer. The spectrum of driver alterations (somatic mutations, CN amplifications, homozygous CN deletions and gene fusions) in BRCA patients prioritized by HIT'nDRIVE. Alteration frequency of the genes is shown on the right and frequency of driver genes predicted per patient is shown on the top panel.

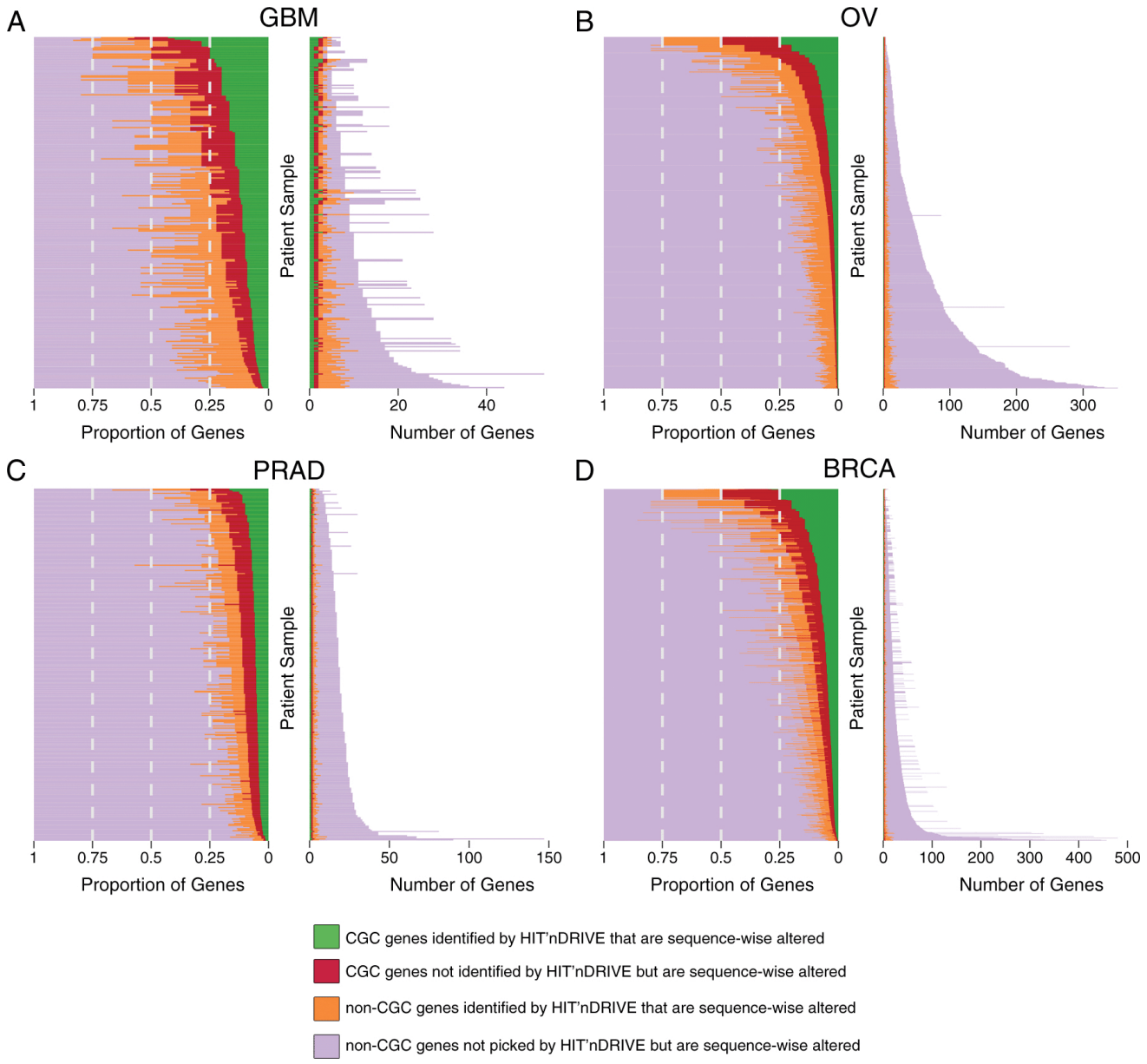


Supplemental Fig. S15. Genomic drivers of Breast Cancer. Circos plot showing the genomic position of the driver genes prioritized by HIT'nDRIVE in BRCA.

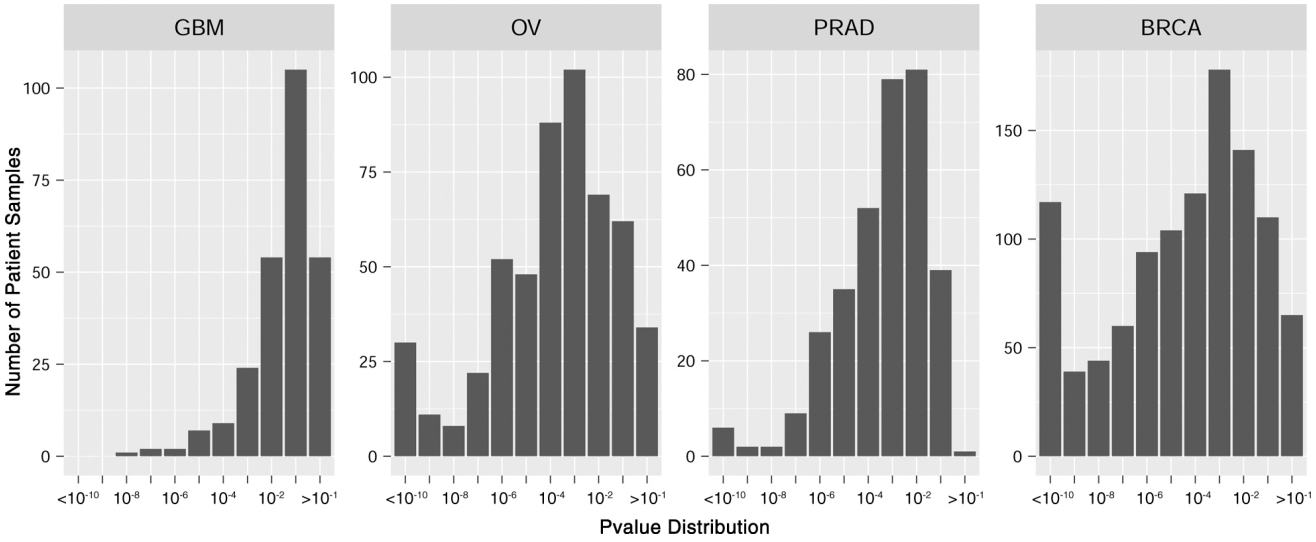


GBM	OV	PRAD	BRCA	
258	526	333	1090	No. of Patients
48	85	40	107	No. of Driver Genes
45	63	33	104	No. of Mutational Driver Genes
21	84	16	76	No. of CNA Driver Genes
6	8	4	35	No. of Fusion Driver Genes
3.37	4.68	2.09	3.08	Mean Driver Events Per Patient
3.00	4.00	2.00	3.00	Median Driver Events Per Patient

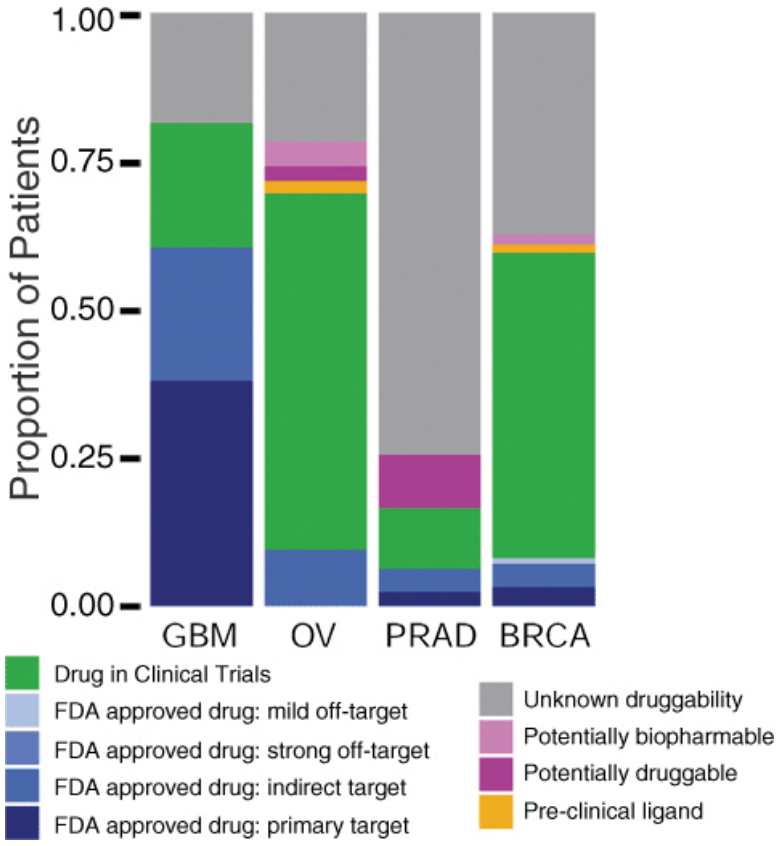
Supplemental Fig. S16. Driver genes distribution across cancer types. The distribution of number of driver genes identified by HIT'nDRIVE in individual cancer cohort. The left most panel shows the distribution of driver genes in a combined cohort of GBM, OV, PRAD and BRCA patients.



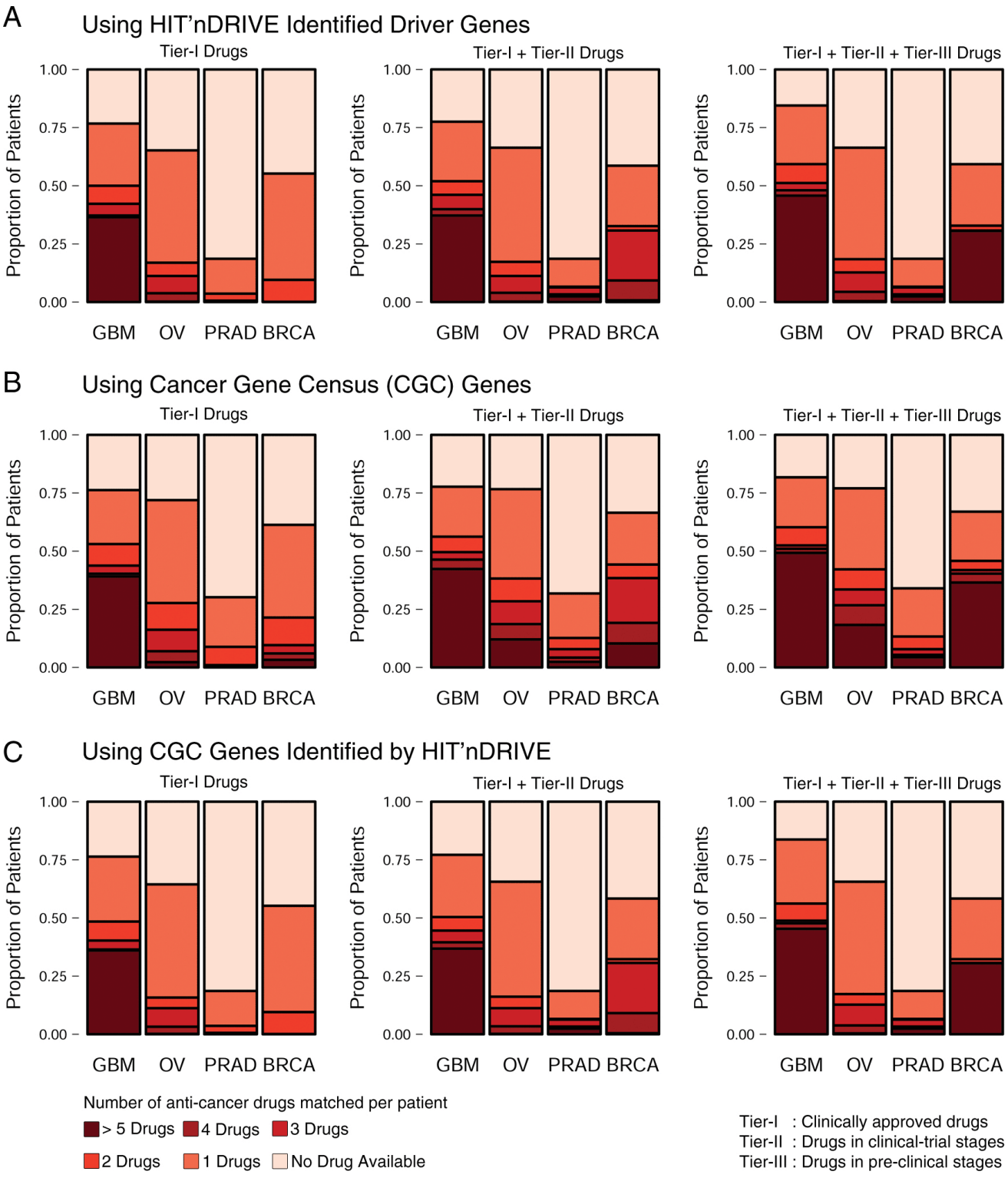
Supplemental Fig. S17. Sequence-wise altered Cancer Gene Census (CGC) genes prioritized by HIT'nDRIVE v.s. that of non-CGC genes, for each patient sample, across four cancer types. Only CGC genes specific to a cancer type is considered here. Green: Cancer specific sequence-wise altered CGC genes prioritized by HIT'nDRIVE; Red: Cancer specific sequence-wise altered CGC genes NOT-prioritized by HIT'nDRIVE; Orange: Sequence-wise altered non-CGC genes prioritized by HIT'nDRIVE; Purple: Sequence-wise altered non-CGC genes NOT-prioritized by HIT'nDRIVE. The right panel depicts absolute numbers and the left panel depicts relative proportions. As can be seen the likelihood of a sequence-wise altered CGC gene to be prioritized by HIT'nDRIVE is much higher than that of a non-CGC gene.



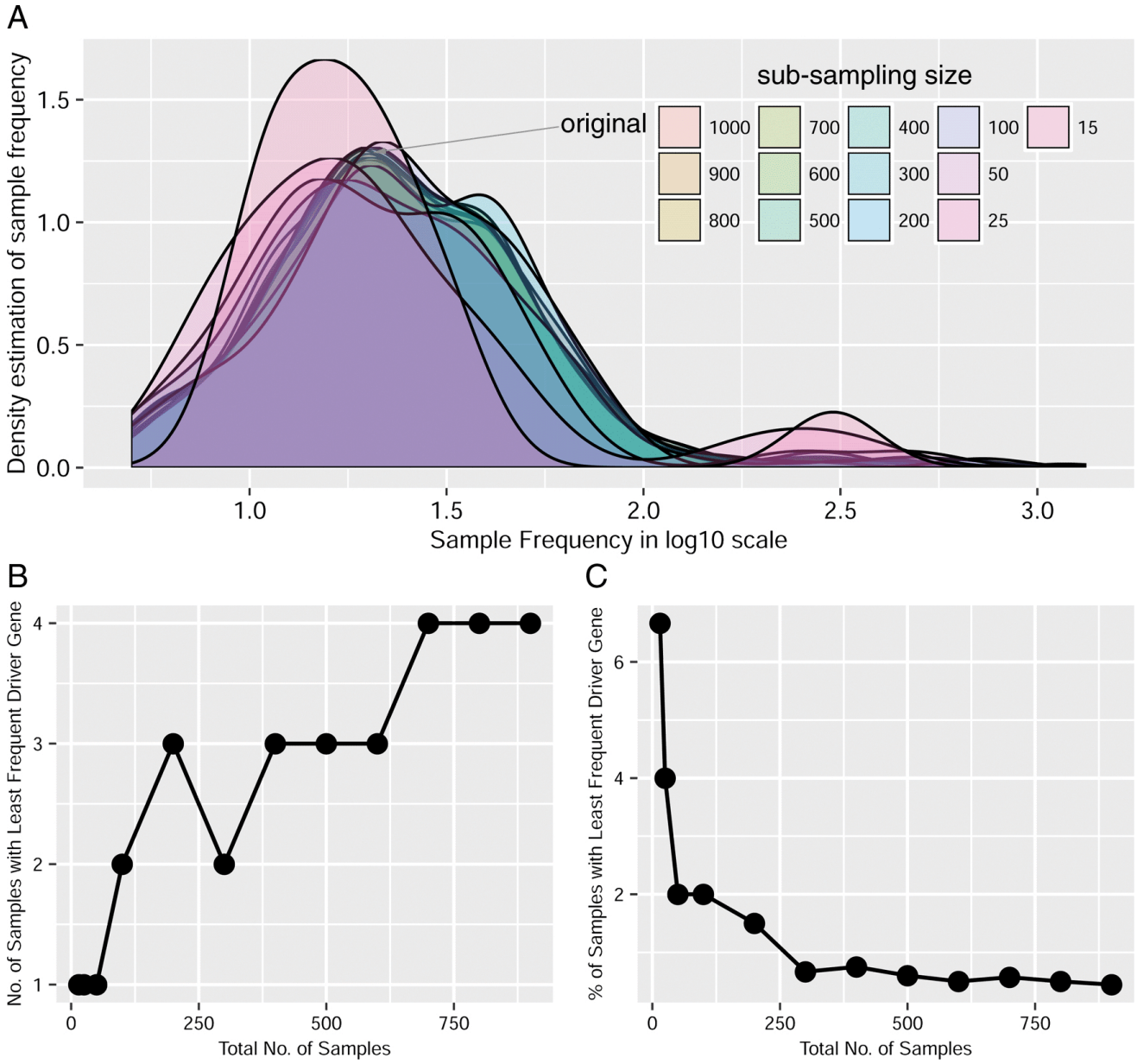
Supplemental Fig. S18. **P-value Distribution of the likelihood of HIT'nDRIVE to pick CGC genes**



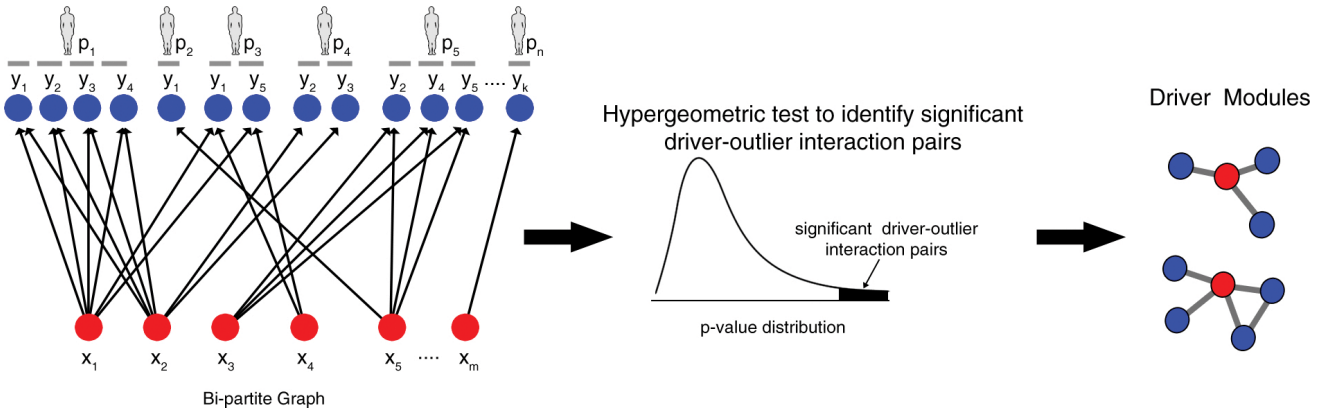
Supplemental Fig. S19. Distribution of patient druggability. Distribution of patients with HIT'nDRIVE predicted driver genes that are target of drugs in different levels of development. The druggability data were obtained as published by (Rubio-Perez et al. 2015).



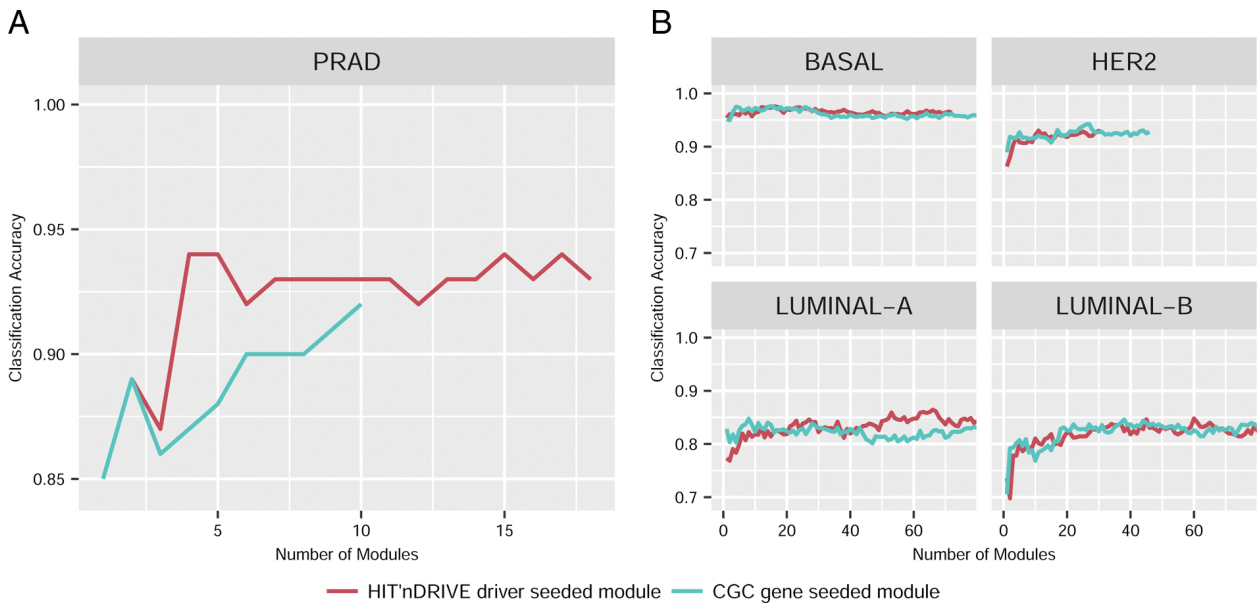
Supplemental Fig. S20. Anti-cancer drugs targeting driver genes predicted by HIT'nDRIVE. We considered the potential driver genes, (A) either predicted by HIT'nDRIVE, (B) or the sequencewise altered CGC genes, or (C) the intersection of CGC and HIT'nDRIVE predicted driver genes for each patient. The figure shows the proportion of patients in which HIT'nDRIVE identified drivers were targets of anti-cancer drugs. Drug-target information were obtained from (Iorio et al. 2016) .



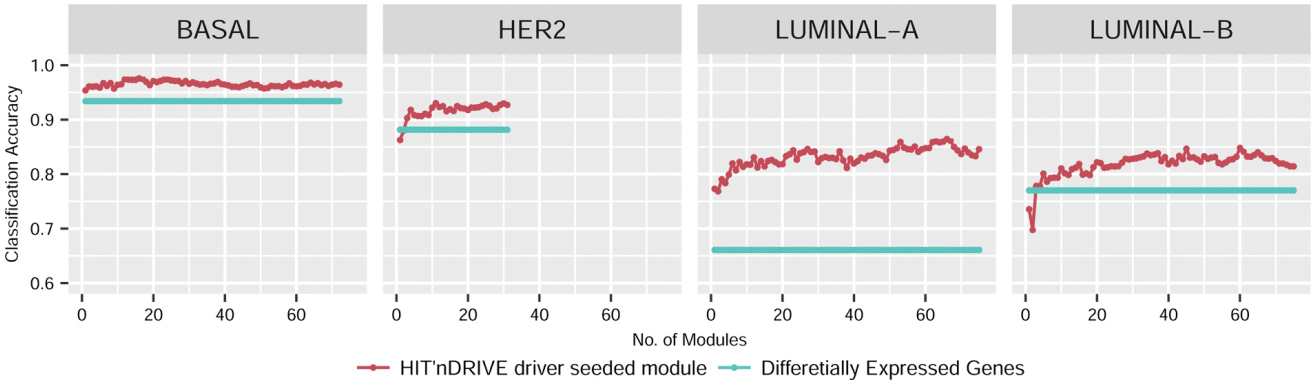
Supplemental Fig. S21. HIT'nDRIVE sensitivity to least frequent driver gene. (A) TCGA-BRCA cohot with 1000 tumors were selected for randomization experiment (labelled here as "original"). Tumor samples of different sample size were sub-sampled such that the sample alteration frequency in the sub-sample population is very similar to that of the original 1000 tumors (see supplementary text for details). HIT'nDRIVE simulation was performed in all of the above sub-sampled tumor populations. The least frequent driver gene was identified for each sub-sampled population. (B-C) Number of samples (and percentage of sub-sampled tumor population) in which the least frequent driver gene is present.



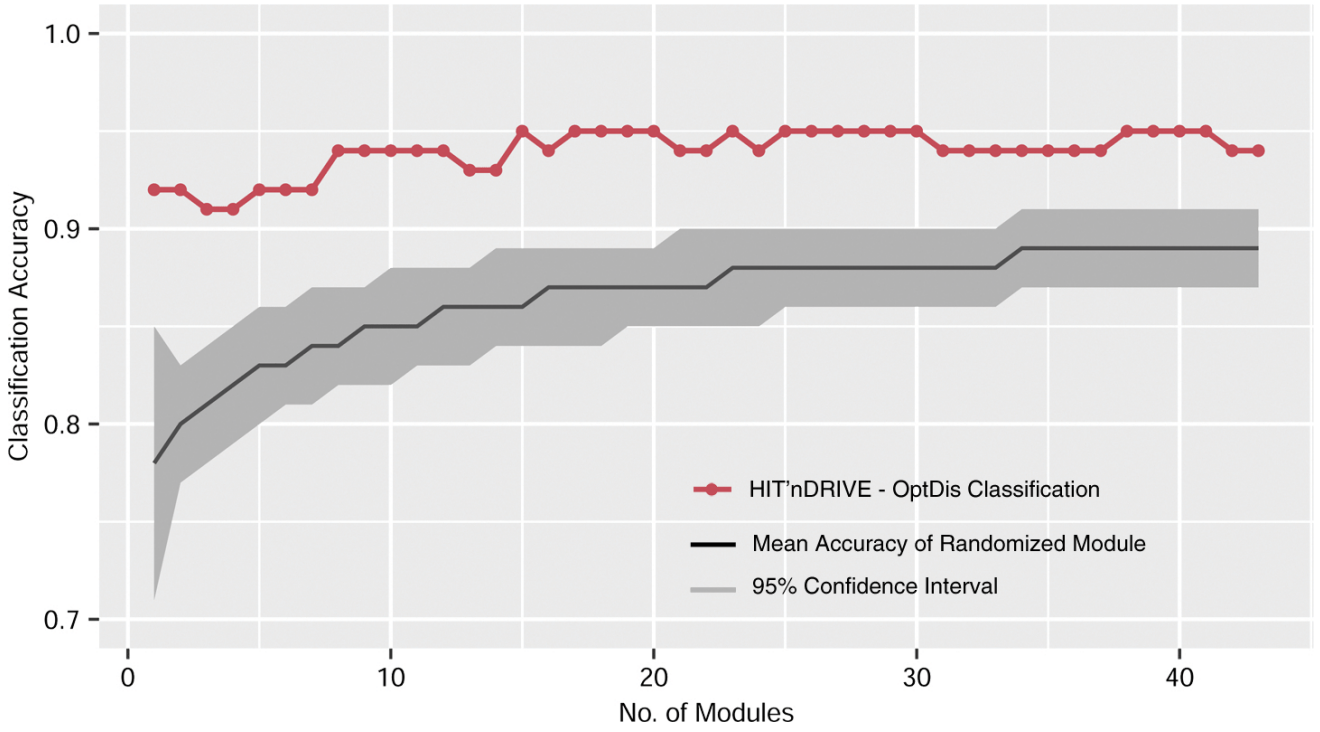
Supplemental Fig. S22. Schematic Diagram of HIT'nDRIVE-unsupervised approach to prioritize driver-modules. Driver-outlier interaction pairs are identified from the bipartite graph. We perform a hypergeometric test to identify significant driver-outlier interaction pairs across the patient cohort (pvalue < 0.001). Each driver-module is seeded with one HIT'nDRIVE identified driver gene, and includes outlier genes with significant driver-outlier interaction pairs.



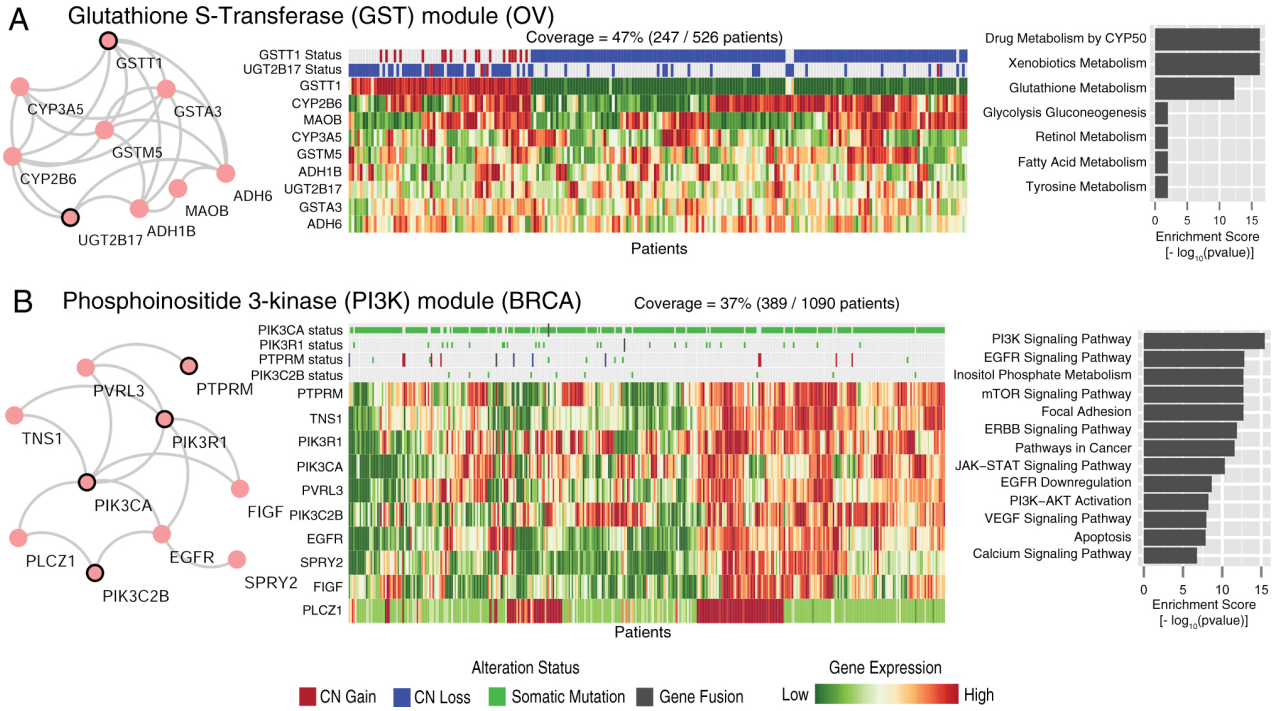
Supplemental Fig. S23. Phenotype Classification using CGC Genes Seeded Modules. Phenotype Classification accuracy of HIT'nDRIVE driver seeded module vs Cancer Gene Census (CGC) genes seeded modules. (A) TCGA-PRAD gene-expression dataset with Tumor and Normal samples. (B) Subtype classification accuracy of HIT'nDRIVE identified driver seeded modules vs CGC BRCA driver seeded modules on TCGA-BRCA cohort with respect to four subtypes of breast cancer (Basal, Her2, Luminal-A and Luminal-B).



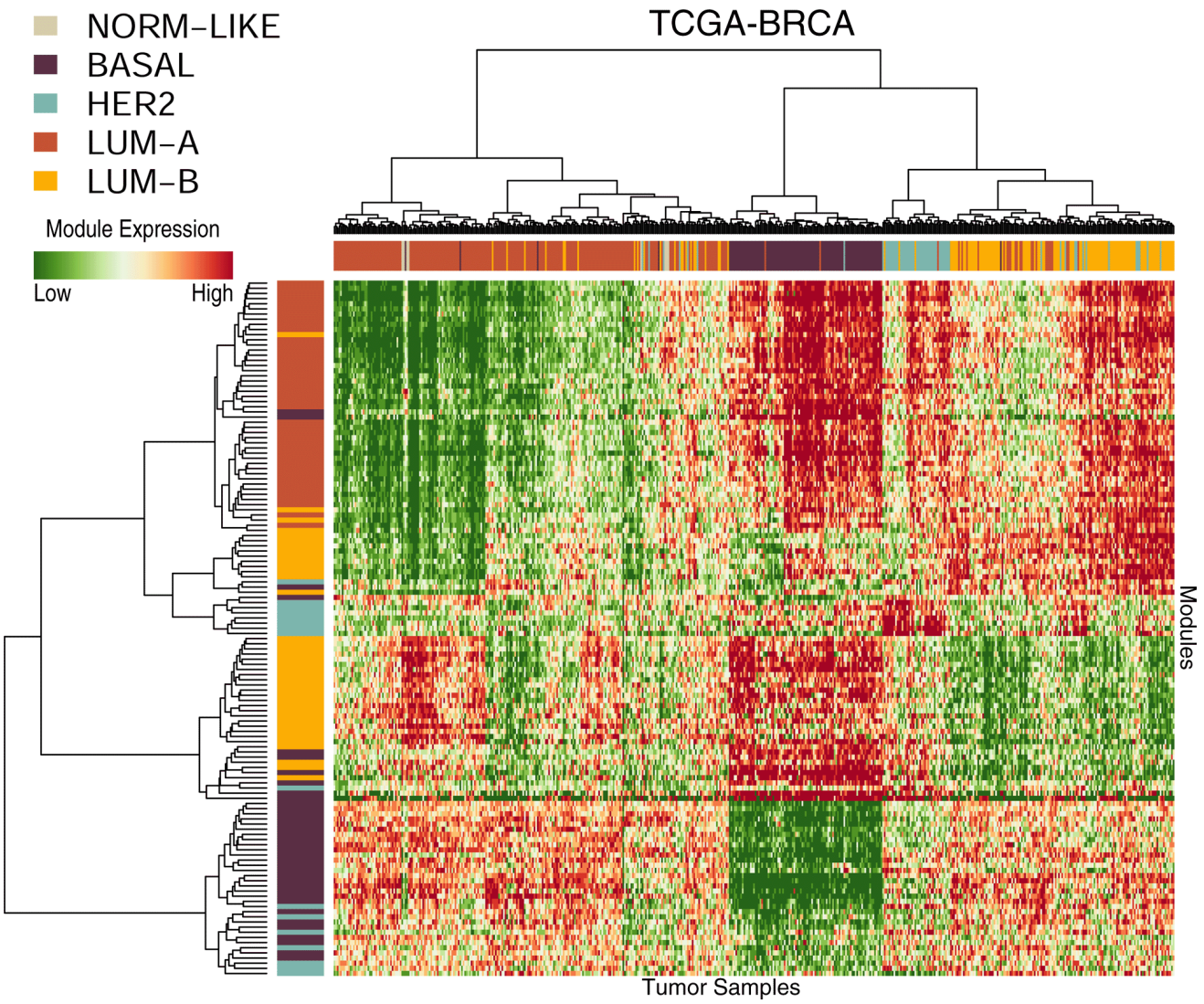
Supplemental Fig. S24. Comparison of phenotype classification accuracy achieved by HIT'nDRIVE-OptDis with that achieved by the best possible nearest neighbour classifier that uses a linear combination of all differentially expressed genes (part of R's caret package). As can be seen, in all subtypes but especially for the Luminal-A subtype HIT'nDRIVE-OptDis provides a much higher classification accuracy.



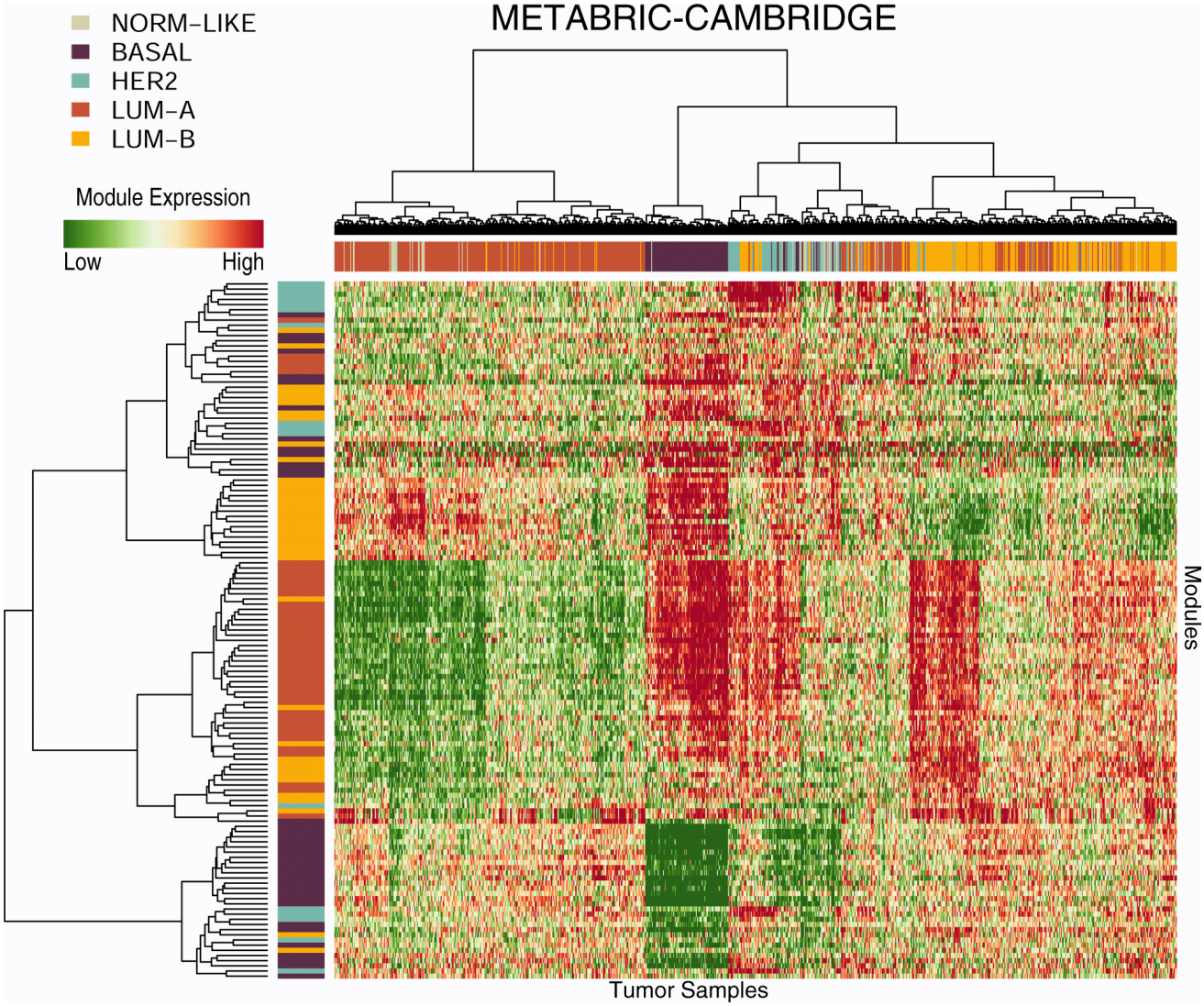
Supplemental Fig. S25. Comparison of HIT'nDRIVE+OptDis based modules against randomly selected modules. Phenotype Classification accuracy of HIT'nDRIVE driver seeded module identified by OptDis in TCGA-PRAD data against classification accuracy using randomly selected modules.



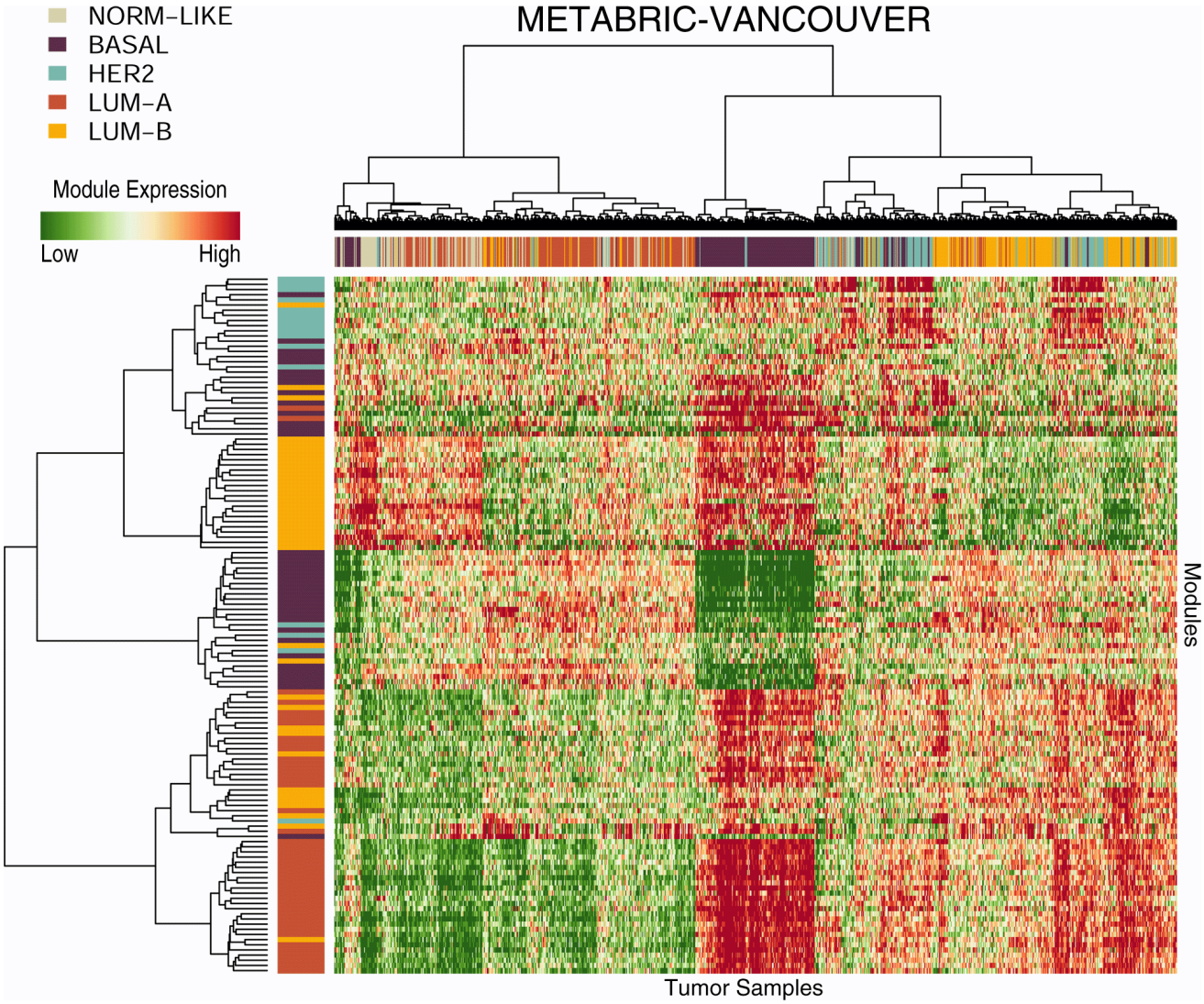
Supplemental Fig. S26. Mutual Exclusivity of Driver Modules. The left panel represents protein-interaction network among the component genes in the module seeded by a driver gene. The network was constructed based on STRING v10 protein-interaction network. Each node represents a gene/protein and edges represents interaction between the connected nodes. The driver gene node is colored in black. The middle panel represents the gene expression heatmap of the driver module genes among the patients in which the respective driver genes have been altered. The matrix on top of the heatmap shows the alteration status of the driver genes. The right panel shows the pathway enrichment of the driver modules. The enrichment test was performed using component genes in the driver module which was tested against different pathway databases (See supplementary methods for details). (A) Glutathione S-Transferase (GST) modules in OV. (B) Phosphoinositide-3-Kinase (PI3K) module in BRCA.



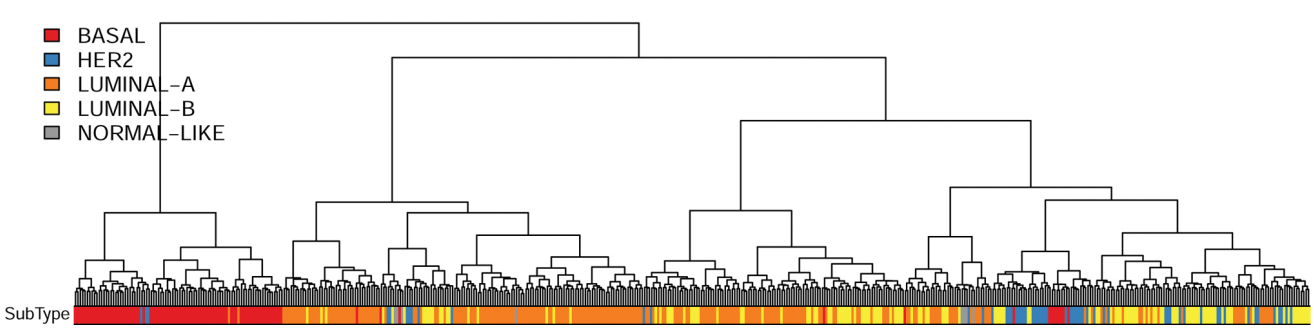
Supplemental Fig. S27. Module Expression Heatmap: TCGA-BRCA Dataset. The heatmap represents the activity-score (i.e. average expression of all component genes in the module) of the driver module. These driver module were discovered to distinguish one breast cancer subtype from the other subtypes. The vertical column represents each patient sample analyzed. The top color-bar represents the breast cancer subtypes. The horizontal row represents each driver module discovered. The color bar on the left of the heatmap represents each breast cancer subtype the module belongs to (i.e. separates that subtypes with rest of the subtypes). The dendrogram was generated using Euclidean distance and Ward's minimum variance method.



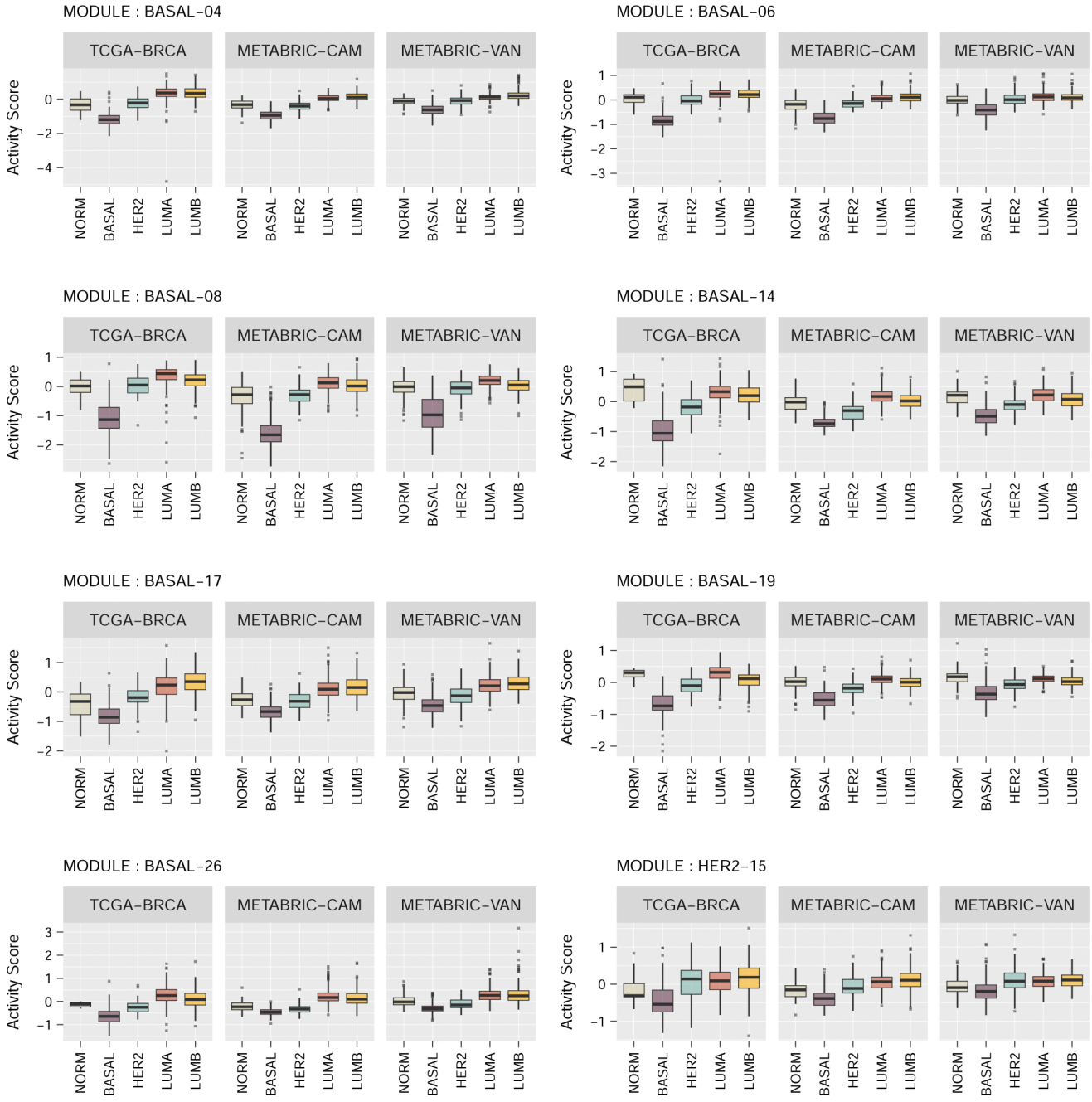
Supplemental Fig. S28. Module Expression Heatmap: METABRIC-CAMBRIDGE Dataset. The heatmap represents the activity-score (i.e. average expression of all component genes in the module) of the driver module. These driver module were discovered to distinguish one breast cancer subtype from the other subtypes. The vertical column represents each patient sample analyzed. The top color-bar represents the breast cancer subtypes. The horizontal row represents each driver module discovered. The color bar on the left of the heatmap represents each breast cancer subtype the module belongs to (i.e. separates that subtypes with rest of the subtypes). The dendrogram was generated using Euclidean distance and Ward's minimum variance method.



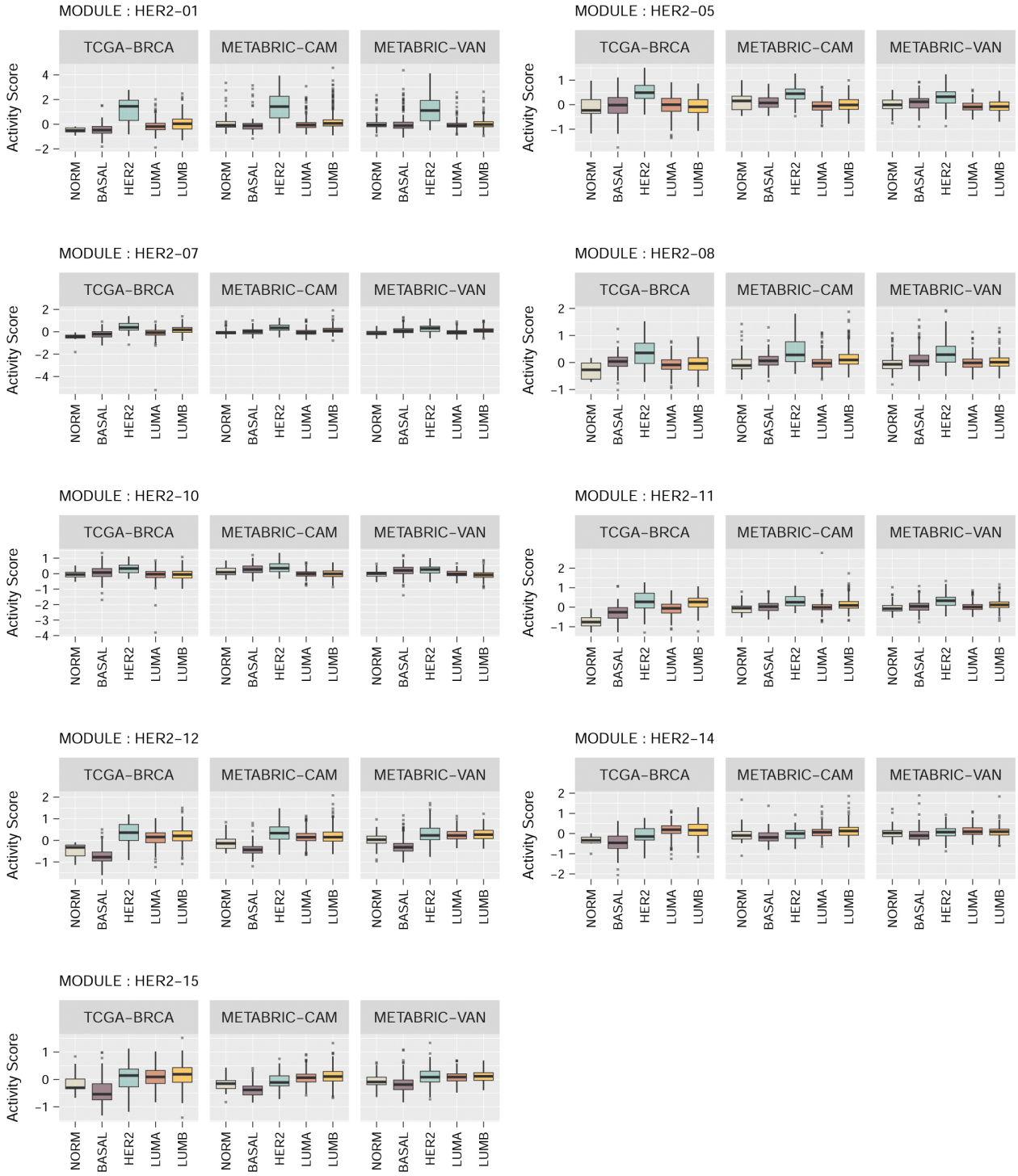
Supplemental Fig. S29. Module Expression Heatmap: METABRIC-VANCOUVER Dataset. The heatmap represents the activity-score (i.e. average expression of all component genes in the module) of the driver module. These driver module were discovered to distinguish one breast cancer subtype from the other subtypes. The vertical column represents each patient sample analyzed. The top color-bar represents the breast cancer subtypes. The horizontal row represents each driver module discovered. The color bar on the left of the heatmap represents each breast cancer subtype the module belongs to (i.e. separates that subtypes with rest of the subtypes). The dendrogram was generated using Euclidean distance and Ward's minimum variance method.



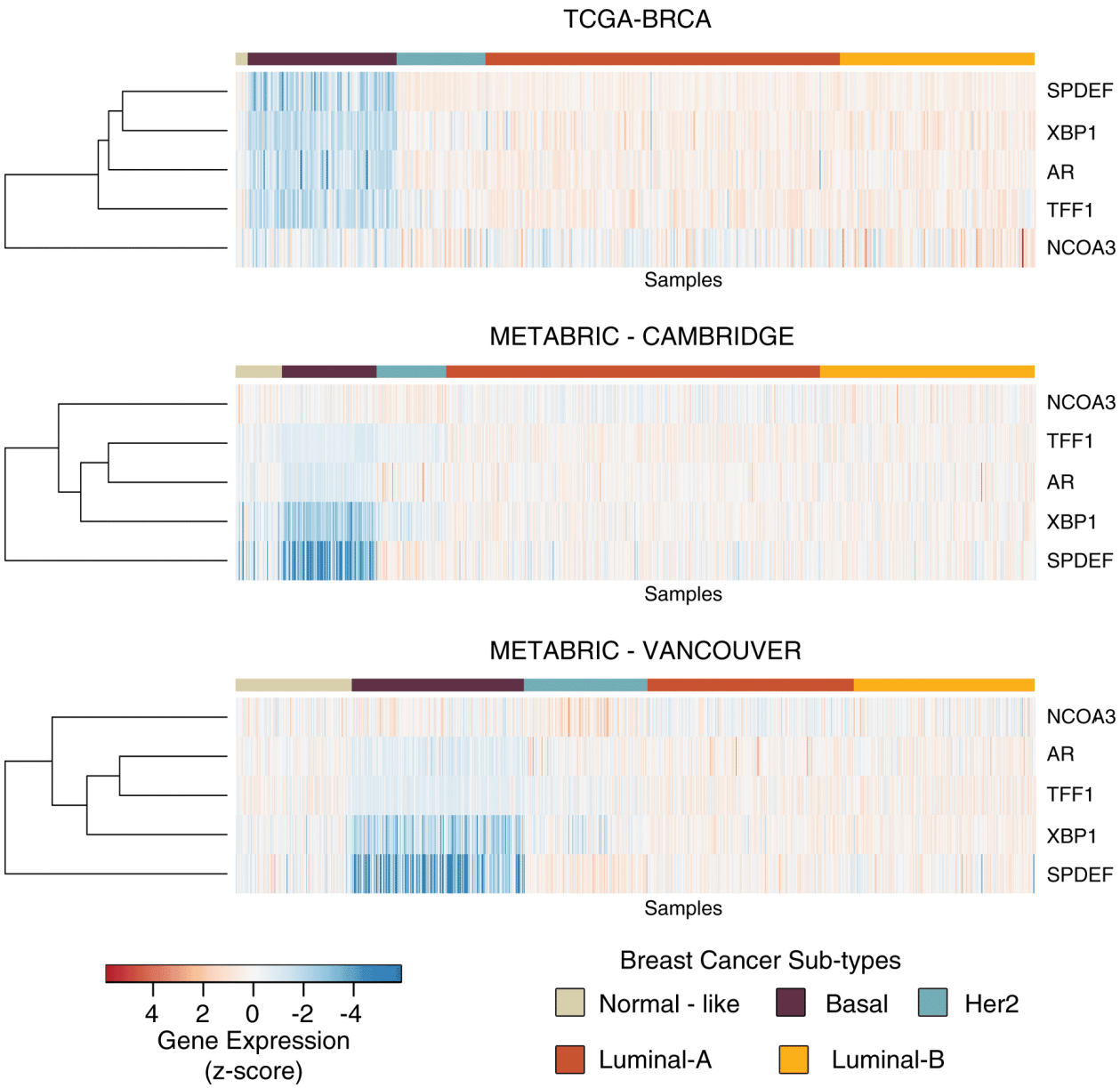
Supplemental Fig. S30. Unsupervised Clustering of BRCA subtypes in TCGA-BRCA cohort. Unsupervised classification of BRCA subtypes based on the gene-expression profiles. The dendrogram was generated using Euclidean distance and Ward's minimum variance method (via hclust, R's hierarchical clustering function). As can be seen, unsupervised clustering can not identify BRCA subtypes well. In particular, LUMINAL-A and LUMINAL-B subtypes are well mixed in the dendogram.



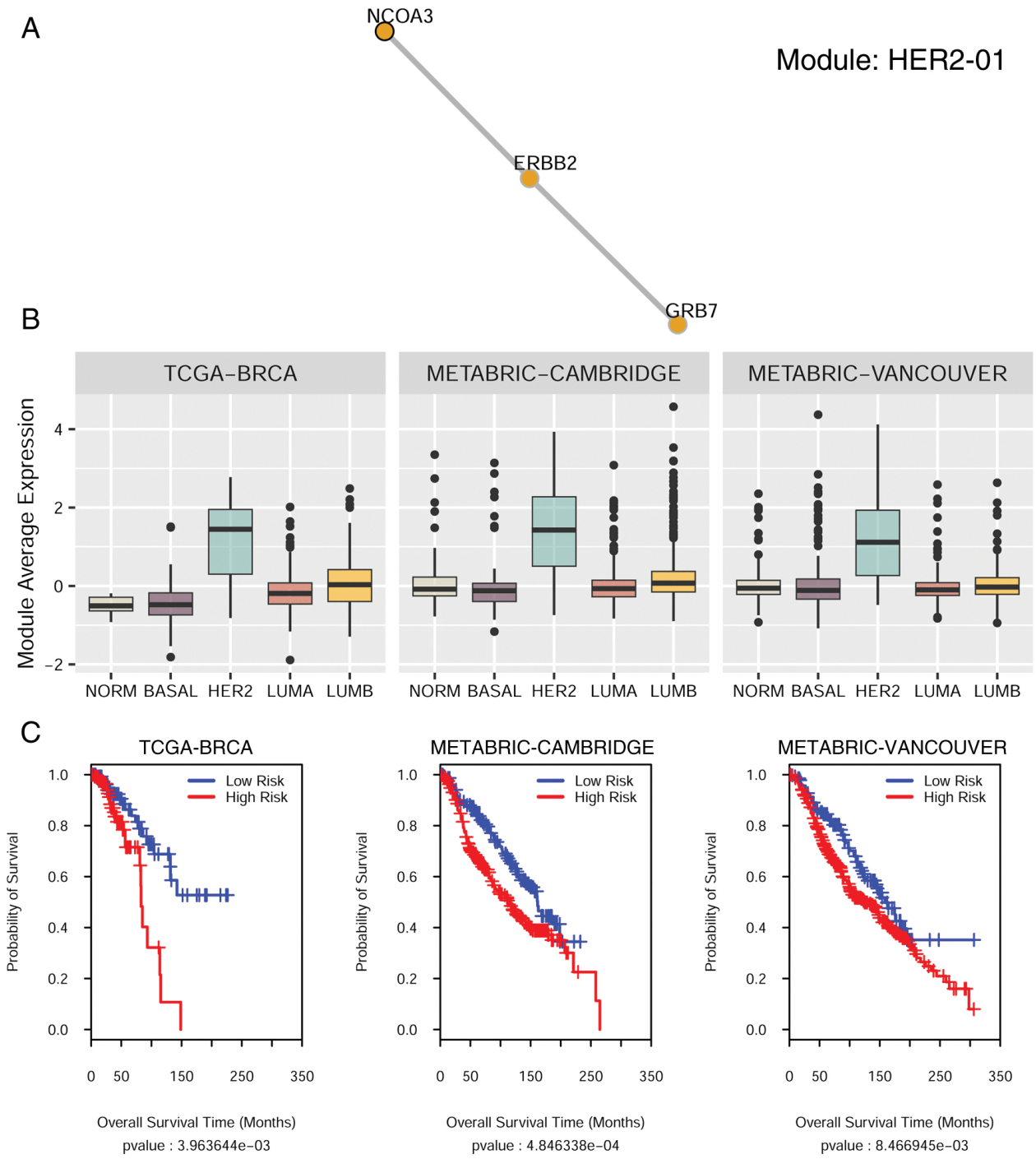
Supplemental Fig. S31. Activity Score of BRCA subtype-specific modules containing *ESR1*. Activity Score of a module represents the average expression of component genes in the module.



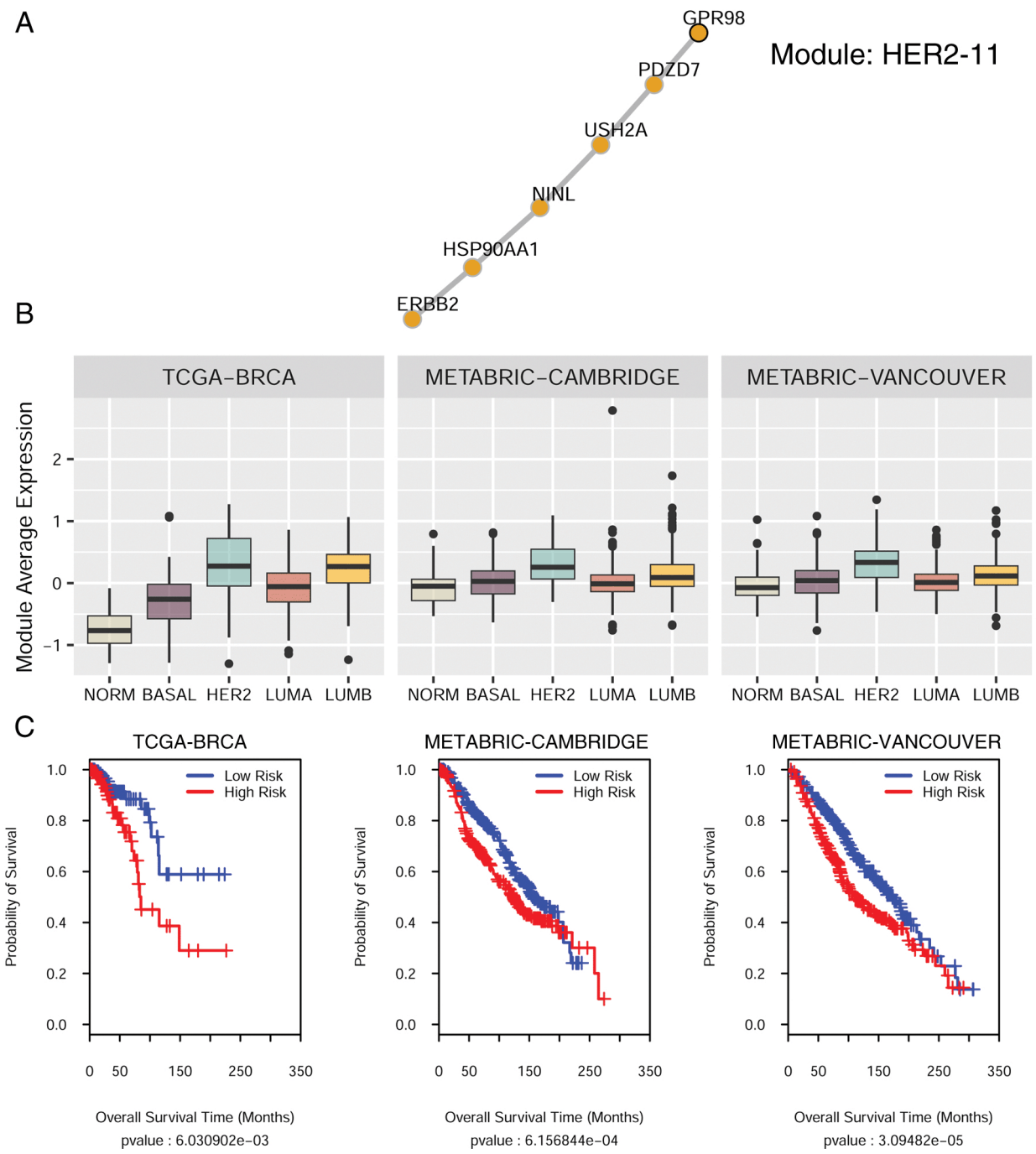
Supplemental Fig. S32. Activity Score of BRCA subtype-specific modules containing *ERBB2*. Activity Score of a module represents the average expression of component genes in the module.



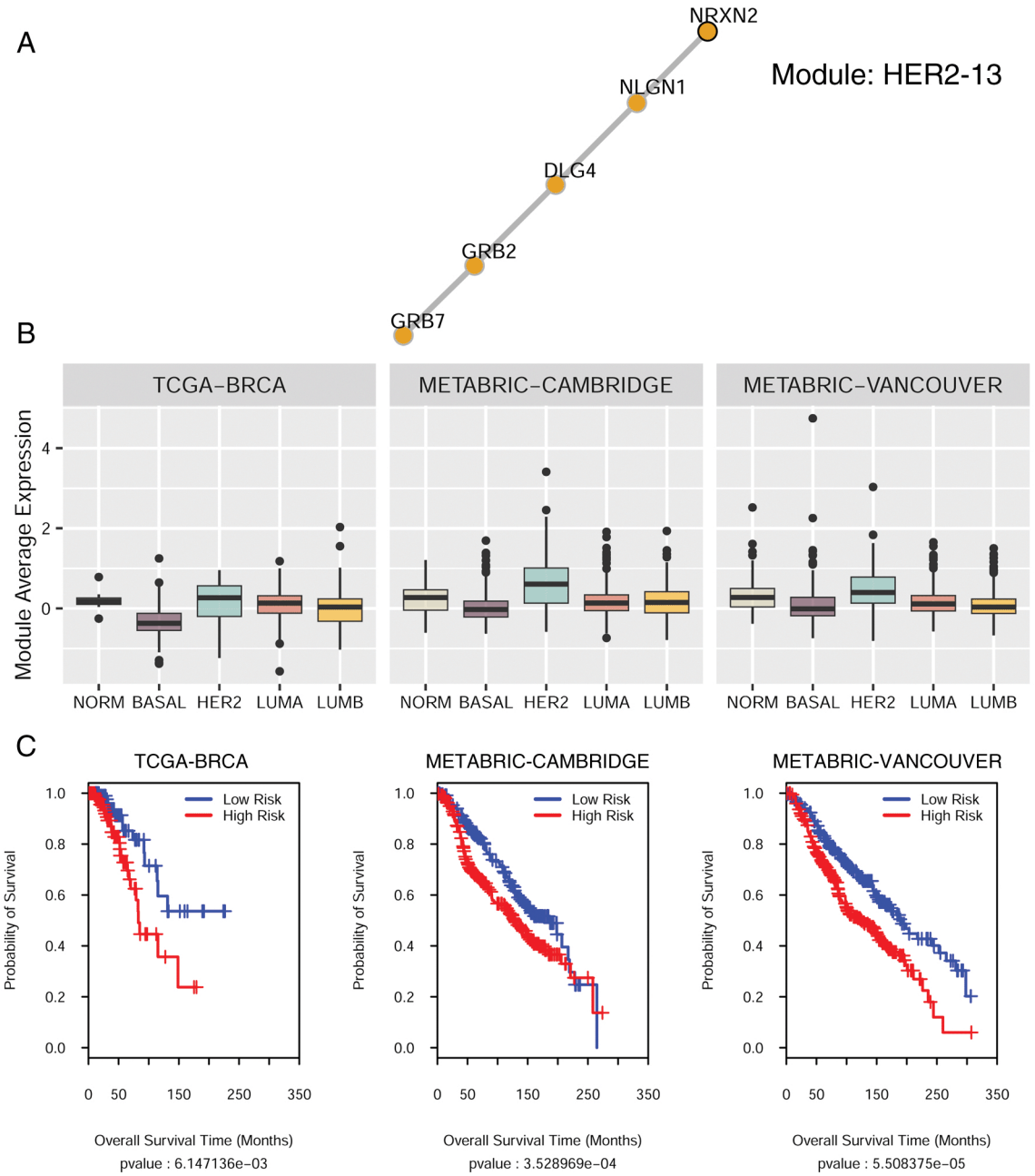
Supplemental Fig. S33. Heatmap of *NCOA3* driver module expression across different BRCA subtypes. The *NCOA3* module contains *NCOA3*, *AR*, *TFF1*, *XBP1* and *SPDEF* as component genes.



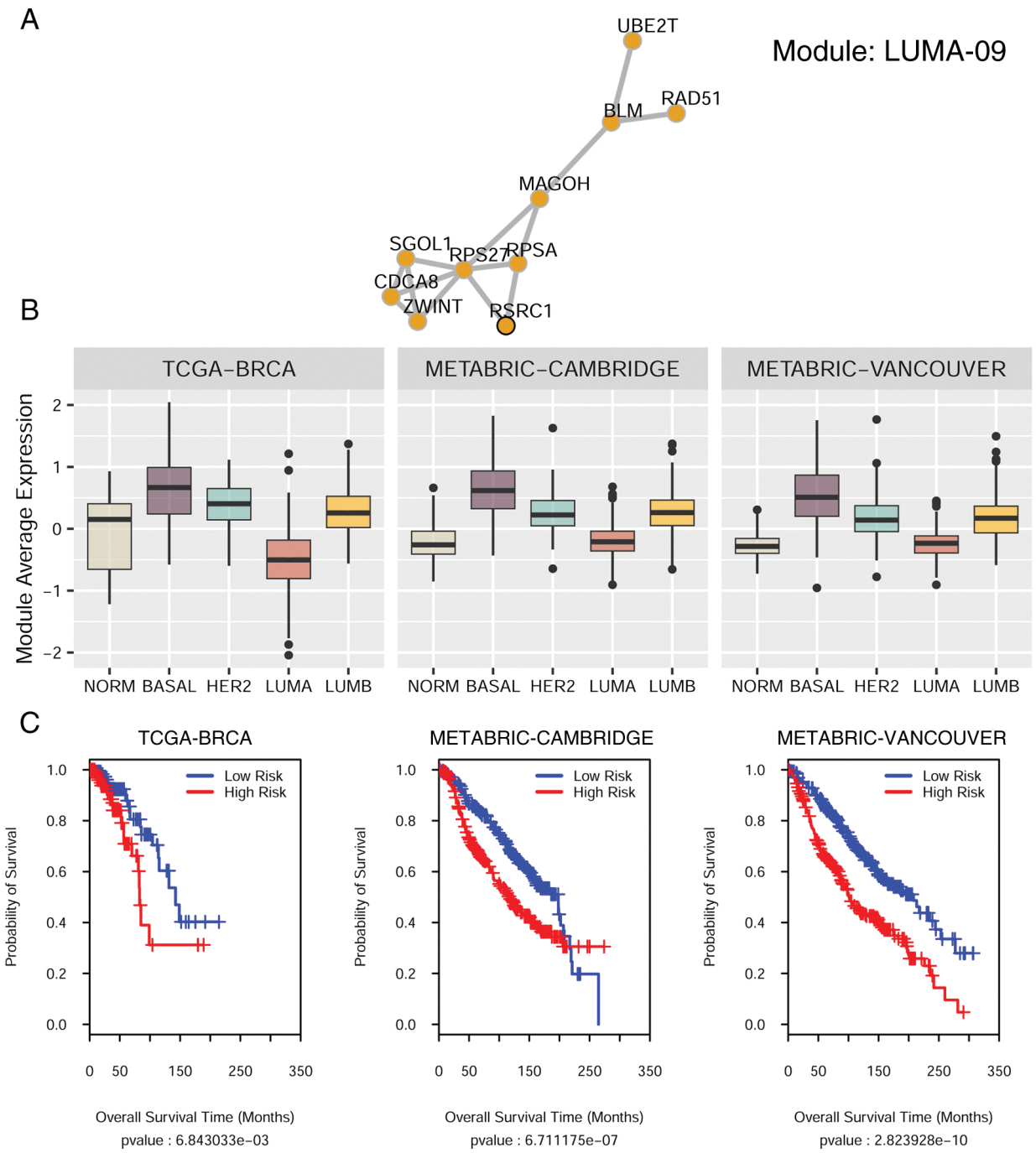
Supplemental Fig. S34. BRCA subtype specific driver module (HER2-01). (A) Module seeded by *NCOA3* that distinguished Her2 subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



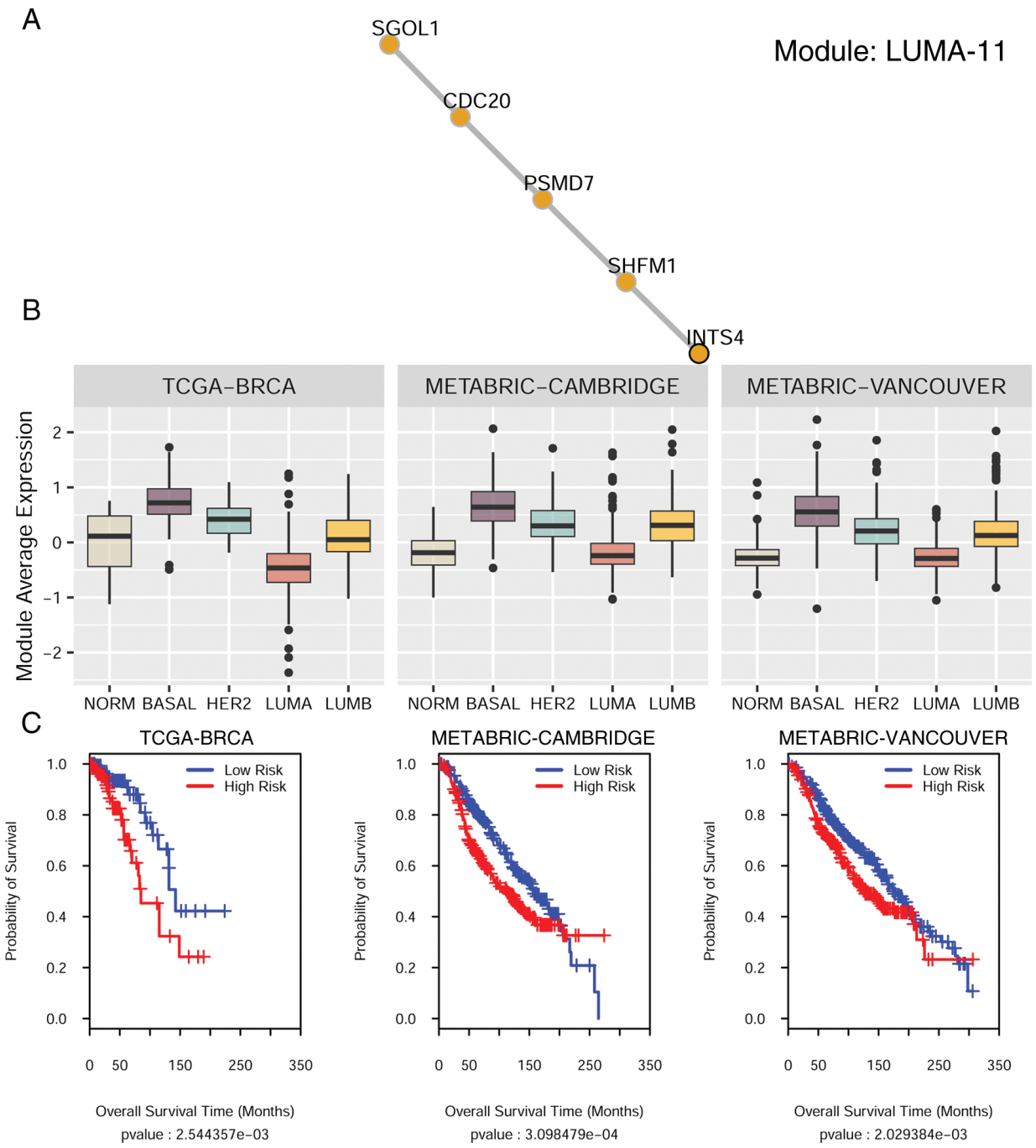
Supplemental Fig. S35. BRCA subtype specific driver module (HER2-11). (A) Module seeded by GPR98 that distinguished Her2 subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



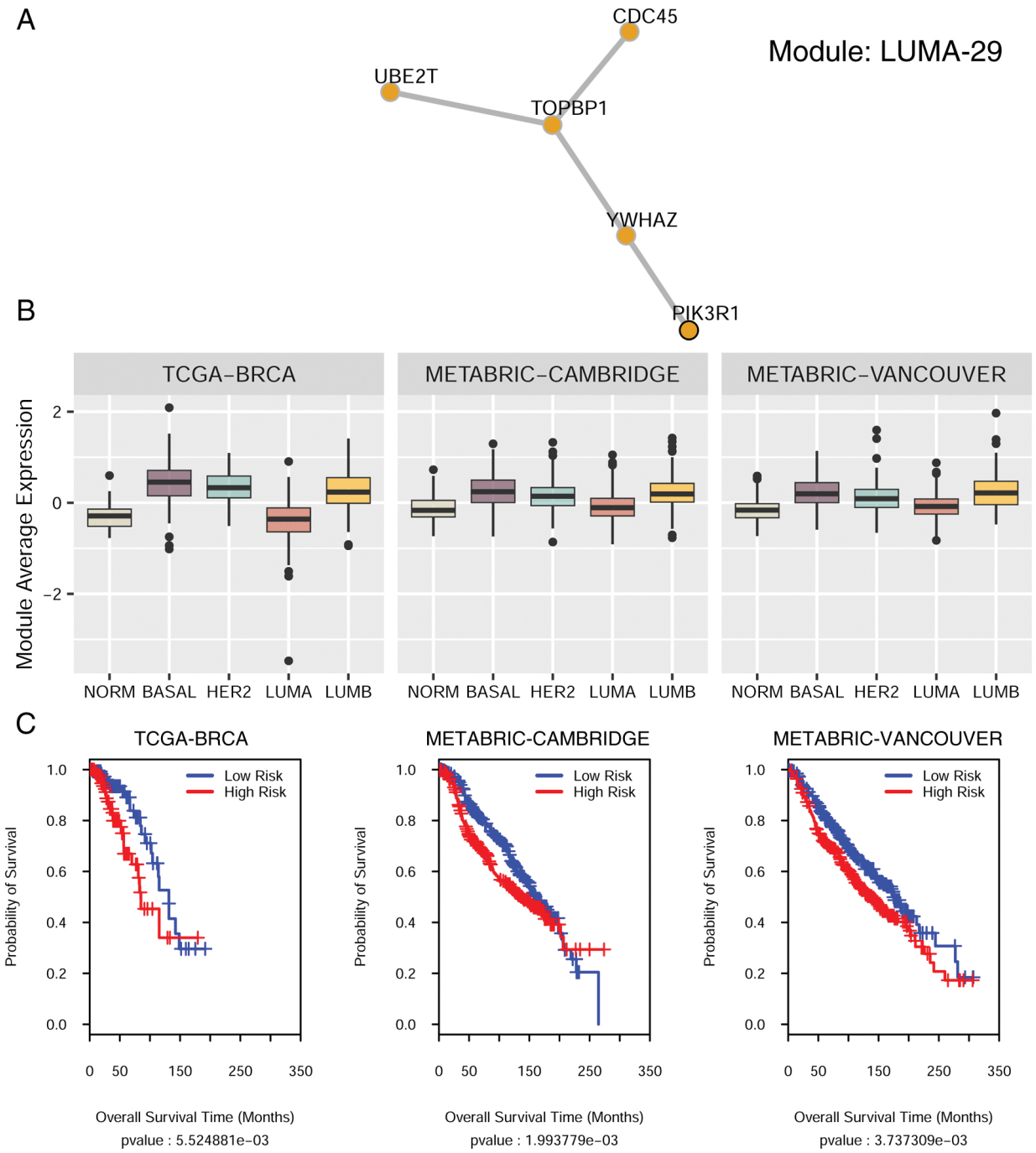
Supplemental Fig. S36. BRCA subtype specific driver module (HER2-13). (A) Module seeded by *NRXN2* that distinguished Her2 subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



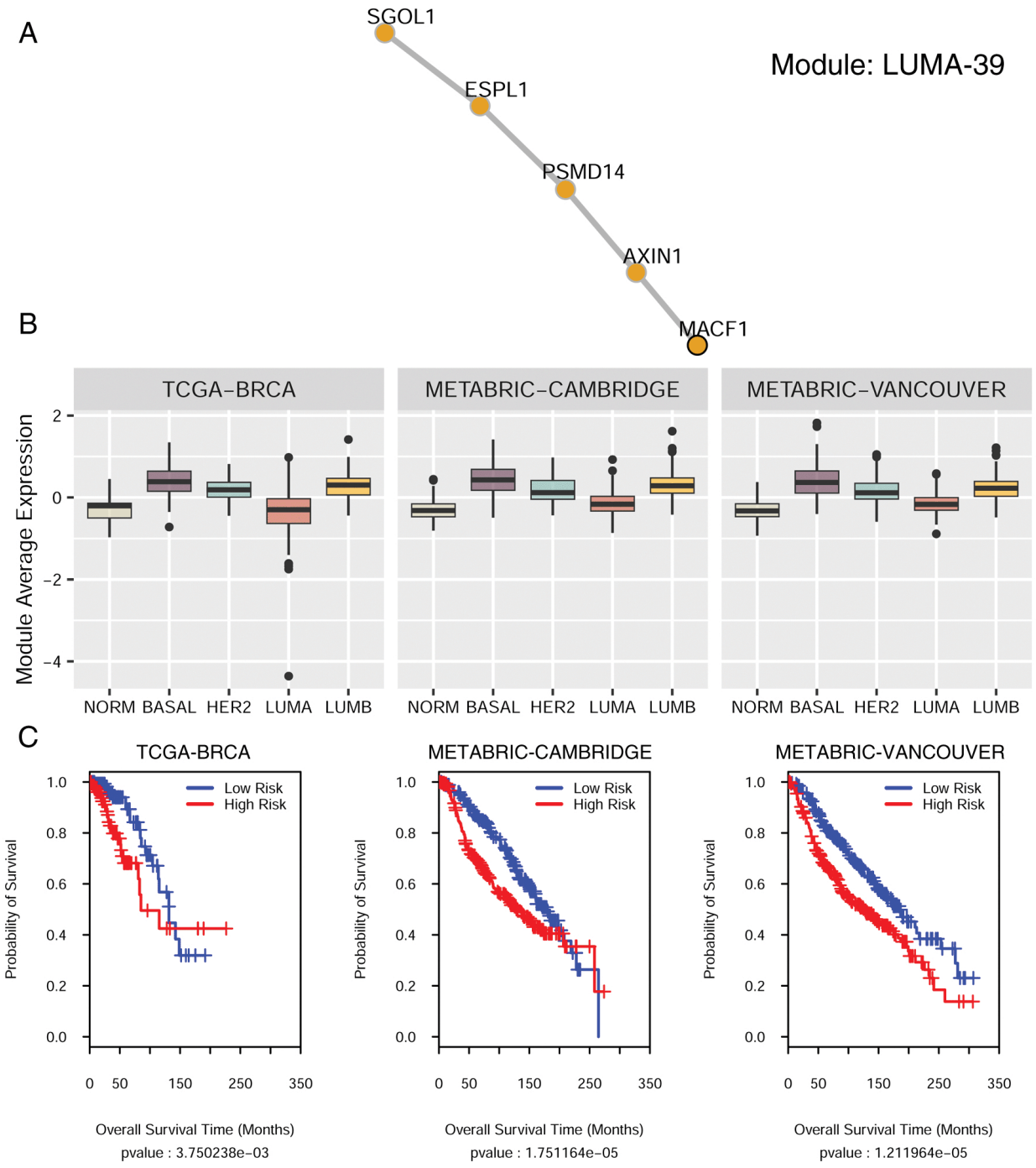
Supplemental Fig. S37. BRCA subtype specific driver module (LUMA-09). (A) Module seeded by *RSRC1* that distinguished Luminal-A subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



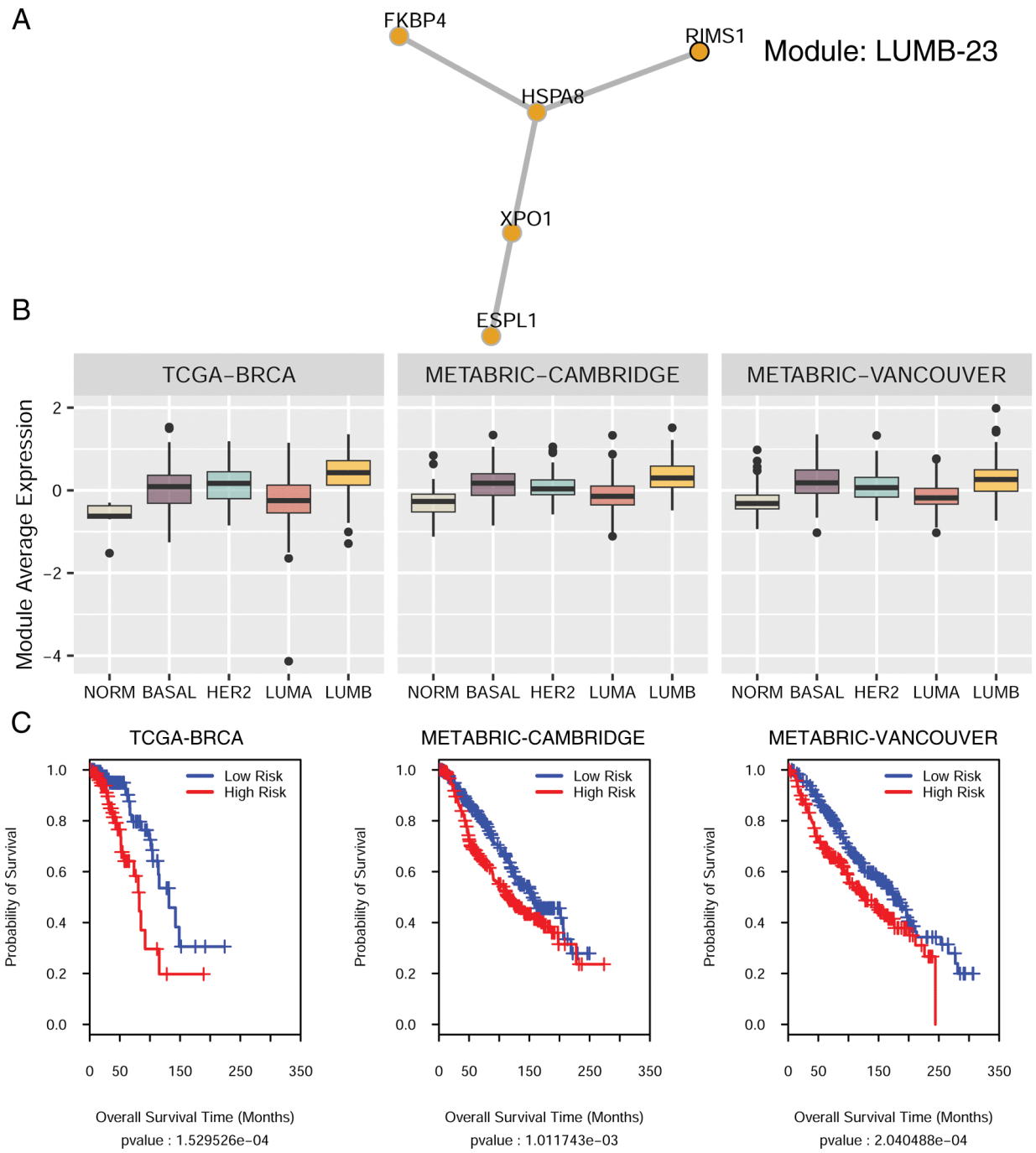
Supplemental Fig. S38. BRCA subtype specific driver module (LUMA-11). (A) Module seeded by *INTS4* that distinguished Luminal-A subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



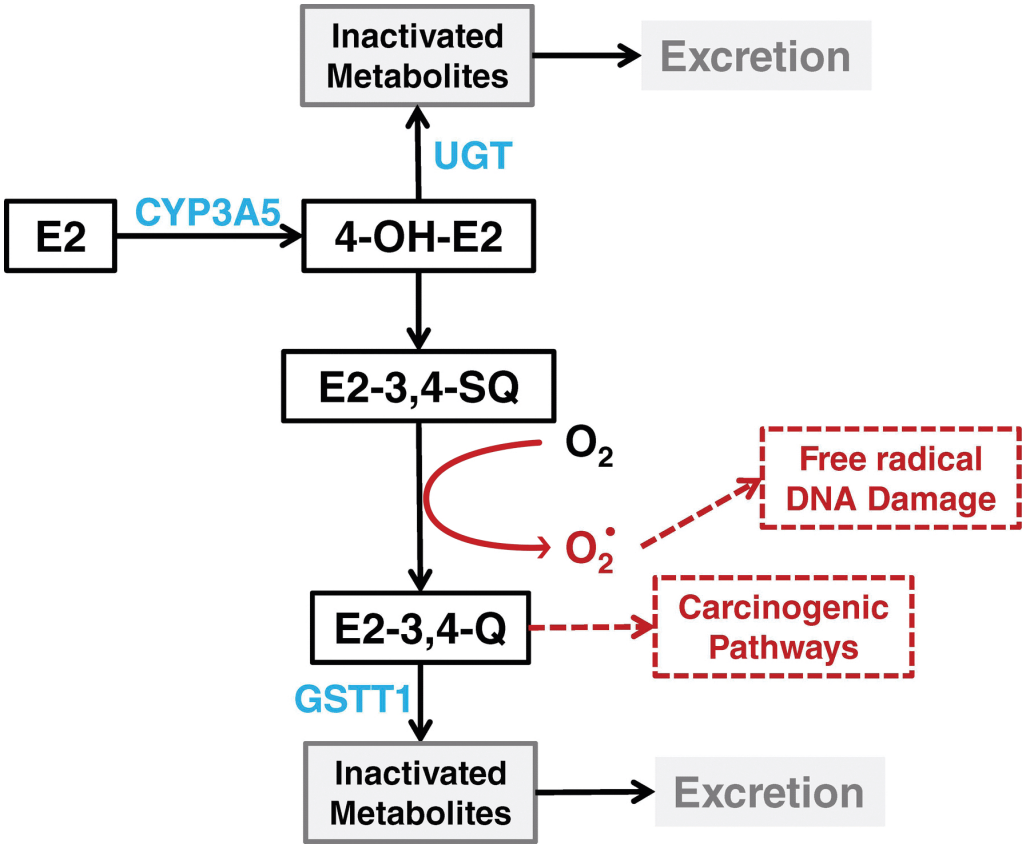
Supplemental Fig. S39. BRCA subtype specific driver module (LUMA-29). (A) Module seeded by *PIK3R1* that distinguished Luminal-A subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



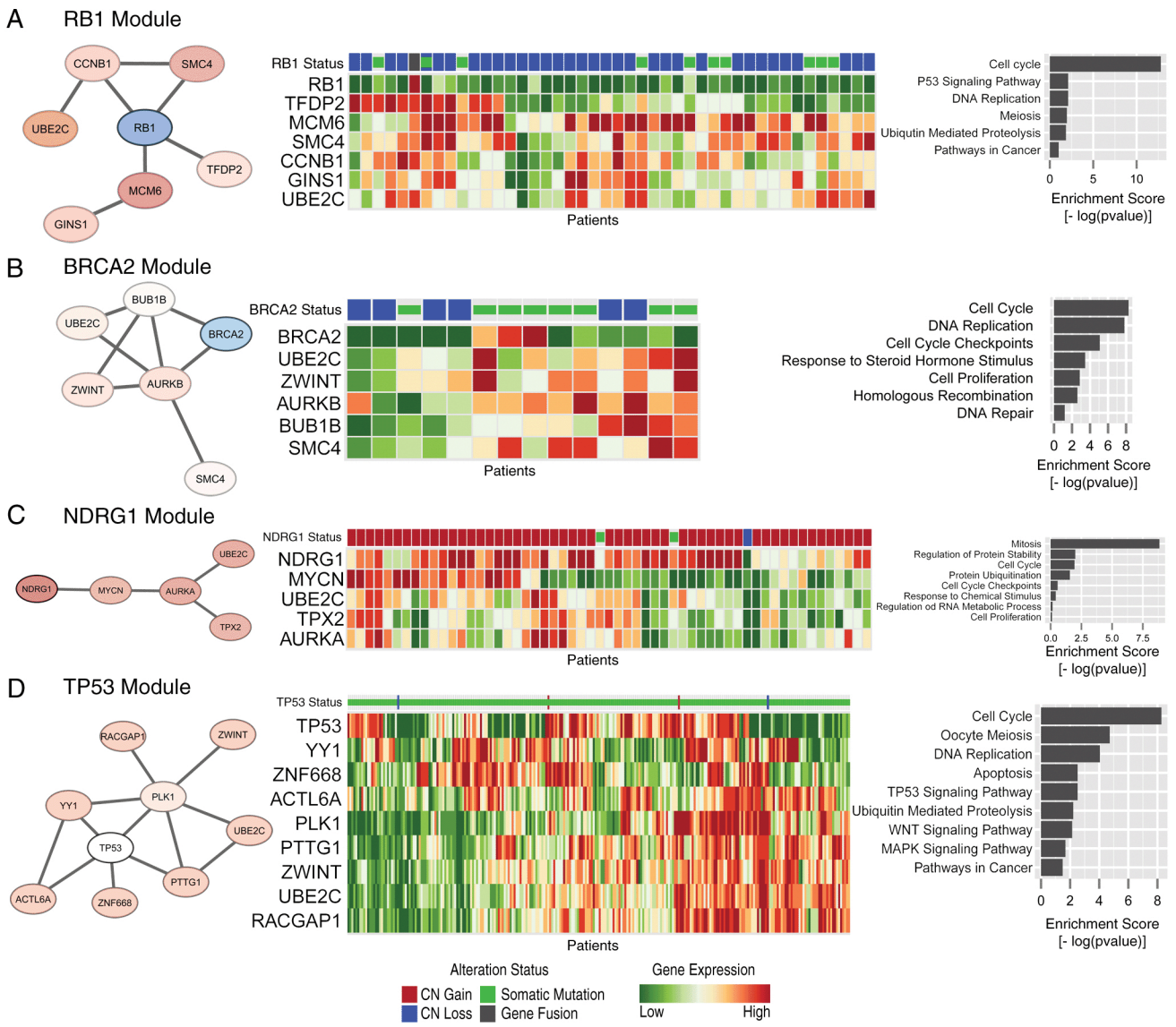
Supplemental Fig. S40. BRCA subtype specific driver module (LUMA-39). (A) Module seeded by *MACF1* that distinguished Luminal-A subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



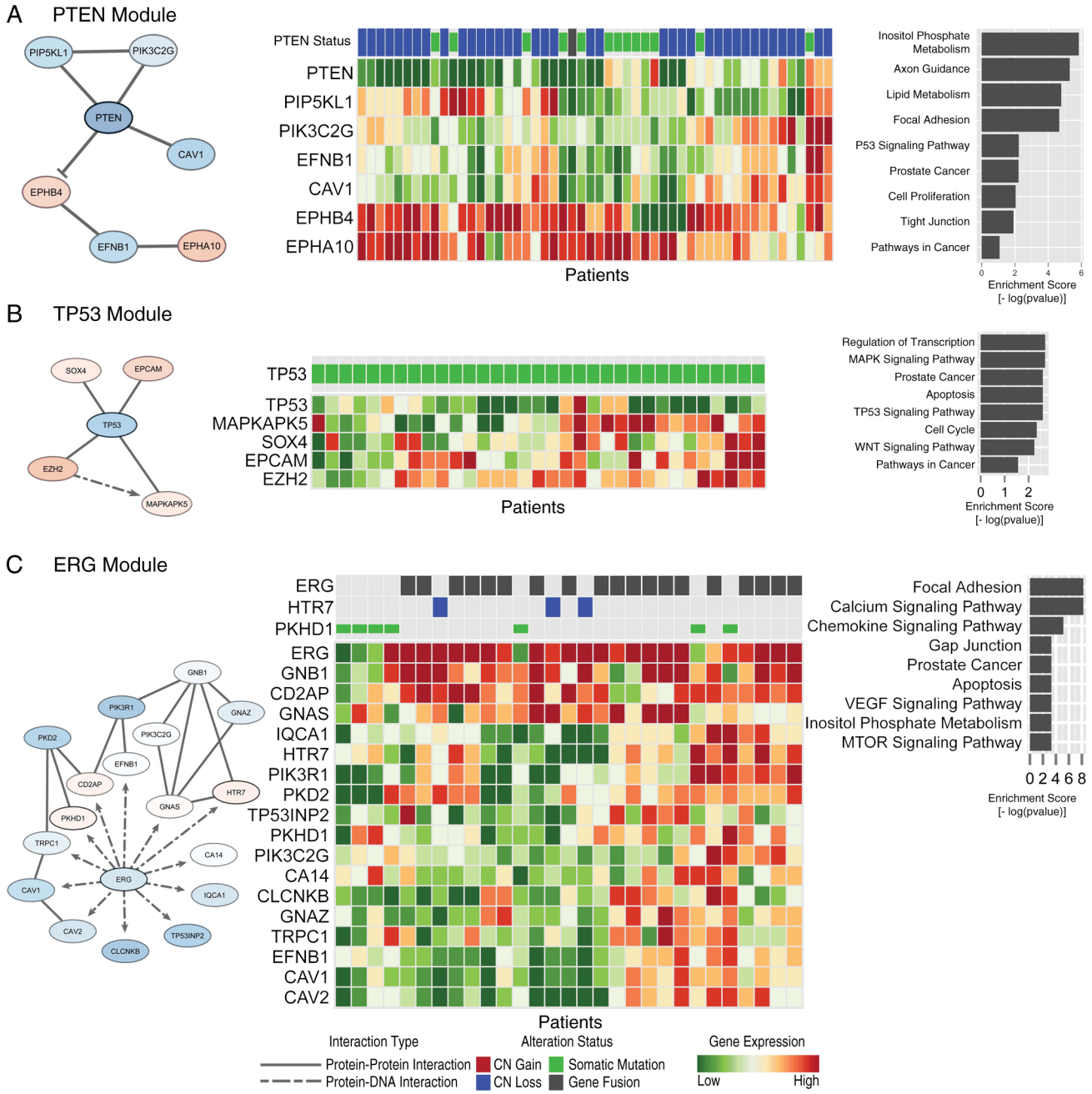
Supplemental Fig. S41. BRCA subtype specific driver module (LUMB-23). (A) Module seeded by *RIMS1* that distinguished Luminal-A subtype from rest of the BRCA subtypes. (B) Activity-score of the module across different BRCA subtypes. (C) Kaplan-Meier plot showing the significant association of the module with patients' clinical outcome in three different datasets



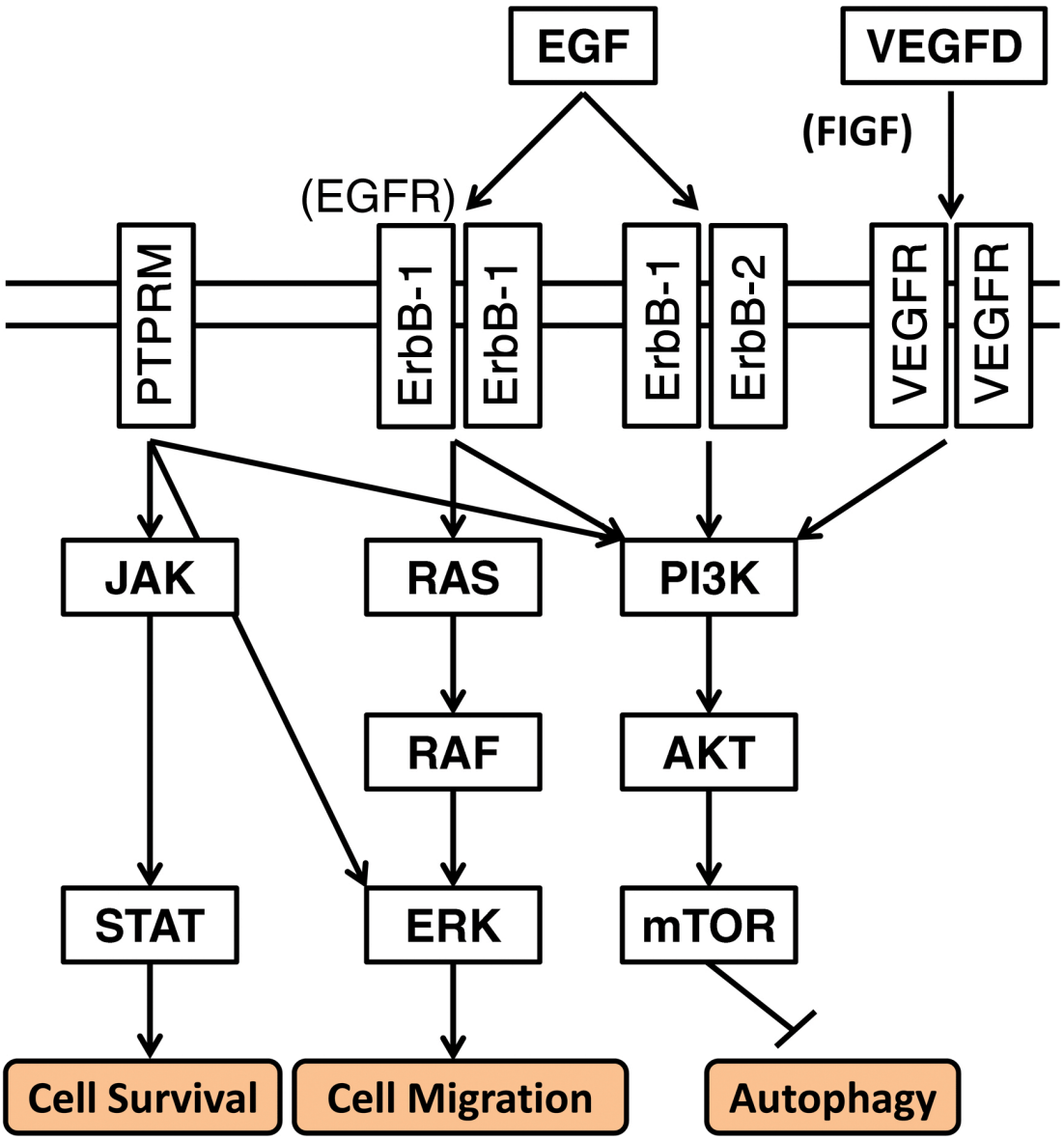
Supplemental Fig. S42. Metabolism of Estrogen. Estradiol (E2) hydrolyation by CYP3A5 leads to a carcinogenic pathway. 4-OH-E2 undergoes metabolic redox cycling to generate free radicals such as superoxide and the chemically-reactive estrogen semiquinone/quinone intermediates. GSTT1 and UGT inactivates estrogen and intermediate compounds, avoiding the formation of carcinogens such as E2-3,4-Q. 4-OH-E2, 4-hydroxyestradiol; E2-3,4-SQ, estradiol-3,4-semiquinone; E2-3,4-Q, estradiol-3,4-quinone; GSTT1, Glutathione S-Transferase Theta 1; UGT, UDP Glucuronosyltransferase.



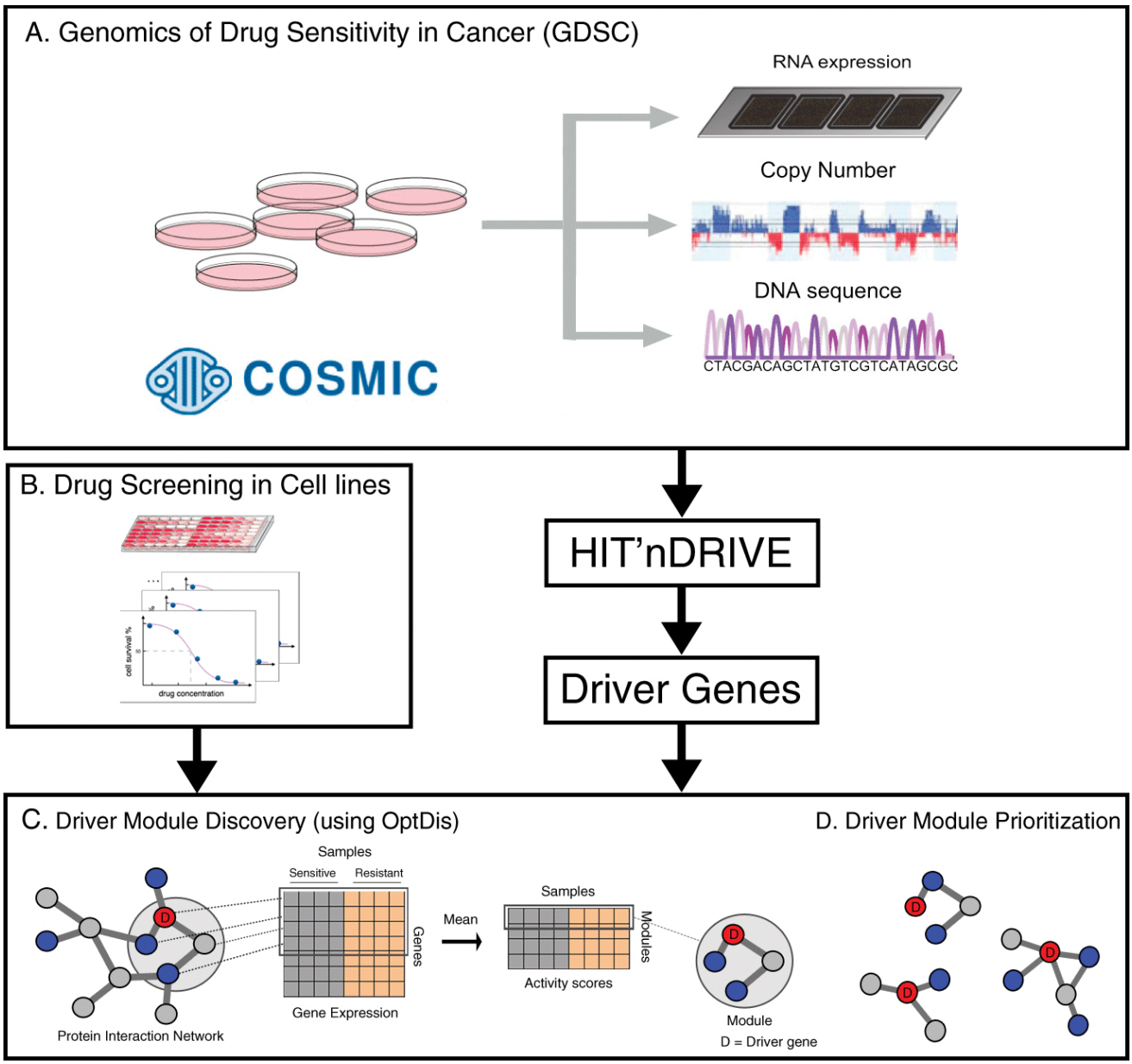
Supplemental Fig. S43. Drivers Modules of Ovarian Cancer. The left panel represents protein-interaction network among the component genes in the module seeded by a driver gene. The network was constructed based on String v10 protein-interaction network. Each node represents a gene/protein and edges represents interaction between the connected nodes. The node color represents the mean gene expression of the gene among the patient samples represented. The driver gene node is colored in black. The middle panel represents the gene expression heatmap of the driver module genes among the patients in with the respective driver gene(s) have been altered. The matrix on top of the heatmap shows the alteration status of the driver gene(s). The right panel shows the pathway enrichment of the driver modules.



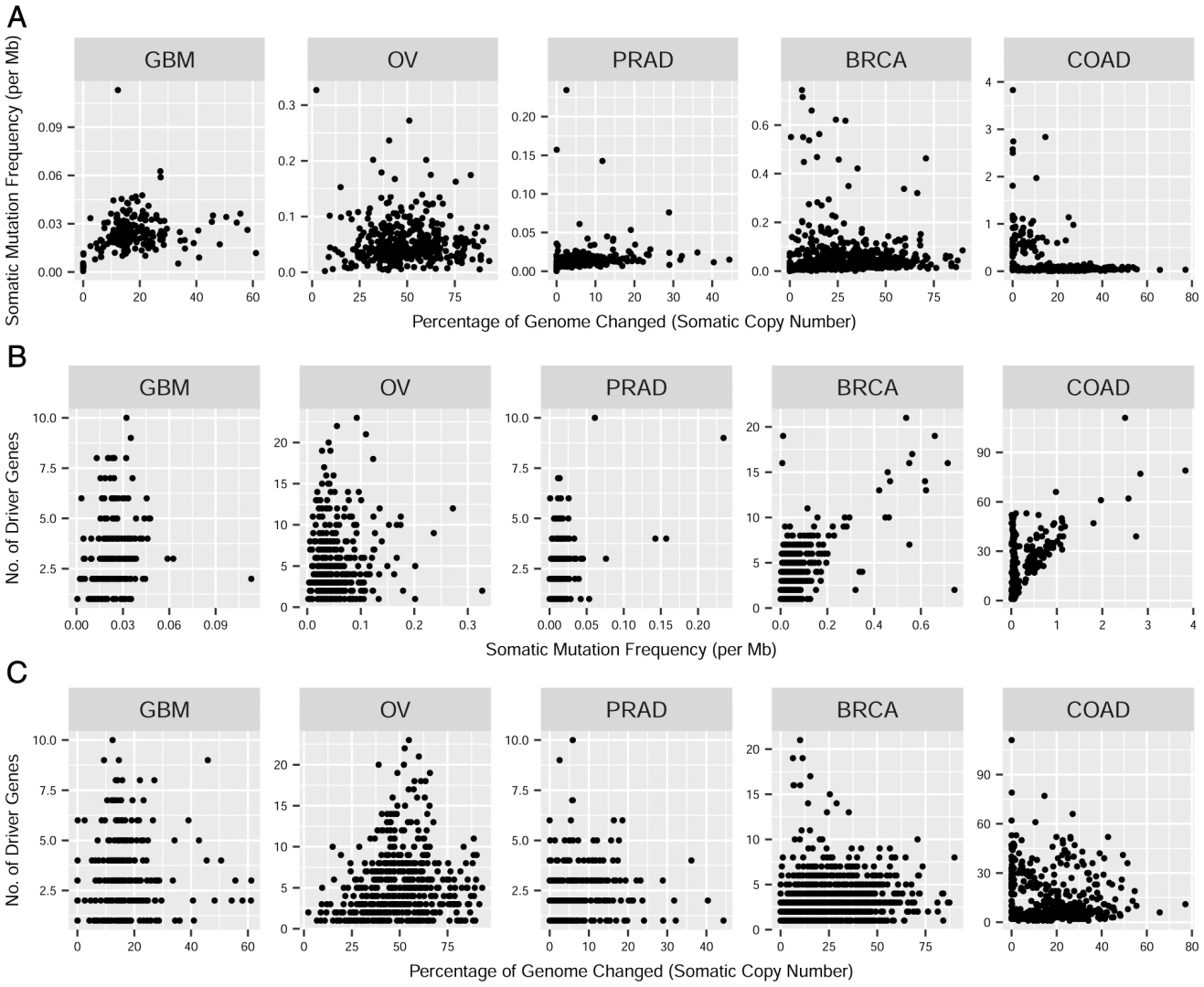
Supplemental Fig. S44. Drivers Modules of Prostate Cancer. The left panel represents protein-interaction network among the component genes in the module seeded by a driver gene. The network was constructed based on String v10 protein-interaction network and ChipSeq based interaction edges for some key transcription factors related to prostate cancer were added. We infer interactions involving transcription factor binding to the promoter region of associated genes. Each node represents a gene/protein and edges represents interaction between the connected nodes. The node color represents the mean gene expression of the gene among the patient samples represented. The driver gene node is colored in black. The middle panel represents the gene expression heatmap of the driver module genes among the patients in with the respective driver gene(s) have been altered. The matrix on top of the heatmap shows the alteration status of the driver gene(s). The right panel shows the pathway enrichment of the driver modules.



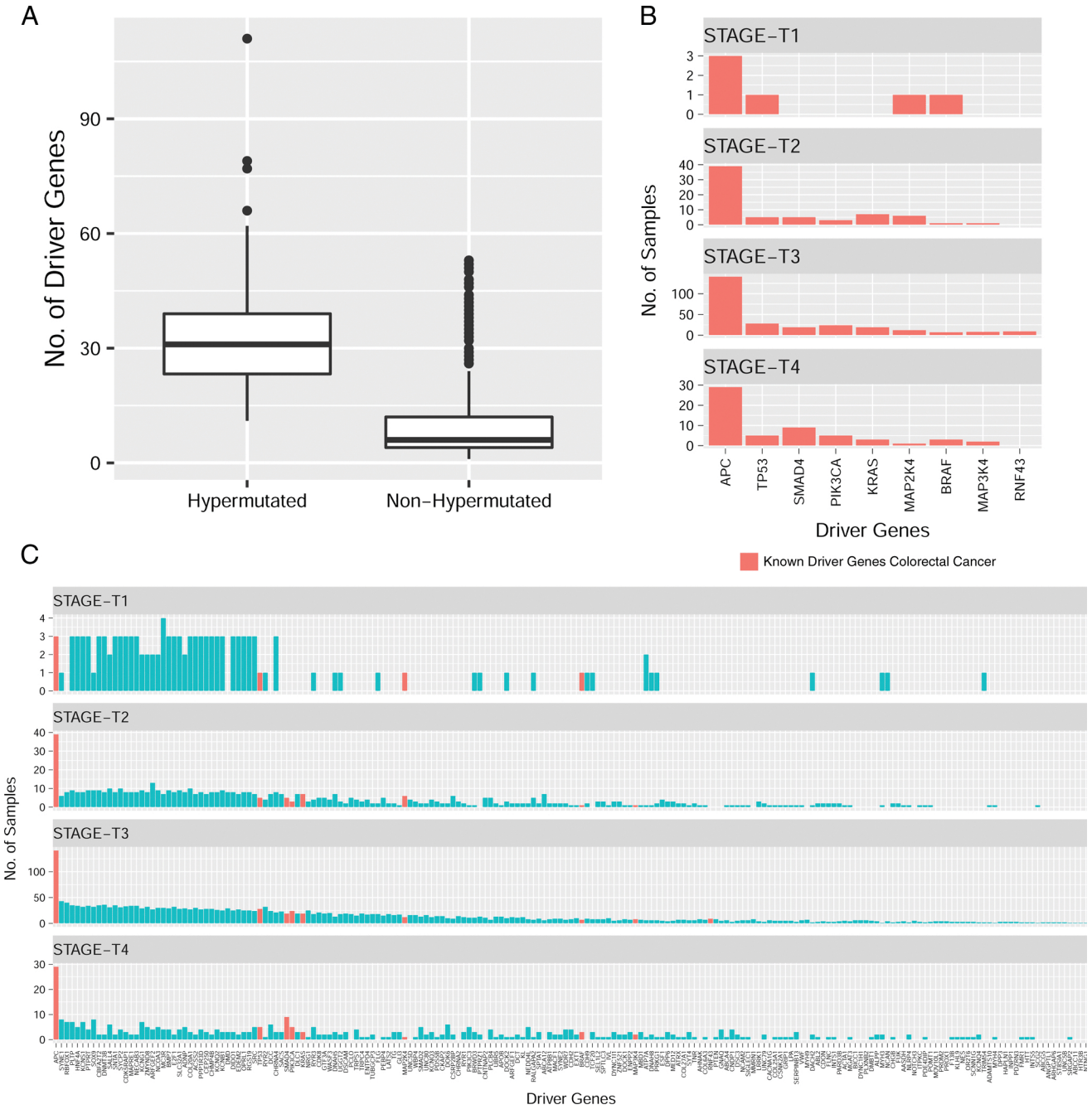
Supplemental Fig. S45. EGFR-PI3K Signaling Pathway.



Supplemental Fig. S46. Overview of Drug Response Analysis using HIT'nDRIVE + OptDis. (A) Somatic mutation, Copy Number Aberration and mRNA Gene Expression data (for four cancer types - GBM, OV, PRAD and BRCA) were obtained from Genomics of Drug Sensitivity in Cancer (GDSC) project (Iorio et al. 2016). We ran HIT'nDRIVE to identify driver genes of individual cancer cell lines. (B) Drug sensitivity data from drug screening of a total of 265 drugs on the above cell lines were obtained from GDSC project. The cell lines were stratified into either sensitive or resistant phenotype. (C-D) The driver genes were used as seeds in the network to identify sub-networks that discriminate between the drug-response phenotypes (i.e. sensitive vs resistant cell lines).



Supplemental Fig. S47. Correlation between the number of driver genes predicted by HIT'nDRIVE with mutation rate and copy-number burden. (A) Correlation between Mutation rate (frequency of somatic mutation per Mb) with copy-number burden (percentage of genome changed calculated using somatic copy number changes). Correlation of the number of driver genes predicted by HIT'nDRIVE with (B) mutation rate and (C) copy-number burden.



Supplemental Fig. S48. HIT'nDRIVE predicted driver genes of Colorectal cancer (TCGA-COAD). (A) Box plot comparing the number of HIT'nDRIVE predicted driver genes in hypermutated and non-hypermutated cases of TCGA-COAD. (B) Recurrent frequency of the known stage-specific driver genes that were also predicted by HIT'nDRIVE in non-hypermutated COAD samples. (C) Recurrent frequency of all driver genes predicted by HIT'nDRIVE in non-hypermutated COAD samples.

References

Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, *et al.* 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, **8**(8):816–24.

Bowen NJ, Walker LD, Matyunina LV, Logani S, Totten Ka, Benigno BB, and McDonald JF 2009. Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. *BMC medical genomics*, **2**:71.

Chatterjee P, Choudhary GS, Sharma A, Singh K, Heston WD, Ciezki J, Klein EA, and Almasan A 2013. PARP Inhibition Sensitizes to Low Dose-Rate Radiation TMPRSS2-ERG Fusion Gene-Expressing and PTEN-Deficient Prostate Cancer Cells. *PLoS ONE*, **8**(4):1–12.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, *et al.* 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403):346–52.

Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, and Tabernero J 2017. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature reviews. Cancer*, **17**(2):79–92.

Ekhart C, Rodenhuis S, Smits PHM, Beijnen JH, and Huitema ADR 2009. An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug treatment. *Cancer Treatment Reviews*, **35**(1):18–31.

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, *et al.* 2016. The Reactome pathway Knowledgebase. *Nucleic acids research*, **44**(D1):D481–7.

Fruman DA and Rommel C 2014. PI3K and cancer: lessons, challenges and opportunities. *Nature reviews. Drug discovery*, **13**(2):140–56.

Gordon V and Banerji S 2013. Molecular pathways: PI3K pathway targets in triple-negative breast cancers. *Clinical Cancer Research*, **19**(14):3738–3744.

Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, Quist MJ, Jing X, Lonigro RJ, Brenner JC, *et al.* 2012. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, **487**(7406):239–43.

Guillemette C, Bélanger A, and Lépine J 2004. Metabolic inactivation of estrogens in breast tissue by UDP-glucuronosyltransferase enzymes: an overview. *Breast cancer research : BCR*, **6**(6):246–254.

Hormozdiari F, Alkan C, Eichler EE, and Sahinalp SC 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, **19**(7):1270–1278.

Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H, *et al.* 2016. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, **166**(3):740–54.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, *et al.* 2009. Human Protein Reference Database–2009 update. *Nucleic acids research*, **37**(Database issue):D767–72.

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, the R Core Team, *et al.* 2016. *caret: Classification and Regression Training*. R package version 6.0-68.

Levin DA, Peres Y, and Wilmer EL 2008. *Markov Chains and Mixing Times*. American Mathematical Society.

Mihail M, Papadimitriou CH, and Saberi A 2006. On certain connectivity properties of the internet topology. *J. Comput. Syst. Sci.*, **72**(2):239–251.

Murat A, Migliavacca E, Gorlia T, Lambiv WL, Shay T, Hamou MF, de Tribolet N, Regli L, Wick W, Kouwenhoven MCM, *et al.* 2008. Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *Journal of clinical oncology*, **26**(18):3015–24.

Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, and Ganesan S, *et al.* 2006. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, **9**(2):121–132.

Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald T, Bunting K, Demichelis F, Melnick A, *et al.* 2012. Oncogene-mediated alterations in chromatin conformation. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(23):9083–9088.

Rosner B 1983. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics)*, **25**(2):165–172.

Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A, and Lopez-Bigas N 2015. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer cell*, **27**(3):382–396.

Sharma NL, Massie CE, Ramos-Montoya A, Zecchini V, Scott HE, Lamb AD, MacArthur S, Stark R, Warren AY, Mills IG, *et al.* 2013. The Androgen Receptor Induces a Distinct Transcriptional Program in Castration-Resistant Prostate Cancer in Man. *Cancer Cell*, **23**(1):35–47.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al.* 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43):15545–50.

Sun L, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, Pastorino S, Passaniti A, Menon J, Walling J, Bailey R, *et al.* 2006. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, **9**(4):287–300.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, *et al.* 2015. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, **43**(D1):D447–D452.

Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, *et al.* 2010. Integrative genomic profiling of human prostate cancer. *Cancer cell*, **18**(1):11–22.

Tetali P 1999. Design of on-line algorithms using hitting times. *SIAM J. Comput.*, **28**(4):1232–1246.

Tew KD, Manevich Y, Grek C, Xiong Y, Uys J, and Townsend DM 2011. The role of glutathione S-transferase P in signaling pathways and S-glutathionylation in cancer. *Free Radical Biology and Medicine*, **51**(2):299–313.

The Cancer Genome Atlas Research Network 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216):1061–8.

The Cancer Genome Atlas Research Network 2011. Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353):609–15.

The Cancer Genome Atlas Research Network 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418):61–70.

The Cancer Genome Atlas Research Network 2015. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**(4):1011–25.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz La, and Kinzler KW 2013. Cancer genome landscapes. *Science (New York, N.Y.)*, **339**(6127):1546–58.

Weigelt B and Downward J 2012. Genomic Determinants of PI3K Pathway Inhibitor Response in Cancer. *Frontiers in Oncology*, **2**(109):1–16.

Yoshihara K, Tajima A, Komata D, Yamamoto T, Kodama S, Fujiwara H, Suzuki M, Onishi Y, Hatae M, Sueyoshi K, *et al.* 2009. Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates ZEB2 in tumor progression and prognosis. *Cancer Science*, **100**(8):1421–1428.

Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, and Verhaak RGW 2014. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, **34**(37):4845–4854.