



HIT'nDRIVE: Patient-specific multi-driver gene prioritization for precision oncology

Raunak Shrestha, Ermin Hodzic, Thomas Sauerwald, et al.

Genome Res. published online July 18, 2017

Access the most recent version at doi:[10.1101/gr.221218.117](https://doi.org/10.1101/gr.221218.117)

P<P Published online July 18, 2017 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

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

Raunak Shrestha^{1,2,*}, Ermin Hodzic^{3,*}, Thomas Sauerwald⁴, Phuong Dao⁵, Kendric Wang², Jake Yeung², Shawn Anderson², Fabio Vandin⁶, Gholamreza Haffari⁷, Colin C. Collins^{2,8}, and S. Cenk Sahinalp^{2,3,9,†}

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

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

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

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

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

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

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

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

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

*Co-first authors

†Correspondence: cenk@sfu.ca

Running Title: Computational prioritization of cancer driver genes

Keywords: Driver gene prediction, random walk facility location, multi-hitting time, driver based phenotype inference, driver module, precision oncology

Corresponding Author:

S. Cenk Sahinalp
School of Computing Science,
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
Email: cenk@sfu.ca

School of Informatics and Computing,
Indiana University,
Bloomington, IN 47405, USA
Email: cenksahi@indiana.edu

Abstract

Prioritizing molecular alterations that act as drivers of cancer remains a crucial bottleneck in therapeutic development. Here we introduce HIT'nDRIVE, a computational method that integrates genomic and transcriptomic data to identify a set of patient-specific, sequence-altered genes, with sufficient collective influence over dysregulated transcripts. HIT'nDRIVE aims to solve the “random walk facility location” (RWFL) problem in a gene (or protein) interaction network, which differs from the standard facility location problem by its use of an alternative distance measure: “multi-hitting time”, the expected length of the shortest random walk from any one of the set of sequence-altered genes to an expression-altered target gene. When applied to 2200 tumors from four major cancer types, HIT'nDRIVE revealed many potentially clinically actionable driver genes. We also demonstrated that it is possible to perform accurate phenotype prediction for tumor samples by only using HIT'nDRIVE seeded driver gene modules from gene interaction networks. In addition, we identified a number of breast cancer subtype-specific driver modules that are associated with patients' survival outcome. Furthermore, HIT'nDRIVE, when applied to a large panel of pan-cancer cell-lines, accurately predicted drug efficacy using the driver genes and their seeded gene modules. Overall, HIT'nDRIVE may help clinicians contextualize massive multi-omics data in therapeutic decision making, enabling widespread implementation of precision oncology.

Introduction

Genomic and transcriptomic alterations are the major contributors of tumorigenesis and progression of cancer. Over the past decade, high-throughput sequencing efforts have provided an unprecedented opportunity to identify such genomic alterations that can lead to changes in gene regulation, protein structure, and function ([Stratton et al. 2009](#)). Genomic and transcriptomic data provide unique and complementary information about a particular tumor, but the translation of “big” molecular data into insightful and impactful patient outcomes is extraordinarily challenging ([Vogelstein et al. 2013](#)). During tumor progression, cancer cells accumulate a multitude of genomic alterations; however most are inconsequential “passenger” alterations that are effectively neutral. Nevertheless, a small fraction provide mission-critical “hallmark” functions and are known as “driver” alterations that modify transcriptional programs and therefore drive and sustain tumor progression ([Stratton et al. 2009](#); [Vogelstein et al. 2013](#); [Greenman et al. 2007](#)). Improving our knowledge on driver alterations, possibly through an integrative analysis of various omics data is critical to better understand cancer mechanisms and select appropriate therapies for specific cancer patients.

There are several computational methods for identifying cancer drivers. However, many of them rely on the

recurrence frequency of single nucleotide variants (SNVs) with respect to the background mutation rate ([Greenman et al. 2006](#); [Youn and Simon 2011](#); [Lawrence et al. 2013](#); [Korthauer and Kendzierski 2015](#)). As a result, these methods are restricted to identifying only highly recurrent mutations as driver events. Recent studies have implicated novel drivers that affect only a small subset of cancer patients. Notable examples include *SPOP* mutations and *CHD1* deletions which are present in less than 20% of prostate cancer patients ([Barbieri et al. 2012](#); [Grasso et al. 2012](#)). Whereas recurrent drivers are hypothesized to initiate carcinogenesis and are therefore present in the majority of tumor cells, rare drivers can arise during tumor evolution and be isolated to a smaller fraction of cells due to clonal expansion ([Greaves and Maley 2012](#); [Ding et al. 2012](#)). These rare driver genes may be functionally important but are likely to be missed by a frequency-based approach.

Perhaps the first computational method to consider large scale genomic alterations as driver events is CONEXIC ([Akavia et al. 2010](#)), which correlates genes with highly recurrent copy number alterations (CNAs) with variation in gene expression profiles within a Bayesian network. Similarly, with no prior knowledge of pathways or protein interactions, MOCA correlates gene mutation information with expression profile changes in other genes ([Masica and Karchin 2011](#)). Suo *et al* ([Suo et al. 2015](#)) also prioritize highly mutated genes that interact with large number of differentially expressed genes in a gene network. Another approach, (Multi) Dendrix ([Leiserson et al. 2013](#)) aims to simultaneously identify multiple driver pathways, assuming mutual exclusivity of mutated genes among patients, using either a Markov chain Monte Carlo algorithm or integer linear programming (ILP). XSEQ ([Ding et al. 2015](#)) uses probabilistic model to compute influence of mutated genes over expression profile changes in other genes by considering direct gene interactions. Finally, MEMo ([Ciriello et al. 2012](#)), identifies sets of proximally-located genes from interaction networks, which are also recurrently altered and exhibit patterns of mutual exclusivity across the patient population.

Simultaneously with the above methods, several approaches were developed outside of cancer research to correlate the presence of casual genes with gene expression. For example, Tu *et al* ([Tu et al. 2006](#)) used a random walk approach on a molecular interaction network to associate causal genes and pathways. Similarly, ResponseNet ([Yeager-Lotem et al. 2009](#); [Lan et al. 2011](#)) relates genetic perturbations to transcriptomic response in yeast model thereby identifying sub-network of regulators mediating the interactions. ResponseNet formulates a minimum-cost flow optimization problem that aims to maximize the flow between the source and target while minimizing the cost of the connecting paths. Similarly, eQTL electrical diagrams (eQED) ([Suthram et al. 2008](#)) integrates expression quantitative trait loci (eQTL) analysis with molecular interaction network using the circuit network model. To the best of our knowledge, NetQTL ([Kim et al. 2011](#)) is the first method to link CNAs to expression profile changes

within an interaction network and connects specific “causal” aberrant genes with potential targets in the interaction network. EPoC ([Jornsten et al. 2011](#)) links CNAs to expression changes in an interaction network assuming steady-state perturbation effects. Similarly, PARADIGM ([Vaske et al. 2010](#)), computes gene-specific inferences using factor graphs to integrate different genomic changes and infer pathways altered in a patient. MAXDRIVER ([Chen et al. 2013](#)) uses maximum information flow to identify potential causal genes (CNAs) in an interaction network.

More recently, HotNet ([Vandin et al. 2011](#)), was the first tool to use a network diffusion approach to compute a pairwise influence measure between the genes in the (protein interaction) network and identify sub-networks enriched for mutations in cancer. TieDIE ([Paull et al. 2013](#)) also uses the diffusion model to identify a collection of pathways and sub-networks that associate a fixed set of driver genes to expression profile changes in other genes. Briefly, the network diffusion approach aims to measure the influence of one node over another by calculating the stationary proportion of a “flow” originating from the starting node that ends up in the destination node. Since this is based on the stationary distribution, the inferences that can be made by the diffusion model are time independent. In that sense, the diffusion approach is very similar to Rooted PageRank, the stationary probability of a random walk originating at a source node, being at a given destination node. Shi *et al* ([Shi et al. 2016](#)) also prioritizes genes based on diffusion score matrix (derived from a tripartite graph of mutations, outliers and patients) rank aggregation. A final method, DriverNet ([Bashashati et al. 2012](#)), also aims to correlate genomic alterations with target genes’ expression profile changes, but only among direct interaction partners; the novel feature of DriverNet is that it aims to find the minimum number of potential drivers that can “cover” targets.

Among the above strategies, the ones based on mutual exclusivity still focus on frequent events. The others, based on “information flow” in gene/protein interaction networks, do not aim to discover cancer drivers, but rather are designed to identify dysregulated sub-networks or modules. In addition, the notion of influence they employ is based on stationary distribution of “information” originating at a particular gene/protein. As a result, none of the available methods aim to identify rare, patient-specific driver events, based on a time dependent notion of influence. Finally, none of the available techniques aim to simultaneously consider different types of genomic alterations as potential drivers.

To address the above challenges, we have developed a novel combinatorial method, HIT’nDRIVE (a preliminary version was presented at the Research in Computational Molecular Biology (RECOMB) conference ([Shrestha et al. 2014](#))). HIT’nDRIVE jointly analyzes genome and transcriptome data for identifying and prioritizing sequence-altered genes as potential cancer drivers. For that HIT’nDRIVE integrates patient-specific genomic alterations

with the associated transcriptome profile, identifying driver genes that dysregulate large portion of each patient's transcriptome. Drawing upon the domain knowledge of molecular interactions presented as a gene/protein interaction network, HIT'nDRIVE uses network topology to derive the influence of one (sequence-altered) gene over another (expression-outlier) gene and aims to identify the most parsimonious set of patient-specific driver genes that have sufficient "influence" over a large proportion of the expression-outliers.

Results

The primary goal of HIT'nDRIVE is to link alterations at the genomic level to changes at transcriptome level through a gene/protein interaction network. Intuitively, it aims to find the *smallest* set of altered genes that can explain most of the observed transcriptional changes in the cohort (see [Methods](#) section for details). In other words, HIT'nDRIVE aims to identify the minimum number of potential driver genes which can cause a user-defined proportion of the downstream expression effects observed. HIT'nDRIVE formulates this as a "Random Walk Facility Location" (RWFL) problem, a combinatorial optimization problem that we introduce in this paper. RWFL generalizes the classical "Facility Location" (FL) problem by changing the notion of distance it uses. Given a network, FL problem defines the distance between a potential driver gene and an outlier gene as the length of the shortest path between them. The RWFL problem, in contrast, uses "hitting time" (or "first passage time") ([Condamin et al. 2007](#); [Liben-Nowell and Kleinberg 2007](#)), the expected length of a random walk between the two nodes, as their distance. Under the use of hitting time, the FL problem completely changes nature: in the classical FL formulation the goal is to associate each outlier gene in the network with exactly one (the closest) driver gene. In the RWFL formulation, each outlier gene is associated with multiple driver genes (whose collective distance to the outlier will no longer be the shortest pairwise distance), forming a many-to-many relation.

As per the standard facility location problem, RWFL is NP-hard, even to approximate. As a result, we reduce it to the weighted multi-set cover (WMSC) problem, for which we give an ILP formulation (see [Methods](#) section for details). Intuitively, in this new formulation, HIT'nDRIVE associates the genomic alterations with transcriptomic changes in the form of a bipartite graph with nodes on one partition representing the set of aberrant genes and nodes on the other partition representing the set of expression-altered genes and each edge has an influence value equal to the inverse pairwise hitting time between the two nodes it connects ([Figure 1A](#)). The WMSC problem on this representation of data asks to find the smallest subset of aberrant genes (as potential drivers) whose total influence (sum of pairwise influence values) over a user defined fraction of expression-altered genes (for each patient) is sufficiently high.

In order to quantitatively assess the genes identified by HIT'nDRIVE, we extended our previously developed

algorithm, OptDis ([Dao et al. 2011](#)), for *de novo* identification of modules of small size inside the interaction network which are seeded by at least one predicted driver. The modules are chosen so that their discriminative power (for phenotype classification) is the greatest among connected sub-networks of similar size that contain the individual predicted driver genes ([Figure 1B-C](#)). We report the classification accuracy based on the identified driver-seeded modules as means of quantitative validation of our results (in the absence of ground truth). We also look at the genes that build the chosen modules (of high classification accuracy) in attempt to identify cancer-related pathways.

We have implemented HIT'nDRIVE in C++ and solved the ILP using IBM CPLEX version 12.5.1. HIT'nDRIVE uses three different user-defined input parameters (see [Methods](#) section for details): (i) α determines the fraction of outliers to be covered overall (across all patients), (ii) β determines the fraction of outliers to be covered in each patient, and (iii) γ controls the fractional lower bound on the sum of the incoming edge weights (influence values). HIT'nDRIVE is robust with respect to the changes in α and β but is somewhat sensitive to the value of γ , 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 too ([Supplemental Fig. S1](#)).

We used STRING v10 ([Szklarczyk et al. 2015](#)) functional protein-interaction network for our analysis. HIT'nDRIVE is not sensitive to topological biases in the network ([Supplemental Fig. S3](#)). Although, the number of driver genes predicted by HIT'nDRIVE differed slightly when different networks were used (and was proportional to the number of nodes in the network), the proportion of overlap between the driver genes predicted on different networks was quite robust ([Supplemental Fig. S4](#), see [Supplemental Results](#) for details).

HIT'nDRIVE predicts frequent as well as infrequent driver genes in multi-omics cancer datasets

We applied HIT'nDRIVE to prioritize driver genes in 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](#)) obtained from the The Cancer Genome Atlas (TCGA) data portal. Only samples with matched genomic alterations (SNVs and/or CNAs and/or gene fusions) and transcriptomic changes (outlier genes from gene-expression profile) were used in our study. We used the fusion prediction calls as reported in the TCGA Fusion gene Data Portal ([Yoshihara et al. 2014](#)).

In GBM, we obtained 48 unique candidate driver genes altered at varying frequencies across 258 GBM patients ([Supplemental Fig. S8-S9](#), [Supplemental Table S1,S5](#)). *EGFR* (36%), *TP53* (29.5%), *PTEN* (28%) and *CHEK2* (26%) were the most frequently altered driver genes in GBM followed by *CDKN2A* (16%), *RBI* (13%), *SEC61G*

(12%). Previous efforts in GBM genome characterization identified amplification in *EGFR*, *PDGFRA*, mutations in *CHEK2*, *TP53*, *PTEN*, *RB1*, *NF1* and deletions in *CDKN2A* to be associated with GBM ([Parsons et al. 2008](#); [The Cancer Genome Atlas Research Network 2008](#); [Verhaak et al. 2010](#)). HIT'nDRIVE prioritized all of the above alterations. Alterations in *EGFR* is characteristic of classical subtype, *NF1* with mesenchymal subtype, *PDGFRA* and *IDH1* with pro-neural subtype of GBM ([Verhaak et al. 2010](#)). Fifteen out of 48 driver genes predicted by HIT'nDRIVE ($p\text{value} = 8 \times 10^{-4}$), were present in Cancer Gene Census (CGC) database ([Futreal et al. 2004](#)), that contains genes for which mutations have been causally implicated in cancer ([Figure 2A](#)). *GSTT1* (deleted in 21 patients), a key player in drug metabolism, was neither found in CGC nor in Catalogue of Somatic Mutations in Cancer (COSMIC) ([Forbes et al. 2017](#)) databases. Twelve GBM driver genes were found to be actionable targets. Actionable genes were extracted from TARGET database ([Van Allen et al. 2014](#)), which contains genes directly linked to a clinical action. In addition to the above list, 6 other driver genes were druggable ([Figure 2B](#)). We extracted the list of druggable genes from Drug-Gene Interaction Database (DGIdb) ([Griffith et al. 2013](#)). Interestingly, around 85% of the patients in GBM cohort harbour at least one actionable driver gene and further 5% of patients have druggable targets ([Figure 2C](#)). HIT'nDRIVE also identified 12 infrequent driver genes, which we define as genes altered in at most 2% of the cases. Among the infrequent genes, *SACS* is known to be associated with neurological functions, *NLRP3* is involved in apoptosis, and *TIAM2* is involved in invasion and metastasis.

The 526 OV patients harboured a total of 85 unique driver alterations ([Supplemental Fig. S10-S11](#), [Supplemental Table S2,S6](#)). *TP53* mutations were prevalent in more than half (58%) of the patients in the cohort. Consistent with the previous findings, we found OV patients to be driven by genomic copy-number changes rather than recurrent point mutations ([Ciriello et al. 2013](#); [Patch et al. 2015](#)). Recurrent somatic CNAs were observed in *GSTT1* (32.3%), *WWOX* (28.1%), *FAM49B* (15.0%), *UGT2B17* (14.6%), *CCNE1* (13.1%), *SLC39A4* (13.1%) and *MYC* (12.5%). Mutations in *TP53*, *BRCA1/2* and loss of *RB1*, *NF1* and *CCNE1* were previously associated with OV ([The Cancer Genome Atlas Research Network 2011](#); [Patch et al. 2015](#)). HIT'nDRIVE revealed 18 CGC driver genes ($p\text{value} = 2 \times 10^{-5}$) ([Figure 2A](#)) among which 13 genes were actionable targets and other 12 genes were at least druggable ([Figure 2B](#)). More than 75% of OV patients harboured at least one actionable targets and additional 6% of patients have druggable target ([Figure 2C](#)). *GSTT1* (altered in 170 patients), in OV, is involved in estrogen and drug metabolism. It was neither found in CGC nor in COSMIC databases. We identified 13 infrequent genes, among which *MAPK1* is known to play an important role in oncogenic pathways in cancer.

HIT'nDRIVE identified 40 driver genes across 333 PRAD patients ([Supplemental Fig. S12-S13](#), [Supplemental Table S3,S7](#)). Copy number loss of *SPECC1L* (23.7%), *STEAP1B* (13%), *WWOX* (10%) and amplification of *NSD1*

(16.2%), *SIRPB1* (16.2%) were the most recurrent events in PRAD patients. We also found recurrent somatic mutation in *MUC4* (11%), *SPOP* (10.5%) and *TP53* (10%). The most common alterations in PRAD genomes are fusion of androgen-regulated promoters with *ERG* and other members of ETS family of transcription factors mainly, TMPRSS2-*ERG* fusions (Tomlins et al. 2005). Since we relied on the gene fusion predictions obtained from TCGA Fusion gene Data Portal (Yoshihara et al. 2014) which analyzed only 178 (out of 333) patients, we observed *ERG* gene fusion in only 5.7% cases. The more recent TCGA publication (The Cancer Genome Atlas Research Network 2015) reported *ERG* fusions in almost half of the patients in the cohort. Moreover, the tools used for gene fusion detection, in the two studies, were different as a result of which we observed much smaller number of *ERG* fusions than reported previously. *SPOP*, *TP53*, *FOXA1* and *PTEN* are the most frequently mutated genes which have been previously associated with prostate cancer (Barbieri et al. 2012). PRAD patients harboured 12 driver genes present in CGC database ($p\text{value} = 9 \times 10^{-4}$) (Figure 2A) out of which 8 driver genes were actionable (Figure 2B). Approximately a quarter of PRAD patients could benefit with actionable targeted therapy (Figure 2C). Moreover an additional 14% of patients harboured druggable genes which warrants deeper investigation of drug repurposing opportunities. *NBPF1* (mutated in 17 patients), is a known tumor suppressor gene known to have neural function and also involved in cell-cycle arrest, was neither found in CGC nor in COSMIC databases. We identified 11 infrequent genes in PRAD among which *IDH1* mutant patients were recently identified as a distinct molecular-subtype of PRAD (The Cancer Genome Atlas Research Network 2015), *NKX3-1* is required for normal prostate tissue development and *CDKN1B* was previously associated with PRAD.

In BRCA, HIT'nDRIVE identified 107 driver genes across 1090 patients (Supplemental Fig. S14-S15, Supplemental Table S4,S8). Somatic mutation of *PIK3CA* (30.5%) and *TP53* (30.2%) were the most recurrent events in BRCA. This was followed by somatic mutation of *CHD1* (11.2%), *GATA3* (10.5%), *MUC16* (6.9%), *MAP3K1* (6.9%) and CNA amplification of *NSD1* (8.7%) and *MED1* (6.9%). BRCA patients harboured 16 genes present in CGC database ($p\text{value} = 9.3 \times 10^{-3}$) (Figure 2A) among which 10 genes were actionable targets (Figure 2B). More than 60% of BRCA patients could benefit with the actionable targeted therapy. Furthermore, additional 11% of BRCA patients harboured at least one of the 19 potentially druggable genes (Figure 2C). *ACACA* (altered in 36 patients mostly from HER2 subtype), involved in fatty-acid metabolism, was neither found in CGC nor in COSMIC databases. We identified 46 infrequent driver genes among which *BRCA2* and *GNAS* have been previously linked to BRCA.

Although the driver events per tumor sample greatly varied, the median number of driver genes among the 2207 tumor samples in all four cancer types was three (Supplemental Fig. S16). Twenty-three percent of 2207 tumor

samples harboured just a single driver gene. Remaining 77% of tumor samples harboured two or more driver events which may either indicate the existence of multiple sub-clonal populations within the tumor, or the presence of collaboration among multiple sequence altered genes in an oncogenic pathway.

To evaluate the sensitivity of HIT'nDRIVE to infrequent driver genes, we performed an *in-silico* experiment using 1000 TCGA-BRCA tumors (referred here as the “original set”). Different subsets of tumor samples with varying sample-size were chosen such that the sub-sample tumor population has similar alteration-frequency distribution to that of the original 1000 tumor samples ([Supplemental Fig. S21A](#), see [Supplemental Methods](#)). HIT'nDRIVE analysis was performed on the chosen subsets of tumor samples independently to identify driver genes and then compared the frequency of driver genes detected. HIT'nDRIVE detected driver genes that were prevalent in just a single patient tumor with a sample size of 15, 25 or 50 tumors representing 6.5, 4, and 2% of the tumor population respectively ([Supplemental Fig. S21B-C](#)). Even when the sample-size was increased to 700-900 samples, HIT'nDRIVE was able to detect driver genes prevalent in just 4 patients representing < 0.5% of the tumor population. This demonstrates that HIT'nDRIVE prioritizes driver genes independent of the selected sample-size and HIT'nDRIVE is very sensitive in detecting infrequent driver genes.

Network properties of cancer driver genes

Centrality of Driver Genes in the Interactome. Cancer driver genes are known to occupy critical positions in the interactome. To check whether HIT'nDRIVE predicted driver genes also occupy similar positions in the interaction network, we used the node degree as a “local measure”, and node betweenness (the number of shortest paths between node pairs that pass through the node) as a “global measure” of centrality. The driver genes predicted by HIT'nDRIVE include a number of well-known high-degree hubs - *TP53*, *EGFR*, *RBI*, *MYC*, *PIK3CA*, *ERG*, *CHD1* that are “central” in the interactome with high degree and high betweenness ([Figure 3A](#)). Although there was very weak correlation between the number of edges (i.e. degree centrality) of a node and the number of samples/patients in which it is identified as a driver, remarkably, each hub gene was typically altered in a large fraction of patients. Because of their centrality perturbations, hub genes are likely to dysregulate several other genes and the associated signaling pathways. Interestingly, HIT'nDRIVE also identified low-degree genes (*IDH1*, *MTAP*, *NF1*, *NRG1*, *NSD1*) that reside in the periphery of the interaction network. In particular, in prostate cancer, there seems to be an inverse correlation between the degree and how often the gene is picked as a driver. Most of these low-degree genes are altered in a small fraction of patients, indicating that HIT'nDRIVE, unlike many other methods, does not primarily return hubs that are altered in a large number of patients but is capable of identifying rare driver genes without trivial topological biases.

As discussed in previous section, we used hitting time to compute the influence of a node in an interaction network. The influence from a source to a target node depends on the topological position of the target node in the network. We observed that the nodes occupying central positions in the network, i.e. with high betweenness centrality, tend to receive more influence than the nodes in the periphery of the network (Pearson correlation coefficient (R) = 0.61) (Figure 3B). This is because the distance between any source node and a central target node (i.e. a hub) is usually very short, implying a low hitting time, and thus high influence of the source node on the target. We also observed negative correlation between a node's total incoming influence and the median outgoing influence (R = -0.54) (Figure 3C). Although, the central hub nodes (for example - *UBC*, *TP53*) are good receivers of influence, when individual influences are considered, they do not contribute a lot.

Influential nodes prioritized as cancer driver genes. Next we examined the influential driver genes that are responsible for driving cancer. For this, we computed the total outgoing influence from each altered gene (which has been chosen as a driver), defined as the weighted sum of all influence values from the source to all outlier genes it is connected to (targets), weighted by the corresponding outlier weights. First we investigated driver genes with high influence values within each cancer type. We observed that on average the total influence of driver genes was higher than that of other altered genes in all cancer types (Figure 3D). *EGFR*, *PTEN*, *CHEK2*, *TP53* and *CDKN2A* were the most influential driver genes in GBM which together exerted 38.5% of the total influence on the GBM patient cohort. In OV, *TP53*, *GSTT1* and *MYC* together exerted 20% of the total influence. Similarly, in PRAD cohort, *SPOP*, *MUC4* and *TP53* were the most influential genes exerting 23.7% of the total influence. *PIK3CA*, *TP53* and *CHD1* were the most influential genes exerting 23% of the total influence on the BRCA patient cohort. Moreover, the gene influence was positively correlated to its alteration frequency (Figure 3E).

We investigated influence of the predicted driver genes within individual patients. Many recurrently altered driver genes had higher influence compared to other driver genes. For example, *EGFR* in GBM; *TP53* in OV; *ERG* in PRAD; *TP53*, *PIK3CA* and *PTEN* in BRCA.

Interestingly, among the highly influential genes there were also less-recurrent but functionally important and actionable driver genes. For example, somatic mutations in ATP binding cassette subfamily B member 1 (*ABCB1*) were influential driver genes in seven GBM patients (Figure 3F). *ABCB1* is a membrane-bound protein present in the endothelial cells of the blood-brain barrier. It harnesses the energy of ATP hydrolysis to drive the unidirectional transport of exogenous and xenobiotic substances (drug compounds) from the cytoplasm to the extracellular space. It is known to transport many anticancer compounds including temozolomide (TMZ), which is used as a first-line treatment for GBM patients. Mutations and over-expression of *ABCB1* in GBM have been associated with resistance

to TMZ ([Lin et al. 2014](#)). It was intriguing that some of these GBM patients had undergone treatment prior to tissue collection and were initially mislabelled as untreated patients. Treatment-induced selection pressure in the drug transporter might be a plausible reason for high influence exerted by *ABCB1*.

Similarly, HIT'nDRIVE predicted *BRAF* as driver genes in eight PRAD patients (6 somatic mutations and 2 gene-fusions) ([Figure 3G](#)). These patients harboured *BRAF* as a highly influential driver gene. None of these patients harboured *BRAF*^{V600E} mutation that is prevalent in cutaneous melanomas, thyroid cancer and many other cancer types. However, *BRAF*^{L597R} can be targeted using MEK inhibitors ([Dahlman et al. 2012](#); [Bowyer et al. 2014](#)). BRAF plays important roles in growth factor signalling pathways, which affects cell division and differentiation. These results serve as proof of concept that HIT'nDRIVE can prioritize functionally relevant cancer driver genes.

Phenotype classification using dysregulated modules seeded with the predicted driver genes

Evaluating computational methods for predicting cancer driver genes is challenging in the absence of the ground truth (i.e. follow-up biological experiments). Therefore, we mainly focused on testing whether our predictions provide insight into the cancer phenotype and improve classification accuracy on an independent cancer dataset. To test association of the driver genes identified by HIT'nDRIVE with the cancer phenotype, as explained in the earlier section, we used the driver gene seeded gene-modules, a set of functionally related genes (e.g. in a signaling pathway), from the protein interaction network, as features for classifying the cancer phenotype. Using OptDis (here referred to as HIT'nDRIVE-OptDis), we identified small connected sub-networks that include (i.e. are seeded by) predicted driver genes in a greedy fashion. More specifically, we prioritized sub-networks (of at most seven genes) iteratively so that in each iteration we identified the sub-networks that maximally discriminates sample phenotypes in a gene-expression matrix, among the sub-networks that share very few genes (at most 20%) with the sub-networks already prioritized.

Furthermore, we have also developed an unsupervised method for module identification (here referred to as HIT'nDRIVE-unsupervised), i.e. one that does not depend on any phenotype information. This unsupervised method seeds each module with one HIT'nDRIVE identified driver gene, and includes outlier genes that it has influence over and co-occurs with significantly across patients ([Supplemental Fig. S22](#)). For this, we perform a hypergeometric test to identify significant driver-outlier interaction (i.e. mutual presence) pairs across the patient cohort ($p\text{value} < 10^{-3}$).

Here we compare HIT'nDRIVE-OptDis and HIT'nDRIVE-unsupervised to another network based driver genes prioritization method - DriverNet ([Bashashati et al. 2012](#)). DriverNet itself does not aim to identify modules that we can use to compare against HIT'nDRIVE-OptDis or HIT'nDRIVE-unsupervised modules. Rather, DriverNet identifies driver genes in an iterative fashion, where in each iteration, DriverNet picks the driver genes which “covers”

the maximum number of uncovered outliers. We use this driver and the outlier genes it covers as the “next” DriverNet module.

We used the set of prioritized sub-networks, i.e. the driver modules, first, to perform binary sample classification: tumor vs normal. For this, we used gene-expression data for each of the four cancer types (GBM, OV, PRAD and BRCA) from TCGA as discovery datasets to calculate the mean gene expression value for each sub-network/driver module, for each patient ([Supplemental Table S10-S13](#)). On these sub-networks, we used the k-nearest neighbour (KNN) classifier (with $k = 1$), to perform classification on both the expression values from TCGA, and additional validation gene-expression datasets ([Fig. S4A-C](#), [Supplemental Table S9](#)). The additional validation datasets were used in order to assess the capability of the modules identified on TCGA cohort, in classifying other cohorts.

For every dataset analyzed, the maximum classification accuracy achieved by HIT’nDRIVE modules (either HIT’nDRIVE-unsupervised or HIT’nDRIVE-OptDis), for any number of modules considered, was higher than that achieved by DriverNet modules ([Figure 4A](#)). Moreover, in most datasets, HIT’nDRIVE methods achieve maximum or near-maximum accuracy using a smaller fraction of modules ([Supplemental Table S14](#)). All three methods achieved perfect or near perfect classification accuracy in TCGA-GBM, TCGA-OV and TCGA-BRCA datasets except for TCGA-PRAD dataset (where the maximum classification accuracy achieved was 90% by HIT’nDRIVE-Unsupervised, 95% by HIT’nDRIVE-OptDis and 86% by DriverNet). Overall, the driver modules (identified in one cohort) were able to distinguish the tumor phenotype from normal very well in validation datasets (on other cohorts) supporting the relevance of the identified driver genes to the cancer phenotype.

Lastly, we compared the classification accuracy of randomly formed modules in the TCGA-PRAD dataset, (which is the single most challenging dataset) compared against modules identified using HIT’nDRIVE-OptDis. To generate the random modules, 60 random sets of genes, of random size (up to 7 genes), were selected 20000 times. Phenotype classification was then performed using KNN ($k = 1$) classifier. HIT’nDRIVE-OptDis modules demonstrated superior classification accuracy compared to randomly selected modules ([Supplemental Fig. S25](#)).

CGC cancer type-specific gene enrichment. Next, we looked into the list of prioritized driver genes by both HIT’nDRIVE and DriverNet and their overlap with the known CGC genes ([Figure 4B](#)). DriverNet selects a much larger number of driver genes, as compared to HIT’nDRIVE, to cover most outlier genes (across all four cancer types) due to its model considering only direct interactions in the network. In particular, in OV and BRCA, the number of HIT’nDRIVE identified driver genes are an order of magnitude smaller than that of DriverNet. Although in GBM and PRAD datasets, the number of driver genes identified by DriverNet is somewhat lower and comparable to that identified by HIT’nDRIVE (primarily because most outliers were filtered out due to sharing no

interaction edge with candidate altered genes), HIT'nDRIVE identified driver genes cover a significantly larger number of outliers. More importantly, even though HIT'nDRIVE identifies a smaller number of driver genes, a larger fraction of these driver genes can be found in CGC database - in comparison to the DriverNet identified driver genes. In fact, even a larger fraction of CGC genes specific to the relevant cancer type can be found among HIT'nDRIVE identified driver genes. Specifically, HIT'nDRIVE predicted four glioblastoma specific CGC genes (*IDH1*, *PDGFRA*, *PIK3CA* and *PIK3R1*) in TCGA-GBM dataset. Among them, *IDH1*, *PDGFRA* and *PIK3CA* were not identified by DriverNet. Similarly, four ovarian cancer specific CGC genes (*BRCA1*, *BRCA2*, *CCNE1* and *MAPK1*) were predicted in TCGA-OV dataset. *CCNE1* was not identified by DriverNet. Five prostate cancer specific CGC genes (*BRAF*, *ERG*, *FOXA1*, *PTEN* and *SPOP*) were predicted in TCGA-PRAD dataset. *BRAF* and *SPOP* were not identified by DriverNet. And seven breast cancer specific CGC genes (*BRCA2*, *CCND1*, *CDH1*, *GATA3*, *MAP3K1*, *PIK3CA* and *TP53*) were predicted in TCGA-BRCA dataset. Among them, *CDH1* and *MAP3K1* were not identified by DriverNet.

Breast cancer subtype classification using driver modules. Our next goal was to classify four major subtypes of breast cancer - Basal, HER2, Luminal-A and Luminal-B. For that purpose, we performed binary classification for each subtype: e.g. Basal vs non-Basal (including the normal samples). This was achieved through the use of HIT'nDRIVE-identified driver genes from TCGA-BRCA as seed genes, with which we identified subtype-specific driver modules from TCGA-BRCA gene-expression data (as described for tumor classification). We respectively obtained 37, 16, 43 and 39 subtype-specific driver modules for Basal, HER2, Luminal-A and Luminal-B subtypes (Supplemental Fig. S27-29, Supplemental Table S15). As described above, using these sub-type specific driver modules as features, we performed independent classification of BRCA subtypes in TCGA-BRCA, METABRIC-Cambridge and METABRIC-Vancouver datasets (Curtis et al. 2012).

Majority of Basal-like tumors constitute Triple-Negative Breast Cancer (TNBC), which are highly aggressive tumors characterized by lack of expression of estrogen receptor 1 (*ESR1*), progesterone receptor (*PGR*) and erb-b2 receptor tyrosine kinase 2 (*ERBB2*). Molecular mechanisms driving TNBC are least understood and hence, no targeted therapies for TNBC yet exists (Bianchini et al. 2016). Interestingly, HIT'nDRIVE seeded driver modules were able to classify Basal-like tumors with much higher accuracy (98%) as compared to other BRCA-subtypes - HER2 (94%), Luminal-A (85%) and Luminal-B (83%) (Figure 5A, Supplemental Table S16). As expected, *ESR1* and *PGR* was highly expressed in Luminal-A/B but not in Basal and HER2 subtypes. Modules containing *ESR1* were consistently down-regulated in Basal subtype and up-regulated in Luminal-A/B subtype whereas module LUMB-03 was up-regulated in Luminal-B subtype (Supplemental Fig. S31). The *ESR1* network neighbourhood

included eleven known transcriptional targets of *ESR1* (*TFF1*, *PGR*, *SLC9A3R1*, *GNAS*, *RARA*, *WWP1*, *WNT5A*, *TCF7L2*, *FKBP4*, *SPRY2*, and *RAD54B*). These results were consistent with previous findings (Dutta et al. 2012). *ERBB2* was expressed only in 9 (of 16) HER2 modules and was the most prominent hub in the large interactome of HER2 modules. All modules containing *ERBB2* were up-regulated in HER2 subtype and module expression pattern were consistent in different BRCA datasets (Supplemental Fig. S32). *PGR* was present in 2 modules (BASAL-26 and HER2-12) both of which were down-regulated in Basal subtype but up-regulated in Luminal-A/B. These results strongly suggest that HIT'nDRIVE can capture subtype-specific driver genes, and the driver-seeded modules we identified can indeed differentiate BRCA subtypes.

Subtype-specific breast cancer driver modules are associated with survival outcome. To test for association of subtype-specific driver modules with patient survival outcome, we developed a risk-score defined as a linear combination of the normalized gene-expression values of the component genes in the module weighted by their estimated univariate Cox proportional-hazard regression coefficients (see Methods section). Based on the risk-score values, patients were stratified into low-risk (risk-score < 33 percentile) and high-risk (risk-score > 66 percentile) groups. 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 cohorts (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.

We first compared driver-seeded modules against driver-gene-free modules that, according to OptDis, have the best discriminative score for the TCGA-BRCA dataset. For each module we calculated three distinct indices: log-rank test pvalue, hazard-ratio (HR) and concordance-index (c-index). We found driver-seeded modules to outperform driver-free modules on all three indices demonstrating that the driver-seeded modules were better correlated with survival (Figure 5B). Motivated by this, we identified the top modules for each of the BRCA subtypes which do well based on all three indices and checked whether they can return meaningful results with respect to survival. We found 9 driver modules significantly associated with patients' survival outcome (p-value < 0.01, hazard-ratio > 1.5 and concordance-index > 0.5) in TCGA-BRCA cohort (Supplemental Table S17). These 9 modules were also significantly associated with patient survival outcome (p-value < 0.01) in two additional cohorts (METABRIC cohorts) (Supplemental Fig. S34-S41). It is interesting to note that two of these modules (BASAL-02 and HER2-01) were seeded by an oncogene - nuclear receptor coactivator 3 (*NCOA3*) driver gene. *NCOA3* driver module was the second-topmost module (Figure 5C) to separate Basal from other subtypes and the top-most module (Supplemental Fig. S34) to separate HER2 subtype. *NCOA3* driver module was down-regulated in Basal subtype and

associated with patients' overall survival (Figure 5D-E). A fraction of breast (and ovarian) cancer patients are known to harbour *NCOA3* mutation, amplification or deletion (Gupta et al. 2016). *NCOA3* alone cannot distinguish the basal subtype. *NCOA3* requires other component genes in the module (*AR*, *XBPI*, *TFF1* and *SPDEF*) to collectively distinguish the basal subtype (Supplemental Fig. S33) which, as per our knowledge, is a novel finding. However, the interaction within the module are well known. *NCOA3* is a coactivator of steroid hormone receptor, *AR* and *ESR1*, and transcriptional target of *XBPI* (Gupta et al. 2016). *NCOA3* is known to stimulate many intracellular signaling pathways that are critical for cancer proliferation and metastasis. The activity of *NCOA3* is known to be associated with reduced responsiveness to tamoxifen in patients (Osborne et al. 2003). *SPDEF* is associated with regulation of *AR* activity (Lehmann et al. 2011).

HIT'nDRIVE seeded driver genes accurately predict drug efficacy

Next, we obtained somatic mutation, copy number aberration and gene expression data of pan-cancer cell lines from Genomics of Drug Sensitivity in Cancer (GDSC) project (Iorio et al. 2016). We used HIT'nDRIVE (Supplemental Fig. S46) to identify driver genes of individual cancer cell lines (Supplemental Table S18). Following up on the premise by Iorio et al. (2016) that potential driver genes (i.e. cancer genes, which include the CGC genes) alone could predict drug efficacy fairly well, the predicted driver genes were used as seeds in the network (STRING v10) to identify sub-networks that discriminate between the drug-response phenotypes (i.e. sensitive vs resistant cell lines). As available in GDSC, 265 different drug treatments were tested on each cell line provided. We present results for 25 cancer types (the remaining 5 cancer types for which only a very limited number of cell lines are available are statistically insignificant and thus have not been used).

Perhaps our most interesting result is that, for many drugs, the top HIT'nDRIVE predicted driver module for a specific cancer type (more specifically, OptDis modules seeded by HIT'nDRIVE identified driver genes, prioritized with respect to drug efficacy) not only includes the drug target but also the associated (downstream) signaling pathway. As importantly, we measured the accuracy of drug-response phenotype classification using HIT'nDRIVE-OptDis for each drug-treatment in different cancer types (Figure 6A, Supplemental Table S19). In most cancer types, HIT'nDRIVE-OptDis correctly predicted the response to more than 25% of the drugs in 95% of the patients or more. Specifically, Stomach adenocarcinoma (STAD) and Chronic Myelogenous Leukemia (LCML) are the cancer types with highest fraction of drugs predicted with an accuracy of 95% or more whereas Liver hepatocellular carcinoma (LIHC) and GBM are the cancer types with the lowest fraction of drugs predicted with the same accuracy. Below we provide some of our observations on three well known/promising cancer drugs for which we obtained high accuracy on specific cancer types.

Gefitinib is a clinically approved (for patients with non-small cell lung cancer) protein kinase inhibitor which selectively inhibits EGFR. Interestingly, in BRCA, *EGFR* copy-number amplification or overexpression primarily activates RAS-RAF-MAPK pathway and PI3K-AKT-mTOR pathway triggering response for cell proliferation, invasion and survival. Using HIT'nDRIVE, *EGFR* was found as a driver gene of BRCA cell lines. Furthermore, *EGFR* seeded driver module was the second highest scoring module to distinguish the drug-response phenotype increasing the classification accuracy to 98% (Figure 6B-C).

Another example, Nutlin-3a is a promising pre-clinical stage compound which inhibits the interaction between MDM2 and TP53 inducing apoptosis. *MDM2* was predicted as a driver gene in OV cell lines by HIT'nDRIVE. *MDM2* seeded module was the top predictor (maximum accuracy 94%) of the drug-response phenotype when treated with Nutlin-3a (Figure 6B,E). Our method predicted many other interacting partners (both as seed or component genes in the module) of *MDM2* and *TP53* which are known to play a critical role in TP53 pathway.

Finally, TMZ is a clinically approved first-line therapy for GBM. ABC transporters (including ABCB1) help to transport TMZ from the extracellular space to the cytoplasm of a cell. TMZ methylates selective nucleotides of DNA triggering DNA repair pathway. MGMT specifically removes the methyl groups from the methylated nucleotides escaping from DNA strand breaks. *MGMT* was predicted as a component gene in the third top-scoring module. Failure to repair DNA strand breaks triggers DNA damage response pathway further activating *TP53* and apoptosis. Interestingly, *TP53* was predicted as the seed of the top scoring module by HIT'nDRIVE-OptDis. Furthermore, another gene in the DNA damage response pathway, *CDKN2A*, seeds another top ranking module, which improves the overall classification accuracy to 97% (Figure 6B,D). Note that both *CDKN2A* and *TP53* are the most frequently altered genes in GBM.

Discussion

In recent years, there has been an unprecedented increase in the multi-dimensional high-throughput data profiling (especially genome and transcriptome) of cancer patients. This has revealed extensive mutational heterogeneity observed in the cancer (sub)types, yielding a long-tailed distribution of mutated genes across the patients, implying the existence of many rare/private driver genes. Thus, there is a great need for computational methods to mine these massive datasets and prioritize clinically actionable driver events to aid treatment modalities using precision oncology.

Here, we have presented a network-based combinatorial method, HIT'nDRIVE, which models the collective effects of sequence altered genes on expression altered genes. HIT'nDRIVE aims to solve the “random-walk facility location” (RWFL) problem on a gene/protein interaction network – which differs from the standard facility location

problem by its use of “hitting time”, the expected minimum number of hops in a random-walk originating from any sequence altered gene (i.e. a potential driver) to reach an expression altered gene, as a distance measure. We introduced the notion of “multi-hitting time” and presented efficient and accurate methods to estimate it based on single-source hitting time in large-scale networks. HIT’nDRIVE reduces RWFL (with multi-hitting time as the distance) to a weighted multi-set cover problem, which it formulates and solves as an ILP.

As a measure of influence, hitting time - the expected length of a random walk between two nodes, or its general version, the multi-hitting time, is quite different from the diffusion-based measures or Rooted PageRank, which are based on asymptotic distributions. We argue that hitting time is a better measure for our purposes as it is: (i) parameter free (diffusion model introduces at least one additional parameter - the proportion of incoming flow “consumed” at a node in each time step), (ii) it is time dependent (while the diffusion model and PageRank measures the stationary behavior) and (iii) it is more robust with respect to small perturbations in the network ([Hopcroft and Sheldon 2007](#)).

In this paper we have demonstrated that (1) HIT’nDRIVE increases our ability to identify potential genomic driver alterations. (2) HIT’nDRIVE prioritizes clinically actionable driver genes many of which happen to be private drivers. This implies that it is possible to replicate the lengthy and costly experimental approaches for detecting driver genes in common tumor types by HIT’nDRIVE - *in-silico*, strongly supporting the biological relevance of HIT’nDRIVE’s algorithmic framework. The fact that a high portion of HIT’nDRIVE prioritized drivers in well studied cancer types overlap with known driver genes increases our confidence in the calls made by HIT’nDRIVE in rarer tumor types for which driver genes are mostly unknown. In fact, the initial results of the Pan Cancer Analysis of Whole Genomes (PCAWG) project reveal that more than 20% of tumors do not have a single (genomically altered) driver gene from CGC (PCAWG, unpublished). HIT’nDRIVE is thus being used for the analysis of PCAWG data to reduce this gap. Results on new driver genes identified by HIT’nDRIVE, especially in rare tumor types will be made available by PCAWG. (3) HIT’nDRIVE prioritizes driver genes present in both the centre and periphery of an interaction network. (4) Our analysis revealed that driver genes have higher collective influence on the transcriptome than other altered genes. Some of these driver genes are central and naturally have high influence, however there are also many non-central driver genes with high influence over other genes in the network. (5) HIT’nDRIVE is especially suitable for identifying such non-central driver genes or infrequent/private drivers. (6) We demonstrated that it is also possible to perform accurate phenotype prediction for tumor samples by only using HIT’nDRIVE implied driver genes and their “network modules of influence” (small sub-networks involving each driver gene where the aggregate expression profile correlates well with the cancer phenotype) as features, providing additional evidence

that these genes may be driving the cancer phenotype. The network modules we identified may provide new insights into the biological mechanisms underlying tumor progression. (7) HIT'nDRIVE can capture subtype specific driver genes and such driver seeded modules can indeed differentiate between different subtypes of a cancer. (8) We have demonstrated that subtype specific driver modules are also associated with patients' survival outcome providing additional evidence that these driver genes have clinical significance. (9) We also demonstrated that HIT'nDRIVE seeded driver genes (more specifically, OptDis modules seeded by HIT'nDRIVE identified driver genes, prioritized with respect to drug efficacy) not only include the drug target but also the associated (downstream) signaling pathway. This provides us the possibility of identifying and clinically targeting multiple genes (not necessarily sequence-wise altered but are nevertheless in the module identified by HIT'nDRIVE) dysregulating critical oncogenic or metabolic pathways.

We also note that targeted therapeutics are being extensively used in clinical trials but the drug response rate is very poor (only ~5% of patients in clinical trials have good response to targeted therapeutics) (Prasad 2016). This is most likely because even if a cancer patient harbours an alteration for which targeted therapeutics are available, we do not know if that alteration is responsible for driving the tumor (Beltran et al. 2015). HIT'nDRIVE could potentially play a key role by prioritizing potential driver alterations from a vast pool of passenger alterations. In our study, we have used drug efficacy data from pan-cancer cell lines in order to demonstrate that the potential genomic drivers (more precisely driver gene seeded modules) of the cell-lines can be used as features to predict drug-efficacy. Following similar procedure in clinical trials, we believe that the application of HIT'nDRIVE to predict drug efficacy would likely improve the drug response rate.

HIT'nDRIVE predicted *ABCB1* as the most influential driver gene in seven TCGA-GBM cases that were treated with TMZ prior to tissue collection. Using GDSC dataset, we demonstrated that HIT'nDRIVE-OptDis can predict mechanisms of drug sensitivity for TMZ and other drugs (Figure 6B-E). Since *ABCB1* was not mutated in any of the GBM cell lines in the analysis, it was not identified as a driver gene of GBM cell lines. However, the top seed driver gene, TP53, is an interaction partner of *ABCB1* (in STRING v10 network). Other seed driver genes and its component genes in the module that are direct interaction partners of *ABCB1* are UBC, CAV1, WDTC1 and DNAH8. ABC transporters (including *ABCB1*) helps to transport TMZ from the extracellular space to the cytoplasm of a cell. On the other hand, DNA damage caused by TMZ activates TP53 thereby dysregulating apoptotic pathways. Thus, the presented analysis demonstrates that the downstream expression changes are, most likely, the manifestation of the selection pressure in *ABCB1* induced by TMZ treatment.

Protein-protein interaction (PPI) networks representing physical interactions now include thousands of proteins

and over a million (undirected) interactions between them. Regulatory networks on the other hand represent gene/protein regulation occurring at multiple levels of biological systems through directed links. Since available regulatory networks are very limited in size and scope, our study focuses on PPI networks. However, HIT'nDRIVE can easily be applied to regulatory networks as they grow in size and scope. In addition, the use of multi-hitting time as a distance measure between two or more driver genes and a target gene enables HIT'nDRIVE to capture synthetic rescue like scenarios; this is ideally suited for undirected PPI networks, but in principle can be extended to regulatory networks in the future.

HIT'nDRIVE is a driver gene prioritization tool that is flexible enough to incorporate different types of -omics data. Both principles under RWFL and HIT'nDRIVE can be utilized to identify the causal genes in different complex disease facing analogous problems to cancer. Finally, we believe that applications of RWFL problem may extend beyond its application to driver gene identification - to influence analysis in social networks, disease networks and others.

Methods

An overview of the HIT'nDRIVE algorithmic framework

HIT'nDRIVE links alterations at the genomic level to changes at transcriptome level through a gene/protein interaction network. More specifically, HIT'nDRIVE identifies the minimum number of potential driver genes which can cause a user-defined proportion of the downstream expression effects observed. We formulate this as a “Random Walk Facility Location” (RWFL) problem, a new combinatorial optimization problem which generalizes the classical “Facility Location” (FL) problem by the use of a novel distance measure. Given a network, FL problem defines the distance between a potential driver gene and an outlier gene as the length of the shortest path between them. The RWFL problem, in contrast, uses “hitting time” (or “first passage time”) ([Condamin et al. 2007](#); [Liben-Nowell and Kleinberg 2007](#)), the expected length of a random walk between the two nodes, as their distance. Under the use of hitting time, the FL problem completely changes nature: in the classical FL formulation the goal is to associate each outlier gene in the network with exactly one (the closest) driver gene. In the RWFL formulation, each outlier gene is associated with multiple driver genes (whose collective distance to the outlier will no longer be the shortest pairwise distance), forming a many-to-many relation. Intuitively, hitting time measures how accessible a particular outlier gene is from potential driver genes. Thus RWFL problem asks to find the smallest set of sequence-altered genes from which one can reach (a good proportion of) outliers within a user defined “multi-hitting time” - the expected length of the shortest random walk originating from any of the sequence altered genes, and ending at an outlier.

As per the standard facility location problem, RWFL is NP-hard. In fact, even the problem of computing the multi-hitting time between a set of nodes in a network and a particular target node is difficult. In what follows, we summarize how we approach this problem within the HIT’nDRIVE framework (see [Supplemental Methods](#) for details).

For simplicity, we first describe how HIT’nDRIVE works on single patient data. Given an interaction network with $\mathcal{X}$ denoting the set of sequence-altered genes (through SNVs or CNAs) and $\mathcal{Y}$ denoting the set of expression-altered genes, HIT’nDRIVE computes the smallest subset of $\mathcal{X}$ whose joint “influence” over (a user defined fraction of) expression-altered genes is sufficiently high (i.e. above a user defined threshold). The influence of a set of (sequence-altered) genes X over an expression-altered gene g is defined as $\frac{1}{MHT(X,g)}$, where $MHT(X,g)$ denotes the multi-hitting time, the expected length of the shortest random walk originating at each one of the genes in X that ends at g . Therefore, HIT’nDRIVE aims to solve the RWFL problem in a network where $\mathcal{X}$ are the “potential facilities” and $\mathcal{Y}$ are the “requests”.

Since RWFL is a computationally hard problem, and cannot be solved in a reasonable amount of time in its original formulation, we reduce the RWFL problem to the weighted multi-set cover (WMSC) problem, for which we give an ILP formulation. Intuitively, in this new formulation, HIT’nDRIVE associates the genomic alterations with transcriptomic changes in the form of a bipartite graph $G_{bip}(\mathcal{X}, \mathcal{Y}, \mathcal{E})$ where $\mathcal{X}$ is the set of aberrant genes, $\mathcal{Y}$ is the set of patient-specific expression-altered genes, and $\mathcal{E}$ is the set of edges. If gene x_i is mutated in a patient p , we set edges between x_i and all of the expression altered genes in the same patient (y_j, p) where the edges are weighted by the inverse pairwise hitting times $w_{ij} := H_{x_i, y_j}^{-1}$ ([Figure 1A](#)). The WMSC problem on this representation of data asks to find the smallest subset of $\mathcal{X}$ (as potential drivers) whose total influence (sum of pairwise influence values) over a user defined fraction of expression-altered genes (for each patient) is sufficiently high.

The reduction from RWFL problem to the WMSC problem is achieved by estimating the multi-hitting time as a function of independent hitting times of the drivers to an outlier, which provides an upper bound on the multi-hitting time. The exact individual hitting times are calculated by a matrix inversion method ([Tetali 1999](#)) (see [Supplemental Methods](#) for details). The resulting WMSC problem can then be formulated as the ILP below, which is efficiently solvable by IBM CPLEX (within minutes) for all data sets we considered.

$$\begin{aligned}
& \min_{x_1, \dots, x_{|\mathcal{D}|}} \sum_i x_i \\
& \text{s.t.} \\
& \forall i, j : x_i = e_{ij} \\
& \forall j : \sum_i e_{ij} w_{ij} \geq y_j \gamma \lambda_j \sum_i w_{ij} \\
& \sum_j y_j \geq \alpha |\mathcal{D}| \\
& \forall p : \arg \beta_{\lambda_j}(y_j) = 1 \\
& x_i, e_{ij}, y_j \in \{0, 1\}
\end{aligned}$$

The above ILP formulation for the WMSC problem introduces binary variables x_i , y_j , e_{ij} , respectively, for each potential driver, expression-alteration event, and edge in the bipartite graph. The objective of the ILP is to minimize the number of drivers (i.e. the sum of x_i values) subject to four constraints. The first constraint ensures that a selected driver contributes to the coverage of each of the expression alteration events it is connected to (in each patient, if multiple patients are available). The second constraint ensures that selected (patient-specific) driver genes contribute enough to cover at least a (γ) fraction of the sum of all incoming edge weights to each expression alteration event. This constraint corresponds to setting an upper bound on our estimate on the inverse of multi-hitting time of the selected (patient specific) drivers on an expression alteration event. The third constraint ensures that the selected driver genes collectively cover at least an α fraction of the set of expression alteration events. And the fourth constraint ensures that for each patient, the top β fraction of the expression altered genes with highest weights (λ_j) are always covered.

As indicated above our ILP formulation for WMSC problem can be generalized to multiple patients with the objective of minimizing the total number of driver genes across all patients, subject to the constraint that a user-defined proportion of outlier genes in each of the patients are covered by the subset of driver genes present in that patient (see [Supplemental Methods](#) for details).

In order to quantitatively assess the genes identified by HIT'nDRIVE, we extended our previously developed algorithm, OptDis ([Dao et al. 2011](#)), for *de novo* identification of modules of small size inside the interaction network which contain (i.e. are seeded by) at least one predicted driver gene. The modules are chosen so that their discriminative power (for phenotype classification) is the greatest among connected sub-networks of similar size that contain the individual predicted driver genes. In general, OptDis performs supervised dimensionality reduction on the set of connected sub-networks. It projects the high dimensional space of all connected sub-networks to a user-specified lower dimensional space of sub-networks such that, in the new space, the samples belonging to the same class are closer and the samples from different class are more distant to each other (i.e. minimize in-class

distance and maximize out-class distance) with respect to a normalized distance measure (typically L_1). Then we use module features (average expression of genes in the module) for phenotype classification (Figure 1B-C). Using such module features, we hope that the classifier in use does not *overfit* on rare driver genes and is able to *generalize* the signal coming from rare drivers to new patients. We report the classification accuracy based on the identified driver-seeded modules as means of quantitative validation of our results (in the absence of ground truth). We also look at the genes that build the chosen modules (of high classification accuracy) in attempt to identify cancer-related pathways.

Datasets and Analysis

We used publically available datasets of four major cancer-types glioblastoma multiforme (GBM), Ovarian serous cystadenocarcinoma (OV), breast adenocarcinoma (BRCA), and prostate adenocarcinoma (PRAD) 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.

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

Copy number aberrations (CNAs): 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 $= 5 \times 10^{-6}$) 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 aberrant genes at the genomic level.

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.

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.

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.

Interaction networks: We used STRING v10 (Szklarczyk et al. 2015) protein-interaction network which contains high confidence functional protein-protein interactions (see [Supplemental Methods](#) for details).

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 (see [Software and Code Availability](#) for details). A cut-off threshold of false discovery rate (FDR) ≤ 0.01 was used to obtain the significantly enriched pathways. Same procedure, as above, is used to assign biological functional to the gene-modules.

Derivation of outlier-genes: We used generalized extreme studentized deviate (GESD) test (Rosner 1983) to obtain the outlier genes (see [Supplemental Methods](#) and [Software and Code Availability](#) for details).

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 $S \geq 33$ 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 cohorts (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).

Software and Code Availability

The hitting-time based influence-matrix generation and HIT'nDRIVE algorithm are implemented in C++. They can be accessed from GitHub (<https://github.com/sfu-compbio/hitndrive>) and as [Supplemental Archive 1](#). GESD test based outlier detection method is available at GitHub (<https://github.com/raunakms/GESD>) and as [Supplemental Archive 2](#). A Fisher's exact test based gene-set enrichment analysis is available at GitHub (<https://github.com/raunakms/GSEA-Fisher>) and as [Supplemental Archive 3](#).

Acknowledgements

We gratefully acknowledge Dr. Shancheng Ren (Second Military Medical University, Shanghai, China) for providing us RNA-seq data for prostate cancer patients. We would like to thank Dr. Oktay Gunluk (IBM T. J. Watson Research Center, NY, USA) and IBM Academic Initiative for providing us free licence to IBM ILOG CPLEX Optimization Studio (CPLEX). This project was funded by: Canadian Cancer Society Research Institute (S.C.S, C.C.C.), Natural Sciences and Engineering Research Council of Canada (NSERC) Discover Grants (S.C.S) and Terry Fox Research Institute (CCC). R.S. is supported by Mitacs Accelerate Awards, Prostate Cancer Foundation - BC Research Award and Canadian Institutes of Health Research (CIHR) Bioinformatics Training Program. E.H. is supported by NSERC-CREATE Computational Methods for the Analysis of the Diversity and Dynamics of Genomes (MADD-Gen) program. The authors thank all members of the Sahinalp's and Collins' labs for helpful

suggestions. The results published here are in part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contribution

R.S., E.H., G. H. and S.C.S conceived the study. R.S and E.H. developed the HIT'nDRIVE algorithm to prioritize driver genes, performed entire analysis and wrote the paper. T.S. provided mathematical proofs for estimating multi-source hitting time. P.D. and K.W. developed OptDis algorithm to identify driver modules. F.V, J.Y. and S.A. helped in data analysis. C.C.C. helped in biological interpretation of the results. All authors contributed to preparation of the manuscript.

Disclosure Declaration

Authors declare no competing financial interests.

References

- Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, et al. 2010. An integrated approach to uncover drivers of cancer. *Cell*, **143**(6):1005–17.
- Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White Ta, Stojanov P, Van Allen E, Stransky N, et al. 2012. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature genetics*, **44**(6):685–9.
- Bashashati A, Haffari G, Ding J, Ha G, Lui K, et al. 2012. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, **13**(12):R124.
- 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.
- Beltran H, Eng K, Mosquera JM, Sigaras A, Romanel A, Rennert H, Kossai M, Pauli C, Faltas B, Fontugne J, et al. 2015. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA oncology*, **1**(4):466–74.

- Bianchini G, Balko JM, Mayer IA, Sanders ME, and Gianni L 2016. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nature reviews. Clinical oncology*, **13**(11):674–690.
- Bowyer SE, Rao AD, Lyle M, Sandhu S, Long GV, McArthur Ga, Raleigh JM, Hicks RJ, and Millward M 2014. Activity of trametinib in K601E and L597Q BRAF mutation-positive metastatic melanoma. *Melanoma research*, **24**(5):504–8.
- Chen Y, Hao J, Jiang W, He T, Zhang X, Jiang T, and Jiang R 2013. Identifying potential cancer driver genes by genomic data integration. *Scientific Reports*, **3**(3538).
- Ciriello G, Cerami E, Sander C, and Schultz N 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, **22**(2):398–406.
- Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, and Sander C 2013. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, **45**(10):1127–1133.
- Condamin S, Bénichou O, Tejedor V, Voituriez R, and Klafter J 2007. First-passage times in complex scale-invariant media. *Nature*, **450**(7166):77–80.
- 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.
- Dahlman KB, Xia J, Hutchinson K, Ng C, Hucks D, Jia P, Atefi M, Su Z, Branch S, Lyle PL, *et al.* 2012. BRAFL597 mutations in melanoma are associated with sensitivity to MEK inhibitors. *Cancer Discovery*, **2**(9):791–797.
- Dao P, Wang K, Collins C, Ester M, Lapuk A, and Sahinalp SC 2011. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, **27**(13):i205–13.
- Ding J, McConechy MK, Horlings HM, Ha G, Chun Chan F, Funnell T, Mullaly SC, Reimand J, Bashashati A, Bader GD, *et al.* 2015. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature communications*, **6**:8554.
- Ding L, Ley TJ, Larson DE, Miller Ca, Koboldt DC, *et al.* 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**(7382):506–10.

- Dutta B, Pusztai L, Qi Y, André F, Lazar V, Bianchini G, Ueno N, Agarwal R, Wang B, Shiang CY, *et al.* 2012. A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. *British journal of cancer*, **106**(6):1107–16.
- Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, *et al.* 2017. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, **45**(D1):D777–D783.
- Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR 2004. A census of human cancer genes. *Nature reviews. Cancer*, **4**(3):177–83.
- 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.
- Greaves M and Maley CC 2012. Clonal evolution in cancer. *Nature*, **481**(7381):306–13.
- Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, *et al.* 2007. Patterns of somatic mutation in human cancer genomes. *Nature*, **446**(7132):153–8.
- Greenman C, Wooster R, Futreal PA, Stratton MR, and Easton DF 2006. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, **173**(4):2187–98.
- Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, Koval J, Das I, Callaway MB, Eldred JM, *et al.* 2013. DGIdb: mining the druggable genome. *Nature methods*, **10**(12):1209–10.
- Gupta A, Hossain MM, Miller N, Kerin M, Callagy G, and Gupta S 2016. NCOA3 coactivator is a transcriptional target of XBP1 and regulates PERK-eIF2 α -ATF4 signalling in breast cancer. *Oncogene*, **35**(45):5860–5871.
- Hopcroft J and Sheldon D 2007. Manipulation-Resistant Reputations Using Hitting Time. In Bonato A and Chung FRK, editors, *Algorithms and Models for the Web-Graph: 5th International Workshop, WAW 2007, San Diego, CA, USA, December 11-12, 2007. Proceedings*, pages 68–81. Springer Berlin Heidelberg.
- 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.
- Jornsten R, Abenius T, Kling T, Schmidt L, Johansson E, Nordling TE, Nordlander B, Sander C, Gennemark P, Funa K, *et al.* 2011. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, **7**(1):486.

- Kim YA, Wuchty S, and Przytycka TM 2011. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS computational biology*, **7**(3):e1001095.
- Korthauer KD and Kendzierski C 2015. MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics*, **31**(10):1526.
- Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, and Yeger-Lotem E 2011. ResponseNet: Revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Research*, **39**(SUPPL. 2):424–429.
- Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts Sa, *et al.* 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**(7457):214–8.
- Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, and Pietenpol JA 2011. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, **121**(7):2750–67.
- Leiserson MDM, Blokh D, Sharan R, and Raphael BJ 2013. Simultaneous identification of multiple driver pathways in cancer. *PLoS computational biology*, **9**(5):e1003054.
- Liben-Nowell D and Kleinberg J 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, **58**(7):1019–1031.
- Lin F, De Gooijer MC, Roig EM, Buil LCM, Christner SM, Beumer JH, WEurdinger T, Beijnen JH, and Van Tellingen O 2014. ABCB1, ABCG2, and PTEN determine the response of glioblastoma to temozolomide and ABT-888 therapy. *Clinical Cancer Research*, **20**(10):2703–2713.
- Masica DL and Karchin R 2011. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer research*, **71**(13):4550–61.
- Osborne CK, Bardou V, Hopp TA, Chamness GC, Hilsenbeck SG, Fuqua SAW, Wong J, Allred DC, Clark GM, and Schiff R, *et al.* 2003. Role of the estrogen receptor coactivator AIB1 (SRC-3) and HER-2/neu in tamoxifen resistance in breast cancer. *Journal of the National Cancer Institute*, **95**(5):353–61.
- Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, *et al.* 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y.)*, **321**(5897):1807–12.

- Patch AM, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S, Nones K, Cowin P, Alsop K, Bailey PJ, *et al.* 2015. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, **521**(7553):489–494.
- Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, and Stuart JM 2013. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, **29**(21):2757–64.
- Prasad V 2016. Perspective: The precision-oncology illusion. *Nature*, **537**(7619):S63–S63. Outlook.
- Rosner B 1983. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, **25**(2):165–172.
- Shi K, Gao L, and Wang B 2016. Discovering potential cancer driver genes by an integrated network-based approach. *Molecular BioSystems*, **12**(9):2921–2931.
- Shrestha R, Hodzic E, Yeung J, Wang K, Sauerwald T, Dao P, Anderson S, Beltran H, Rubin MA, Collins CC, *et al.* 2014. HIT’nDRIVE: Multi-driver gene prioritization based on hitting time. In Sharan R, editor, *Research in Computational Molecular Biology: 18th Annual International Conference, RECOMB 2014, Pittsburgh, PA, USA, April 2-5, 2014, Proceedings*, pages 293–306. Springer International Publishing.
- Stratton MR, Campbell PJ, and Futreal PA 2009. The cancer genome. *Nature*, **458**(7239):719–24.
- 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.
- Suo C, Hrydziusko O, Lee D, Pramana S, Saputra D, Joshi H, Calza S, and Pawitan Y 2015. Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics*, **31**(16):2607.
- Suthram S, Beyer A, Karp RM, Eldar Y, and Ideker T 2008. eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular systems biology*, **4**(162):162.
- 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.

- Tetali P 1999. Design of on-line algorithms using hitting times. *SIAM J. Comput.*, **28**(4):1232–1246.
- 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.
- Tomlins Sa, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, *et al.* 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)*, **310**(5748):644–648.
- Tu Z, Wang L, Arbeitman MN, Chen T, and Sun F 2006. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, **22**(14):489–496.
- Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, *et al.* 2014. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature medicine*, **20**(6):682–8.
- Vandin F, Upfal E, and Raphael BJ 2011. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology*, **18**(3):507–22.
- Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, *et al.* 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**(12):i237–45.
- Verhaak RGW, Hoadley Ka, Purdom E, Wang V, Qi Y, *et al.* 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, **17**(1):98–110.
- 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.

- Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, Auluck PK, Geddie ML, Valastyan JS, Karger DR, *et al.* 2009. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, **41**(3):316–323.
- 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.
- Youn A and Simon R 2011. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**(2):175–81.

Figure Legends

Figure 1. Overview of HIT'nDRIVE algorithmic framework. (A) HIT'nDRIVE integrates sequence-wise changes in genome with expression changes in transcriptome obtained from patients' tumor samples. The influence values derived from the protein interaction network indicate how likely a driver gene influences its downstream target genes in the network. (B) The predicted driver genes are used as seeds to discover modules of genes that discriminate between the sample phenotypes using OptDis. (C) Based on this the driver modules are ranked and thus prioritized.

Figure 2. Summary of driver genes prioritized by HIT'nDRIVE. (A) Distribution of predicted driver genes in cancer genes databases. CGC database contains genes for which mutations have been causally implicated in cancer. Genes curated in CGC database represents likely drivers of cancer. COSMIC is a comprehensive database of somatic mutations that have been reported in different cancers. However, every gene present in COSMIC database may not represent drivers of cancer. (B) Distribution of driver genes in druggable genes databases. Actionable genes in cancer therapy were derived from TARGET database. List of druggable genes were extracted from DGI database. (A-B) The numbers in the panel represent the number of genes in respective categories. (C) Distribution of patient druggability. Patient druggability was accessed using information in TARGET and DGI databases. The numbers in the panel represent the number of patients in respective categories.

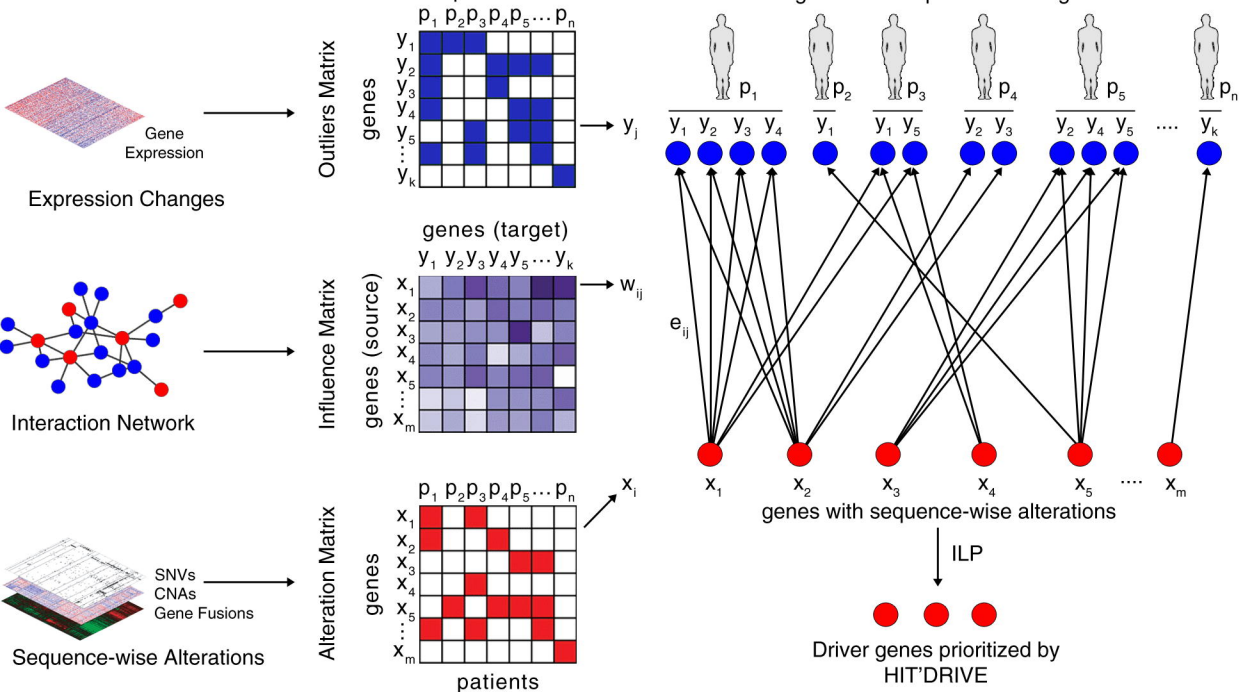
Figure 3. Network properties of driver genes. (A) The centrality of the predicted drivers in STRING v10 network. The size of the circles is proportional to the alteration frequency of the driver gene. The color scale represents the total influence of the driver gene on the expression outliers. (B) Correlation between influence and centrality. Each dot represents a target node receiving certain amount of influence from all source nodes in the network. A lowess regression line is represented in blue. (C) Correlation between incoming and outgoing influence of a node. Each dot represents a node in the network and the color scale represents its betweenness centrality. A linear regression line is represented in blue. (D) Boxplot of the total influence of driver genes predicted by HIT'nDRIVE on the expression outliers compared to that of other altered genes (genes not predicted as drivers). (E) Correlation between gene influence and its alteration frequency in the respective patient cohort. (F) Relative influence of driver genes in each patient in GBM cohort with mutation in *ABCB1*. (G) Relative influence of driver genes in each patient in PRAD cohort with mutation in *BRAF*. (All gene influence values have been multiplied by 10^5 before log transformation.)

Figure 4. Phenotype classification using driver-seeded modules. (A) Phenotype (tumor vs normal) classification accuracy in gene-expression datasets of different cancer-types using three different methods - HIT'nDRIVE-unsupervised (left panel), HIT'nDRIVE-OptDis (middle panel) and DriverNet (right panel). (B) Comparison of HIT'nDRIVE with DriverNet.

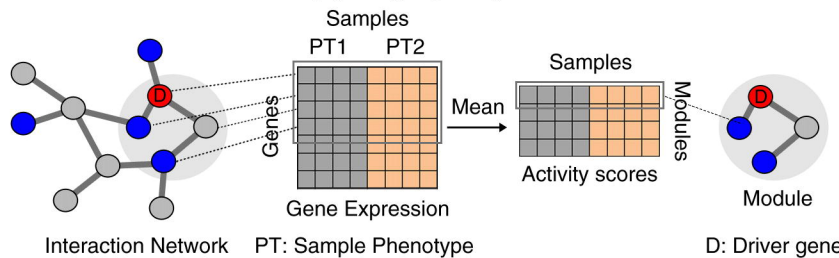
Figure 5. BRCA subtype classification using driver modules. (A) Performance accuracy of classifying different subtypes for breast cancer using activity-score of subtype specific driver modules as features in three distinct datasets. (B) Box plot comparing subtype specific driver-seeded modules and driver-free modules with respect to three distinct measures - log-rank test pvalue, hazard-ratio (HR) and concordance-index (c-index). (C) A BRCA subtype specific driver module (BASAL-02) seeded by *NCOA3* that distinguished Basal subtype from rest of the BRCA subtypes. (D) Activity-score of BASAL-02 module across different BRCA subtypes. (E) Kaplan-Meier plot showing the significant association of BASAL-02 module with patients' clinical outcome in the three datasets considered.

Figure 6. Drug efficacy predicted by HIT'nDRIVE seeded driver genes. (A) Accuracy of drug-response phenotype classification for all 265 drugs used in GDSC study across 25 cancer types (the remaining 5 cancer types for which only a very limited number of cell lines have been made available are statistically insignificant and thus have not been used). The classification accuracy for each drug on each cancer type is measured based on the collective use of at most 10 best discriminating modules, i.e. the accuracy is maximized across the range of 1 to 10 (best discriminating) modules. Note that many of the drugs were not tested on all cancer types; in fact for the vast majority of cancer types only a handful of drugs were tested. (B) Classification accuracy of modules that distinguish the drug-response phenotypes after treatment with Gefitinib in BRCA cell-lines (top-panel), Temozolomide in GBM cell-lines (middle-panel), and Nutlin-3a in OV cell-lines (bottom-panel). Important genes identified in the modules and involved in the dysregulated signalling pathways have been highlighted. (C-E) The figures represent the dysregulated signalling pathways in the respective drug perturbation.

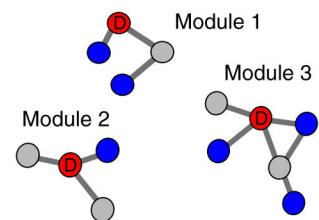
A Driver Gene Prioritization



B Driver Module Discovery (using OptDis)

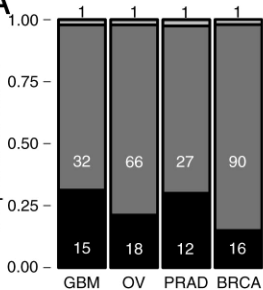


C Driver Module Prioritization



A

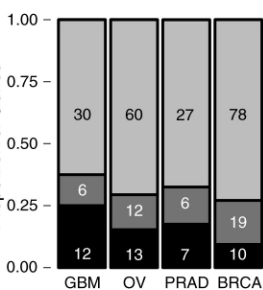
Proportion of Genes



Genes not present in CGC/COSMIC
 Genes mutated in cancer (COSMIC)
 Cancer genes in CGC

B

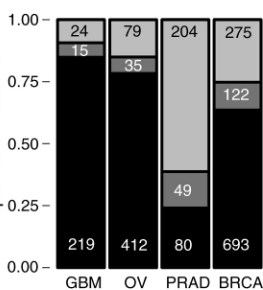
Proportion of Genes



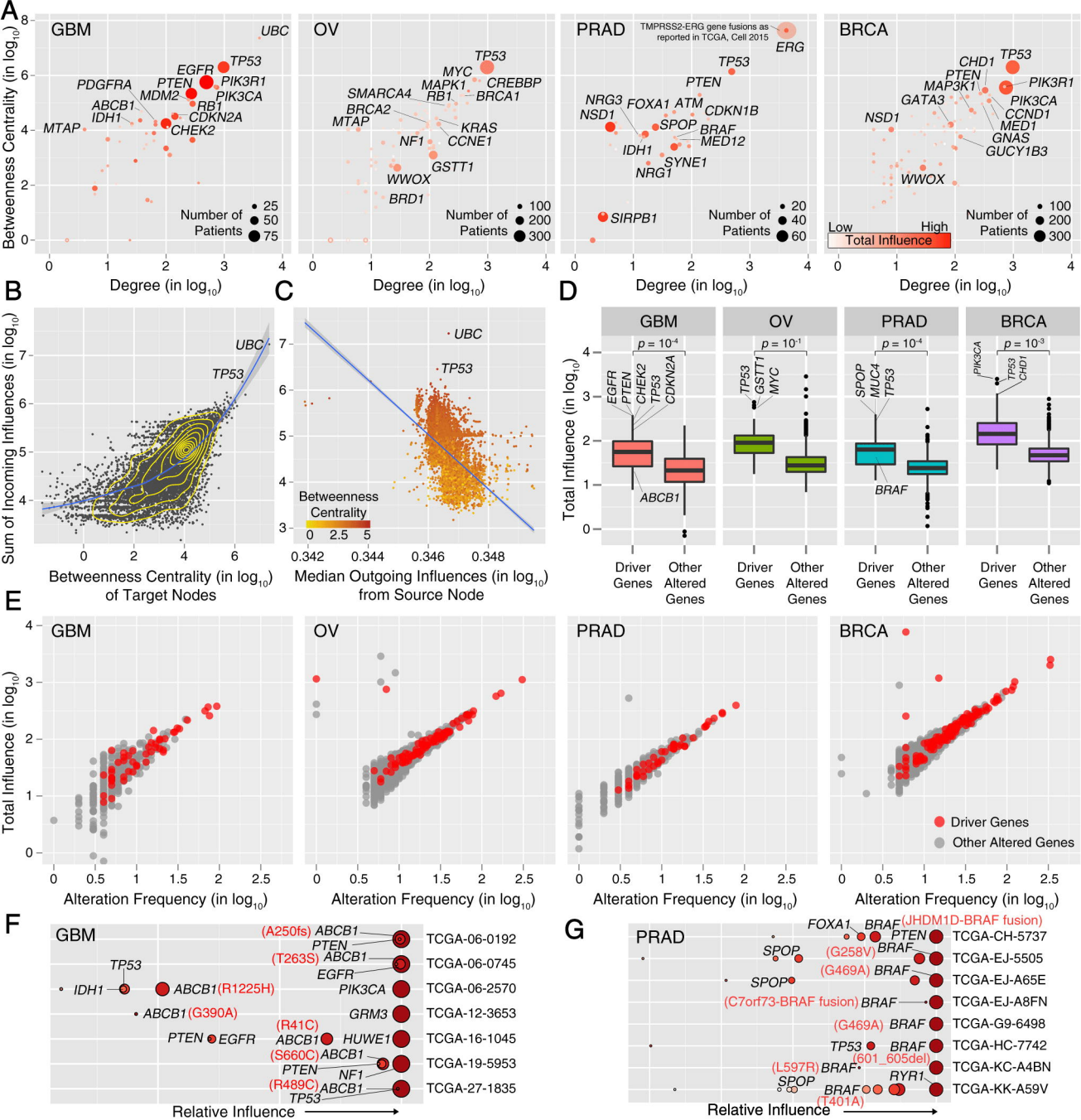
Undruggable genes
 Druggable genes
 Actionable genes

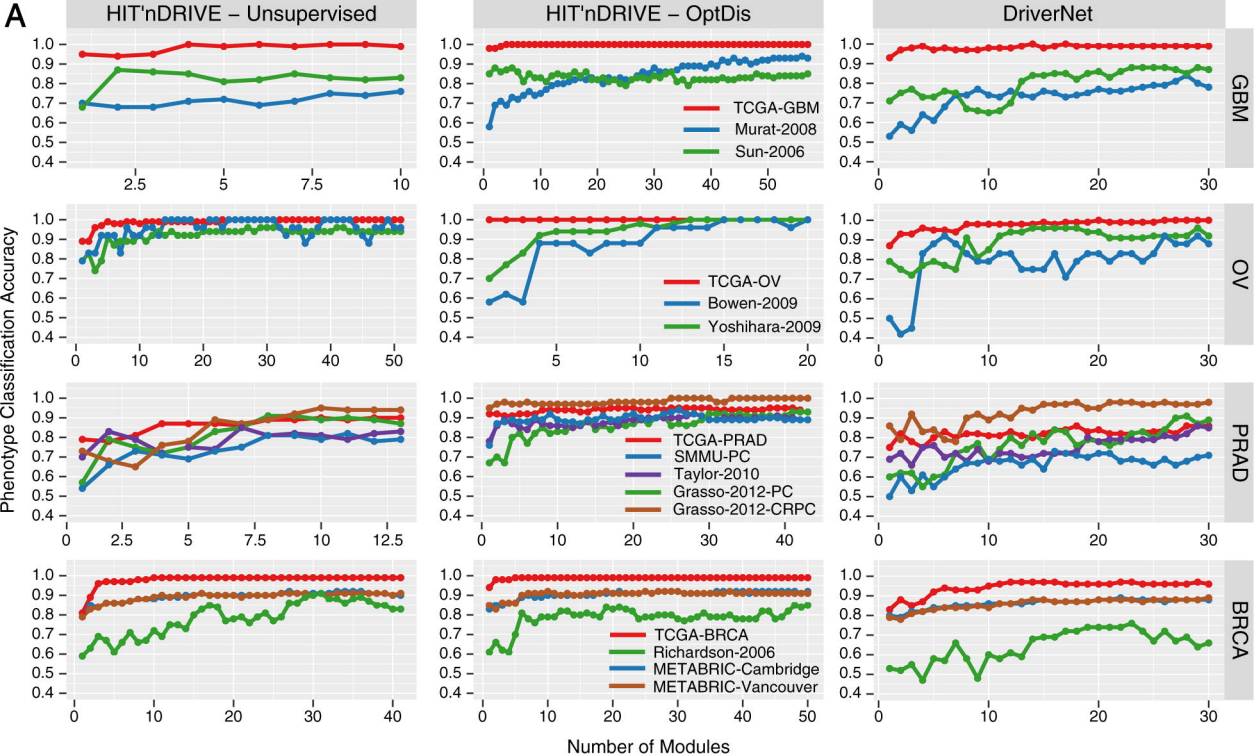
C

Proportion of Patients



Patient with undruggable genes
 Patient with druggable gene(s)
 Patient with actionable gene(s)





B

