PANOPLY: A computational method for identification of promising drugs for a patient based on multidimensional data

Krishna R. Kalari[1,$,*], Jason P. Sinnwell[1,$], Kevin J. Thompson[1], Xiaojia Tang[1], Erin E. Carlson[1], Jia Yu[4], Peter T. Vedell[1], James N. Ingle[3], Richard M. Weinshilboum[4], Liewei Wang[4], Judy C. Boughey[2], Matthew P. Goetz[3,4], Vera Suman[1,*].

[1]Department of Health Sciences Research, [2] Department of Surgery, [3] Division of Medical Oncology, [4] Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester 55905 USA

$ Co-first authors

* To whom correspondence should be addressed.

Corresponding authors:

Krishna R Kalari, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA; Tel: 507-538-4602; Email: Kalari.Krishna@mayo.edu

Vera J Suman, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA; Tel: 507-538-4602; Email: suman@mayo.edu

ABSTRACT

The majority of cancer patients receive treatments that are minimally informed by omics data. Our goal was to develop a precision medicine computational framework (PANOPLY**: P**recision c**a**ncer ge**no**mic re**p**ort: single samp**l**e inventor**y**) to identify and prioritize drug targets and cancer therapy regimens. The PANOPLY approach integrates clinical data with germline and somatic features obtained from multi-omics platforms, and apply machine learning, and network analysis approaches in the context of the individual patient and matched controls. The PANOPLY employs multiple steps including i) selection of matched controls, ii) preprocessing of data and identification of driver mutations and altered genes in the patient, iii) identification of suitable drugs using the driver-gene network and random forest analysis, and iv) a multi-omics case report of the patient with prioritization of anti-cancer drugs. The PANOPLY framework can

be executed on a stand-alone virtual machine and is also available for download as an R package. We applied PANOPLY to multiple publicly accessible multi-omics tumor datasets with survival data available. We also applied the method to an institutional breast cancer neoadjuvant chemotherapy study which collected clinical and genomic data as well as patient-derived xenografts (PDXs) to investigate the prioritization offered by PANOPLY. We found that that the prioritized drug, olaparib, was more effective than placebo at treating the tumor in the chemotherapy resistant PDX model ($P < 0.05$). Further studies are ongoing to determine the efficacy of PANOPLY using additional PDXs. In summary, PANOPLY prioritizes drugs based on both clinical and multi-omics data, and it can aid oncologists in their decision-making for therapy in the individual patient.

INTRODUCTION

There has been substantial progress in the fight against Cancer; however, cancer remains the second leading cause of death in the United States [1]. A major focus of cancer research has been the identification of oncogenic drivers and the development of drugs that selectively target those driver events. This approach has led to the development of agents that have been shown to successfully target the driver mutational events such as: trastuzumab, which targets *HER2*+ breast cancer [2,3], imatinib, which inhibits the *BCR-ABL*l tyrosine kinase produced by the Philadelphia translocation in chronic myelogenous leukemia [4], vemurafenib for the treatment of *BRAF* V600E mutant malignant melanoma [5], agents targeting *EGFR* mutations non-small cell lung carcinoma and [6], and crizotinib for non-small cell lung cancer with *ALK* rearrangements [7].

The mapping of the human genome has opened the door to the exploration of the tumor and environmental features to uncover the drivers of cancer and its resistance to treatment. Currently, many commercial platforms for gene sequencing (e.g. Foundation One) have been developed to identify these mutations and are commonly used in clinical practice. However,

most of these platforms are focused on detecting a limited number of gene abnormalities in specific genes and do not include comprehensive “omic” data analysis. For many tumor types, this leads to the inability to link mutational drivers with druggable targets. This was evident in the prospective SAFIR01/UNICANCER clinical trial, in which 407 patients were screened for copy number alterations and mutations in PIK3CA and AKT1 alterations and treatment could only be personalized for 13% (55 of 407) of patients with metastatic breast cancer. Moreover, only 4 of the 43 patients who received the recommended targeted treatment had an objective tumor response [8].

A pressing need exists to develop a personalized treatment plan to treat a patient’s unique tumor and germline profile. Development of databases and algorithms that synthesize multiple complex forms of omics data to identify and prioritize drug targets and pathways are necessary for identification of the promising drugs to treat cancer. Databases such as MD Anderson’s Personalized Cancer Therapy [9], Vanderbilt’s My Cancer Genome [10], the Broad Institute’s TARGET [11], The Cancer Genome Atlas (TCGA) [12], and the Catalogue of Somatic Mutations in Cancer (COSMIC) [13] contain information on the frequency of alterations in thousands of patients with cancer. Programs such as DriverNet [14], IntOGen [15], analyze a single type of omics data, such as somatic mutations, to identify potential driver genes. Other programs, such as XSeq [16], OncoRep [17], OncoIMPACT [18], and iCAGES [19] integrate data on somatic mutations and/or copy number alterations (CNAs), and gene expression. Although integrating these data types represents substantial progress toward the full molecular characterization needed for precision cancer care, no comprehensive method for integrating multi-omics and clinical data has yet been developed and validated for selecting the most compatible agents for a given patient’s –omics profile. Thus, a method is urgently needed that synthesizes multiple forms of -omics data to build gene networks and identify drugs that can target dysregulated sections of those networks.

We have developed a method called PANOPLY (**P**recision c**a**ncer ge**no**mic re**p**ort: single sampl**e** inventor**y**) that identifies the somatic or germline alterations found in a cancer patient as well as performs a comprehensive multi-omics network analysis to select a set of agents appropriate to target an individual's tumor. The results are summarized in a report which the patient's medical oncology team can use to choose the particular agent to be administered. In brief, PANOPLY uses machine learning and knowledge-driven network analysis to analyze gene expression, Copy Number Alterations (CNA), germline and somatic single nucleotide variants (SNVs), and expressed SNV data to identify patient specific alterations driving oncogenesis and then prioritize the drugs which target the networks and pathways associated with these cancer-driving alterations. Herein, we describe the PANOPLY and provide examples using both institutional and publicly available data sets where PANOPLY was used to identify drugs for individual patients and subgroups of patients who do not respond to traditional chemotherapy. We then validated these predictions by testing one of the PANOPLY-recommended drugs in a patient with chemo-resistant triple-negative breast cancer (TNBC) using patient-derived xenografts (PDXs). PANOPLY is freely available as an R package at http://bioinformaticstools.mayo.edu/research/panoply/

MATERIALS AND METHODS

PANOPLY: Overview and Design

PANOPLY was designed as a means to identify and rank FDA-approved anti-cancer drugs likely to benefit a patient based on their multi-omics profile. This approach requires germline and somatic tumor data from the patient (case) as well as a set of patients (controls or matched controls) with a similar clinical profile to the case but differs with regard to the phenotype of interest (response to a given treatment, survival, etc.). PANOPLY identifies the differences and similarities in germline and somatic features of the case and their controls and uses these

results to determine the treatment options based on druggable targets found in case but not in its controls. The results are summarized in a report that can aid clinicians in their treatment decision-making.

A high-level overview of PANOPLY workflow is shown in Figure 1, and the details of each step are provided below. The workflow can be applied for multi-omics data analysis. However, it is also designed to work if only gene expression data is available. After preprocessing the data, the PANOPLY workflow can run in less than two minutes consuming less than 2G of RAM in an R session.

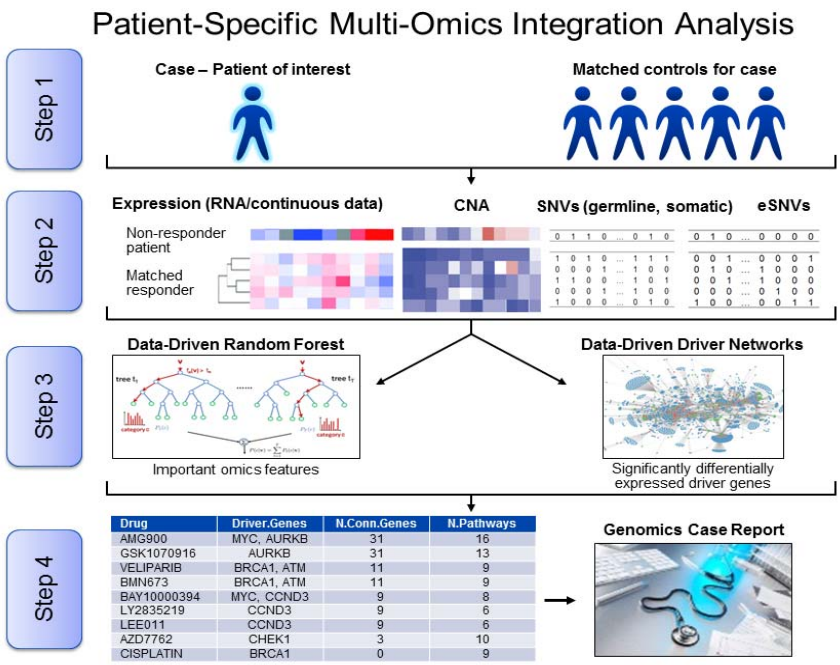


Figure 1. High-level overview of PANOPLY. Step 1: A patient of interest is compared with matched controls. Step 2: For each subject, expression data, copy number alteration (CNA), single nucleotide variant (SNV), and expressed single nucleotide variant (eSNV) data will be preprocessed to identify case-specific driver alterations. Step 3: Multi-omics data will be provided to random forest and network analysis methods to identify the top prioritized drugs to target genes that are driving oncogenesis in the patient. Step 4: A genomics case report listing the drugs to prioritize, based on their ability to target driver mutations and their dysregulated gene networks, is generated for researchers and clinicians.

PANOPLY's Curated Knowledge Data Set

PANOPLY relies on a curated knowledge data set consisting of (a) a set of cancer genes (tumor suppressors, oncogenes, and germline cancer associated genes), (b) a list of drugs that target those cancer genes, and (c) a gene-gene interaction network consisting of the cancer genes and their downstream connections, created from the human gene-gene interaction database (Reactome) [20]. To build this data set, first, we identified a total of 844 cancer genes from 11,941

genes implicated in cancer and pooled from various publications and cancer databases [21-23]. (See Section 1.1 of the Supplementary Material and Supplemental_Figure_S1). Using the Reactome gene interaction network, 429 of the 844 cancer genes were found to have a direct interaction with 8,358 of the 11,941 cancer-related genes, resulting in 13,913 gene-gene connections. These gene-gene connections are referred to as the curated cancer network (CCN). The 36 downstream genes that can be targeted from 429 cancer gene list were added to set of cancer target genes (CTGs). From publicly available drug databases and recent publication [24], 374 FDA-approved drugs were identified that targeted 475 CTGs. Additional details can be found in Supplementary Material section 1.1.

PANOPLY Methods and Workflow

Step 1 – Selection of matched controls. For the case of interest, controls of the same age and sex with a similar clinical profile to the case but differing with regard to the phenotype of interest (response, survival, etc.) must be identified. For example, matched controls are selected to have the same tumor subtype or disease as the case, the same sex, and similar age (+/- 5 years), tumor size, tumor grade, nodal status, *etc*. Based on simulation studies, it is recommended that at least four controls be matched to each case. When the case is drawn from an existing data set, its matched controls will typically be obtained from the same study population. If no control samples exist for a study, as would often be the case in the clinic, a matched-control set could be chosen from The Cancer Genome Atlas (TCGA) and then there is a module within PANOPLY to normalize the data to adjust for batch effects.

Step 2 – **Pre-processing of data and identification of driver mutations and genes**. In this step, PANOPLY (a) performs pre-processing protocols that prepare multi-omics data for analyses and (b) the list of curated cancer genes described in Supplementary Section 1.1 will

be applied to compare the molecular profile of a case with the matched controls to identify driver mutations and genes.

PANOPLY modules exist to process continuous gene expression and protein expression data and dichotomous genomic events data, such as CNVs, germline, and somatic SNVs and expressed SNVs. If CNA data are continuous, PANOPLY will convert CNA data in gain/loss categories. Below is an outline of the selection of candidate driver genes (a similar process is used for protein expression data):

- Conditional quantile normalization (CQN) of the gene expression data is conducted using the R-package CQN [25] to correct for GC-content per gene, gene length, and across-sample variability.
- For each gene *i*, the mean $\mu_{R,i}$ and standard distribution σ_R of the expression of that gene (on log2 scale) is determined among the matched controls. Assuming the distribution of the gene expression in the matched controls follows a Normal ($\mu_{R,i}$ σ_R), the percentile where the expression of gene *i* in the case $E_{C,i}$ falls on this Normal ($\mu_{R,i}$ σ_R) is determined. If the percentile is 95th or higher (5th or lower), then gene *i* is considered a driver gene. This threshold can be adjusted downward or upward. As the majority of targeted agents are inhibitors, PANOPLY currently focuses on identifying over-expressed genes. Additional details are provided in the Gene Expression description of the Supplementary Materials 1.6 section.
- Somatic variants were called using SomaticSniper (version 1.0.0.1-19) [26], JointSNVMix2 (version 0.8b2) [27], or MuTect (version 1.1.4) [28], and annotated with SnpEFF (version 3.0c) [29].

- Copy number losses/gains are identified using Pattern CNV and using a cutoff to call a loss or gain based on tumor percentage (see Supplementary Material Section 1.2 Copy Number Events for how the cutoff is determined).
- For a somatic mutation to be identified as a causal gene or driver, the variant or copy number alteration (CNA) must be present in the tumor but not in the blood of the case, and not a frequently observed mutation in matched-control tumors or matched-control blood samples.
- Rare germline mutations identified in the cases that are non-synonymous with allele frequency <= 0.05 are also considered as cancer drivers.

<u>Step 3</u> – Identification of suitable drugs using driver-gene networks and random forest analysis

Drug Network Test: PANOPLY identifies drugs that target the set of driver genes (identified in Step 2) and their effects on gene networks and relevant biological networks and pathways for those driver genes.

For each of the 374 FDA-approved anti-cancer drugs in our knowledge base, the R package GAGE [30] is employed to determine whether genes directly targeted by the drug are differentially expressed between the case and its controls. This identifies genes that are perturbed (e.g. over-expressed) in the case, and also determines which drug is most likely to be effective in the patient. Specifically, for each control, the difference in expression between the case and control is determined for each gene, and the mean and standard deviation of these differences is calculated. For each gene within a network, a z-test is used to test whether the mean difference in expression is different from that of the entire gene set. The resulting p-values for each of the network genes are combined into a network p-value, such that

$$\mathrm{X} = -\frac{1}{M}\sum_{i}^{M} \log P_i\,, \qquad and\ DNT.Pval(x) = P(X > x) \sim Gamma(M, 1)$$

where P_i is the individual p-value of network gene *i*, *M* is the size of the matched-control set, and X (the network test statistic) is the negative log sum of *M* independent p-values corrected for the number of genes in the network. The network test follows a *Gamma* (*M*, 1) distribution under the null hypothesis.

Drug Meta-Test: PANOPLY analyzes whether genes from the CCN that are directly affected by the case's driver genes are differentially expressed between the case and matched controls. Here, *g* represents the driver-gene network, which consists of the driver gene and its direct downstream genes from the CCN. Z-scores are calculated as stated above. Stouffer's Z-score method [31] is then applied to these Z-scores summed across the gene network to obtain an overall Z-score, $Z_g = \frac{1}{N_g}\sum_i^{N_g} z_i$ which follows a multivariate normal (MVN) distribution.

A measure of differential gene expression across the case-specific driver gene networks that are targeted by a drug D is calculated as the weighted sum of Z_g where $w_g = \sum_{j \in g} q_j / N_g$, N_g is the number of genes in driver-gene network g, and q_j indicates whether member gene *j* is targeted directly by the drug D. Thus, w_g represents the effect of the drug on the driver-gene network *g*. For drug D, the meta-statistic across all driver-gene networks is then calculated by the equation

$$Meta.Z_D = \sum \sqrt{w_g} Z_g.$$

To evaluate the significance of $Meta.Z_D$, a simulation approach is undertaken [32]. We center Z_g from the driver-gene network test to zero, such that $Z_g \sim MVN(0, \mathrm{V})$. We then create $Z'_{g,k}$ = $r_k * Z_g$, where $r_k \sim \mathrm{Normal}(0,1)$, such that $Z'_{g,k} \sim MVN(0, \mathrm{V})$ and V is the covariance matrix. Simulation-based DMT statistics are calculated from $Z'_{g,k}$ to get $Meta.Z'_{D,k}, k = 1 \ldots K$. The DMT simulation-based p-value for drug D is the number of times $Meta.Z'_{d,k}$ exceeds the observed $Meta.Z_D$ divided by K, or equivalently,

$$DMT.pval_D = \sum_K 1_{Meta.Z'_{D,k} > Meta.Z_D} / K.$$

Drug Ranking Scores: To rank drugs' predicted ability to inhibit overexpression of genes in the case's driver gene network a P.score is calculated using an approach by Brown [33] that is, $P.score_D = -log_{10}DNT.pval_D - log_{10}DMT.pval_D$. The most effective drugs for a case are those with the highest P.score, or equivalently low DNT and DMT p-values. Through empirical simulations, we recommend that P.score above 1.0 indicate the drug can efficiently and specifically target affected CCN genes in a case.

Random Forest Score: An additional approach was undertaken for feature reduction and identification of effective drugs, namely, random forest (RF) method for feature selection [34]. Taking the multi-omics, case-specific genomic events identified in Step 2, driver-gene features are fed to the RF model as a case-control classification problem. The importance score for each feature or event is acquired after decision trees have been constructed 1,000 times. The variable importance score is used to rank drugs. For drug *D,* $RF.score_D = \frac{1}{N.D}\sum_i^{N.D} RF_i$, where *N.D* is the total number of driver genes targeted by the drug, and RF_i is the variable importance score of each gene feature *i* given by the RF algorithm. Variable importance scores are generally low, with higher numbers meaning higher variable importance. Therefore, an RF.score above 0.001 indicates an important feature for the case.

Each score provides complementary information. If we consider the simple scenario in which a drug completely inhibits overexpression of the genes in the network (100% drug-target inhibition), the P.score indicates the significance of the sub networks associated with the target driver genes (therapeutic breadth), while, the RF.score indicates how well an identified target driver gene or alteration can differentiate the case compared and matched controls.

Step 4–Integrated multi-omics PANOPLY case report. In this last step, patient-specific treatment data and visualizations are provided in a comprehensive genomics case report. We provide an example report in the Supplementary Material for a cancer patient in Section 1.6. Briefly, the report contains clinical data, a list of driver genes in the case, and a list of drugs recommended for the case based on those driver genes, along with visualizations of druggable CCN alterations. The list of prioritized drugs contains information on targetable driver genes, pathways that include the targetable driver genes, and the P.score and RF.score for each drug. The top prioritized drugs perform well using both approaches, with an RF.score in the upper quartile and a P.Score above the median of all drugs for the case.

Use cases of PANOPLY method

To demonstrate PANOPLY's potential value, we first performed false positive and true positive rate evaluations using simulated data. Second, we applied PANOPLY to publicly available (TCGA) data sets including –omics and survival data to see whether it could successfully identify alternative drugs for patients with refractory colon and breast cancers. Third, we applied PANOPLY to an in-house breast cancer dataset (referred to as BEAUTY study [35]) to see whether it could successfully identify alternative drugs for patients with chemoresistant TNBC. Finally, we used patient-derived xenografts (PDXs) to validate the effectiveness of the top drug predicted by PANOPLY in a patient with chemoresistant TNBC.

Mining multiple PANOPLY case reports

PANOPLY is primarily intended for single-case genomic analysis. PANOPLY uses clustering techniques to identify a subset of patients and drugs that could fit into an on-going bucket trial. Prioritized drugs generated by PANOPLY are aggregated for the cases using percentile ranking, as implemented in the dplyr R package. Rankings are based on the combination of the

RF.score and the aggregated network score P.score. Non-negative matrix factorization (NMF) method [36] are implemented to identify patient cohorts, with similarly prioritized drug profiles. Clusters are evaluated, assessing the cophenetic and average silhouette scores. Drugs are assessed using Kim's method (the default method for the NMF package) to select the most delineating drugs. A word cloud plot is generated with R package using the protein targets of the clusters of drugs associated with the top 10% of delineating drugs.

Validation of PANOPLY's Drug Predictions With Patient-Derived Xenografts

To validate PANOPLY's drug predictions, the effectiveness of the top drug predicted for breast cancer patient from BEAUTY study (BC-051-1-1) with TNBC whose tumor did not respond to neoadjuvant paclitaxel and anthracycline/cyclophosphamide treatment was investigated using a PDX model. PDX tumors created with pre- and post-treatment samples were studied in female NOD-SCID mice. The tumors were grown to 200–250 mm^3, and then the mice were randomized (7 per group) into olaparib (15 mg/kg, once daily) or vehicle treatment groups. Mouse tumor size and body weight were measured twice per week, and mice were euthanized after 12 days of treatment. The difference in volume between drug- and vehicle-treated tumors was assessed using Wilcoxon rank-sum tests.

False Positive Rate (FPR) identification

We performed simulations to evaluate the performance of the DNT and DMT tests. Two non-cancer breast tissue data sets from different RNA-Seq platforms were utilized. We obtained dbGap permission to analyze the gene expression data from (phs000644.v1.p1) 20 microdissected normal, healthy breast tissue samples collected by the Susan G. Komen for the Cure Tissue Bank, sequenced using the SOLID-RNA platform. The normalized gene expression data was provided by Dr. Milan Radovich from Indiana University [37]. We then used the mean and covariance structure of the expression for 6234 genes (using curated cancer network

genes) from these 20 healthy breast samples, and randomly generated multivariate normal (MVN) data expression for patient sets of size M+1, where we assigned first M patients as the number of matched 'controls' (M=2, 4, 8, or 16) and the last patient as the "case". PANOPLY analysis steps were carried out for each case-control set and the simulation was repeated 500 times for each M. The second data set we utilized for our simulations was generated from 113 breast-cancer normal adjacent tissue samples from TCGA that were sequenced using Illumina HiSeq2000 platform. The RNA-Seq data were processed through the MAP-RSeq workflow [38] the gene expression counts were obtained and normalized with CQN normalization [25], and quality control checks were performed. The final data set contained gene expression for 8245 genes on 95 different samples. The simulation involved taking each patient one at a time and randomly selecting their controls from the remaining 95 patients and running the PANOPLY analysis. This process was repeated for control sets of size 2, 4, 8, and 16.

The factor that influences the false positive rate (the number of drugs falsely identified as promising) is the overlap of gene sets defined by the cancer gene networks and the cancer drugs. Supplemental_Figure_S2 shows the 429 cancer gene networks clustered by the similarity of their corresponding network genes. There are large overlaps among the 429 individual cancer gene networks. For example, the CCN genes of FGFR2 (n=42) and FGFR4 (n=42) networks share 41 common genes. In addition, as shown in the sub-figure of Figure S2 on these genes, FGFR3 (n=38) network shares 37 genes in common with FGFR2, FGFR4, and FGFR1 (n=51) networks. Similarly, Figure S3 shows the similarity of drugs by the genes they target. The sub-figure shows an example that is common in cancer drugs, with three drugs that target the same four genes (BRCA1, BRCA2, ATM, and ATR), with a fourth drug, olaparib, targeting a fifth gene, PARP1. To evaluate the effect of the gene set overlap has on the false positive rates, we performed simulations over three different sets of gene and drug networks as described below and summarized in Table 1.

S1 Complete Cancer Driver Genes and Drugs: Full set of 374 cancer drugs and 429 cancer gene networks

S2 Drivers-Only, Complete Drugs: This scenario is closest to what we would see in a patients PANOPLY analysis (after Step 3 driver gene identification method is applied) after identifying a subset of 429 cancer gene networks that are driven by a mutational event. We simulated a random set of 30 cancer driver genes that had a mutational event (e.g. copy number gain) along with overexpression of 20% curated cancer genes using the gene expression data. We then test the driver genes networks to show the effect on the false positive rate. The number of drugs tested in this scenario is the full set of 374 drugs.

S3. Reduced Cancer Drivers and Drugs: We applied a clustering-based reduction of cancer gene networks and drugs to define sets with reduced gene set overlap. From the clustering of 429 cancer gene networks in Supplemental_Figure_S2, we chose a similarity cutoff of 0.9 (red dashed line) and selected the cancer gene network within each branch with the most number of genes. For example, if FGFR1, FGFR2, FGFR3, and FGFR4 were the cancer gene networks in their branch, FGFR1 (n=51) would represent that entire branch of cancer gene networks because it has the most genes in its CCN. The total number of cancer genes in the dendrogram at a cutoff (binary similarity measure) of 0.9 is 175 cancer gene networks. Similarly, Supplemental_Figure_S3 shows dendrogram of drugs and their target gene sets following the same steps; the cutoff of 0.9 reduces the set to 117 drugs with reduced overlap of gene sets they target.

Table 1. False positive rate simulation scenarios

Scenario	Cancer networks Tested	Drugs Tested
S1: Complete Cancer Driver Genes and Drugs	429	374
S2: Drivers-Only, Complete Drugs	120 (30 CN gain + 90 expressed genes)	374
S3: Reduced Cancer Drivers and Drugs	175 (reduced by binary clustering)	117 (reduced by binary clustering)

True positive rate (TPR) identification

We next evaluated the performance of the DNT and DMT to find true positives, that is, to find drugs that would target over-expressed genes in a controlled simulation with "turned-on" cancer mechanisms. To create these scenarios, we took the simulated 500 sets of size M+1, (M=2,4,8,and16) previously generated and replaced the 'case' with an observation with a specific set of genes having a value larger 2 or 3 standard deviation higher than the mean expression from the original 20 healthy breast tissue samples [37]. The sets of genes were chosen to emulate conditions for which the olaparib (FDA approved targeted therapy for patients with hereditary *BRCA1* or *BRCA2* mutations) would be a strong drug candidate. We describe below the sets of genes, and which of the DNT or DMT are expected to perform well at identifying olaparib as a top drug for that set.

Set A: The overexpressed genes {*BRCA1, BRCA2, ATM, ATR, PARP1*} are all directly targeted by olaparib. The DNT is expected to perform better than the DMT; DNT evaluates the expression changes of these five genes in the case and matched controls. However, the DMT is a meta-test and is an aggregate of any driver gene network containing any of those five genes, which is over 30 networks ranging from sizes 4 to 285 genes.

Set B: The overexpressed genes {*BRCA2, ESR1, RAD51, RAD51C*} are the complete set of genes in the BRCA2 cancer gene network. Effectively, this scenario mimics a "turned on" signal associated with a *BRCA2* mutation, for which the DMT is expected to perform well since the BRCA2 network would contribute more to the meta test for olaparib than the other 30-plus networks because it is a smaller network, and therefore have a stronger weight in the DMT compared to DNT.

Set C: The overexpressed genes {*BRCA1, RAD21, FANCG, FANCI, SUMO2, H2AFX, ATM*} contain a mixture of genes that belong to two larger gene networks, *BRCA1* (*size*=84) and *ATM* (*size*=42), both of which are targeted by olaparib and over-expressed themselves. This scenario is expected to do moderately well with both DNT and DMT drug tests.

Since olaparib is not in the reduced drug set in simulation scenario with Set C, we only evaluate true positive rates for simulation scenarios S1 and S2.

RESULTS

PANOPLY Workflow: A Genome-Guided Approach to Identify the Right Drugs

We developed PANOPLY to determine alternative drugs for patients with cancer using a genome-guided approach. Prior-knowledge databases for genes and drugs are used to perform network and RF analyses of the features, as shown in Figure 2. The workflow is fully automated, and the drugs for an individual case are prioritized based on a scoring method that incorporates both types of analyses, and the results for each case are compiled in a report for the oncologist.

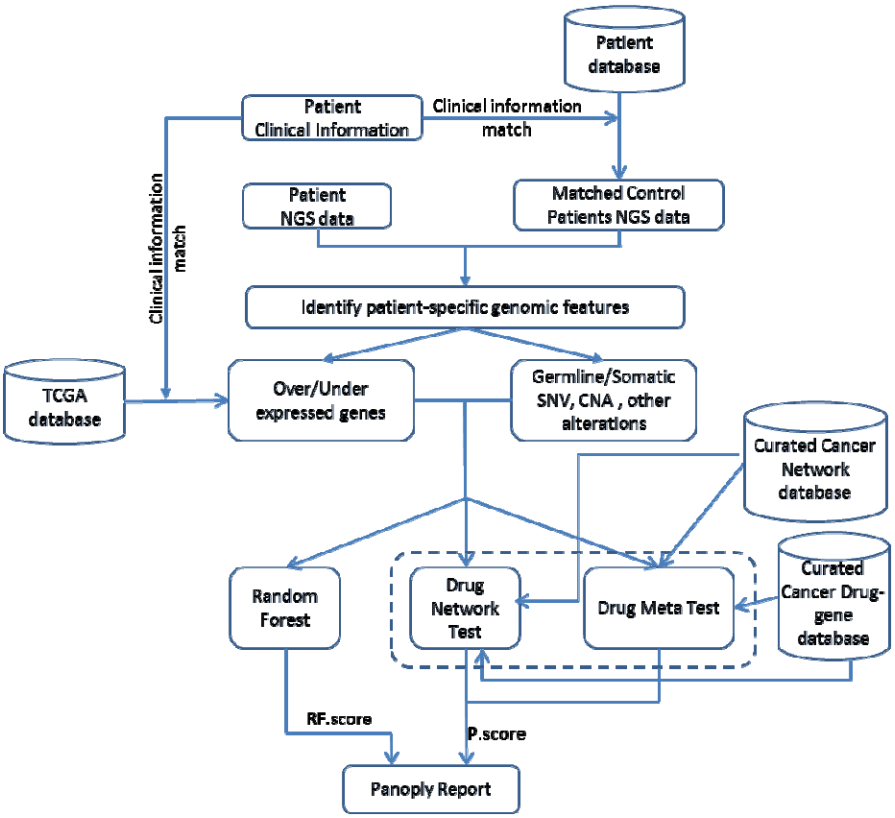


Figure 2. Detailed PANOPLY flowchart. Description of databases and tests conducted by the PANOPLY are indicated in detail in the flowchart. Each module of input features analyzed and databases used by the PANOPLY framework are represented in the diagram.

Type I Error Rate and Power Evaluations

Investigation of False Positive Rate (Type I error)

We compared the DNT and DMT p-values against nominal levels of α = 0.05 and 0.01, whereas the Drug P-Score we compare against the empirical 95th and 99th percentile of P.score from two simulated independent uniform distributions. Supplemental_Figure_S4 shows the false positive rates derived from the two data sets (with the case set *N*=1 and matched control set sizes *M*=2, 4, 8, and 16 and the cancer networks and drugs scenarios laid out in Table 1. The false positives rates are similar across matched control set sizes, with slightly higher rates at the

smallest matched set sizes of two. False positive rates of the drug network test (DNT, darkest bar) are adequate for Complete Cancer Driver Genes and Drugs scenario (S1) and Drivers Only- Complete Drugs Scenario (S2), but below the nominal rates in Reduced Cancer Drivers and Drugs scenario (S3) as indicated in Table 2. False positive rates for the drug meta test (DMT, middle bar) are highly inflated for S1, but near the nominal rate in the reduced sets in S2 and S3 (Table 2). The DMT false positive rates in S3 are even closer to nominal rates for the multivariate normal (MVN) simulated set, but are still inflated for the TCGA samples (second row in Supplemental_Figure_S4). The inflated false positive rate is reduced when the overlap of driver genes is reduced, but still above the nominal rates when the overlap is reduced strongly in both the driver genes and the drugs.

Investigation of True Positive Rate (Statistical Power)

We quantify power as the proportion of simulated datasets for which the DNT and DMT p-values are significant at $\alpha = 0.05$, and when the drug is in the top promising ten ranked drugs for each test. The Drug P-Score power is the proportion of simulated datasets for which the P-Score for olaparib exceeds the empirical 95th percentile of P-Score from two simulated independent uniform distributions and the proportion for which olaparib is in the top ten drugs ranked by P-Score. We justify using the metric of the drug appearing in the top 10 drugs for a simulated case as important to ensure that even if the inflated false positive rates means too many drugs have a small p-value, at least the top drugs contain olaparib, and is not drowned out by other drugs that may target some of the genes in the set of genes.

Table 2. Simulation of three scenarios to estimate FPR and TPR using SOLID and TCGA normal adjacent data.

	SOLID, healthy normals (M=8)	SOLID, healthy normals / TCGA normal adjacent (M=8)		
Scenario	**FPR (DNT/DMT)**	**TPR(Set A)**	**TPR(Set B)**	**TPR(Set C)**
S1: Complete Cancer Driver Genes and Drugs	5% / 18%	40% / 35%	5% / 50%	5% / 30%
S2: Drivers-Only, Complete Drugs	5% / 9%	5% / 32%	5% / 45%	5% / 25%
S3: Reduced Cancer Drivers and Drugs	5% / 8%	NA	NA	NA

Supplemental_Figure_S5 shows the true positive rate results for spiked-in expression in gene sets A, B, and C (rows) at the two standard deviation level, with simulation scenarios S1 and S2 in the columns. We observe within each of the six sub-figures, power increases as matched set sizes (M) go from 2 to 4 to 8, but the gains level off between *M*=8 and M=16 (Table 2 consists of the values for M=8). Figure S6 shows the same six sub-figures for simulations, but with spiked-in expression at three standard deviations higher than the mean for sets A, B, and C. Focusing on scenario S2 and M=8, we see the true positive rates in set A for DNT and DMT increase from 60% and 40% to 90% and 45%, respectively. Similarly, we see true positive rates in Set C for DNT and DMT increase from 10% and 40% to 25% and 45%, respectively. We conclude that the DNT (darkest bar) performs well in for gene set A as expected, but were surprisingly low in Sets B and C, with measurable power only in Supplemental_Figure_S6. We see a slightly higher power for S1 than S2, but S2 is preferred because of the greatly-reduced false positive rates. The DMT slightly underperformed the DNT for gene Set A, but strongly outperforms the DNT for gene Sets B and C as over-expressed with power levels. The true positive rates with the top-10 are similar to the tests evaluated at α=0.05, which validates that olaparib is both statistically significant and one of the top-10 drugs found in those simulations.

Examples: Patients with Refractory Cancer

PANOPLY drug predictions for a patient with refractory colon cancer

We downloaded clinical, CNA, and normalized gene expression data for TCGA colon adenocarcinoma (COAD) patients directly from Synapse (SAGE reference) via R. As no somatic mutation data were available for the Synapse COAD module, we restricted the analysis to male patients age 45-70 with Stage II or higher disease, who had both CNA and normalized RNA gene expression data. Four cases that had a recurrent tumor and died within 155 days of initial diagnosis were chosen. For each case, five controls with same clinical characteristics as the case but who had remained disease-free for more than 1000 days (three years) were chosen.

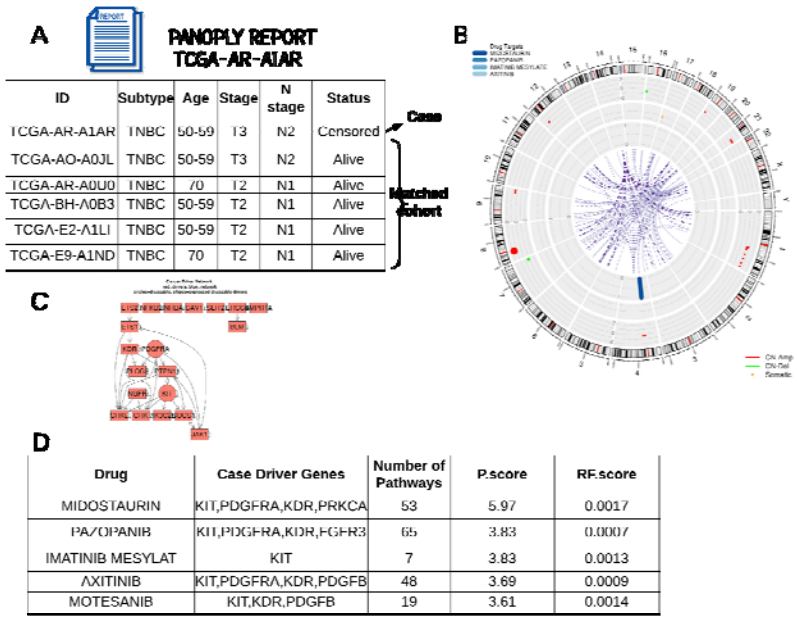
In here, we describe one colon cancer case's molecular profile (TCGA-AA-3488) compared to controls. The case displayed 645 driver genes: 226 CNAs not present in the case germline sample or matched controls tumor or germline samples; in addition, 419 genes were over-expressed in the case tumor sample relative to the matched control samples. Of the 226 CNAs, 113 were amplifications, and 113 were deletions. Of those *BRAF, CDKN2A, CDKN2B, FGF10, IL7R, INHBA, JAK2, KEL, MAFA, NTRK1, NTRK2, PIK3CA, PRSS1, RBL1, SKIL, SMO, SOX2* and *SPTA1* cancer-related genes were both amplified and overexpressed.

Of the case's 645 driver genes, 53 genes are differentially expressed between the case and matched controls and can be targeted by antineoplastic drugs. Based on the network and random forest analysis of driver genes and gene expression data, PANOPLY ranked the following drugs lestaurtinib (*JAK2, NTRK1, NTRK2*), LY2784544 (*JAK2*), GDC-0032(*PIK3CA*), NVP-BGT226 (*MTOR, PIK3CA*), regorafenib (*MAPK11, RAF1, BRAF, KRAS, KIT, FGFR1/2,PDGFRA/B, ARAF, KDR, EPHA2, ABL1, NTRK1, CYP3A4, CYP2C8/9, CYP2B6, CYP2C19,ABCB1, ABCG2, UGT1A1/9,FLT1/4, RET, TEK and DDR2*) and ARQ736 (BRAF) as significant for patient TCGA-AA-3488 with significant P.score and RF.score (1.5.a section in

Supplementary Material and Supplemental_Table_S1).As shown in Supplemental_Table_S1, the six drugs target a number of driver genes and pathways. The details of the genes that are specifically altered in the case (*PIK3CA, JAK2, BRAF, NTRK1, NTRK2, MTOR, KIT, KRAS, and FLT4*) and can be targeted by a drug are listed in Supplemental_Table_S2.

PANOPLY drug predictions for a patient with refractory breast cancer

The somatic mutations, gene expression and CNA data of breast cancer patients were obtained



ID	Subtype	Age	Stage	N stage	Status
TCGA-AR-A1AR	TNBC	50-59	T3	N2	Censored
TCGA-AO-A0JL	TNBC	50-59	T3	N2	Alive
TCGA-AR-A0U0	TNBC	70	T2	N1	Alive
TCGA-BH-A0B3	TNBC	50-59	T2	N1	Alive
TCGA-E2-A1LI	TNBC	50-59	T2	N1	Alive
TCGA-E9-A1ND	TNBC	70	T2	N1	Alive

Case

Matched cohort

Drug	Case Driver Genes	Number of Pathways	P.score	RF.score
MIDOSTAURIN	KIT,PDGFRA,KDR,PRKCA	53	5.97	0.0017
PAZOPANIB	KIT,PDGFRA,KDR,FGFR3	65	3.83	0.0007
IMATINIB MESYLAT	KIT	7	3.83	0.0013
AXITINIB	KIT,PDGFRA,KDR,PDGFB	48	3.69	0.0009
MOTESANIB	KIT,KDR,PDGFB	19	3.61	0.0014

Figure 3: PANOPLY report. A) Case and matched cohort samples selected for PANOPLY analysis B) visualization of unique genomic features (Red and green –CNA, blue somatic, inner circle connections of driver genes C) Network analysis shows top driver genes (somatic, overexpressed and CNA) connections unique to this patient. The network consists of driver genes (red circle if druggable, red square if not) that are both over-expressed and druggable.D) Top drugs predicted to target the TNBC cancer for TCGA-AR-A1AR.

from the TCGAbiolinks data portal via R. A patient with TNBC (TCGA-AR-A1AR) who survived only 1.4 years after diagnosis was matched to five controls: TNBC patients who survived for over 3.5 years and were of similar age and had similar clinical characteristics (Figure 3A). The case tumor sample displayed 175 genomic alterations (CNAs and somatic mutations) and 183 over-expressed genes compared with samples from the matched controls. Circos plot shows

somatic mutations, copy number amplifications, and deletions that are specific to the case are represented in Figure 3B. As shown in the Figure 3B, chromosomes 1 and eight consists of huge amplifications in the case along with a somatic *TP53* mutation. Of the 258 alterations, TCGA-AR-A1AR consists of *ABL1, NUP214, PDGFRA, RAB14, RAC1*, and *SET* genes both amplified and overexpressed; *RPS24* gene consists of copy number loss and is up-regulated; *TP53* gene consists of somatic mutation and a copy number loss. The rest of the alterations in the other genes for this case consists of one type of alteration event. The details of the alteration events for this case are provided as a detailed report at http://bioinformaticstools.mayo.edu/research/panoply/

Based on the network and random forest analysis of driver genes, PANOPLY ranked the drugs for this patient. The Figure 3C presents the significant network of the case TCGA-AR-A1AR, consisting of only driver genes (red circles if druggable, red squares if they are not druggable, and ellipse if genes are over-expressed and druggable). The most compatible drugs, the driver genes, the number of pathways targeted by a drug, P.score, and RF.score are presented Figure 3D. The top drug Midostaurin can target four (*KDR, KIT, PDGFRA*, and *PRKCA*) out of 18 genes in the drug-gene network. Of the four driver genes, *KDR, PDGFRA*, and *KIT* genes have copy number amplification and overexpression of genes compared to its matched controls and TCGA breast normal adjacent samples; *PRKCA* gene is downregulated compared to matched controls and TCGA breast normal adjacent samples. The other recommended drugs target two highly connected cancer genes *FGFR3* and *PDGFB*. As shown in 1.5.b section of Supplemental_Table_S3, the five drugs target a number of driver genes and pathways. The details of the genes that are uniquely altered in the case (*KDR, KIT, PDGFRA, FGFR3, PDGFB*, *PRKCA)* and can be targeted by a drug are listed in Supplemental_Table_S3.

PANOPLY Drug Predictions for Patients with Chemoresistant, Triple-Negative Breast Cancer in the BEAUTY study Clinical, genomics and drug response data was available from a

neoadjuvant clinical trial [39] of 16 weeks of paclitaxel and anthracycline/cyclophosphamide +/- anti-Her2 therapy in women with Stage I-III breast cancer (referred to as BEAUTY trial). Patients with no residual disease in the breast found at surgery were considered to have chemo-sensitive tumors, and patients with the residual disease in the breast were considered to have chemo-resistant tumors. Raw data from the RNAseq and whole-exome sequencing of diagnostic biopsy from a cohort of 17 BEAUTY patients with chemoresistant TNBC and 22 age-matched BEAUTY patients with chemo-sensitive basal TNBC were processed through Mayo Clinic best-practice bioinformatics pipelines [38,40] and used in this present analysis to investigated PANOPLY's ability to mine individual case reports and compare groups of cases and controls.Of the 374 drugs that were investigated, 30 drugs were removed from the analysis as they were not listed as the prioritized drug by PANOPLY for any of the 17 cases. NMF clustering of the samples and drugs was performed with the prioritized scores from 344 drugs. Based upon the cophenetic and average silhouette scores, two clusters were selected to be optimal for 17 cases. The percentile ranking of top 10% (35/344) drugs were aggregated per sample cluster using the median score and presented as a heatmap (Figure 4A). The target genes of the drug clusters were collated, and a word cloud was generated with the targets (Figure 4B). As shown in Figure 4A, the cluster 1 consists of nine samples; the patients in that group primarily consist of kinase-inhibitors as their top prioritized drugs (drugs=16) and the drugs in that cluster can predominantly target the PIK3CA-mTOR-AKT signaling pathways. The cluster 2 from Figure 4A consists of a set of prioritized drugs (drugs=19) for eight patients, as shown in Figure 4B these drugs can primarily target genes associated with cell cycle control, specifically targeting the histone deacetylases (*HDAC1*) and the aurora kinases A and B.

As described in Supplemental_Figure_S3, the drug, and intended protein target annotation contains inherent redundancies, created by variations of drug formulations and Pharmaceutical branding. Therefore, using binary clustering (Supplemental_Figure_S3), we have reduced the

number of drugs from 374 to 117, with the most distinct targets. To investigate further the drugs driving the NMF clustering, we identified a total of 3/35 drugs (drugs= 16+19 =35 from two clusters) after removing the most correlated drugs and target genes in the two clusters;

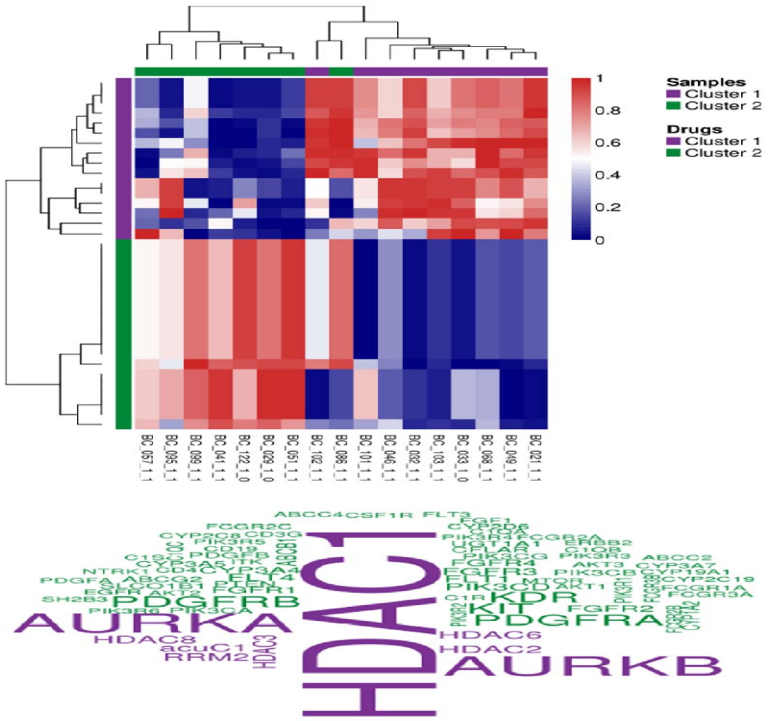


Figure 4. **Clustering and wordcloud plots of 17 TNBC chemotherapy resistant patients**. A) Two-way hierarchical clustering of the top 25% of the drugs predicted by PANOPLY for 17 basal TNBC patients. The heatmap shows that there are two sample and drug clusters are implicated in the NMF clustering analysis. B) Word cloud of the target genes from the two drug clusters predicted by the NMF analysis.

specifically, there was one drug (CAL-101, targeting *PIK3CD*) in cluster 1 and two drugs (Vorinostat, targeting histone deacetylases and Triapine, targeting RRM2) in cluster 2.

A similar analysis was performed for the five LAR patients (BC016, BC040, BC056, BC096, BC116) with chemoresistance TNBC disease and compared them to basal chemosensitive age-matched controls. The PANOPLY analysis identified PF-04691502 as the cell cycle inhibiting drug targeting *PTEN, PIK3R1*, and *PIK3CA* driver genes for the BC016 patient. For patient B040, the PANOPLY identified IMATINIB as the top drug targeting tyrosine kinases (CYP proteins, platelet derived growth factor receptors A and B, solute carrier family 22 members 1

and 2, ATP binding cassette subfamily members, etc). Similarly, for patient BC056, the PANOPLY identified DOVITINIB (target genes – KIT, FGFR 1-4, PDGFR A and B, *CFLAR, KDR*, and *NTRK1*). Patient BC096 was the only sample which specifically targeted hormone signaling with the Ethinyl Estradiol prioritized by PANOPLY. Finally, for patient BC116 the PANOPLY prioritized XL765 as promising drug, this drug target *PTEN, MTOR* and *PIK3R1* genes.

Validation of a PANOPLY-Prioritized Drug using PDX models from a Chemoresistant Breast Cancer patient

Next, we used PANOPLY to predict and validate promising drugs for a chemo-resistant TNBC

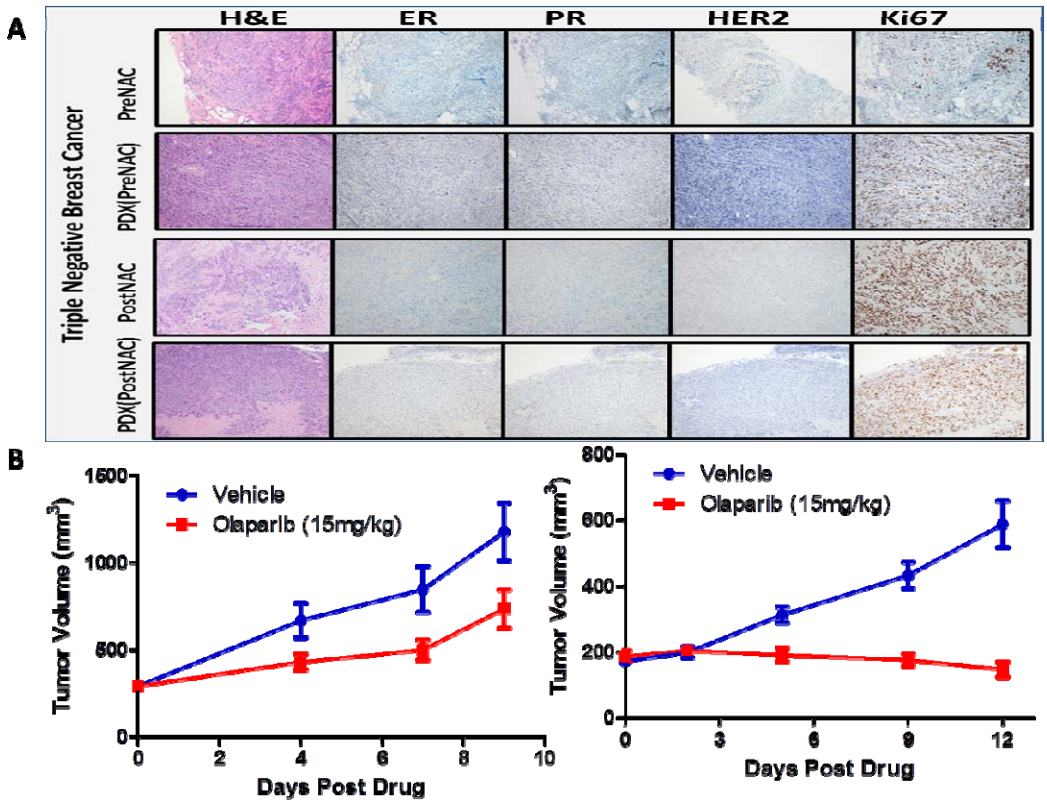


Figure 5. Patient-derived xenografts (PDXs) validate PANOPLY's prediction that olaparib is an effective treatment for a patient with chemoresistant TNBC (BC_051_1_1). A) The top panel shows histological stains of the patient's tumor and PDXs (pre or post-treatment). B) Cytotoxicity data shows the PDXs response to the top predicted drug olaparib compared to the no treatment (the left plot shows the olaparib drug response data from pre-treatment mice, whereas the data from the right shows the data from post-treatment PDX models. Both the datasets were generated using the Vehicle as controls).

patient from the BEAUTY cohort. The PANOPLY report for this patient (BTN051) is available at http://bioinformaticstools.mayo.edu/research/panoply/. Somatic mutations, germline mutations, CNA, gene expression, and expressed mutation data from the case and nine age-matched patients with chemosensitive TNBC was obtained for PANOPLY analysis (Supplemental_Table_S4). In the case, somatic alterations were detected in nine genes (non-synonymous mutations in *FOXD3, TP53, DCC, CFTR,* stop gain in *BRCA1* and frameshift mutations in *PTEN, PALB2*, and *NF1*). In addition, 111 copy number amplifications (affecting *IRF4, MYC*, and other known cancer genes; and 119 copy number deletions (PANOPLY case report **T**N_Beauty_Patient.pdf is comprehensively described in in Supplementary Material 1.6 section) were detected. As shown in the circos plot in Supplemental_Figure_S7, orange indicates somatic mutations (9) and the inner, blue dots represent 18 rare germline variants that are unique to the case (18). Based on P.score and RF.score (network Supplemental_Figure_ S8 and random forest analysis), PANOPLY indicated that olaparib was the most promising drug for this patient (from Supplementary Material 1.6 section Table S5). Olaparib targets the patient's driver genes *BRCA1* (which harbored a stop-gain somatic mutation and a deletion), *BRCA2* (upregulation and deletion, possible LOH), *ATR* and *ATM* (deletion), and *PARP1* (overexpression), as well as 43 biological pathways that these genes are involved Supplemental_Table_S6. Additional details of the promising drugs and driver genes are presented in Supplemental_Table_S5 and Supplemental_Table_S6. Although olaparib was predicted as an alternate drug for the patient based on molecular data from the primary tumor, we have also tested the efficacy of the drug in both pre and post-NAC PDX models. Figure 5A shows histologic images from the patient's tumor and a corresponding PDX, both from the pre-NAC and post-NAC) time points. Pre-NAC, the tumor was an invasive ductal TN carcinoma. Pre- and post-NAC, the patient tumor, and its corresponding PDX had similar morphologic features and a triple negative staining pattern (Figure 5A). These findings suggest that the PDX are representative of the patient's tumor, both pre-NAC, and post-NAC.

Pre-NAC PDXs were implanted into 14 non-skid mice. Mice were randomized to either olaparib or vehicle when the tumor had grown to 200–250 mm^3, mouse tumor size, and body weights were measured twice per week. Twelve days after starting treatment, the mice were sacrificed. This experiment was repeated using post-NAC PDX's. For both the pre-NAC and post-NAC PDXs, tumor volume at day 12 was significantly lower in the olaparib group than in the vehicle group (Wilcoxon rank sum test P=0.04 and p <0.001 respectively; Figure 5B). At the post-NAC results shows promise that this approach may be successful in identifying an effective therapy. Unfortunately, seven months after adjuvant anthracycline chemotherapy ended, the patient experienced distant brain progression followed by systemic recurrence in the lung and bone and subsequent death. Because PANOPLY was still under development at the time of her recurrence, the drug identified by PANOPLY was not provided to the patient in a timely manner.

DISCUSSION

In creating PANOPLY, our goal was to develop a flexible workflow capable of analyzing multiple forms of -omics and clinical data to identify driver genes, their effects on gene networks and the drugs capable of targeting those altered pathways in patients with cancer. Here, using expression, CNA, somatic and germline alterations data from publicly available and in-house data sets, we demonstrated that PANOPLY holds promise in identifying agents capable of targeting driver gene–induced changes (from gene expression, somatic mutations, germline mutation, and CNA data), both for individual patients and for subgroups of patients that share a cancer subtype. This is evident in PANOPLY's prioritization that olaparib was promising for a patient with chemoresistant TNBC; that agent was confirmed to reduce tumor size when applied to that patient's xenografts.

PANOPLY currently analyzes the patient's molecular and clinical data together along with the knowledge databases such as Reactome, DGI-db, and others for sophisticated drug-gene

network analysis. This represents a substantial advancement of our algorithm relative to existing programs, such as XSeq [16], OncoRep [17], and OncoIMPACT [18], which can only integrate molecular data such as somatic mutations or CNAs, and gene expression. Currently, in the clinical setting, oncologists are using genomics reports from a limited target panel for precision care. Those reports do not include any additional analysis to prioritize genes or pathways nor do they provide any information on drugs and druggable targets.

Working closely with clinicians, basic scientists, and pharmacologists, we have developed PANOPLY to integrate molecular, clinical and drug data to prioritize targets and facilitate individualized treatment for the patients. PANOPLY provides a personalized list of prioritized drugs along with the literature links on the effectiveness of drugs in that cancer types based on the clinical and molecular profile. The oncologist will still have to go through the list and refine drug selection based on the inherent clinical knowledge, ancillary clinical trials, and insurance coverage availability. However, a few limitations of PANOPLY are that the method still cannot delineate the clinical effectiveness of a similar class of drugs. We plan to accomplish this in future by bringing in drug knowledge such as chemical structure, molecular size, and drug dosage. Our method is highly dependent on drug-gene target annotations that are heavily biased by product literature and databases. Moreover, clinical translation of the PANOPLY is constrained by the cost effectiveness in developing PDXs for drug validation.

PANOPLY is a flexible framework that can integrate many other types of –omics data, including structural variants, circular RNAs, long non-coding RNAs, and fusion data with modifications to its code. PANOPLY's framework can also be extended to the metastatic setting; additional clinical data such as prior exposure to drugs, features of primary and recurrence disease would be required. Additional genomics data sets are needed to modify the existing approach for metastatic tumors.

As more is learned about the molecular underpinning of cancer using various resources, we plan to expand our knowledge base to improve our PANOPLY's predictions. We will also build an individualized medicine cancer database with reports and tools to search our PANOPLY knowledge base easily. We have validated PANOPLY's predictions in a single patient at present using PDXs, in future, we plan to validate PANOPLY-predicted drugs in PDXs derived from additional patients. As several other groups and we have shown, PDX models faithfully represent tumor biology [41 42-44], so these results should provide insight into PANOPLY's reliability. However, PANOPLY's predictions will need to be tested in human trials. To this end, multi-omics data are being collected prospectively in the context of a Mayo Clinic study designed to identify biomarkers of response to palbociclib and endocrine therapy. In this study, PANOPLY predictions will be provided to clinicians to guide subsequent treatments following progression on this therapy. With modifications to its computational framework, PANOPLY can also be applied to study other diseases.

In conclusion, our results indicate that combining multiple sources of -omics and clinical data to predict promising agents for a patient or groups of patients with cancer is feasible. With further validation, PANOPLY can be a tool to help clinicians in their decision-making process.

ACKNOWLEDGEMENTS

KRK, JPS, KJT, XT, EEC, JY, PTV, JNI, RMW, LW, JCB, MPG, and VJS are funded in part by the Mayo Clinic Center for Individualized Medicine. KRK, RMW, LW, MPG, JNI, and VJS are funded partially by the Mayo Clinic Breast Specialized Program of Research Excellence (SPORE) (P50CA116201) and MPG, VJS are funded by the Mayo Comprehensive Cancer Center Grant (P30CA 15083-43). KRK, JS, KJT, XT, EEC, and PTV are funded by the Division of Biostatistics and Informatics at the Mayo Clinic.

REFERENCES

1. CDC: Leading causes of death, 2015

2. Perez EA, Reinholz MM, Hillman DW, et al: HER2 and chromosome 17 effect on patient outcome in the N9831 adjuvant trastuzumab trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 28:4307-15, 2010

3. Perez EA, Dueck AC, McCullough AE, et al: Predictability of adjuvant trastuzumab benefit in N9831 patients using the ASCO/CAP HER2-positivity criteria. Journal of the National Cancer Institute 104:159-62, 2012

4. Gambacorti-Passerini C, Antolini L, Mahon F, et al: Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. J Natl Cancer Inst 103:553-61, 2011

5. Chapman P, Hauschild A, Robert C, et al: Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 364:2507-16, 2011

6. Mok T, Wu Y, Thongprasert S, et al: Gefitinib or caboplatin-paclitaxel in pulmonary adenocarcinoma N Engl J Med 361:947-57, 2009

7. Kwak E, Bang Y, Camidge D, et al: Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 363:1693-703, 2010

8. André F, Bachelot T, Commo F, et al: Comparative genomic hybridsation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). Lancet Oncol 15:267-74, 2014

9. Meric-Bernstam F, Johnson A, Holla V, et al: A decision support framework for genomically informed investigational cancer therapy. Journal of the National Cancer Institute 107, 2015

10. Yeh P, Chen H, Andrews J, et al: DNA-Mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): a catalog of clinically relevant cancer mutations to enable genome-directed anticancer therapy. Clinical cancer research : an official journal of the American Association for Cancer Research 19:1894-901, 2013

11. Van Allen EM, Wagle N, Stojanov P, et al: Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nature medicine 20:682-8, 2014

12. Cerami E, Gao J, Dogrusoz U, et al: The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery 2:401-4, 2012

13. Forbes SA, Beare D, Gunasekaran P, et al: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research 43:D805-11, 2015

14. Bashashati A, Haffari G, Ding J, et al: DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome biology 13:R124, 2012

15. Gundem G, Perez-Llamas C, Jene-Sanz A, et al: IntOGen: integration and data mining of multidimensional oncogenomic data. Nature methods 7:92-3, 2010

16. Ding J, McConechy MK, Horlings HM, et al: Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nature communications 6:8554, 2015

17. Meissner T, Fisch KM, Gioia L, et al: OncoRep: an n-of-1 reporting tool to support genome-guided treatment for breast cancer patients using RNA-sequencing. BMC medical genomics 8:24, 2015

18. Bertrand D, Chng KR, Sherbaf FG, et al: Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic acids research 43:e44, 2015

19. Dong C, Guo Y, Yang H, et al: iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. Genome medicine 8:135, 2016

20. Croft D, Mundo AF, Haw R, et al: The Reactome pathway knowledgebase. Nucleic acids research 42:D472-7, 2014

21. Peter T. Vedell EWK, Gavin R. Oliver, Krishna R. Kalari, Steven N. Hart, Jean-Pierre A. Kocher, Hugues Sicotte Identification of Genes Implicated in Tumorigenesis for Interpretation of Tumor Genomes Individualizing Medicine Conference. Rochester, MN, USA, 2013

22. FoundationMedicine,

23. Hart SN, Ellingson MS, Schahl K, et al: Determining the frequency of pathogenic germline variants from exome sequencing in patients with castrate-resistant prostate cancer. BMJ open 6:e010332, 2016

24. Cheng L, Schneider BP, Li L: A bioinformatics approach for precision medicine off-label drug drug selection among triple negative breast cancer patients. Journal of the American Medical Informatics Association : JAMIA 23:741-9, 2016

25. Hansen KD, Irizarry RA, Wu Z: Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13:204-16, 2012

26. Larson DE, Abbott TE, Wilson RK: Using SomaticSniper to Detect Somatic Single Nucleotide Variants. Current protocols in bioinformatics 45:15 5 1-8, 2014

27. Roth A, Ding J, Morin R, et al: JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 28:907-13, 2012

28. do Valle IF, Giampieri E, Simonetti G, et al: Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC bioinformatics 17:341, 2016

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

30. Luo W, Friedman MS, Shedden K, et al: GAGE: generally applicable gene set enrichment for pathway analysis. BMC bioinformatics 10:161, 2009

31. Stouffer SA, Suchman, E.A., DeVinney, L.C., Star, S.A. &, Williams RMJ: The American Soldier: Adjustment during Army Life. Princeton., Princeton University Press, 1949

32. Lin DY: On rapid simulation of P values in association studies. American journal of human genetics 77:513-514, 2005

33. Brown MB: A method for combining non-independent, one-sided tests of significance. Biometrics 31:987-992, 1975

34. Breiman L: Random forests. Machine Learning 45:5-32, 2001

35. Matthew P. Goetz M, Krishna R. Kalari, PhD, Vera J. Suman, PhD, Ann M. Moyer, MD, Jia Yu, PhD, Daniel W. Visscher, MD, Travis J. Dockter, MS, Peter T. Vedell, PhD, Jason P. Sinnwell, MS, Xiaojia Tang, PhD, Kevin J. Thompson, PhD, Poulami Barman, MS, Douglas W. Mahoney, MS, Erin Carlson, BA, Steven N. Hart, PhD, Ping Yin, MD, Bo Qin, PhD, Sara J. Felte, BA, Sarah A. McLaughlin, MD, Alvaro Moreno-Aspitia, MD, John A Copland, Donald W. Northfelt, MD, Richard. Gray, MD, Katie Hunt, MD, Amy Conners, MD, Hugues Sicotte, PhD, Jeanette E. Eckel-Passow, PhD, Jean-Pierre Kocher, PhD, James N. Ingle, MDa, Marissa. Ellingson, MS, Michelle McDonough, MD, Eric D. Wieben, PhD, Richard Weinshilboum, MD, Liewei Wang, MD, PhD, and Judy C. Boughey, MD: Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer. JNCI in Press, 2017

36. Gaujoux R, Seoighe C: A flexible R package for nonnegative matrix factorization. BMC bioinformatics 11:367, 2010

37. Radovich M, Clare SE, Atale R, et al: Characterizing the heterogeneity of triple-negative breast cancers using microdissected normal ductal epithelium and RNA-sequencing. Breast Cancer Research and Treatment 143:57-68, 2014

38. Kalari KR, Nair AA, Bhavsar JD, et al: MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. BMC bioinformatics 15:224, 2014

39. Goetz MP, Kalari KR, Suman VJ, et al: Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer. JNCI: Journal of the National Cancer Institute 109, 2017

40. Asmann YW, Middha S, Hossain A, et al: TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. Bioinformatics 28:277-8, 2012

41. Hou J, Wang L: FKBP5 as a selection biomarker for gemcitabine and Akt inhibitors in treatment of pancreatic cancer. PloS one 7:e36252, 2012

42. Swinnen JV, Beckers A, Brusselmans K, et al: Mimicry of a cellular low energy status blocks tumor cell anabolism and suppresses the malignant phenotype. Cancer Res 65:2441-8, 2005

43. Liu B, Fan Z, Edgerton SM, et al: Metformin induces unique biological and molecular responses in triple negative breast cancer cells. Cell Cycle 8:2031-40, 2009

44. Iliopoulos D, Hirsch HA, Struhl K: Metformin decreases the dose of chemotherapy for prolonging tumor remission in mouse xenografts involving multiple cancer cell types. Cancer Res 71:3196-201, 2011