

UC Davis

UC Davis Previously Published Works

Title

Systematic Establishment of Robustness and Standards in Patient-Derived Xenograft Experiments and Analysis

Permalink

<https://escholarship.org/uc/item/4wx0v318>

Journal

Cancer Research, 80(11)

ISSN

0008-5472

Authors

Evrard, Yvonne A
Srivastava, Anuj
Randjelovic, Jelena
et al.

Publication Date

2020-06-01

DOI

10.1158/0008-5472.can-19-3101

Peer reviewed



Published in final edited form as:

Cancer Res. 2020 June 01; 80(11): 2286–2297. doi:10.1158/0008-5472.CAN-19-3101.

Systematic Establishment of Robustness and Standards in Patient-Derived Xenograft Experiments and Analysis

Yvonne A. Evrard¹, Anuj Srivastava², Jelena Randjelovic³, The NCI PDXNet Consortium, James H. Doroshow⁴, Dennis A. Dean II³, Jeffrey S. Morris⁵, Jeffrey H. Chuang^{2,6}

¹Leidos Biomedical Research, Inc, Frederick National Laboratory for Cancer Research, Frederick, MD

²The Jackson Laboratory for Genomic Medicine, Farmington, CT

³Seven Bridges Genomics, Cambridge, MA

⁴Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD

⁵The University of Texas M.D. Anderson Cancer Center, Houston, TX

⁶University of Connecticut Health Center, Farmington, CT

Abstract

Patient-Derived Xenografts (PDXs) are tumor-in-mouse models for cancer. PDX collections, such as the NCI PDXNet, are powerful resources for preclinical therapeutic testing. However, variations in experimental and analysis procedures have limited interpretability. To determine the robustness of PDX studies, the PDXNet tested temozolomide drug response for three pre-validated PDX models (sensitive, resistant, and intermediate) across four blinded PDX Development and Trial Centers (PDTCs) using independently selected SOPs. Each PDTc was able to correctly identify the sensitive, resistant, and intermediate models, and statistical evaluations were concordant across all groups. We also developed and benchmarked optimized PDX informatics pipelines, and these yielded robust assessments across xenograft biological replicates. These studies show that PDX drug responses and sequence results are reproducible across diverse experimental protocols. In addition, we share the range of experimental procedures that maintained robustness, as well as

Person to whom reprint requests should be addressed: Jeffrey H. Chuang, The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT USA 06032, Tel: 860-837-2473, jeff.chuang@jax.org.

All authors in this publication are part of the NCI PDXNet Consortium. Additional contributing members are: Baylor College of Medicine, Houston, TX (Salma Kaochar, Michael T. Lewis, Nicolas Mitsiades); Frederick National Laboratory for Cancer Research, Frederick, MD (Li Chen, Rajesh Patidar); The Jackson Laboratory for Genomic Medicine, Farmington, CT (Peter N. Robinson, Zi-Ming Zhao); The Jackson Laboratory, Bar Harbor, ME (Carol J. Bult, Michael Lloyd, Steven Neuhauser, Xing Yi Woo); National Cancer Institute, Investigational Drug Branch, Bethesda, MD (Jeffrey A. Moscow); Seven Bridges Genomics, Inc., Cambridge, Charlestown, MA (Brandi Davis-Dusenbery, Jack DiGiovanna, Christian Frech, Ryan Jeon, Nevena Miletic, Jacqueline Rosains, Isheeta Seth, Tamara Stankovic, Adam Stanojevic); University of California School of Medicine, Davis, CA (Luis Carvajal-Carmona, Moon Chen, Chong-Xian Pan); The University of Texas M.D. Anderson Cancer Center, Houston, TX (Huiqin Chen, Michael Davies, Bingliang Fang, Min Jin Ha, Funda Meric-Bernstam, Jack Roth); University of Utah Huntsman Cancer Institute, Salt Lake City, UT (Sasi Arunachalam, David Nix, Alana L. Welm, Bryan E. Welm); Washington University School of Medicine in St. Louis, St. Louis, MO (Sherri Davies, Li Ding, Ramaswamy Govindan, Shunqiang Li, Cynthia Ma, Brian A. Van Tine); The Wistar Institute, Philadelphia, PA (Meenhard Herlyn, Andrew Kossenkov, Vito Rebecca, Jayamanna Wickramasinghe, Min Xiao)

standardized cloud-based workflows for PDX exome-seq and RNA-Seq analysis and for evaluating growth.

Keywords

Xenograft Experiments; PDX; PDXNet; PDTC; Temozolomide

Introduction

Patient-Derived Xenografts (PDX) are *in vivo* preclinical models in which human cancers are engrafted into a mouse for translational cancer research and personalized therapeutic selection (1–4). Prior studies have shown that treatment responses of tumor-bearing mice usually reflect the responses in patients (5,6). PDXs have been used successfully for preclinical drug screens (4,5), to facilitate the identification of potential biomarkers of drug response and resistance (4,7), to select appropriate therapeutic regimens for individual patients (8), and to measure evolutionary processes in cancer in response to treatment (9). At the genomic level, engrafted human tumors have been shown to retain most genomic aberrations from the original patient tumor (8,10). These successes have led to the development of a number of PDX collections in both academia and industry (5,11,12) for use in preclinical testing.

Despite these successes, important questions remain for the use of PDXs as a model system for treatment response. The reproducibility of treatment response has not been well-evaluated because research teams often perform experiments in models that are not used by other groups. Variations in engraftment, dosing, and response assessment protocols also frustrate comparisons of results. Moreover, intratumoral heterogeneity, genetic drift and selection during tumor collection, engraftment, and xenograft passaging can result in genomic variation among primary tumor samples and derived xenografts (10,13). Whether such variation impacts the accuracy of PDXs as a preclinical model has been unclear. Resolution of this issue requires not only controlled treatment replicates but also standardized PDX-specific sequence analysis pipelines to robustly identify genomic aberrations. Progress on these topics is important to the overall field of cancer patient-derived models, as analogous concerns pertain for organoids and other 3D culture systems.

To resolve such questions for the use of PDXs in precision medicine, the US National Cancer Institute has supported a consortium of PDX-focused research centers, the NCI PDXNet. Here we in the PDXNet consortium report the results of experiments to test the robustness of PDX treatment responses across different research centers, using temozolomide treatment on three models because of prior data on their temozolomide responses from the NCI Patient Derived Models Repository (PDMR). We report on replicate evaluations across four additional PDX Development and Trials Centers (PDTC) using blinded treatment and response evaluation protocols. Simultaneously, we have performed exome and RNA sequencing at each center to determine biological and technical stability of genomic characterizations of samples from each center. These sequence analyses have been performed with optimized analysis pipelines chosen based on an extensive new

benchmarking of pipelines from each center on synthetic sequence sets. Finally, we have statistically analyzed the cohort growth curves for each model in each research center using five separate metrics. These studies allow us to answer whether PDXs have sufficiently robust behaviors to withstand variations in experimental procedures, response measurement algorithms, genomic variation among replicates, and alternative sequence analysis protocols. We also report effective SOPs for experimental procedures, pipelines for statistical assessment of response, and sequence analysis workflows. We expect these standards to advance the use of PDXs and other in vivo models in cancer precision medicine, a critical need for the evaluation of PDX results in the context of moving novel therapeutics or therapeutic combinations to the clinic.

Methods

Animal Models

Three PDX models were selected based solely on their temozolomide responsiveness. They were 625472–104-R (colon adenocarcinoma), 172845–121-T (colon adenocarcinoma), and BL0293-F563 (urothelial/bladder cancer). Cryopreserved PDX tumor fragments were shipped from the PDMR to the individual PDTCs including Huntsman Cancer Institute/Baylor College of Medicine (HCI-BCM), MD Anderson Cancer Center (MDACC), Washington University-St. Louis (WUSTL), and The Wistar Institute/University of Pennsylvania/MDACC (WIST)., implanted for initial expansion and then passaged for the preclinical study. Briefly, cryopreserved PDX material was prepared into implantation size pieces as outlined in Table 1. The PDX material plus a drop of Matrigel (BD BioSciences, Bedford, MA.) was then implanted subcutaneously in NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG) host mice. Mice were housed in sterile, filter-capped polycarbonate cages, maintained in a barrier facility on a 12-hour light/dark cycle, and were provided sterilized food and water, ad libitum. Animals were monitored weekly for tumor growth. The initial passage of material was grown to approximately 1000–2000 mm³ calculated using the following formula: tumor volume (mm³) = (tumor length x [tumor width]²)/2 (14). Tumor material was then harvested, a portion cryopreserved, and the remainder implanted into NSG host mice for the preclinical drug study. Related patient data, clinical history, representative histology and short-tandem repeat profiles for the PDX models can be found at <https://pdmr.cancer.gov>; model BL0293-F563 was originally developed by The Jackson Laboratory (tumor model TM00016, <http://tumor.informatics.jax.org/mtbwi/pdxSearch.do>).

Preclinical Studies

Specific tumor staging size, implantation method, and cohort size at the PDMR and each PDTC are outlined in Table 1 based on each site's standard practices. In general, tumors were staged to a preselected size (weight = 100–200 mm³). Tumor-bearing mice were randomized before initiation of treatment and assigned to each group. Body weight was monitored 1–2 times weekly and tumor size was assessed 2–3 times weekly by caliper measurement. For all sites, drug studies were performed at passage 3 for 625472–104-R, passage 4 for 172845–121-T, and passage 6 for BL0293-F563 (passage 0 = first implanted host). Temozolomide (NSC 362856) was obtained from the Developmental Therapeutics Program, NCI and administered at the times and doses indicated in Table 1. Animals were

sacrificed when the tumors reached an individual PDTC's animal welfare endpoint or a maximum tumor size; if tumor growth delay was observed a tertiary endpoint was used by some sites (Table 1).

Ethics Statement

The Frederick National Laboratory for Cancer Research (location of the PDMR) is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International and follows the USPHS Policy for the Care and Use of Laboratory Animals. All the studies were conducted according to an approved animal care and use committee protocol in accordance with the procedures outlined in the "Guide for Care and Use of Laboratory Animals" (National Research Council; 1996; National Academy Press; Washington, D.C.).

All patients and healthy donors gave written informed consent for study inclusion and were enrolled on institutional review board-approved protocols of record for the sites that developed the PDX models (DCTD, NCI and The Jackson Laboratory). The study was performed in accordance with the precepts established by the Helsinki Declaration. The study design and conduct complied with all applicable regulations, guidances, and local policies and was approved by the institutional review board of record for each PDTC.

Statistical Analysis of Tumor Growth Data

There is not a single consensus in literature in terms of which endpoint to use to measure tumor response in PDX models. There are a number of potential options. Rather than considering just one, our strategy was to consider a wide range of potential analytical strategies, each of which captures different aspects of the response and has its own strengths and weaknesses. Analytical strategies for evaluating tumor growth data include Percent Change in tumor volume (ΔV_t , normalized relative to starting volume before treatment), Area under the tumor growth curve up to time t ($aAUC_t$), Adjusted area under the curve ($aAUC_{max}$), RECIST criteria ($RECIST_{t,c}$). Metrics computed to evaluate antitumor activity of the treatment group compared to the control group include Tumor Growth Inhibition (TGI_t) and Progression-free Survival (PFS_δ) (See Supplementary Materials 1 and 2, Supplementary Table 1 and Supplementary Figures 1–10 for details, including percentages and parameters used to classify tumor response). Here, we compare and contrast these metrics in this pilot study and assess the robustness of sensitivity assessments across different analytical strategies, with the goal of making recommendations for the broader community. Towards this goal, we built an R analysis pipeline that computes all of the following measures as well as generates a set of useful graphical summaries.

One-way ANOVA or two-sample t-tests were performed to test the difference of tumor Volume changes (ΔV_t) at day $t=21$ between treatment and control groups as appropriate, and similar analyses were done for the AUC measures. Fisher's exact test was performed to test the association between treatment and drug response (non-PD vs. PD). The log-rank test was used to compare PFS distributions between treatment and control groups. All of the analysis was implemented using R.

We have developed an R markdown script that can be used to automatically run these analyses and produce summary plots given the input data is formatted as described in Supplementary Materials 1. Email cgc@sbgenomics.com to request the R script that we freely share with this publication for other researchers to use to analyze their PDX data.

Computational Workflows

All analyses were performed on the Cancer Genomics Cloud (CGC, <https://cgc.sbgenomics.com/>) (15) with workflows and tools implemented using Common Workflow Language. Human and mouse data were aligned to GRCh38 and mm10 assemblies, respectively. All workflows are available in the Temozolomide Pilot Workflows Project on the CGC). CGC users can request access to the workflows by emailing cgc@sbgenomics.com.

Human-mouse read deconvolution

We compared several tools for mouse-human read deconvolution. These were Xenome (v1.0.0) (16), BBSplit (v37.93) (<https://sourceforge.net/projects/bbmap/>), Disambiguate (v1.0; commit c52402a) (<https://sourceforge.net/projects/bbmap/>), ICRG (17), and XenofilteR (v1.5) (18). For the WES data benchmark and the RNA-seq benchmark, we respectively used experimental WES series and RNA-seq data to simulate human-mouse mixture for evaluation. For tools requiring aligned data inputs (BAM Files), BWA-Mem was used for alignment. Only reads unambiguously classified as human by a tool were labeled “human.” All other reads were considered “not human” for the true/false positive/negative calling. See Supplementary Materials 3 for additional details.

Tumor-normal WES variant calling

Five tumor-normal WES data analysis workflows from PDXNet research groups were tested on the benchmark sets, as detailed in (Supplementary Table 2 and 3, Supplementary Figures 11–13), with the goal of evaluating the accuracy in the presence of variable mouse contamination, coverage, and VAF. Starting from FASTQ data the workflows performed mouse-human disambiguation, alignment, and variant calling with one or more somatic variant callers (Mutect (19,20), VarScan (21), Strelka (9), Manta (22) and Pindel (23)). Two simulated whole exome-seq datasets were used in the benchmark for the tumor-normal variant calling workflow. The first dataset (DN) was prepared by researchers from HCI-BCM and consisted of data based on two normal samples, variants from ClinVar spiked in and with 10 and 50 % mouse contamination. The second dataset (BS) was NA12878 WES data contaminated with 10% mouse reads which was spiked with BamSurgeon [i] at 0.05, 0.1, 0.2, and 0.3 VAF using both the ClinVar variant set used for DN, variants from TCGA BRCA SNPs combined, and with indels from the ClinVar set (BS-BRCA). For all the submitted workflows, default parameters were used as specified by the workflow authors. See Supplementary Materials 4–6 for additional details. All workflows are accessible through the CGC upon request.

Tumor-only WES variant and CNV calling

Because a substantial number of PDXs among the broader research community lack matched normal DNA, we also developed a workflow for tumor-only mutation calling (Supplementary Figure 14). Preprocessing steps include quality control filtering, removing adaptors, mouse reads were removed with xenome, trimmed reads were aligned to human genome (build GRCh38.p5), duplicate reads were removed with PicardTools, and BaseRecalibrator from the Genome Analysis Tool Kit (GATK) v4.0.5.1 (24,25) was used to adjust the quality of raw reads. Variants were called in Mutect2 using the Exome Aggregation Consortium (26) database lifted over to GRCh38 as a germline reference with the allele frequency of samples not in reference set to 0.0000082364. Variant calls were then filtered using GATK FilterMutectCalls v 4.0.5.1. See Supplementary Materials 6 for additional details. Workflow is available from the CGC upon request.

To call copy number, we built a pooled normal reference using CNVkit v0.9.3 (27) from the three samples that used the same exome-seq capture kit and with sex matching. Afterward we used CNVkit to call the CNV segments from each sample using the pooled normal reference. MDACC samples exhibited low mean target coverage so we turned on the --drop-low-coverage option in CNVkit to reduce the noise in the CNV profile.

RNA-seq expression calling

Because the disambiguation of mouse and human reads was sharp for both DNA and RNA data, we did not expect expression calling workflows to have issues specific to PDXs. Therefore, we dockerized only one PDX RNA-seq expression workflow (Supplementary Materials 7, Supplementary Figure 14) that was submitted by The Jackson Laboratory (JAX). The transcriptomes of hg38 and NOD (based on the mm10 mouse genome) were used to construct the xenome (version 1.0.0) (16) indices (k=25), and then reads were classified as human, mouse, both, neither or ambiguous at default xenome parameters. Reference indices for the alignment were built by rsem-prepare-reference using ENSEMBL annotation (version GRCh38.91) for STAR aligner (version 2.5.1b) (28). Human-specific reads were mapped to reference indices using STAR, and expression estimates were computed using rsem-calculate-expression v1.2.31 (29) at default parameters. Picard CollectRnaSeqMetrics: (broadinstitute.github.io/picard/picard-metric-definitions.html) was used to calculate the post-alignment mapping statistics. An implementation of this workflow has been deployed on the CGC.

Comparisons of xenograft sequence data across PDTCs

Each PDTc submitted WES and RNA-seq data from untreated xenografts that had been successfully grown in mice at the respective sites (Supplementary Tables 4–7, Supplementary Figures 15–22). These data are available through the Sequencing Read Archive at accession number PRJNA608267. Groups were asked to submit xenograft sequence data according to their standard practices, without pre-specification of the sample passage number or the sequencing protocol. In the intersection analysis, only variants with allele frequency > 0.2 were retained. We note that MDACC had fewer calls that passed the allele frequency filter in comparison to other centers. This is because MDACC provided samples had mean target coverage ~30X whereas samples from other centers were

sequenced to a depth of ~150X (Supplementary Table 7). We also analyze mutational differences in cancer-related genes, using the CancerMine database: <http://bionlp.bcgsc.ca/cancermine/>. We listed the top 15 genes, by citation count, associated with each of the terms cancer driver, oncogene and tumor suppressor from the database and then combined these to get 33 unique cancer genes (Supplementary Figures 17–19).

For the copy number comparisons, the copy number alteration (CNA) segments obtained from CNVkit using a pooled normal were median-centered and visualized in IGV v2.4.13 (30). To determine the overall concordance of the CNA between each pair of samples, we first intersected the CNA segments for each pair of samples and then binned them into 100kb-windows using Bedtools v2.26.0 (31).

RNA-seq data provided by each center were generated using different kits and protocols, and the data from HCI-BCM was sequenced in single end mode (Supplementary Table 6). Sequence data were analyzed with the ‘PDXnet RNA Expression Estimation’ and the ‘PDXnet RNA Expression Estimation – SE’ workflows on the CGC. RNA expression estimates were downloaded from CGC for additional analyses. The single-end data provided by HCI-BCM yielded estimates of RNA expression that were twice as high when compared to the paired-ended sample provided by other centers due to differential handling of paired-end and single-end data by RSEM (29) tool. To eliminate the biases in the count estimation across centers, HCI-BCM, estimated transcript counts were divided in half. From the mapping stats and from automatic library type detection algorithm in the tool Salmon, we noted that RNA-Seq library generated at MDACC are non-directional though the sequencing protocol used is for directional library thus we decided to consider MDACC library as non-directional during the analysis.

Results

Study design and treatment results

A critical, yet unresolved, question that motivated the inception of the PDXNet was what the inter-laboratory reproducibility of PDX drug studies would be across centers with independently established practices for preclinical testing, i.e. how much standardization would be needed to run large-scale, multicenter preclinical studies. To address this question, the NCI Patient Derived Models Repository (PDMR) reviewed preclinical studies performed by the Biological Testing Branch (BTB/DCTD/NCI), which has performed numerous in vivo studies with PDX models. The PDMR selected three PDX models with non-published known responses to temozolomide for an inter-laboratory reproducibility pilot. The three PDX models selected were 625472–104-R (colon adenocarcinoma, non-responsive model), 172845–121-T (colon adenocarcinoma, intermediate response), and BL0293-F563 (urothelial/bladder cancer, complete response). Patient data, clinical history, and representative histology and sequence data can be found at <https://pdmr.cancer.gov>.

For the study set-up (Figure 1a), the four PDTs – Huntsman Cancer Institute/Baylor College of Medicine (HCI-BCM), MD Anderson Cancer Center (MDACC), Washington University-St. Louis (WUSTL), and The Wistar Institute/University of Pennsylvania/MDACC (WIST) – were directed to use their standard preclinical study set-up (Figure 1b)

and monitoring SOPs (Figure 1c) and to use literature searches to determine temozolomide dosing and schedule. Each group also performed exome and RNA-seq of untreated tumors that had been successfully engrafted (Figure 1d). All PDTCs were kept blinded to which models were temozolomide sensitive or resistant and to all other groups' preclinical study set-ups. In addition, none of the PDTCs had previous experience with temozolomide; so the reference doses/schedules would need to be determined independently at each center. The exceptions to blinding were that all PDTCs were required to use NSG host mice and implant PDX material subcutaneously. In addition, the PDTCs used drug prepared by the Developmental Therapeutics Clinic (DTP/NCI) to ensure that there were no variations in manufacture.

The laboratory SOPs for the preclinical study set-ups were collated by the PDMR (Table 1). While all centers staged tumors to between 100–200 mm³, implantation methodologies varied. Three groups directly implanted ~1 mm³ PDX fragments into each host mouse, one group minced a ~1 mm³ PDX fragment into a slurry for implantation, and one dissociated PDX material and implanted 3–5 × 10⁶ cells per host (For each model all hosts had the same number of cells injected in all control and treated animals. Variation was only across models). Comparison of vehicle control growth curves for all groups demonstrated overall similar growth kinetics of the models at each site irrespective of the implantation methodology used (Supplementary Figure 1).

Each PDTC independently researched published literature to select a temozolomide dosing and schedule for its site, with key references noted: HCI-BCM (32–34), MDACC (35), WUSTL (34,36–39), and WIST (40,41). While diverse literature was considered, all sites selected a 50 mg/kg dose and one of two different dosing schedules. These schedules were either daily temozolomide treatment for 5 days followed by 23 days of rest (28-day cycle) or 5 days of treatment followed by 2 days of rest (7-day cycle); 1–4 cycles were used (Table 1).

Overall, all sites reported similar responses irrespective of the methodology, dosing, or schedule used (Figure 2a–o), with especially strong concordance in the non-responsive and complete response model results, as detailed quantitatively below. If the drug x model combination had been performed as part of an exploratory study, these independent experiments would likely yield similar decisions about treatment efficacy. The intermediate response models showed more variation in growth across centers. The intermediate cases were also more clearly affected by the variability in SOP end-point times, one of the biggest variations among methodologies (Table 1). For example, some groups sacrificed all mice once the vehicle control group reached a threshold volume, while other groups ended after a defined length of time after the last dosing. This resulted in some studies observing strong tumor inhibition through the end of study, while others observed regrowth after initial inhibition (Supplementary Figure 2). Nevertheless, the similarities in response indicated that the existing range of methodologies is sufficient and robust enough to capture the critical cases of strong response and non-response. After discussion of these results, the PDXNet Consortium has agreed on a standard of continued monitoring of all cohorts where response is observed for at least 1.5–2 cycle lengths beyond the last dosing cycle, provided animal health end-points are not reached. Detailed quantitative comparisons and statistical analysis are addressed in the next section.

Statistical Robustness of PDX Treatment Response

Statistical approaches for evaluating cohort drug response—A challenge of evaluation of PDX response is that there is still no standard statistical approach for analysis of tumor response for PDX growth data. Common measures of tumor size include percent change in volume from baseline to a fixed time end-point; area under the tumor growth curve; tumor growth inhibition, defined as the ratio of the average tumor size at a given time point relative to control; and time to progression, a potentially censored end-point measuring time from baseline until growth to a certain multiple of baseline. Classification of growing PDX tumors into RECIST-like categories (42) (Complete Response-CR, Partial Response-PR, Stable Disease-SD, and Progressive Disease-PD) is another assessment that has the advantage of congruence with clinical trials, but it can be strongly dependent on category thresholds that do not analogize straightforwardly with patient primary tumors. Each of these measures has their own strengths and limitations. For example, the percent change from baseline is intuitive, interpretable, and unlike RECIST does not require specification of a cut point. In contrast to the area under the curve (AUC) approaches it does not use all of the tumor time course information but only the first and last points. Here we consider all of these measures and assess concordance of results across analytical strategies as well as across growth data from each center.

PDX tumor volume analysis software—We have devised an automated analysis script in R that, given data in a prespecified format and a time point of interest, will automatically plot the tumor growth curves and group mean curves, compute all of these statistical measures and their associated plots, and produce an annotated .html report in R markdown that serves as a complete summary of the results (see Methods). In the supplementary materials (Supplementary Materials 1 and Supplementary Table 1), we describe a standard format for the recorded data that is compatible with our analysis scripts and we also provide instructions for researchers to use this script to analyze their own data. We believe that this automated script can enhance reproducibility and transparency of analyses and can be revised and adapted as a standard analysis script for general use.

Comparisons across statistical methods

We statistically assessed drug response for the measures mentioned above across all research groups. Table 2 contains the p-values for assessing treatment vs. control differences for each of the statistical tests (see Methods). Figure 3 shows associated plots from the HCI-BCM studies for each of the three models (Figure 3, **columns**) for several data representations and statistical evaluation approaches (Figure 3, **rows**). Associated plots for drug response at other sites i.e. MDACC, WUSTL, PDMR and WIST are shown in Supplementary Figures 3, 4, 5, and 6, respectively. Overall, we found that assessments of drug response were robust across research groups, with particularly decisive evaluations for the non-responsive and responsive models. The various analytical methods (Supplementary Figures 7, 8, 9 and 10) also gave results consistent with one another, with a few exceptions noted below. However, the intermediate group was difficult to classify. For the intermediate group most of the statistical measures showed clear difference from control, but the results were inconsistent for RECIST criteria.

RECIST yielded qualitatively similar ordering of the models as the other methods, but it had the lowest power and showed considerable variability across cut points, complicating its use. The percent change in tumor size and area under the curve measures largely agreed and showed good statistical power. The tumor growth inhibition measure also yielded consistent results. The natural statistical test is whether this ratio is less than 1, but this should be accompanied by an assessment of the clinical significance of the effect size, since it is possible to have a small p-value with minimal inhibition in a preclinical study, e.g. 10% or 20%, that might not ultimately correspond to a clinical response. We recommend statistical testing vs. control while accompanied by an assessment of clinical significance that may depend on the context.

Cloud Workflows for PDX Sequence Analysis

Robust sequence analysis pipelines are essential for understanding cancer genetics from PDX models. While prior PDX pipelines have been published, e.g. (13,43), it can be time-consuming for researchers to implement and evaluate other groups' methods. To address this problem, five PDXNet teams provided sequence analysis workflows for PDX exome-seq mutation calling, and the PDXNet Data Commons and Coordinating Center (PDCCC) dockerized these for co-localized application and sharing with the research community via the National Cancer Institute Cancer Genomics Cloud (CGC). The Seven Bridges Genomics team in the PDCCC also independently evaluated each of these pipelines. Each submitting group also specified parameters as part of the workflow submission. Evaluations were performed on simulated benchmark mixtures of human and mouse reads with various mouse/human read ratios and variant allele frequencies (see Methods).

Benchmarking of human-mouse read disambiguation

We first compared the efficacy of the five pipelines (Supplementary Table 2) for human-mouse read disambiguation using a series of simulated benchmark WES and RNA-Seq datasets. The simulated WES and RNA-Seq datasets were used to test the five commonly used human-mouse read deconvolution tools: BBSplit, Xenome, Disambiguate, Xenofilter, and ICRG. All tools achieved >99 % precision for both WES and RNA-Seq benchmarks (Figure 4a). Xenofilter showed the lowest recall (96.60 % and 89.63 % recall in WES and RNA-seq benchmarks, respectively), whereas BBSplit showed the best overall performance i.e. highest precision without any loss in recall (99.87 % and 99.64 % precision in WES and RNA-seq benchmarks, respectively).

Benchmarking of WES analysis pipelines

We next compared WES results generated by the five pipelines including variant calling and the effectiveness of mouse-human disambiguation. For this analysis, two simulated benchmark datasets were created, with two levels of mouse contamination (10% and 50%) and a range of variant allele frequencies (VAFs) - 0.025, 0.05, 0.1, 0.2, and 0.3, with spike-ins of point mutations and indels (See Methods). For performance metrics, we used precision/recall (across SNPs, INS, DELs) and pseudo-ROC curves (see Methods). We observed minimal impact of different percentages of mouse contamination on the performance of the five workflows (Supplementary Table 3). The overall best performing workflow, Workflow 2, is shown in Figure 4b and performance results across workflows are

shown in Figure 4c. When analyzing variant caller performances, we observed that MuTect2 (used in Workflows 2 and 4) performed consistently well across all samples for all the tested VAF levels. Supplementary Figure 11 shows SNP performance across 0.05 and 0.3 VAFs for BS-DN dataset across different coverage values (although we only show 2 VAF levels, the caller performed well across all VAF levels tested i.e. 0.05 – 0.3); however, indel recall decreased at lower VAFs. VarScan2 (used in Workflows 3 and 4) called only a small number of variants at lower VAFs as evident from the very low recall values. We also observed marked differences in performance of two VarScan2 PDTC workflows, e.g. the DN dataset when processed through workflow 3 at low VAFs i.e. at 0.025 and 0.05 VAF had SNP precision values of 0% and 1.71%, respectively, and when processed through workflow 4 had SNP precision values of 2.16% and 12.4%. The difference in performance between workflows 3 and 4 is possibly due to the fact that in workflow 3 VarScan2 was run independently, whereas in workflow 4 the final calls are a union of VarScan2 and Mutect2 calls. Recall was good at higher VAFs, but precision varied. For example, the DN dataset when processed through workflow 3 at 0.2 and 0.3 VAF had SNP precision values of 98.43% and 99.13%, respectively, and when processed through workflow 4 had SNP precision values of 33.04% and 45.03%. Strelka2 (part of workflows 1 and 5) was the most aggressive caller, achieving considerable recall even at the lowest VAFs tested. However, Strelka2 performance varied between the two workflows that used it, i.e. workflow 1 and workflow 5, possibly because workflow 1 used the recommended settings for running Strelka (combining it with Manta), whereas workflow 5 ran Strelka independently. We observed similar trends in the pseudo-ROC curves consistent with results described above.

PDXNet Exome, RNA-seq, and CNV workflows

According to the achieved precision and recall values across SNPs, INS, and DELs (F1 statistic), Workflow 2 was the best performing WES workflow for PDX data. Consequently, we recommend using Workflow 2 for somatic calling in PDX tumor-normal paired WES samples. As the other workflows (Supplementary Figures 12 and 13) may be suited for other datasets we are releasing all workflows on the CGC. In addition, we are releasing a tumor-only exome-seq variant calling pipeline, an RNA-seq expression pipeline, and a CNV calling from exome-seq pipeline (See Methods and Supplementary Figure 14). The tumor-only exome-seq, RNA-seq, and CNV calling pipelines were used to analyze samples from each PDTC in the temozolomide experiments.

Robustness of PDX Sequence Evaluations

To test the robustness of these sequence analysis workflows, we applied them to PDX samples from the temozolomide study. Each PDTC generated an independent biological sample of an untreated PDX for each of the three patient models. They then sequenced these independently and submitted the sequence data to the coordinating center.

Variant Calls from Exome-Seq

FASTQ files from whole exome sequencing were obtained from the four PDTCs (MDACC, HCI-BCM, WUSTL, and WIST). Each center provided WES and RNA sequencing data from the PDX models: 625472–104-R, 172845–121-T, and BL0293-F563. No matched normal data were available for these models.

The WES data were analyzed with the optimal WES pipeline that was modified to take into account the lack of normal DNA, i.e. the ‘PDX WES Tumor-Only: Mutect2’ workflow. The exome capture kits used by each center covered different regions and total amounts of the genome (Supplementary Table 4), resulting in disparate variant calls among centers. The length of the genome covered by the intersection of the capture loci across all groups was 33.71Mb. Filtering out variants from non-intersecting regions or with low allele frequencies (AF<5%) made the average number of variant calls across centers for each model comparable (Figure 4d, Supplementary Table 5, Supplementary Figure 15), though centers with lower sequencing depth had fewer calls meeting the QC threshold. A distribution of allele frequencies for calls meeting the QC threshold for each sample across each center is shown in Supplementary Figure 16. Mutations in cancer genes showed similarities across centers (Supplementary Figures 17–19), though there were variations related to sequencing depth and allele frequency, e.g. the lower depth of the MD Anderson samples resulted in fewer variant calls. When found, mutations appeared at similar AFs across centers, and shared mutations tended to have higher AFs. These results indicate that, although our chosen pipeline is an improvement over prior ones, increased sequencing depth would still be valuable.

Copy number calls from Exome-Seq

We called the copy number for each sample using CNVkit with a pooled normal approach (27) (Supplementary Figure 20). Overall, we observed similar profiles among samples from the same model. The most apparent difference between samples was an overall shift relative to the baseline. As such, comparing absolute copy number gain and loss calls between samples remains challenging. Supplementary Figure 21 shows the Pearson correlation coefficients between samples. We observed higher Pearson coefficients (>0.746) for pairwise comparisons for samples of the same tumor among the HCI-BCM, WUSTL, and WIST PDTCs, compared to samples of different tumors. On the other hand, the MDACC profiles were noisier due to lower coverage, despite using the “drop low coverage” option in CNVkit, and we were unable to identify strong correlations between samples of the same tumor for MDACC.

Expression calls from RNA-Seq

Data provided by each PDTC were generated using different RNA-seq protocols (Supplementary Table 6) and were analyzed with the rsem-1–2–31-workflow-with-star-aligner (single-end data) and rsem-1–2–31-workflow-with-star-aligner-pe (paired-end data) workflows on the CGC, with small adjustments based on single vs. paired end sequencing or directionality parameters (see Methods). To account for differences in library size, data were normalized by Trimmed Mean of M-values (TMM), and further converted to count per million (CPM) with the R package edgeR (29,44). Following normalization and CPM conversion, significant batch effects were still present in these data (Supplementary Figure 22). To correct for batch effect among centers, median polish by center was applied to TMM normalized CPM data as implemented in the MBatch R package (github.com/MD-Anderson-Bioinformatics/MBatch). Following batch correction, samples tended to cluster by model rather than sample, though with some exceptions (Figure 4e).

Discussion

Our work demonstrates the robustness of PDXs as a model system for studying cancer drug response. In particular, we have demonstrated the experimental robustness of PDX response for three different models even among research groups blinded to the expected response and who followed independently developed preclinical protocols. As has been published numerous times, reproducibility of experimental results is a confounding factor in the ability to build on previously published data (45–47). These results demonstrate that in the context of a cytotoxic agent, even when groups are not told what experimental protocol to use, PDXs can yield accurate and consistent treatment responses. Even given these results we feel that it is important to standardize preclinical methodologies and analyses tools so that data can be compared across the PDTs over time. For example, one change that will be implemented at all sites will be to monitor tumor volume changes for at least 1.5–2 cycle lengths beyond the last dosing cycle to assess durability of response. It is also important to recognize that different classes of drugs, more heterogeneous tumors, as well as some histologies may have wider variation in reproducibility or response; standardization of methodologies will help minimize the experimental variables that may affect interpretation of the data.

While prior studies have also investigated the robustness of PDX drug response, they have not included comparisons across research groups. For example, Izumchenko et al (6) demonstrated similar responses between 92 patients and matched xenografts, and Gao et al (5) and Townsend et al (48) showed that 1×1×1 (animal, model, treatment) xenograft experiments were predictive of response in larger cohorts, including for resistance mechanisms to MAPK inhibition in melanoma (5) and to MDM2 inhibition in hematologic malignancies (48). However, such results may depend on the chosen treatment protocols. Our findings further show that PDX treatment results can be robust enough to withstand protocol variations and blinding. Moreover, this work extends prior investigations to standardize statistical analysis of PDX growth data (49) by showing that statistical analyses can tolerate a wide range of variations in experimental protocols and statistical parameters.

In addition, we have developed standardized PDX sequence analysis pipelines for tumor-normal variant calling, tumor-only variant calling, and RNA-seq expression calling. We have provided these as public tools on the CGC, making them easily accessible for other researchers and applicable to the broad data collections shared on the CGC. Not only have these pipelines been tested on extensive benchmark datasets, but we have also applied the tumor-only variant calling and RNA-seq pipelines to sequence data generated across the PDTs in the temozolomide study. These give similar results across the groups, demonstrating both the efficacy of the pipelines and the minor sequence evolution from PDX to PDX during the process of generating test cohorts across groups.

Importantly, we have also developed biostatistical analysis workflows for tumor volume data, which we are releasing here as well. Our results show a high level of concordance among the various biostatistical analysis strategies, but with some caveats. The RECIST criteria is heavily threshold dependent, has lower statistical power, and less consistent with results from the other strategies. Since each strategy has its own strengths and weaknesses, we recommend testing multiple strategies for PDX analyses. It is also important to consider

clinical as well as statistical significance, considering effect sizes to be sure any effect is of sufficient magnitude to be meaningful, a determination that may depend on the clinical context. Classifying PDX volume data into meaningful patient-analogous categories of complete response, stable disease and partial response remains challenging, though this may become possible as datasets with paired clinical and PDX response data increase. In the meantime, our automated analysis scripts, which collate the results and analytical steps into an automated report, provide a standard tool for the PDX field, and future PDXNet volume data will be released in a data format consistent with these scripts. We encourage others to follow the volume data standards we have developed here, which will assist in the quantitative application of PDX treatment data for predicting the efficacy of drugs in patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Support for this work included funding provided by the NIH to the PDXNet Data Commons and Coordination Center (NCI U24-CA224067), to the PDX Development and Trial Centers (NCI U54-CA224083, NCI U54-CA224070, NCI U54-CA224065, NCI U54-CA224076, NCI U54-CA233223, and NCI U54-CA233306) and to the National Cancer Institute Cancer Genomics Cloud (HHSN261201400008C and HHSN261201500003I).

Supporting grants or fellowships:

Support for this work included funding provided by the NIH to the PDXNet Data Commons and Coordination Center (NCI U24-CA224067), to the PDX Development and Trial Centers (NCI U54-CA224083, NCI U54-CA224070, NCI U54-CA224065, NCI U54-CA224076, NCI U54-CA233223, NCI U54-CA233306) and to support for the Cancer Genomics Cloud (HHSN261201400008C and HHSN261201500003I).

Conflict of Interest

F. Meric-Bernstam reports receiving commercial research grants from Novartis, AstraZeneca, Calithera, Aileron, Bayer, Jounce, CytoMx, eFFECTOR, Zymeworks, PUMA Biotechnology, Curis, Millennium, Daiichi Sankyo, Abbvie, Guardant Health, Takeda, Seattle Genetics, and GlaxoSmithKline as well as grants and travel related fees from Taiho, Genentech, Debiopharm Group, and Pfizer. She also served as a consultant to Pieris, Dialectica, Sumitomo Dainippon, Samsung Bioepis, Aduro, Origimed, Xencor, Jackson Laboratory, Zymeworks, Kolon Life Science, and Parexel International, and advisor to Inflection Biosciences, GRAIL, Darwin Health, Spectrum, Mersana, and Seattle Genetics. A.L.W and B.E.W receive a portion of royalties if University of Utah licenses certain PDX models to for-profit entities. M.T.L is a founder of, and equity stake holder in, Tvardi Therapeutics Inc., a founder of, and limited partner in, StemMed Ltd., and a Manager in StemMed Holdings LLC. He also receives a portion of royalties if Baylor College of Medicine licenses certain PDX models to for-profit entities.

References

1. Tentler JJ, Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, et al. Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol*. 2012;96:338–50.
2. Cho S, Kang W, Han JY, Min S, Kang J, Lee A, et al. An Integrative Approach to Precision Cancer Medicine Using Patient-Derived Xenografts. *Mol Cells*. 2016;39:77–86. [PubMed: 26831452]
3. Byrne AT, Alf erez DG, Amant F, Annibali D, Arribas J, Biankin AV, et al. Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nat Rev Cancer* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2017;17:254. [PubMed: 28104906]
4. Krepler C, Sproesser K, Brafford P, Beqiri M, Garman B, Xiao M, et al. A comprehensive patient-derived xenograft collection representing the heterogeneity of melanoma. 2017;21:1953–67.

5. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;21:1318. [PubMed: 26479923]
6. Izumchenko E, Paz K, Ciznadija D, Sloma I, Katz A, Vasquez-Dunddel D, et al. Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors. *Ann Oncol*. 2017;28:2595–605. [PubMed: 28945830]
7. Dong G, Mao Q, Yu D, Zhang Y, Qiu M, Dong G, et al. Integrative analysis of copy number and transcriptional expression profiles in esophageal cancer to identify a novel driver gene for therapy. *Sci Rep* [Internet]. Nature Publishing Group; 2017;7.
8. Garralda E, Paz K, López-Casas PP, Jones S, Katz A, Kann LM, et al. Integrated next-generation sequencing and avatar mouse models for personalized cancer treatment. *Clin Cancer Res*. 2014;20:2476–84. [PubMed: 24634382]
9. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* [Internet]. 2018;15:591–4. [PubMed: 30013048]
10. Ben-David U, Ha G, Tseng Y-Y, Greenwald NF, Oh C, Shih J, et al. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat Genet* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2017;49:1567. [PubMed: 28991255]
11. Doroshow J Abstract IA12: NCI's patient-derived cancer models repository. *Clin Cancer Res*. 2016;22:IA12–IA12.
12. Krupke DM, Begley DA, Sundberg JP, Richardson JE, Neuhauser SB, Bult CJ. The mouse tumor biology database: A comprehensive resource for mouse models of human cancer. *Cancer Res*. 2017;77:e67–70. [PubMed: 29092943]
13. Bruna A, Rueda OM, Greenwood W, Batra AS, Callari M, Batra RN, et al. A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell* [Internet]. The Authors; 2016;167:260–274.e22. [PubMed: 27641504]
14. Teicher B, Plowman J, Dykes D, Hollingshead M, Simpson-Herren L, Alley M. *Human Tumor Xenograft Models in NCI Drug Development*. Totowa, NJ: Humana Press Inc; 1997.
15. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The cancer genomics cloud: Collaborative, reproducible, and democratized - A new paradigm in large-scale computational research. *Cancer Res*. 2017;77:e3–6. [PubMed: 29092927]
16. Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, et al. Xenome-a tool for classifying reads from xenograft samples. *Bioinformatics*. 2012;28:172–8.
17. Callari M, Batra AS, Batra RN, Sammut SJ, Greenwood W, Clifford H, et al. Computational approach to discriminate human and mouse sequences in patient-derived tumour xenografts. *BMC Genomics*. *BMC Genomics*; 2018;19:19. [PubMed: 29304755]
18. Kluijn RJC, Kemper K, Kuilman T, de Ruiter JR, Iyer V, Forment J V., et al. XenofilteR: Computational deconvolution of mouse and human reads in tumor xenograft sequence data. *BMC Bioinformatics*. *BMC Bioinformatics*; 2018;19:1–15. [PubMed: 29291722]
19. Van der Auwera GA. Somatic variation discovery with GATK4. *Am Assoc Cancer Res*. 2017.
20. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* [Internet]. Nature Publishing Group; 2013;31:213–9. [PubMed: 23396013]
21. Wilson RK, Mardis ER, McLellan MD, Koboldt DC, Shen D, Zhang Q, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22:568–76. [PubMed: 22300766]
22. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32:1220–2. [PubMed: 26647377]
23. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25:2865–71. [PubMed: 19561018]

24. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;43:491. [PubMed: 21478889]
25. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303. [PubMed: 20644199]
26. M L, KJ K, EV M, KE S, E B, T F, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2017;536:285–91.
27. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016;12:1–18.
28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21. [PubMed: 23104886]
29. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinforma Impact Accurate Quantif Proteomic Genet Anal Res*. 2014;41–74.
30. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14:178–92. [PubMed: 22517427]
31. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2. [PubMed: 20110278]
32. U.S. FDA. Temozolomide (Temodar) [New Drug Approval Package #21–029]. US Food Drug Adm.
33. Hirst TC, Vesterinen HM, Sena ES, Egan KJ, MacLeod MR, Whittle IR. Systematic review and meta-analysis of temozolomide in animal models of glioma: Was clinical efficacy predicted. *Br J Cancer* [Internet]. Nature Publishing Group; 2013;108:64–71. [PubMed: 23321511]
34. Keir ST, Maris JM, Reynolds CP, Kang MH, Kolb EA, Gorlick R, et al. Initial Testing (Stage 1) of Temozolomide by the Pediatric Preclinical Testing Program. *Pediatr Blood Cancer*. 2013;60:783–90. [PubMed: 23335050]
35. Middlemas DS, Stewart CF, Kirstein MN, Poquette C, Friedman HS, Houghton PJ, et al. Biochemical correlates of temozolomide sensitivity in pediatric solid tumor xenograft models. *Clin Cancer Res*. 2000;6:998–1007. [PubMed: 10741727]
36. Kitange GJ, Carlson BL, Schroeder MA, Grogan PT, Lamont JD, Decker PA, et al. Induction of MGMT expression is associated with temozolomide resistance in glioblastoma xenografts. *Neuro Oncol*. 2009;11:281–91. [PubMed: 18952979]
37. Stacchiotti S, Tortoreto M, Bozzi F, Tamborini E, Morosi C, Messina A, et al. Dacarbazine in solitary fibrous tumor: A case series analysis and preclinical evidence vis-à-vis temozolomide and antiangiogenics. *Clin Cancer Res*. 2013;19:5192–201. [PubMed: 23888069]
38. Stevens MFG. Chapter 5 - Temozolomide: From Cytotoxic to Molecularly Targeted Agent In: Neidle S, editor. *Cancer Drug Des Discov* (Second Ed. San Diego: Academic Press; 2014 page 145–64.
39. Nair AB, Jacob S. A simple practice guide for dose conversion between animals and human. *J basic Clin Pharm* [Internet]. 2016;7:27–31. [PubMed: 27057123]
40. Plowman J, Waud W, Koutsoukos A, Rubinstein L, Moore T, Grever M. Preclinical Antitumor Activity of Temozolomide in Mice: Efficacy against Human Brain Tumor Xenografts. *Cancer Res*. 1994;4:3793–9.
41. Viel T, Schelhaas S, Wagner S, Wachsmuth L, Schwegmann K, Kuhlmann M, et al. Early Assessment of the Efficacy of Temozolomide Chemotherapy in Experimental Glioblastoma Using [18F]FLT-PET Imaging. *PLoS One*. 2013;8.
42. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* [Internet]. Elsevier Ltd; 2009;45:228–47. [PubMed: 19097774]
43. Woo XY, Srivastava A, Graber JH, Yadav V, Sarsani VK, Simons A, et al. Bioinformatics workflows for genomic analysis of tumors from Patient Derived Xenografts (PDX): challenges and guidelines. *bioRxiv* [Internet]. 2018;414946.

44. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26:139–40. [PubMed: 19910308]
45. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* [Internet]. Nature Publishing Group; 2011;10:712–3. [PubMed: 21892149]
46. Ioannidis JPA. Why Most Clinical Research Is Not Useful. *PLoS Med*. 2016;13:1–10.
47. Collins AT, Lang SH. A systematic review of the validity of patient derived xenograft (PDX) models: The implications for translational research and personalised medicine. *PeerJ*. 2018;2018:1–22.
48. Townsend EC, Murakami MA, Christodoulou A, Christie AL, Köster J, DeSouza TA, et al. The Public Repository of Xenografts (ProXe) enables discovery and randomized phase II-like trials in mice. *Cancer Cell*. 2016;29:574–86. [PubMed: 27070704]
49. Mer AS, Ba-alawi W, Smirnov P, Wang YX, Brew B, Ortmann J, et al. Integrative pharmacogenomics analysis of patient-derived xenografts. *Cancer Res* [Internet]. 2019;canres.0349.2019.

Statement of Significance

The PDXNet Consortium shows that Patient-Derived Xenografts (PDXs) drug responses and sequence results are reproducible across diverse experimental protocols, establishing the potential for multi-site preclinical studies to translate into clinical trials.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

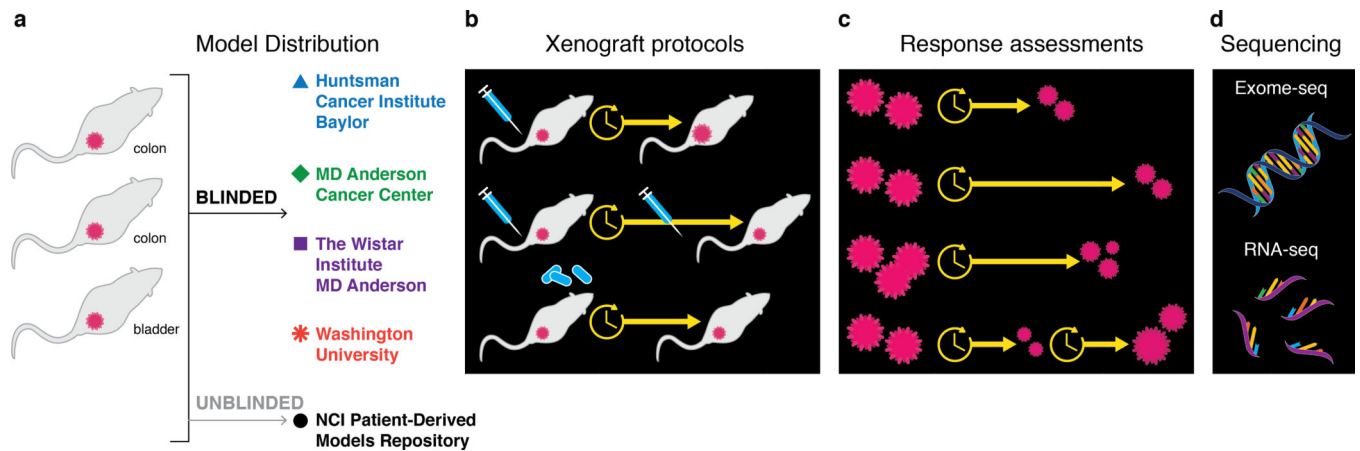
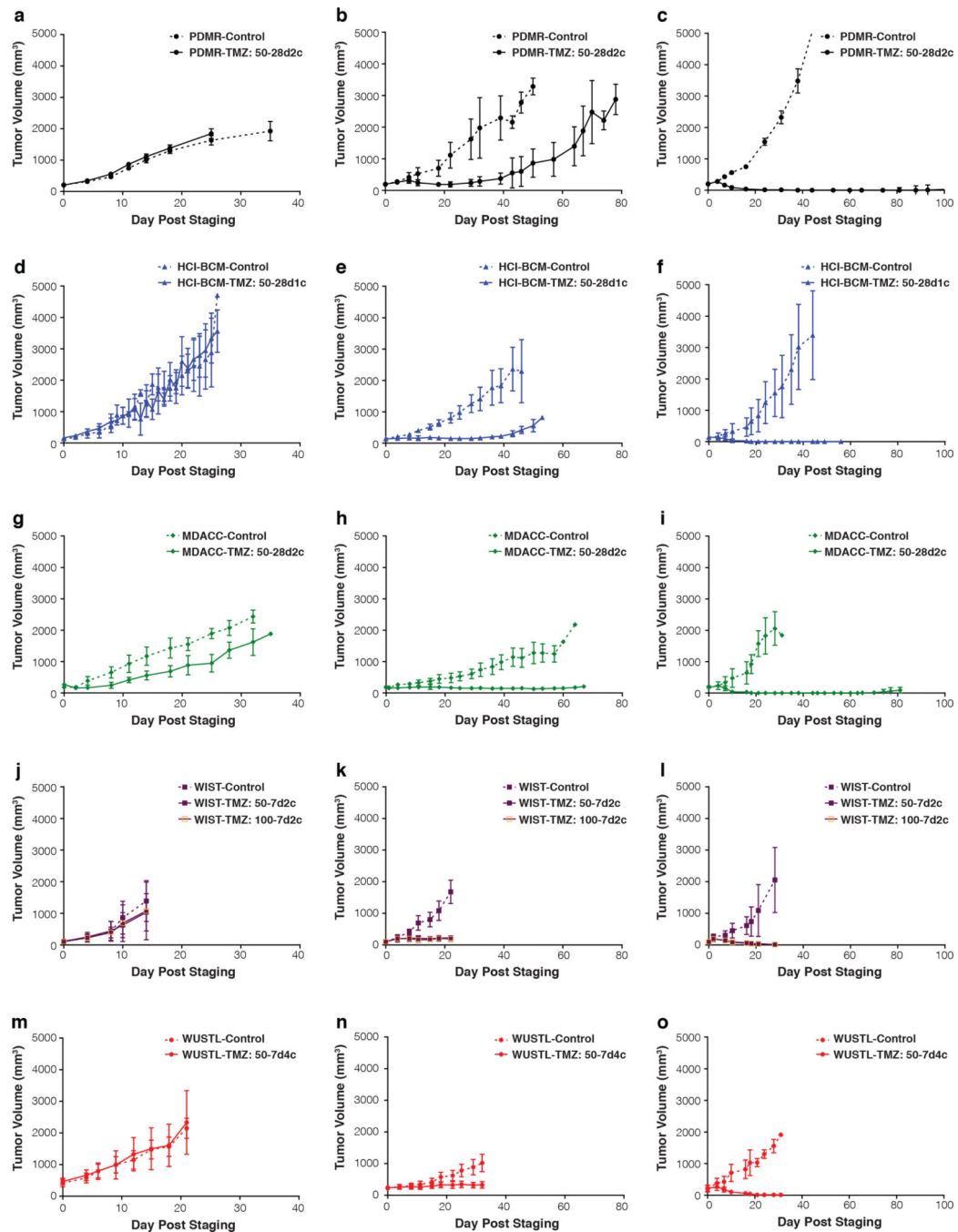


Figure 1.

a) Three models were distributed for experimentation to 4 centers: Huntsman Cancer Institute/Baylor College of Medicine (HCI-BCM), MD Anderson Cancer Center (MDACC), Washington University-St. Louis (WUSTL), and The Wistar Institute/University of Pennsylvania/MDACC (WIST). These three centers were chosen based on prior results on temozolomide treatment response obtained by the NCI Patient-Derived Models Repository (PDMR). **b)** Each of the three models were treated with temozolomide by the 4 centers under blinded protocols. **c)** Treatment responses were comparatively assessed under several biostatistical protocols. **d)** Sequence data were collected by each center and assessed.

**Figure 2.**

Comparison of PDX tumor volume control and temozolomide treatment arms at the PDMR (a-c), HCI-BCM (d-f), MDACC (g-i), WIST (j-l), and WUSTL (m-o). Model 625472-104-R (a, d, g, j, m), 172845-121-T (b, e, h, k, n), and BL0293-F563 (c, f, i, l, o). Axes are held constant for comparison between studies. Dashed lines, vehicle control groups, Solid lines, temozolomide treatment groups. Median \pm SD. For statistical assessments, see Figure 3 and Table 2. (PDMR - NCI Patient-Derived Models Repository, HCI-BCM - Huntsman Cancer Institute/Baylor College of Medicine, MDACC - MD Anderson Cancer Center, WUSTL -

Washington University-St. Louis, and WIST - The Wistar Institute/University of Pennsylvania/MDACC.)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

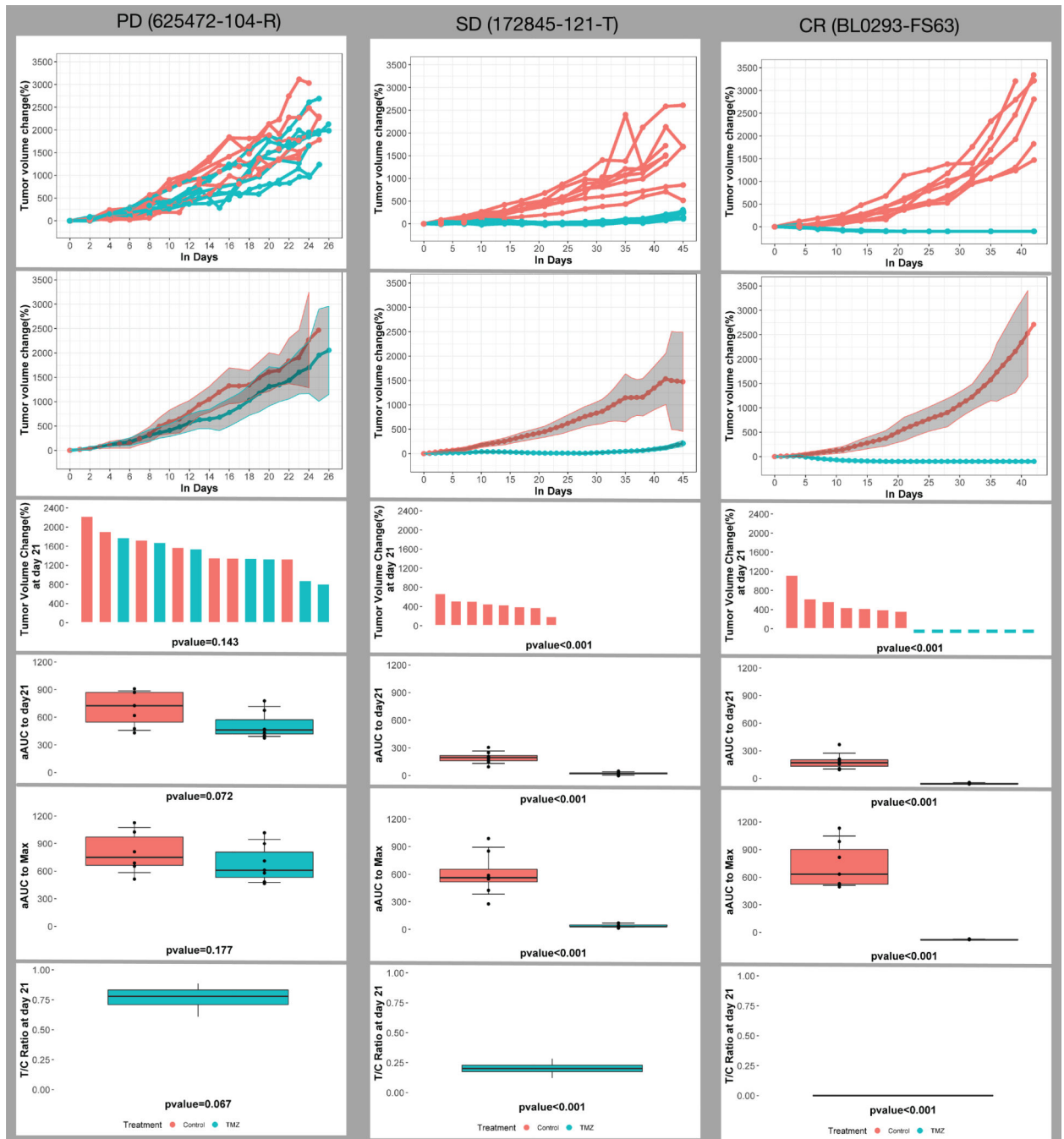


Figure 3. Analytical Summaries, HCI-BCM Study. Analytical results from HCI-BCM study for progressive model (625472–104-R), stable disease model (172845–121-T) and complete response model (BL0293-F563) (columns 1, 2, and 3 respectively), with interpolated individual curves (row 1), mean curves for treatment and control with 95% confidence bands (row 2), waterfall plots demonstrating V_{21} (row 3), boxplots of $aAUC_{21}$ (row 4) and $aAUC_{max}$ (row 5) for treatment and control, and a boxplot of TGI_{21} (row 6), along with p-values comparing treatment to control for each measure.

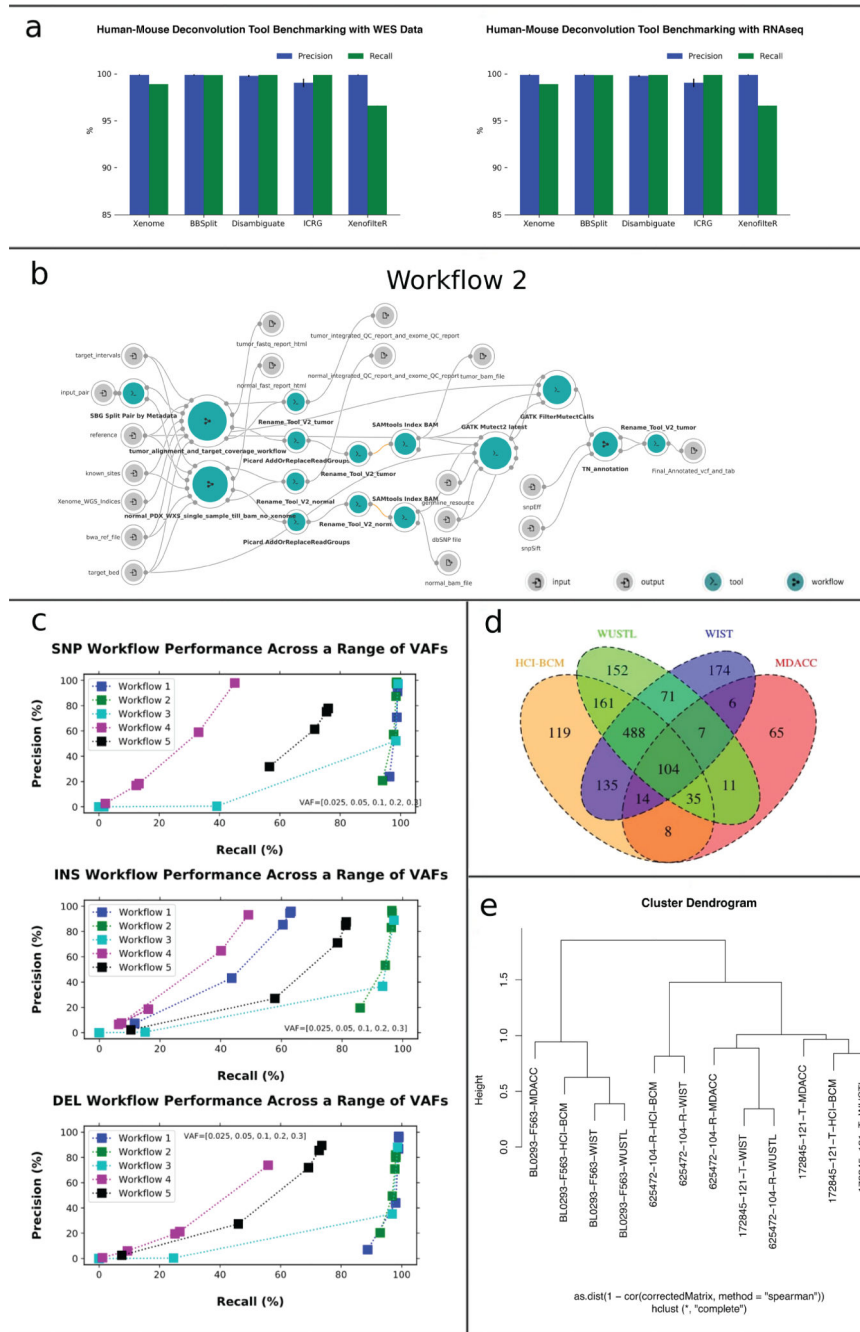


Figure 4. Workflow Benchmarking and Analysis Summary. a) Panel A shows results of the evaluation of mouse-human disambiguation tools (Xenome, BBSplit, Disambiguate, ICRG, XenofilteR). Each figure shows precision (blue) and recall (green) for a simulated data. Left figure shows results of mouse disambiguation for whole exome data. Right figure shows results of mouse disambiguation for RNA-seq data. b) The panel shows the wiring diagram for the whole exome workflow selected to process data for this study. The selected workflow was selected from 5 workflows submitted by the PDTCS. Wiring diagrams for submitted






whole exome workflows submitted by the PDX Development and Trials Centers. Wiring diagrams include nodes and connections. Nodes depict inputs - , outputs - , tools - , and workflows - . Connections between nodes depict that input to a node is from the output of another node. Orange nodes -  identify a tool or a workflow with an available update. c) Panel shows performance evaluations of five workflows submitted by the PDTC. Each workflow was evaluated by SNP (top), INS (middle), and DEL (bottom) with a range of variant allele frequencies (0.025, 0.05, 0.3, 0.2, 0.3). Each plot shows recall and precision respectively on the x and y axis. Results for each of the workflow are shown with the same color: Workflow 1- blue, Workflow 2 – green, Workflow 3- light blue, Workflow 4 – purple, and Workflow 5 – black d) A Venn diagram showing the overlap in high-quality variant calls for model **BL0293-F563** by model using intersected array and removing lower allele frequency (AF) calls. e) Dendrogram of median polish by center (by MBatch) using TMM normalized count per million values. For d) and e), HCI-BCM -Huntsman Cancer Institute/ Baylor College of Medicine, MDACC -MD Anderson Cancer Center, WUSTL -Washington University-St. Louis, and WIST -The Wistar Institute/University of Pennsylvania/MDACC.

Table 1.

Comparison of preclinical study set-ups and end-points at the PDMR and individual PDTs for the temozolomide Reproducibility Pilot.

	PDMR	HCI-BCM	MDACC	WUSTL	WIST
Implantation					
<i>Implantation Type</i>	~1mm ³ Fragment	~1mm ³ Fragment	~1mm ³ Fragment	~3.0 × 10 ⁶ cells, dissociated*	<1 mm ³ fragments in slurry, ~150uL of slurry implanted
<i>Implantation Site</i>	Subcutaneous, single flank	Subcutaneous single flank	Subcutaneous single flank	Subcutaneous single flank	Subcutaneous single flank
<i>Staging Site (mm³)</i>	200	100–200	200	200	100
<i>Cohort Size</i>	8	8	10	10	8
Dosing and Schedule					
<i>Temozolomide Dose (mg/kg)</i>	50	50	50	50	50 and 100
<i>Schedule</i>	QDx5 28d cycle	QDx5 28d cycle	QDx5 28d cycle	QDx5 7d cycle	QDx5 7d cycle
<i>Number of cycles of Treatment</i>	2	1	2	4	2
<i>Route of administration</i>	Oral	Oral	Oral	Oral	Intraperitoneal
Study End -Points					
<i>A</i>	Animal Health	Animal Health	Animal Health	Animal Health	Animal Health
<i>B</i>	Max. tumor size, 4000 m ³	Max. tumor size, 4000 m ³	Max. tumor size, 1600–2000 m ³	Max. tumor size, 1500 m ³	Max. tumor size, 1500 m ³
<i>C</i>	300 days, if Max. TV not reached	0.5 cycles after last dose	When Control TV, 1600–2000 mm ³	4 weeks after last dose	When Control arm TV, 1500 mm ³

PDMR - NCI Patient-Derived Models Repository, HCI-BCM - Huntsman Cancer Institute/Baylor College of Medicine, MDACC - MD Anderson Cancer Center, WUSTL - Washington University-St. Louis, and WIST - The Wistar Institute/University of Pennsylvania/MDACC.

Abbreviations: QDx5 (Once daily for 5 days), TV (tumor volume).

*. This is an average across WUSTL models. The numbers of implanted cells per mouse for each model are: BL0293-F563: 4.0×10^6 ; 172845-121-T: 2.6×10^6 ; and 625472-104-R: 2.5×10^6 .

Table 2.

Statistical tests of treatment vs. control difference Statistical tests of treatment vs. control difference.

PD (Model 625472-104-R)											
Site	v_{21}	$aAUC_{21}$	$aAUC_{max}$	TGI_{21}	$RECIST_{21}$						
					-95,-50,10	-95,-30,20	-95,-30,50	-95,-30,100	-95,-50,50	-95,-50,100	
MDACC	0.163	0.236	0.448	0.067	1.000	1.000	1.000	1.000	1.000	1.000	
WUSTL	0.918	0.376	0.470	0.538	1.000	1.000	1.000	1.000	1.000	1.000	
HCI-BCM	0.143	0.072	0.177	0.814	1.000	1.000	1.000	1.000	1.000	1.000	
PDMR	0.404	0.756	0.501	0.751	1.000	1.000	1.000	1.000	1.000	1.000	
SD (Model 172845-121-T)											
Site	v_{21}	$aAUC_{21}$	$aAUC_{max}$	TGI_{21}	$RECIST_{21}$						
					-95,-50,10	-95,-30,20	-95,-30,50	-95,-30,100	-95,-50,50	-95,-50,100	
MDACC	<.001	0.003	<.001	<.001	0.048	0.048	0.008	0.048	0.008	0.048	
WUSTL	<.001	<.001	<.001	<.001	1.000	0.474	0.211	<.001	0.211	<.001	
HCI-BCM	<.001	<.001	<.001	<.001	0.200	0.026	<.001	<.001	<.001	<.001	
PDMR	<.001	<.001	<.001	<.001	0.003	<.001	<.001	<.001	<.001	<.001	
WIST*	<.001	<.001	<.001	<.001	1.000	1.000	1.000	0.200	1.000	0.200	
	<.001	<.001	<.001	<.001	1.000	1.000	0.200	0.026	0.200	0.026	
CR (Model BL0293-F563)											
Site	v_{21}	$aAUC_{21}$	$aAUC_{max}$	TGI_{21}	$RECIST_{21}$						
					-95,-50,10	-95,-30,20	-95,-30,50	-95,-30,100	-95,-50,50	-95,-50,100	
MDACC	<.001	<.001	<.001	<.001	<.001	0.008	0.008	0.008	0.008	0.008	
WUSTL	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	
HCI-BCM	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	
PDMR	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	
WIST*	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	
	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	

PDMR - NCI Patient-Derived Models Repository, HCI-BCM - Huntsman Cancer Institute/Baylor College of Medicine, MDACC - MD Anderson Cancer Center, WUSTL - Washington University-St. Louis, and WIST - The Wistar Institute/University of Pennsylvania/MDACC.

This table presents the p-values reported for various analytical measures, including change from baseline to 21 days (v_{21}), adjusted area under the curve for 21 days ($aAUC_{21}$), adjusted area under the curve until last measurement ($aAUC_{max}$), tumor growth inhibition at day 21 (TGI_{21}),

and RECIST criteria for various choices of boundaries between CR/PR, PR/SD, and SD/PD given by (**c1**, **c2**, **c3**). For RECIST, p-values are testing PD vs. not PD.

* For WIST, the first row is for temozolomide treated at 50mg/kg and the second row is 100 mg/kg.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript