Generation of TNBC PDX

Gene expression analyses of 93 TNBC PDXs (29657 unique genes/probes) was performed to identify six TNBC subtypes including 2 basal-like (BL1 and BL2), an immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem–like (MSL), and a luminal androgen receptor (LAR) subtype (Supplemental Figure 1) as described previously (*1*). Six to ten-week-old female NSG mice were obtained from The Jacksons Laboratory (https://www.jax.org) and were used for engraftment of human tissue. Mice were anesthetized with isoflurane. An inverted Y−shaped incision was made along the thoracic−inguinal region to expose the mammary glands. Two-to-four million tumor cells mixed with Matrigel in a volume of 30 µl were injected into the 4th inguinal mammary fat pad. The skin was gathered, and the incision closed with wound clips. Following engraftment, tumor growth in PDX mice was monitored.

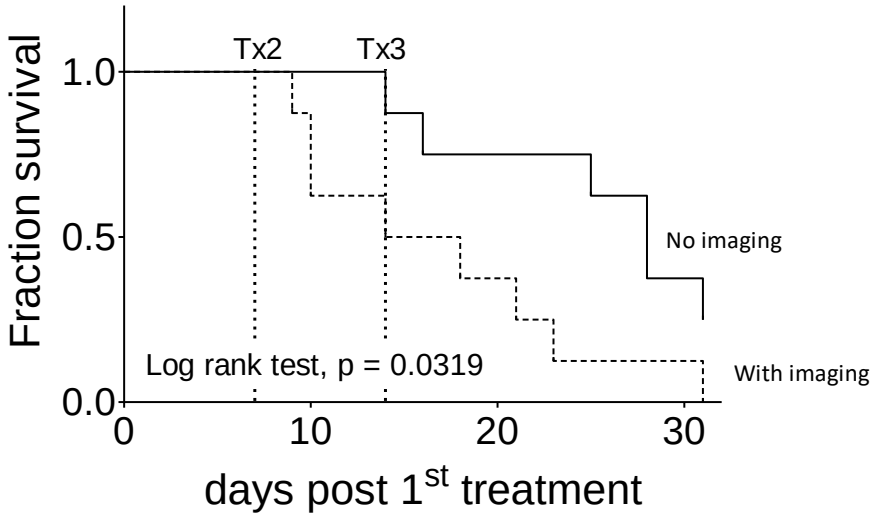
Supplemental Figure 1 is on separate page at end of Supplemental document.

Supplemental Figure 1. Heatmap of the correlation matrix among the 93 PDXs. The heatmap of the correlation matrix among the 93 PDXs was generated with row side color bar indicating subtype (light blue: BL1, dark blue: BL2, red: IM, green: LAR, orange: M, black: MSL, grey: UNS). The column side color bar indicating the PDX lines. The result showed that the PDX of the same lines are highly correlated with each other and mostly belonged to the same TNBC subtype.

Preclinical Studies

Three distinct experiments were carried out. In the first experiment, test-retest studies were performed on consecutive days (Day 1 vs Day 2) to assess the reproducibility of PET image

metrics. Typically, N=8-12 PDX mice for each TNBC subtypes were used in the study. PDX mice were imaged as per the imaging protocol described below. Care was taken to repeat the exact conditions on Day 1 and Day 2 including scanner utilized. Total of 46 PDX mice were used in this cohort. In the second experiment, a separate cohort was used to assess the impact of animal handling/imaging on survival using the study design depicted in Figure 1A. To that end, a separate cohort (N=8; N=16 total) of PDX was administered treatment weekly, but no imaging was performed. Our results suggested that repeat imaging impacted survival (Supplemental Figure 2), and for that reason we excluded +11d imaging time point from the study design. Previous studies have reported that animal handling has dramatic effects on biodistribution and image metrics of FDG uptake (*2*). This observation has broad implications in developing best practices for therapeutic imaging studies, as it suggests that in designing preclinical therapeutic-imaging protocols, the complexity of a combined therapeutic-imaging study should be kept minimal as to not impact the overall objectives of a given investigation.



Supplemental Figure 2. Kaplan–Meier survival curves for two cohort of PDX mice (N=8-10 per group) with weekly combined therapy Docetaxel (20mg/ kg I.P.)/Carboplatin

(50mg/kg I.P). One group (dashed line) was imaged at baseline, +4d, and +11d post baseline (per imaging protocol depicted in Scheme 1C). The second cohort (solid line) was not imaged. Caliper measurements were performed bi-weekly. We observed significant differences in the survival of PDX. For this reason, we dropped the +11d imaging time point from our study design.

The third experiment involved a therapeutic arm with imaging. The study design of the therapeutic arm is depicted in Figure 1A. Preclinical imaging was performed at baseline and +4 days following therapy. In all therapeutic studies, the docetaxel (20mg/kg I.P.)/carboplatin (50mg/kg I.P) was administered at baseline (following imaging) and weekly for a period of four weeks. Tumor volumes were measured bi-weekly as a surrogate measure of response to therapy.

Preclinical PET/CT Imaging

Four hours prior to imaging session, food was removed from metabolism cages while water was given ad libitum. Mice were anesthetized with 2-2.5% isoflurane by inhalation via an induction chamber. Anesthesia was maintained throughout the imaging session by delivering 1%–1.5% isoflurane via a custom-designed nose cone. A heat lamp was used to maintain body temperature. Mice were injected with ^{18}FDG (6.66 – 8.14 MBq) by tail vein immediately before a 0-60 min dynamic small animal PET acquisition. Small animal PET images were acquired on the microPET Focus 220 scanner (Concorde Microsystems Inc., Knoxville, TN) or on the Inveon microPET/CT scanner (Siemens Medical Solutions, Washington D.C.), while the CT images were acquired with the Inveon. CT-based attenuation correction was used. PET scanners are cross-calibrated as per the established standard operating procedures outlined at https://c2ir2.wustl.edu/.

Image Analysis

We evaluated thresholds of SUV_{max} ranging from 100% to 0% in 5% increments for each tumor (i_{th}). Please note, threshold of 100% of SUV_{max} amounts to SUV_{max}, and at the limit of threshold of 0%, the resulting image metrics defines SUV_{mean}. Thus, we evaluated 21 image metrics. These 21 image metrics were evaluated for the whole tumor and using a single highest intensity slice of the tumor. In addition, we calculated peak measures of $4mm^3$, $14mm^3$, and $33mm^3$. Thus, overall 43 imaging metrics were evaluated.

Image Histogram Reproducibility Analysis (IHRA). IHRA was performed as percent threshold of SUV_{max}. At 100% threshold, SUV_{100} corresponds to high intensity voxels (or SUV_{max}). At the limit, as the threshold reaches 0%, SUV_0 is identical to SUV_{mean}. Image voxels were used to compute mean SUV above a given threshold. At each percent threshold (Th varies from 100% to 0%), the SUV_{Th} is calculated as percent of SUV_{max}, i.e., $SUV_{Th}=Th*SUV_{max}/100$. SUV_{th} represents the mean of the voxels with SUV greater than SUV_{Th}. At Th=25% for example, the mean of voxels >= $SUV_{25}=0.25*SUV_{max}$ is calculated. Therefore, as the %Th decreases, the volume of the tumor region under consideration increases with the addition of lower intensity voxels. At each threshold, the mean of the voxels at the threshold is computed by taking the average over all the voxels in the defined tumor region/threshold. This process is repeated for the whole tumor as well as for the metabolically active tumor region in each mouse that is being investigated for tumor reproducibility studies.

Analysis of single slice. In an effort to facilitate analysis, results obtained from whole tumor analysis were compared to those obtained from single slice (SS). A single slice (SS) with the maximum mean activity over the slice (the hottest slice) was selected for processing to investigate the reproducibility of the data. Here also, I_{th} was used to define the tumor region and the hottest

slice data was processed following the same procedure as discussed earlier in the case of whole tumor volume data to compute different thresholds of interest.

SUV_{peak} analyses. SUV_{peak} denotes the mean of all the voxels of in a sphere centered at the hottest voxel. Three different spherical volumes of ~ 4 mm^3 (SUV_{P4}), 14 mm^3 (SUV_{P14}), and 33 mm^3 (SUV_{P33}) were considered corresponding to spheres of radius of 1, 2, 3 voxels. The SUV_{Peak} values were further investigated in the reproducibility and treatment response studies; first to compute the limits of agreement (LOA) and later to evaluate their performance in assessing the response to therapy.

Evaluation of preclinical PERCIST (µPERCIST). The tumor threshold based on the PERCIST criteria (*3*) is provided by Th=*α*[mean concentration of liver ROI] + β*[standard deviation of liver ROI]*. Liver ROIs were determined 50-60min post injection of FDG. Optimization of α and β entails maximizing Lin's concordance correlation coefficient (LCC) while minimizing the repeatability coefficient (RC) (which would minimize the 95% CI, hence maximize reproducibility); thus the objective function to maximize is the ratio of LCC (*4*) to the RC (defined in Supplemental Statistics section). A range of values for α and β were evaluated and optimized. Implementation of *µPERCIST* relies on evaluation of SUV_{peak} which was described earler.

Statistical Analysis

Reproducibility Statistics. Let Δ denote the within mouse difference between the measurements, and N denote the number of paired measurements. The standard deviation for the mean difference is calculated using Eq. 1, and the within-mouse standard deviation (wSD), using Eq. 2.

$$dsd = \sqrt{\frac{\sum(\Delta_i)^2}{N}} \qquad \text{Eq. 1}$$

$$wSD = \frac{dsd}{\sqrt{2}} \quad \text{Eq. 2}$$

The 95% confidence limits (95% CL) in the BA plots are the limits of agreement (LOA) defined as the mean difference ± the repeatability coefficient (RC), defined in Eq.3. These limits are independent of the sample size so that the results from an individual test-retest experiment is expected to fall within these boundaries 95% of the time.

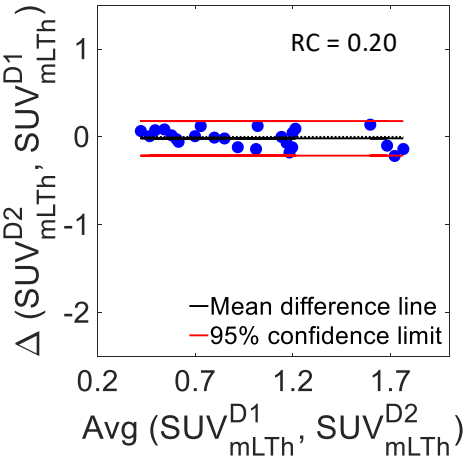
$$RC = 1.97 \times \sqrt{2} \times wSD = 2.77 \times wSD \quad \text{Eq. 3}$$

Two methods for assessing reproducibility were used, Lin's concordance correlation coefficient (LCC) (*4*) and Bland-Atlman plots (BA) (*5*). The LCC, being the product of the Pearson correlation coefficient (PCC) and the bias correction factor (BCF), accounts for both precision and accuracy. The method outlined in Watson and Petrie (*6*) was followed to calculate these metrics. The procedure used to calculate the statistical parameter for the BA plots are summarize in Galbraith (*7*) and Raunig (*8*).

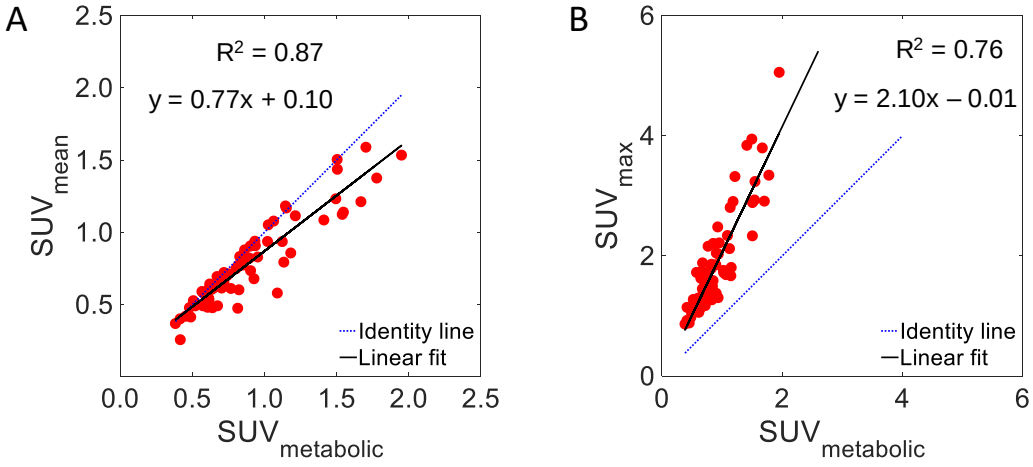
Performance analysis of image metrics response to therapy. Sensitivity, the number of positive responses that are correctly classified as positive; Specificity, the number of negative responses that are correctly classified as negative; Precision, the probability that a prediction of positive is actually positive; Negative predictive value (NPV), the probability that a prediction of negative is actually negative; Accuracy, the fraction of correct prediction to the total number of observation; and F-score, the harmonic mean of precision and sensitivity, are the standard performance binary classification metrics used to assess the response to the therapy (*9,10*). The evaluations were categorized as; True Positive (TP) when outcome was a positive response (True) and SUV change also predicted a positive response (True). False Negative (FN) when outcome truth was

a positive result (True), but SUV change predicted a non-response (False). True Negative (TN) when outcome showed a nonresponse (False), and the SUV change also predicted a nonresponse. False Positive (FP) when outcome is a nonresponse, but the SUV change predicts positive response (True). If an image metrics of response to therapy was within the LOA, it was considered indistinguishable from metric variability. In that scenario, the image metric was not used in calculating performance.

SUPPLEMENTAL RESULTS



Supplemental Figure 3. Tumor SUV BA plot for optimized liver threshold.



Supplemental Figure 4. Correlation between SUV_{mean} and SUV_{max} to SUV of metabolic tumor ($SUV_{metabolic}$).

The performance of image metrics is tabulated in Supplemental Table 1. The accuracy of image metrics by subtype is tabulated in Table S2.

Supplemental TABLE 1A. Performance of imaging metrics to predict response to therapy for data points outside the LOA.

SUV metric	Sensitivity	Specificity	Precision	NPV	Accuracy	F-Score	Uncertain Fraction (%)
ΔSUV_{max}	1.00	0.25	0.57	1.00	0.63	0.73	45
ΔSUV_{mean}	0.92	0.22	0.63	0.67	0.64	0.75	24
ΔSUV_{25}	0.91	0.22	0.59	0.67	0.60	0.72	31
ΔSUV_{mean} (SS)	0.71	0.11	0.56	0.20	0.48	0.63	21
ΔSUV_{25} (SS)	0.91	0.25	0.63	0.67	0.63	0.74	34
ΔSUV_{P4}	1.00	0.29	0.62	1.00	0.67	0.77	48
ΔSUV_{P14}	0.91	0.25	0.63	0.67	0.63	0.74	34
ΔSUV_{P33}	0.89	0.25	0.57	0.67	0.59	0.69	41

Supplemental TABLE **1B**. Performance of imaging metrics to predict response to therapy for all data points (29 samples)

SUV metric	Sensitivity	Specificity	Precision	NPV	Accuracy	F score
ΔSUV_{max}	0.74	0.30	0.67	0.38	0.59	0.70
ΔSUV_{mean}	0.84	0.30	0.70	0.50	0.66	0.76
ΔSUV_{25}	0.79	0.30	0.68	0.43	0.62	0.73
ΔSUV_{mean} (SS)	0.74	0.10	0.61	0.17	0.52	0.67
ΔSUV_{25} (SS)	0.74	0.30	0.67	0.38	0.59	0.70
ΔSUV_{P4}	0.79	0.30	0.68	0.43	0.62	0.73
ΔSUV_{P14}	0.89	0.30	0.71	0.60	0.69	0.79
ΔSUV_{P33}	0.79	0.30	0.68	0.43	0.62	0.73

All performance metrics range from 0 (no prediction) to 1 (high). Data points within the LOA were penalized for uncertainly by exclusion from the analysis. Refer to supplementary Table S1 for performance analysis inclusive of data points within the LOA. Uncertain Fraction is the percent of studies within the LOA which were not used in prediction (due to uncertainty) for each image metric.

Supplemental TABLE 2A: Accuracy of prediction by PDX subtype* (for data points outside the LOA)

WHIM	SUV_{mean}	SUV_{25}	SUV_{max}	SUV_{mean} (SS)	SUV_{25} (SS)	SUV_{P4}	SUV_{P14}	SUV_{P33}	SUV_{P64}
IM	1.00	1.00	1.00	0.50	1.00	1.00	1.00	1.00	1.00
BL1	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00
BL2	0.55	0.50	0.50	0.50	0.50	0.44	0.44	0.38	0.38
M	0.33	0.33	0.50	0.33	0.50	1.00	0.50	0.50	0.50

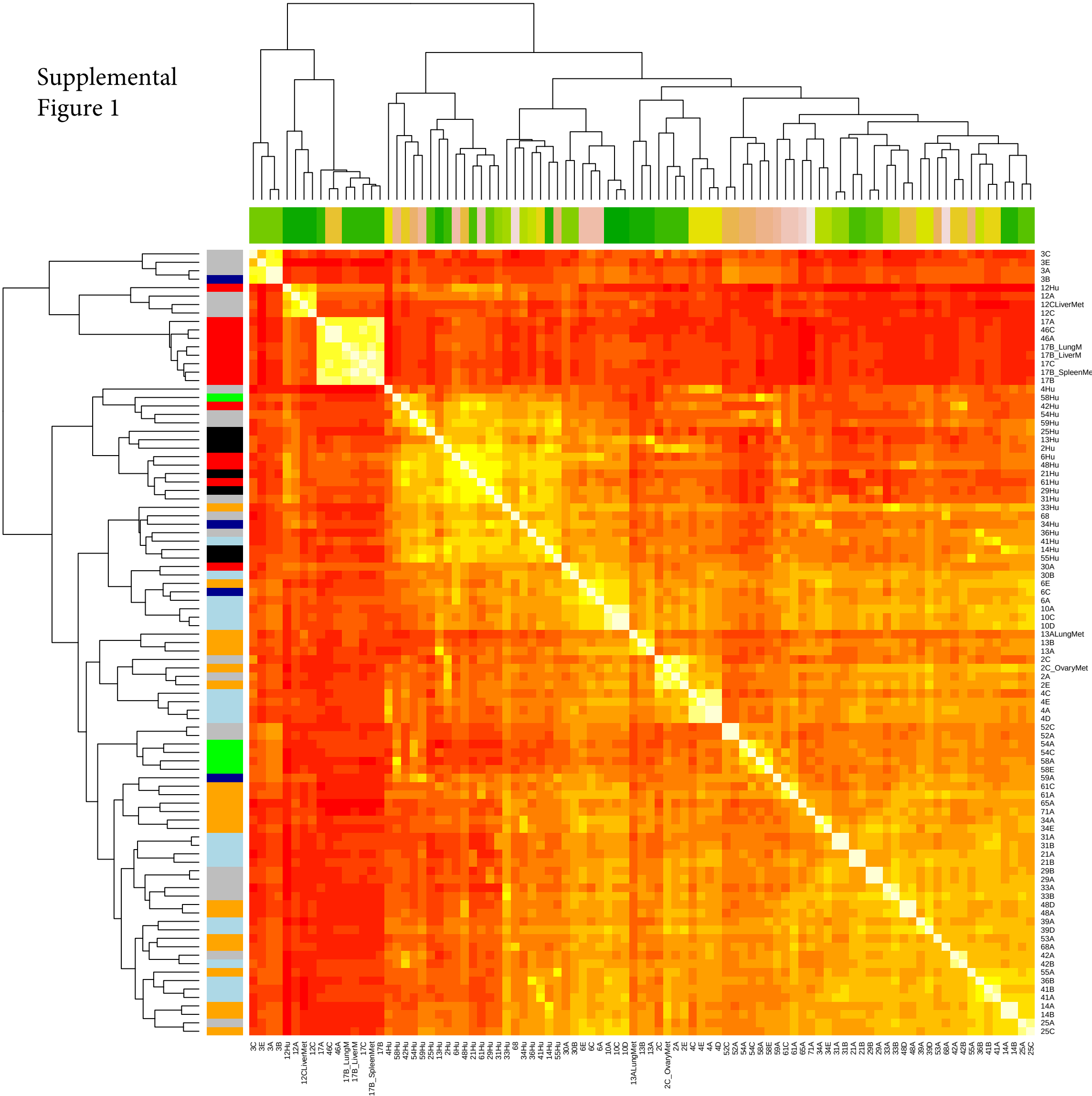
*Accuracy = (TP + TN)/(P+N); LAR PDX are not included due to low sample count (N=4).

Supplemental TABLE 2B: Accuracy for each PDX* (for data points inclusive of within LOA)

PDX	SUV_{mean}	SUV_{25}	SUV_{max}	SUV_{mean} (SS)	SUV_{25} (SS)	SUV_{P4}	SUV_{P14}	SUV_{P33}	SUV_{P64}
IM	0.75	0.75	0.63	0.63	0.50	0.75	0.88	0.75	0.75
BL1	1.00	0.75	1.00	0.50	1.00	1.00	1.00	1.00	1.00
BL2	0.62	0.62	0.54	0.54	0.62	0.54	0.62	0.54	0.54
M	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33

*Accuracy = (TP + TN)/(P+N); LAR PDX are not included due to low sample count (N=4).

Supplemental
Figure 1



References

1. Lehmann BD, Bauer JA, Chen X, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* 2011;121:2750-2767.

2. Fueger BJ, Czernin J, Hildebrandt I, et al. Impact of animal handling on the results of 18F-FDG PET studies in mice. *J Nucl Med.* 2006;47:999-1006.

3. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50 Suppl 1:122S-150S.

4. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45:255-268.

5. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135-160.

6. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology.* 2010;73:1167-1179.

7. Galbraith SM, Lodge MA, Taylor NJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR Biomed.* 2002;15:132-142.

8. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res.* 2015;24:27-67.

9. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology.* 2016;4:320-330.

10. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10:e0118432.