Anurag et al., Cancer Discovery, 2022
Supplementary Figure legends

Fig S1: Cohort characteristics.

A) Correlation between average tumor content as assessed by immunohistochemistry (IHC) and mRNA-protein correlation for all baseline samples for which we had paired mRNA and protein data. Samples indicated by teal points were excluded from this study due to poor tumor content and poor mRNA-protein correlation (tumor content <45%).

B) Pairwise correlation analysis depicting correlations across (green) and within (red) TMT-plexes at the protein (left) and phosphopeptide level (right). A common reference sample from our prospective BRCA study (PMID: 33212010) was measured in each TMT plex as part of this study and showed consistently high correlations across all 8 TMT plexes. Pairwise within-plex correlations of tumor samples are shown in red.

C) Gene-wise mRNA-protein correlations ordered from most negative to most positive by Spearman rank correlation coefficient are shown in the top panel. Signed log10 p-values were used as input for GSEA using the KEGG pathway database. Top pathways enriched for genes that had positive correlation between mRNA and protein are shown in the middle panel while top pathways enriched for genes with lower correlation are shown in the bottom panel.

D) Gene function prediction performance quantified by AUROC shows co-expression networks based on proteomic profiles outperformed RNA-based networks for predicting KEGG pathway membership. Dotted lines indicate 10% increase or decrease in prediction performance.

E) Genotoxic stress sites (sites induced by nocodazole, UV, and ionizing radiation) and ATM and CDK1/2 target sites are acutely induced by chemotherapy. Volcano plot shows results from Post-Translational Modification-Set Enrichment Analysis (PTM-SEA; PMID: 30563849) using the signed (by direction of change) -log10 p-values from paired Wilcoxon signed tests comparing phosphosite levels in on-treatment (cycle 1 day 3) tumors to matching baseline tumors (n = 13).

F) Proportion plot showing distribution of samples obtained from baseline tumors grouped by PAM50 (Basal, HER2, LumA, LumB), TNBC subtyping (BL1-basal like 1, BL2-basal like 2, IM-immunomodulatory, LAR-luminal androgen receptor, M-mesenchymal, MSL-mesenchymal stem like, UNS-unspecified tumors) and Race (AA-african american, C-caucasian, Others). The left panel shows all samples, and the right panel shows samples segregated into two groups based on pCR. P-values were obtained using Fisher's exact test.

G) Heatmap showing distribution of COSMIC mutational signatures in thebaseline (pre-treatment) samples. Signatures 6 and 10, which are associated with MMR and POLE mutations respectively, were higher in RCB II/III categories (ANOVA test, $p<0.05$).

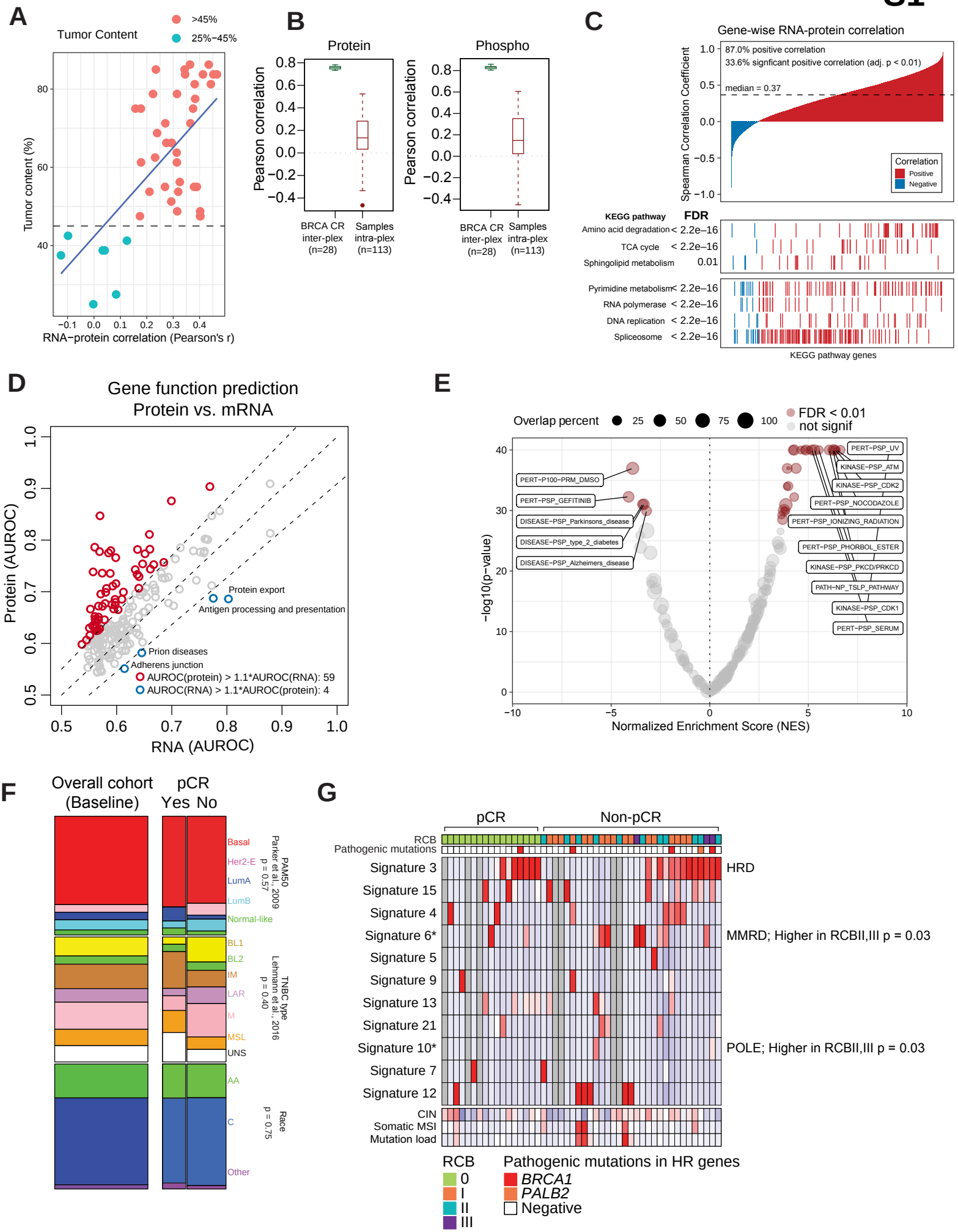

A
Tumor Content
>45%
25%−45%
Tumor content (%)
RNA−protein correlation (Pearson's r)
B
Protein
Phospho
Pearson correlation
BRCA CR inter-plex (n=28)
Samples intra-plex (n=113)
C
Gene-wise RNA-protein correlation
87.0% positive correlation
33.6% signficant positive correlation (adj. p < 0.01)
median = 0.37
Spearman Correlation Coefficient
Correlation
Positive
Negative
KEGG pathway
FDR
Amino acid degradation < 2.2e−16
TCA cycle < 2.2e−16
Sphingolipid metabolism 0.01
Pyrimidine metabolism < 2.2e−16
RNA polymerase < 2.2e−16
DNA replication < 2.2e−16
Spliceosome < 2.2e−16
KEGG pathway genes
D
Gene function prediction
Protein vs. mRNA
Protein (AUROC)
RNA (AUROC)
Protein export
Antigen processing and presentation
Prion diseases
Adherens junction
AUROC(protein) > 1.1*AUROC(RNA): 59
AUROC(RNA) > 1.1*AUROC(protein): 4
E
Overlap percent
25
50
75
100
FDR < 0.01
not signif
−log10(p−value)
Normalized Enrichment Score (NES)
PERT−P100−PRM_DMSO
PERT−PSP_GEFITINIB
DISEASE−PSP_Parkinsons_disease
DISEASE−PSP_type_2_diabetes
DISEASE−PSP_Alzheimers_disease
PERT−PSP_UV
KINASE−PSP_ATM
KINASE−PSP_CDK2
PERT−PSP_NOCODAZOLE
PERT−PSP_IONIZING_RADIATION
PERT−PSP_PHORBOL_ESTER
KINASE−PSP_PKCD/PRKCD
PATH−NP_TSLP_PATHWAY
KINASE−PSP_CDK1
PERT−PSP_SERUM
F
Overall cohort (Baseline)
pCR
Yes
No
PAM50
Parker et al., 2009
p = 0.57
Basal
Her2-E
LumA
LumB
Normal-like
TNBC type
Lehmann et al., 2016
p = 0.40
BL1
BL2
IM
LAR
M
MSL
UNS
Race
p = 0.75
AA
C
Other
G
pCR
Non-pCR
RCB
Pathogenic mutations
Signature 3
Signature 15
Signature 4
Signature 6*
Signature 5
Signature 9
Signature 13
Signature 21
Signature 10*
Signature 7
Signature 12
CIN
Somatic MSI
Mutation load
HRD
MMRD; Higher in RCBII,III p = 0.03
POLE; Higher in RCBII,III p = 0.03
RCB
0
I
II
III
Pathogenic mutations in HR genes
BRCA1
PALB2
Negative

Fig S2: PDL1 IHC, distribution of Rb phosphorylation, and association of Rb protein and drug sensitivity in cell lines.

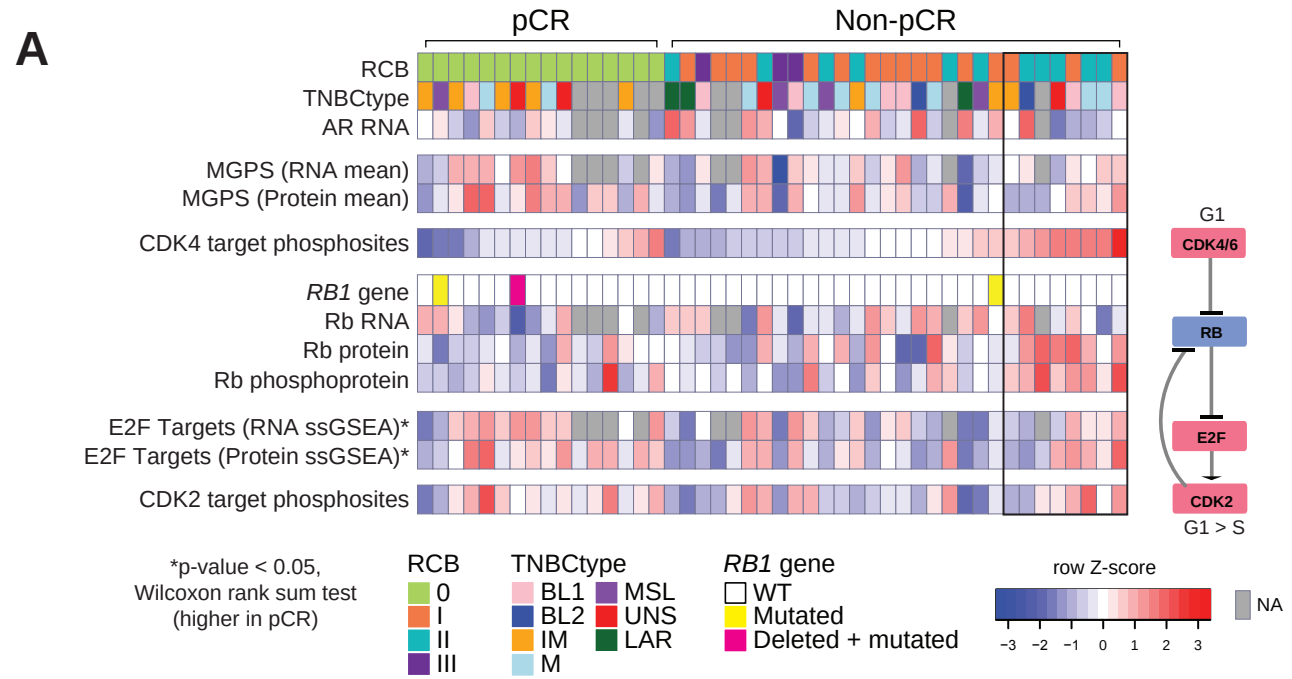
A) While cell cycle features are elevated in pCR tumors, a subset of non-pCR tumors have elevated levels of CDK4 target sites and Rb levels. Heatmap shows RNA- and protein-based multi-gene proliferation scores (MGPS; PMID: 28045625; PMID: 12058064), single sample PTM-SEA scores for CDK4 and CDK2 target sites (PMID: 30563849), ssGSEA scores for E2F targets (PMID:26771021; PMID: 19847166), and other proteogenomic features for genes regulating the G1/S transition of the cell cycle (see pathway on right). Box outlines samples from non-pCR tumors with high CDK4 activity and Rb phosphorylation. Asterisks indicate $p<0.05$ by Wilcoxon rank sum test comparing non-pCR to pCR tumors.

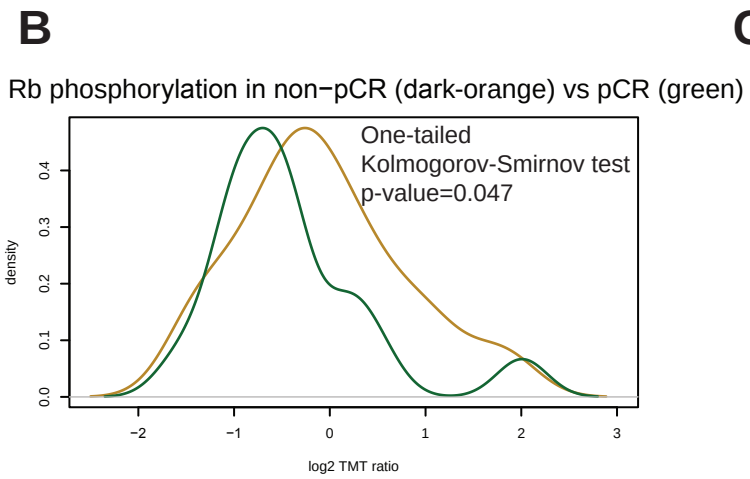
B) The distribution of Rb phosphorylation levels in non-pCR tumors (brown) is overlapping with but also shifted towards higher levels than in pCR tumors (green). Density plots for each group are shown.

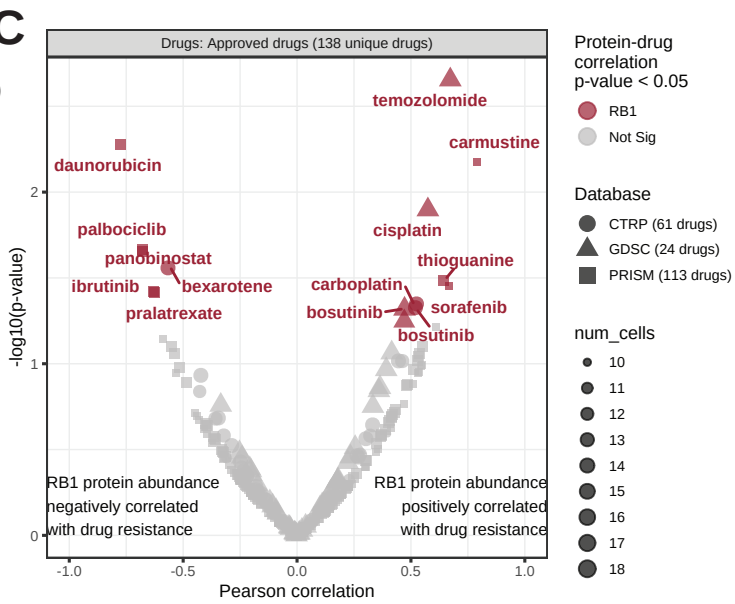
C) Plot showing Pearson correlations between Rb protein by TMT profiling in TNBC cell lines and responses from all approved drugs in CTRP (Cancer Therapeutics Response Portal), GDSC (Genomics of Drug Sensitivity in Cancer), and PRISM (Profiling Relative Inhibition Simultaneously in Mixtures) databases from the DepMap (Dependency Map) resource. Note that, of all Rb protein-drug response correlations, platinum compounds such as cisplatin and carboplatin, as well as an alkylating agent that induces DNA damage, temozolomide, were among the most positively correlated pairs (higher Rb protein in TNBC cell lines with increasing resistance to those drugs) while response to palbociclib was negatively correlated (higher Rb protein in TNBC cell lines with decreasing resistance to palbociclib).

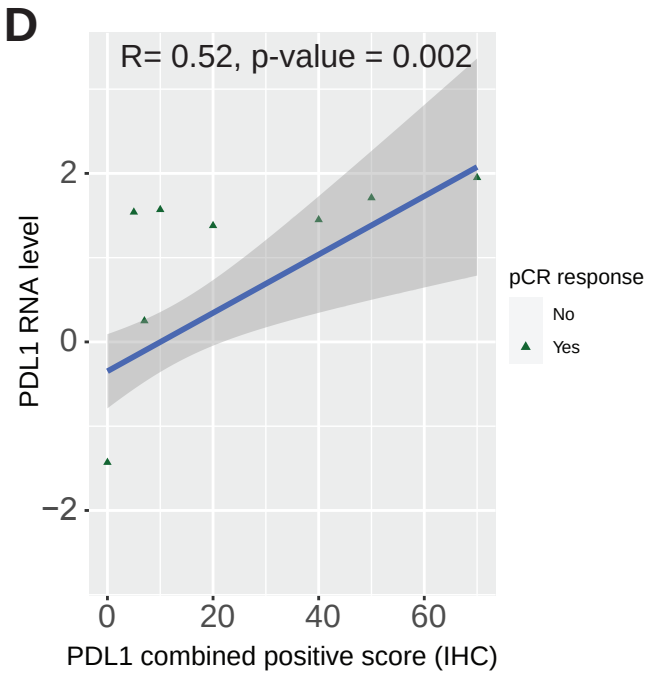
D)Scatter plots showing correlation between PDL1 IHC levels with PDL1 RNA levels across tumors.

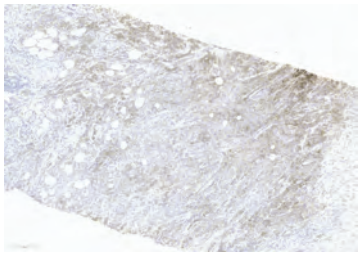
E-F) Representative images for patient samples with high (A) and low (B) PDL1 IHC staining.


A
pCR
Non-pCR
RCB
TNBCtype
AR RNA
MGPS (RNA mean)
MGPS (Protein mean)
CDK4 target phosphosites
RB1 gene
Rb RNA
Rb protein
Rb phosphoprotein
E2F Targets (RNA ssGSEA)*
E2F Targets (Protein ssGSEA)*
CDK2 target phosphosites
G1
CDK4/6
RB
E2F
CDK2
G1 > S
*p-value < 0.05,
Wilcoxon rank sum test
(higher in pCR)
RCB
0
I
II
III
TNBCtype
BL1
BL2
IM
M
MSL
UNS
LAR
RB1 gene
WT
Mutated
Deleted + mutated
row Z-score
−3 −2 −1 0 1 2 3
NA


B
Rb phosphorylation in non−pCR (dark-orange) vs pCR (green)
One-tailed
Kolmogorov-Smirnov test
p-value=0.047
density
log2 TMT ratio


C
Drugs: Approved drugs (138 unique drugs)
temozolomide
carmustine
daunorubicin
palbociclib
panobinostat
ibrutinib
bexarotene
pralatrexate
cisplatin
thioguanine
carboplatin
bosutinib
sorafenib
bosutinib
-log10(p-value)
RB1 protein abundance
negatively correlated
with drug resistance
RB1 protein abundance
positively correlated
with drug resistance
Pearson correlation
Protein-drug
correlation
p-value < 0.05
RB1
Not Sig
Database
CTRP (61 drugs)
GDSC (24 drugs)
PRISM (113 drugs)
num_cells
10
11
12
13
14
15
16
17
18


D
R= 0.52, p-value = 0.002
PDL1 RNA level
PDL1 combined positive score (IHC)
pCR response
No
Yes


E

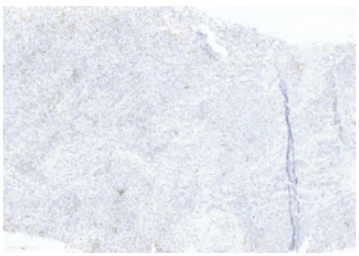

F

Fig S3: Metabolic multi-omic signature and results from PTM-SEA.

A) The multi-omics metabolic gene signature identified in this study was further investigated in patients treated with carboplatin and paclitaxel in the BrighTNess clinical trial. The mean mRNA expression score for this signature was significantly higher in residual disease. P-value was calculated using the Wilcoxon rank sum test.

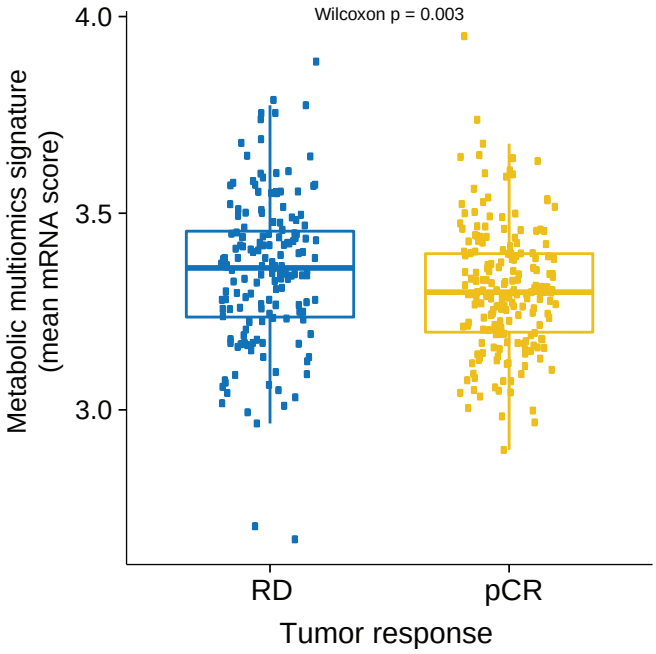

Wilcoxon p = 0.003
Metabolic multiomics signature
(mean mRNA score)
4.0
3.5
3.0
RD
pCR
Tumor response

Fig S4: Proteogenomic prioritization of candidates driving chemo-resistance in TNBC.

A) Chromosome-wise differences in the SCNA landscape in non-pCR and pCR samples are shown as the top two tracks. Amplification and deletion events are indicated by red and blue respectively. The bottom panels show differential mRNAs and proteins up- and down-regulated in non-pCR compared to pCR cases arranged based on their chromosomal location to align with the SCNA tracks.

B) Venn-diagram showing differential (non-pCR vs. pCR) mRNA and proteins located on cytoband 8q21.3.

C) Pircos (proteogenomic circos) plots for chromosome 19 showing signed -log10 p-values for mRNA and protein based on Wilcoxon rank-sum tests comparing non-pCR to pCR cases and frequency of amplifications and deletions (CNV) in non-pCR and pCR. Genes annotated in blue on the outermost track represent those to be low at the CNV, mRNA, and protein level from (C).

D) Boxplot showing comparing RNA expression of DNA-repair genes located on 19q13.31-33 in the previously published BrighTNess clinical trial (Treatment Arm C, paclitaxel).

E) Kaplan-Meier (KM) curve depicting metastasis free survival of patients with TNBC breast cancer from the Hatzis dataset. Tumors are categorized into high and low LIG1 based on mean RNA expression cutoff.

F) Boxplot showing LIG1 mRNA levels to be significantly lower in baseline TNBC tumors resistant (SD + PD) to cisplatin monotherapy vs sensitive (CR + PR) by wilcoxon test. Response Evaluation Criteria in Solid Tumors (RECIST) as follows: PD, progressive disease; SD, stable disease; PR, partial response; CR, complete response. Data is derived from GSE18864.

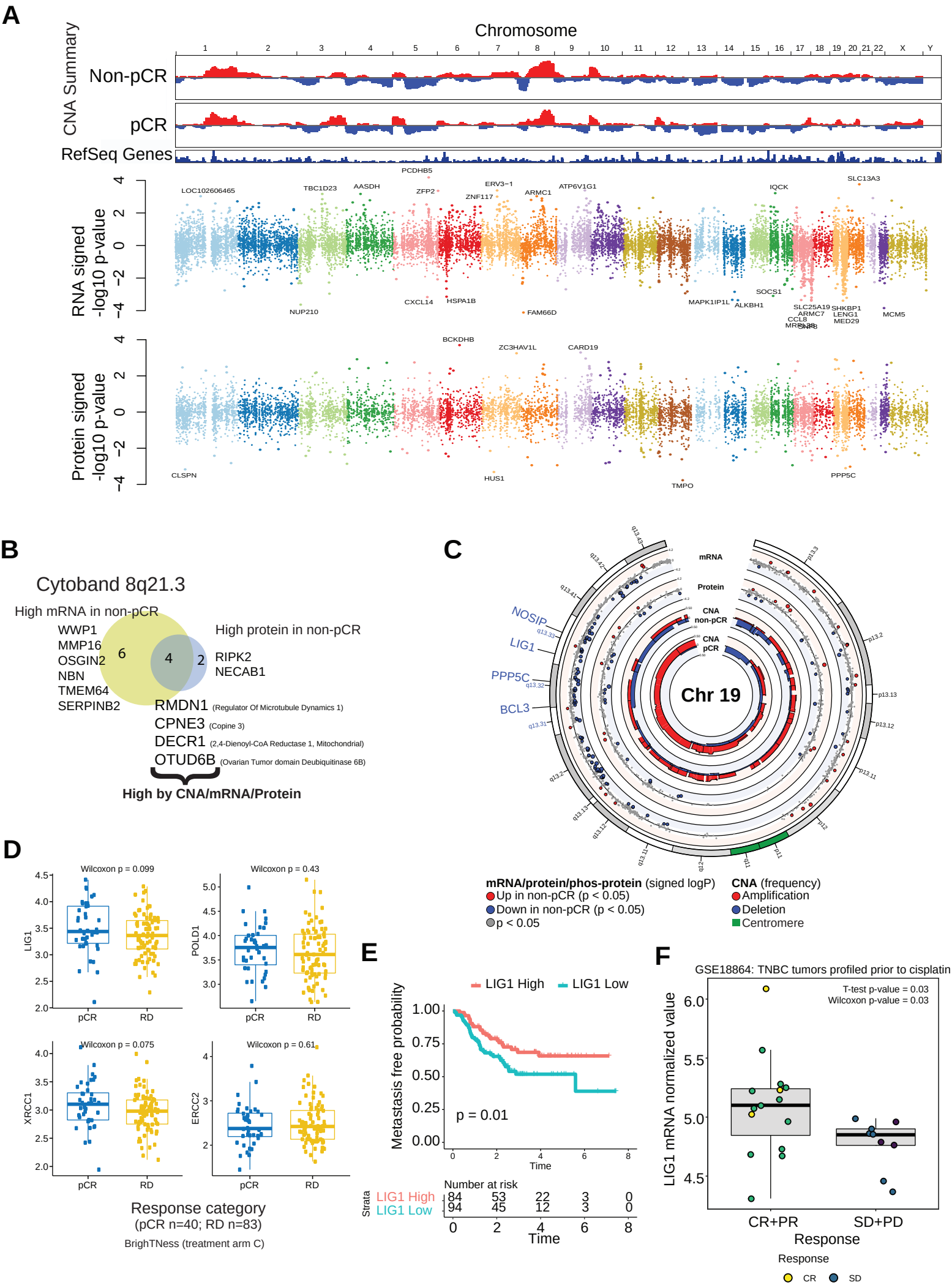

A
Chromosome
CNA Summary
Non-pCR
pCR
RefSeq Genes
RNA signed -log10 p-value
Protein signed -log10 p-value
PCDHB5
LOC102606465
TBC1D23
AASDH
ZFP2
ZNF117
ERV3-1
ARMC1
ATP6V1G1
IQCK
SLC13A3
NUP210
CXCL14
HSPA1B
FAM66D
MAPK1IP1L
ALKBH1
SOCS1
SLC25A19
ARMC7
SHKBP1
LENG1
MED29
MCM5
CCL8
BCKDHB
ZC3HAV1L
CARD19
CLSPN
HUS1
TMPO
PPP5C
B
Cytoband 8q21.3
High mRNA in non-pCR
WWP1
MMP16
OSGIN2
NBN
TMEM64
SERPINB2
6
4
2
High protein in non-pCR
RIPK2
NECAB1
RMDN1 (Regulator Of Microtubule Dynamics 1)
CPNE3 (Copine 3)
DECR1 (2,4-Dienoyl-CoA Reductase 1, Mitochondrial)
OTUD6B (Ovarian Tumor domain Deubiquitinase 6B)
High by CNA/mRNA/Protein
C
mRNA
Protein
CNA non-pCR
CNA pCR
Chr 19
NOSIP
LIG1
PPP5C
BCL3
mRNA/protein/phos-protein (signed logP)
Up in non-pCR (p < 0.05)
Down in non-pCR (p < 0.05)
p < 0.05
CNA (frequency)
Amplification
Deletion
Centromere
D
Wilcoxon p = 0.099
Wilcoxon p = 0.43
Wilcoxon p = 0.075
Wilcoxon p = 0.61
LIG1
POLD1
XRCC1
ERCC2
pCR
RD
Response category
(pCR n=40; RD n=83)
BrighTNess (treatment arm C)
E
LIG1 High
LIG1 Low
Metastasis free probability
p = 0.01
Time
Number at risk
Strata
LIG1 High 84 53 22 3 0
LIG1 Low 94 45 12 3 0
F
GSE18864: TNBC tumors profiled prior to cisplatin
T-test p-value = 0.03
Wilcoxon p-value = 0.03
LIG1 mRNA normalized value
CR+PR
SD+PD
Response
CR
SD
PR
PD

Fig S5: Features associated with *LIG1* genomic status.

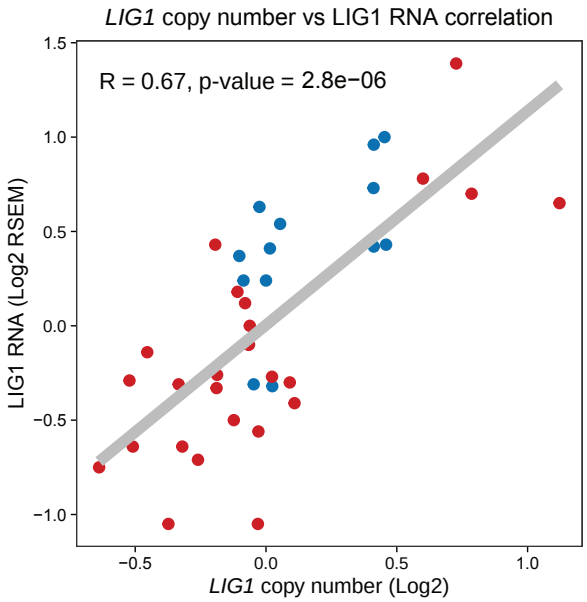
A-B) Plot showing results for Spearman rank correlation between *LIG1* copy number and LIG1 mRNA expression (A) and LIG1 protein level (B). The samples are colored according to treatment response.

C) Boxplot showing immune stimulatory scores across *LIG1* copy number and chemotherapy response categories. P-values are calculated using the Wilcoxon rank-sum test.

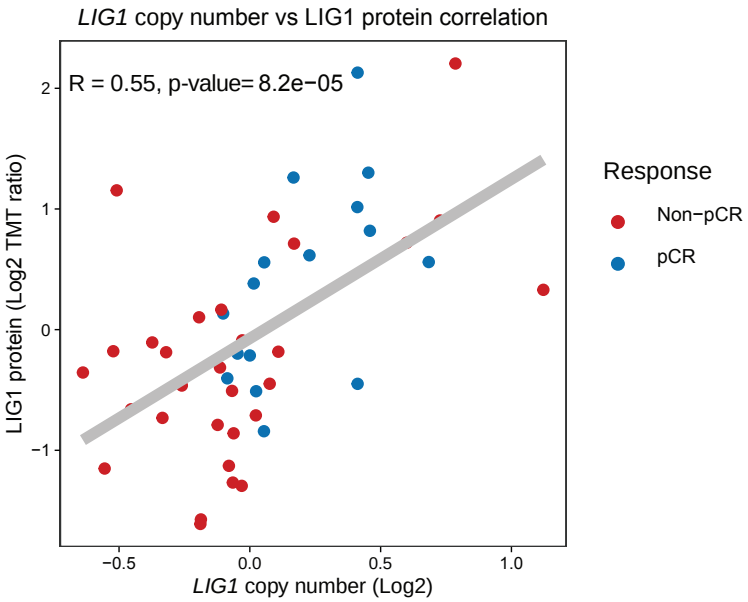
D) LIG1 loss tumors have higher CDK1/2 activity than non-loss tumors. Volcano plot shows results from Post-Translational Modification-Set Enrichment Analysis (PTM-SEA; PMID: 30563849) using the signed -log10 p-values from Wilcoxon rank sum tests comparing phosphosite levels in LIG1 loss (GISTIC = -1) tumors to tumors with GISTIC = 0-2 as input.

E) Heatmap showing ssPTMSEA scores for significantly differential kinases and pathways shown in D. *LIG1* single copy loss and RCB class are shown as additional rows.

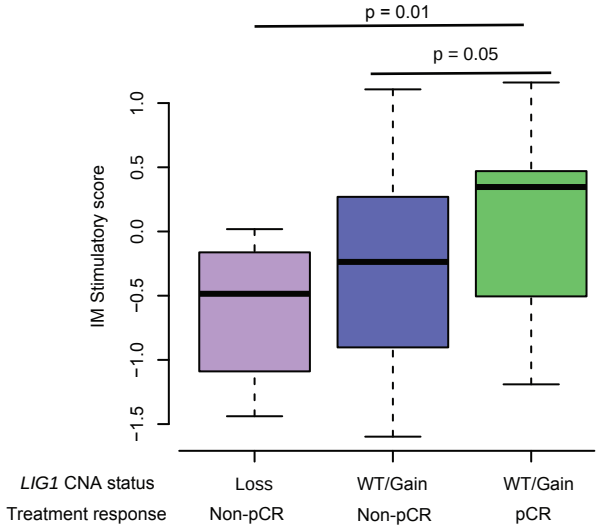
A


LIG1 copy number vs LIG1 RNA correlation
R = 0.67, p-value = 2.8e−06
LIG1 RNA (Log2 RSEM)
LIG1 copy number (Log2)

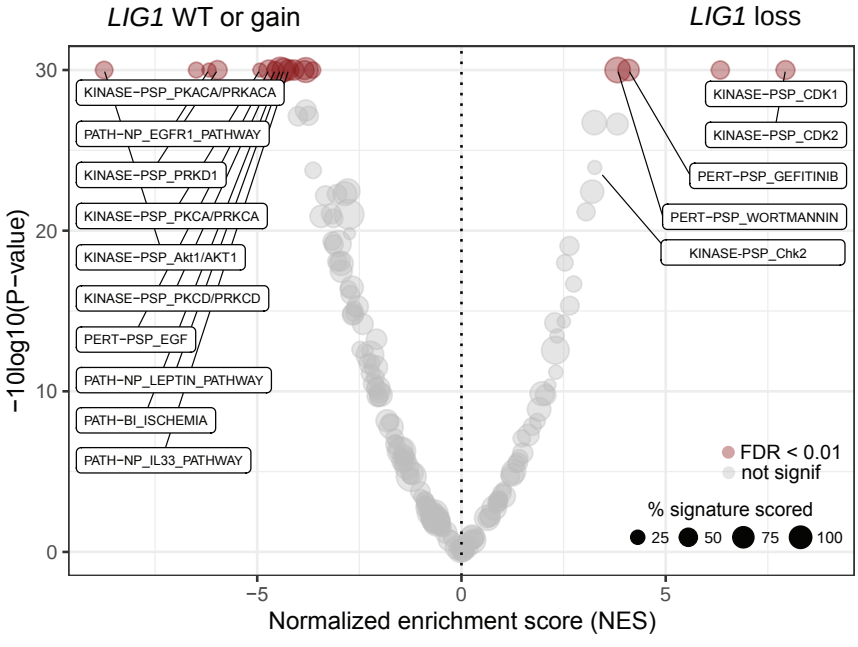
B


LIG1 copy number vs LIG1 protein correlation
R = 0.55, p-value= 8.2e−05
LIG1 protein (Log2 TMT ratio)
LIG1 copy number (Log2)
Response
Non−pCR
pCR

C


p = 0.01
p = 0.05
IM Stimulatory score
LIG1 CNA status
Loss
WT/Gain
WT/Gain
Treatment response
Non-pCR
Non-pCR
pCR

D


LIG1 WT or gain
LIG1 loss
KINASE−PSP_PKACA/PRKACA
PATH−NP_EGFR1_PATHWAY
KINASE−PSP_PRKD1
KINASE−PSP_PKCA/PRKCA
KINASE−PSP_Akt1/AKT1
KINASE−PSP_PKCD/PRKCD
PERT−PSP_EGF
PATH−NP_LEPTIN_PATHWAY
PATH−BI_ISCHEMIA
PATH−NP_IL33_PATHWAY
KINASE−PSP_CDK1
KINASE−PSP_CDK2
PERT−PSP_GEFITINIB
PERT−PSP_WORTMANNIN
KINASE-PSP_Chk2
FDR < 0.01
not signif
% signature scored
25
50
75
100
−10log10(P−value)
Normalized enrichment score (NES)

E

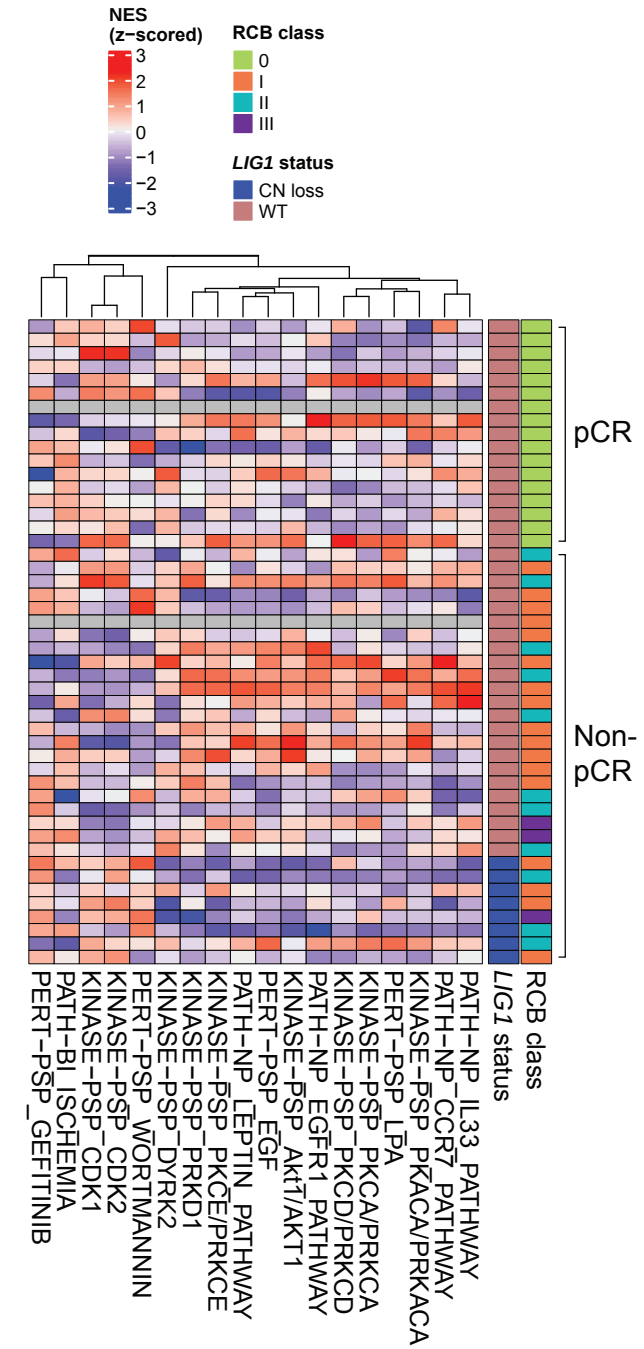

NES
(z−scored)
RCB class
0
I
II
III
LIG1 status
CN loss
WT
pCR
Non-pCR
PERT−PSP_GEFITINIB
PATH−BI_ISCHEMIA
KINASE−PSP_CDK1
KINASE−PSP_CDK2
PERT−PSP_WORTMANNIN
KINASE−PSP_DYRK2
KINASE−PSP_PRKD1
KINASE−PSP_PKCE/PRKCE
PATH−NP_LEPTIN_PATHWAY
PERT−PSP_EGF
KINASE−PSP_Akt1/AKT1
PATH−NP_EGFR1_PATHWAY
KINASE−PSP_PKCD/PRKCD
KINASE−PSP_PKCA/PRKCA
PERT−PSP_LPA
KINASE−PSP_PKACA/PRKACA
PATH−NP_CCR7_PATHWAY
PATH−NP_IL33_PATHWAY
LIG1 status
RCB class

Fig S6: LIG1 association with advanced TNBC disease in preclinical models and independent cohorts.

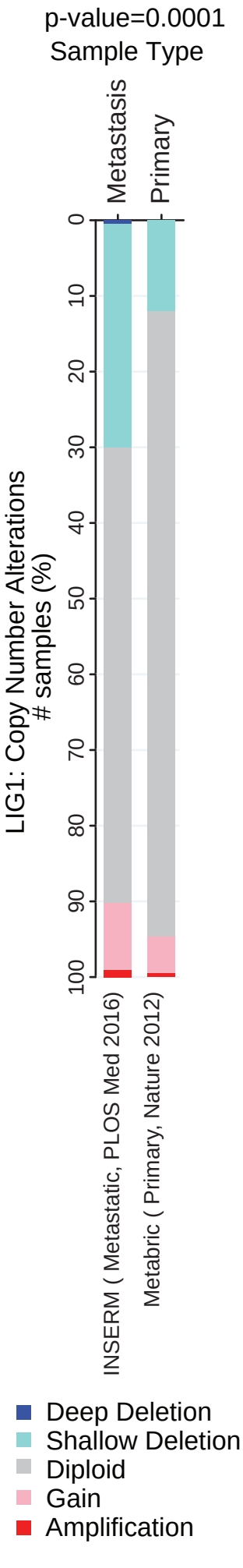
A) Proportion plot showing distribution of LIG1 copy number alterations in TCGA-Breast (primary disease) and INSERM (metastatic disease) cohorts. P-value was obtained using Fisher's exact test.

B) Quantification by densitometry using the same western blots for LIG1, POLD1 and XRCC1, along with GAPDH as a control, in WHIM 68, 74 and 75 as shown in Fig. 6A. Normalized LIG1, POLD1, and XRCC1 protein abundance from each PDX was calculated by dividing levels of each protein by their corresponding GAPDH signals. These values were then normalized to WHIM68 (primary disease).

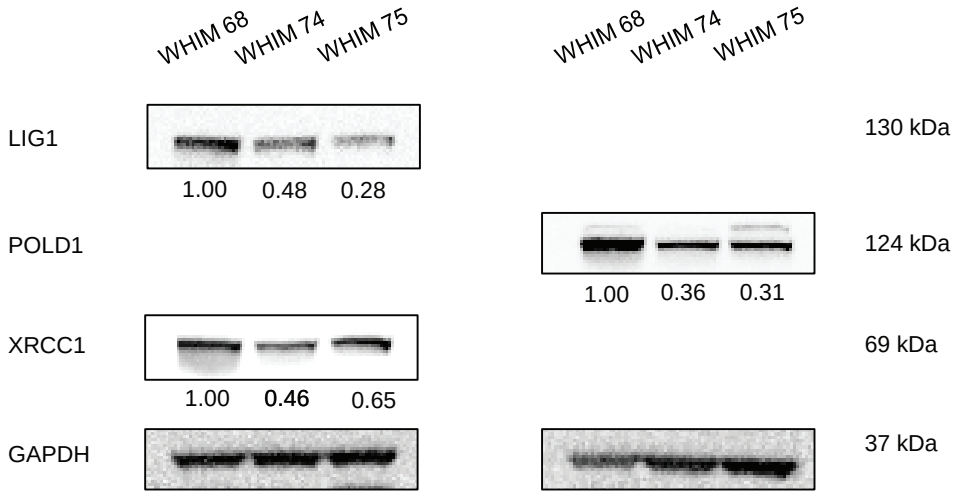
C) Tumor volumes in PDX models treated with 4 weekly cycles of vehicle or 20 mg/kg docetaxel. P-values derived from a general linear model within each PDX were computed using estimated mean log2 fold changes in tumor volume at Day 28 vs Day 0 for each treatment arm.

D) Boxplots showing LIG1 mRNA levels in TNBCs PDXs categorized into complete response (CR) and non-CR groups. After 4 weeks of docetaxel treatment, CR was defined as PDXs with non-palpable tumors and non-CR defined for PDXs with residual tumors with measurable dimensions. Wilcoxon rank sum test was used to compare the two groups.

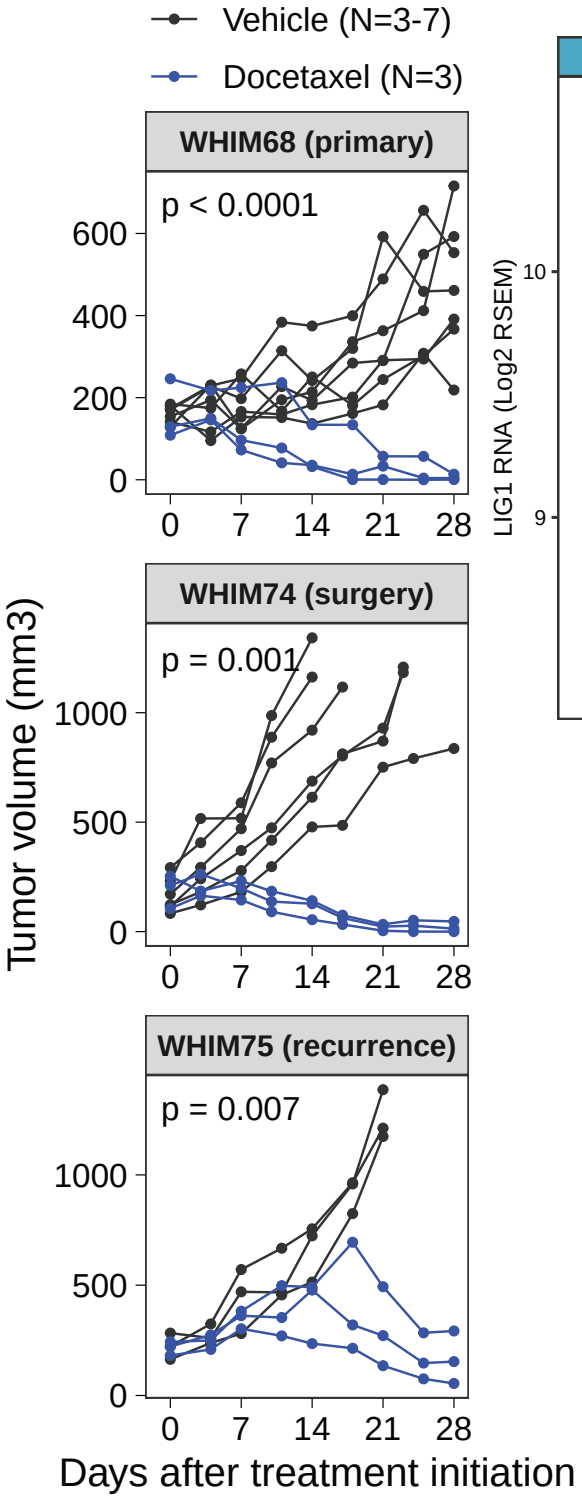
A


p-value=0.0001
Sample Type
Metastasis
Primary
LIG1: Copy Number Alterations
samples (%)
0
10
20
30
40
50
60
70
80
90
100
INSERM (Metastatic, PLOS Med 2016)
Metabric (Primary, Nature 2012)
Deep Deletion
Shallow Deletion
Diploid
Gain
Amplification

B


WHIM 68
WHIM 74
WHIM 75
WHIM 68
WHIM 74
WHIM 75
LIG1
130 kDa
1.00
0.48
0.28
POLD1
124 kDa
1.00
0.36
0.31
XRCC1
69 kDa
1.00
0.46
0.65
GAPDH
37 kDa

C


Vehicle (N=3-7)
Docetaxel (N=3)
WHIM68 (primary)
p < 0.0001
WHIM74 (surgery)
p = 0.001
WHIM75 (recurrence)
p = 0.007
Tumor volume (mm3)
0
200
400
600
500
1000
0
7
14
21
28
Days after treatment initiation

D

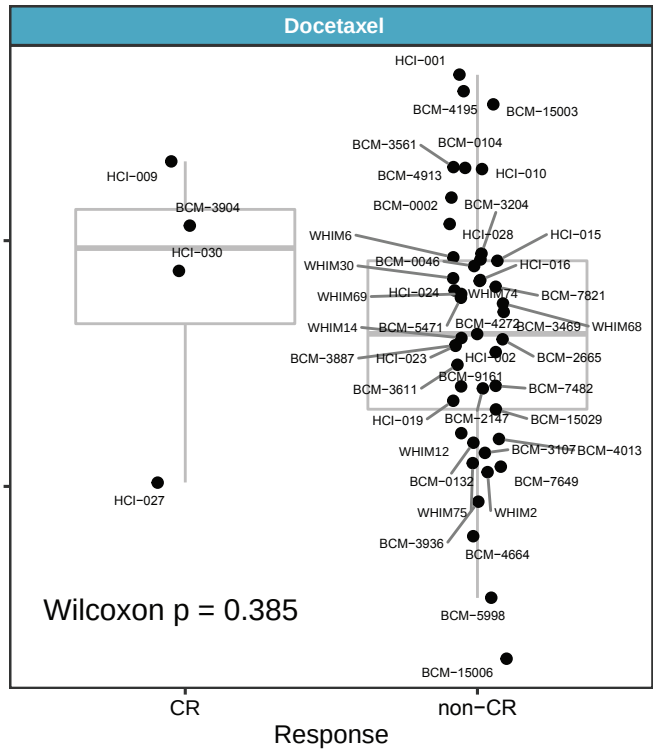
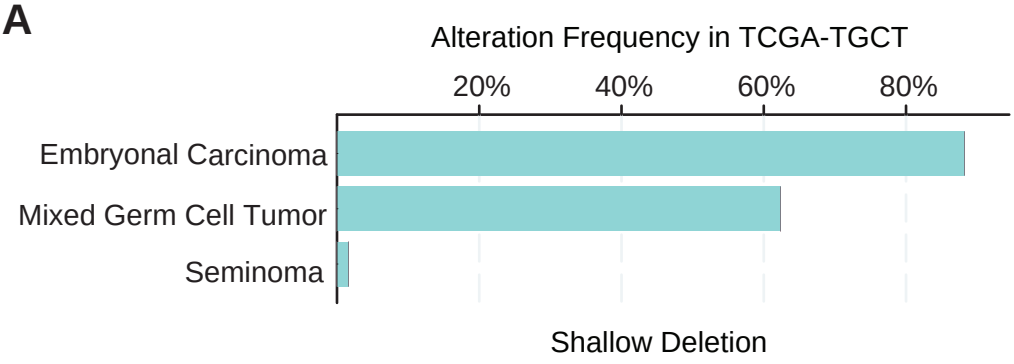

Docetaxel
LIG1 RNA (Log2 RSEM)
10
9
HCI-001
BCM-4195
BCM-15003
BCM-3561
BCM-0104
HCI-009
BCM-4913
HCI-010
BCM-3904
BCM-0002
BCM-3204
HCI-030
WHIM6
HCI-028
HCI-015
WHIM30
BCM-0046
HCI-016
WHIM69
HCI-024
WHIM74
BCM-7821
WHIM14
BCM-5471
BCM-4272
BCM-3469
WHIM68
BCM-3887
HCI-023
HCI-002
BCM-2665
BCM-9161
BCM-3611
BCM-7482
HCI-019
BCM-2147
BCM-15029
WHIM12
BCM-3107
BCM-4013
BCM-0132
BCM-7649
HCI-027
WHIM75
WHIM2
BCM-3936
BCM-4664
BCM-5998
BCM-15006
Wilcoxon p = 0.385
CR
non−CR
Response

Fig S7:*LIG1* loss across TCGA Pan-Cancer cohort.

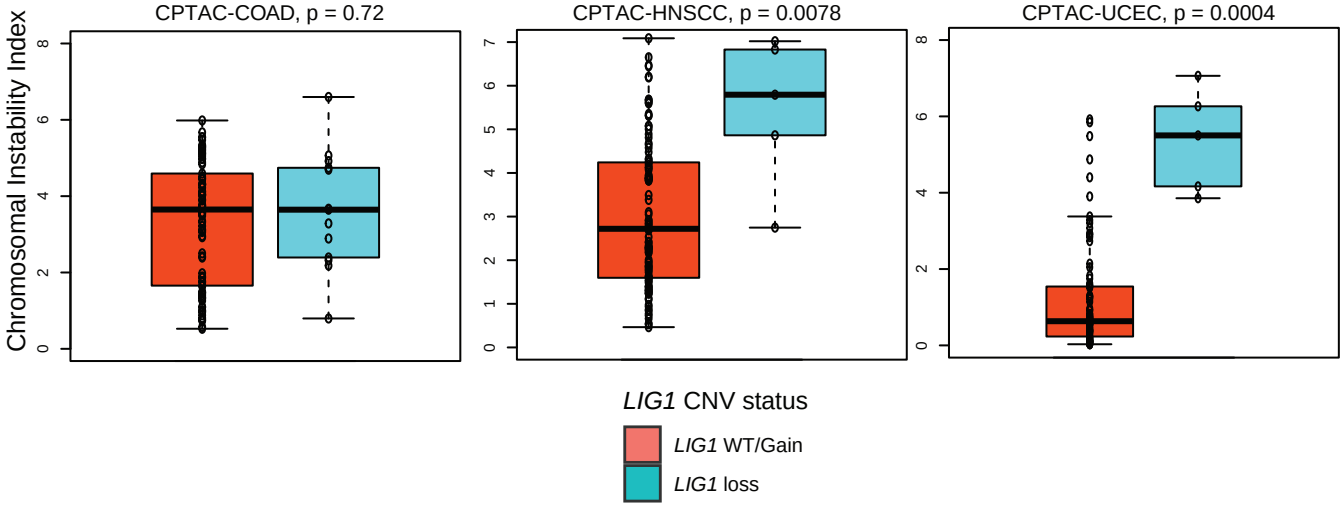
A) LIG1 single copy loss (shallow deletion, GISTIC <= -1) frequency in Testicular Germ Cell Tumors (TGCT) subtypes. The TCGA-TGCT dataset was used for this analysis.

B) Chromosomal Instability (CIN) scores in *LIG*1 WT/gain and loss in 3 CPTAC cancer cohorts. HSCC: Head and Neck Squamous Cell Carcinoma (PMID: 33417831) and UCEC: Endometrial Carcinoma (PMID: 32059776), and COAD: Colon Cancer PMID: 31031003). Tumors with LIG1 loss had significantly (Wilcoxon test) higher levels of CIN scores than tumors without LIG1 loss. Boxplots show the interquartile range (IQR) for the Chromosomal Instability index (CIN) for each group with median marked in center. Whiskers indicate 1.5x IQR. P-values are from Wilcoxon rank sum tests comparing *LIG1* loss tumors (GISTIC = -1) to tumors with GISTIC = 0-2.

A


Alteration Frequency in TCGA-TGCT
20%
40%
60%
80%
Embryonal Carcinoma
Mixed Germ Cell Tumor
Seminoma
Shallow Deletion

B


CPTAC-COAD, p = 0.72
CPTAC-HNSCC, p = 0.0078
CPTAC-UCEC, p = 0.0004
Chromosomal Instability Index
LIG1 CNV status
LIG1 WT/Gain
LIG1 loss

Supplementary Table legends

Table S1: Sample metadata

Sample specific metadata for the core needle biopsies with high tumor content greater than 45%. Samples that are part of the proteomics tandem mass tag (TMT) based experimental design with tumor content < 45% have been excluded from this table.

Table S2: Genomic datasets

A. Sample-wise somatic mutations in Mutation Association (MAF) format

B. Copy number log2 ratios

C. GISTIC thresholded absolute copy number

D. Upper quartile normalized RSEM data median centered by gene.

Table S3: Proteomics datasets

A. Median-MAD normalized log2-transformed protein data matrix (sample/CR ratios)

B. Gene-centric global proteome data matrix derived from Table S3A. Protein-level data were aggregated to gene-level using the sample/CR ratio assigned to the dominant isoform per gene symbol (similar to the razor-peptide approach).

C. Median-MAD normalized log2-transformed phosphoproteomic data matrix (sample/CR ratios)

D. Gene-centric phosphoproteomic data matrix derived from Table S3C. Phosphosite-level sample/CR ratios were aggregated to gene-level using mean ratio.

Table S4: Differential analysis results

A. Differential mRNA statistics (paired wilcoxon test) comparison on-treatment and baseline samples

B. Differential protein statistics (paired wilcoxon test) comparison of-treatment and baseline samples

C. Differential phospho-protein statistics (paired wilcoxon test) comparison of-treatment and baseline samples

D. Differential phospho site statistics (paired wilcoxon test) comparison of-treatment and baseline samples

Table S5: Differential expressed mrna, protein, phosphoprotein comparing non-pCR and pCR tumors.

A. Differential mRNA statistics (Wilcoxon test) comparison non-pCR and pCR high tumor content baseline samples

B. Differential protein statistics (Wilcoxon test) comparison non-pCR and pCR high tumor content baseline samples

C. Differential phospho-protein statistics (Wilcoxon test) comparison non-pCR and pCR high tumor

D. Differential phosphosite statistics (Wilcoxon test) comparison non-pCR and pCR high tumor content baseline samples content baseline samples

Table S6: Sample wise ssGSEA normalized enrichment score for Hallmark pathways.

A. ssGSEA scores derived from RNA, protein and phsophoprotein data for high tumor content (>=45%) baseline samples

B. PTM-SEA score derived from phosphosite level data

Table S7: LIG1 levels and treatment response in TNBC PDX models

Supplementary Data and Methods

Immunohistochemistry:

For immunohistochemistry (IHC) cut tissue sections (5mm) on charged glass slides were baked for 10-12 hours at 58oC in a dry slide incubator, deparaffinized in xylene and rehydrated via an ethanol step gradient. For CD3, heat-induced antigen retrieval steps were performed at pH 9.0. The primary antibody was incubated at room temperature for 1 hour [CD3 (polyclonal, Dako, 1:100)] followed by a standard chromogenic staining protocol with the Envision Polymer-HRP anti-mouse/3,30diaminobenzidine (DAB, Dako) process. Slides were counterstained in Harris hematoxylin. PDL1 staining was performed using the PD-L1 IHC 22C3 pharmDx kit per regulated protocol on the Dako Autolink-48 platform (SK006; Agilent). Pathology slide scoring was performed using established professional guidelines for TNBC, when appropriate. All immunohistochemistry results were evaluated against positive and negative tissue controls.

Genomic analysis:

Whole exome sequencing (WES): Tumor DNA was extracted from fresh-frozen biopsies and matched leukocyte germline DNA from blood samples. WES data was generated for 59 unique baseline DNA samples using the Illumina platform. For this, paired-end libraries were constructed as described previously(1) with the following modifications. Samples were barcoded at ligation step using Illumina unique dual barcodes adapters (Cat# 20022370) and were amplified 6-8 cycles using the Library Amplification Ready-mix containing KAPA HiFi DNA Polymerase (Kapa Biosystems, Inc). For capture enrichment, libraries were pooled in equimolar ratios in groups of 10 and were hybridized in solution to the HGSC VCRome 2.1 design (2). To this design, exome coverage across >3,500 clinically relevant genes that are previously <20X (~2.72Mb) was supplemented. Enriched libraries were sequenced on the NovaSeq 6000 instrument using the S4 reagent kit (300 cycles) to generate 2x150bp paired-end reads. For these 110134 samples, on average, 11.01 Gb of unique sequence data was generated with 97.3% of the bases in the exome design coverage to 20x read depth or greater.

RNA-Seq data: Transciptome data was generated for 60 samples in this study. For this, strand-specific, poly-A+ RNA-seq libraries for sequencing on the Illumina platform were prepared as

previously described (3). Briefly, poly-A+ mRNA was extracted from 1 μg total RNA, followed by fragmentation and first strand cDNA synthesis. The resultant cDNA was end-repaired, A-tailed and ligated with Illumina Dual barcode adapters. Libraries were sequenced on NovaSeq 6000 instruments using the S4 reagent kit (300 cycles) to generate 2x150bp paired-end reads. Between 59.96 and 112.62M total reads were generated for these 60 samples. The average strand-specificity and rRNA rate was 97.04% and 1.79% respectively. The transcripts for 22868 to 27856 genes were detected in these samples.

The paired-end reads were mapped to the human genome version GRCh38.d1.vd1 (From GDC) using STAR-2.7.1a. Gene expression estimation was performed using RSEM-1.3.1, and RSEM and FPKM values were upper-quartile normalized. Unless otherwise noted, gene median-centered log2-transformed RSEM values were used for the analyses presented here.

Somatic and copy number variant calling: Somatic variants were called using paired tumor and blood normal from WES data. Tools used for somatic variant calling are Strelka2, Mutect2, CARNAC, and Pindel (v 0.2.5b9). Variants reported by these tools were filtered using GATK VariantFiltration (v 3.8.0) with parameters window 35, cluster 3, FS > 30.0, and QD < 2.0. We kept SNVs called by any 2 callers among Strelka2, Mutect2 and indels called by any 2 callers among Mutect2, CARNAC, and Pindel. To merge SNVs and indels, we applied a 10X coverage cutoff for both tumor and normal sequence depth. We also filtered somatic SNVs and indels by a minimal variant allele frequency (VAF) of 0.02. Then annovar (v 04.16.2018) was used to annotate remaining variants. Somatic mutations were called comparing tumor DNA against matched blood normal DNA using WES data. Similarly, germline mutations were called by comparing normal WES against the reference genome. Hg19.UCSC.add_miR.140312.refgene was used to map the copy number information to genes. COSMIC mutational signature scores for every sample were estimated using deconstructSigs (4).

For somatic copy number alteration analysis, bam files were processed by the CopywriteR package (5) to derive log2 tumor-to-normal copy number ratios, and the circular binary segmentation (CBS) algorithm (6) implemented in the CopywriteR package was used for the copy number segmentation, with the default parameters. Chromosomal instability for each chromosome in each sample was inferred from the segmentation data using a weighted-sum approach in which the absolute values of the log2 ratios of all segments within a chromosome were weighted by the segment length and summed up (7). The genome-wide chromosome instability index (CIN) was

derived by adding up the instability scores for all 22 autosomes in each sample. Next, we used GISTIC2 (8) to retrieve gene-level copy number values and call significant copy number alterations in the cohort. A threshold of +/-0.3 was applied to log2 copy number ratio to identify gene-wise gain or loss of copy number, respectively. Each gene of every sample was assigned a thresholded copy number level that reflects the magnitude of its deletion or amplification. These are integer values ranging from −2 to 2, where 0 means no amplification or deletion of magnitude greater than the threshold parameters described above. Amplifications are represented by positive numbers: 1 means amplification above the amplification threshold; 2 means amplification larger than the arm level amplifications observed in the sample. Deletions are represented by negative numbers: −1 means deletion beyond the threshold; −2 means deletions greater than the minimum arm-level copy number observed in the sample.

For the Pancancer analysis GISTIC value +/- 2 exceed the high-level thresholds for amplifications/deep deletions, and those with +/- 1 exceed the low-level thresholds but not the high-level thresholds. The low-level thresholds are just the 'ampthresh' and 'delthresh' noise threshold input values to GISTIC (typically 0.1 or 0.3) and are the same for every thresholds.

Proteomics data generation and analysis:

Proteomic sample preparation: Samples were prepared for proteomic analysis as described in a previous microscaled proteogenomic study (9). Protein lysates in 8 M urea were reduced with 1 mM DTT for 45 min, then alkylated with iodoacetamide (IAA) for 45 min protected from light. Before digestion, urea was diluted to a final concentration of 2 M with 50 mM Tris-HCl pH 8.5. Protein lysates were then treated with endopeptidase LysC (Promega) at a 1:40 enzyme mass to BCA-estimated protein mass ratio, followed by overnight treatment with trypsin (Promega) at a 1:30 ratio. Both digestions were performed at 25C. To desalt the digestion mixture, peptides were acidified to 1% formic acid (FA), and purified with a 50 mg tC18 Sep-Pak cartridge (Waters). Peptides were eluted with 50% acetonitrile/0.1% FA. Peptide concentration was measured by Nanodrop (Thermo) using 280 nm absorbance. To evaluate peptide sample quality and digestion efficiency, 0.5 ug peptides from each sample were run on a nLC1200 coupled to Q-Exactive + LC-MS instrumentation (Thermo). The remainder of the eluted peptides were then snap-frozen and dried with a vacuum centrifuge.

Basic reverse phase fractionation and phosphoenrichment

For basic reverse phase fractionation, ~330 ug of peptides were dissolved in 500 uL of 5 mM ammonium formate and 5% acetonitrile using an offline Agilent 1260 LC with a 30 cm long, 2.1 mm inner diameter C18 column, running at 200 uL per minute. Peptides were separated across 72 fractions, which were then concatenated into 18 fractions. 2 ug peptides from each of the 18 fractions were set aside for whole proteome analysis, of which 0.67 ug was analyzed from each fraction. For phospho-enrichment, the 18 fractions were further concatenated to 6 fractions (~50 ug per fraction).

Phosphopeptides were enriched using Fe^{3+} immobilized metal affinity chromatography (IMAC) as previously described (9). Ni-NTA (Qiagen) beads were washed three times with HPLC grade water followed by incubation with 100 mM EDTA (Sigma) to strip nickel from the beads. Beads were washed three more times with HPLC grade water and incubated in 10 mM $FeCl_3$ (Sigma) for 30 min. Fe^{3+}-loaded beads were resuspended in a 1:1:1 solution of methanol, acetonitrile, and 0.01% acetic acid in water. Dried peptides were resuspended to a final volume of 500 uL with 50% acetonitrile and 0.1% trifluoroacetic acid (TFA) followed by 100% acetonitrile and 0.1% TFA for a final concentration of 80% acetonitrile. Each reconstituted fraction was added to 20 uL of 50% bead slurry and incubated while rotating end-over-end at room temperature for 30 min. Beads were then spun down and supernatant removed. The beads were transferred to a conditioned C18 stage-tip in 200 uL of 80% acetonitrile and 0.1% TFA. Phosphopeptides were eluted from the beads with 500 mM potassium phosphate, pH 7, onto the C18, washed with 1% formic acid, and eluted into an autosampler vial with 50% acetonitrile and 0.1% FA.

Proteomic data acquisition and processing

Proteome and phosphoproteome data acquisition was performed with a Proxeon nLC-1200 coupled to Thermo Lumos instrumentation. For proteomic analysis, peptides were run on a 110 min gradient with 86 min of effective gradient (6 to 30% buffer B containing 90% ACN and 0.1%FA). For phosphoproteomics analysis, two injections were run per fraction: a first injection over a 90 min gradient with 70 min of effective gradient (6 to 30% buffer B containing 90% ACN and 0.1% FA), and a second injection over a 140 min gradient with 120 min of effective gradient (6 to 35% buffer B containing 90% ACN and 0.1% FA). The acquisition parameters for the 110 min proteome and 90 min phosphoproteome methods are: MS1 resolution = 60,000, MS1 injection time = 50 s, MS2 resolution = 50,000, MS2 injection time = 105 s, AGC 6e4. For the 140 min

phosphoproteome injections, the MS2 injection time was increased to 250 ms, and the MS2 AGC decreased to 5e4. A cycle time of 2 s was used for all methods.

Raw files were searched against the human (clinical cores) or humanRefSeq protein databases complemented with 553 small-open reading frames (smORFs) and common contaminants (Human: RefSeq.20171003_Human_ucsc_hg38_cpdb_mito_259contamsnr_553smORFS.fasta), using Spectrum Mill (Broad Institute) as previously described in detail (9). Lysine was required to be fully TMT-labeled while N-termini were allowed to be under-labeled for TMT quantification using the "Full, Lys only" option. Carbamidomethylation of cysteines was set as a fixed modification, and N-terminal protein acetylation, oxidation of methionine (Met-ox), de-amidation of asparagine, hydroxylation of proline, TMT over-labeling of serine, threonine, and tyrosine, and cyclization of peptide N-terminal glutamine and carbamidomethylated cysteine to pyroglutamic acid (pyroGlu) and pyro-carbamidomethyl cysteine were set as variable modifications. For phosphoproteome analysis, phosphorylation of serine, threonine, and tyrosine were allowed as additional variable modifications, while de-amidation of asparagine and hydroxylation of proline were disabled. Trypsin Allow P was specified as the proteolytic enzyme with up to 4 missed cleavage sites allowed. For proteome analysis, the allowed precursor mass shift range was −18 to 326 Da. For phosphoproteome analysis, the range was −18 to 272 Da, to allow for up to 3 phosphorylations and 2 Met-ox per peptide. Precursor and product mass tolerances were set to ±20 ppm. For core biopsy analysis, peptide FDR limits were set to 0.8% for charge states 2-4 and 0.4% for charge states 5-6, and for PDX analysis peptide FDR limits were set to 0.6% for 2-4 and 0.3% for 5-6, employing a target-decoy approach using reversed protein sequences (PMID: 20013364). For PDX analyses, the subgroup-specific (SGS) option in Spectrum Mill was enabled as previously described (9) . This allowed better dissection of proteins of human and mouse origin. If specific evidence for both human and mouse peptides from an orthologous protein were observed, then peptides that cannot distinguish the two (shared) were ignored. However, the peptides shared between species were retained if there was specific evidence for only one of the species, thus yielding a protein group with a single subgroup attributed to only the single species consistent with the specific peptides.

PDX proteomics data generation and analysis

For the PDX experiment, cryopulverized PDX tumor tissues were lysed and digested as described above. 50ug peptides were dissolved in 200ul 50 mM HEPES, pH 8.5 and labeled with 400ug of

TMT reagent. TMT sample generation, basic reverse fractionation and proteomic analysis was performed identical to that of clinical core biopsies. Raw files were searched against the human and mouse (PDX samples) UniProt protein databases complemented with 553 small-open reading frames (smORFs) and common contaminants (Human and mouse: UniProt.human.mouse.20171228.RIsnrNF.553smORFs.264contams.fasta) using Spectrum Mill subgroup-specific (SGS) option. For PDX analyses, the subgroup-specific (SGS) option in Spectrum Mill was enabled as previously described (9). This allowed better dissection of proteins of human and mouse origin. If specific evidence for both human and mouse peptides from an orthologous protein were observed, then peptides that cannot distinguish the two (shared) were ignored. However, the peptides shared between species were retained if there was specific evidence for only one of the species, thus yielding a protein group with a single subgroup attributed to only the single species consistent with the specific peptides.

ory, immune inhibitory, and human leukocyte antigen (HLA) (10).

Immunoblotting

Fresh frozen WHIM68, WHIM74, and WHIM75 tumors were cryopulverized (Covaris CP02) then lysed in RIPA buffer (10 mM Tris-Cl [pH=8.0], 1 mM EDTA, 0.5 mM EGTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 140 mM NaCl, 10 mM PMSF, 1 ug/mL pepstatin, phosSTOP phosphatase inhibitor table [Roche], and cOmplete EDTA-free protease inhibitor tablet [Roche]) for 20 min. on ice. 30 µg of clarified lysates were heated at 90°C for 10 min. before loading on 4-12% SDS-PAGE gels (Invitrogen) and electroblotted onto PVDF membranes (Bio-Rad). Membranes were blocked in 5% milk for 1h hour at RT followed by incubation of primary antibodies at 4°C overnight: LIG1 (cat# 18051-1-AP, ProteinTech, 1:1000), POLD1 (cat# 15656-1-AP, ProteinTeech, 1:1000), XRCC1 (cat# ab134056, Abcam, 1:1000). GAPDH (cat# sc-47724, Santa Cruz Biotechnology, 1:4000) was used as a loading control. Proteins were visualized by incubation with Cytiva Amersham ECL Select Western Blot Detection reagent (Fisher Scientific) and images were captured with a ChemiDoc Imaging System (Bio-Rad).

Supplemental Method references:

1. Rokita JL, Rathi KS, Cardenas MF, Upton KA, Jayaseelan J, Cross KL, *et al.* Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design. Cell Rep **2019**;29(6):1675-89 e9 doi 10.1016/j.celrep.2019.09.071.
2. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, *et al.* Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol **2011**;12(7):R68 doi 10.1186/gb-2011-12-7-r68.
3. Peters TL, Kumar V, Polikepahad S, Lin FY, Sarabia SF, Liang Y, *et al.* BCOR-CCNB3 fusions are frequent in undifferentiated sarcomas of male children. Mod Pathol **2015**;28(4):575-86 doi 10.1038/modpathol.2014.139.
4. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol **2016**;17:31 doi 10.1186/s13059-016-0893-4.
5. Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, *et al.* CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol **2015**;16:49 doi 10.1186/s13059-015-0617-1.
6. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics **2007**;23(6):657-63 doi 10.1093/bioinformatics/btl646.
7. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, *et al.* Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell **2019**;177(4):1035-49 e19 doi 10.1016/j.cell.2019.03.030.
8. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol **2011**;12(4):R41 doi 10.1186/gb-2011-12-4-r41.
9. Satpathy S, Jaehnig EJ, Krug K, Kim BJ, Saltzman AB, Chan DW, *et al.* Microscaled proteogenomic methods for precision oncology. Nat Commun **2020**;11(1):532 doi 10.1038/s41467-020-14381-2.
10. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, *et al.* The Immune Landscape of Cancer. Immunity **2019**;51(2):411-2 doi 10.1016/j.immuni.2019.08.004.