|  |  |  |  |
| --- | --- | --- | --- |
| **R Package** | **Version** | **Use** | **Reference (URL included if published in paper)** |
| R | 4.1.0 | Computing language | R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. |
| Seurat | 4.0.2 | Single-cell (sc) object | Hao and Hao et al. Integrated analysis of multimodal single-cell data. bioRxiv (2020) [Seurat V4] <https://www.biorxiv.org/content/10.1101/2020.10.12.335331v1> |
| Revelio | 0.1.0 | Sc cell cycle annotation | <https://www.embopress.org/doi/full/10.15252/msb.20209946> |
| plyr | 1.8.6 | Mapvalues | Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL <http://www.jstatsoft.org/v40/i01/>. |
| viridis | 0.6.1 | Sc color palette | Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer (2021). Rvision - Colorblind-Friendly Color Maps for R. R package version 0.6.1. |
| HGNChelper | 0.8.1 | Check symbols for Revelio cell cycle markers | Levi Waldron and Markus Riester (2019). HGNChelper: Identify and Correct Invalid HGNC Human Gene Symbols and MGI Mouse Gene Symbols. R package version 0.8.1.  <https://f1000research.com/articles/9-1493> |
| ggplot2 | 3.3.5 | All plots | H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.  <https://www.springer.com/gp/book/9780387981413> |
| dplyr | 1.0.6 | Data manipulation | Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. |
| escape | 1.2.0 | Sc gene set enrichment | Nick Borcherding and Jared Andrews (2021). escape: Easy single cell analysis platform for enrichment. R package version 1.2.0. |
| GSEABase | 1.54.0 | Sc GSEA data structure | Martin Morgan, Seth Falcon and Robert Gentleman (2021). GSEABase: Gene set enrichment data structures and methods. R package version 1.54.0. |
| ggridges | 0.5.3 | Ridge plot | Claus O. Wilke (2021). ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.3. |
| GSVA | 1.40.0 | ssGSEA | Hänzelmann, S., Castelo, R. and Guinney, A. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics, 14:7, 2013.  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-7> |
| ComplexHeatmap | 2.8.0 | PDX heatmap | Gu, Z. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.  <https://academic.oup.com/bioinformatics/article/32/18/2847/1743594> |
| circlize | 0.4.12 | Used to make ComplexHeatmap annotation | Gu, Z. (2014) circlize implements and enhances circular visualization in R. Bioinformatics.  <https://academic.oup.com/bioinformatics/article/30/19/2811/2422259> |
| ggpubr | 0.4.0 | Stats on ggplot | Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. |
| reshape2 | 1.4.4 | Reformat dfs | Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL <http://www.jstatsoft.org/v21/i12/>. |
| pheatmap | 1.0.12 | All other heatmaps | Raivo Kolde (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. |
| RColorBrewer | 1.1.2 | RdBu palette | Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1.2. |

**Heatmap of mouse DDR protein abundance**

Protein abundance values of the genes curated in the “DNA double-strand break repair” pathway (R-HSA-5693532; Reactome Database ID Release 63) was queried in our previously published mouse mammary proteomic dataset5. The heat map created using the pheatmap R package (v 1.0.12) shows the z-scores of LFQ-adjusted IBAQ values in basal, luminal progenitor (LP), and luminal mature (LM) cell types from estrogen plus progesterone (E+P) samples.

**Human breast cancer cell-lines**

All 7 human breast cancer cell-lines used in this study were purchased from either ATCC or DSMZ. MDA-MB-231 (ATCC® HTB-26™), Hs 578T (ATCC® HTB-126™), EFM-192A (ACC 258), and EVSA-T (ACC 433) were cultured in DMEM media + 10% FBS. HCC1395 (ATCC® CRL-2324™), BT-549 (ATCC® HTB-122™), and HCC1187 (ATCC® CRL-2322™) were cultured in RPMI media + 10% FBS. All cell-lines were incubated at 37°C, 5% CO2.

**Human breast cancer cell-line engraftment**

Each human breast cancer cell-line in PBS was mixed with Matrigel® (Corning, 356231) in 1:1 v/v ratio. A total of 10 μl cell-Matrigel mixture containing 5×105 cells was injected directly into the right inguinal mammary gland of 6-7-week-old virgin female NSG mice using a Hamilton syringe. Xenograft tumours were monitored twice a week starting 7 days after engraftment. Tumour dimensions were measured with a Vernier caliper two times a week and tumour volume (mm3) was estimated by 0.5 × (minimum diameter in mm)2 × (maximum diameter in mm) from day 7 post-injection until the end of the study. Mice were sacrificed when humane endpoints were reached (tumour volume >1500 mm3, cumulative clinical score >8 or a max score for any animal condition, or limb paralysis).

***In vivo* drug testing**

1 g of olaparib (MedChemExpress, HY-10162) was dissolved in 10 mL of DMSO and stored at -80°C. A fresh aliquot of 100 mg/mL olaparib stock was diluted 1:10 in the vehicle solution (10% w/v 2-hydroxypropyl-β-cyclodextrin in PBS) at the time of injection. Once MDA-MD-231 and HCC1187 tumors reached a volume of 100-200 mm3, mice were randomized into two groups and were treated with either vehicle control (10% DMSO in the vehicle solution) or 100 mg/kg olaparib daily via intraperitoneal injections until they reached an endpoint as described above.

**Human proteomic analysis**

The abundance values of proteins in the three mammary epithelial populations: basal (BC), luminal progenitor (LP), and luminal mature (LM) was extracted from our previously published human mammary epithelial proteome6. The proteomic dataset contains log2-transformed iBAQ-adjusted LFQ values representing protein abundance which were adjusted for batch effects using the ComBat function in the ‘sva’ R package (v 3.30.1). Only samples from premenopausal patients were taken into account (n=6 for each BC, LP, LM cell types).

**Volcano plot and Pathway analysis on the human DDR proteins**

Of the 276 curated human DDR genes39, 127 proteins were detected in the mammary proteomic dataset based on matching by gene symbol. Of 127 DDR proteins, proteins were defined as highly expressed or enriched in one cell type if they met the fold-change (FC) and statistical significance cut-offs compared to the other two cell types. Enriched proteins had a log2FC > 0 and a p-value < 0.05 (paired t-test; p-value was adjusted for multiple testing via FDR) and were visualized in a volcano plot. The top 5 “upregulated” and “downregulated” proteins were based on rank values from the summation of p-value and fold-change ranks. The protein with the lowest p-value was given a rank of 1, while the highest earned the rank of r, which represents the number of total proteins (i.e. 127). For upregulated, the protein with the most positive log2FC value was given a rank of 1, while the least positive (log2FC > 0) value was given a rank of r. For downregulated, the protein with the most negative log2FC value was given a rank of 1, while the least negative (log2FC < 0) value was given a rank of r. The gene names of all enriched proteins were visually labelled in volcano plots. Pathway analysis of all enriched proteins/genes for each cell type queried KEGG 2019 terms using Enrichr (<https://amp.pharm.mssm.edu/Enrichr/>). The top 5 enriched pathways was determined based on the ‘combined’ score and were visualized in bar charts.

**Generation of global human and mouse lineage signatures**

The total or global mammary proteome consisted of 6034 (human) and 4695 (mouse) annotated proteins. Basal, luminal progenitor and luminal mature lineage signatures were acquired by searching for proteins enriched in one cell type compared to the other two cell types with a fold-change > 5 in human or fold-change > 3 in mouse and p < 0.05 in one-way ANOVA in conjunction with Tukey’s test.

**Enrichment of signatures in breast cancer subtypes**

Human and mouse lineage signatures in breast cancer expression profiles from METABRIC via single-sample Gene Set Expression Analysis (ssGSEA) using the ‘GSVA’ R package (v 1.30.0). The scores were categorized by PAM50 plus Claudin-low subtypes for each signature and assessed for significance using a one-way ANOVA and Tukey’s multiple comparisons test.

TCGA breast cancer RNA-seq and somatic mutation data were obtained following approval by the Data Access Committee (project #11689). The results published here are partly based upon data generated by TCGA managed by the NCI and NHGRI at <http://cancergenome.nih.gov>.

**Correlations to COSMIC somatic mutational signatures**

Thirty different somatic mutational signatures from COSMIC (Catalogue Of Somatic Mutations In Cancer) were defined using the mutSignatures package in R48. The signature scores were computed using the ssGSEA algorithm72 with standard parameters and using all genes included in each set. Spearman’s correlation coefficients and p-values were computed in R.

**Enrichment in human breast cancer cell-lines**

Enrichment of human or mouse lineage signatures in 50 human breast cancer cell-lines from the Genomics of Drug Sensitivity in Cancer database was determined using the ssGSEA algorithm. For mouse signatures, gene symbols were converted to human homologs using the ‘biomaRt’ R package (v 2.38.0). Top cell-lines were defined as having the greatest differences between basal and luminal progenitor signatures in ssGSEA enrichment scores.

**Correlation to breast cancer cell-line drug sensitivity screening**

Pearson correlations were performed to measure the association between enrichment scores for each cell type and IC50 values of 23 DDR-related drugs (targeting ‘DNA replication’ or ‘Genome integrity’) in human breast cancer cell-lines from the Genomics of Drug Sensitivity in Cancer portal49. Only the top ten cell-lines that were enriched for each basal and luminal progenitor signature were considered. Pearson’s correlation coefficients and p-values were computed in R. The IC50 data was from: <https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html>

**Acquisition/making of human breast cancer PDX (ask Arvind/Mitchell)**

**Enrichment of signatures in human breast cancer PDX**

Enrichment of human lineage signatures in the transcriptomes of 48 human breast cancer patient-derived xenograft models was determined using the ssGSEA algorithm. The heatmap depicting enrichment scores with annotations for response to talazorapib (ie. angle) and PAM50 subtype was created using the R packages ComplexHeatmap (v 2.8.0) and circlize (v 0.4.12).

Linear models were fitted using the lm function and QR factorization method in R. Pearson correlation was used to measure the linear dependence between angle and ssGSEA values for each lineage. ~~The graphs were created using the ggplot2 R package’s geom\_smooth layer (v 3.3.5).~~ Point shapes depicted PAM50 subtypes: basal, Normal and other (HER2, Luminal A, Luminal B, unknown).

Qualitative drug responses were determined for each PDX model replicate and scored according to the RECIST drug response metric system. Replicates that were recorded as complete response (CR) and partial response (PR) were categorized as responders (“Yes”), while stable disease (SD) and progressive disease (SD) were considered non-responders (“No”). Next, a boxplot was created using ggplot2 that showed basal cell lineage enrichment in PDX models and two-tailed Wilcoxon signed-rank test was performed using the ggpubr R package (v 0.4.0) to determine statistical significance between responders and non-responders.

**Annotation of cell cycle phase in human single-cell RNAseq data**

Our previously published human mammary single-cell transcriptomic dataset was interrogated (number of cells: total 6808; basal cell (BC) 1272; luminal progenitor (LP) 3364; mature luminal (ML) 2171) [REF: Nat Met paper]. Further processing was performed in the R statistical environment (v 4.1.0). Cell cycle annotations per cell using the counts matrix were determined using the Revelio R package (0.1.0). ~~The counts matrix from the Seurat object was fed into the createRevelioObject function and then annotated using the getCellCyclePhaseAssignInformation function.~~ The ccPhase parameter described discrete cell cycle phases for each cell (number of cells: M.G1 803; G1.S 1111; S 1206; G2 1192; G2.M 495). The number of cells per cell cycle phase was used to create a bar graph using ggplot2 package (v 3.3.5).

The Seurat R package (v 4.0.2) was used to create uniform manifold approximation and projection (UMAPs) with reduction umapharmony. The FeaturePlot function was used to project GMNN expression over the UMAP.

**Enrichment of DDR gene sets in human single-cell RNAseq data**

Gene lists for HR and NHEJ were derived from Reference: Supp table from "Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas". The RAD51 gene set consists of "RAD51B", "RAD51C", "RAD51D", the RAD51 genes found in the single cell dataset. Gene lists were converted to GeneSet objects using the GSEABase R package (1.54.0). The cells were stratified into 2 groups based on Revelio cell cycle phase annotation S/G2/M (ccPhase: S, G2.M, G2) and M/G1/S (ccPhase: M.G1, G1.S). Enrichment of the gene sets was calculated using the enrichIt function in escape R package (v 1.2.0). Ridge plots of enrichment scores were created using the ridgeEnrichment and geom\_density\_ridges from the escape and ggridges (v 0.5.3) R packages respectively.

The Seurat object was subsetted to GMNN > 0 (Number of cells: Total 1116; BC 125; LP 525; ML 466). Statistical significance of enrichment of HR and RAD51 sets between the three mammary epithelial cell populations was determined one-way ANOVA in conjunction with Tukey’s test.