

Drug combination prediction for cancer treatment using disease-specific drug response profiles and single-cell transcriptional signatures

Daniel Osorio , Parastoo Shahrouzi, Xavier Tekpli, Vessela N Kristensen, Marieke L Kuijjer 

Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway • Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway • Department of Clinical Molecular Biology and Laboratory Science (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, Norway • Department of Biochemistry and Developmental Biology, University of Helsinki, Helsinki, Finland • iCAN Flagship in Digital Precision Cancer Medicine, University of Helsinki, Helsinki, Finland

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

eLife Assessment

This research addresses an **important** and timely topic in cancer treatment, as the authors present a novel computational tool, 'retriever,' which has the potential to revolutionize personalized cancer treatment strategies by predicting effective drug combinations for triple-negative breast cancer. The strength of the evidence presented is **solid**, as evidenced by the systematic testing of 152 drug response profiles and 11,476 drug combinations.

<https://doi.org/10.7554/eLife.102442.1.sa3>

Abstract

Developing novel cancer treatments is a challenging task that can benefit from computational techniques matching transcriptional signatures to large-scale drug response data. Here, we present '*retriever*,' a tool that extracts robust disease-specific transcriptional drug response profiles based on cellular response profiles to hundreds of compounds from the LINCS-L1000 project. We used *retriever* to extract transcriptional drug response signatures of triple-negative breast cancer (TNBC) cell lines and combined these with a single-cell RNA-seq breast cancer atlas to predict drug combinations that antagonize TNBC-specific disease signatures. After systematically testing 152 drug response profiles and 11,476 drug combinations, we identified the combination of kinase inhibitors QL-XII-47 and GSK-690693 as the topmost promising candidate for TNBC treatment. Our new computational approach allows the identification of drugs and drug combinations targeting specific tumor cell types and subpopulations in individual patients. It is, therefore, highly suitable for the development of new personalized cancer treatment strategies.

Introduction

Developing new drugs and pharmacological regimens to treat complex diseases such as cancer is an expensive and challenging task (1). Several computational techniques that use structural interactions (2), transcriptional signatures (3–5), biological networks perturbations (6, 7), and data mining (8) are available to aid in the discovery of new treatment regimens which could improve patient management (9). Methods based on transcriptional signatures use the observed changes in gene expression profiles between samples from patients affected with the disease under study compared to samples from control subjects, and match these to the response profiles of cell lines that act as surrogates for the disease to different compounds. These data can be combined to identify compounds that may antagonize the disease's transcriptional changes, returning it to a more healthy-like state (5). Finding specific transcriptional response profiles to drug candidates, as well as robust disease-associated transcriptional alterations, is therefore crucial for such approaches to produce reliable predictions.

To obtain precise disease-specific transcriptional signatures that can be targeted by pharmacological compounds, one needs to account for transcriptional variability between patients by including samples from multiple individuals (10). This allows for the identification of a target set of genes that is constitutively expressed with the same magnitude and pattern across patients/cells. Targeting such a consistently expressed set of genes is thought to increase the effectiveness and safety of the treatment (11). Transcriptional signatures are measurable at a variety of resolutions, ranging from low resolution, as in the complex mixture of cells present within tissues, to high resolution, such as single-cells (12). The Cancer Genome Atlas (TCGA) project provides RNA-sequencing (RNA-seq) data for tumors and adjacent tissues (13). While such tissue-level data have been useful in the identification of cancer subtypes and prognostic signatures, the identification of accurate transcriptional signatures between disease and healthy states has been hindered by the wide heterogeneity of cell types found in tumor tissues (14).

This issue can be overcome by using single-cell RNA-seq data (15). A drawback of single-cell RNA-seq experiments is that, due to the higher cost of profiling of single cells, often fewer biological replicates are available. However, by combining multiple experiments, sample sizes of cells with a desired molecular phenotype can be increased, improving characterization of the transcriptional changes observed in disease (16). In addition, the construction of pseudo-bulk profiles (the sum of the expression values of a gene across all cells obtained from the same individual) from single-cell data makes it possible to account for gene expression variability among patients (17, 18).

The LINCS-L1000 project published transcriptional profiles of several cell lines, treated with hundreds of compounds at various concentrations and time points of drug exposure (4). These profiles have previously been used to identify potential drugs that can be repurposed to treat a variety of diseases, including cancer (3). Additionally, the project provides an interactive portal in which users can interrogate whether an up or down-regulated gene set of interest overlaps with transcriptional drug response signatures. The portal then returns a ranked list of compounds that are likely to have an inverse effect on disease-associated gene expression levels (19). However, these predictions are not based on robust tissue or disease-specific transcriptional profiles and may therefore over or underestimate the potential effect of drugs on specific diseases. Additionally, the LINCS-L1000 web portal returns an exhaustive list of independent experiments with matching inverse patterns, rendering it difficult to identify compounds that induce stable transcriptional responses in cell lines derived from the same disease across multiple time points and drug concentrations. Thus, being able to extract robust disease-specific transcriptional drug response signatures that are consistent at different time points, drug concentration, or cell line from the

profiles provided by the LINCS-L1000 project would significantly improve drug prioritization and accelerate the identification of new pharmacological options for personalized treatment of cancer (20 [↗](#)).

Once both the disease profile and the disease-specific response to multiple compounds are available, rank-based correlation analysis can be used to quantitatively identify compounds that can revert the transcriptional changes that distinguish diseased samples from healthy ones (5 [↗](#), 9 [↗](#)). Nonetheless, monotherapy in cancer is highly susceptible to the development of resistance following an initial response to treatment (21 [↗](#)). Combination therapy, or the simultaneous administration of multiple drugs to treat a disease, has evolved into the standard pharmacological regimen for treating complex diseases such as cancer. Combination therapy prevent tumor evolution and help inhibit the development of drug resistance in cancer, thereby improving patient survival (22 [↗](#)).

The *in silico* prediction of responses to drug combinations is an active topic of research in computational biology (23 [↗](#)–25 [↗](#)). Recently, Pickering (2021) found that the transcriptional response signatures to 856, 086 unique two-drug combinations could be predicted based on the 1, 309 drug response profiles to compounds tested by the Connectivity Map Build 2 project (parent of the LINCS-L1000 project) (26 [↗](#)). After analyzing 148 independent studies involving 257 treatment combinations from the Gene Expression Omnibus (GEO) database, it was found that averaging the expression profiles of individual treatments provides 78.96 percent accuracy in predicting the direction of differential expression for the combined treatment (27 [↗](#)). However, different from the Connectivity Map project—where the generation of combinatorial response profiles is possible thanks to the availability of single response profiles for each individual compound (28 [↗](#))—the LINCS-L1000 project does not provide robust single drug response profiles, making the generation of meaningful combinatorial profiles not yet possible.

Here, we present *retriever*, a tool that uses correlation analysis and hierarchical collapsing to extract robust disease-specific transcriptional drug response signature profiles that are consistent across time, concentration, and cell line, from data provided by the LINCS-L1000 project. We integrated these transcriptional drug response profiles with a single-cell RNA-seq signature representing the transcriptional changes observed in triple-negative breast cancer cells compared to healthy breast epithelial cells. We used these two signatures—the transcriptional changes induced by the disease and the cellular responses to drugs—to prioritize drugs and to predict novel drug combinations suitable for the treatment of triple-negative breast cancer. The transcriptional changes associated with TNBC were computed from a single-cell breast cancer RNA-seq atlas built by us after compiling singlecell RNA-seq data obtained from 36 publicly available healthy breast and breast cancer samples. After systematically testing drug response profiles and combinations at both the population and individual patient levels, we identified QL-XII-47 and GSK690693 as the most promising combination of kinase inhibitors for treating TNBC. In addition to recommending drug combinations, the profiles returned by *retriever* allow for the characterization of possible mechanisms of action of the identified compound(s) to reverse the disease's transcriptional profile towards a healthy-like state.

Methods

Single-cell RNA-seq datasets

We collected publicly available single-cell RNA-seq count matrices for healthy breast tissues and breast cancer samples from multiple sources (see Data Availability). Each sample's gene identifiers (IDs) were translated into current gene symbols using the dictionary of gene ID synonyms provided by ENSEMBL (29 [↗](#)). Datasets were loaded into R and integrated into a single 'Seurat' object (30 [↗](#)). Data were then subjected to quality control keeping only cells with a library size of

at least 1,000 counts and within the 95 percent confidence interval of the prediction of the mitochondrial content ratio and detected genes in proportion to the cell's library size. We also removed all cells that had mitochondrial proportions greater than 10% (31). We then normalized, scaled, and reduced the dimensionality of the data through Principal Component Analysis (PCA) using the default functions and parameters included in Seurat. Data integration was performed using Harmony (32). UMAP projections of the data were extracted using the top 20 dimensions returned by Harmony (33). Cell clustering was performed using the functions included in Seurat for this purpose, with a resolution of 0.01, using UMAP embeddings as the source for the nearest neighbor network construction.

Cell type assignation

Assignation of cell types was performed based on the expression of the marker genes reported by Wu et al. (34, 35). Epithelial cell subtypes were assigned using the Nebulosa package by querying cells expressing *ESR1*, *PGR*, and *ERBB2* receptors (36).

Single-cell RNA-seq differential expression analysis

Transcriptional signatures of triple-negative breast cancer epithelial cells compared to healthy epithelial cells were quantified by differential expression using MAST (37). We compared (Supplementary Fig. S1 and Fig. S2) all epithelial cells from cancer patients (5,730 cells, 31,73%) against those from healthy donors (12,323 cells, 68.26%). In addition, we compared the cluster of patient-derived epithelial cells depleted of hormone and growth factor receptor expression (enrichment for *ESR1*[−], *PGR*[−], and *ERBB2*[−] cells—2,998 cells, 16.6%) against all healthy epithelial cells (12,323 cells, 68.3%), against the cluster of epithelial cells depleted of hormone and growth factor expression (6,117 cells, 33.9%) derived from healthy individuals, and against the cluster of epithelial cells enriched for triple-positive cells (*ESR1*⁺, *PGR*⁺, and *ERBB2*⁺) from healthy individuals (6,206 cells, 34.4%). We selected the comparison of the cluster enriched with triple-negative cancer cells with the cluster enriched for triplepositive healthy epithelial cells as the disease-specific transcriptional profile, as this profile showed the highest Spearman Correlation Coefficient (ρ) with tissue-level fold-changes computed on TCGA breast cancer data.

Pseudo-bulk analyses

Pseudo-bulk profiles were computed by summing up all expression values for the same gene across cells of the same individual. We applied this procedure to the triple-negative epithelial cells derived from patients against the triple-positive epithelial cells from healthy individuals. The computed pseudo-bulk profiles from diseased and healthy individuals were then compared using DESeq2 (38).

TCGA breast cancer dataset analysis

Upper-quantile FPKM-transformed RNA-seq data at the tissue level as well as the associated metadata were downloaded from the TCGA breast cancer (TCGA-BRCA) project using the 'TCGAbiolinks' Bioconductor package (39). Samples labeled as 'healthy' and 'basal' were collected and differential expression between these groups was evaluated using DESeq2 (38).

Comparisons between data-types

The $\log_2$ fold-changes of expression between cancer and control samples at the single-cell level, pseudo-bulk level and tissue level were used to validate the computed signatures describing the transcriptional changes in triple-negative breast cancer compared to healthy epithelial cells using the Spearman correlation coefficient (40).

Application of *retriever* to triple-negative breast cancer

The transcriptional response profiles of triple-negative breast cancer cell lines that were treated with different concentrations and treatment lengths of different compounds available in the LINCS-L1000 project were collected using the ‘ccdata’ Bioconductor package (28 [4](#)). The LINCS-L1000 project provides these profiles for three different TNBC cell lines (BT20, HS578T, and MDAMB231). We used these data to compute the diseasespecific transcriptional response profiles.

Generation of drug combination transcriptional response profiles

All possible two-drug combinations were generated by averaging the computed disease-specific transcriptional response profiles.

Drug repurposing analysis

Using the set of overlapping genes between the LINCS-L1000 project and differential expression analysis from single-cell RNA-seq data, we performed Spearman correlation analyses with independent disease-specific drug response profiles and drug combination profiles. We also calculated 95% confidence intervals, p-values, and falsediscovery rates (FDR) (41 [4](#)). Compounds were then ranked based on their correlation coefficients, from negative to positive. For individual sample analysis, we compared the cells from each sample to a consistent set of control cells and performed Spearman correlation analyses with both the independent disease-specific drug response profiles and the drug combination profiles.

Mechanisms of action prediction

To identify possible mechanisms of action for drugs or drug combinations, we took the disease-specific transcriptional response profiles and used these as input for GSEA, with the MSigDB Hallmarks gene set as reference to compute the enrichment of pathways as well as the directionality of the effect (42 [4](#)). Pathways with $FDR < 0.05$ were considered significant.

Cell Lines and Cell Culture

Human breast adenocarcinoma cell lines DU4475 and BT20 were purchased from the American Type Culture Collection (ATCC), while CAL120 was obtained from Leibniz-Institut DSMadeno Z-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH. In all cases, the provider supplied an authentication certificate. To ensure the integrity of the cell lines, routine mycoplasma testing was conducted using the Lonza MycoAlert™ kit (LT07-418). None of the cell lines used in this study appeared in the database of commonly misidentified cell lines maintained by the International Cell Line Authentication Committee and NCBI Biosample. The cell lines were not passaged more than 7 times, with the highest passage used being passage 14. DU4475 cell line was cultured in RPMI-1640 media (ThermoFisher Cat:A1049101) supplemented with 10% Fetal Bovine Serum (FBS; ThermoFisher Cat:10500064) and 1% penicillin–streptomycin (ThermoFisher; 10,000 U/mL). CAL120 cell line was maintained in DMEMHigh glucose media (ThermoFisher Cat:11965084) with the same supplements, while BT20 cell line was cultured in EMEM (ThermoFisher Cat:31095029) media with identical supplementation.

Cell Viability Assays

BT20 and CAL120 cells were initially seeded at a density of 4000 and 2500 cells/well, respectively, in a 96-well plate and allowed to incubate overnight. After 24 hours, the culture media was replaced with media containing predetermined concentrations of selected drugs, namely BMX/BTK Inhibitor II, QL-XII-47 (CAS 1469988-75-7, Sigma Cat:5307580001), and GSK-690693 (Sigma Cat:SML04285MG), and incubated for an additional 72 hours. For the suspended cell line DU4475, cells were not counted, and drugs were added from the initiation of culturing. Cell viability for the drug assay was evaluated using the ATP-based CellTiter-Glo® 3D Cell Viability Assay kit (Promega

Cat:7570). The luminescence signal of each well was measured using a bottom-reading luminescent plate reader (SpectraMax M5 Multimode microplate reader). In each independent experiment, a minimum of three technical replicates were employed, and at least three independent experiments were conducted.

Results

The *retriever* algorithm

The *retriever* algorithm extracts disease-specific drug response profiles using three steps ([Figure 1](#)). The first step summarizes the cellular responses at different time points after the application of the drug, the second step summarizes the responses at different concentrations, and the third step summarizes the responses across different cell lines. This provides robust, disease-specific transcriptional response profiles based on the responses observed in all cell lines used as surrogates of a specific disease.

In the first step, *retriever* takes the response profiles of a given cell line to the same compound under the same concentration, but at different time points, and averages these. Then, the descriptive power of the extracted profile to represent the transcriptional response to a drug at a given concentration in the cell line, consistent across time, is evaluated using Spearman's correlation coefficient.

The averaged profile is returned if the correlation with the computed profile is larger than a user-defined threshold (default $\rho = 0.6$). This threshold can take values between 0 and 1, representing the percentage of agreement between the cellular responses to the same compound. If the threshold is too close to one, very few averaged profiles will be returned due to strong filtering, and if it is defined to close to zero, very noisy profiles (similar to simply averaging all the available profiles) will be returned. The original drug response profiles that do not reach this threshold are removed. The averaged profile is then recomputed using all response profiles that reached the threshold. This procedure ensures the removal of aberrant or insufficient cellular responses. Only averaged signatures of at least two profiles are used in the second step.

In the second step, *retriever* takes the stable time-consistent signature profiles of a compound at a particular concentration in the same cell line. To summarize the response at different concentrations, it applies the same procedure described in the first step over the averaged profiles. The profiles returned by the second step are stable transcriptional response profiles that are consistent across time and drug concentration in a specific cell line.

The last step extracts disease-specific drug response profiles by again applying the procedure described in step 1 to the stable response profiles to the compound in all disease-specific cell lines available in the LINCS-L1000 project. The profiles returned by the third step are robust disease-specific response signatures representing transcriptional changes to a specific compound.

Single-cell RNA-seq atlas of breast samples

To showcase how *retriever* can be used to prioritize drugs, we applied it to single-cell data from breast cancer and healthy breast tissues.

For this, we compiled 36 publicly available single-cell RNA-seq count matrices from breast samples (26 diseased and 10 healthy). In total, we combined 109,097 cells into a single Seurat object, maintaining the sample of origin metadata. Following quality control (see Methods section), 77,384 cells were retained for further analysis ([Figure 2A](#)), of which 30,790 were derived from

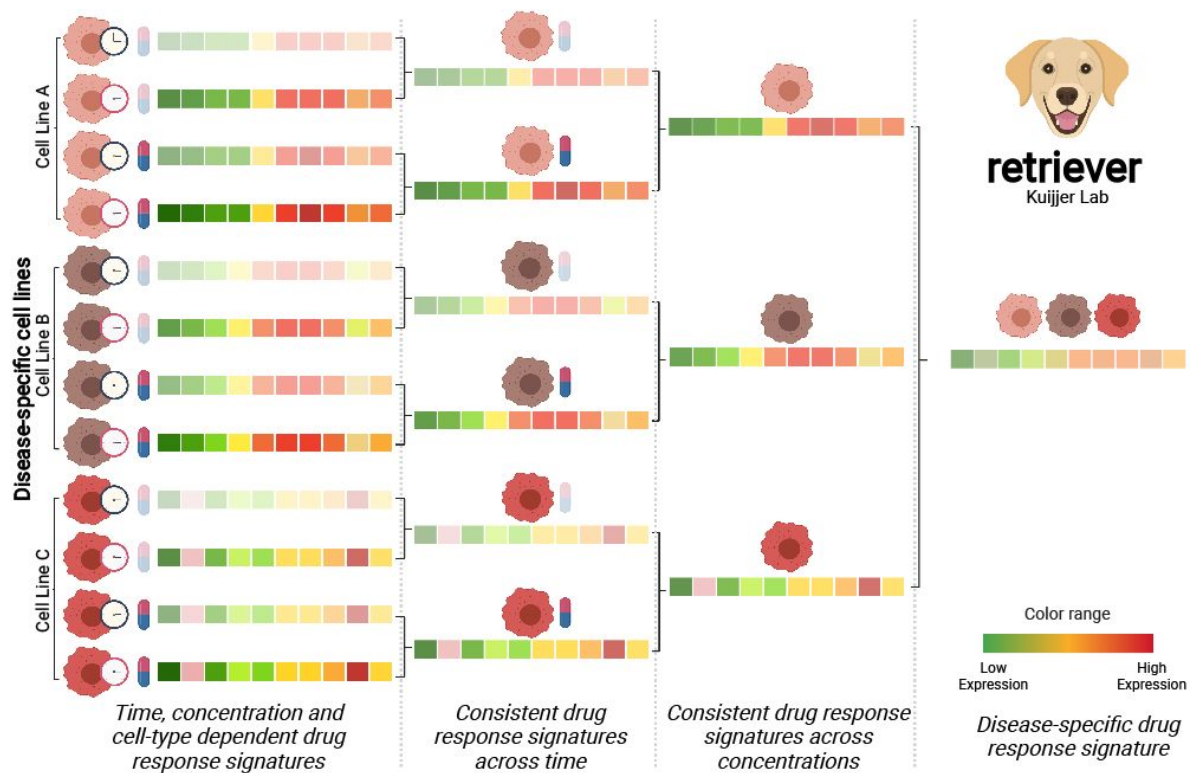


Fig. 1.

Overview of the *retriever* algorithm.

retriever generates disease-specific transcriptional drug response signatures by merging transcriptional signatures over time, concentration, and cell-type. These signatures can then be matched to single-cell or bulk expression profiles to predict drugs and drug combinations most likely to be effective in treating a disease.

healthy (left panel in [Figure 2B](#)) and 46, 594 from cancer samples (middle panel in [Figure 2B](#)). Cell types were assigned to the nine identified clusters in the low dimensional representation ([Figures 2C](#), [S3](#)) using markers reported by Wu et al. ([34](#), [35](#)).

Transcriptional signatures associated with triple-negative breast cancer at the single-cell level

To identify the transcriptional signature that best represents the changes associated with triple-negative breast cancer, we first identified the single cells with a triple-negative phenotype within the epithelial cluster of cells. To do so, we queried cells expressing ESR1, PGR, and ERBB2 receptors, which allowed us to assign a cluster of cells enriched for the expression of the three receptors (Right panel in [Figure 2B](#)); cells belonging to the cluster enriched for triplepositive cells will be further referred as “triple-positive-like” cells in the remainder of this work. While cells belonging to the other cluster will be referred to as “triple-negative-like” cells.

We compared the triple-negative-like breast cancer cells with different healthy epithelial cell subpopulations, including all epithelial cells, cells belonging to the triple-negative cluster, and cells belonging to the triple-positive cluster of epithelial cells. We then compared fold changes to those derived from the population level-data from TCGA (see Methods section). The most representative profile was the one comparing triple-negative-like epithelial cancer cells with healthy triple-positive-like epithelial cells. Note that, while approximately half of all healthy epithelial cells are assigned to the triple-positive cluster, these receptors are expressed in only 20% of these cells, and at lower levels compared to breast cancer cells ([Supplementary Fig. S1](#)).

As a result, we used MAST to perform differential expression analysis at the single cell level between triple-negative-like epithelial cancer cells ($n = 2, 998$) and healthy triple-positivelike epithelial cells ($n = 6, 206$), identifying 205 differentially expressed genes (106 upregulated and 99 downregulated) with absolute $\log_2$ fold-changes larger than 1 and false discovery rate lower than 0.05 ([Figure 3A](#)).

After testing using both the hypergeometric test through Enrichr ([43](#)) and gene set enrichment analysis (GSEA) using the Hallmark MSigDB signatures as reference gene sets ([42](#)), we found that the differentially expressed genes were associated with the activation of *Oxidative Phosphorylation Pathway*, *Interferon Alpha Response*, as well as *Myc Targets*, and the downregulation of *TNF α Signaling via NF- κ B*, *Estrogen Response*, *UV Response Up*, *Apoptosis*, *Hypoxia*, *Unfolded Protein Response*, *IL-6/JAK/STAT3 Signaling*, *Inflammatory Response*, and the *Androgen Response* pathway ([Figure 3C](#), [Supplementary Table S1](#)). Many of these pathways are known to be highly associated with TNBC molecular phenotype. For example, activation of both the oxidative phosphorylation pathway and of Myc targets is associated with a worse outcome of the disease ([44](#), [45](#)), and the crosstalk between the interferon alpha response pathway and NF- κ B is associated with drug resistance and tumor progression in this malignancy ([46](#)). The downregulation of the estrogen response pathway as well as apoptosis and hypoxia are markers of TNBC ([47](#)). In addition, downregulation of androgen and inflammatory responses is associated with worse prognosis and higher chemotherapy responsiveness respectively ([48](#), [49](#)).

We found 6, 149 genes expressed in both the single-cell RNAseq datasets and the bulk RNA-seq from the TCGA-BRCA project. Spearman correlation coefficients between the expression changes computed at the single-cell level (Left panel in [Figure 3A](#)), pseudo-bulk level (Middle panel in [Figure 3A](#)) and tissue level (Right panel in [Figure 3A](#)), revealed a monotonic positive association between these different levels ([Figure 3B](#)), supporting the descriptive power of the

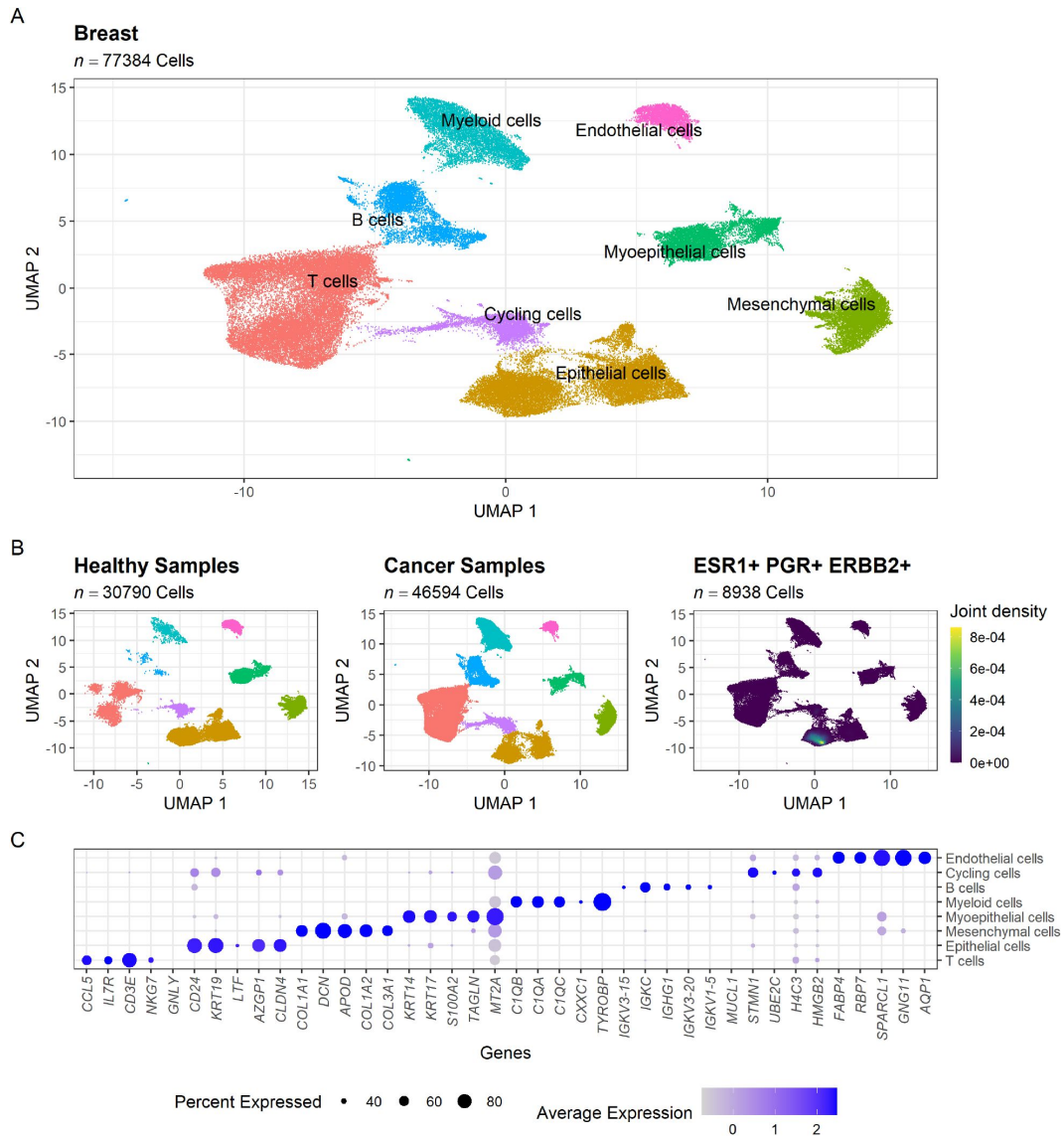


Fig. 2.

Single-cell atlas of breast samples.

(A) UMAP projection of the integrated 77,384 cells from 36 breast samples. **(B)** UMAP projection of the healthy (left) and cancer (center) cells, and visualization of triple-positive epithelial cells in the cancer samples. **(C)** Dotplot representing the normalized expression level and percentage of cells expressing the top five differentially expressed genes for each cell type.

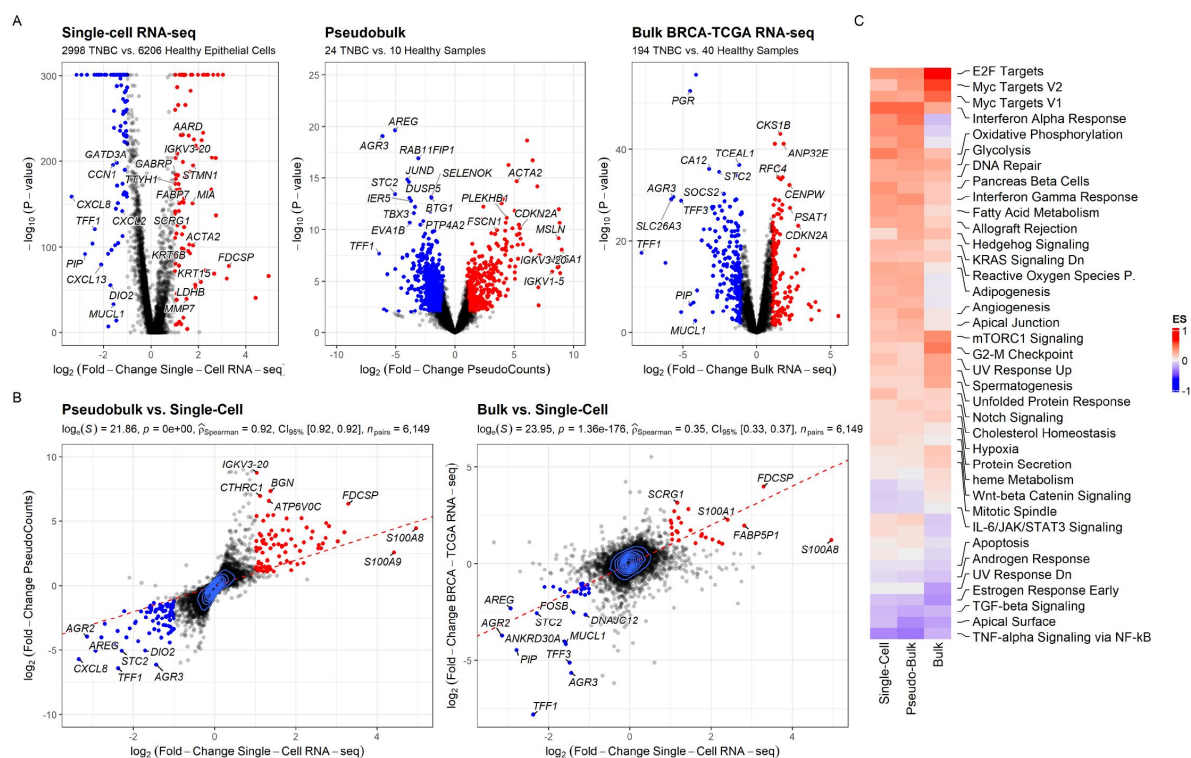


Fig. 3.

Transcriptional changes identified in TNBC cells.

(A) Volcano plots report the difference between TNBC and healthy samples, on the left based on single-cell RNA-seq count data computed using MAST, in the middle using the pseudo bulk measures in each single-cell RNA-seq sample, and on the right using the bulk RNA-seq data from the BRCA-TCGA project. Each dot represents a gene. Dots are color coded, in red if the $\log_2$ fold-change is larger than 1 and in blue if the $\log_2$ fold-change is smaller than -1. (B) Comparisons of the transcriptional changes associated with TNBC at the single-cell, pseudo bulk and tissue level. Each dot represents a gene. Dots are color coded as in Figure 2A. (C) Single-sample GSEA Enrichment Score (ES) for the transcriptional changes between TNBC and healthy cells at different levels of resolution. Labeled pathways are those that show same trend (positive or negative ES) in the three data types.

computed differential single-cell expression signature to describe transcriptional changes associated with TNBC, and ensuring that the selected cell type as well as the pseudo-bulk samples are high-resolution descriptors of the diseased tissues.

Applying *retriever* to extract TNBC-specific transcriptional drug-response signatures

We collected 4, 899 response profiles measured in TNBC cell lines from the LINCSL1000 project using the ‘ccdata’ package available in Bioconductor (28 [↗](#)). These profiles correspond to the expression change of 1000 genes in response to 205 compounds on average at four different concentrations (0.08 μ M, 0.4 μ M, 2 μ M, and 10 μ M), and at two time points (6 and 24 hours), in three different TNBC cell lines (BT20, HS578T, and MDAMB231).

An illustration of the step-by-step process of constructing a generalized response profile to a compound across three TNBC cell lines is presented in **Figure 4** [↗](#) using the QL-XII-47 compound as a case example.

1. Computing time-consistent generalized response profiles

In the first step (**Figure 4A** [↗](#)), we computed time-consistent generalized response profiles. To achieve this, we took the response profile of the MDAMB231 cell line to QL-XII-47 under the same concentration at two different time points (6 and 24 hours) and averaged them. When we compared the transcriptomes of the cell line exposed at two different time points, we found little or no correlation (large boxes labeled in gray) for each concentration (0.08 μ M, 0.4 μ M, 2 μ M, and 10 μ M). However, we found that their averaged profile displays high predictive power (Spearman Correlation Coefficient (ρ) > 0.65 in all cases) of the cellular response to the compound, independent of the time of treatment. When present, we removed the profiles that did not correlate with the averaged profile above 0.6 (A total of 19.72% of the profiles). This procedure allows to identify concentrations in which the compounds induces a different, aberrant or insufficient cellular responses to the drug and also allows us to compute a single robust response profile to the compound.

2. Computing timeand concentration-consistent response profiles

In the second step (**Figure 4B** [↗](#)), we extract time and concentration-consistent responses. For this purpose, we averaged the time-consistent profiles computed in the first step. As before, we removed the profiles that did not correlate with the averaged profile above 0.6, such as the profile present in the first panel of **Figure 4B** [↗](#), which showed a different cellular response at the concentration (A total of 16.91% of the profiles computed in Step 1). We then recomputed the profile for the compound in the MDAMB231 cell line with all profiles that had correlation coefficients above the threshold.

3. Generating disease-specific drug response profiles

In the last step (**Figure 4C** [↗](#)), we processed the outputs of the second step to extract a generalized response profile to a compound across the three TNBC cell lines. To do this, we performed independently steps 1 and 2 for QL-XII-47 in the other two TNBC cell lines available in the LINCSL1000 project (BT20, HS578T) and averaged these profiles.

After summarizing the time points, concentrations, and cell lines for all the 205 compounds tested in TNBC cell lines, we ended with robust transcriptional response profiles of 152 (74.14%) compounds (**Supplementary Table 1** [↗](#)) that are specific for TNBC. The remaining 53 compounds were removed, as they did not show a stable response across time points, concentrations, and cell lines.

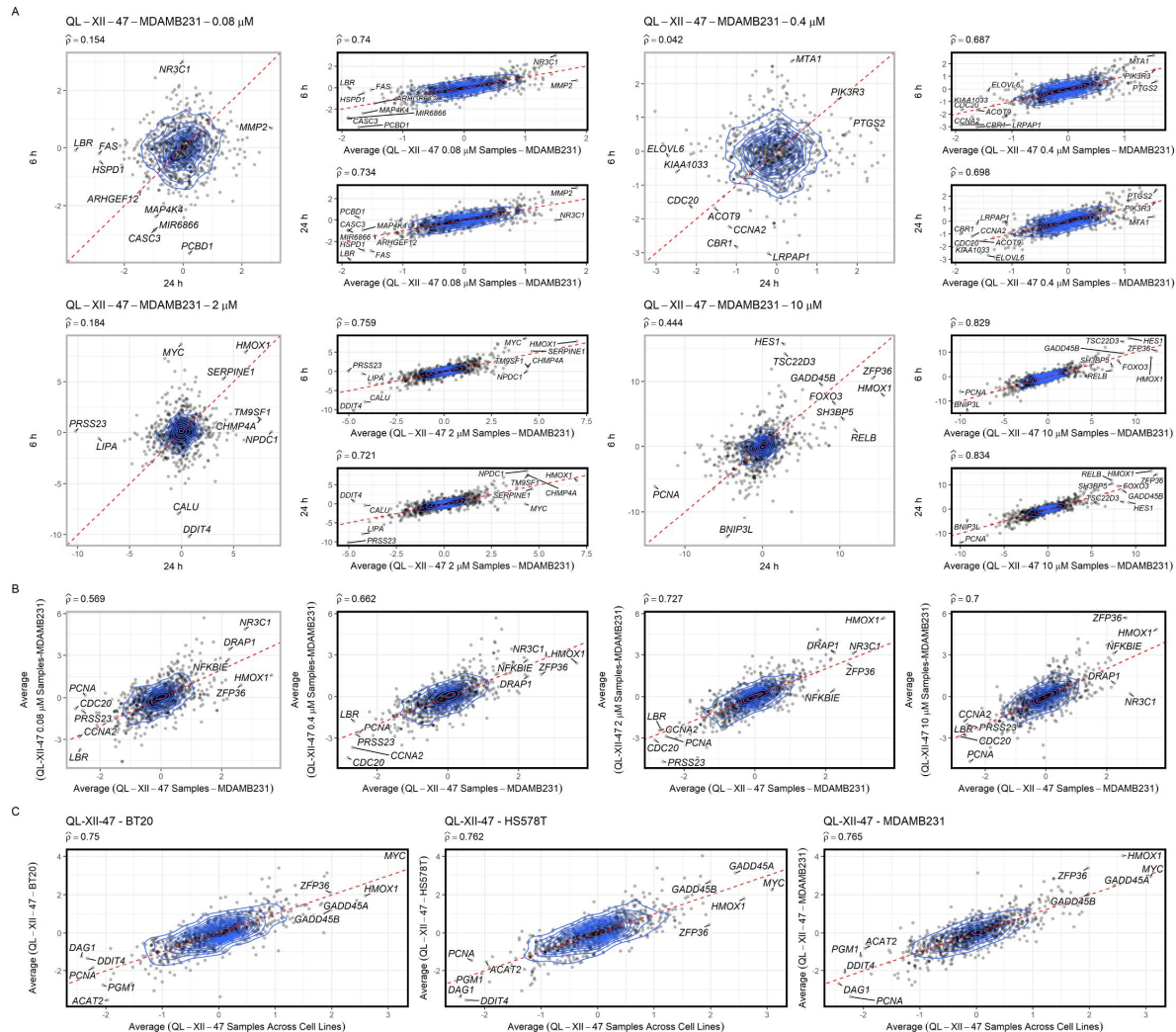


Fig. 4.

Case example showing the construction of a single transcriptional response profile to a compound across three TNBC cell lines.

Frames of the scatterplots displaying the relationship between the profiles and the averaged profiles are color coded, in black if the computed Spearman correlation coefficient ($\hat{\rho}$) is larger than 0.6 or in gray otherwise. **(A)** Generation of a time-consistent response profile. **(B)** Generation of a time and concentration-consistent response profile. **(C)** Generation of a time, concentration, and cell-line-consistent response profile.

Having a single generalized disease-specific response profile for each compound allowed us to (1 [↗](#)) generate compound combinations *in-silico* as candidates for the treatment of TNBC (Supplementary Table 2), (2 [↗](#)) to rank compounds and compound combinations based on their predicted potential to reverse the disease state towards a healthy-like state (Supplementary Table 3 and 4), and to (3 [↗](#)) interrogate the computed profiles using rank-based gene set enrichment analysis (GSEA) to predict possible mechanisms of action of the compounds when used to treat TNBC.

Our method to extract disease-specific drug response profiles is implemented in the ‘*retriever*’ R package (see Data Availability section) and allows the computation of disease-specific transcriptional drug response signatures for different cancer types available in the LINCS-L1000 project.

Ranking compound candidates for the treatment of TNBC

We used the 152 TNBC-specific transcriptional response profiles extracted with *retriever* to compute 11, 476 response profiles to nonredundant drug combinations. To do so, we calculated the averaged effect sizes of Spearman correlation between the response profiles and the differential single-cell transcriptional signatures and then ranked the compounds and compound combinations likely to specifically antagonize the TNBC-specific transcriptional signatures. Since Spearman correlation is a rank-based approach, this method can quantitatively identify inverse effects induced by drugs using the response profiles and disease-associated transcriptional signatures. Thus, we computed the Spearman coefficients, 95% confidence intervals, and p-values corrected for multiple testing (FDR) for all 11, 628 profiles (Supplementary Tables 3 and 4). We then ranked the drug combinations by their correlation coefficients, from negative to positive. A negative value represents the expected inverse effect of the drug against the expression changes observed in the disease towards a healthy-like state.

We found that QL-XII-47 is the most promising compound for reversing the transcriptional profile of TNBC back to a healthy-like state. Followed by Torin-2, Torin-1, QL-X-138, and WYE-125132. This top candidate remains consistent even when evaluating each disease signature for individual diseased samples included in the generated atlas. Specifically, QL-XII-47 emerged as the top candidate in 11 of the 25 ($P = 9.99 \times 10^{-25}$; Top panel in **Figure 6** [↗](#)) analyzed samples (one sample was excluded due to insufficient cells for differential expression analysis after quality control).

QL-XII-47 is a highly effective and selective Bruton tyrosine kinase (BTK) inhibitor that covalently modifies Cys481 of the protein. QL-XII-47 has an IC_{50} of 7 nM for inhibiting BTK kinase activity and induces a G1 cell cycle arrest in Ramos cells (B lymphocytes from Burkitt lymphoma), which is associated with significant degradation of the BTK protein. It was also shown that, at sub micromolar concentrations, QL-XII-47 inhibits the proliferation of B-cell lymphoma cell lines (50 [↗](#)). Furthermore, other BTK inhibitors have shown to reduce TNBC cell viability (51 [↗](#)). In addition, independent validation of sensitivity to QL-XII-47 has been tested before in two other TNBC cell lines (HC70 and HCC1806) that were not included in the LINCS-L1000 project. Low toxicity (lower than Torin-2) and more than 50% reduction of the growth rate (GR_{max}) was found for both cell lines (52 [↗](#)).

Through enrichment analysis, we found that QL-XII-47 may act in TNBC through activation of the *TNF-alpha Signaling via NF-κB*, *Hypoxia-associated genes*, the *Inflammatory Response*, the *IL-2/STAT5 Signaling pathway*, and by deactivating genes involved in *Fatty Acid Metabolism* (FDR < 0.05 in all cases, **Fig 5A** [↗](#)).

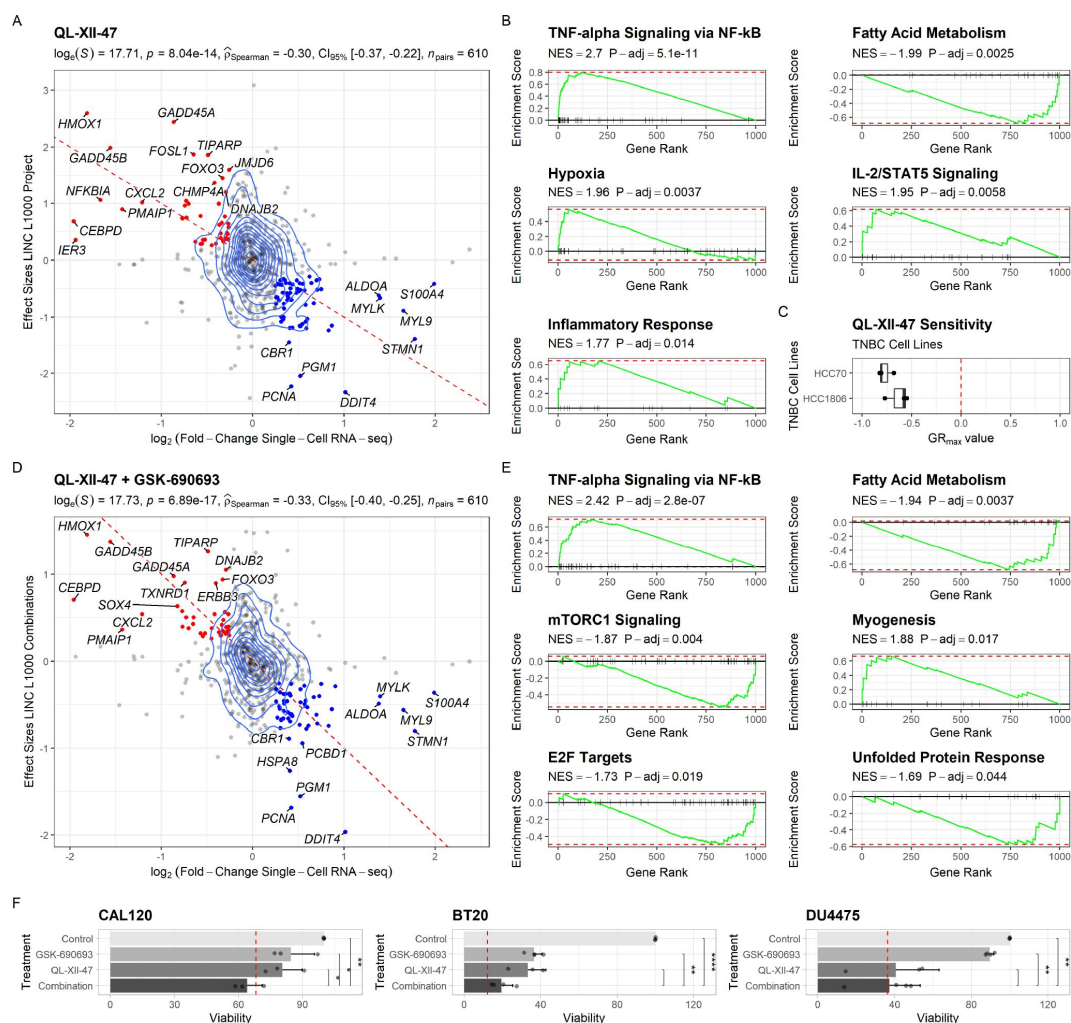


Fig. 5.

Correlation analysis and mechanism of action prediction for the disease-specific drug response profiles in TNBC.

(A) Spearman correlation analysis between the expression changes associated with TNBC and the disease-specific drug response signature of QL-XII-47. Each dot represents a gene. Dots are color coded, in red if they are expected to be upregulated by the drug, in blue if they are expected to be downregulated by the drug, and in gray if no significant change is expected. The red line represents the perfect match between both profiles. Density lines reflect the number of dots in each section of the plot. (B) Mechanisms of action predicted for QL-XII-47 in TNBC based on the enrichment of biological pathways in the disease-specific transcriptional drug response signature. (C) Independent sensitivity evaluation of the effect of QL-XII-47 in two other TNBC cell lines. (D) Spearman correlation analysis between the expression changes associated with TNBC and the combination signature of QL-XII-47 and GSK-690693. Each dot represents a gene. Dots are color coded as in (A). The red line represents the perfect match between both profiles. Density lines reflect the number of dots in each section of the plot. (E) Mechanisms of action predicted for the mixture of QL-XII-47 and GSK-690693 based on the enrichment of their combination drug response signature. (F) Comparisons of cellular viability among three distinct breast cancer cell lines following treatment with 0.6 μM QL-XII-47, 0.8 μM GSK-690693, and the combination of both compounds under same concentration. In red, is the expected viability under the drug effect additive scenario. P values were calculated using a one-sided t-test: * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$.

Ranking compound combination candidates for the treatment of TNBC

After testing combinations of diseasespecific drug response profiles, we found that the combination of QL-XII-47 with GSK-690693 – a pan-AKT kinase inhibitor that reduces tumor cell proliferation and induces tumor cell apoptosis (53 [↗](#), 54 [↗](#)) – was the best-performing drug combination to revert TNBC signatures. This combination consistently ranked as the best across individual diseased samples in our atlas. Specifically, QL-XII-47 and GSK-690693 emerged as the top combination in 4 of the 25 analyzed samples ($P = 5.76 \times 10^{-17}$; see bottom panel in **Figure 6** [↗](#)).

Gene set enrichment analysis predicted that this combination may act through the activation of *TNF-alpha signaling via NF-κB* and *Myogenesis*, and the deactivation of *Fatty Acid Metabolism*, *mTORC1 Signaling*, *E2F targets*, and the *Unfolded Protein Response* (FDR < 0.05 in all cases). These pathways have previously been associated with the inhibition of triplenegative breast cancer growth and metastasis (55 [↗](#)–57 [↗](#)), and thus highlight the potential of our computational approach to prioritize drugs by integrating cell-type specific profiles obtained from single-cell RNA-seq data sets with disease-specific transcriptional drug response profiles.

Moreover, we performed experimental validation of the identified top combination in three different breast cancer cell lines (CAL120, BT20, and DU4475). We compared the cellular viability (Supplementary Table 6) of the cells under baseline conditions and after treatment with 0.6 μM QL-XII-47, 0.8 μM GSK690693, and the combination of both compounds under the same concentration. We found that in all cases, as predicted by *retriever*, QL-XII-47 (top hit) displayed the strongest effect compared to GSK-690693. Additionally, we found that both compounds displayed an additive effect in two (BT20 and CAL120) of the three cell lines tested supporting *retriever*'s potential to identify useful drug combinations from cellular transcriptional responses to individual compounds.

Discussion

Single-cell RNA-seq provides an unprecedented resolution to characterize cellular heterogeneity in cancer. Compared with cell lines, the cells from tumors characterized by single-cell RNA-seq do not exhibit metabolic adaptation due to culturing. In advantage to bulk RNA-seq, single-cell RNA-seq can provide a more accurate signature of the changes observed in specific cells relevant to the disease under study (14 [↗](#), 58 [↗](#)). Thus, leveraging the information provided by this new technology is important to accelerate the identification of new personalized treatments that target specific subpopulations of cells in a tissue (59 [↗](#)).

However, just as important as characterizing the signatures of the disease in affected cells at the highest possible resolution is to characterize the response profiles of drugs that may help to reverse disease-associated transcriptional changes towards a healthy state. Large consortia including the LINCS-L1000 project provide transcriptional phenotypes at different time points after applying hundreds of compounds to multiple types of cell lines at different concentrations. Even though this information is valuable, its direct application to pathophysiological scenarios is limited due to the difficulty of extracting consistent response profiles in these large data sets (60 [↗](#)).

Here, we introduced *retriever*, a tool to extract robust diseasespecific drug response profiles from the LINCS-L1000 project. *retriever* mines the information associated with each cell line in the project to select subsets of cell lines that are surrogates of a defined disease, after which it extracts

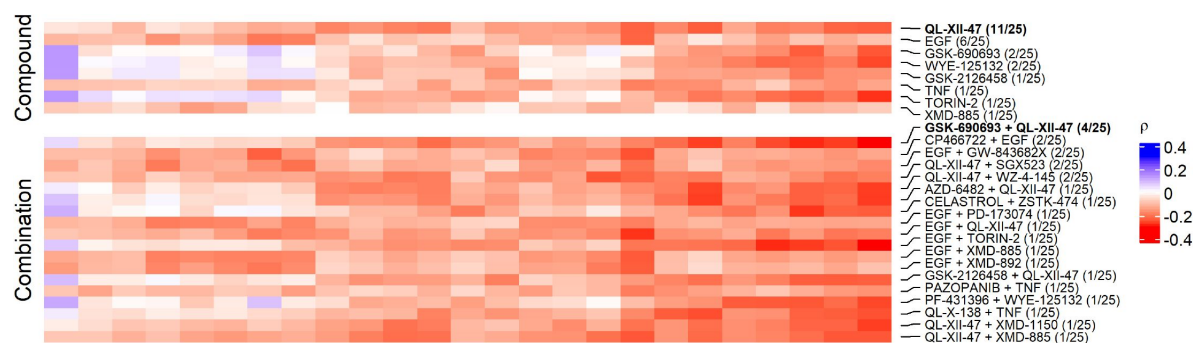


Fig. 6.

Heatmap displaying the Spearman correlation coefficients for the top-ranking compounds and compound combinations in each sample within the generated atlas.

The compounds or combinations that ranked 1st in at least one sample are included. The number of times each compound or combination ranked 1st across all samples is indicated in parentheses.

disease-specific response profiles. By doing this, *retriever* maximizes the robustness of the transcriptional signatures that are used to rank candidate compounds for the identification of new treatment strategies. With *retriever*, we hypothesize that if a signature is robust across multiple cell lines representing the same disease, it will likely also be more robust across individuals, as cell lines exhibit genetic and metabolic differences.

We showcased *retriever*'s strength by retrieving time-, dose-, and cell-type consistent transcriptional drug response signatures of TNBC cells from the LINCS-L1000 project and combining it with single-cell RNA transcriptome profiles obtained from a large single-cell breast cancer atlas, which we used to predict novel drug combinations against TNBC. This identified a combination of two kinase inhibitors predicted to be effective together acting on important biological pathways in breast cancer, and thereby antagonize the TNBC-specific signature.

The prediction of drug combinations with *retriever* that we present here, although it is relatively simple and relies on the hypothesis of independent mechanisms of action for each compound, has been established as accurate (predicting the right directionality of the change) in analyses performed on profiles from the Connectivity Map project (28 [↗](#)). Nevertheless, our computational approach only allows us to rank compounds and compounds combinations that are suitable for the development of treatments based on the negative correlation of the drug response profiles and the disease signatures. Other promising compounds and drug combinations still need to be validated experimentally to define the right concentration and evaluate if they indeed exhibit a synergistic mechanism of action. Their toxicity across cell lines and, after preclinical tests, potential adverse events in *in-vivo* models and cancer patients also need to be evaluated.

Thanks to the multiple cell lines available in the LINCS-L1000 project, our approach can be replicated in at least 13 other cancer types, including: *adult acute monocytic leukemia*, *adult acute myeloid leukemia*, *cecum adenocarcinoma*, *colon (adeno)carcinoma*, *endometrial adenocarcinoma*, *lung adenocarcinoma*, *large cell lung carcinoma*, *small cell lung carcinoma*, *melanoma*, *ovarian mucinous adenocarcinoma*, *prostate carcinoma*, and *triple-negative breast cancer* among other variants (such as local or metastatic tumors for some cancer types), as soon as single-cell RNA-seq data for those cancer types and healthy tissues become available. Considering the increase in single-cell RNA-seq studies being published since the development of the technique, we expect this to be possible for most cancer types represented in the LINCS-L1000 project in the near future (61 [↗](#)).

Finally, our approach can also be applied to disease profiles derived from a single patient to recommend personalized treatment strategies. We envision this to be one of the potentially most impactful applications of *retriever* in computationally-informed precision medicine once single-cell transcriptional profiles become a standard in the clinic. In addition, while we have showcased *retriever*'s strength in predicting drug combinations for TNBC in general, the method can be applied to single-cell RNAseq data derived from individual tumors, opening up the road for applications in precision medicine.

Supplementary Material

Supplementary Figures

Fig. S1.

Comparisons of *ESR1*, *PGR*, and *ERBB2* expression levels across epithelial cells in cancer and healthy tissues.

TP stands for Triple-Positive, and TN for Triple-Negative.

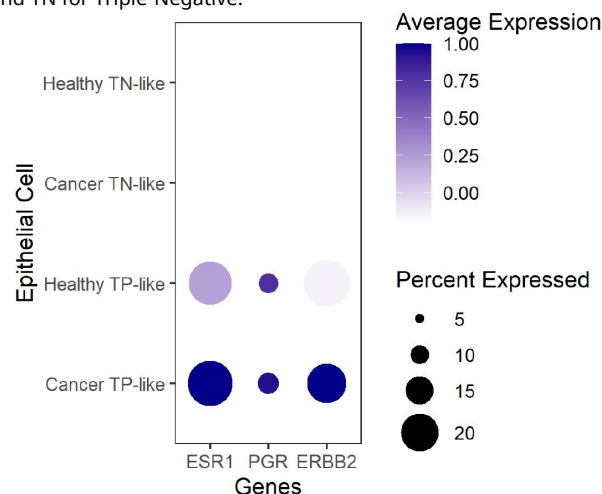
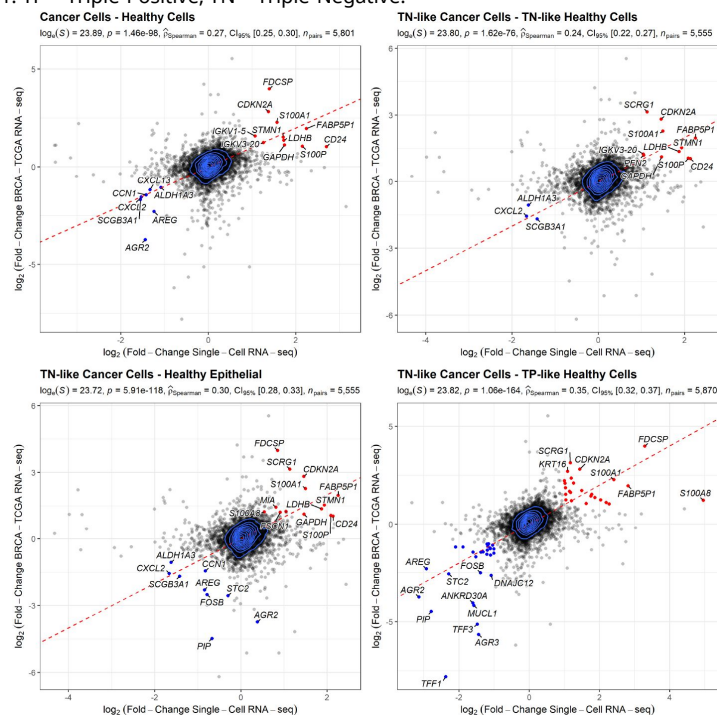


Fig. S2.

Comparisons of the transcriptional changes associated with TNBC at the single-cell and tissue level across different subpopulations of cells.

Each dot represents a gene. Dots are color coded, in red if the $\log_2$ fold-change is larger than 1 and in blue if the $\log_2$ fold-change is smaller than -1. TP = Triple-Positive, TN = Triple-Negative.



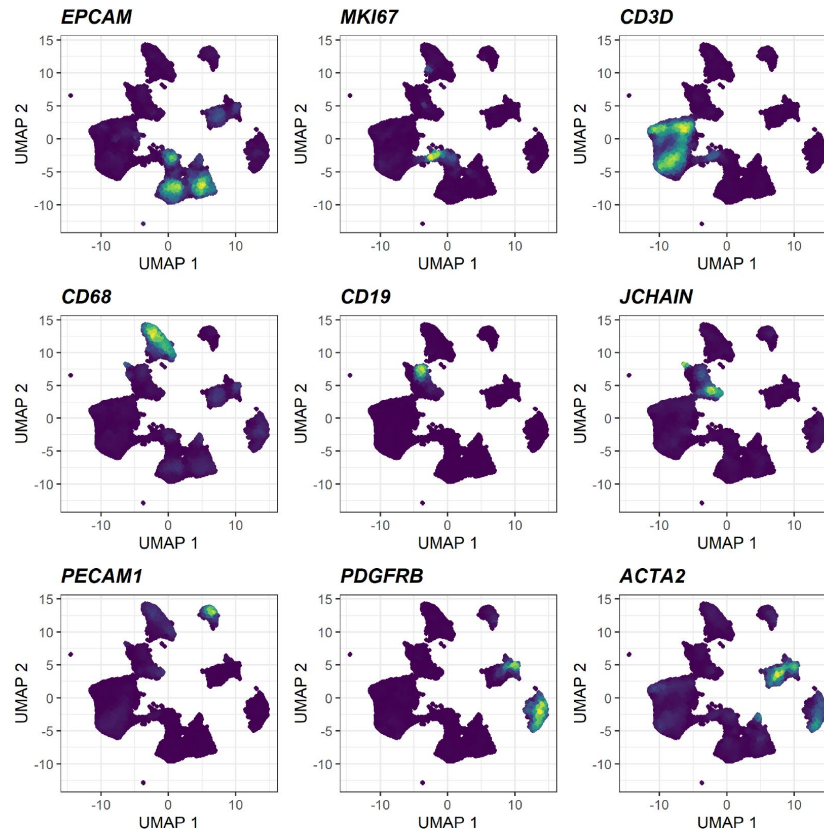


Fig. S3.

Expression of markers for epithelial cells (*EPCAM*), proliferating cells (*MKI67*), T cells (*CD3D*), myeloid cells (*CD68*), B cells (*MS4A1*), plasmablasts (*JCHAIN*), endothelial cells (*PECAM1*), mesenchymal cells (fibroblasts/perivascular-like cells; *PDGFRB*), and muscular cells (*ACTA2*).

Supplementary Tables

Cancer Hallmark	FDR	Leading Genes
Oxidative Phosphorylation	1.76×10^{-4}	<i>LDHB</i> , <i>IDH2</i> , <i>SLC25A5</i> , <i>ATP5F1C</i> , <i>COX6C</i> , <i>PHB2</i> , <i>ATP6V0C</i> , and <i>UQCRLH</i>
Interferon Alpha Response	6.10×10^{-3}	<i>CD74</i> , <i>IFITM1</i> , <i>IFI27</i> , and <i>LY6E</i>
Myc Targets	4.24×10^{-2}	<i>RPLP0</i> , <i>PABPC1</i> , <i>PHB2</i> , and <i>PPIA</i>
TNF α Signaling via NF- κ B	2.95×10^{-19}	<i>EDN1</i> , <i>BTG1</i> , <i>GADD45B</i> , <i>TSC22D1</i> , <i>CEBPD</i> , <i>CXCL1</i> , <i>SOD2</i> , <i>AREG</i> , <i>CXCL2</i> , <i>NFKBIA</i> , <i>MARCKS</i> , <i>KLF6</i> , <i>MAFF</i> , <i>CCL2</i> , <i>FOSB</i> , <i>CCN1</i> , <i>IL6ST</i> , <i>IER5</i> , <i>ATF3</i> , and <i>IER3</i>
Estrogen Response	6.54×10^{-8}	<i>UGCG</i> , <i>KRT18</i> , <i>ELOVL5</i> , <i>STC2</i> , <i>HSPB8</i> , <i>SLC39A6</i> , <i>TFF3</i> , <i>PMAIP1</i> , <i>TFF1</i> , <i>IL6ST</i> , and <i>AREG</i>
UV Response Up	6.54×10^{-8}	<i>NFKBIA</i> , <i>BTG1</i> , <i>HNRNPU</i> , <i>HMOX1</i> , <i>FOSB</i> , <i>TMBIM6</i> , <i>IL6ST</i> , <i>SOD2</i> , <i>CXCL2</i> , and <i>ATF3</i>
Apoptosis	6.54×10^{-8}	<i>WEE1</i> , <i>KRT18</i> , <i>GADD45B</i> , <i>TNFRSF12A</i> , <i>PMAIP1</i> , <i>HMOX1</i> , <i>EMPI1</i> , <i>SOD2</i> , <i>ATF3</i> , and <i>IER3</i>
Hypoxia	3.45×10^{-7}	<i>KLF6</i> , <i>BTG1</i> , <i>HSPA5</i> , <i>STC2</i> , <i>MAFF</i> , <i>HMOX1</i> , <i>ADM</i> , <i>CCN1</i> , <i>ATF3</i> , and <i>IER3</i>
Unfolded Protein Response	9.36×10^{-5}	<i>HSPA5</i> , <i>STC2</i> , <i>DNAJB9</i> , <i>CCL2</i> , <i>ATF3</i> , and <i>ATF4</i>
IL-6/JAK/STAT3 Signaling	2.59×10^{-4}	<i>TNFRSF12A</i> , <i>HMOX1</i> , <i>CXCL1</i> , <i>CXCL13</i> , and <i>IL6ST</i>
Inflammatory Response	1.57×10^{-3}	<i>NFKBIA</i> , <i>EDN1</i> , <i>KLF6</i> , <i>CXCL8</i> , <i>CCL2</i> , and <i>ADM</i>
Androgen Response	4.62×10^{-3}	<i>TSC22D1</i> , <i>ELOVL5</i> , <i>DNAJB9</i> , and <i>SLC38A2</i>

Table S1.

Hallmark MSigDB signatures associated with the differentially expressed genes observed in triple-negative breast cancer (TNBC).

False Discovery Rate (FDR) computed using Gene Set Enrichment Analysis (GSEA).

Data Availability

All the data and code required to replicate the analysis as well as the figures and tables are available at <https://github.com/dosorio/L1000-TNBC>. The *retriever* package is available at <https://github.com/kuijjerlab/retriever>. The following datasets were used to construct the single-cell RNA-seq breast atlas used in this study:

1. Tabula sapiens wild-type mammary gland. Data from (62) Quake, Stephen R., and Tabula Sapiens Consortium. “*The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors.*” *bioRxiv* (2021). Accessed through: <https://tabula-sapiensportal.ds.czbiohub.org/>
2. Wild-type data from seven individuals reported by (63) Nguyen, Quy H., *et al.* “*Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity.*” *Nature communications* 9.1 (2018): 1-12. GEO accession code: GSE113197.
3. Wild-type data from five individuals reported by (64) BhatNakshatri, Poornima, *et al.* “*A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells.*” *Cell Reports Medicine* 2.3 (2021): 10021955. GEO accession code: GSE164898.
4. Triple-negative breast cancer data from 9 individuals from (34) Wu SZ, Al-Eryani G, Roden DL, Junankar S *et al.* *A single-cell and spatially resolved atlas of human breast cancers.* *Nat Genet* 2021 Sep;53(9):1334-134756. GEO accession code: GSE176078
5. Breast cancer data from 17 individuals from the ArrayExpress database accession code: E-MTAB-8107. Described in (65) Qian, Junbin, *et al.* “*A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling.*” *Cell research* 30.9 (2020): 745-76257.

Acknowledgements

This work was supported by the Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (187615, M.L.K and D.O), the Norwegian Cancer Society (214871, M.L.K.), the Norwegian Research Council (313932, M.L.K), and the Marie SkłodowskaCurie Postdoctoral Scientia Fellows program from the Faculty of Medicine at the University of Oslo, co-funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions Grant (801133, D.O and P.S).

Additional files

Table S2. Computed robust disease-specific response profiles to 152 compounds across three triple-negative breast cancer cell lines (TNBC). [↗](#)

Table S3. Computed 11, 476 compound combinations response profiles across three triple-negative breast cancer cell lines (TNBC). [↗](#)

Table S4. Ranked compounds based on their predicted potential to reverse the disease state (TNBC) towards a healthy state. [↗](#)

Table S5. Ranked compound combinations based on their predicted potential to reverse the disease state (TNBC) towards a healthy state. [↗](#)

Table S6. Cellular viability measurements among three distinct breast cancer cell lines following treatment with 0.6 μ M QL-XII-47, 0.8 μ M GSK-690693, and the combination of both compounds under same concentration. [↗](#)

Table S7. Ranked compounds for each sample based on their predicted potential to reverse the disease state (TNBC) towards a healthy state. [↗](#)

Table S8. Ranked compound combinations for each sample based on their predicted potential to reverse the disease state (TNBC) towards a healthy state. [↗](#)

Note

This reviewed preprint has been updated to correct the affiliations.

References

1. Sleire L., Forde H. E., Netland I. A., Leiss L., Skeie B. S., Enger P. O. 2017) **Drug repurposing in cancer** *Pharmacol Res* **124**:74–91 <https://doi.org/10.1016/j.phrs.2017.07.013>
2. Hodos R. A., Kidd B. A., Shameer K., Readhead B. P., Dudley J. T. 2016) **In silico methods for drug repurposing and pharmacology** *Wiley Interdiscip Rev Syst Biol Med* **8**:186–210 <https://doi.org/10.1002/wsbm.1337>
3. Liu T. P., Hsieh Y. Y., Chou C. J., Yang P. M. 2018) **Systematic polypharmacology and drug repurposing via an integrated l1000-based connectivity map database mining** *R Soc Open Sci* **5**:181321 <https://doi.org/10.1098/rsos.181321>
4. Subramanian A., Narayan R., Corsello S. M., Peck D. D., Natoli T. E., Lu X., Gould J., Davis J. F., Tubelli A., Asiedu J. K., Lahr D. L., Hirschman J. E., Liu Z., Donahue M., Julian B., Khan M., Wadden D., Smith I. C., Lam D., Liberzon A., Toder C., Bagul M., Orzechowski M., Enache O. M., Piccioni F., Johnson S. A., Lyons N. J., Berger A. H., Shamji A. F., Brooks A. N., Vrcic A., Flynn C., Rosains J., Takeda D. Y., Hu R., Davison D., Lamb J., Ardlie K., Hogstrom L., Greenside P., Gray N. S., Clemons P. A., Silver S., Wu X., Zhao W. N., Read-Button W., Wu X., Haggarty S. J., Ronco L. V., Boehm J. S., Schreiber S. L., Doench J. G., Bittker J. A., Root D. E., Wong B., Golub T. R. 2017) **A next generation connectivity map: L1000 platform and the first 1,000,000 profiles** *Cell* **171**:1437–1452 <https://doi.org/10.1016/j.cell.2017.10.049>
5. Wang Y., Yella J., Jegga A. G. 2019) **Transcriptomic data mining and repurposing for computational drug discovery** *Methods Mol Biol* **1903**:73–95 https://doi.org/10.1007/978-1-4939-8955-3_5
6. Ben Guebila M., Lopes-Ramos C. M., Weighill D., Sonawane A. R., Burkholz R., Shamsaei B., Platig J., Glass K., Kuijjer M. L., Quackenbush J. 2021) **Grand: a database of gene regulatory network models across human conditions** *Nucleic Acids Res* <https://doi.org/10.1093/nar/gkab778>
7. Ozdemir E. S., Halakou F., Nussinov R., Gursoy A., Keskin O. 2019) **Methods for discovering and targeting druggable protein-protein interfaces and their application to repurposing** *Methods Mol Biol* **1903**:1–21 https://doi.org/10.1007/978-1-4939-8955-3_1
8. Su E. W., Sanger T. M. 2017) **Systematic drug repositioning through mining adverse event data in clinicaltrials.gov** *PeerJ* **5**:e3154 <https://doi.org/10.7717/peerj.3154>
9. Avalos-Moreno M., Lopez-Tejada A., Blaya-Canovas J. L., Cara-Lupianez F. E., Gonzalez-Gonzalez A., Lorente J. A., Sanchez-Rovira P., Granados-Principal S. 2020) **Drug repurposing for triple-negative breast cancer** *J Pers Med* **10** <https://doi.org/10.3390/jpm10040200>
10. Whitehead A., Crawford D. L. 2005) **Variation in tissue-specific gene expression among natural populations** *Genome Biol* **6**:R13 <https://doi.org/10.1186/gb-2005-6-2-r13>
11. Simonovsky Eyal, Schuster Ronen, Yeger-Lotem Esti 2019) **Large-scale analysis of human gene expression variability associates highly variable drug targets with lower drug effectiveness and safety** *Bioinformatics* **35**:3028–3037

12. Levitin H. M., Yuan J., Sims P. A. 2018) **Single-cell transcriptomic analysis of tumor heterogeneity** *Trends Cancer* **4**:264–268 <https://doi.org/10.1016/j.trecan.2018.02.003>
13. Consortium International Cancer Genome, Hudson T. J., Anderson W., Artez A., Barker A. D., Bell C., Bernabe R. R., Bhan M. K., Calvo F., Eerola I., Gerhard D. S., Gutmacher A., Guyer M., Hemsley F. M., Jennings J. L., Kerr D., Klatt P., Kolar P., Kusada J., Lane D. P., Laplace F., Youyong L., Nettekoven G., Ozenberger B., Peterson J., Rao T. S., Remacle J., Schafer A. J., Shibata T., Stratton M. R., Vockley J. G., Watanabe K., Yang H., Yuen M. M., Knoppers B. M., Bobrow M., Cambon-Thomsen A., Dressler L. G., Dyke S. O., Joly Y., Kato K., Kennedy K. L., Nicolas P., Parker M. J., Rial-Sebbag E., Romeo-Casabona C. M., Shaw K. M., Wallace S., Wiesner G. L., Zeps N., Lichter P., Biankin A. V., Chabannon C., Chin L., Clement B., de Alava E., Degos F., Ferguson M. L., Geary P., Hayes D. N., Hudson T. J., Johns A. L., Kasprzyk A., Nakagawa H., Penny R., Piris M. A., Sarin R., Scarpa A., Shibata T., van de Vijver M., Futreal P. A., Aburatani H., Bayes M., Botwell D. D., Campbell P. J., Estivill X., Gerhard D. S., Grimmond S. M., Gut I., Hirst M., Lopez-Otin C., Majumder P., Marra M., McPherson J. D., Nakagawa H., Ning Z., Puente X. S., Ruan Y., Shibata T., Stratton M. R., Stunnenberg H. G., Swerdlow H., Velculescu V. E., Wilson R. K., Xue H. H., Yang L., Spellman P. T., Bader G. D., Boutros P. C., Campbell P. J., et al. 2010) **International network of cancer genome projects** *Nature* **464**:993–8 <https://doi.org/10.1038/nature08987>
14. Potter S. S. 2018) **Single-cell rna sequencing for the study of development, physiology and disease** *Nat Rev Nephrol* **14**:479–492 <https://doi.org/10.1038/s41581-018-0021-7>
15. Kolodziejczyk A. A., Kim J. K., Svensson V., Marioni J. C., Teichmann S. A. 2015) **The technology and biology of single-cell rna sequencing** *Mol Cell* **58**:610–20 <https://doi.org/10.1016/j.molcel.2015.04.005>
16. Osorio D., Kuijjer M. L., Cai J. J. 2021) **rpanglaodb: an r package to download and merge labeled single-cell rna-seq data from the panglaodb database** *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab549>
17. Zimmerman K. D., Espeland M. A., Langefeld C. D. 2021) **A practical solution to pseudoreplication bias in single-cell studies** *Nat Commun* **12**:738 <https://doi.org/10.1038/s41467-021-21038-1>
18. Squair J. W., Gautier M., Kathe C., Anderson M. A., James N. D., Hutson T. H., Hudelle R., Qaiser T., Matson K. J. E., Barraud Q., Levine A. J., Manno G. La, Skinnider M. A., Courtine G. 2021) **Confronting false discoveries in single-cell differential expression** *Nat Commun* **12**:5692 <https://doi.org/10.1038/s41467-021-25960-2>
19. Caroli J., Dori M., Bicciato S. 2020) **Computational methods for the integrative analysis of genomics and pharmacological data** *Front Oncol* **10**:185 <https://doi.org/10.3389/fonc.2020.00185>
20. Qiu Y., Lu T., Lim H., Xie L. 2020) **A bayesian approach to accurate and robust signature detection on lincs l1000 data** *Bioinformatics* **36**:2787–2795 <https://doi.org/10.1093/bioinformatics/btaa064>
21. Gonda T. J., Ramsay R. G. 2015) **Directly targeting transcriptional dysregulation in cancer** *Nat Rev Cancer* **15**:686–94 <https://doi.org/10.1038/nrc4018>
22. Chatterjee N., Bivona T. G. 2019) **Polytherapy and targeted cancer drug resistance** *Trends Cancer* **5**:170–182 <https://doi.org/10.1016/j.trecan.2019.02.003>

23. Jeon M., Kim S., Park S., Lee H., Kang J. 2018) **In silico drug combination discovery for personalized cancer therapy** *BMC Syst Biol* **12**:16 <https://doi.org/10.1186/s12918-018-0546-1>
24. Liu H., Zhang W., Zou B., Wang J., Deng Y., Deng L. 2020) **Drugcomdbb: a comprehensive database of drug combinations toward the discovery of combinatorial therapy** *Nucleic Acids Res* **48**:D871–D881 <https://doi.org/10.1093/nar/gkz1007>
25. Celebi R., Bear O. th, Walk Don't, Movva R., Alpsoy S., Dumontier M. 2019) **In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data** *Sci Rep* **9**:8949 <https://doi.org/10.1038/s41598-019-45236-6>
26. Lamb J., Crawford E. D., Peck D., Modell J. W., Blat I. C., Wrobel M. J., Lerner J., Brunet J. P., Subramanian A., Ross K. N., Reich M., Hieronymus H., Wei G., Armstrong S. A., Haggarty S. J., Clemons P. A., Wei R., Carr S. A., Lander E. S., Golub T. R. 2006) **The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease** *Science* **313**:1929–35 <https://doi.org/10.1126/science.1132939>
27. Pickering A. 2021) **ccmap: Combination connectivity mapping** *Bioconductor* <https://doi.org/10.18129/B9.bioc.ccmap>
28. Pickering A. 2021) **ccdata: Data for combination connectivity mapping (ccmap) package** *Bioconductor* <https://doi.org/10.18129/B9.bioc.ccdata>
29. Howe K. L., Achuthan P., Allen J., Allen J., Alvarez-Jarreta J., Amode M. R., Armean I. M., Azov A. G., Bennett R., Bhai J., Billis K., Boddu S., Charkhchi M., Cummins C., Da Rin Fioretto L., Davidson C., Dodiya K., El Houdaigui B., Fatima R., Gall A., Garcia Giron C., Grego T., Guijarro-Clarke C., Haggerty L., Hemrom A., Hourlier T., Izuogu O. G., Juettemann T., Kaikala V., Kay M., Lavidas I., Le T., Lemos D., Gonzalez Martinez J., Marugan J. C., Maurel T., McMahon A. C., Mohanan S., Moore B., Muffato M., Oheh D. N., Paraschas D., Parker A., Parton A., Prosovetskaia I., Sakthivel M. P., Salam A. I. A., Schmitt B. M., Schuilenburg H., Sheppard D., Steed E., Szpak M., Szuba M., Taylor K., Thormann A., Threadgold G., Walts B., Winterbottom A., Chakiachvili M., Chaubal A., De Silva N., Flint B., Frankish A., Hunt S. E., Gr I. Isley, Langridge N., Loveland J. E., Martin F. J., Mudge J. M., Morales J., Perry E., Ruffier M., Tate J., Thybert D., Trevanion S. J., Cunningham F., Yates A. D., Zerbino D. R., Flicek P. 2021) **Ensembl 2021** *Nucleic Acids Res* **49**:D884–D891 <https://doi.org/10.1093/nar/gkaa942>
30. Hao Y., Hao S., Andersen-Nissen E., Mauck W.M., Zheng S., Butler A., Lee M. J., Wilk A. J., Darby C., Zager M., Hoffman P., Stoeckius M., Papalexi E., Mimitou E. P., Jain J., Srivastava A., Stuart T., Fleming L. M., Yeung B., Rogers A. J., McElrath J. M., Blish C. A., Gottardo R., Smibert P., Satija R. 2021) **Integrated analysis of multimodal single-cell data** *Cell* **184**:3573–3587 <https://doi.org/10.1016/j.cell.2021.04.048>
31. Osorio D., Cai J. J. 2021) **Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell rna-sequencing data quality control** *Bioinformatics* **37**:963–967 <https://doi.org/10.1093/bioinformatics/btaa751>
32. Korsunsky I., Millard N., Fan J., Slowikowski K., Zhang F., Wei K., Baglaenko Y., Brenner M., Loh P. R., Raychaudhuri S. 2019) **Fast, sensitive and accurate integration of single-cell data with harmony** *Nat Methods* **16**:1289–1296 <https://doi.org/10.1038/s41592-019-0619-0>

33. Becht E., McInnes L., Healy J., Dutertre C. A., Kwok I. W. H., Ng L. G., Ginhoux F., Newell E. W. 2018) **Dimensionality reduction for visualizing single-cell data using umap** *Nat Biotechnol* <https://doi.org/10.1038/nbt.4314>
34. Wu S. Z., Al-Eryani G., Roden D. L., Junankar S., Harvey K., Andersson A., Thennavan A., Wang C., Torpy J. R., Bartonicek N., Wang T., Larsson L., Kaczorowski D., Weisenfeld N. I., Uytingco C. R., Chew J. G., Bent Z. W., Chan C. L., Gnanasambandapillai V., Dutertre C. A., Gluch L., Hui M. N., Beith J., Parker A., Robbins E., Segara D., Cooper C., Mak C., Chan B., Warriar S., Ginhoux F., Millar E., Powell J. E., Williams S. R., Liu X. S., O'Toole S., Lim E., Lundeborg J., Perou C. M., Swarbrick A. 2021) **A single-cell and spatially resolved atlas of human breast cancers** *Nat Genet* **53**:1334–1347 <https://doi.org/10.1038/s41588-021-00911-1>
35. Wu Sunny Z, Roden Daniel L, Wang Chenfei, Holliday Holly, Harvey Kate, Cazet Aurélie S, Murphy Kendelle J, Pereira Brooke, Al-Eryani Ghamdan, Bartonicek Nenad, et al. 2020) **Stromal cell diversity associated with immune evasion in human triple-negative breast cancer** *The EMBO journal* **39**:e104063
36. Alquicira-Hernandez J., Powell J. E. 2021) **Nebulosa recovers single cell gene expression signals by kernel density estimation** *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab003>
37. Finak G., McDavid A., Yajima M., Deng J., Gersuk V., Shalek A. K., Slichter C. K., Miller H. W., McElrath M. J., Prlic M., Linsley P. S., Gottardo R. 2015) **Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data** *Genome Biol* **16**:278 <https://doi.org/10.1186/s13059-015-0844-5>
38. Love M. I., Huber W., Anders S. 2014) **Moderated estimation of fold change and dispersion for rnaseq data with deseq2** *Genome Biol* **15**:550 <https://doi.org/10.1186/s13059-014-0550-8>
39. Colaprico A., Silva T. C., Olsen C., Garofano L., Cava C., Garolini D., Sabedot T. S., Malta T. M., Pagnotta S. M., Castiglioni I., Ceccarelli M., Bontempi G., Noushmehr H. 2016) **Tcgbiolinks: an r/bioconductor package for integrative analysis of tcga data** *Nucleic Acids Res* **44**:e71 <https://doi.org/10.1093/nar/gkv1507>
40. Spearman C. 1987) **The proof and measurement of association between two things. by c. spearman, 1904** *Am J Psychol* **100**:441–71
41. Benjamini Yoav, Hochberg Yosef 1995) **Controlling the false discovery rate: a practical and powerful approach to multiple testing** *Journal of the Royal statistical society: series B (Methodological)* **57**:289–300
42. Liberzon Arthur, Birger Chet, Thorvaldsdóttir Helga, Ghandi Mahmoud, Mesirov Jill P, Tamayo Pablo 2015) **The molecular signatures database hallmark gene set collection** *Cell systems* **1**:417–425
43. Xie Zhuorui, Bailey Allison, Kuleshov Maxim V, Clarke Daniel JB, Evangelista John E, Jenkins Sherry L, Lachmann Alexander, Wojciechowicz Megan L, Kropiwnicki Eryk, Jagodnik Kathleen M, et al. 2021) **Gene set knowledge discovery with enrichr** *Current protocols* **1**:e90
44. Evans K. W., Yuca E., Scott S. S., Zhao M., Paez Arango N., Cruz Pico C. X., Saridogan T., Shariati M., Class C. A., Bristow C. A., Vellano C. P., Zheng X., Gonzalez-Angulo A. M., Su X., Tapia C., Chen K., Akcakanat A., Lim B., Tripathy D., Yap T. A., Di Francesco M. E., Draetta G. F., Jones P., Heffernan T. P., Marszalek J. R., Meric-Bernstam F. 2021) **Oxidative phosphorylation is a**

metabolic vulnerability in chemotherapy-resistant triple negative breast cancer *Cancer Res* <https://doi.org/10.1158/0008-5472.CAN-20-3242>

45. Horiuchi D., Kusdra L., Huskey N. E., Chandriani S., Lenburg M. E., Gonzalez-Angulo A. M., Creasman K. J., Bazarov A. V., Smyth J. W., Davis S. E., Yaswen P., Mills G. B., Esserman L. J., Goga A. 2012) **Myc pathway activation in triple-negative breast cancer is synthetic lethal with cdk inhibition** *J Exp Med* **209**:679–96 <https://doi.org/10.1084/jem.20111512>
46. Provance O. K., Geanes E. S., Lui A. J., Roy A., Holloran S. M., Gunewardena S., Hagan C. R., Weir S., Lewis-Wambi J. 2021) **Disrupting interferon-alpha and nf-kappab crosstalk suppresses ifitm1 expression attenuating triple-negative breast cancer progression** *Cancer Lett* **514**:12–29 <https://doi.org/10.1016/j.canlet.2021.05.006>
47. Jamdade V. S., Sethi N., Mundhe N. A., Kumar P., Lahkar M., Sinha N. 2015) **Therapeutic targets of triple-negative breast cancer: a review** *Br J Pharmacol* **172**:4228–37 <https://doi.org/10.1111/bph.13211>
48. Gerratana L., Basile D., Buono G., De Placido S., Giuliano M., Minichillo S., Coinu A., Martorana F., De Santo I., Del Mastro L., De Laurentiis M., Puglisi F., Arpino G. 2018) **Androgen receptor in triple negative breast cancer: A potential target for the targetless subtype** *Cancer Treat Rev* **68**:102–110 <https://doi.org/10.1016/j.ctrv.2018.06.005>
49. Oshi M., Newman S., Tokumaru Y., Yan L., Matsuyama R., Endo I., Takabe K. 2020) **Inflammation is associated with worse outcome in the whole cohort but with better outcome in triple-negative subtype of breast cancer patients** *J Immunol Res* **2020**:5618786 <https://doi.org/10.1155/2020/5618786>
50. Wu H., Wang W., Liu F., Weisberg E. L., Tian B., Chen Y., Li B., Wang A., Wang B., Zhao Z., McMillin D. W., Hu C., Li H., Wang J., Liang Y., Buhrlage S. J., Liang J., Liu J., Yang G., Brown J. R., Treon S. P., Mitsiades C. S., Griffin J. D., Liu Q., Gray N. S. 2014) **Discovery of a potent, covalent btk inhibitor for b-cell lymphoma** *ACS Chem Biol* **9**:1086–91 <https://doi.org/10.1021/cb4008524>
51. Campbell R., Chong G., Hawkes E. A. 2018) **Novel indications for bruton's tyrosine kinase inhibitors, beyond hematological malignancies** *J Clin Med* **7** <https://doi.org/10.3390/jcm7040062>
52. Chopra Sameer S, Jenney Anne, Palmer Adam, Niepel Mario, Chung Mirra, Mills Caitlin, Sivakumaren Sindhu Carmen, Liu Qingsong, Chen Jia-Yun, Yapp Clarence, et al. 2020) **Torin2 exploits replication and checkpoint vulnerabilities to cause death of pi3k-activated triple-negative breast cancer cells** *Cell systems* **10**:66–81
53. Altomare D. A., Zhang L., Deng J., Di Cristofano A., Klein-Szanto A. J., Kumar R., Testa J. R. 2010) **Gsk690693 delays tumor onset and progression in genetically defined mouse models expressing activated akt** *Clin Cancer Res* **16**:486–96 <https://doi.org/10.1158/1078-0432.CCR-09-1026>
54. Levy D. S., Kahana J. A., Kumar R. 2009) **Akt inhibitor, gsk690693, induces growth inhibition and apoptosis in acute lymphoblastic leukemia cell lines** *Blood* **113**:1723–9 <https://doi.org/10.1182/blood-2008-02-137737>
55. Wang W., Zhang R., Wang X., Wang N., Zhao J., Wei Z., Xiang F., Wang C. 2020) **Suppression of kif3a inhibits triple negative breast cancer growth and metastasis by repressing rb-e2f**

signaling and epithelial-mesenchymal transition *Cancer Sci* **111**:1422–1434 <https://doi.org/10.1111/cas.14324>

56. Sulaiman A., McGarry S., Lam K. M., El-Sahli S., Chambers J., Kaczmarek S., Li L., Addison C., Dimitroulakos J., Arnaout A., Nessim C., Yao Z., Ji G., Song H., Liu S., Xie Y., Gadde S., Li X., Wang L. 2018) **Co-inhibition of mtorc1, hdac and esr1alpha retards the growth of triple-negative breast cancer and suppresses cancer stem cells** *Cell Death Dis* **9**:815 <https://doi.org/10.1038/s41419-018-0811-7>
57. Camarda R., Zhou A. Y., Kohnz R. A., Balakrishnan S., Mahieu C., Anderton B., Eyob H., Kajimura S., Tward A., Krings G., Nomura D. K., Goga A. 2016) **Inhibition of fatty acid oxidation as a therapy for myc-overexpressing triple-negative breast cancer** *Nat Med* **22**:427–32 <https://doi.org/10.1038/nm.4055>
58. Lagziel Shoval, Gottlieb Eyal, Shlomi Tomer 2020) **Mind your media** *Nature Metabolism* **2**:1369–1372
59. Wiedmeier Julia E, Noel Pawan, Lin Wei, Von Hoff Daniel D, Han Haiyong 2019) **Single-cell sequencing in precision medicine** In: *Precision Medicine in Cancer Therapy* Springer pp. 237–252
60. Safikhani Zhaleh, Smirnov Petr, Freeman Mark, El-Hachem Nehme, She Adrian, Rene Quevedo, Goldenberg Anna, Birkbak Nicolai J, Hatzis Christos, Shi Leming, et al. 2016) **Revisiting inconsistency in large pharmacogenomic studies** *F1000Research* **5**
61. Svensson Valentine, Eduardo da, Veiga Beltrame, Pachter Lior 2020) **A curated database reveals trends in single-cell transcriptomics** *Database* **2020**
62. The Tabula Sapiens Consortium, Quake Stephen R 2021) **The tabula sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors** *bioRxiv*
63. Nguyen Q. H., Pervolarakis N., Blake K., Ma D., Davis R. T., James N., Phung A. T., Willey E., Kumar R., Jabart E., Driver I., Rock J., Goga A., Khan S. A., Lawson D. A., Werb Z., Kessenbrock K. 2018) **Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity** *Nat Commun* **9**:2028 <https://doi.org/10.1038/s41467-018-04334-1>
64. Bhat-Nakshatri P., Gao H., Sheng L., McGuire P. C., Xuei X., Wan J., Liu Y., Althouse S. K., Colter A., Sandusky G., Storniolo A. M., Nakshatri H. 2021) **A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells** *Cell Rep Med* **2**:100219 <https://doi.org/10.1016/j.xcrm.2021.100219>
65. Qian J., Olbrecht S., Boeckx B., Vos H., Laoui D., Etlioglu E., Wauters E., Pomella V., Verbandt S., Busschaert P., Bassez A., Franken A., Bempt M. V., Xiong J., Weynand B., van Herck Y., Antoranz A., Bosisio F. M., Thienpont B., Floris G., Vergote I., Smeets A., Tejpar S., Lambrechts D. 2020) **A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling** *Cell Res* **30**:745–762 <https://doi.org/10.1038/s41422-020-0355-0>

Author information

Daniel Osorio*

Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway

ORCID iD: [0000-0003-4424-8422](https://orcid.org/0000-0003-4424-8422)

For correspondence: daniecos@uio.no

*These authors contributed equally to this work

Parastoo Shahrouzi*

Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway

*These authors contributed equally to this work

Xavier Tekpli

Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway

Vessela N Kristensen

Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway, Department of Clinical Molecular Biology and Laboratory Science (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, Norway

Marieke L Kuijjer

Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway, Department of Biochemistry and Developmental Biology, University of Helsinki, Helsinki, Finland, iCAN Flagship in Digital Precision Cancer Medicine, University of Helsinki, Helsinki, Finland

ORCID iD: [0000-0001-6280-3130](https://orcid.org/0000-0001-6280-3130)

For correspondence: marieke.kuijjer@ncmm.uio.no

Editors

Reviewing Editor

Yu Zhao

Institute of Radiation Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin, China

Senior Editor

Caigang Liu

Shengjing Hospital of China Medical University, Shenyang, China

Reviewer #1 (Public review):

Summary:

Identifying drugs that target specific disease phenotypes remains a persistent challenge. Many current methods are only applicable to well-characterized small molecules, such as those with known structures. In contrast, methods based on transcriptional responses offer broader applicability because they do not require prior information about small molecules.

Additionally, they can be rapidly applied to new small molecules. One of the most promising strategies involves the use of "drug response signatures"-specific sets of genes whose differential expression can serve as markers for the response to a small molecule. By comparing drug response signatures with expression profiles characteristic of a disease, it is possible to identify drugs that modulate the disease profile, indicating a potential therapeutic connection.

This study aims to prioritize potential drug candidates and to forecast novel drug combinations that may be effective in treating triple-negative breast cancer (TNBC). Large consortia, such as the LINCS-L1000 project, offer transcriptional signatures across various time points after exposing numerous cell lines to hundreds of compounds at different concentrations. While this data is highly valuable, its direct applicability to pathophysiological contexts is constrained by the challenges in extracting consistent drug response profiles from these extensive datasets. The authors use their method to create drug response profiles for three different TNBC cell lines from LINCS.

To create a more precise, cancer-specific disease profile, the authors highlight the use of single-cell RNA sequencing (scRNA-seq) data. They focus on TNBC epithelial cells collected from 26 diseased individuals compared to epithelial cells collected from 10 healthy volunteers. The authors are further leveraging drug response data to develop inhibitor combinations.

Strengths:

The authors of this study contribute to an ongoing effort to develop automated, robust approaches that leverage gene expression similarities across various cell lines and different treatment regimens, aiming to predict drug response signatures more accurately. The authors are trying to address the gap that remains in computational methods for inferring drug responses at the cell subpopulation level.

Weaknesses:

One weakness is that the authors do not compare their method to previous studies. The authors develop a drug response profile by summarizing the time points, concentrations, and cell lines. The computational challenge of creating a single gene list that represents the transcriptional response to a drug across different cell lines and treatment protocols has been previously addressed. The Prototype Ranked List (PRL) procedure, developed by Iorio and co-authors (PNAS, 2010, doi:10.1073/pnas.1000138107), uses a hierarchical majority-voting scheme to rank genes. This method generates a list of genes that are consistently overexpressed or downregulated across individual conditions, which then hold top positions in the PRL. The PRL methodology was used by Aissa and co-authors (Nature Comm 2021, doi:10.1038/s41467-021-21884-z) to analyze drug effects on selective cell populations using scRNA-seq datasets. They combined PRL with Gene Set Enrichment Analysis (GSEA), a method that compares a ranked list of genes like PRL against a specific set of genes of interest. GSEA calculates a Normalized Enrichment Score (NES), which indicates how well the genes of interest are represented among the top genes in the PRL. Compared to the method described in the current manuscript, the PRL method allows for the identification of both upregulated and downregulated transcriptional signatures relevant to the drug's effects. It also gives equal weight to each cell line's contribution to the drug's overall response signature.

The authors performed experimental validation of the top two identified drugs; however, the effect was modest. In addition, the effect on TNBC cell lines was cell-line specific as the identified drugs were effective against BT20, whose transcriptional signatures from LINCS were used for drug identification, but not against the other two cell lines analyzed. An incorrect choice of genes for the signature may result in capturing similarities tied to experimental conditions (e.g., the same cell line) rather than the drug's actual effects. This

reflects the challenges faced by drug response signature methods in both selecting the appropriate subset of genes that make up the signature and in managing the multiple expression profiles generated by treating different cell lines with the same drug.

<https://doi.org/10.7554/eLife.102442.1.sa2>

Reviewer #2 (Public review):

Summary:

In their study, Osorio and colleagues present 'retriever,' an innovative computational tool designed to extract disease-specific transcriptional drug response profiles from the LINCS-L1000 project. This tool has been effectively applied to TNBC, leveraging single-cell RNA sequencing data to predict drug combinations that may effectively target the disease. The public review highlights the significant integration of extensive pharmacological data with high-resolution transcriptomic information, which enhances the potential for personalized therapeutic applications.

Strengths:

A key finding of the study is the prediction and validation of the drug combination QL-XII-47 and GSK-690693 for the treatment of TNBC. The methodology employed is robust, with a clear pathway from data analysis to experimental confirmation.

Weaknesses:

However, several issues need to be addressed. The predictive accuracy of 'retriever' is contingent upon the quality and comprehensiveness of the LINCS-L1000 and single-cell datasets utilized, which is an important caveat as these datasets may not fully capture the heterogeneity of patient responses to treatment. While the in vitro validation of the drug combinations is promising, further in vivo studies and clinical trials are necessary to establish their efficacy and safety. The applicability of these findings to other cancer types also warrants additional investigation. Expanding the application of 'retriever' to a broader range of cancer types and integrating it with clinical data will be crucial for realizing its potential in personalized medicine. Furthermore, as the study primarily focuses on kinase inhibitors, it remains to be seen how well these findings translate to other drug classes.

<https://doi.org/10.7554/eLife.102442.1.sa1>

Author response:

Reviewer 1:

Summary:

Identifying drugs that target specific disease phenotypes remains a persistent challenge. Many current methods are only applicable to well-characterized small molecules, such as those with known structures. In contrast, methods based on transcriptional responses offer broader applicability because they do not require prior information about small molecules. Additionally, they can be rapidly applied to new small molecules. One of the most promising strategies involves the use of "drug response signatures"-specific sets of genes whose differential expression can serve as markers for the response to a small molecule. By comparing drug response signatures with expression profiles characteristic of a disease, it is possible to identify drugs that modulate the disease profile, indicating a potential therapeutic connection.

This study aims to prioritize potential drug candidates and to forecast novel drug combinations that may be effective in treating triple-negative breast cancer (TNBC). Large consortia, such as the LINCS-L1000 project, offer transcriptional signatures across various time points after exposing numerous cell lines to hundreds of compounds at different concentrations. While this data is highly valuable, its direct applicability to pathophysiological contexts is constrained by the challenges in extracting consistent drug response profiles from these extensive datasets. The authors use their method to create drug response profiles for three different TNBC cell lines from LINCS.

To create a more precise, cancer-specific disease profile, the authors highlight the use of single-cell RNA sequencing (scRNA-seq) data. They focus on TNBC epithelial cells collected from 26 diseased individuals compared to epithelial cells collected from 10 healthy volunteers. The authors are further leveraging drug response data to develop inhibitor combinations.

Strengths:

The authors of this study contribute to an ongoing effort to develop automated, robust approaches that leverage gene expression similarities across various cell lines and different treatment regimens, aiming to predict drug response signatures more accurately. The authors are trying to address the gap that remains in computational methods for inferring drug responses at the cell subpopulation level.

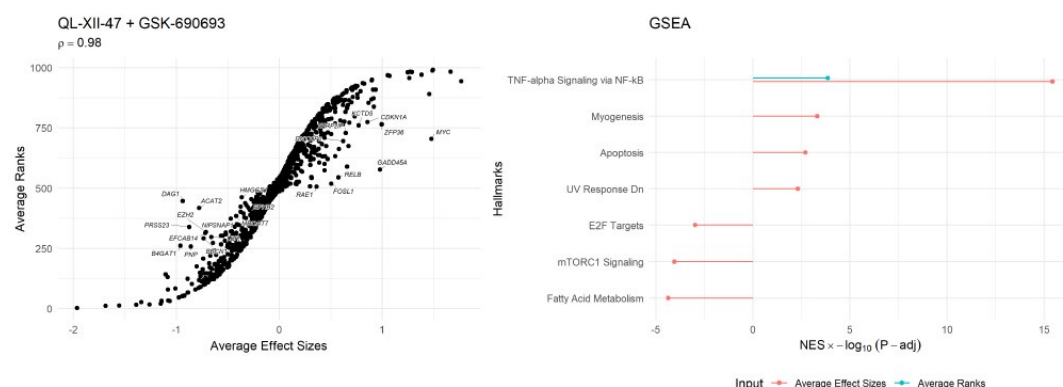
Weaknesses:

One weakness is that the authors do not compare their method to previous studies. The authors develop a drug response profile by summarizing the time points, concentrations, and cell lines. The computational challenge of creating a single gene list that represents the transcriptional response to a drug across different cell lines and treatment protocols has been previously addressed. The Prototype Ranked List (PRL) procedure, developed by Iorio and co-authors (PNAS, 2010, doi:10.1073/pnas.1000138107), uses a hierarchical majority-voting scheme to rank genes. This method generates a list of genes that are consistently overexpressed or downregulated across individual conditions, which then hold top positions in the PRL. The PRL methodology was used by Aissa and co-authors (Nature Comm 2021, doi:10.1038/s41467-021-21884-z) to analyze drug effects on selective cell populations using scRNA-seq datasets. They combined PRL with Gene Set Enrichment Analysis (GSEA), a method that compares a ranked list of genes like PRL against a specific set of genes of interest. GSEA calculates a Normalized Enrichment Score (NES), which indicates how well the genes of interest are represented among the top genes in the PRL. Compared to the method described in the current manuscript, the PRL method allows for the identification of both upregulated and downregulated transcriptional signatures relevant to the drug's effects. It also gives equal weight to each cell line's contribution to the drug's overall response signature.

The authors performed experimental validation of the top two identified drugs; however, the effect was modest. In addition, the effect on TNBC cell lines was cell-line specific as the identified drugs were effective against BT20, whose transcriptional signatures from LINCS were used for drug identification, but not against the other two cell lines analyzed. An incorrect choice of genes for the signature may result in capturing similarities tied to experimental conditions (e.g., the same cell line) rather than the drug's actual effects. This reflects the challenges faced by drug response signature methods in both selecting the appropriate subset of genes that make up the signature and managing the multiple expression profiles generated by treating different cell lines with the same drug.

We appreciate the reviewer’s thoughtful feedback and their suggestion to refer to the Prototype Ranked List (PRL) manuscript. Unfortunately, since this methodology for the PRL isn’t implemented in an open-source package, direct comparison with our approach is challenging. Nonetheless, we investigated whether using ranks would yield similar results for the most likely active drug pairs identified by *retriever*. To do this, we calculated and compared the rankings of the average effect sizes provided by *retriever*. Although the Spearman ($\rho = 0.98$) correlation coefficient was high, we observed that key genes are disadvantaged when using ranks compared to effect sizes. This difference is particularly evident in the gene set enrichment analysis, where using average ranks identified only one pathway as statistically significantly enriched. The code to replicate these analyses is available at <https://github.com/dosorio/L1000-TNBC/blob/main/Code/>.

Author response image 1.



Given the similarity in purpose between *retriever* and the PRL approach, we have added the following statement to the introduction: “Previously, this goal was approached using a majority-voting scheme to rank genes across various cell types, concentrations, and time points. This approach generates a prototype ranked list (PRL) that represents the consistent ranks of genes across several cell lines in response to a specific drug.”

Regarding the experimental validation, we believe there is a misunderstanding about the evidence we provided. We would like to clarify that we used three different TNBC cell lines: CAL120, BT20, and DU4475. It’s important to note that CAL120 and DU4475 were not included in the signature generation process. Despite this, we observed effects that exceeded the additive effects expectations, particularly in the CAL120 cell line (Figure 5, Panel F).

Reviewer 2:

Summary:

In their study, Osorio and colleagues present ‘retriever,’ an innovative computational tool designed to extract disease-specific transcriptional drug response profiles from the LINCS-L1000 project. This tool has been effectively applied to TNBC, leveraging single-cell RNA sequencing data to predict drug combinations that may effectively target the disease. The public review highlights the significant integration of extensive pharmacological data with high-resolution transcriptomic information, which enhances the potential for personalized therapeutic applications.

Strengths:

A key finding of the study is the prediction and validation of the drug combination QL-XII-47 and GSK-690693 for the treatment of TNBC. The methodology employed is robust, with a clear pathway from data analysis to experimental confirmation.

Weaknesses:

However, several issues need to be addressed. The predictive accuracy of 'retriever' is contingent upon the quality and comprehensiveness of the LINCS-L1000 and single-cell datasets utilized, which is an important caveat as these datasets may not fully capture the heterogeneity of patient responses to treatment. While the in vitro validation of the drug combinations is promising, further in vivo studies and clinical trials are necessary to establish their efficacy and safety. The applicability of these findings to other cancer types also warrants additional investigation. Expanding the application of 'retriever' to a broader range of cancer types and integrating it with clinical data will be crucial for realizing its potential in personalized medicine. Furthermore, as the study primarily focuses on kinase inhibitors, it remains to be seen how well these findings translate to other drug classes.

We thank the reviewer for their thoughtful and constructive feedback. We appreciate your insights and agree that several important considerations need to be addressed.

We recognize that the predictive accuracy of *retriever* depends on the LINCS-L1000 and single-cell datasets. These resources may not fully represent the complete range of transcriptional responses to disease and treatment across different patients. As you mentioned, this is an important limitation. However, we believe that by extrapolating the evaluation of the most likely active compound to each individual patient, we can help address this issue. This approach will provide valuable insights into which patients in the study are most likely to respond positively to treatment.

On the *in-vitro* validation of drug combinations, we agree that while promising, these results are not sufficient on their own to establish clinical efficacy. Additional *in-vivo* studies will be essential in assessing the therapeutic potential and safety of these combinations, and clinical trials will be an important next step to validate the translational impact of our findings.

Lastly, we appreciate the reviewer's comment about the focus of our study on kinase inhibitors. This result was unexpected, as we tested the full set of compounds from the LINCS-L1000 project. We agree that exploring other top candidates, including different drug classes, will be important for assessing how broadly *retriever* approach can be applied.

<https://doi.org/10.7554/eLife.102442.1.sa0>