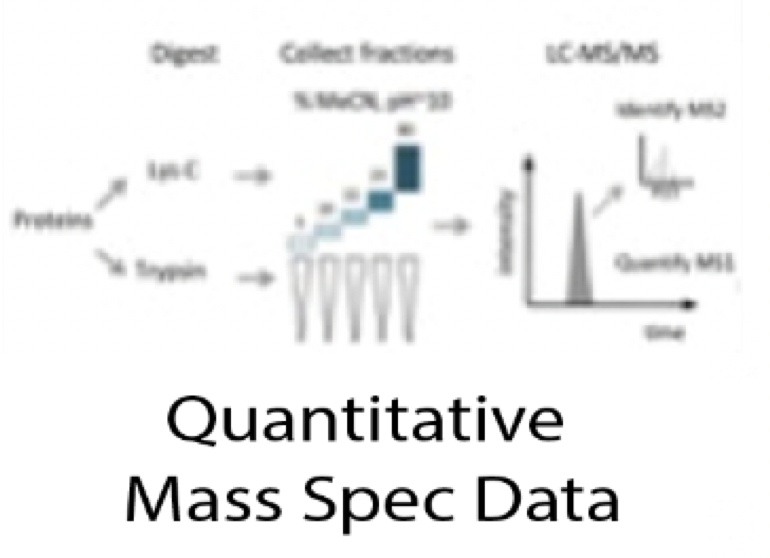
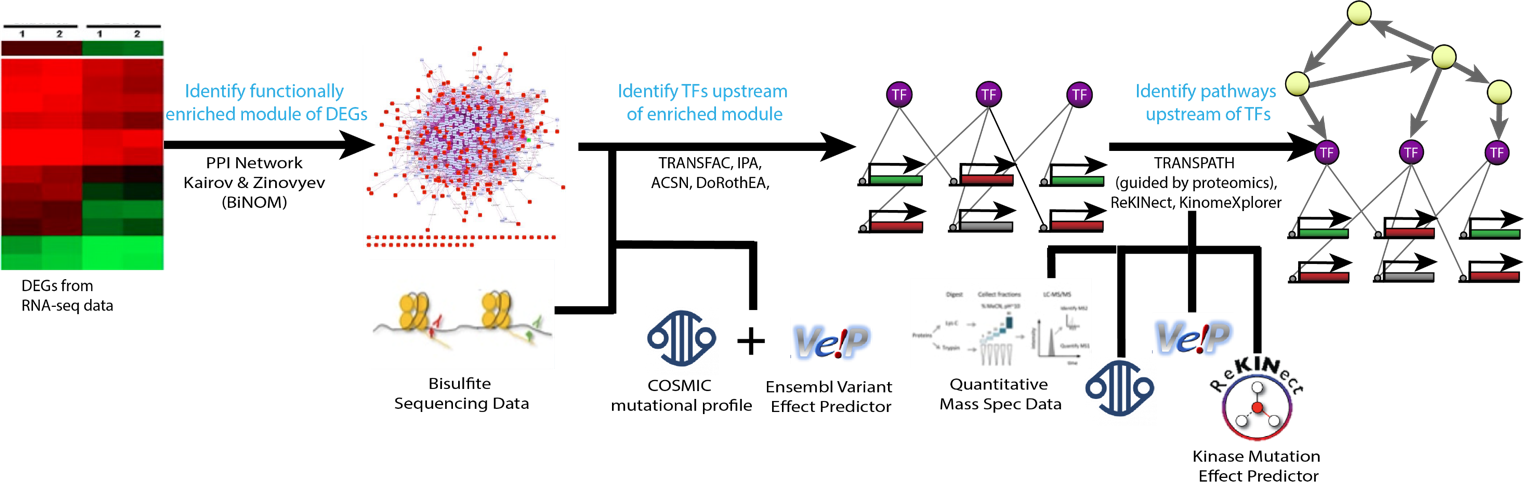
**METHODS**

**I. Reconstruction of the Tumorigenic Intracellular Signaling Network With Multi-Omics Data**



**Figure 1.** Pipeline for the reconstruction of the CL TNBC intracellular signaling network using multi-omics data.

*Data Collection*

Pair-wise RNA-seq data for the CL TNBC cell line MDA-MB-231 and the normal breast cell line MCF10A was taken from [29458327], which is deposited at Gene Expression Omnibus under the accession number GSE96860. Each cell line has two biological and two technical replicates. Bisulfite data for MDA-MB-231 was obtained from [24176112] deposited in Gene Expression Omnibus (GEO) accession number GSE42944. SNV and CNV profiles have been obtained for MDA-MB-231 from Catalogue of Somatic Mutations in Cancer (COSMIC) Cell Lines Project under sample ID COSS905960 [30371878]. Label free quantitative mass spectrometry for MDA-MB-231 was obtained from [25892236], deposited in proteomicsDB under the accession number PRDB004167.

*RNA-seq Data Processing and Differential Expression*

The raw RNA-seq files were sent to Psomagen for a differential expression analysis. The Trimmomatic program was used to drop adaptor sequences and bases with base quality lower than three from the ends, and the sliding window method was used to remove bases that do not qualify for window size 4 and mean quality 15. Reads with length shorter than 36bp were also dropped to produce trimmed data. The trimmed reads were mapped to hg19 using HISAT2, a program that handles spliced reads through Bowtie2 aligner, and known transcripts were assembled with StringTie. If more than one read count for the eight samples was zero, the gene was excluded from the analysis. This includes the removal of all genes with copy number loss as denoted by COSMIC CNV data. Reads were normalized with Relative Log Expression normalization and differential expression was calculated using a Wald test with a negative binomial distribution in the DESeq2 R package.

*Ranking Differentially Expressed Genes*

Differentially expressed genes (DEGs) were filtered for a False Discovery Rate (FDR) less than 0.05, and a P-value less than 0.001. Two different methods were used to rank the DEGs. First, by ascending p-value from the differential expression analysis. Second, a pseudo-z-score adapted from Ruthi et al. was computed for each gene and ranked by its descending absolute value [25195578]. The pseudo-z-score of the expression of node is:

Where is the average expression of gene in the cancerous cell line, is the average expression of gene in the normal cell line, and is the standard deviation of the 4 replicates of the expression of gene from the normal cell line. This score reflects the cancerous expression level of a gene compared to the distribution of expression in the normal samples, and is related to the probability that the expression level in the tumor is from the same distribution as is in the normal tissue.

*Identification of Functionally Related Differentially Expressed Genes*

While classical network construction methods use the first *n* ranked differentially expressed genes selected by a statistical approach as the first layer of the network, this pipeline uses functionally related genes as the first network layer [23055628]. This method developed by Kairov et al. allows for the consideration of biological properties when selecting which genes to include in the network as opposed to a statistical cutoff, therefore capturing functional relationships that classical methods cannot. Thus, these functionally related differentially expressed genes (FunDEGs) form a functional core between DEGs, but are not necessarily the most differentially expressed. The method stems from percolation in graph theory, where one can estimate the expected size of the largest connected component given a graph and randomly selected nodes. For some number of ranked nodes, most of them are included in the largest connected component (LCC). Thus, if the first ranked nodes () form a largest connected component larger than expected by random, their connections are non-random, and they form a functionally related core.

The method is implemented in the BiNOM Cytoscape 2.8.3 plugin with the Human Protein Reference Database (HPRD) as the interaction graph and works as follows [23453054]. First, *k* top-ranked DEGs are mapped into the HPRD. Assume () of these DEGs are found in the HPRD. The size of the largest connected component (LCC) formed by these genes is denoted Then, genes are randomly sampled from the HPRD, preserving the connectivity distribution of the ranked DEGs, and the size of the LCC, , is recorded. This step is repeated 10,000 times and the mean size of the LCCs () is calculated. The percolation score, , is computed for :

This process is repeated for values of from 50 to 1000 in increments of 50. The optimal () is chosen as the smallest such that decreases as increases for [23055628]. This analysis was run separately on the DEGs ranked by p-value and the DEGs ranked by pseudo z-score. Housekeeping genes, except for VIM because it is a marker of EMT, were filtered from the functionally related core and the resultant LCC was selected as the set of First Order FunDEGs [19534766].

After identifying First Order FunDEGs, BiNOM was used to identify Second Order FunDEGs. This step identifies nodes in the HPRD that are not included in the First Order FunDEGs, but have known interactions with those genes. Thus, the genes in the Second Order FunDEGs need not be differentially expressed, but are functionally related to those that are, and could play important roles in tumorigenesis. Housekeeping genes and genes that are not expressed in MDA-MB-231 were removed from this set of genes and the LCC was selected as the set of Second Order FunDEGs.

To determine which set of genes to use for network construction (The first *k* ranked DEGs, First Order FunDEGs, or Second Order FunDEGs), a weighted sums metric was employed. This compares the contents of each gene set to Claduin-Low related genes, Breast Diseases Ontology genes from DOLite in the GeneAnswers R package, and hallmarks of cancer associated with EMT and innate immune response from the atlas of cancer signaling network as these hallmarks are claudin-low specific [28249905, 19478018, 26192618]. The list with the highest weighted sum was selected as the first layer of the network.

*Identification of Transcription Factors*

The next component of the intracellular signaling network is transcription factors (TFs) that transcribe the FunDEGs. These were identified with the Analyze Promotors workflow using TRANSFAC and MatchTM in GeneXplain. TRANSFAC is a manually curated database of eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles [10592259].It stores DNA binding patterns as positional weight matrices (PWMs) that can be used to identify putative transcription factor binding sites (TFBS) with the MatchTM algorithm [12824369]. The algorithm takes a DNA sequences as input and uses two scores- the matrix similarity score (MSS) and the core similarity score (CSS) - to measure the quality of the match between the sequence and the PWM. Matches that have a MSS and CSS higher than pre-defined cutoff values are reported as putative TFBSs [12824369]. The Analyze Promotors workflow compares putative TFBSs within a specified promotor window size between a “Yes” set and a “No” set of genes. The yes-no ratio is calculated for each TF to indicate the proportion of targets of the TF regulates in the “Yes” set compared to the “No” set.

For this analysis, PWMs for human TF binding sites with a p-value<0.001 were used. The “Yes” set was the list of FunDEGs and the “No” set was the list of expressed housekeeping genes. The workflow was run with five different promotor window start sights: 500bp, 1000bp, 1500bp, 2000bp, and 2500bp upstream of the transcription start sight. The results were filtered for TFs with a yes/no ratio > 1.5 and a p-value less than 0.05. Additionally, Ingenuity Pathway Analysis (IPA) upstream regulator analysis was used to identify TF – FunDEG interactions that have been experimentally validated in any tissue or breast cancer sample. IPA contains a manually curated database of experimentally validated cause-effect relationships between genes, proteins, chemical compounds, and microRNA. The upstream regulator analysis algorithm takes a statistical approach to identify regulators whose network connections to FunDEGs are unlikely to occur in a random model [24336805]. These results were filtered for interactions with a p-value less than 0.001. The intersection of the TRANSFAC and IPA results at each promotor window size were filtered against gene and protein expression data for MDA-MB-231 and unmethylated cpg regions from bisulfite data. A weighted sums metric for the lists of TFs from each window size was calculated the same way as previously described, and coverage of FunDEGs was checked. The TF list with the highest weighted sum covering all FunDEGs was selected.

Although TRANSFAC provides putative TFBS, it does not provide information regarding the sign of the interaction between a TF and FunDEG. Those interactions that were not included in the IPA output were mapped into OmniPath, a network of high quality, manually curated signaling pathways using Cytoscape [31886476]. Only pathways with confidence level A and B were considered. The sign of the relationship for any remaining TF – FunDEG interactions that were not in OmniPath was inferred using Pearson Correlation.

Mutations in regulatory regions can influence TF binding affinity. To identify mutations in regulatory regions of the FunDEGs, the COSMIC SNV data for MDA-MB-231 was filtered for SNVs in regulatory regions as determined by the Ensembl Variant Effect Predictor (VEP) [27268795]. If there are mutations in regulatory regions, their impact on TF-binding affinity is predicted as described in [28333948]. In short, each SNV in a regulatory region is analyzed for its ability to modulate the sequence to PWM match score. SNVs that increase the score compared to the wildtype sequence are potential gain-of-binding site events, while SNVs that decrease the score compared to the wildtype sequence are potential loss-of-binding site events. This can be done in GeneXplain using the Mutation effect on sites analysis workflow. Necessary modifications to TF-FunDEG interactions are modified manually according to an increase or decrease in binding affinity.

*Identification of Master Regulators*

The final layer of the network consists of master regulators (MRs) – genes, proteins, or complexes that regulate the TFs in the network. Pathways in which the TFs participate were identified using the Master Regulators with Context Genes with Weights pipeline in GeneXplain and the manually curated TRANSPATH database [12519957]. This workflow uses a specified radius and the Dijkstra algorithm for shortest paths on the TRANSPATH database to identify upstream molecules regulating the TFs. Proteomics data for MDA-MB-231 was used as context to guide the shortest paths algorithm towards nodes with protein expression. The workflow weights the context genes based off of the average protein intensity for each gene by decreasing the cost of the edges around genes with higher levels of protein expression, thereby attracting the algorithm to those molecules [29900117]. The analysis was run with radii ranging from 5 to 10, indicating how far upstream from the TFs to search for MRs. The smallest radius such that all the TFs have an upstream MR was chosen. Resultant master regulators were filtered for a z-score > 1.5, an FDR < 0.05, and to include only those that have protein expression. Nodes in MR pathways without gene expression data were removed from the network so long as their removal did not create a source node.

SNVs have the capability to inactivate, activate, or modify the specificity of a protein’s kinase domain which may cause profound signaling changes within a cell. To determine the impact of kinase mutations in our network, the Ensembl VEP ProteinSeqs plugin was used to obtain reference (wild-type) and mutated protein transcripts. Then, the ReKINect software was used to predict and classify kinase modifying mutations in the SNV data for MDA-MB-231 [26388441]. ReKINect employs a database of all known human kinase domains, 111 SH2 domains, and 149,838 phosphorylation sites and will predict the functional impact of the SNV. For mutations affecting protein kinase function, the downstream targets of the mutated kinases were determined with KINSpect, using the ReKINect database [26388442]. KinomeXplorer was used to investigate upstream kinase specificity changes for mutated proteins [24874572]. The identified mutated pathways are added to or removed from the network manually, as they are not included in signaling databases.

*Addition/Removal of Mutated Pathways*

Mutations are a critical factor in the development of cancer. After considering the effect of mutations in TF binding regions and kinase signaling pathways, the effect of any other mutated network node must be considered manually. Ensembl Variant Effect Predictor (VEP) was used to provide information regarding the location and consequence of each SNV [27268795]. Mutations that lie in protein coding regions were filtered to include only those that modify the protein amino acid sequence (missense, splice-site variation, frameshift, deletion, stop gained). A literature curation of the gene variants that cause protein transcript alterations was done to explore their impact on TNBC tumorigenesis.If a gene variant causes gain of function activities that affect a specific TNBC pathway, the interaction and pathway were added to the network manually as appropriate, since they are not included in signaling databases.