# Interactive Visualization of METABRIC Breast Cancer Genomic and Clinical Data Using Dash

Ebubechukwu Jack-Davies
Department of Computing
University of connecticut
Email: ebubejackdavies@gmail.com

***Abstract*—Breast cancer is a heterogeneous disease driven by a complex interplay of clinical, genomic, and treatment factors. Modern datasets such as METABRIC contain hundreds of genomic features per patient, making it difficult to explore patterns and generate hypotheses using static tables or single-purpose plots [4]. This work presents an interactive web-based dashboard, implemented in Python using Dash and Plotly, for exploring clinical, mutation, and mRNA expression data from the METABRIC breast cancer cohort. The dashboard integrates high-level summaries, mutation profiles, mRNA expression analysis, and co-occurrence/co-expression networks into a unified interface. Techniques such as binary mutation encoding, correlation analysis, principal component analysis (PCA), clustering, and Kaplan–Meier survival curves are combined to support both exploratory analysis and interpretation of high-dimensional data [5], [6], [8]. We describe the dataset structure, analysis pipeline, and visualization design choices, and we discuss how this tool can be extended to include additional data types and more advanced models. The system demonstrates how interactive visualization can make complex genomic datasets more accessible to researchers and clinicians [12].**

***Index Terms*—Breast cancer, METABRIC, data visualization, Dash, mutation analysis, mRNA expression, PCA, survival analysis.**

## I. Introduction

Breast cancer remains one of the most prevalent and deadly cancers worldwide. Modern high-throughput technologies have made it possible to measure hundreds of genomic features per tumor, including gene-level mutation status and mRNA expression profiles, in addition to traditional clinical variables such as age, tumor size, grade, and receptor status [1], [2]. While these rich datasets enable powerful downstream modeling, they are challenging to explore and interpret without appropriate visualization and interaction tools.

In this project, we design and implement an interactive dashboard for the well-known METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset [4]. The goal is to allow a user to:

- Summarize clinical characteristics of the cohort;
- Explore distributions of common driver gene mutations [3], [11];
- Inspect mRNA expression patterns and dimensionality-reduced structure;
- Examine co-mutation and co-expression patterns between genes [10];
- Relate genomic features to survival outcomes via Kaplan–Meier curves [8].

The dashboard is built using Python, Pandas, and Plotly Dash, and is structured into multiple tabs (Overview, Mutation, mRNA Expression, and Co-occurrence/Co-expression). Each tab focuses on a different subset of questions while reusing a consistent data model.

The contributions of this work are:

1) A unified pipeline that preprocesses METABRIC clinical, mutation, and mRNA data into a form suitable for interactive visualization;
2) A set of visualization components for summarizing high-dimensional mutation and mRNA expression data, including PCA, clustering, heatmaps, and network views [5], [6], [10];
3) An assessment of the strengths and limitations of these techniques when applied to high-dimensional genomic data, and a discussion of how the dashboard could be improved in future iterations.

The rest of the paper is organized as follows. Section II describes the data and its structure. Section III explains the technologies and analysis techniques used. Section IV presents the analysis results and visualizations, including references to example figures from the dashboard. Section V provides a discussion of the findings and possible extensions. Section VI concludes the paper and outlines future directions.

## II. Data

The dashboard is built around the METABRIC RNA and mutation dataset, which combines clinical variables, mutation calls, and gene-level expression measurements for approximately 1900 patients and several hundred columns [4].

### A. Clinical Variables

The first part of the dataset consists of clinical and pathological variables, including:

- **Patient identifiers:** e.g., `patient_id`;
- **Demographics and tumor characteristics:** age at diagnosis (`age_at_diagnosis`), tumor size (`tumor_size`), lymph node involvement (`lymph_nodes_examined_positive`), histologic grade (`neoplasm_histologic_grade`), and breast surgery type (`type_of_breast_surgery`);
- **Receptor and subtype information:** estrogen receptor status (`er_status_measured_by_ihc`, `er_status`),

HER2 status (her2_status_measured_by_snp6, her2_status), and molecular subtype (pam50_+_claudin-low_subtype), as well as integrative cluster labels (integrative_cluster) [1], [2];

- **Treatment variables:** binary indicators for chemotherapy, hormone therapy, and radiotherapy;
- **Outcome variables:** overall survival status (overall_survival), survival time in months (overall_survival_months), and in some cases cancer-specific event information.

These variables are typically categorical or low-dimensional numeric and are used in the Overview and Mutation tabs for grouping and stratification.

### B. Mutation Data

Mutation features are encoded in columns whose names end with the suffix _mut. There are roughly 170 such mutation columns, each corresponding to a gene (e.g., tp53_mut, pik3ca_mut, gata3_mut, cdh1_mut) [3], [11]. The raw values can be represented in different string or numeric forms, so the pipeline standardizes them by:

1) Casting values to strings and stripping whitespace;
2) Treating a defined sentinel set (e.g., "0", "0.0", empty strings, "nan", "None") as absence of mutation;
3) Encoding all other values as 1 (mutation present).

This results in a binary mutation matrix of shape $(N_{\text{patients}}, N_{\text{mutation genes}})$ that is used for computing mutation frequencies, generating mutation heatmaps, performing co-mutation analysis, and constructing network visualizations.

### C. mRNA Expression Data

The dataset also contains hundreds of continuous expression features for genes including well-known breast cancer genes such as *BRCA1*, *BRCA2*, *TP53*, *GATA3*, and many others. These expression features are typically stored as z-scores relative to a reference population [1], [2].

In practice, we treat this block as a high-dimensional continuous matrix, with each row corresponding to a patient and each column corresponding to the expression of one gene. The dashboard uses this matrix for:

- Visualizing distributions of selected genes;
- Computing correlation matrices between genes;
- Running PCA to reduce dimensionality [5];
- Running K-means clustering in the PCA space [6].

### D. Survival Information

Survival analysis in the dashboard is powered by two key columns:

- overall_survival_months: survival or follow-up time;
- overall_survival: indicators of whether the patient died during the follow-up period (derived from textual encodings such as "dead" or "deceased").

The code constructs survival event variables by mapping several textual encodings of death to a binary event indicator, with remaining values treated as censored [8], [9].

## III. Technologies and Methods

### A. Technology Stack

The dashboard is implemented in Python with the following components:

- **Dash:** provides the web application framework, layout system, and callback mechanism for interactivity;
- **Plotly:** used to generate interactive plots, including bar charts, box plots, heatmaps, scatter plots, and network-like visualizations;
- **Pandas and NumPy:** used for data loading, cleaning, and transformation into numeric matrices;
- **Scikit-learn:** used for standardization, PCA, and K-means clustering in the mRNA Expression tab [5], [6];
- **Lifelines:** used to fit Kaplan–Meier survival curves and produce survival plots in the Overview and Mutation tabs.

The application code is organized into modular files such as overview.py, mutation.py, mrna.py, and co.py, each responsible for registering callbacks and assembling the visualizations for a specific tab.

### B. Data Preprocessing

Key preprocessing steps include:

- **Mutation binarization:** as described in Section II, mutation columns are converted into clean 0/1 matrices using a sentinel set to capture missing or "no mutation" values;
- **Handling missing values:** for expression and clinical variables, missing values are either ignored (e.g., when computing correlations or PCA) or excluded on a per-plot basis, rather than imputed, to maintain interpretability;
- **Standardization for PCA and clustering:** because gene expression scales can differ, the mRNA matrix is standardized to zero mean and unit variance before applying PCA and clustering;
- **Survival event construction:** if a more specific cancer-event field is present, it is used; otherwise, a derived event indicator from overall_survival is constructed [8].

### C. Dimensionality Reduction and Clustering

To deal with the high dimensionality of the mRNA expression data, the dashboard uses PCA [5]:

- A user-selectable number of principal components is computed;
- A scatter plot shows patients in the space of the first two (or more) components;
- A bar chart reports the variance explained by each principal component.

After obtaining PCA scores for each patient, K-means clustering is applied in the reduced space (using up to a fixed number of components). The user can select the number of clusters $k$, and the dashboard:

- Assigns each patient to a cluster;
- Colors the PCA scatter plot by cluster label;
- Reports a silhouette score to give a rough indication of cluster separation [6].

This approach is chosen over clustering in the full gene space because PCA reduces noise and redundancy, makes visualization (e.g., 2D scatter plots) feasible, and provides a compromise between explainability (via variance explained) and interpretability.

### D. Correlation, Co-occurrence, and Networks

The dashboard also computes:

- Mutation correlation matrices based on binary mutation vectors;
- mRNA expression correlation matrices between genes;
- Co-mutation and co-expression networks, where nodes are genes and edges connect genes whose pairwise correlation magnitude exceeds a user-defined threshold [10].

These networks are visualized with Plotly, using node-link diagrams where node size or color can reflect mutation frequency or other gene-level summaries, and edge style can encode sign and magnitude of correlation.

### E. Survival Analysis

The survival analysis functions use Kaplan–Meier estimators and related methods [8], [9] to:

- Plot overall survival for the full cohort in the Overview tab;
- Plot stratified survival curves for subsets defined by mutation status (e.g., mutated vs. wild-type *GATA3*) in the Mutation tab.

This allows the dashboard to visually compare survival distributions between genomic subgroups.

## IV. Analysis Results and Visualizations

In this section, we describe the major visual components of the dashboard and the kinds of patterns they reveal. Screenshots from the running Dash app can be inserted into the figures referenced below.

### A. Overview Tab

The Overview tab provides high-level context for the cohort, including:

- Summary cards for key statistics such as number of patients, median age, median tumor size, and distribution of subtypes;
- Distributions of key clinical variables (e.g., PAM50 subtypes, ER status, treatment frequencies);
- A Kaplan–Meier curve showing overall survival for the cohort [8].

### B. Mutation Tab

The Mutation tab focuses on gene-level mutation patterns. It includes:

- A bar chart of the top $N$ most frequently mutated genes, computed from the binary mutation matrix;
- Stratified plots showing how mutation frequencies differ by PAM50 subtype, ER status, or treatment categories;
- Survival curves comparing patients with and without a particular mutation.

Mutations in genes such as *TP53*, *PIK3CA*, and *GATA3* typically appear among the most frequent, and the visualizations allow users to see how these mutations are distributed across receptor subtypes and treatment groups [3], [11].

### C. mRNA Expression Tab

The mRNA tab enables exploration of the high-dimensional gene expression data:

- A single-gene expression view using box or violin plots, optionally stratified by subtype or receptor status;
- A correlation heatmap for a selectable subset of genes;
- A PCA variance plot showing how much variance each principal component captures;
- A 2D PCA scatter plot where each point represents a patient;
- K-means clustering in PCA space, visualized by coloring points according to cluster labels [5], [6].

### D. Co-mutation and Co-expression Tab

The Co-occurrence tab integrates both mutation and expression relationships:

- A co-mutation heatmap showing pairwise correlations between mutation patterns of the top $N$ most frequently mutated genes;
- A co-mutation network where nodes are genes and edges represent strong pairwise correlations;
- A co-expression heatmap and co-expression network for mRNA expression correlations [10], [12].

## V. Discussion

The dashboard demonstrates how interactive visualization can help manage and interpret a dataset with hundreds of genomic features and multiple clinical variables. By organizing the interface into logical tabs and reusing a consistent data model, users can move from high-level clinical summaries to focused mutation and expression analyses [12].

### A. Handling High Dimensionality

One of the central challenges is the high dimensionality of the mRNA expression data. Directly visualizing all features at once is not practical. The combination of PCA, correlation heatmaps, and clustering addresses this problem by:

- Reducing dimensionality to a small set of principal components with interpretable variance explained;
- Allowing users to focus on a subset of genes (e.g., top mutated genes or genes of interest) for correlation analysis;
- Providing unsupervised clusters that may correspond to biological subgroups, which can be compared to existing labels such as PAM50 subtypes [1], [2].

However, dimensionality reduction introduces its own trade-offs. PCA is linear and may not fully capture non-linear structure. Also, clusters found in PCA space may be sensitive to preprocessing and gene selection. Non-linear methods such as t-SNE or UMAP could better capture complex manifolds in expression data but are harder to interpret and more computationally demanding [7].

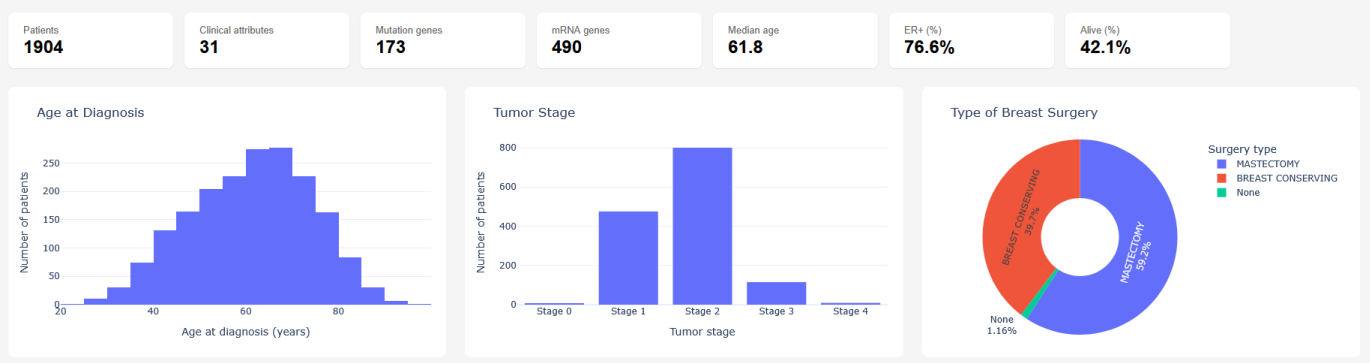


Fig. 1. Overview dashboard with cohort summary cards and surgery distributions.

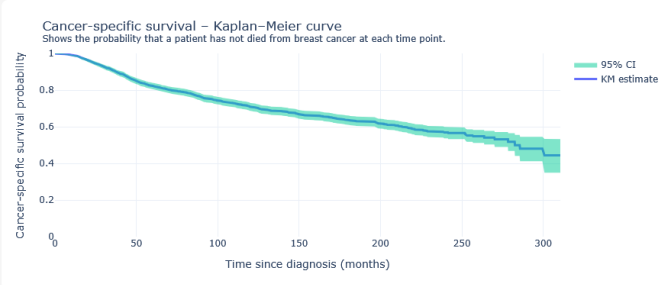


Fig. 2. Kaplan–Meier curve of overall survival for the METABRIC cohort.

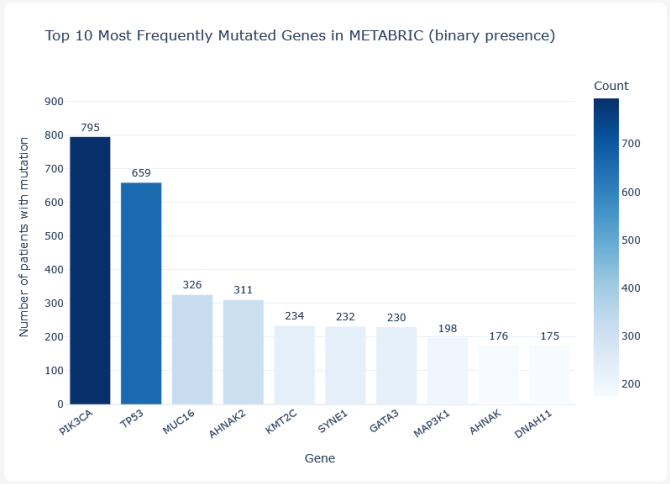


Fig. 3. Top 10 most frequently mutated genes in the METABRIC cohort.

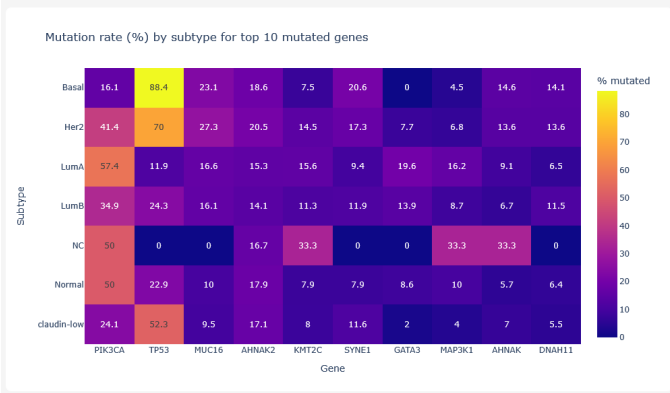


Fig. 4. Heatmap of the top 10 most mutated genes' subtypes.

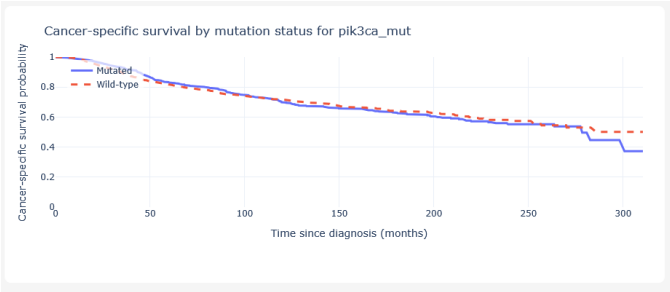


Fig. 5. Kaplan–Meier survival curves comparing patients with and without a selected mutation.

## B. Interpretation of Mutation and Expression Patterns

The mutation and co-mutation visualizations highlight genes like *TP53*, *PIK3CA*, and *GATA3*, which are known to be important in breast cancer [3], [11]. The dashboard allows users to see how often these genes are mutated, how their mutation patterns co-occur across patients, and whether mutations correlate with differences in survival.

The mRNA expression views complement this by showing whether strongly co-mutated genes are also co-expressed, or whether expression relationships differ from mutation patterns. For example, a pair of genes may frequently co-mutate but not show strong co-expression, which could suggest different biological mechanisms. Co-expression networks similar to weighted correlation network analysis (WGCNA) can reveal gene modules and hub genes that may be biologically meaningful [10].

## C. Limitations and Future Work

Despite its strengths, the current system has several limitations:

- The pipeline uses fixed methods (PCA, K-means, correlation) and does not expose alternative dimensionality reduction techniques such as t-SNE or UMAP [7];

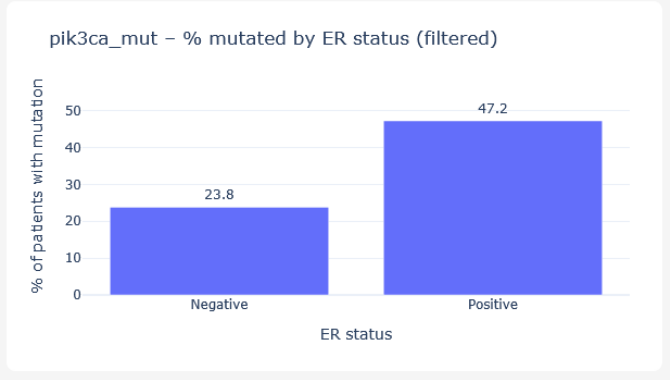


Fig. 6. Mutation frequencies stratified by ER status (e.g., ER-positive vs. ER-negative).

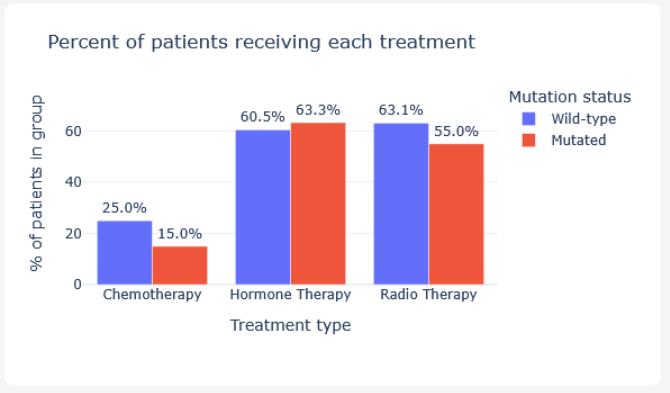


Fig. 7. Distributions of chemotherapy, hormone therapy, and radiotherapy use in the cohort, stratified by mutation status.

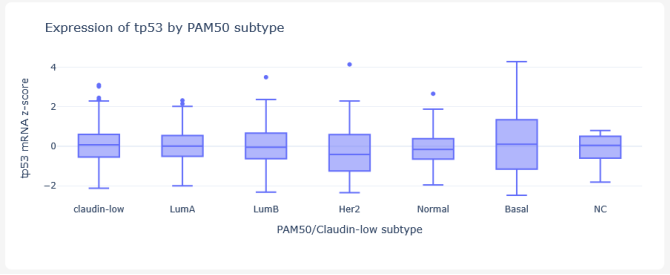


Fig. 8. Expression of a single gene by subtype.

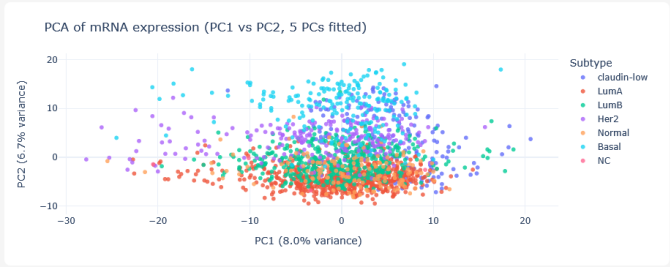


Fig. 9. PCA scatter plot of mRNA expression for patients.

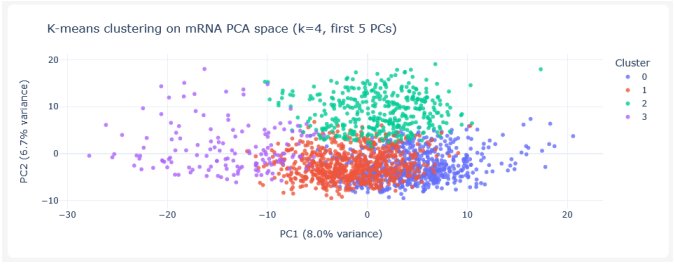


Fig. 10. PCA scatter plot of mRNA expression for patients colored by K-means cluster.

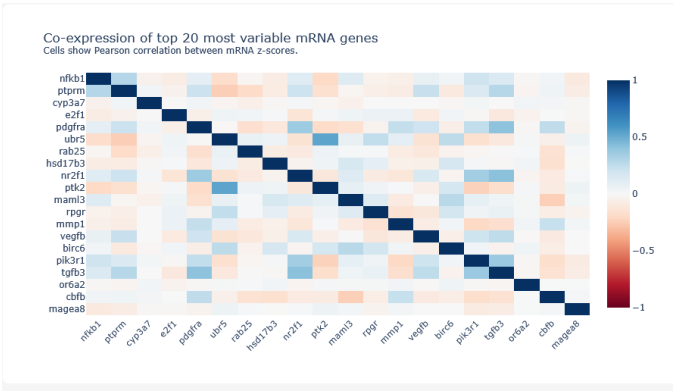


Fig. 11. Correlation heatmap of selected mRNA expression genes.

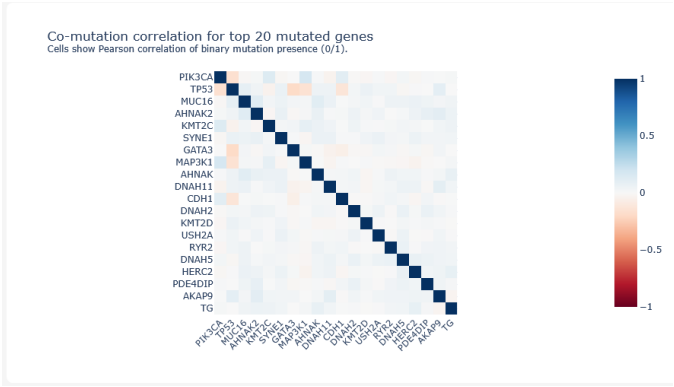


Fig. 12. Co-mutation heatmap for the top $N$ mutated genes.

- The dashboard focuses on exploration rather than predictive modeling; no survival regression or risk modeling is implemented [9];
- Missing values are largely handled via dropping, which may reduce sample size for some analyses;
- Computing correlation matrices and clustering on the fly for large gene subsets can be computationally expensive.

If the project were restarted or extended, several improvements could be made:

- Add options for non-linear methods (e.g., t-SNE, UMAP) to better capture complex structure in expression data [7];
- Incorporate Cox proportional hazards models or other survival models to quantify associations between genomic

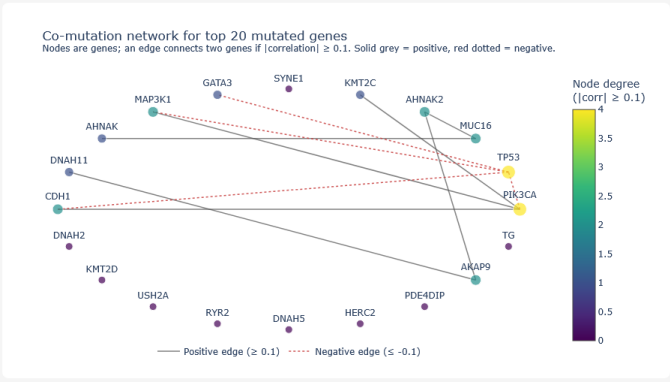


Fig. 13. Co-mutation network; nodes are genes, edges represent strong pairwise correlations.

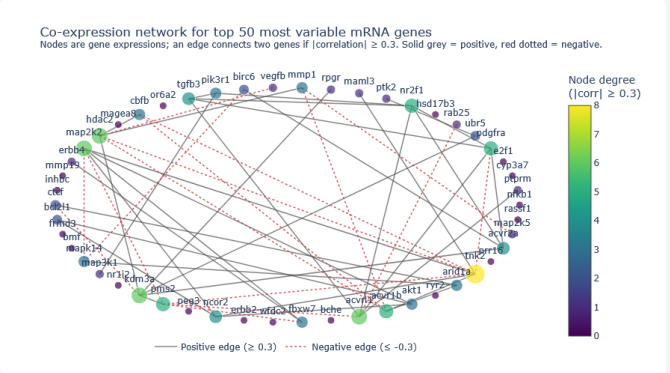


Fig. 14. Co-expression network for a subset of mRNA genes showing correlated expression patterns.

features and outcomes [9];

- Allow users to select genes based on external annotations (e.g., pathway membership) or upload their own gene lists;
- Overlay treatment regimens, response information, or additional clinical endpoints onto PCA and clustering plots;
- Add export features for subsets of patients or genes for downstream analysis.

## VI. Conclusion

This work presents an interactive Dash-based dashboard for exploring clinical, mutation, and mRNA expression data from the METABRIC breast cancer cohort. By integrating multiple views—overview summaries, mutation distributions, mRNA expression analysis, PCA, clustering, and co-mutation/co-expression networks—the system helps users navigate a high-dimensional dataset and generate hypotheses about underlying biological patterns [1], [2], [4].

The project demonstrates that even with relatively simple modeling techniques, carefully designed visualizations can reveal complex relationships in genomic data. Future work could incorporate more advanced models, additional datasets, and more flexible interaction patterns to further support clinical and translational research [12].

## References


[1] C. M. Perou et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, 2000.

[2] T. Sørlie et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *PNAS*, vol. 98, no. 19, pp. 10869–10874, 2001.

[3] S. Banerji et al., "Sequence analysis of mutations and translocations across breast cancer subtypes," *Nature*, vol. 486, pp. 405–409, 2012.

[4] C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, pp. 346–352, 2012.

[5] I. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, 2016.

[6] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[7] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[8] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *JASA*, vol. 53, no. 282, pp. 457–481, 1958.

[9] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society, Series B*, vol. 34, pp. 187–220, 1972.

[10] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, 2008.

[11] M. S. Lawrence et al., "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, pp. 214–218, 2013.

[12] A. Lex et al., "StratomeX: Visual analysis of cancer subtypes and their genomic correlates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2546–2555, 2013.