Data Mining Diverse Compendia of Triple Negative Breast Cancer Samples for Improved Tumor Subtyping

Jacob M. Luber[1,2], Joan Malcom[1], Adam Lavertu[1], Kwangbom Choi[1], Matthew A. Hibbs[1,2], Joel Graber[1], Carol J. Bult[1]
[1]The Jackson Laboratory, Bar Harbor, Maine 04609 USA
[2]Department of Computer Science, Trinity University, San Antonio, Texas 78212 USA

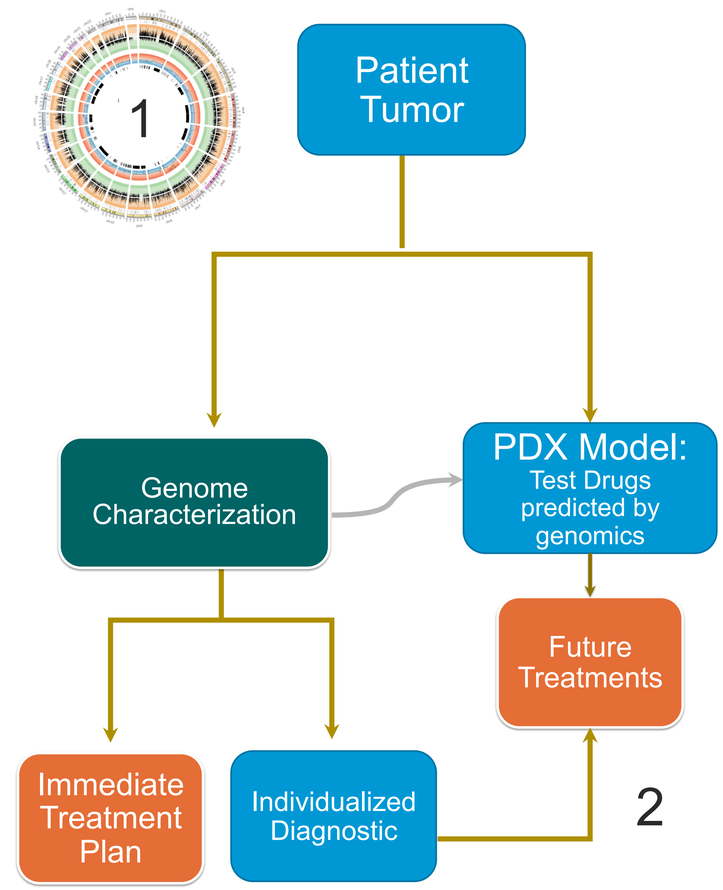
TRINITY UNIVERSITY
FOUNDED 1869

The Jackson Laboratory
Leading the search for tomorrow's cures

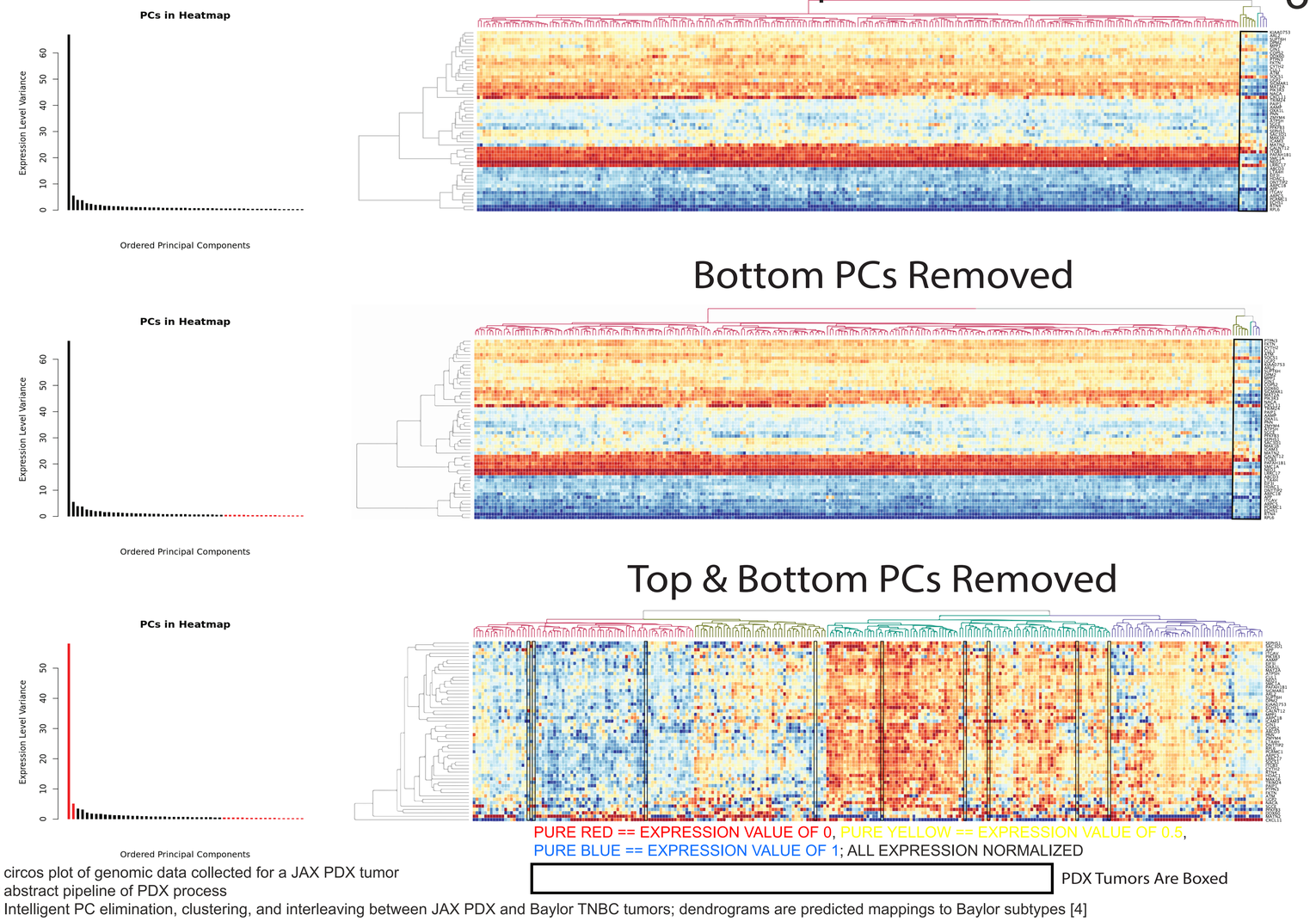
ABSTRACT

Integrating cancer genomic data from diverse experimental sources (e.g. in vitro, in vivo, or ex vivo biological samples) and platforms (e.g. microarrays, RNA-seq) poses specific technical challenges due to batch effects, tumor heterogeneity, and ambiguous subtyping. Triple Negative Breast Cancer (TNBC) is especially difficult to treat because of rapid propagation and clonal variance. A potential avenue for improving TNBC treatment is to determine more finely grained subtypes of TNBCs that respond differently to treatments. While a large corpus of TNBC data is available from heterogeneous data sources, conventional unsupervised analysis approaches cannot distinguish between noise related to experimental and platform disparities and biologically relevant signal. We have built a computational framework that uses statistically sound data mining techniques to analyze and compare gene expression signatures gathered from multiple studies of TNBC including patient derived xenograft (PDX) models, cell lines, and primary patient biopsies. By utilizing these methods our pipeline elucidates better potential subtyping on diverse compendia of TNBC data. Our results demonstrate the challenges of heterogeneous data integration, but also reveal potential insights into techniques that can allow comparisons across samples generated by multiple labs using different experimental protocols and data collection platforms. In addition to comparisons among TNBC subtypes, we were also able to cluster and compare TNBC data with other tumor types from 462 PDX models to assess more global patterns and trends, possibly related to tumor microenvironment and the PDX process. Continuing work focuses on correlating TNBC subtypes with tumor response profiles for different therapeutics.

OVERVIEW OF THE PDX PROCESS



EFFECT OF INTELLIGENT PRINCIPAL COMPONENT ELIMINATION ON CLUSTERING



PURE RED == EXPRESSION VALUE OF 0, PURE YELLOW == EXPRESSION VALUE OF 0.5,
PURE BLUE == EXPRESSION VALUE OF 1; ALL EXPRESSION NORMALIZED

PDX Tumors Are Boxed

1 circos plot of genomic data collected for a JAX PDX tumor
2 abstract pipeline of PDX process
3 Intelligent PC elimination, clustering, and interleaving between JAX PDX and Baylor TNBC tumors; dendrograms are predicted mappings to Baylor subtypes [4]

TRY OUR PDX PC ELIMATION ON ANY SET OF GENES:

jacobluber.com/tnbc

HOW IT WORKS

(1) We spin up an Amazon EC2 Instance geographically close to your location (according to IP).
(2) You enter a set of custom human genes (or select a pre-curated set), and two variance elimination bounds.
(3) Our back end processes uses your input gene set to pull expression data, remaps probes, normalizes each dataset, then maps gene names to GRCh38.
(4) Our back end takes this data, performs SVD on it---removing principal components according to your selected variance contribution bounds, then hierarchically clusters a reprojection of the data.
(5) Our front end renders this reprojection with D3.js as an interactive heatmap.

DATA WRANGLING

Utilizing the canonical singular value matrix decomposition (SVD) $Xm*n=Um*n\Sigma n*nV^{T}n*n$ we initially explored established methods for signal balancing in genomic data previously applied to simpler species where correlations are calculated between genes' coefficients in U rather than re-projected to an approximation of X [1]. This approach proved unwieldy for TNBC data, but it did elucidate that PCs contributing the highest variance likely significantly correspond (in part) to host species, microenvironment, and immune response--removing these as well as the lowest ranked PCs improves signal. Variance contribution bounds for PCs to be eliminated are provided. The NP-Hard problem of ideal gene subset curation was dealt with by expanding gene sets grouped as tumor milestones in the KEGG breast cancer pathway [2] with predicted functional relationships in mammary epithelium from GIANT [3]. Achieving interleaving (as some of the curated gene sets show) yields the possibility of mapping drug treatment response from PDX samples to similar human tumors, mapping subtype categorizations from human tumors to PDX samples, and categorizing both in terms of clusterings related to post transcriptional regulation mapped out in KEGG [2].

REFERENCES

[1] Hibbs et. al., Exploring the functional landscape of gene expression: directed search of large microarray compendia, Bioinformatics, 2007.
[2] KEGG microRNAs in Cancer.
[3] Greene et. al., Understanding multicellular function and disease with human tissue-specific networks.
[4] Burstein et. al., Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer, Clin Cancer Res, 2015.