Interactive Meta-analysis Framework for Biomarker Identification in Breast Cancer subtypes.

Milton Yutaka Nishiyama Junior, Ana Claudia Oliveira Carreira, Nathan de Oliveira Nunes

Instituto Butantan, Instituto Butantan

Abstract

Breast cancer is the most diagnosed type of cancer among women in the world, classified into 4 subtypes (HER2, LumA, LumB, Triple negative-TNBC), with TNBC being the most aggressive and lethal. We have developed an interactive meta-analysis framework to identify new common and exclusive molecular markers and signaling pathways in each study of breast cancer subtypes and in non-tumoral (MCF10A) and tumoral (Hs578T) cell lines with alteration of the expression of the CD90/Thy-1 gene. The interactive framework based on the shiny R package provides an intuitive and easy approach to compare multiple studies using gene expressions profiles (RNA-seq) in different conditions and able to parallelize the high-throughput data analysis. This framework is based on the steps and methods: M1: Load of data, followed by inter (RLE / TMM / UQ) and Intra (ComBat) sample normalization. M2: Classification of samples in subtypes/clusters. M3: Differentially expressed genes (DEGs) identification (edgeR), M4: Qualitative (ReactomePA) and Quantitative (FGSEA) pathway enrichment analysis, M5: Relevance and significance for genes (Ranking Product (RP)) and Pathways (Fisher / Stouffer methods) and M6: Connectivity between studies (TF-IDF and average precision). There are two meta-analysis approaches, one based on gene expression by M3 and p-value from enriched pathways by M4, followed by the meta-analysis methods in M5, and establishing the similarity matrix in M6, to measure the connectivity and average precision between the studies. The best normalization methods were established to correct bias and errors within and between samples. The classification approaches allowed us to cluster the samples in a biologically meaningful way. The DEGs or enrichment analysis provided genes/pathways information for further downstream meta-analysis. The methods of RP for the genes and Fisher for the pathways increased the identification in more than 80% of the TNBC canonical genes and pathways and confirmed the expected associations between the studies. Important and already described genes in the literature such as CGA, PLUNC and SMR3B have been identified in TNBC by our approach. Among the highlighted pathways, we found ECM and GPCR as one of most relevant between all breast cancer subtypes. We have developed a framework that can be applied to any set of studies, integrating a robust set of statistical methods and bioinformatics approaches in an user-friendly visualization tool aiming to identify the common, exclusive and promising genes and pathways for our studies (MCF10A/CD90+ and Hs578T/shCD90) compared to Breast cancer subtypes.

Funding: CNPQ

Link to Video: