Glioblastoma Multiforme Subtype Classification: Integrated Analysis of Protein and Gene Expression Data

Durairaj Renu¹, Vadiraja Bhat², Mona Al-Gizawiy³, Carolina B. Livi², Stephen Madden², Christine A. Miller², Michael Rosenberg², Kathleen Schmainda³, Prateek Singh¹, Pramila Tata¹, Shama P. Mirza³

¹Strand Life Sciences, Bangalore, India; ²Agilent Technologies, Inc. Santa Clara, CA; ³Medical College of Wisconsin, Milwaukee, WI
Overview

- Integrated analysis of genomics and proteomics data used to improve on the molecular signatures for tumor classification
- Data analysis approach to identify signature genes and proteins for stratification of Glioblastoma Multiforme subtypes

Introduction

Stratifying cancer patients based on their molecular characteristics helps to identify the effective therapy for each patient. Glioblastoma Multiforme (GBM) is the most common and highly malignant brain tumor in adults. Studies by The Cancer Genome Atlas (TCGA) Research Network identified four distinct molecular subtypes for GBM based on gene expression profiling, classifying tumors into Proneural, Classical, Mesenchymal, and Neural molecular subtypes.

We performed an integrated analysis of mRNA expression data from the TCGA study and untargeted protein expression data from an independent set of GBM samples. Our analysis identified a core subset of the GBM subtype classification signature which clearly differentiates the known subtypes in the larger TCGA cohort.

Methods

Proteomics analysis of 10 GBM specimens and 10 epilepsy controls was performed using a label-free approach by LC/Q-TOF MS. The LC-MS data was processed using Agilent Spectrum Mill software against UniProt human database with results validated at 1% FDR. The mRNA expression dataset in the form of unified expression values for 173 GBM tumors were obtained from the TCGA portal. Hierarchical clustering of mRNA profiles was performed using the Euclidean distance metric and Ward’s linkage rule. The data was mean centered and scaled prior to performing Principal Component Analysis. Sample-sample correlation analysis was performed using the Pearson similarity metric. All analyses reported in this study were performed using the GeneSpring multi-omic bioinformatics software package.
Results

Label free Proteomics Analysis

The protein database search of the untargeted protein expression data identified about 14K unique proteins across the entire data set. Statistical analysis performed on the proteomics data identified proteins that are measured both in tumor and normal samples and differentially expressed (T-test unpaired, Benjamini Hochberg corrected p-value ≤0.05 and Fold change ≥2.0). The identified differential proteins indicated an ability to classify the tumors in the Principal Component Analysis and sample-sample correlation analysis that were performed.

Tumor subgroups demonstrated using differential proteins, both sample correlation and PCA display the heterogeneity within tumors

![Figure 1: Separation between control (-E) and tumor (-T) samples based on proteins differentially expressed between control and tumor samples (A) Sample-Sample correlation (B) PCA](image)

mRNA Expression based Molecular Subtypes in GBM

![Figure 2: (A) Molecular subtypes in GBM defined by Verhaak et al. validated in GeneSpring. (B) Sample quality metadata visualized as a heat map](image)
Combined Analysis of Protein and Gene Expression

The differential proteins were compared against the subtype specific genes from the TCGA study. This integration identified core signature of genes and proteins.

Classification Performance of Core Subset Identified

Figure 3: Overlap between proteins from discovery differential analysis and subtype specific genes from TCGA

Figure 4: Classification performance of genes identified using proteomics integration to original signature from TCGA demonstrated using Principal Component Analysis.
Summary

- Identified a set of differential proteins for the tumor samples from label-free proteomics data
- Correlation Analysis and PCA on the differential proteins identified the heterogeneity within tumor samples
- Clustering Analysis and PCA on mRNA profiles from TCGA confirmed the subtype specific genes
- Genomics data quality and sample bias assessed using metadata analysis
- Venn analysis of differential proteins and subtype specific genes identified core subset of common genes and proteins
- Core subset genes identified using proteomics integration classified tumors in the larger TCGA genomics cohort

Acknowledgement

The GeneSpring team at Strand Life Sciences, Bangalore, India. Special thanks to Shyam Kalakoti, Srikanthi Ramachandrula and Vinod Kumar.

Reference

Verhaak et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17: 98–110

Abstract - ThP 459

Strand Life Sciences Pvt. Ltd, 5th Floor, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024.

Agilent Technologies, 5301 Stevens Creek Blvd. Santa Clara, CA, 95051 United States.