# Quiescent stem cell marker genes in glioma gene networks are sufficient to distinguish between normal and glioblastoma (GBM) samples.

**Shradha Mukherjee** ( smukher2@gmail.com )

**Method Article**

**Keywords:** stem cells, quiescence, proliferation, glioblastoma, cancer, network analysis, differential gene expression, biomarkers, bioinformatics

**Posted Date:** July 3rd, 2020

**DOI:** https://doi.org/10.21203/rs.3.pex-977/v1



**Version of Record:** A version of this preprint was published at Scientific Reports on July 2nd, 2020. See the published version at https://doi.org/10.1038/s41598-020-67753-5.

# Abstract

Grade 4 glioma or GBM has poor prognosis and is the most aggressive grade of glioma. Accurate diagnosis and classification of tumor grade is a critical determinant for development of treatment pathway. Extensive genomic sequencing of gliomas, different cell types, brain tissue regions and advances in bioinformatics algorithms, have presented an opportunity to identify molecular markers that can complement existing histology and imaging methods used to diagnose and classify gliomas. 'Cancer stem cell theory' purports that a minor population of stem cells among the heterogeneous population of different cell types in the tumor, drive tumor growth and resistance to therapies. However, characterization of stem cell states in GBM and ability of stem cell state signature genes to serve as diagnostic or prognostic molecular markers are unknown. In this work, two different network construction algorithms, Weighted correlation network analysis (WGCNA) and Multiscale Clustering of Geometric Network (MEGENA), were applied on publicly available glioma, control brain and stem cell gene expression RNA-seq datasets, to identify gene network regulatory modules associated with GBM. Both gene network algorithms identified consensus or equivalent modules, HuAgeGBsplit_18 (WGCNA) and c1_HuAgeGBsplit_32/193 (MEGENA), significantly associated with GBM. Characterization of HuAgeGBsplit_18 (WGCNA) and c1_HuAgeGBsplit_32/193 (MEGENA) modules showed significant enrichment of rodent quiescent stem cell marker genes (GSE70696_QNPbyTAP). A logistic regression model built with eight of these quiescent stem cell marker genes (GSE70696_QNPbyTAP) was sufficient to distinguish between control and GBM samples. This study demonstrates that GBM associated gene regulatory modules are characterized by diagnostic quiescent stem cell marker genes, which may potentially be used clinically as diagnostic markers and therapeutic targets in GBM.


# Introduction

**Overview:** This is a comprehensive bioinformatics pipeline for doing differential gene expression analysis, network analysis and characterization of genes/network modules of interest. Modules of interest maybe Disease or other trait associated modules or genes. Characterization includes GO Biological Process, Pathway analysis etc.

**Please cite:** This pipeline was generated and used for the publication with the following article,

Mukherjee, S.*# Quiescent stem cell marker genes in glioma gene networks are sufficient to distinguish between normal and glioblastoma (GBM) samples. Scientific Reports (accepted 2020).


Shradha Mukherjee*#

Independent Researcher

Address correspondence to: Shradha Mukherjee, PhD; Garland Avenue, Downtown Los Angeles, Los Angeles, CA 90017, U.S.

Email: smukher2@gmail.com

* First author and # Corresponding author

# Reagents

**Input:** RNA-seq gene expression data (fastq files, or bam files or htseq-count files) or Microarray gene expression data

**Output:** Gene network modules associated with trait of interest (disease) and their Gene Ontology (GO), pathway analysis and gene network module enrichment of deferentially expressed genes.

# Equipment

Windows, Linux or Mac computers with internet.

# Procedure

**Please cite:** I hope you find these research findings and bioinformatics pipelines useful. Please help by citing as follows. Thank you.

Mukherjee, S.*# Quiescent stem cell marker genes in glioma gene networks are sufficient to distinguish between normal and glioblastoma (GBM) samples. Scientific Reports (accepted 2020).

Shradha Mukherjee*#

Independent Researcher

Address correspondence to: Shradha Mukherjee, PhD; Garland Avenue, Downtown Los Angeles, Los Angeles, CA 90017, U.S.

Email: smukher2@gmail.com

* First author and # Corresponding author

**Please download protocol i.e. codes from 'Supplementary Files':** The following 5 folders are provided with files and codes:

**Step1 RNAseqASUcyverseClusterResultsScript/

Contains representative or sample code for analysis of RNA-seq data from rat, mouse and human. Glioblastoma RNA-seq was from human only, while quiescent/proliferative stem cells RNA-seq were from rat, mouse and human. The starting file to run these codes is fastq/bam files to produce htseq counts rest of the DEG, WGCNA and MEGENA pipelines.

**Step2 DEGNoCuffQbyTStemCellspairs/

Contains the codes, input files and results for entire analysis of the quiescence vs proliferation differential gene expression analysis with 3 methods (Limma, edgeR and simple comparison of means) starting from htseq count files. The htseq count files were obtained in Step1 above.

The minimum files, folders and .Rmd codes required to reproduce the results are provided. After downloading these if the user runs the .Rmd codes in RStudio using command 'knit to html' or 'Run all' it should reproduce same results.

**Step3 HuDisGBsplitWGCNA/

Contains the codes, input files and results for WGCNA analysis of htseq files obtained in Step1 above. Also contains code for characterization of modules with enrichR such as GO analysis; and contains code for Linear Regression/Modeling (to distinguish control and cancer samples) and Survival analysis.

The minimum files, folders and .Rmd codes required to reproduce the results are provided. After downloading these if the user runs the .Rmd codes in RStudio using command 'knit to html' or 'Run all' it should reproduce same results.

**Step4 HuDisGBsplitMEGENA/

Contains the codes, input files and results for MEGENA analysis of htseq files obtained in Step1 above. Also contains code for characterization of modules with enrichR such as GO analysis;

The minimum files, folders and .Rmd codes required to reproduce the results are provided. After downloading these if the user runs the .Rmd codes in RStudio using command 'knit to html' or 'Run all' it should reproduce same results.

**Step5 PreservationAnalysis/

Contains the codes, input files and results for comparison of MEGENA analysis and WGCNA analysis outputs obtained in Step3 and Step4 above.

The minimum files, folders and .Rmd codes required to reproduce the results are provided. After downloading these if the user runs the .Rmd codes in RStudio using command 'knit to html' or 'Run all' it should reproduce same results.

**Alternatively download link on github:**
https://github.com/smukher2/GithubScientificReportsGlioblastomaStemApril2020

# Troubleshooting

**Acknowledgements and additional resources:**

1) SVA: Dr. Jeff Leek http://jtleek.com/genstats/inst/doc/02_13_batch-effects.html and https://www.bioconductor.org/packages/release/bioc/vignettes/sva/inst/doc/sva.pdf

2) WGCNA: Dr. Jeremy Miller https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/JMiller/

3) WGCNA Preservation: Dr. Steve Horvath and Dr. Peter Langfelder https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/ModulePreservation/Tutorials/MiniTutorial-MouseLiver.pdf

4) WGCNA Hub Genes: Dr. Steve Horvath and Dr. Peter Langfelder https://pdfs.semanticscholar.org/5e42/e2185c54874277794395e5825808e5f5709c.pdf

5) MEGENA: Dr. Won-Min Song and Dr. Bin Zhang https://rdrr.io/github/songw01/MEGENA/f/vignettes/MEGENA_pipeline_02202020.Rmd

6) edgeR: https://github.com/smukher2/StemCells_RNAseq_Expression_edgeR_visualization_GO and https://web.stanford.edu/class/bios221/labs/rnaseq/lab_4_rnaseq.html

7) limma: https://kasperdanielhansen.github.io/genbioconductor/html/limma.html and https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html

Users new to SVA, WGCNA, MEGENA and Differential Gene Expression methods are encouraged to refer to the above mentioned turotials and pipelines. Other acknowledgements have been added to the best of our knowledge as #comment in the code/pipeline itself.

## Time Taken

Variable as it depends on number of samples, number of variables and user.

## Anticipated Results

1) SVA+LM approach reduces batch effects in RNA-seq datasets.

2) Network analysis to reveal (here GBM) associated modules.

3) Differential gene expression analysis to identify marker genes (here proliferative and quiescent stem cell marker genes).

4) Enrichment of marker genes (here proliferative and quiescent stem cell marker genes) underly gene network modules (here GBM modules).

5) Gene Ontology (GO) annotation of marker genes enriched gene network modules (here GBM modules enriched with quiescent stem cell marker genes).

## References

**Please refer to associated article for complete list of references.**

Mukherjee, S.*# Quiescent stem cell marker genes in glioma gene networks are sufficient to distinguish between normal and glioblastoma (GBM) samples. Scientific Reports (accepted 2020).

Shradha Mukherjee*#

Independent Researcher

Address correspondence to: Shradha Mukherjee, PhD; Garland Avenue, Downtown Los Angeles, Los Angeles, CA 90017, U.S.

Email: smukher2@gmail.com

* First author and # Corresponding author

# Acknowledgements

**Acknowledgements:** This work was conducted by the author using free open access resources, personal resources and personal funds from 2017 to 2019. This work is self-owned and not affiliated with any organization/university. The author would like to express deepest gratitude to Springer Nature and Scientific Reports for their generous support. Other open source and free resources utilized for this work are acknowledged below.

For datasets: Glioblastoma RNA-seq datasets, SRP027383 and SRP091303, were obtained from NCBI SRA. Other RNA-seq datasets for control brains (GSE67333, GSE64810, GSE100297 and GSE53697) and stem cells (GSE68270, GSE70696, GSE99777, GSE93991 and GSE114574), were obtained from NCBI GEO. The author is most grateful to NCBI SRA https://www.ncbi.nlm.nih.gov/sra and NCBI GEO https://www.ncbi.nlm.nih.gov/geo/ for storing and making these datasets open access. The author also appreciates the organizations and investigators who generously submitted these datasets for sharing to NCBI GEO and NCBI SRA.

For full text publications access: The author is most grateful to NIH/NLM (U.S. National Institute of Health's National Library of Medicine) for open-access to full-text scientific publications on www.ncbi.nlm.nih.gov/pmc/ PubMed Central.

For computational pipelines: Citations and descriptions of all computational pipelines have been given under methods sections. Additionally, the author would like to add special thanks to Dr. Steve Horvath, Professor at Department of Human Genetics and Biostatistics, University of California Los Angeles (UCLA) and Dr. Jeffrey T. Leek, Professor at Department of Biostatistics, Johns Hopkins for their publicly available computational pipelines on WGCNA/enrichment/Preservation analyses and SVA/Regression analyses, respectively. Dr. Won-Min Song and Dr. Bin Zhang from Department of Genetics and Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai for their publicly available computational MEGENA pipeline on github. Dr. Steve Horvath, Dr. Jeffrey T. Leek, Dr. Won-Min Song and Dr. Bin Zhang's computational tutorials were integral to the author's learning and for execution of the work done in this paper.

For computational platform: The computation was performed on personal laptops (MacBook Pro and Windows OS) with high-speed internet connection and computational tools. RNA-seq datasets were mapped using NSF supported publicly available Cyverse Discovery Environment.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DEGNoCuffQbyTStemCellspairs.zip
- HuDisGBsplitMEGENA.zip
- RNAseqASUcyverseClusterResultsScript.zip
- PreservationAnalysis.zip