A multistep computational procedure to identify candidate master Transcriptional Regulators (TRs) of glioblastoma (GBM)

Michelangelo Cordenonsi (✉ michelangelo.cordenonsi@unipd.it)
Silvio Bicciato (✉ silvio.bicciato@unimore.it)
Stefano Piccolo (✉ piccolo@bio.unipd.it)

Method Article

Keywords: gene regulatory network, master Transcriptional Regulators, single-cell RNA-seq analysis

Posted Date: December 9th, 2020

DOI: https://doi.org/10.21203/rs.3.pex-1230/v1

Abstract

We describe a multistep computational procedure to identify candidate master Transcriptional Regulators (TRs) of glioblastoma (GBM) from glioblastoma single cell gene expression profiles and tissue-specific transcription factors.

Introduction

To identify candidate master transcriptional regulators of glioblastoma, we assembled a 3-step computational workflow named "Rhabdomant" that: (i) reconstructs the gene regulatory network (GRN) from GBM single cell gene expression profiles and GBM-specific transcription factors; (ii) anchors regulatory interactions to the GBM epigenetic landscape, through the association of target genes to transcriptional regulators (TRs) that are potentially bound to their cis-regulatory elements; (iii) scores of candidate master TRs from the differential enrichment of their regulons in different cell states.

Reagents

Equipment

Procedure

Step 1: network reconstruction. To reconstruct the GBM regulatory network, we used ARACNe-AP[1]. ARACNe-AP requires in input a gene expression matrix and a predefined list of transcriptional regulators. As gene expression data, we used the single cell transcriptomes (normalized gene counts) of all 883 neoplastic cells from Darmanis et al.[2]. To limit the effects of scRNA-seq data sparsity in GRN reconstruction[3], we removed from the input matrix genes detected only in a limited number of cells (normalized counts >0 in at least 100 cells, i.e., slightly over the number of cells in which a gene is detected on average in this dataset). The filter retained a total of 9,845 genes. As candidate transcriptional regulators (TRs), we used a list of transcription factors whose DNA-binding motifs were found enriched in the open chromatin regions of brain cancers and their dedicated transcriptional co-factors. Briefly, taking advantage of the epigenetic analyses provided by a recent large-scale ATAC-seq profiling of several human tumor types, including GBM[4], we carried out a DNA binding motif enrichment analysis using the HOMER *findMotifsGenome.pl* function on the open chromatin regions of GBMs, and then selected the transcription factors whose DNA-binding motifs were highly enriched (FDR<0.0001) in at least 10% of these genomic regions. We excluded few factors (i.e., BATF, JUNB, PRDM1), since they are known to work mainly as transcriptional repressors. We manually curated this list with partner transcriptional co-factors (e.g., NOTCHs for RBPJ, etc) for a total of 151 TRs (Supplementary Table 1). ARACNe-AP was run with 100 bootstrap iterations setting the DPI (data processing inequality) tolerance to 0 and the MI (mutual information) P value threshold to 10^{-8}. This resulted a single cell GBM network comprising 32,016 interactions between 70 TRs and 6,079 genes.

Step 2: anchoring of regulatory interactions to GBM epigenetic landscape. Since the network inferred in the first step is based only on GBM single cell expression data, we intersected the gene interactomes of the GBM network with epigenetic data to eliminate spurious associations generated by the noise intrinsic to scRNA-seq data and to retain only biologically relevant associations between TRs and putative target genes. Specifically, we first analyzed the epigenetic data provided by Corces et al.[4] using the HOMER *annotatePeaks.pl* function to map the TRs potentially able to bind each of the ATAC-seq peak of GBMs. Then, we used the associations between each ATAC-seq peak and its target genes (Data S2 of Corces et al.[4]) to map each TR to a set of putative target genes (Supplementary Table 2). Finally, we intersected this set of putative direct target genes with the interactomes of the GBM network, thus removing upstream regulators, indirectly interacting genes, and false positives and retaining only the lists of target genes with binding motif for a TR in their cis-regulatory elements (regulons).

Step 3: scoring of candidate master TRs. To identify master regulators of GSCs, we applied the VIPER algorithm[5] to the GBM regulons resulting from Step 2. VIPER uses the expression level of genes that are most directly regulated by a given protein, such as the targets of a transcription factor, to infer the activity of the transcription factor on a set of samples. We used the VIPER algorithm as implemented in the *msviper* function of the Bioconductor *viper* package (v. 1.16.0) in R 3.5.0. The *msviper* requires in input i) a gene expression signature resulting from the comparison of two types of samples, ii) a null model composed by a set of signatures obtained after permuting the samples at random, and iii) a gene regulatory network. To reduce the effect of outlier samples on the gene expression signature, we performed the *msviper* analysis with bootstrap, i.e., imputing the gene expression signature as a matrix of bootstrapped signatures. Specifically, we generated the bootstrapped signature matrix comparing, in 100 bootstrap iterations, the expression profiles of GSC versus GDC cells with the *bootstrapTtest* function. We defined the null model as a set of gene expression signatures obtained through the *ttestNull* function by randomly shuffling 1,000 times the samples among the GSC and GDC sets and we used the GBM regulons from Step 2 as the gene network. Finally, we used the *bootstrapmsviper* function to integrate the regulator activity results of the *msviper* analysis across all bootstrapped iterations. Based on VIPER analysis, we defined as candidate master TRs those regulomes (i.e., the TR and its regulon) with an FDR ≤ 0.05 and a number of target genes >70. This resulted in 27 candidate MRs, of which 15 were candidate master TRs of the GSC state and 7 were candidate master TRs of the DGC state.

References

1 Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics 32, 2233-2235 (2016).

2 S. Darmanis *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. Cell reports 21, 1399-1410 (2017).

3 Blencowe, M., Karunanayake, T., Wier, J., Hsu, N. & Yang, X. Network Modeling Approaches and Applications to Unravelling Non-Alcoholic Fatty Liver Disease. Genes (Basel) 10 (2019).

4 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. Science 362 (2018).

5 Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nature genetics 48, 838-847 (2016).

Supplementary Files

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

- supplTable1.xlsx
- supplTable1.xlsx
- supplTable2.xlsx
- supplTable2.xlsx