– SUPPLEMENTARY MATERIAL B –

Validation of the updated glioma TCGA labels

Given the retrospective approach of our study, the reliability of the newly assigned labels has been evaluated through analysis of correspondent omics data from the TCGA glioma samples considered in this study.

When considering the proposed updated labels derived from the employment of Method-2021, enhanced performance has been observed in comparison with original TCGA classification, in various aspects, such as network inference, variable selection, classification, and survival modelling [21, 22]. In addition, the updated labels provide a significant improvement in the distinction of the glioma types resulting from the projection of the TCGA samples on a lower-dimensional space based on omics data. Figure SB1 shows, per row, the results of the Uniform Manifold Approximation and Projection (UMAP) [23] computed by considering as input RNA-Sequencing (RNA-Seq), DNA-methylation (DNA-Meth), and microRNA (miRNA) data. Every point represents a glioma patient projected in a 2D space by the UMAP method. For each omic, the points have the same distribution, with colours representing different diagnostic labels, which vary following TCGA labels' annotations, Method-2016, and Method-2021 (per column). Based on the graphical distributions of the samples in the two-dimensional space, we observe a globally accurate assignment of the oligoastrocytoma samples achieved by Method-2016, as most of the points are correctly distributed into astrocytoma-oligodendroglioma classes. By comparing the UMAP outcomes from RNA-Seq data (Figure SB1, first row), it is evident the presence of LGG patients aggregated in the GBM area. According to our Method-2021, all of the referred samples change their diagnostic label to GBM, resulting in a more consistent graphical distribution. This improvement is even more evident if considering DNA-Meth data (Figure SB1, second row), which leads to a nearly ideal separation of the three patient groups based on Method-2021.

Both RNA-Seq and DNA-Meth data highlight that GBM cases can be easily distinguished from astrocytoma and oligodendroglioma, while the latter are more closely distributed. The similarity between the two LGG classes (astrocytoma and oligodendroglioma) is supported by their molecular biomarker characterisation, i.e., they both exhibit IDH-mutation. To further investigate the separation among these patients, we considered miRNA data as additional omics, which is exclusively available for LGG (Figure SB1, third row). In this case, both the proposed Method-2016 and Method-2021 contribute to a more refined distinction between astrocytoma and oligodendroglioma, resulting in more homogeneous groups, though without a clearly evident separation.

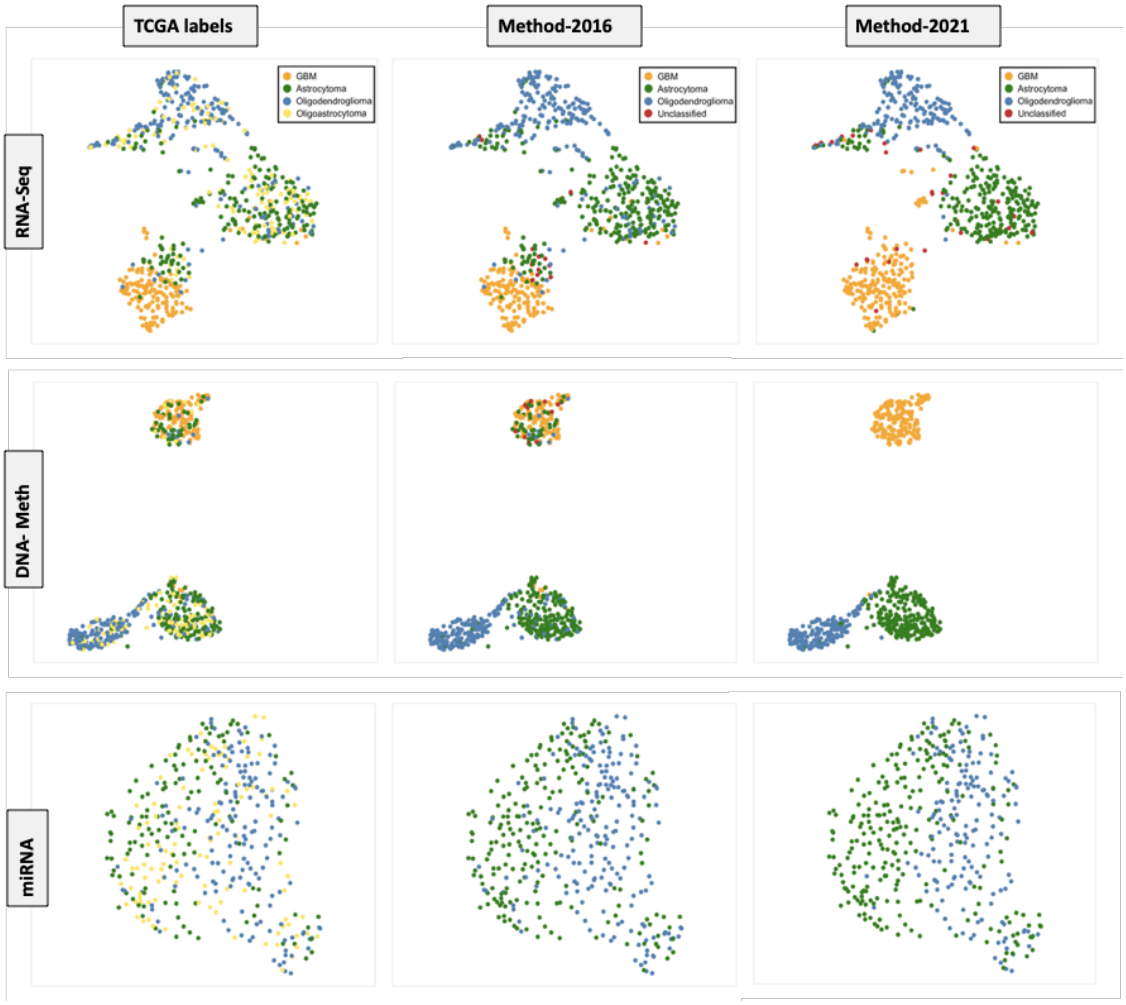

Figure SB1 – UMAP projection of TCGA glioma patients based on multi-omics TCGA data. Rows refer to different omics data, while columns correspond to diagnostic labels based on TCGA annotation, Method-2016, and Method-2021. For each omics, the point distribution is maintained, while changing the associated labels according to the different methods. Samples are coloured according to glioma type: GBM in orange, Astrocytoma in green, Oligodendroglioma in blue; Oligoastrocytoma (Mixed glioma) in yellow; Unclassified samples in red. Note: each omics dataset accounts for a different set of patients, with significant inter-omics overlap. This shows at unclassified patients not being present in all omics-method sets. *Abbreviations: RNA-Seq, Ribonucleic acid-Sequencing data; DNA-Meth, Deoxyribonucleic acid-Methylation data; miRNA, microRNA data.*