Title: Normal tissue content impact on the GBM molecular classification

AUTHORS: Rodrigo Madurga1,2,3, Noemí García-Romero1,2, Beatriz Jiménez1,2, Ana Collazo1,2, Francisco Pérez-Rodríguez1,2, Aurelio Hernández-Laín4, Carlos Fernández-Carballal5, Ricardo Prat-Acín6, Massimiliano Zanin7,8, Ernestina Menasalvas8,9, Ángel Ayuso-Sacido10, 11, 12*.

1. Fundación de Investigación HM Hospitales, HM Hospitales, Madrid, Spain.

2. Instituto de Investigación Sanitaria HM Hospitales (IISHM), HM Hospitales, Madrid, Spain.

3. Faculty of Experimental Sciences, Universidad Francisco de Vitoria, Madrid, Spain

4. Unidad Multidisciplinar de Neurooncología, Hospital Universitario 12 de Octubre, Madrid, Spain.

5. Servicio de Neurocirugía, Hospital General Universitario Gregorio Marañón, Madrid, Spain.

6. Departamento de Neurocirugía, Hospital Universitario la Fe, Valencia, Spain.

7. Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain.

8. Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain.

9. ETS de Ingenieros Informáticos, Universidad Politécnica de Madrid, Madrid, Spain

10. Fundación Vithas, Vithas Hospitals, Madrid, Spain.

11. Formerly: Fundación de Investigación HM Hospitales, HM Hospitales, Madrid, Spain.

12. Facultad de Medicina (IMMA), Universidad San Pablo-CEU, Madrid, Spain.

*. Corresponding author: +34 686966904, ayusosacido@gmail.com

Authors description:

Rodrigo Madurga is a postdoc researcher in Biostatistics and Bioinformatics at Fundación de Investigación HM Hospitales and Professor at the Faculty of Experimental Sciences of the Universidad Francisco de Vitoria. His research focuses on computational and statistical modeling in cancer genomics.

Noemí García-Romero is a postdoc resercher in Molecular Biology at Fundación de Investigación HM Hospitales. Her research focuses on brain tumors, particularly on Glioblastoma.

Beatriz Jiménez is an oncologist physician specialized in brain tumors at HM Hospitales.

Ana Collazo is an oncologist physician specialized in brain tumors at HM Hospitales.

Francisco Pérez-Rodríguez is a pathologist at HM Hospitales.

Aurelio Hernández-Laín is a pathologist at Hosptal Universitario 12 de octubre.

Carlos Fernández-Carballal is an oncologist physician specialized in brain tumors at Hospital Universitario Gregorio Marañón.

Ricardo Prat-Acín is an oncologist physician specialized in brain tumors at Hospital Universitario la Fe.

Massimiliano Zanin is a researcher at the Institute for Cross-Disciplinary Physics and Complex Systems, Spain. His main topics of interest are Complex Networks and Data Science.

Ernestina Menasalvas is a full professor at the Department of Computer Systems Languages and Software Engineering at the Faculty of Computer Science of Universidad Politecnica de Madrid. Her subject area is Data Mining.

Ángel Ayuso-Sacido is The Head of the Brain Tumour Laboratory, Scientific Director at

Fundación de Investigación Vithas Hospitales and Professor at the Medial School of CEU San Pablo University

Abstract

Molecular classification of glioblastoma has enabled a deeper understanding of the disease. The four-subtype model (including Proneural, Classical, Mesenchymal and Neural) has been replaced by a model that discards the Neural subtype, found to be associated with samples with a high content of normal tissue. These samples can be misclassified preventing biological and clinical insights into the different tumor subtypes from coming to light.

In this work, we present a model that tackles both the molecular classification of samples and discrimination of those with a high content of normal cells.

We performed a transcriptomic in silico analysis on GBM samples (n = 810) and tested different criteria to optimize the number of genes needed for molecular classification. We used gene expression of normal brain samples (n = 555) to design an additional gene signature to detect samples with a high normal tissue content. Microdissection samples of different structures within GBM (n = 122) have been used to validate the final model. Finally, the model was tested in a cohort of 43 patients and confirmed by histology.

Based on the expression of 20 genes, our model is able to discriminate samples with a high content of normal tissue and to classify the remaining ones. We have shown that taking into consideration normal cells can prevent errors in the classification and the subsequent misinterpretation of the results. Moreover, considering only samples with a low content of normal cells, we found an association between the complexity of the samples and survival for the three molecular subtypes.

Introduction

Glioblastoma (GBM) is the most lethal brain tumor with a median overall survival (OS) of 15 months and incidence rate of 3-4 new diagnosed cases per 100,000 population [1,2].

In the last decade, there has been increasing interest in the molecular classification of GBM [3]. In 2010, a four-subtype classification model (Proneural (PN), Classical (CL), Mesenchymal (ME) and Neural (NE)) was proposed, based on the expression levels of 840 genes [4]. This classification has widely been used to analyze differences in treatment response patterns of different GBM subgroups [5,6].

However, when an unsupervised clustering of the samples was performed against tumoral-related genes, a three cluster GBM classification was obtained by analyzing the expression of 150 genes [7]. Notably, each cluster was strongly associated with one group of the four-subtype model, except for the NE subtype. One possibility is that this subtype includes samples with a high content of normal cells. In fact, at the infiltrative margins of GBM, normal cells have been found to far outnumber tumoral ones and the NE subtype is associated with this region [8].

This abundance of normal cells is a problem for transcriptional classification of samples, as long as RNA expression levels are affected by tumor purity [9]. The content of normal cells in a sample affects transcriptional classification, complexity of the sample measured by the simplicity score [7] and intratumoral heterogeneity, among others. Different algorithms can calculate a purity score of tumor samples from CNV data [10] or from gene expression signatures [11]. However, few models are able to tackle both the molecular classification of GBM and the tumor purity of the sample.

In the present work, we develop a model that classifies GBM samples as PN, CL or ME and provides us with information about the abundance of normal cells in the samples based on the expression of 20 genes. This model not only integrates the molecular and purity classification of the samples but does so in a cost-effective way.

Materials and methods

Gene expression data processing and normalization

GBM IDH wt cohort (IDHWT)

Gene expression data from GBM patients with known IDH status were collected from TCGA [12], GlioVis [13] and from the Gene Expression Omnibus. Affymetrix data sets were normalized using robust multi-array average normalization (RMA) followed by quantile normalization as implemented in the 'affy' package for R/Bioconductor [14]. Affymetrix data sets consisted in TCGA (n = 528) [12], GSE4271 (n = 76) [15] and GSE36245 (n = 46) [16]. Additionally, RNA-seq data were downloaded from GSE48865 (n = 100) [17] and GSE121720 (n = 60). Collected RNA-seq had been mapped to the hg19 human genome and log transformed. Because data sets were generated on different platforms and by different labs, we used ComBat to address the strong batch effects expected from such variable data sources [18]. We applied aggregation workflow, as described elsewhere [19], to select the probe set that represents each gene in each data set. Once different data sets had been aggregated, IDH wt and $CIMP^-$ samples were filtered obtaining a final cohort of 551 samples. Where the CIMP status was unknown, this was determined by the support vector machine, using the TCGA cohort as training data set [13]. The final cohort was divided in a training cohort (n = 367) and a validation cohort (n = 184).

Histology cohort (HIS)

Solid surgical tissue samples were obtained from patients operated in HM Hospitales, Madrid, Spain; Hospital General Universitario Gregorio Marañon, Madrid, Spain and Hospital Universitario la Fe, Valencia, Spain. A total of 43 GBM IDH wt patients were analyzed by qRT-PCR.

NormalBrain cohort (NB)

Normalized microarray (custom-designed Agilent 8x60K) gene expression from 6 human brains were downloaded from the ALLEN Human Brain Atlas (http://human.brain-map.org/static/download). One hundred samples of cerebral cortex were randomly selected from each brain. Outliers were removed using principal component analysis (PCA), which consisted in considering the first two principal components and marking all the samples with a distance greater than 2.5 as outliers. After outlier removal, to avoid batch effects we applied ComBat [18] to the NTB cohort using the TCGA cohort (including 10 normal samples) as a reference. Finally, the NTB cohort was divided into NTB-training (n = 370) and NTB-validation (n = 185).

Ivy GAP cohort (IVYGAP)

The Ivy Glioblastoma Atlas Project (Ivy GAP) analyzed the transcriptome of different anatomical structures from 10 different tumors [20]. The normalized read counts of these 122 samples were downloaded from the Ivy GAP portal (http://glioblastoma.alleninstitute.org/static/download.html) and log transformed.

Differentially expressed genes

Differentially expressed genes were identified using the R/Bioconductor package 'multtest' [21] with 5000 bootstrap iterations and using FDR as method to control type one error rate. The significance level was set to 0.05.

Reduction of the number of genes in the gene signature

The 50-gene signatures proposed elsewhere [7] were used for the subtype classification of the IDHWT training set in PN, CL or ME. These results were used as a reference in the following analysis. Afterwards, the gene signature of each subtype was reduced, one gene at a time, to a 2-gene signature. For each step, 1000 different combinations of signature genes were randomly generated and used for the subtype classification of the training cohort. The overlap of these classifications with the 50-gene signature classification was estimated, as well as the accuracy and precision achieved by the reduced gene signatures for each subtype.

Along with the randomly selected genes, three different criteria were used to rank genes inside each of the three gene-signatures, keeping the top ranked genes in each step of gene removal. These criteria were difference and relative difference in gene expression and statistical significance, measured by the value of the statistic, obtained when comparing the expression levels between subtypes. The overlap of the classification obtained from these criteria was compared to that obtained from the randomly selected gene signatures for each size of gene signature.

Molecular classification

Two methods for the classification of GBM samples have been used through this work:

Single sample gene set enrichment analysis (ssGSEA)

Single sample gene set enrichment analysis defines an enrichment score for a gene signature, in this case representative of a subtype, for each sample within a dataset. The process starts with the rank-normalization and rank-order of gene expression for a given sample. A statistic is then calculated from the difference between the cumulative empirical distribution functions (ECDF) of the gene signature and the remaining genes

[7]. A null distribution of the enrichment scores for each signature is obtained as follows. A large number of virtual samples (> 10,000) are generated assigning to each gene the expression level of the same gene in a randomly selected sample in the dataset. Null distributions are used to give an empirical p value to the enrichment scores obtained for each sample in the dataset [7]. A given sample is classified with the subtype with the lowest empirical p value.

Centroid-based classification

Verhaak's classification is based on a 210 gene signature for each of the four subtypes [4]. ClaNC software [22] was used to assign a category to the training samples based on Verhaak's gene signature.

CNA and mutations from the TCGA cohort

CNA and mutational information about samples from the TCGA cohort were downloaded from the TCGA repository using TCGAbiolinks package from Bioconductor [23].

Survival analysis

The optimal cutoff of the simplicity score for the survival analysis was determined by the *get.cutoff()* function described elsewhere [24]. The method used for the cutoff optimization was survival significance, were a Cox proportional hazard model is fitted to the dichotomized simplicity score and the survival variable. The point with the most significant split (measured by the log rank test) is defined as the optimal cutoff.

Histology

Six tissue microarrays (TMAs) were constructed from 32 Formalin Fixed Paraffin Embedded (FFPE) tissues using an arraying instrument (GALILEO CK 3500). From each tissue block a total of three tissue cores were made with a diameter of 0.6-1 mm. Then,

TMA blocks were cut at 4 µm and stained with hematoxylin & eosin (H&E). The H&E stained tissue blocks were evaluated by a pathologist for tumoral cells density, abundance of pathogenic blood vessels and presence of necrosis.

RNA isolation and qRT-PCR

Total RNA was isolated using RNeasy Mini or Micro kit (QIAGEN) following the manufacturer's recommendations. cDNA synthesis (High-Capacity cDNA Reverse Transcription Kit; Applied BioSystems) was performed from one µg of RNA. An optical 384-well plate equipped with an ABI PRISM 7900 HT sequence detection system (Applied Biosystems) was used for the performance of qRT-PCR reactions using SYBR Green. Two housekeeping genes were used to normalize data, the primers used for each gene can be seen in Supplementary Table S1. Because the ssGSEA classification is based on the ranking of genes rather than on absolute expression, we used the Ct value of the genes, which were scaled within the same sample:

$$Z = \frac{r_i - 1}{n_{genes} - 1}$$

Statistical analysis

All the statistical analyses have been performed using R software. The statistical tests applied are indicated in the text. When relevant, p values were adjusted using Benjamin and Hochberg method.

Results

Transcriptomic data aggregation

As explained in the Materials and Methods, the initial cohort, comprised of 810 patients, was filtered to discard IDH mt or GCIMP$^+$ patients, which are already known to have favorable clinical outcomes [25,26]. A total of 86 patients were IDH mutant and 163

were IDH unknown. It is worth mentioning that IDH status of Phillips' cohort (GSE121720) were obtained elsewhere [27]. After filtering, a final cohort of 551 IDH wt / GCIMP$^-$ patients was obtained and divided, randomly, into two cohorts: IDHWT training cohort (n = 367) and IDHWT validation cohort (n = 184), as shown in Figure 1A.

5-Genes signature for molecular Glioma subtypes.

The IDHWT training cohort was classified into PN, CL and ME as proposed elsewhere [7]. The simplicity score proposed in the cited work was estimated. This gives a value between 0 and 1, with high values corresponding to samples activating a single subtype and low values for samples activating multiple subtypes. Afterwards, to performed a clean comparison between subtypes, samples with a simplicity score higher than 0.95 (n = 121) were selected and a differential expression analysis was performed between subtypes for the 150 genes involved in the classification process. Because the genes for the original classification were selected on the basis of differences in gene expression between groups [7], we used the results of the previous analysis to rank the gene signatures according to differences and relative difference in mean expressionand statistical significance between subtypes.

The three rank criteria were used separately to remove the last gene from the gene signature of each subtype at a time.The reduced gene signatures were used to classify the IDHWT training cohort and the overlap with the original classification was estimated. Additionally, 1,000 randomly ranked gene lists were generated and used to evaluate the overlap with the original classification, in order to generate a null distribution for each size of the gene signature.

To delucidate if the above mentioned rank criteria generate reduced gene signatures with better performance than random, we used the overlap mean and standard

deviation obtained from the null distribution to scale the results obtained with the three different ranking criteria (see Figure 1B and SI Appendix Figure S1). Colored symbols represent the scaled overlap of the difference in mean ranking criterion and dashed line represent the mean Z-score obtained for each criterion. It can be appreciated that differences in means criterion reach higher Z-scores for different gene signature sizes than the other two criteria. It is worth mentioning that with just 5 genes per subtype an overlap higher than 90% is obtained and that when the number of genes is reduced to 2 the overlap is still higher than 80%, with Z-scores higher than 2 in both cases (Figure 1B). Comparing the classification by subtype, using 5 genes per subtype the true positive ratio (TPR) reached values higher than 90% for the CL and ME subtypes, and as high as 83% for the PN subtype. On the other hand, the true negative ratio (TNR) reached values higher than 93% for the three subtypes (SI Appendix Figure S2A-C).

Considering these results, we decided to reduce the number of genes per subtype to the minimum for which an overlap with the original classification of at least 90% is achieved. Therefore, we used the difference in means expression rank criteria to reduce the gene signature of each subtype to 5 genes. At this point, we now have a 15 gene signature that classifies samples into 3 subtypes (5-gene signatures for each subtype). This gene signature seems to preserve an overlap close to 90% with the original signature of 150 genes (50-gene signature for each subtype) and requires 1/10th of the molecular information used by the previous method.

Prediction of samples with high content of normaltissue

As mentioned above, the classification proposed by Wang et al. [7] stratifies samples into 3 subtypes with a high correspondence with 3 of the 4 subtypes previously proposed by Verhaak et al. [4]. The remaining subtype, called Neural (NE), was

hypothesized to include samples with a high content of normal tissue [7]. Under this assumption, we asked ourselves whether the presence of normal tissue could affect the classification of the samples. Therefore, we used the centroid-based classification, proposed by Verhaak et al. [4], to classify the training cohort and study the distribution of the simplicity score, obtained as proposed by Wang et al. [7], in the four subtypes. Figure 1C shows that the NE subtype is enriched for lower simplicity scores (Wilcoxon test, $p <= 0.001$). Because the simplicity score is based on the distances to the dominant subtype and between non-dominant subtypes [7], a low value can be obtained either if a non-dominant subtype is close to the dominant one, indicative of a complex sample; or if the dominant subtype is weakly activated, which may occur when the expression levels of the tumoral cells are masked by a high content of normal cells [9].

We started to search for genes of the NE signature that were overexpressed in the NE subtype compared with the non-neural samples, with a simplicity score higher than 0.95. This analysis gave 46 overexpressed genes (see Supplementary Table S2). We then used the NTB-training cohort (see Material and Methods) and analyzed the differential expression of the 46 overexpressed genes between normal tissue and non-neural samples with high simplicity scores ($ss > 0.95$). The analysis resulted in 35 overexpressed genes in normal tissue (Supplementary Table S2). We then selected the top 5 of these genes in relation to differences in mean expression and formed a fourth gene signature (*CCK*, *CRYM*, *SERPINI1*, *KCNK1* and *GPR22*). We used Enrichr [28,29] to analyze which tissues were enriched in this new gene signature. From the ARCHS4 Tissues library six different structures, all of them from brain, were significantly enriched (p value < 1e-4, q value < 0.001) (see Supplementary Table S3). This new gene signature was added to

the previously reduced gene signatures (Figure 1A) with the intention of detecting samples with a high content of normal tissue.

We used our 20-gene based signature to classify the samples into the different subtypes, or as normal tissue abundant (NT). Figure 1D shows the distribution of the simplicity scores from Wang's classification, for the different subtypes obtained with the reduced gene signature. It can be observed that the NT is enriched for lower simplicity scores (Wilcoxon test, $p < 1e\text{-}4$) as was also found for the NE subtype (Figure 1C). Although the result is similar to that obtained for the NE subtype, we still need to prove that samples classified as NT have a high content of normal brain cells. In parallel, it is worth proving that the addition of a fourth subtype did not alter the overlap when compared with the 50-gene signature classification.

Validation of the model

We performed splitting iteration 1000 times to generate different validation cohorts. Each of these cohorts were used to study the overlap between the classification performed by the reduced gene signature and the original gene signature (Figure 2A). If the samples classified as NT are not considered, a mean overlap of 89 % and 2% standard deviation are achieved.

When the results were analyzed by subtype, we found that our classification model, with 5-gene signatures, is an excellent predictor of the results that would be obtained by the 50-gene signature model as can be appreciated in the ROC space shown in Figure 2B. For all the subtypes, the model reaches a high sensitivity: 0.94 ± 0.04, 0.89 ± 0.04 and 0.87 ± 0.03 for PN, CL and ME respectively, and high specificity: 0.95 ± 0.02, 0.94 ± 0.02 and 0.95 ± 0.02 for PN, CL and ME respectively. As a measurement of the accuracy for each subtype we estimate the F1-score, with values of 0.90 ± 0.03, 0.89 ± 0.03 and 0.89

± 0.02 for PN, CL and ME respectively. In synthesis, the results presented here show that a significant reduction in the number of genes used for the classification does not dramatically affect the performance of the classification.

To confirm that the reduced gene signature gives valuable information we used the TCGA cohort, for which genomic data is available, to study the incidence of genomic alterations in the different subtypes. Figure 2C shows that the characteristic genomic alterations for each subtype are still found when samples are classified using the reduced gene signature.

Wang et al. [7] found that the ME subtype shows a reduced OS for single subtype activated samples, that is, samples with a simplicity score higher than 0.99 (~20% of the samples). We observed that the simplicity score calculated from the empirical p values obtained from our model was lower than that obtained from Wang's model. Only four samples had a simplicity score higher than 0.99. However, if we select the top 20% of the simplicity scores of the samples the same result is obtained as shown in Figure 2D (log rank test, $p = 0.03$).

NT associates with abundance in normal brain cells

Once the reduced gene signatures have been proven to be a good predictor of the original classification, it is time to address whether or not NT is identifying samples with a high content of normal cells. We used the ABSOLUTE method [10], which gives a tumoral purity score based on copy number variation data to study tumoral purity of the samples in the different subtypes. The TCGA cohort was used for this analysis as it is the only one for which CNV data were available. Additionally, we used the ESTIMATE method [11], which gives a tumoral purity score based on the enrichment scores obtained for an immunological and a stromal gene signature, for the same purpose.

330 Figure 3A shows the ABSOLUTE and ESTIMATE purity scores obtained for the TCGA
331 cohort. As reported previously [7], PN and CL subtypes showed higher scores than ME
332 for both algorithms (Wilcoxon test, $p < 1e\text{-}15$). This result was shown to be due to the
333 higher infiltration of immunological cells occurring in ME tumors [7]. Our results showed
334 that PN and CL subtypes also obtained a higher ABSOLUTE purity score than NT
335 (Wilcoxon test, $p = 1e\text{-}5$) indicating a higher content of normal brain cells in NT samples.
336 However, no significant differences were observed for the ESTIMATE purity score.
337 Therefore, there was no increase in the amount of immunological or stromal cells. This
338 result was confirmed by repeating the analysis for the IDHWT validation cohort. The
339 results can be seen in Figure 3B, with no significant differences between PN and CL
340 subtypes compared with NT, although the ME subtype shows significantly lower values
341 (Wilcoxon test, $p = 1.9e\text{-}14$).
342 These results are in line with the hypothesis that NT are samples with a high content of
343 normal cells. This hypothesis is also supported by the classification performed on the
344 NTB-validation cohort. Of the 185 normal cortex samples, 184 (99.5 %) were classified
345 as NT and only 1 sample was classified as tumoral, in this case PN.
346 To test NT in a cancer context we used the IVYGAP cohort, which consists of different
347 tumoral structures obtained from GBM biopsies subjected to laser microdissection.
348 Briefly, the IVYGAP cohort is composed of five different structures as defined elsewhere
349 (http://help.brain-map.org/display/glioblastoma/Documentation): cellular tumor (CT)
350 has the most core tumor with a tumor cell to non-tumor cell ratio between 100/1 and
351 500/1, microvascular proliferation (MVP) are regions characterized by two or more
352 vessels sharing common vessel walls, pseudopalisading cells around necrosis (PAN),
353 which are generally found in the core tumor, infiltrating tumor (IT) that corresponds to

the intermediate region between the cellular tumor and the leading edge and has a tumor cell to non-tumor cell ratio of 10-20/100; and the leading edge (LE) that is the boundary of the tumor with a tumor cell to non-tumor cell ratio of 1-3/100. We classified the 122 samples of the IVYGAP cohort using the original 50-gene signatures and the reduced 5-gene signatures. The fraction of each subtype by structure is shown in Figure 3C. Interestingly, all the LE structures were classified as NT as well as almost 75% of the IT samples. Most of these structures were classified as PN or CL by the 50-gene signatures. Besides, only one CT sample was classified as NT showing that the gene-signature detects samples with a high content of normal cells with high precision. It is worth mentioning that for CT, MVP and PAN structures, the results were highly coincident between the 5-gene and 50-gene signatures, with CT corresponding mainly to CL or PN and PAN and MVP mostly to ME, as reported elsewhere [20].

We used the results of the 50-gene signatures to obtain the simplicity scores of each sample. Figure 3D shows that LE is significantly enriched for lower simplicity scores in comparison to the other structures (Wilcoxon test, $p < 0.02$). This result confirms that the presence of a high content of normal cells in a sample can affect interpretation of the results.

The ssGSEA classification system performed a random permutation of the experimental data to generate a null distribution, obtaining a p value for the association of a sample to each subtype. Figure 3E shows the p value for NT obtained from samples of different structures. LE, the structure with the lowest content of tumoral cells, has significantly lower p values (Wilcoxon test, $p < 1e\text{-}7$). The p values increase slightly for IT and show median values close to 0.8 for structures found in the core tumor. Therefore, the

association with the NT subtype increases with the content of normal brain cells in the sample.

Two pathologists independently classified the HIS cohort according to: tumoral cell density, abundance of pathogenic blood vessels and presence of necrosis. Simultaneously, we used the expression levels obtained by qRT-PCR to classify the same cohort (Supplementary Table S4). Comparing the results, we found that of the 7 samples classified as infiltrating tumor by the pathologists, 5 (71%) were now classified as NT. The specificity of the NT class was 82%. We also observed that samples classified as CL and PN were indistinguishable in relation to the histological parameters. However, samples with an absence of pathogenic blood vessels and necrosis were mostly not classified as ME (80%). Figure 3F shows representative H&E stained histological images for each subtype.

Taken together, these results show that samples classified as NT have a lower tumoral cell density which is not due to immune cell infiltration, and that NT associates with samples or tumor regions with low cellularity.

Survival analysis

To study the clinical relevance of the simplicity score obtained by our model, we classified all the samples from the IDHWT cohort and discarded those that fell into the NT group. We, then, evaluated the optimal cutoff for the simplicity score considering the hazard ratio for each subtype using the *get.cutoff()* function described elsewhere [24]. Results can be seen in Supplementary Figure S3. Using the corresponding cutoff to divide samples into the PN and CL subtypes we observed a significantly higher survival for the simpler samples. The difference in median survival was found to be around 8 months in the PN subtype (HR = 0.54, 95% CI 0.32 – 0.93, log rank p-value = 0.02, Figure

4A) and 15.6 months for the CL subtype (HR = 0.38, 95%CI 0.21 – 0.7, log rank p-value = 0.001, Figure 4B). The opposite was observed for the ME subtype in that the simpler samples presented a poorer survival, with a difference in median survival of 2 months (HR = 2.18, 95% CI 1.23 – 3.88, log rank p-value = 0.007, Figure 4C).

Discussion

We developed a model based on the expression of 20 genes for the molecular classification of GBM samples. This model can detect samples with a high content of normal tissue, classifying them as NT, and also classifies the samples into PN, CL or ME. Although it uses only 5 genes per subtype, our model showed an overlap of 87% with the 50 gene per subtype model proposed elsewhere [7]. It also detects the main characteristic genetic alterations of the different subtypes [4,25] and the difference in OS compared with subtypes for simpler samples [7]. These results show that molecular classification of GBM can be performed in a cost-effective way and we hope that this model will encourage researchers and physicians to use this classification method more frequently in the future.

The NT gene signature shows 71% sensitivity and 82% specificity in the HIS cohort. On the other hand, for microdissected samples of specific tumoral regions, where the above-mentioned variability is absent, we found that NT classification is strongly associated with samples from the boundary region of the tumor (83.7% sensitivity, 98.7% specificity). This result is consistent with the association found between this region and the NE subtype [8,20]. It is noteworthy that the strength of the association between the different tumoral structures and NT, measured as an empirical p value, increases with the percent of normal cells. It remains to be tested whether or not the

empirical p value of NT can be used as an estimator of the percentage of normal brain cells in the sample. Further analyses are required to establish this.

We observed that NT samples are associated with low simplicity scores. Simplicity score was proposed as an estimate of the complexity of the GBM sample [7], where a low simplicity score indicates that the samples do not present a unique predominant subtype. Our results show that samples with a high content of normal cells can lead to an erroneous classification of a sample, considering it as one with high tumoral complexity, when actually it corresponds to a sample with low tumor purity. In fact, single cell RNA-seq analyses reveal that the subtype of a bulk tissue sample coincides with the subtype of the dominant cell population in the sample [30,31]. Moreover, it is important to know if the sample used for RNA extraction comes from the boundary of the tumor, because a low tumoral purity can alter gene expression measurements [9].

The incorporation of NT as a quality parameter of samples that are going to be classified brings important clinical and biological advantages, as discussed below. From a clinical point of view, Gill et al. suggested that the boundary of the tumor has to be classified as this is the tumoral region which cannot usually be resected [8]. The model we propose here not only classifies the infiltrating tumor mainly as NT, but also indicates the molecular subtype of the tumor as the second dominant group in that sample, which in 88% of the cases matches the molecular subtype of the cellular tumor. Therefore, when the piece of tissue used for molecular classification comes from the infiltrating tumor, our model can detect the molecular subtype of the tumor at that time. However, if the sample has a high content in normal cells and is classified as NT, parameters like the simplicity score should not be taken into consideration and special care should be taken in the interpretation of experiments like gene expression measurements.

448 It is also relevant that when different cellular tumor sections of the same patients were
449 analyzed we found that they were either PN or CL for all sections from the same patient.
450 Puchalski et al. maintained that CT were PN, CL or NE [20]. Their results agree with ours
451 if we consider NE to not be a real subtype. As reported in the cited work, we found that
452 microvascular proliferation regions and regions around necrosis were mainly ME. In the
453 same line, the ME subtype was reported to express markers of hypoxia and
454 microvascularity [32].
455 We can, therefore, regard GBM as a PN or CL tumor that evolves to ME in response to
456 different inputs, i.e. hypoxia [32]. Initially, the mesenchymal transition occurs in small
457 regions of the tumor, but these grow and eventually become the predominant subtype
458 (see SI Appendix Figure S4). This hypothesis is in line with our survival results. We
459 observed that complex samples, those with no clear dominant subtype, showed a worse
460 survival for PN and CL tumors; that is, when the ME regions of the tumor begin to grow
461 the complexity of the tumor increases with the corresponding survival consequence.
462 However, complex samples showed longer survivals for ME tumors, that is, a complex
463 ME sample is the continuation of the evolution of complex PN or CL samples, but when
464 the tumor becomes mainly ME it shifts to a low complexity ME sample with a worse
465 survival. Further analyses are needed to confirm this hypothesis.
466 Sottoriva et al. reported the intratumoral heterogeneity of GBM after observing that 6
467 out of 10 patients present regions of the same tumor with 2 or 3 different molecular
468 subtypes [33] Nevertheless, recovering the idea that NE is not actually a tumor subtype,
469 their results will be transformed into 5 out of 10 patients with 2 different molecular
470 subtypes: 2 cases with ME and PN and 3 cases with ME and CL. The proposed hypothesis
471 is in line with the shift of the ME subtype upon glioma recurrence [7,15]. Several works

have studied the PN-ME transition [34,35] but there is no evidence for a CL-ME transition.

We believe that our classification model, which takes into consideration tumor samples with a high content of normal tissue, will help to provide clinical insight into the different molecular subtypes of GBM and to better understand their biology.

Conclusions

In summary, we have developed a model which tackles both the classification of GBM samples into PN, CL or ME, and the detection of a high content of normal cells in a sample. The model shows an overlap of over 85% with the one proposed by Wang et al. and only requires the expression levels of 20 genes, making it a cost-effective alternative to other molecular classification models. The ability of our model to detect samples with high content of normal cells has been tested on microdissected regions of different GBM biopsies as well as on bulk tumor samples, contrasting the model results with the histological examination by two experts. We show the importance of determining the content of normal cells in GBM samples. Otherwise, normal tissue expression patterns can mask the expression patterns of other tumor types in the samples. This can lead to a misinterpretation of the results as we show with the simplicity score but can also affect the conclusions of tumor heterogeneity studies, among others.

Key points

- In this work, we present a cost-effective model based on the expression of 20 genes, which can classify GBM samples into Proneural, Classical and Mesenchymal subtypes.

- The model incorporates a quality parameter that detects samples with a high content of normal tissue, preventing errors in the classification and interpretation of the results in clinical practice.
- Our results show that considering the abundance of normal cells in a sample can shed light on the interpretation of survival, tumor evolution or tumor heterogeneity.
- As the expression of 20 genes can be measured by qRT-PCR we believe that a greater volume of GBM samples will be classified and reported in the future.

Funding

This work was supported by grants from the 'Fondo de Investigaciones Sanitarias' (FIS) [PI17-01489], the Miguel Servet Program [CP11/00147] del Instituto de Salud Carlos III (AAS), and the Ministerio de Economía y Competitividad–FEDERER [RTC-2016-4990-1]. RM was supported by "Convocatoria de ayudas para la contratación de investigadores predoctorales e investigadores postdoctorales cofinanciadas por Fondo Social Europeo a través del Programa Operativo de Empleo Juvenil y la Iniciativa de Empleo Juvenil (YEI)".

Author contributions

R.M. participated in experimental design, performed the in-silico analysis and wrote the manuscript. N.G.R participated in experiments and revised the manuscript. A.H.L. and F.P.R. performed the histological examination of the samples. M.Z. and E.M. participated in the computational analysis, design and revision. B.J., A.C., C.F.C. and R.P provided patients samples. A.A.S participated in experimental design and was responsible for the financial support, and edited, revised and approved the final version of the manuscript.

Aknowledgments

We thank Victor González Rumayor for helping with the Tissue Microarray logistic and Marcos García Lorenzo from the Modeling and Virtual Reality Group of the Rey Juan Carlos University for helping with the computational equipment.

We want to particularly acknowledge the patients in this study for their participation and to the HGM BioBank integrated in RETICS, National Network Biobanks, and collaborating Centers for the generous gifts of clinical samples used in this work. The HGM BioBank, integrated in National Network Biobanks, is supported by Instituto de Salud Carlos III, Spanish Science and Innovation Ministry (Grant nº PT17/0015/0042)".

References

1. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 2016; 131:803–820

2. Ostrom QT, Gittleman H, Fulop J, et al. CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. Neuro-oncology 2015; 17 Suppl 4:iv1–iv62

3. Lee E, Yong RL, Paddison P, et al. Comparison of glioblastoma (GBM) molecular classification methods. Seminars in Cancer Biology 2018;

4. Verhaak R, Hoadley KA, Purdom E, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17:98–110

5. Chen R, Smith-Cohn M, Cohen AL, et al. Glioma Subclassifications and Their Clinical Significance. Neurotherapeutics 2017; 14:284–297

6. Erdem-Eraslan L, Bent MJ van den, Hoogstrate Y, et al. Identification of Patients with Recurrent Glioblastoma Who May Benefit from Combined Bevacizumab and CCNU Therapy: A Report from the BELOB Trial. Cancer Res 2016; 76:525–534

7. Wang Q, Hu B, Hu X, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell 2017; 32:42-56.e6

8. Gill BJ, Pisapia DJ, Malone HR, et al. MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. Proc National Acad Sci 2014; 111:12550–12555

9. Ridder D de, Linden CE van der, Schonewille T, et al. Purity for clarity: the need for purification of tumor cells in DNA microarray studies. Leukemia 2005; 19:618–627

10. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology 2012; 30:413

11. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications 2013; 4:2612

12. Network TCGAR, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013; 45:1113

13. Bowman RL, Wang Q, Carro A, et al. GlioVis data portal for visualization and analysis of brain tumor expression datasets. Neuro-oncology 2016; 19:139–141

14. Gautier L, Cope L, Bolstad BM, et al. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004; 20:307–315

15. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 2006; 9:157–173

16. Sturm D, Witt H, Hovestadt V, et al. Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. Cancer Cell 2012; 22:425–437

17. Bao Z-S, Chen H-M, Yang M-Y, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res 2014; 24:1765–1773

18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118–127

19. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015; 21:nm.3967

20. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. Science 2018; 360:660–663

21. Pollard KS, Dudoit S, Laan MJ van der. Multiple Testing Procedures: the multtest Package and Applications to Genomics. Springer 2005;

22. Dabney AR. ClaNC: point-and-click software for classifying microarrays to nearest centroids. Bioinformatics 2006; 22:122–123

23. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research 2016; 44:e71–e71

24. Budczies J, Klauschen F, Sinn BV, et al. Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization. Plos One 2012; 7:e51862

25. Brennan CW, Verhaak R, McKenna A, et al. The Somatic Genomic Landscape of Glioblastoma. Cell 2013; 155:462–477

587 26. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular Profiling Reveals Biologically
588 Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell 2016; 164:550–563
589 27. Lai A, Kharbanda S, Pope WB, et al. Evidence for Sequenced Molecular Evolution of
590 IDH1 Mutant Glioblastoma From a Distinct Cell of Origin. J Clin Oncol 2011; 29:4482–
591 4490
592 28. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set
593 enrichment analysis web server 2016 update. Nucleic Acids Res 2016; 44:W90–W97
594 29. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list
595 enrichment analysis tool. Bmc Bioinformatics 2013; 14:128
596 30. Darmanis S, Sloan SA, Croote D, et al. Single-Cell RNA-Seq Analysis of Infiltrating
597 Neoplastic Cells at the Migrating Front of Human Glioblastoma. Cell Reports 2017;
598 21:1399–1410
599 31. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral
600 heterogeneity in primary glioblastoma. Science 2014; 344:1396–1401
601 32. Jin X, Kim LJY, Wu Q, et al. Targeting glioma stem cells through combined BMI1 and
602 EZH2 inhibition. Nat Med 2017; 23:1352–1361
603 33. Sottoriva A, Spiteri I, Piccirillo SGM, et al. Intratumor heterogeneity in human
604 glioblastoma reflects cancer evolutionary dynamics. Proc National Acad Sci 2013;
605 110:4009–4014
606 34. Segerman A, Niklasson M, Haglund C, et al. Clonal Variation in Drug and Radiation
607 Response among Glioma-Initiating Cells Is Linked to Proneural-Mesenchymal Transition.
608 Cell Reports 2016; 17:2994–3009

609 35. Halliday J, Helmy K, Pattwell SS, et al. In vivo radiation response of proneural glioma
610 characterized by protective p53 transcriptional program and proneural-mesenchymal
611 shift. Proc National Acad Sci 2014; 111:5248–5253

612 Figure legends

613 Figure 1 (at Results/Transcriptomic data aggregation)

614 Figure1: A) Schematic representation of the process followed to generate the new
615 model including data aggregation, reduction in the number of genes per gene signature
616 and the incorporation of a fourth gene signature to detect samples with a high content
617 of normal tissue. B) Z-score and overlap achieved when reducing the number of genes
618 per subtype for the rank criterion based on differences in mean expression (red). Mean
619 and standard deviation (black) used to obtain the Z-score belong to the null distribution,
620 obtained from the results of a thousand randomly ordered genes. C) Comparison of the
621 distribution of the simplicity score for the four subtypes obtained from Verhaak's
622 classification [4]. P values are obtained from the comparison to the neural (NE) subtype
623 using Wilcoxon test. D) Comparison of the distribution of the simplicity score for the four
624 subtypes obtained by our model. P values are obtained from the comparison to the high
625 content in normal cells (NT) group using Wilcoxon test.

626 Figure 2 (at Results/ Validation of the model)

627 Figure 2: A) Comparison of the classification obtained by Wang's [7] model and from our
628 model. B) ROC space showing the sensitivity and specificity obtained by our model for
629 each subtype using Wang's model as gold standard. Error bars represent the standard
630 deviation. Values in parenthesis correspond to the F1 score of each subtype. C)
631 Frequency of somatic genomic alterations for each subtype. Significance values were

obtained applying the Chi-squared test. D) Survival curves between subtypes for samples with the top 20% simplicity score.

Figure 3 (at Results/ NT associates with abundance in normal cells)

Figure 3: A) Tumor purity of TCGA-IDH-WT samples determined by ABSOLUTE and ESTIMATE, respectively. The difference in tumor purity between subtypes was evaluated by the Wilcoxon test. B) Tumor purity of the validation cohort determined by ESTIMATE. The difference in tumor purity between subtypes was evaluated by the Wilcoxon test. C) Frequency of the molecular classification for samples obtained by microdissection from different tumor structures. The results from Wang's classification are shown in gray and the results from our model in red. D) Simplicity score obtained by Wang's model for the different tumor structures. P values are obtained for the comparison of each structure against the leading edge evaluated by Wilcoxon test. E) Empirical p values associated with the NT group obtained for the different tumor structures. P values are obtained for the comparison of each structure against the leading edge evaluated by Wilcoxon test. F) Representative H&E stained histological images for samples that belong to each molecular subgroup as classified by our model. Scale bars represent 200 μm.

Figure 4 (at Results/Survival analysis)

Figure 4: Survival curves between optimal cut-off of the simplicity score within proneural (A), classical (B) and mesenchymal (C) samples.

Supplementary material legends

Table S1

Table S1: Primers used for qRT-PCR.

Table S2

Table S2: Differentially expression analysis of the NE genes for the NE vs non-NE subtypes and for tumor vs non-tumor tissue.

Table S3

Table S3: Results obtained from the enrichment analysis of the 5 gene signature for NT against tissue databases using Enrichr.

Table S4

Table S4: Histological examination of the TMAs from tissue samples and the classification obtained by the model based on qRT-PCR measurements.

Figure S1

Figure S1: Z-score and overlap achieved when reducing the number of genes per subtype for the rank criterion based on relative differences in mean expression (A) and on the statistic obtained from the differential expression analysis (B). Mean and standard deviation (black) used to obtain the Z-score belong to the null distribution, obtained from the results of a thousand randomly ordered genes.

Figure S2

Figure S2: True positive rate and true negative rate obtained from different gene signature sizes for proneural (A), classical (B) and mesenchymal (C) subtypes using Wang et al. model as gold standard of the molecular classification. Black dots represent the mean and standard deviation of 1,000 randomly ranked gene lists. Three different criteria to rank the genes were used: mean differences (red dots), fold change (blue dots) and statistical significance of the mean differences (green dots).

Figure S3

Figure S3: Hazard ratio optimal cutoff selection of the simplicity score for proneural (A), classical (B) and mesenchymal (C). Continuous line represents the hazard ratio and dashed lines the 95% confidence interval.

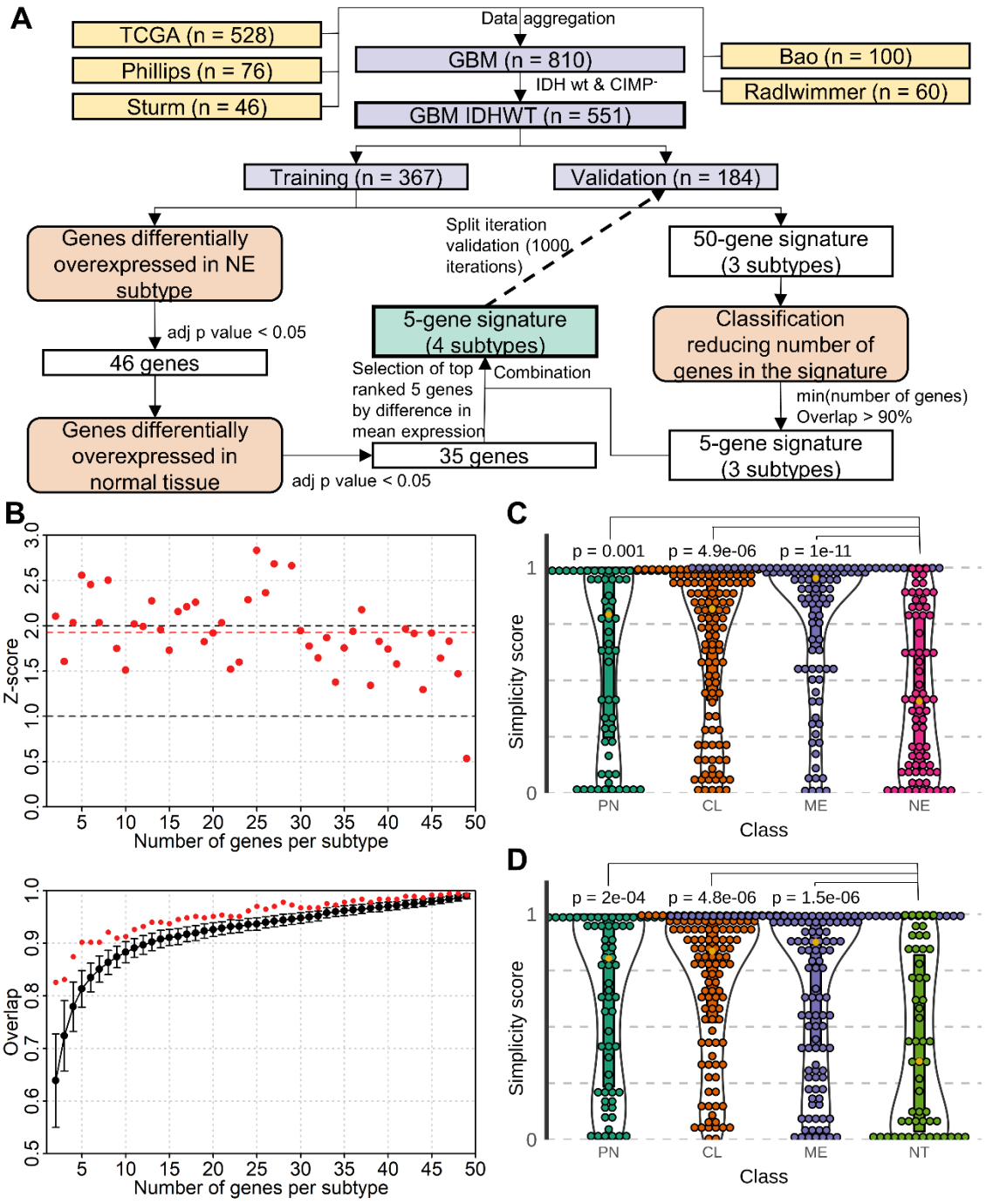
Figure S4

Figure S4: Schematic model of the evolution of a GBM tumor from a cellular tumor (proneural or classical) to a mesenchymal tumor.

Methods S1

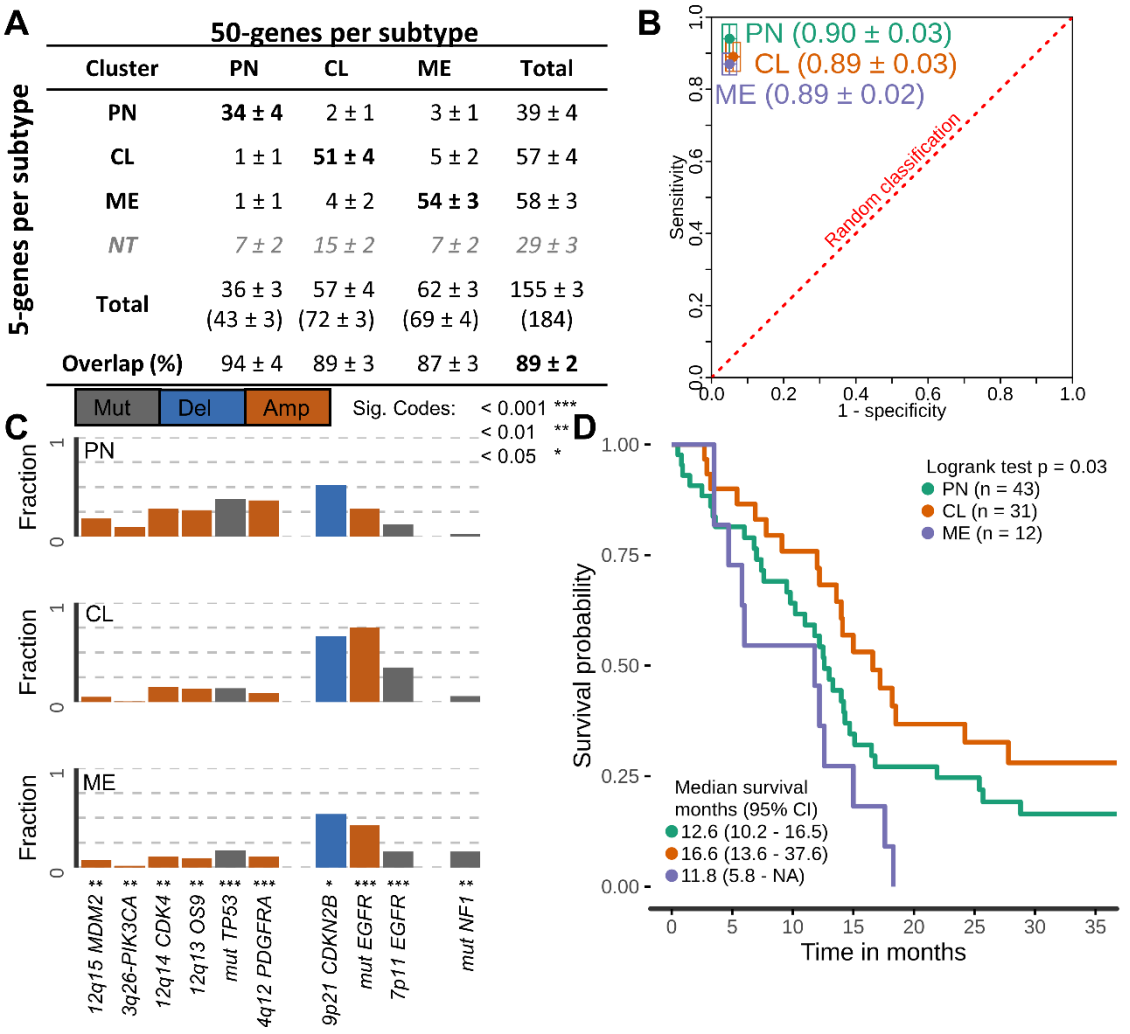
Methods S1: Blocks of code used for the development of the model.

687 Figure 1



688

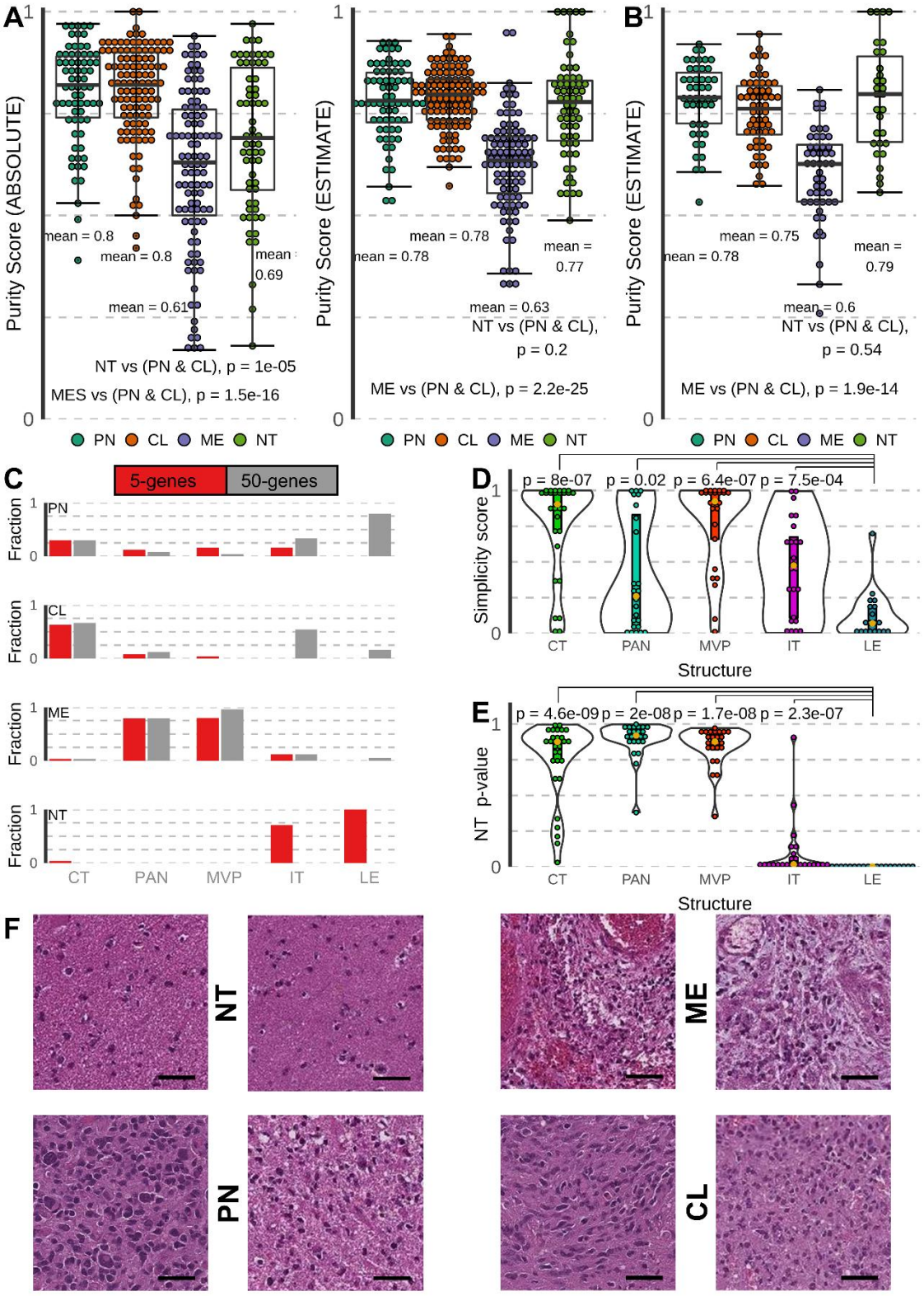
689 Figure 2



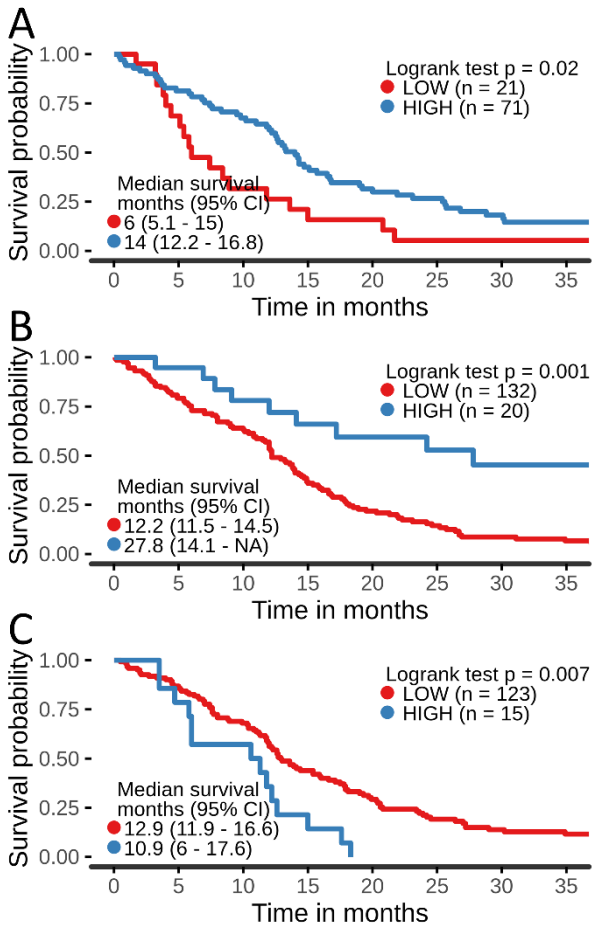
690

691

Figure 3


A
Purity Score (ABSOLUTE)
mean = 0.8
mean = 0.8
mean = 0.61
mean = 0.69
NT vs (PN & CL), p = 1e-05
MES vs (PN & CL), p = 1.5e-16
PN CL ME NT
Purity Score (ESTIMATE)
mean = 0.78
mean = 0.78
mean = 0.63
mean = 0.77
NT vs (PN & CL), p = 0.2
ME vs (PN & CL), p = 2.2e-25
PN CL ME NT
B
Purity Score (ESTIMATE)
mean = 0.78
mean = 0.75
mean = 0.6
mean = 0.79
NT vs (PN & CL), p = 0.54
ME vs (PN & CL), p = 1.9e-14
PN CL ME NT
C
5-genes
50-genes
Fraction
PN
CL
ME
NT
CT PAN MVP IT LE
D
Simplicity score
p = 8e-07 p = 0.02 p = 6.4e-07 p = 7.5e-04
CT PAN MVP IT LE
Structure
E
NT p-value
p = 4.6e-09 p = 2e-08 p = 1.7e-08 p = 2.3e-07
CT PAN MVP IT LE
Structure
F
NT
ME
PN
CL

694 Figure 4



695

696

697 Table S1

Human primers	Forward 5´------3´	Reverse 5´------3´
GPR17	GGGCTTGTGATGGCTACAAT	CTGCCTTCAGGGTCTTTCTG
CA10	ATCCCACCTCAGTGAAATGC	TCATGAAGAAAGGGCCAATC
UGT8	TACTCTCCCACCAGGAGCTG	CCTTTTAACGGCAACATCGT
HRASLS	TCTTCTCATTCCTGGGCTTG	TTCCTCCTCCCAAATTCCTT
RAB33A	GGAAGGTGCAGAAACTGGAG	GGAAACAAGCAGGTGTCAG
ELOVL2	TCTTACCAAAGTGCGTTCCA	CTCCCTCCTTGCCATACAGA
MLC1	CGTAACAGCAGGAGCATGG	TCTGGTCAGGTCCAGAGAGC
SLC4A4	TCAAGACACAGACACGCACA	GGGACTCTGTCTGGAGGTCA
CDH4	GGACACCTGTCACCCTGAGT	GAGAGTGTCCTGGGGTTTGA
FGFR3	TGCCCCTCAGAGACTGAAAT	TCCGTTGTACCAGCCTTTTC
LUM	TGGAGCCAAATGTTATGCAG	GAAAGGCCGCTGTACCATAA
PI3	GCAAGAGCCAGTCAAAGGTC	TTCTTGATTCCTGGGCAGTC
SLPI	CATATGGAGGAGGCTCTGGA	TCTTGAAAGCCTGCTGTGTG
CYP1B1	CTCCTGTGGAAGGCAGAGAA	TCCCCAACTCTTGTCACCTC
NNMT	ACCTTGCAGTGCCTCACTTT	CAAGCAATCTGTCTGCCTCA
CCK	TACATGGGCTGGATGGATTT	GTGAGGTGTGTGGTTGCACT
CRYM	GAATGGCAGTGGAAGACACA	GGGACTGGACTCCCTCATTT
SERPINI1	GACGAGTCATGCATCCTGAA	CCAGTTGCAAACATAATGTGC
KCNK1	CTGCAAACCATTGAGCGTAG	TGGGGTCACAGCTTCTTTGT
GPR22	CTCCCATTCTGGAAATCAACA	GCCAAGTCCCAACACAATTT

698

699

700 Table S2

	NEURAL VS NON-NEURAL		BRAIN VS TUMOR	
Gene	**Difference in mean expression**	**Adjusted p value**	**Difference mean expression**	**Adjusted p-value**
CCK	1,449	4,0E-04	6,128	< 1E-04
CRYM	1,306	2,8E-03	5,6	< 1E-04
SERPINI1	1,166	2,0E-03	4,521	< 1E-04
KCNK1	0,895	4,4E-02	4,43	< 1E-04
GPR22	0,461	6,4E-03	4,032	< 1E-04
HPCAL4	0,516	1,4E-02	3,797	< 1E-04
CPNE6	0,342	5,6E-03	3,13	< 1E-04
CA4	0,534	4,4E-03	2,421	< 1E-04
UROS	0,441	1,7E-02	2,067	< 1E-04
KCNJ3	0,184	4,0E-04	2	< 1E-04
DHRS9	1,183	< 1E-04	1,93	< 1E-04
MDH1	0,345	4,0E-04	1,771	< 1E-04
ANXA3	0,94	4,0E-04	1,583	< 1E-04
CRYZL1	0,518	< 1E-04	1,311	< 1E-04
MGST3	0,48	< 1E-04	1,234	< 1E-04
SNCG	0,184	4,0E-04	1,173	< 1E-04
ACYP2	0,669	< 1E-04	1,168	< 1E-04
YPEL5	0,451	< 1E-04	1,161	< 1E-04
CLCA4	0,422	4,6E-02	1,052	< 1E-04
PEX11B	0,248	1,2E-02	0,947	< 1E-04
ADD3	0,655	< 1E-04	0,918	< 1E-04
MYBPC1	1,431	4,0E-04	0,884	< 1E-04
CASQ1	0,4	6,4E-03	0,884	< 1E-04
SEPW1	0,39	1,6E-03	0,851	< 1E-04
CRBN	0,388	2,0E-03	0,666	< 1E-04
ANXA7	0,578	< 1E-04	0,654	< 1E-04
TMEM144	0,513	1,7E-02	0,59	< 1E-04
TCEAL1	0,394	4,0E-04	0,547	< 1E-04
COX5B	0,346	3,2E-03	0,547	< 1E-04
TTC1	0,3	4,0E-04	0,533	< 1E-04
GUK1	0,381	< 1E-04	0,528	< 1E-04
PEX19	0,245	1,9E-02	0,373	< 1E-04
IMPA1	0,369	2,2E-02	0,32	< 1E-04
RBKS	0,699	< 1E-04	0,258	7,2E-03
MAT2B	0,421	< 1E-04	0,204	7,2E-03
CRYL1	0,703	4,0E-04	NS	NS
SEPP1	0,701	< 1E-04	NS	NS
MRPL49	0,532	< 1E-04	NS	NS
LYRM1	0,503	2,4E-03	NS	NS
TSNAX	0,447	4,0E-04	NS	NS
ATP5L	0,366	< 1E-04	NS	NS
AKR7A3	0,345	6,4E-03	NS	NS
SNX11	0,336	4,0E-04	NS	NS
CCDC121	0,334	2,4E-03	NS	NS
ATP5F1	0,223	1,3E-02	NS	NS
NSL1	0,206	4,4E-02	NS	NS

701

702

703 Table S3

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combine score
1	BRAIN (BULK)	0.00002074	0.002240	8.64	93.12
2	CEREBRAL CORTEX	0.00002074	0.001120	8.64	93.12
3	CINGULATE GYRUS	0.00002074	0.0007467	8.64	93.12
4	DENTATE GRANULE CELL	0.00002074	0.0005600	8.64	93.12
5	DORSAL STRIATUM	0.00002074	0.0004480	8.64	93.12
6	SUPERIOR FRONTAL GYRUS	0.00002074	0.0003734	8.64	93.12
7	ATRIUM	0.01294	0.1997	5.18	22.52
8	HEART (BULK TISSUE)	0.01294	0.1747	5.18	22.52
9	PREFRONTAL CORTEX	0.01294	0.1553	5.18	22.52
10	VENTRICLE	0.01294	0.1398	5.18	22.52

704

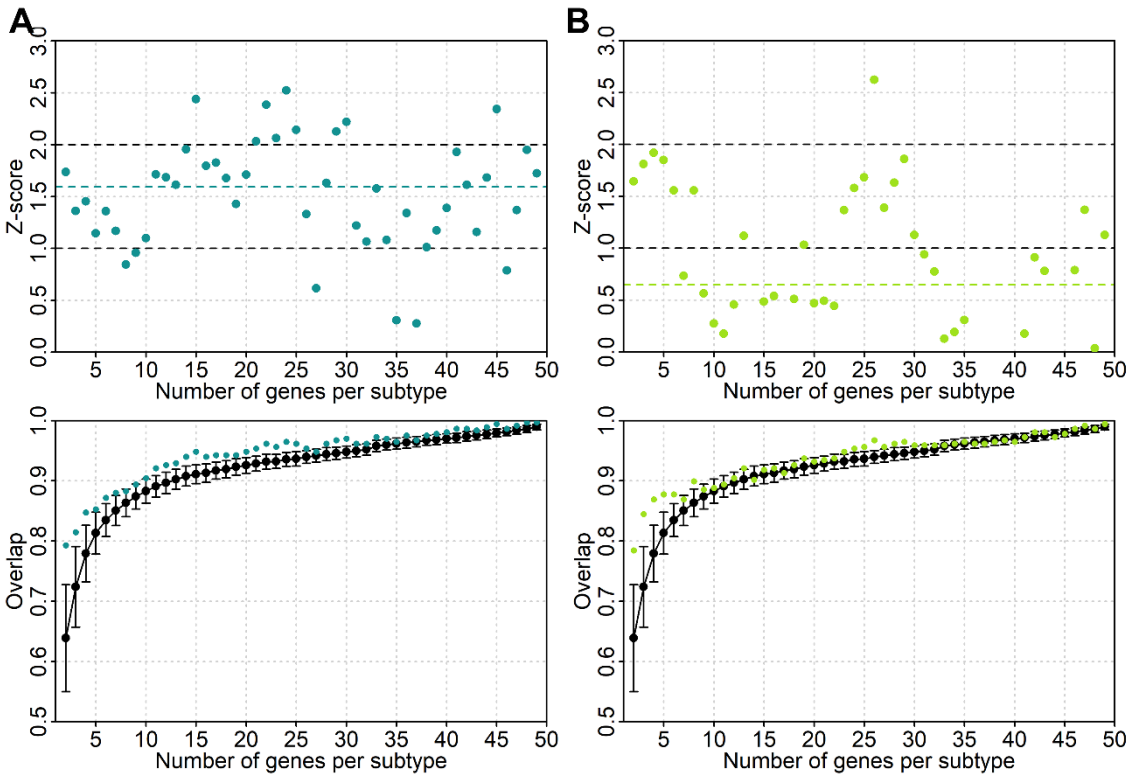
705

706 Table S4

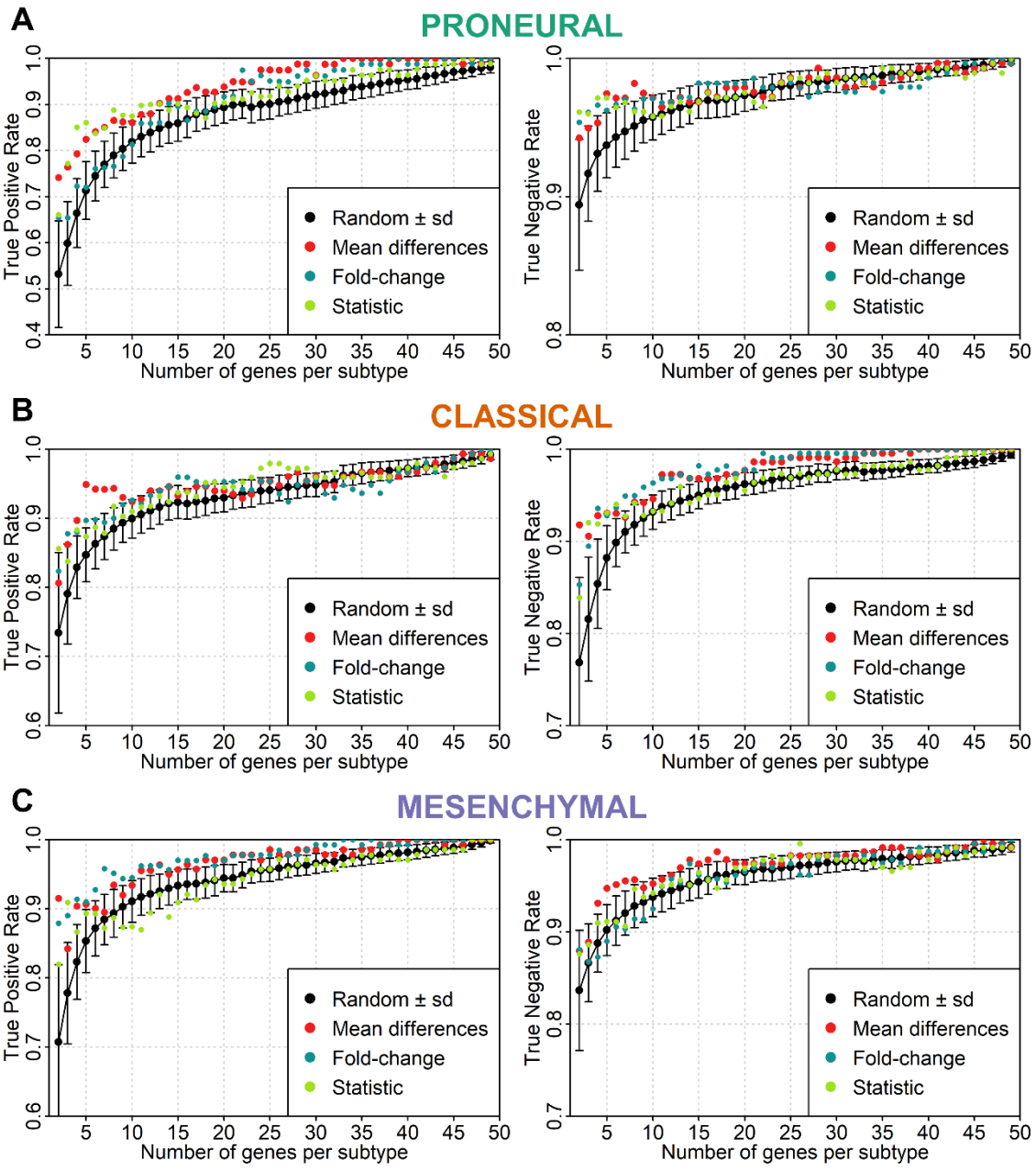
Sample	Cellularity	Pathogenic blood vessels	Necrosis	Molecular classification
S1	MODERATE / MODERATE	YES / YES	NO / NO	ME
S2	LOW / LOW	YES / YES	YES / NO	CL
S3	MODERATE / HIGH	YES / YES	NO / NO	ME
S4	LOW / LOW	NO / NO	NO / NO	NT
S5	MODERATE / MODERATE	YES / YES	YES / NO	ME
S6	LOW / LOW	YES / YES	YES / NO	ME
S7	HIGH / HIGH	YES / YES	NO / NO	ME
S8	MODERATE / MODERATE	YES / YES	YES / NO	ME
S9	LOW / LOW	NO / NO	NO / NO	PN
S10	MODERATE / MODERATE	YES / YES	YES / NO	ME
S11	LOW / LOW	YES / YES	NO / NO	NT
S12	LOW / LOW	YES / YES	YES / NO	CL
S13	MODERATE / MODERATE	YES / NO	YES / NO	CL
S14	MODERATE / MODERATE	YES / YES	YES / NO	ME
S15	LOW / LOW	NO / NO	NO / NO	PN
S16	HIGH / MODERATE	YES / YES	YES / NO	ME
S17	HIGH / HIGH	NO / NO	NO / NO	PN
S18	MODERATE / MODERATE	YES / NO	NO / NO	ME
S19	LOW / LOW	NO / NO	NO / NO	NT
S20	LOW / LOW	NO / NO	NO / NO	NT
S21	LOW / HIGH	YES / YES	YES / NO	NT
S22	MODERATE / MODERATE	NO / NO	NO / NO	ME
S23	HIGH / HIGH	YES / NO	NO / NO	NT
S24	HIGH / HIGH	YES / YES	NO / NO	ME
S25	HIGH / HIGH	YES / NO	NO / NO	ME
S26	LOW / LOW	NO / NO	NO / NO	NT
S27	HIGH / HIGH	NO / NO	NO / NO	CL
S28	MODERATE / MODERATE	YES / NO	YES / YES	NT
S29	HIGH / MODERATE	YES / NO	YES / NO	ME
S30	LOW / HIGH	NO / NO	YES / NO	PN
S31	HIGH / MODERATE	YES / YES	YES / YES	CL
S32	HIGH / MODERATE	YES / NO	YES / NO	CL
S33	LOW / MODERATE	NO / NO	NO / NO	NT
S34	HIGH / MODERATE	YES / YES	NO / YES	CL

707

Figure S1


A
B
Z-score
Number of genes per subtype
Overlap
Number of genes per subtype

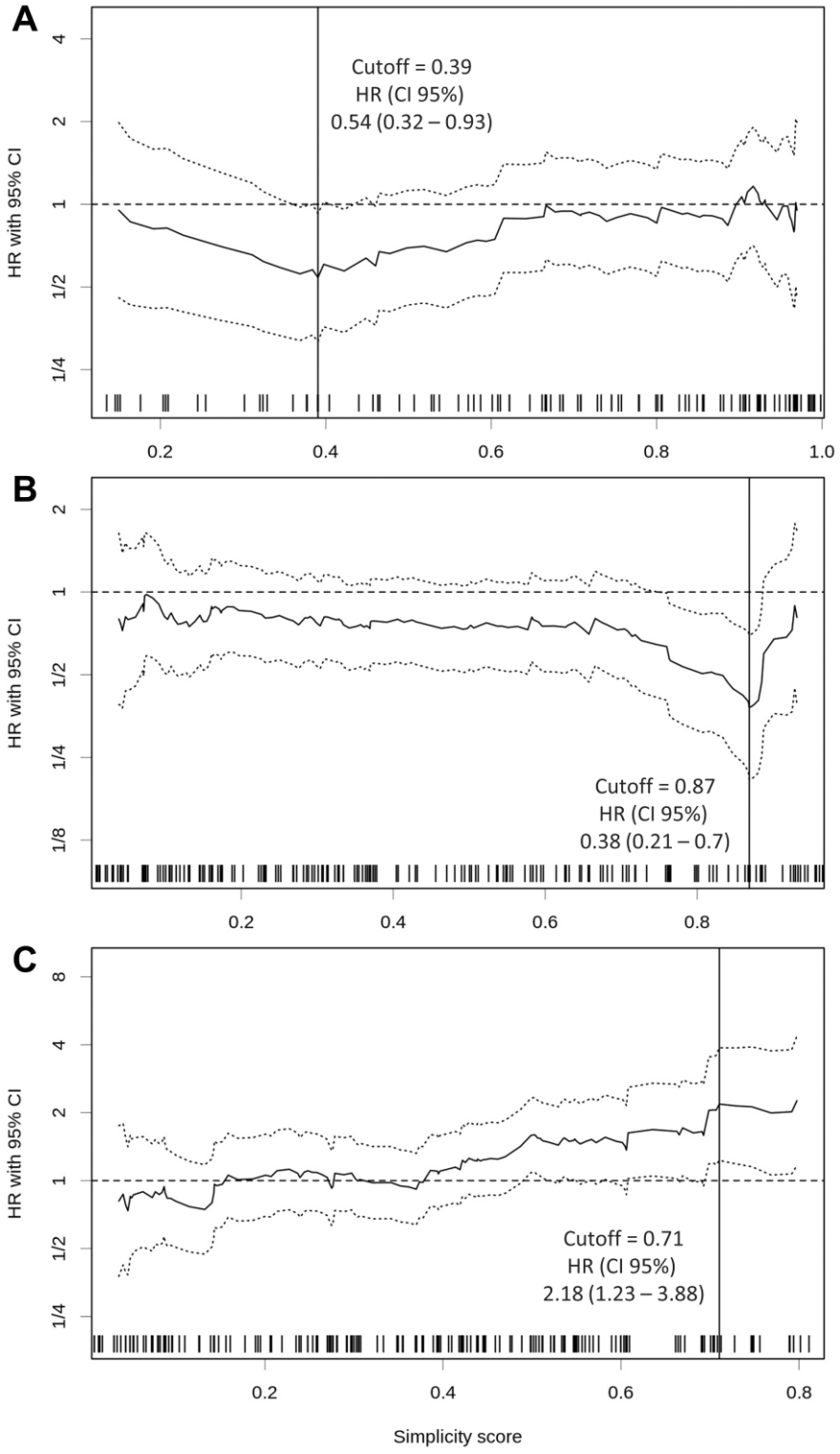
711 Figure S2



712

713

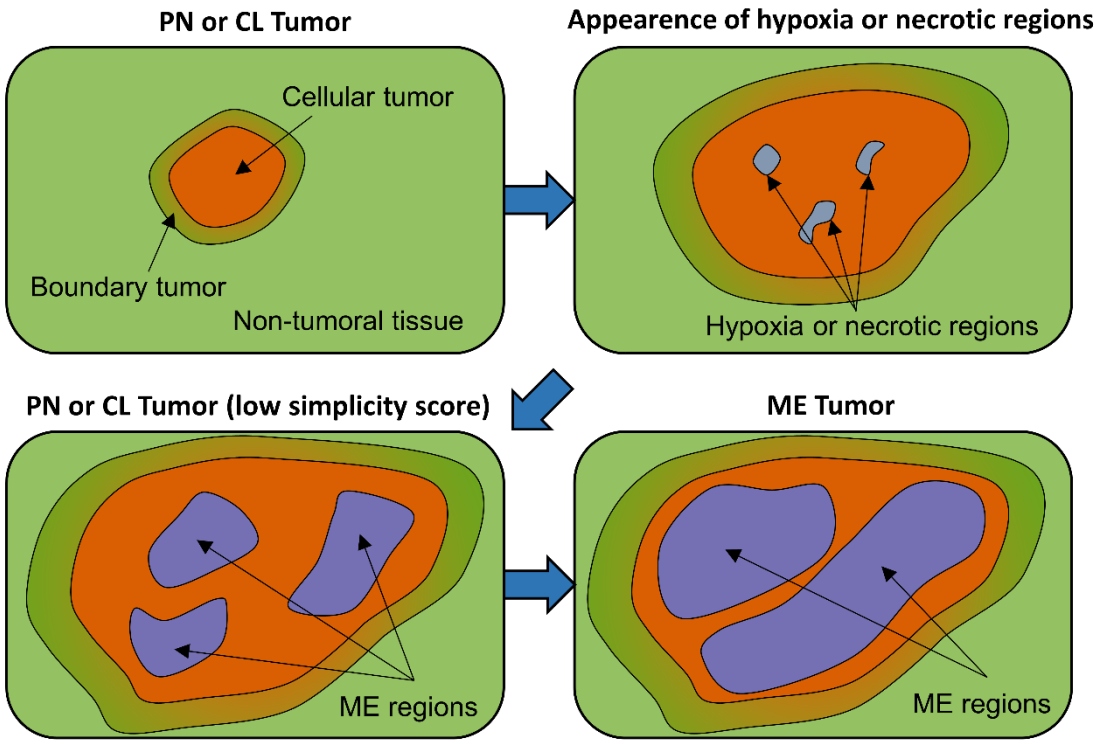
714 Figure S3



715

716

Figure S4


PN or CL Tumor
Cellular tumor
Boundary tumor
Non-tumoral tissue
Appearence of hypoxia or necrotic regions
Hypoxia or necrotic regions
PN or CL Tumor (low simplicity score)
ME regions
ME Tumor
ME regions