Supplementary Information for:

Improved Grading of human astrocytic brain tumours by artificial neural network analysis of gene expression microarray data.

Lawrence P. Petalidis, Anastasis Oulas, Magnus Backlund,Matthew T. Wayland, Lu Liu, Karen Plant, Lisa Happerfield, Tom C. Freeman , Panayiota Poirazi & V. Peter Collins.

Calibration and leave-one-out cross-validation of the ANN models

Calibration refers to the process of optimization of network weights and parameters during training in order to achieve the desired output from a given network input. To calibrate ANN models, predefined training samples from the dataset under investigation were used to train the network. Training was achieved using a leave-one-out cross-validation procedure whereby one sample was left out and all necessary training and feature (gene/probe set) selection was performed on the remaining samples. It is important to note that for a true predictive Leave-one-out cross-validation assessment, gene screening and selection must be performed repeatedly for every leave-one-out run**.** In each of the leave-one-out runs, the analysis leads to a different subset of genes. These subsets are highly redundant but nevertheless differ in some genes, reflecting sample variability and inherent heterogeneity in expression profiles. Previous studies [1],[2] have selected genes using all the training samples, a choice of method that will most likely bias the validation results causing an over-estimation of the validation performance and consequently lead to over-fitting of the data. Over-fitting refers to the inability of the trained models to generalize between the training set and the test set. Even more ambiguous results are obtained when test results are included in the gene selection procedure[3]. In the present study it was assured that every time a sample was left out for validation; training and gene selection were performed on the remaining samples. The left-out-sample was subsequently classified using the genes selected, weights and network parameters saved during training.

Training Methodology and Classification[a]

Leave-one-out cross-validation was attempted using increasing numbers of genes, as ranked by signal-to-noise method (S2N) (see Gene Selection) and the procedure was halted at a maximum of 20 genes. A preprocessing step (or standardization step) was also performed whereby the selected expression profiles of genes were standardized to zero mean and unit standard deviation. We found that optimum and most consistent results were achieved by executing the Leave-one-out cross-validation procedure 100 times (100 *epochs*) for every training sample and then taking an average of the results. Neural networks allow for the capacity to store trained network models, like the ones produced during Leave-one-out cross-validation, allowing these models to later be tested on "unseen data", or data that the network has not been trained with. The total number of trained models depends on the number of training samples in the training model. Hence, for every model type the total number of trained models is 100 x N_i – where N_i is the number of training samples.

The classification problem was split into (k(k-1))/2 different two-class/grade problems where k is the number of classes/grades (***all-pairs*** approach – Fig B).

[a] The term *classification* refers to the general terminology applicable to machine learning algorithms, such as ANNs. For the purposes of this study *classification* can refer to categorization of different classes of tumours, grades of tumours (grading) and/or normal vs. tumour tissue. The term *class* is used in a general context and may represent a particular tumour grade, as is the case with the astrocytic tumour dataset.

Consequently the network was used to obtain 3 different types of models or binary classifiers.

For each validation sample in the two class/grade problem under consideration, an average over 100 model outputs was computed. Samples were classified as class 0 or class 1, depending on whether the average was closer to 0 or 1. The threshold for this decision was kept at 0.5.

The use of an aggregate of neural networks to obtain an average classification is a very powerful tool in achieving better classification performance. A better statistical interpretation of the results is obtained especially when dealing with complex datasets such as microarray data.

One of the advantages of artificial neural networks is that they can be trained to classify samples into multiple classes or grades by increasing the number of output neurons in the network final layer. However, this may have drawbacks when attempting to classify smaller numbers of classes/grades (true for the astrocytic tumour dataset) as neurons will compete with each other with a substantial favoritism towards neurons for the larger classes/grades. In order to adopt a neural network that significantly and successfully recognizes the boundaries of multiple classes/grades, it is more intuitive to split the data into multiple binary problems and use these to train multiple networks and then obtain an average committee voting form these trained "experts". So in the case of the astrocytic tumour datasets rather than simply increasing the number of output neurons to match the number of classes or grades, we use an ***all-pairs*** approach instead. This approaches is usually used to enable a binary classifier applicable to multi-class problems, and have been recently used in support vector machines [2]. It is anticipated [1] that splitting a problem into different sub-tasks and obtaining a mixture of *experts*, may improve classification substantially.

Finally the trained models from Leave-one-out cross-validation that produced the best results for a specific number of selected genes were later used for testing.

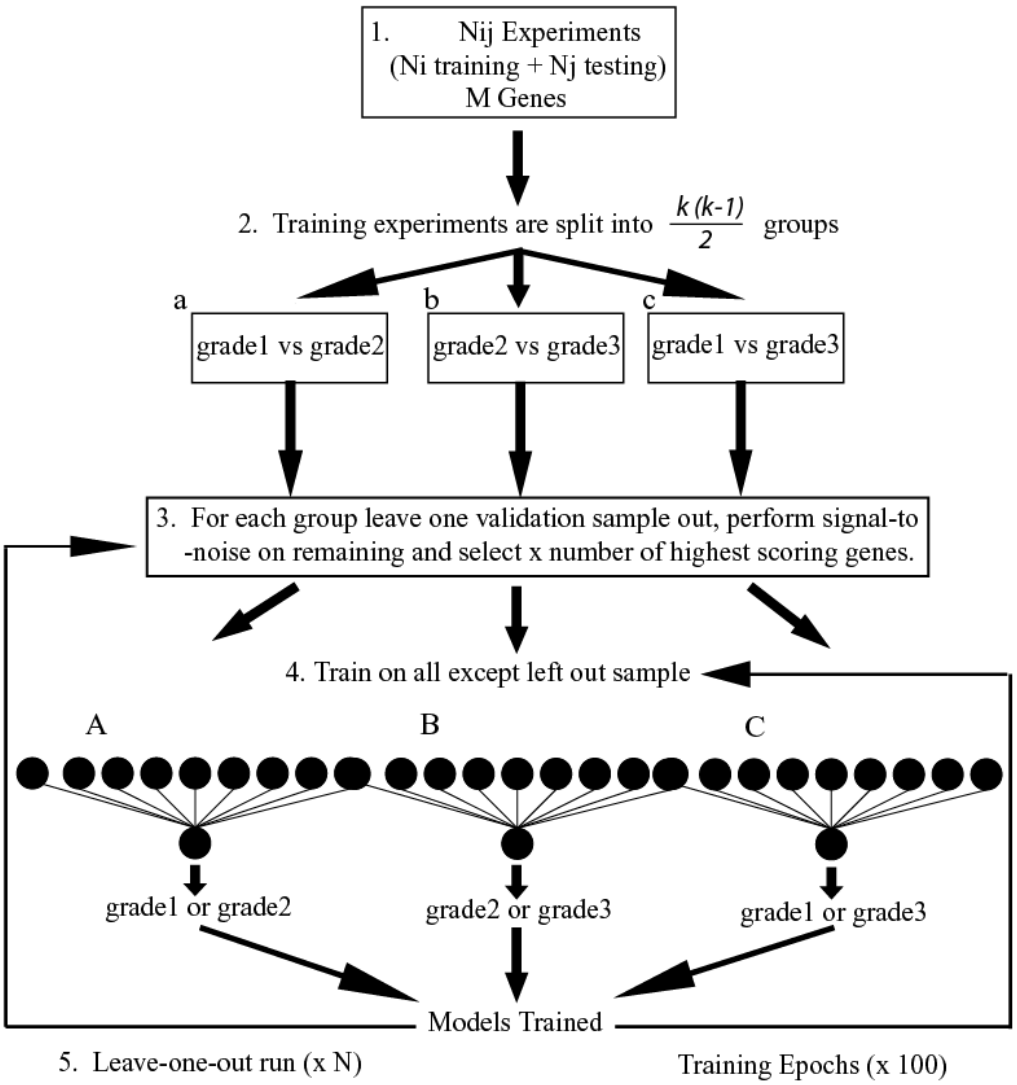


Figure A. Schematic illustration of the *all-pairs* approach using the proposed artificial neural network on a three grade problem such as the one faced for the astrocytoma dataset. The entire dataset was split into a test set and a training set (1). Next, the test experiments were set aside and the training experiments were partitioned into (k(k-1))/2 groups, in this case 3, where k is the number of grades(2): *a – grade1 vs. grade2, b - grade2 vs. grade3* and *c – grade1 vs. grade3*. Each of these groups was used to train three different types of models – *A, B and C* respectively. A leave-one-out cross-validation procedure was undertaken for validation and calibration of the networks. A number of *x* genes were selected using signal-to-noise, where *x* can take the value of any even number between 2 and 20 (3). For each left out sample, the ANN models were calibrated using as input; the *x* highest scoring genes selected from the remaining samples and as output; the tumor grade. The 9-gene case (x =9) is displayed here as there are 9 input nodes (4). For each model type *(A, B and C)* the calibration was optimized with 100 iterative cycles (epochs). The next training sample in line was then left out and the entire training process repeated (5). For each left out sample 100 models were calibrated, resulting in a total of 100 x N_{ic} trained models for every group (where c corresponds to the different grades c=1,2,3, and N_{ic} is the number of training samples for each grade). The test experiments were subsequently classified using all the calibrated/trained models.

Testing

Testing was performed by selecting the corresponding gene indexes from the test set, standardizing the test data and then propagating every test sample through all the trained models. The genes corresponding to the different trained models may differ as a different sample was left out during Leave-one-out cross-validation and gene selection performed on the remaining samples. Hence, during testing the specific genes corresponding to a given model were selected form the test set every time a test sample was propagated through each individual model. Then an average was taken over the network outputs. If the average output from the trained models is more than 0.5 the sample was classified to class 1, if the average output was less than 0.5 then the sample was considered to be classified to class 0. If the average output from the trained models was exactly 0.5 the test sample was considered to be undetermined. This method classifies each test sample to one of the two classes present in the trained models. For the all-pairs approach used here in order to finally place the test sample to one of the 3 categories, a committee vote is taken from the 3 types of models. For each class there are 2 relevant models which distinguish it from the others. However, since we do not know which these are; the outputs from all binary classifiers were recorded, and a sample was classified to the class that the majority of the outputs agree on. In cases where all 3 predictions were in disagreement, the sample was placed to the class denoted by the model with the most significant output value.

Test tumour ID	GB vs AA Model A	Model A Voting	AA vs A Model B	Model B Voting	GB vs A Model C	Model C Voting	**Committee Vote**	**His/path Grade**
GB135	0.001	ANGIO	-	-	-	-	ANGIO	**GB**
GB136	0.0168	ANGIO	-	-	-	-	ANGIO	**GB**
GB81	0.0092	ANGIO	-	-	-	-	ANGIO	**GB**
GB82	0.1688	ANGIO	-	-	-	-	ANGIO	**GB**
GB84	0.0379	ANGIO	-	-	-	-	ANGIO	**GB**
GB87	0.0001	ANGIO	-	-	-	-	ANGIO	**GB**
GB44	0.0524	ANGIO	-	-	-	-	ANGIO	**GB**
GB154	0.9771	DIFFER	0.0508	INTER	0.0671	INTER	INTER	**GB**
GB153	0.0092	ANGIO	-	-	-	-	ANGIO	**GB**
GB126	0.06	ANGIO	-	-	-	-	ANGIO	**GB**
GB35	0.043	ANGIO	-	-	-	-	ANGIO	**GB**
GB50	0.0135	ANGIO	-	-	-	-	ANGIO	**GB**
GB49	0.0025	ANGIO	-	-	-	-	ANGIO	**GB**
GB130	0.0029	ANGIO	-	-	-	-	ANGIO	**GB**
GB238	0.017	ANGIO	-	-	-	-	ANGIO	**GB**
GB103	0.0099	ANGIO	-	-	-	-	ANGIO	**GB**
GB3	0.001	ANGIO	-	-	-	-	ANGIO	**GB**
GB101	0.0014	ANGIO	-	-	-	-	ANGIO	**GB**
GB245	0.381	ANGIO	-	-	-	-	ANGIO	**GB**
A36	0.9729	DIFFER	0.8113	LOWER	0.7479	LOWER	LOWER	**A**
A30	0.9986	DIFFER	0.9153	LOWER	0.8828	LOWER	LOWER	**A**
AA105	1.0	DIFFER	0.0172	INTER	0.539	LOWER	INTER	**AA**
AA106	0.9987	DIFFER	0.7499	LOWER	0.361	INTER	INTER	**AA**
AA15	0.9999	DIFFER	0.3098	INTER	0.8289	LOWER	INTER	**AA**
AA14	0.9512	DIFFER	0.053	INTER	0.3594	INTER	INTER	**AA**
AA13	0.9754	DIFFER	0.1722	INTER	0.3444	INTER	INTER	**AA**

Table A. Results from the propagation of 26 test samples through each of our 3 trained model types A, B and C. Individual model and committee votes are shown for each sample. The committee vote was taken as the vote that the majority of the models concluded upon. Each type of model placed each test sample in one of the two tumour grades it had been trained to recognize. For example, type A models gave a score <0.5 if the sample in question had an expression profile similar to *ANGIO* (GB) and a score >0.5 if the sample's expression profile was closer to *DIFFER* (AA + A). Only *DIFFER* samples received a follow-up grading to differentiate the *INTER* (AA) and the *LOWER* (A) subtypes. Samples were assigned the *LOWER* subtype only if both *LOWER* trained models (B and C) graded them as *LOWER*, otherwise they were assigned to the INTER subtype. The highlighted row corresponds to the single tumour whose histopathological diagnosis and ANN grading did not concur.

Gene Selection

Informative genes were selected using the signal-to-noise (S2N) method described in [4]. Genes with S2N score near 1 or -1 were termed most informative.

$$S2N = \frac{(\mathrm{Avg}_1 - \mathrm{Avg}_2)}{(\sigma_1 + \sigma_2)}$$

Where: - Avg_1 and σ_1 is the mean and standard deviation along all samples of class 1 and similarly Avg_2 and σ_2 for samples of class 2.

For comparison, differentially expressed genes between the three astrocytic tumour grades were also detected in a pairwise fashion using empirical Bayesian analysis implemented in the LIMMA (Linear Models for Microarray Data) package. False discovery rate was controlled within LIMMA using the method of Benjamini & Hochberg [5]. Only probe sets with intensities over 50 units in all 65 samples, geometric mean tumour group expression changes larger than 2-fold and an associated Bayesian p-value less than 0.001 were considered differentially expressed in any of the three pairwise tumour group comparisons made (GB-AA, GB-A and A-AA).

Mathews Correlation

Mathews correlation measure allows for a more significant measure of the Leave-one-out cross-validation performance as it takes into account the sensitivity as well as the specificity of the results. Sensitivity is a measure of the how many samples are classified as positive and are really positive (*true positives – tp*) versus the number of samples that are positive but classified as negatives (*false negatives - fn*). While specificity is a measure of the samples that are classified as negative and are really negative (*true negatives – tp*) versus the number of samples that are negative but classified as positive (*false positives - fn*). In medical assays these definitions often have a more meaningful interpretation: Sensitivity refers to the proportion of people with disease who have a positive test result. Specificity refers to the proportion of people without disease who have a negative test result. However in our case specificity and sensitivity are arbitrarily assigned, where, sensitivity refers to class 1 and specificity refers to class 0; in order to reveal a more meaningful interpretation of our results. Mathews correlation combines both terms into a single measure.

$$Sensitivity = \frac{tp}{tp + fn}$$

$$Specificity = \frac{tn}{tn + fp}$$

$$Mathews\ Correlation = \frac{\sqrt{(tp+fn)(tp+fp)(tn+fp)(tn+fn)}}{(tptn - fpfn)}$$

Error Margin (EM) Score

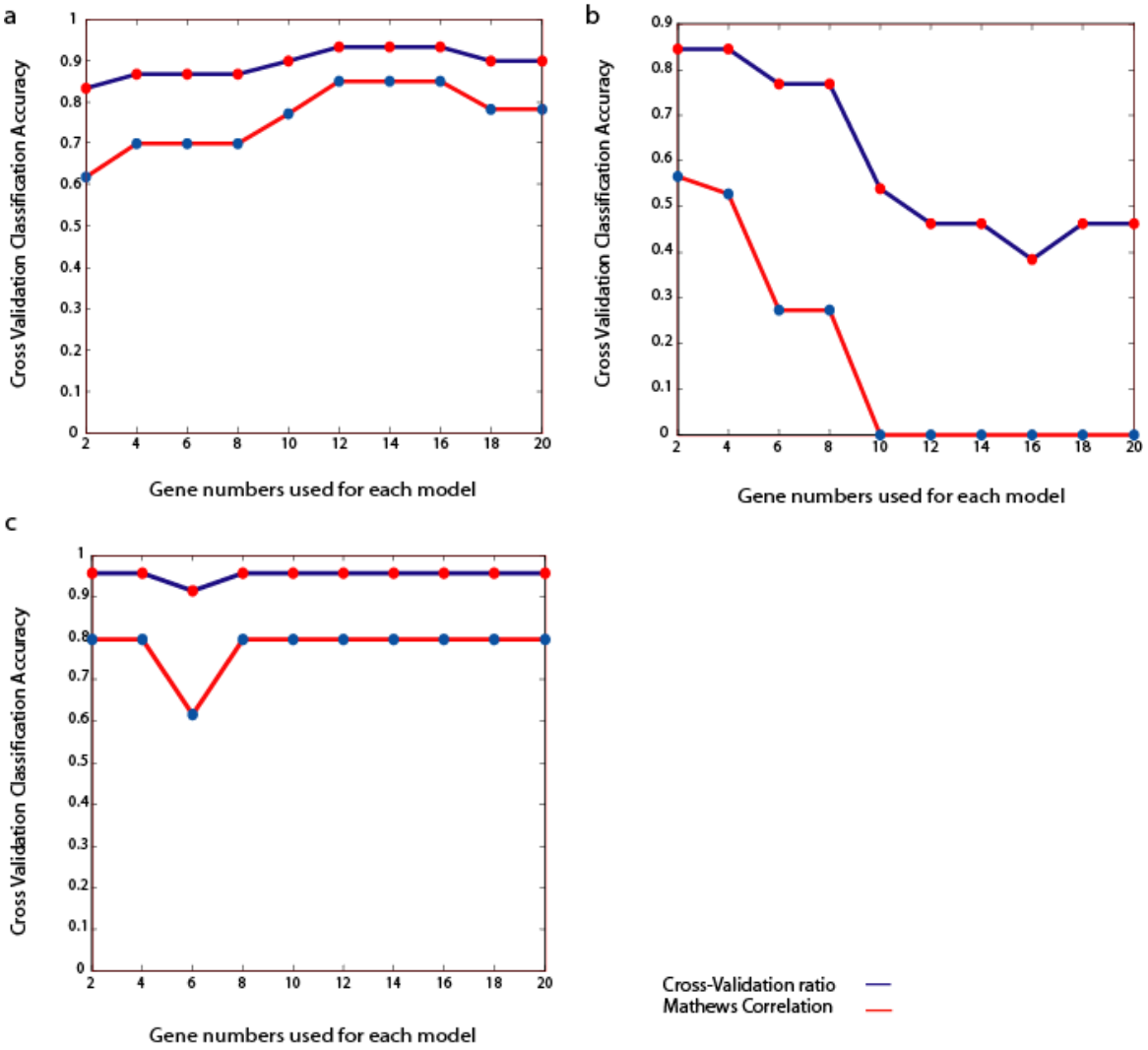
In cases where leave-one-out cross-validation models generate the same percentage accuracy and no distinction can be made by Mathews Correlation. We used the error margin score to decide on which model to finally use for testing. The error margin score simply calculates the difference of the network outputs from the decision boundary (0.5) for all the misclassified ***validation*** samples (*m)* and sums them all up. Hence for conflicting cases the Leave-one-out cross-validation model with the lowest error margin score will be selected for testing.

$$EM = \sum_{i=1}^{n} |m_i - 0.5| \qquad \text{where } n = fp+fn$$

Leave-one-out cross-validation plots for astrocytic tumor grading

For the *GB vs. AA* task Leave-one-out cross-validation reached its optimum success

rate when gene sets of 12, 14 and 16 were selected during every leave-one-out run. A Leave-one-out cross-validation score of 93.33% was achieved, corresponding to 2 mis-graded[b] samples (AA76 and GB133) out of 30 (Fig. C, a). In this binary grading task percentage accuracy and Mathews Correlation was the same for all three gene sets, hence we selected the Leave-one-out cross-validation model to use for testing based on the network outputs (see *Error Margin Score*). Gene sets of 16 had the smallest error margin score (data not shown). For the *AA vs. A* task (Fig. C, b), the top Leave-one-out cross-validation accuracy was 84.62%, corresponding to 2 mis-graded samples out of 13. This accuracy was achieved when gene sets of 2 and 4 with the



highest signal-to-noise score were selected during every leave-one-out run. However looking at Mathews Correlation it was evident that the accuracy was more significant for gene sets of 2 (two mis-graded, one from each grade, A26 and AA92, versus two mis-graded, both from grade "A", as was the case for the 4 gene sets). For the *GB vs. A* task (Fig. C, c) the top Leave-one-out cross-validation score was 95.65, corresponding to 1 mis-graded training sample (A9) out of 23. This accuracy was achieved when gene sets of 2, 4, 8, 10, 12, 14, 16, 18 and 20 with the highest signal-to-noise score were selected during every leave-one-out run. As in the *GB vs. AA* task there was no significant difference depicted by Mathews Correlation, however the

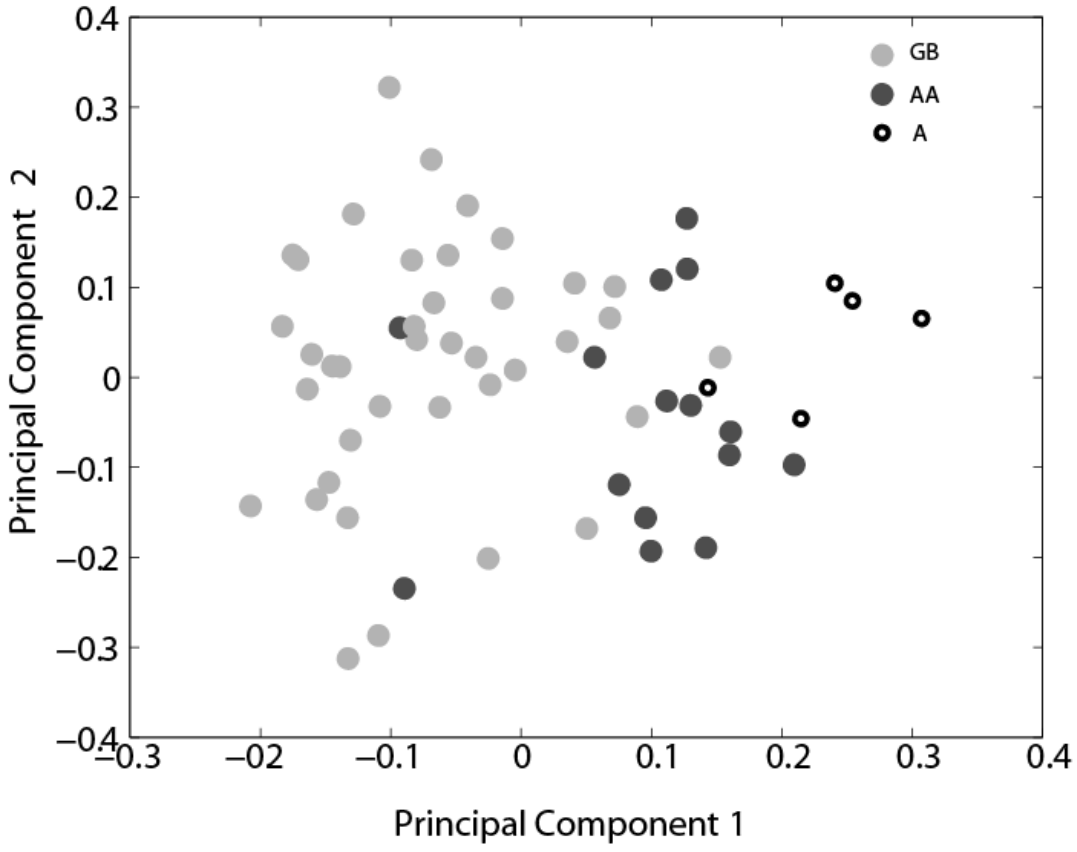
[b] The term 'mis-graded' is used to characterize samples where there is disagreement between network output and initial histopathological diagnosis.

model with the 2 gene sets had the lowest error margin score (data not shown); hence it was selected for testing.

Figure B. Leave-one-out cross-validation plots for three different types of models trained. a (GB vs. AA) - The plot displays the Leave-one-out cross-validation performance for group 'a' training set. As is evident Leave-one-out cross-validation reaches its optimum success rate when gene sets of 12, 14 and 16 with the highest signal-to-noise score were selected during every leave-one-out run. Error margin score concluded on the 16 gene set model to be used for testing. **b (AA vs. A)** - the plot displays the Leave-one-out cross-validation performance for group 'b' training set. Both the Leave-one-out cross-validation ratio and Mathews correlation reveal that Leave-one-out cross-validation reached its optimum success rate when gene sets of 2 with the highest signal-to-noise score were selected during every leave-one-out run. **c (GB vs. A)** - Leave-one-out cross-validation performance for group 'c' training set. Optimum success rate was achieved when gene sets of 2, 4, 8, 10, 12, 14, 16, 18, 20 with the highest signal-to-noise score were selected during every leave-one-out run. The 2 gene model was selected for testing because it had the lowest error margin score.

PCA analysis

a



b

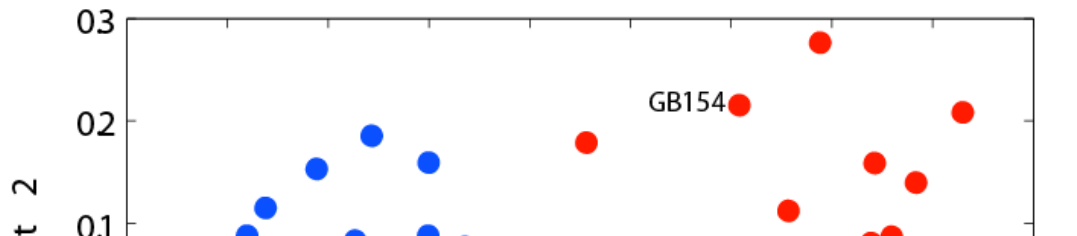


Figure C. Principal component analysis before and after gene selection. Analysis of the first two principal components for all 33 training and 26 test samples: **(a)** using expression values for 22,382 probe sets. Tissue samples are colour-coded (grey, GB; black, AA; white, A), **(b)** using the 59 classifier genes selected during training. Colour coding is according to the 3 molecular subtypes which best characterize the samples (blue: ANGIO, red: INTER, orange: LOWER). Results reflect the ANN outputs with only one GB sample (red: GB154 – labeled with text) seen to group together with the rest of the INTER samples (red).

Hierarchical clustering

Clustering Network Outputs

Training via leave-one-out cross validation resulted in a total of 6600 trained network models (3000: GBvsAA; 2300: GBvsA; 1300: AAvsA). By propagating all 59 training and test tumour samples through these ANN models, 6600 different outputs are obtained for every sample. This results in a matrix 6600-by-59 that was subsequently used in our clustering efforts in order to improve our ANN visualization and attain a more informative representation of results. Furthermore; in this form, results can be easily compared to clustering of genes expression values.

Malignancy grading of additional samples difficult to grade histopathologically

Test tumour ID	GB vs AA Model A	Model A Voting	AA vs A Model B	Model B Voting	GB vs A Model C	Model C Voting	**Committee Vote**	**Hist/path Class**
PA68	0.0428	ANGIO	0.0834	-	0.0056	-	ANGIO	**PA**
PA67	0.7619	DIFFER	0.723	LOWER	0.0067	INTER	INTER	**PA**
AA29	0.0025	ANGIO	0.134	-	0.1916	-	ANGIO	**AA**
AA86	0.0621	ANGIO	0.1451	-	0.0086	-	ANGIO	**AA**
AA93	0.0041	ANGIO	0.0052	-	0.0072	-	ANGIO	**AA**
AA49	0.0253	ANGIO	0.2941	-	0.0384	-	ANGIO	**AA**

Table B. ANN grading of the 6 additional samples that proved to be difficult to grade by histopathological analysis. These include the two Piloytic Astrocytoma tumour samples (PA68 and PA67) as well as the 4 difficult Anaplastic Astrocytoma samples (AA29, AA86, AA93 and AA49). Sample PA67 was assigned to a different tumour grade by each of the models, but since the two *LOWER* trained model did not agree in their grading it was assigned to the *INTER* subtype.

Grading of independent dataset using across-array gene classifiers

Test tumour Index	ANGIO/ DIFFER Model 1	Model 1 Voting	INTER / LOWER Model 2	Model 2 Voting	**Overall Vote**	**His/path Grade**
1	0.4746	ANGIO	-	-	ANGIO	**GB**
2	0.998	DIFFER	0.2283	INTER	INTER	**AA**
3	0.4486	ANGIO	-	-	ANGIO	**GB**
4	0.9967	DIFFER	0.045	INTER	INTER	**GB**
5	0.7472	DIFFER	0.037	INTER	INTER	**GB**
6	0.0209	ANGIO	-	-	ANGIO	**GB**
7	0.9961	DIFFER	0.5934	LOWER	LOWER	**A**
8	0.0113	ANGIO	-	-	ANGIO	**GB**
9	0.0165	ANGIO	-	-	ANGIO	**GB**
10	0.0106	ANGIO	-	-	ANGIO	**GB**
11	0.0847	ANGIO	-	-	ANGIO	**GB**
12	0.0133	ANGIO	-	-	ANGIO	**GB**
13	0.0229	ANGIO	-	-	ANGIO	**GB**
14	0.1739	ANGIO	-	-	ANGIO	**GB**
15	0.0249	ANGIO	-	-	ANGIO	**GB**
16	0.0044	ANGIO	-	-	ANGIO	**GB**
17	0.0207	ANGIO	-	-	ANGIO	**GB**
18	0.9838	DIFFER	0.8986	LOWER	LOWER	**A**
19	0.9936	DIFFER	0.0556	INTER	INTER	**AA**
20	0.0573	ANGIO	-	-	ANGIO	**GB**
21	0.9999	DIFFER	0.4409	INTER	INTER	**AA**
22	0.0655	ANGIO	-	-	ANGIO	**GB**
23	0.1514	ANGIO	-	-	ANGIO	**GB**

Table C. Independent data set analysis of 23 test samples from the Shai et al, 2003[6] dataset using model types 1 and 2 (see text). Individual model and overall voting is shown for each sample. Type 1 models gave a score <0.5 if the sample in question had an expression profile similar to *ANGIO* (GB) and a score >0.5 if the sample's expression profile was closer to *DIFFER* (lower grades II and III). Similarly; type 2 models score tumours resembling *INTER* (AA) with a value <0.5 and tumours resembling *LOWER* (A) with a value >0.5. Highlighted rows correspond to test samples with overall votes that did not agree with original histopathological diagnosis. The gene names of the cross–chip gene classifiers are TYPE1 Genes: LDHA, PEA15, LGALS1, TIMP1, PLAT, ZMYND11, EMP3, NAP1L3, USH1C, DAG1, PDGFA, LGALS3, HNRPH3, CRYAB, EFEMP2, KIAA1279, RPL22, PDPN, TYPE2 genes: SCP2, B2M

Survival Analysis using independent datasets and across-array gene classifiers

ANGIO/ DIFFER	INTER / LOWER	INTER / LOWER	Survival Days	Freije subtyping	ANN subtypeing	Histo/Path
0	-	-	1089	SC2	ANGIO	GBM 1043
0	-	-	420	SC2	ANGIO	GBM 1354
0	-	-	293	SC2	ANGIO	GBM 1398
0	-	-	153	SC2	ANGIO	GBM 1469
0	-	-	1031	SC2	ANGIO	GBM 1516
0	-	-	683	SC2	ANGIO	GBM 1675
0	-	-	723	SC2	ANGIO	GBM 1798
0	-	-	325	SC2	ANGIO	GBM 1902
0	-	-	396	SC2	ANGIO	GBM 2013
0	-	-	396	SC2	ANGIO	GBM 2015
0	-	-	298	SC2	ANGIO	GBM 2079
0	-	-	203	SC2	ANGIO	GBM 2098
0	-	-	7	SC2	ANGIO	GBM 597
0	-	-	185	SC2	ANGIO	GBM 604
0	-	-	112	SC2	ANGIO	GBM 660
0	-	-	356	SC2	ANGIO	GBM 697
0	-	-	188	SC2	ANGIO	GBM 712
0	-	-	53	SC2	ANGIO	GBM 749
0	-	-	182	SC2	ANGIO	GBM 931
0	-	-	418	SC2	ANGIO	GBM 976
0	-	-	443	SC2	ANGIO	MIXED III 799
0.01	-	-	71	SC2	ANGIO	GBM 1495
0.01	-	-	588	SC2	ANGIO	GBM 1667
0.01	-	-	279	SC2	ANGIO	GBM 1900
0.01	-	-	237	SC2	ANGIO	GBM 2017
0.02	-	-	95	SC2	ANGIO	GBM 2158
0.05	-	-	389	SC2	ANGIO	GBM 1905
0.05	-	-	302	SC2	ANGIO	GBM 585
0.05	-	-	412	SC2	ANGIO	GBM 636
0.06	-	-	64	SC2	ANGIO	GBM 1414
0.06	-	-	186	SC2	ANGIO	GBM 824
0.08	-	-	224	SC2	ANGIO	GBM 1342
0.24	-	-	90	SC2	ANGIO	GBM 1032
0.25	-	-	126	SC2	ANGIO	GBM 1022
0.25	-	-	56	SC2	ANGIO	GBM 995

0.33	-	-	927	SC1	ANGIO	GBM 1681
0.4	-	-	96	SC1	ANGIO	GBM 1423
0.43	-	-	506	SC1	ANGIO	GBM 706
0.49	-	-	43	SC2	ANGIO	GBM 746
0.54	0.53	0.14	98	SC2	INTER	GBM 932
0.7	0.41	0.25	877	SC2	INTER	ASTRO III 1704
0.71	0.67	0.27	140	SC2	INTER	GBM 782
0.74	0.76	0.2	961	SC1	INTER	GBM 1656
0.82	0.95	0.06	286	SC1	INTER	GBM 839
0.96	0.81	0.11	569	SC1	INTER	GBM 938
0.98	0.36	0.32	1098	SC1	INTER	GBM 2166
0.99	0.78	0.36	1114	SC1	INTER	ASTRO III 1425
0.99	0.47	0.1	85	SC2	INTER	GBM 1511
0.99	0.6	0.3	664	SC1	INTER	MIXED III 886
1	0.36	0.34	858	SC1	INTER	ASTRO III 1723
1	0.52	0.36	2516	SC1	INTER	ASTRO III 587
1	0.42	0.4	2125	SC1	INTER	ASTRO III 671
1	0.23	0.22	1557	SC1	INTER	ASTRO III 672
1	0.84	0.35	1830	SC1	INTER	ASTRO III 747
1	0.49	0.09	1247	SC1	INTER	GBM 1038
1	0.48	0.64	1088	SC1	INTER	GBM 1478
1	0.54	0.36	1022	SC1	INTER	GBM 1521
1	0.29	0.27	780	SC2	INTER	GBM 1745
1	0.93	0.39	223	SC2	INTER	GBM 2028
1	0.8	0.38	861	SC1	INTER	MIXED III 1721
1	0.6	0.1	2185	SC1	INTER	MIXED III 664
0.99	0.86	0.53	442	SC1	LOWER	ASTRO III 659
1	0.71	0.61	850	SC1	LOWER	MIXED III 615
1	0.7	0.82	1918	SC1	LOWER	MIXED III 713
1	0.57	0.7	1474	SC1	LOWER	MIXED III 912

Table D. Independent data set analysis of 65 samples from the Freije et al, 2004 [7] dataset using ANN models trained with our 59 gene classifiers.(see text). The initial grading of the models classified the 65 tumour samples into two groups (*ANGIO* and *DIFFER*), which, resembled the "Survival Clusters" (SC1 and SC2) obtained in the Freije study with 86.15% similarity.

ANGIO/ DIFFER	**INTER/ LOWER**	**INTER/ LOWER**	**SURVIVAL WEEKS**	**PHILLIPS SUBTYPEING**	**ANN SUBTYPEING**	**Histo/Path**
0.14	-	-		*Mes*	ANGIO	IV
0.15	-	-	51	*Prolif*	ANGIO	IV with necrosis
0.15	-	-	65	*Mes*	ANGIO	IV with necrosis
0.16	-	-	59	*Mes*	ANGIO	IV with necrosis
0.16	-	-	70	*Prolif*	ANGIO	IV with necrosis
0.17	-	-		*Mes*	ANGIO	IV
0.19	-	-		*Prolif*	ANGIO	IV
0.19	-	-		*Mes*	ANGIO	IV
0.19	-	-		*Prolif*	ANGIO	IV
0.2	-	-	32	*Prolif*	ANGIO	IV with

						necrosis
0.2	-	-	12	*Prolif*	ANGIO	IV with necrosis
0.21	-	-	236	*Mes*	ANGIO	IV with necrosis
0.21	-	-	16	*Prolif*	ANGIO	IV with necrosis
0.21	-	-	154	*Mes*	ANGIO	IV with necrosis
0.21	-	-	47	*Mes*	ANGIO	III
0.22	-	-	181	*Mes*	ANGIO	IV with necrosis
0.22	-	-	55	*Prolif*	ANGIO	IV with necrosis
0.22	-	-	33	*Mes*	ANGIO	IV with necrosis
0.23	-	-	77	*Mes*	ANGIO	IV with necrosis
0.23	-	-	3	*Mes*	ANGIO	IV with necrosis
0.24	-	-	95	*Prolif*	ANGIO	IV with necrosis
0.24	-	-		*Mes*	ANGIO	IV
0.25	-	-	313	*Mes*	ANGIO	IV with necrosis
0.25	-	-		*Mes*	ANGIO	IV
0.26	-	-	131	*Mes*	ANGIO	IV with necrosis
0.26	-	-	56	*Mes*	ANGIO	IV with necrosis
0.27	-	-	62	*Prolif*	ANGIO	IV with necrosis
0.27	-	-	210	*Prolif*	ANGIO	IV with necrosis
0.27	-	-	41	*Prolif*	ANGIO	IV with necrosis
0.28	-	-	311	*Mes*	ANGIO	IV with necrosis
0.28	-	-		*Mes*	ANGIO	IV
0.28	-	-		*Mes*	ANGIO	IV
0.29	-	-	79	*Prolif*	ANGIO	IV with necrosis
0.29	-	-		*Mes*	ANGIO	III
0.3	-	-	32	*Prolif*	ANGIO	IV with necrosis
0.3	-	-	62	*Mes*	ANGIO	IV with necrosis
0.31	-	-	238	*Prolif*	ANGIO	IV with necrosis
0.31	-	-	91	*Prolif*	ANGIO	IV with necrosis
0.32	-	-		*Prolif*	ANGIO	IV
0.33	-	-	59	*Prolif*	ANGIO	IV with necrosis
0.34	-	-	57	*Prolif*	ANGIO	IV with necrosis
0.34	-	-	53	*Mes*	ANGIO	IV without necrosis
0.35	-	-	131	*Mes*	ANGIO	IV with

						necrosis
0.35	-	-	106	*Mes*	ANGIO	IV with necrosis
0.35	-	-	53	*Mes*	ANGIO	IV without necrosis
0.36	-	-		*Mes*	ANGIO	IV
0.37	-	-	70	*Prolif*	ANGIO	IV with necrosis
0.4	-	-	53	*Mes*	ANGIO	IV with necrosis
0.42	-	-	17	*PN*	ANGIO	III
0.43	-	-		*Mes*	ANGIO	IV
0.44	-	-	34	*PN*	ANGIO	IV with necrosis
0.47	-	-		*Mes*	ANGIO	IV
0.48	-	-	62	*PN*	ANGIO	IV with necrosis
0.49	-	-	46	*PN*	ANGIO	III
0.51	0.58	0.24	125	*Mes*	INTER	IV with necrosis
0.52	0.11	0.86	111	*Prolif*	INTER	IV with necrosis
0.54	0.09	0.1	52	*Prolif*	INTER	IV with necrosis
0.56	0.3	0.56	467	*PN*	INTER	III
0.56	0.26	0.69	460	*PN*	INTER	III
0.58	0.28	0.31	39	*Prolif*	INTER	IV with necrosis
0.6	0.18	0.39	32	*Mes*	INTER	IV with necrosis
0.61	0.16	0.88	57	*Mes*	INTER	IV with necrosis
0.62	0.19	0.24	300	*PN*	INTER	III
0.64	0.17	0.13	97	*Mes*	INTER	IV with necrosis
0.65	0.14	0.8		*Prolif*	INTER	IV
0.65	0.13	0.35	242	*Prolif*	INTER	IV with necrosis
0.66	0.08	0.16	277	*PN*	INTER	IV without necrosis
0.66	0.28	0.15	145	*PN*	INTER	IV without necrosis
0.67	0.25	0.32	174	*PN*	INTER	III
0.7	0.14	0.8	33	*Prolif*	INTER	IV with necrosis
0.73	0.58	0.54		*PN*	LOWER	IV
0.73	0.07	0.54		*Mes*	INTER	III
0.76	0.34	0.28	86	*PN*	INTER	IV
0.76	0.77	0.99	383	*PN*	LOWER	IV without necrosis
0.79	0.36	0.54		*PN*	INTER	III
0.79	0.4	0.29	73	*PN*	INTER	III
0.8	0.16	0.2		*PN*	INTER	IV
0.8	0.18	0.31	108	*PN*	INTER	IV
0.82	0.31	0.63		*PN*	INTER	III
0.87	0.91	0.9		*PN*	LOWER	IV
0.88	0.37	0.5		*PN*	INTER	IV with necrosis

0.89	0.52	0.37	175	*PN*	INTER	IV with necrosis
0.89	0.9	0.85	146	*PN*	LOWER	III
0.9	0.65	0.85	477	*PN*	LOWER	III
0.92	0.36	0.18		*PN*	INTER	III
0.93	0.27	0.66	97	*Prolif*	INTER	IV with necrosis
0.95	0.65	0.97	150	*PN*	LOWER	III
0.96	0.16	0.25	123	*Prolif*	INTER	IV without necrosis
0.96	0.43	0.21	81	*PN*	INTER	IV
0.97	0.91	0.62	203	*PN*	LOWER	III
0.97	0.52	0.19	316	*PN*	INTER	III
0.97	0.65	0.93	445	*PN*	LOWER	IV with necrosis
0.97	0.55	0.12	41	*PN*	INTER	III
0.98	0.44	0.59	210	*PN*	INTER	IV with necrosis
0.98	0.72	0.8	325	*PN*	LOWER	IV with necrosis
0.99	0.97	0.64	357	*PN*	LOWER	III
0.99	0.55	0.17	102	*PN*	INTER	III
0.99	0.7	0.57	115	*PN*	LOWER	III
0.99	0.76	0.71	244	*PN*	LOWER	III
0.99	0.39	0.21	322	*PN*	INTER	III

Table E. Analysis of Phillips et 2006 samples

	PA tumours graded as *ANGIO* and *INTER*		**AA tumours difficult to diagnose histologically and graded as ANGIO**				**GB tumour graded as *INTER***
Tumour ID	PA68	PA67	AA29	AA49	AA86	AA93	GB154
Age at Operation (yrs) /gender	3/male	26/Female	60/Female	70/female	45/male	57/female	31/male
Survival following operation (Days)	3586 Alive at end of follow-up	5074 Alive at end of follow-up	704	unknown	717	535	710
Clinical and histopathological information	Primary tumour; PA by radiology, histology, Cerebellar tumour	Primary tumour; PA clinically, cystic, histology, Cerebellar tumour	Primary tumour; GB suspected but criteria not fulfilled (suspicion of necrosis)	Treated (irradiation and chemotherapy) recurrent tumour. Necrosis present – due to treatment?	Treated (irradiation and chemotherapy) recurrent tumour. Necrosis present – due to treatment?	Primary tumour; Diagnosis AA.	Primary tumour; Histologically GB. Microvascular proliferation and necrosis present.
Gene data							
EGFR	+/+	+/+/+[1]	+/+/+	+/amp	+/+	+/amp	+/+
CDKN2A	+/+	+/+/+	+/-	- /-	- /N71K[3]	- /-	+/+
CDKN2B	+/+	+/+/+	+/-	- /-	+/-	- /-	+/+
p14ARF	+/+	+/+/+	+/-	- /-	- /L86V[3]	- /-	+/+
PTEN	+/+	+/+/+	+/X404S[2]	- / W274G[3]	+/-	+/-	+/+
CDK4	+/+	+/+/+	+/+	+/+	+/+	+/+	+/amp
MDM2	+/+	+/+/+	+/+	+/+	+/+	+/+	+/+
RB1	+/+	+/+/+	+/+	+/+	+/-	+/+	+/-
TP53	+/+	+/+	+/+	+/+	- /K320del[4] /K320del[4]	+/+	+/E171del[5]

+ = one wild type allele. - = loss of one allele. amp = amplification of an allele. [1] 3 copies of gene (trisomy of chromosome) [2]Stop codon mutated leading to an additional 8 amino acids. [3] Amino acid substitution. [4] Deletion of 16bp with frame shift. [5] Deletion of one base with frame shift.

Table I. All available annotation for the single tumour for which ANN grading differed from the original histopathological diagnosis (GB154), the 4 AA tumours known to be histopathologically difficult to grade and 2 PA tumours. Note: a) The excellent survival of patients PA68 and PA67 which were graded as *ANGIO* and *INTER* by our ANN. According to current WHO criteria, these tumours were found compatible with a grade I PA diagnosis (see text). b) Genotype analysis for 9 genes known to be involved in diffuse adult astrocytoma tumourigenesis including *CDKN2A/CDKN2B/p14*^{ARF}, *TP53*, *RB1*, *PTEN*, *EGFR*, *MDM2*and *CDK4*.

The jar file for the ANN can be downloaded at:
http://www.imbb.forth.gr/people/poirazi/software.html

References

1. Linder, R., et al., *The "Subsequent Artificial Neural Network" (SANN) Approach Might Bring More Classification Power To ANN-based Microarray Analysis.* BMC Bioinformatics, 2004. **20**(18): p. 3544-3552.
2. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* PNAS, 2001. **98**(26): p. 15149-15154.
3. Xu, Y., et al., *Artificial Neural networks and Gene Filtering Distinguishes Between Global Gene Expression Profiles of Barrett's Esophagus and Esophageal Cancer.* Cancer Research, 2002. **62**: p. 3493-3497.
4. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science, 1999. **286**(15): p. 531-537.
5. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Stat Med, 1990. **9**(7): p. 811-8.
6. Shai, R., et al., *Gene expression profiling identifies molecular subtypes of gliomas.* Oncogene, 2003. **22**(31): p. 4918-23.
7. Freije, W.A., et al., *Gene Expression Profiling of Gliomas Strongly Predicts Survival.* Cancer Res, 2004. **64**(18): p. 6503-6510.