

**SUPPLEMENTARY INFORMATION TO  
MATHEMATICAL MODELING OF NOISE AND DISCOVERY OF GENETIC  
EXPRESSION CLASSES IN GLIOMAS**

HM Fathallah-Shaykh, M Rigen, L-J Zhao, K Bansal, B He, HH Engelhard, L Cerullo,  
K Von Roenn, R Byrne, L Munoz, GL Rosseau, R Glick, T Lichtor, and E DiSavino

**ONCOGENE (2002) 21:7164-7174**

## MATHEMATICAL ANALYSIS

*Symbols And Notations:* Bolded letters represent matrices and vectors. \* and • denote multiplication and inner product, respectively.  $\langle \mathbf{b}_1, \dots, \mathbf{b}_n \rangle$  refers to the space defined by a basis  $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ .

### *Replicate Ratios*

The experiments of the 1.9K microarrays are designed to yield 4 replicate spots with probe switching. The experiments of the 1.7K microarrays (see Figure 1 of the paper) generate 2-12 replicate spots with probe switching. The data of the  $n$  replicate spots,  $2 \leq n \leq 12$ , are ‘prepared’ as described in the Methods section of the paper and the column vectors of the output are assembled to constitute the columns of the matrices  $\mathbf{Q}_j(19,200 \times n)$  corresponding to tumors  $j$ ,  $1 \leq j \leq 35$ ,  $2 \leq n \leq 12$ . The gene row vectors of  $\mathbf{Q}_j$  contain the  $\log_2(\text{expression ratios})$  of the  $n$  replicate spots;  $j$  refers to the tumor sample;  $n$  to the number of replicate ratios.

### *Filtering functions*

Figure 1 shows the per cent false positive and per cent false negative after applications of the filtering functions  $f_2, \dots, f_4, \dots, f_{12}$  to spots  $(a_{i1}, a_{i2})$ ,  $\dots$ ,  $(a_{i1}, a_{i2}, a_{i3}, a_{i4})$ ,  $\dots$ , and  $(a_{i1}, a_{i2}, \dots, a_{i12})$ , respectively.

**Definition.** Let  $ST_j$  be the set of row vectors  $(a_{i_1}, a_{i_2}, \dots, a_{i_n})$  in  $Q_j$ ,  $1 \leq i \leq 19,200$ ,  $1 \leq j \leq 35$ ,  $2 \leq n \leq 12$ ;  $a, i, j, n$ , refer to  $\log_2(\text{expression ratio})$ , genes, tumors, and number of replicate ratios, respectively. Let  $A_2$  be  $[-5.6, -0.48] \cup \{0\} \cup [0.48, 5.6]$ . The functions  $f_n : ST_j \rightarrow A_2$ ,  $2 \leq n \leq 12$ , are defined as:  $f_n((a_{i_1}, a_{i_2}, \dots, a_{i_n})) = \text{mean}((a_{i_1}, a_{i_2}, \dots, a_{i_n}))$  only if all the following 3 statements are true: 1) the elements of  $\{a_{i_1}, \dots, a_{i_n}\}$  are either (a) *all* positive and different than 0, or (b) *all* negative and different than 0, 2) all elements of  $\{a_{i_1}, \dots, a_{i_n}\}$  are located outside the interval  $[-0.48, 0.48]$  ( $\log_2(1.4) = 0.48$ ), and 3) a minimum of 75% of the spots corresponding to  $\{a_{i_1}, \dots, a_{i_n}\}$  are not flagged manually. Otherwise,  $f_n((a_{i_1}, a_{i_2}, \dots, a_{i_n})) = 0$

### *Mathematical Modeling*

Let  $E_{11}$ ,  $E_{12}$ ,  $E_{13}$ ,  $E_{14}$  of dimension (19,200X35) contain the unfiltered  $\log_2(\text{expression ratios})$  of the 4 unfiltered replicate spots (illustrated in Figure 2 of the paper); their rows and columns refer to genes and tumors, respectively. Let the matrix  $E$  (19,200X35) contain the data after application of  $f_4$  to the 4 replicate  $\log_2(\text{expression ratios})$  of  $E_{11}$ ,  $E_{12}$ ,  $E_{13}$ ,  $E_{14}$  (see cartoon in Figure 2). To model the noise, we assemble the matrices  $N_r$ ,  $1 \leq r \leq 4$ . Let  $S_0$  be the set of genes whose row-vectors in  $E$  are  $= 0$ ;  $S_0$  contains 9,155 genes. The matrices  $N_r$  (9,155X35) assemble the vectors

corresponding to the elements of  $S_0$  in  $\mathbf{E}_{1r}$ ,  $1 \leq r \leq 4$ , respectively. The 4  $\mathbf{N}_r$  matrices are linearly transformed by singular value decomposition:

$$\mathbf{N}_r (9,155 \times 35) = \mathbf{U}_r (9,155 \times 35) * \mathbf{S}_r (35 \times 35) * \mathbf{V}_r^T (35 \times 35) \quad (1.1)$$

The columns of  $\mathbf{V}_r$  represent the eigengene vectors;  $\mathbf{S}$  is a diagonal matrix containing the eigenvalues that reflect the “eigenexpression” levels or the amount of information carried by the corresponding eigengenes. The matrices  $\mathbf{V}_r$  are orthogonal:

$$\mathbf{V}_r * \mathbf{V}_r^T = \mathbf{I} \quad \text{and} \quad \mathbf{V}_r^T * \mathbf{V}_r = \mathbf{I} \quad (1.2)$$

$\mathbf{I}$  (35X35) is the identity matrix. Let  $\{\mathbf{n}_{r1}, \dots, \mathbf{n}_{r35}\}$  be the sets containing the column vectors of  $\mathbf{V}_r$ ,  $1 \leq r \leq 4$ . Equation (1.2) implies that  $\{\mathbf{n}_{r1}, \dots, \mathbf{n}_{r35}\}$  are orthonormal bases (eigenbases) of 35-dimensional spaces defined by  $\mathbf{V}_r$ . Let  $\mathbf{v}$  and  $\mathbf{w}$  be 2 vectors in space with angle  $\theta$  between them, their inner product:

$$\mathbf{v} \bullet \mathbf{w} = \|\mathbf{v}\| * \|\mathbf{w}\| * \cos \theta \quad (1.3)$$

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \bullet \mathbf{v}} \quad (1.4)$$

$\|\mathbf{v}\|$  is the norm or length of  $\mathbf{v}$ .  $\|\mathbf{w}\| \cdot \cos \theta$  is the projection or coordinate ( $m$ ) of  $\mathbf{w}$  onto  $\mathbf{v}$ . Therefore,

$$m = \frac{(\mathbf{v} \cdot \mathbf{w})}{\sqrt{\mathbf{v} \cdot \mathbf{v}}} \quad (1.5)$$

The row vectors of  $\mathbf{N}_r$  are projected (eigenprojections) onto the 3-dimensional subspaces  $\langle \mathbf{n}_{r1}, \mathbf{n}_{r2}, \mathbf{n}_{r3} \rangle$  of  $\mathbf{V}_r$ . Because  $\|\mathbf{n}_{rk}\| = 1$ ,  $1 \leq k \leq 35$ , the element at position  $(i, k)$  in the matrix **PROJ**:

$$\mathbf{PROJ} = \mathbf{N}_r (9,155 \times 35) * \mathbf{V}_r (35 \times 35)$$

represents the coordinate of the projection of the  $i$ th gene row vector of  $\mathbf{N}_r$  onto the  $k$ th eigengene of  $\{\mathbf{n}_{r1}, \dots, \mathbf{n}_{rk}, \dots, \mathbf{n}_{r35}\}$ . The eigendistance from the origin of a vector

$\mathbf{v}(m_1, \dots, m_n)$  in the  $n$ th dimensional space =  $\sqrt{\sum_{i=1}^n m_i^2}$ . This eigendistance may also be

computed as the norm of the expression vectors (row vectors) of the  $\mathbf{N}_r$  matrices.

### *Molecular Classification And Coefficients of Variance*

Let  $\mathbf{T}$  be the matrix containing the expression of the 108 genes (columns) that are resistant to the application of both filters (see Figure 5 of the paper). The matrix  $\mathbf{T}(35 \times 108)$  (its rows correspond to tumors) is transformed by singular value

decomposition to yield the orthogonal matrix  $\mathbf{V}_T(108 \times 108)$  of dimension = 108, defined by its eigenbasis  $\{\mathbf{t}_1, \dots, \mathbf{t}_{108}\}$ :

$$\mathbf{T}(35 \times 108) = \mathbf{U}_T(35 \times 108) * \mathbf{S}_T(108 \times 108) * \mathbf{V}_T^T(108 \times 108)$$

Therefore,

$$\mathbf{T} * \mathbf{V}_T = \mathbf{U}_T * \mathbf{S}_T \quad (1.6)$$

The tumor row vectors of  $\mathbf{T}$  are projected onto  $\langle \mathbf{t}_1, \dots, \mathbf{t}_{108} \rangle$ . Because the 36<sup>th</sup> to 108<sup>th</sup> eigenvalues of  $\mathbf{S}_T$  are equal to 0, the first 35 eigenvectors of  $\{\mathbf{t}_1, \dots, \mathbf{t}_{108}\}$  ‘carry 100% of the information.’ In addition, equation (1.6) implies that the distance between the projections of any 2 row-vectors of  $\mathbf{T}$  onto  $\langle \mathbf{t}_1, \dots, \mathbf{t}_{108} \rangle$  is equal to the distance that separates their projections in  $\langle \mathbf{t}_1, \dots, \mathbf{t}_{35} \rangle$ . The distance between 2 vectors  $\mathbf{v}(v_1, \dots, v_{108})$  and  $\mathbf{w}(w_1, \dots, w_{108})$  is:

$$\|\mathbf{v} - \mathbf{w}\| = \sqrt{(\mathbf{v} - \mathbf{w}) \bullet (\mathbf{v} - \mathbf{w})} = \sqrt{\sum_{i=1}^{108} (v_i - w_i)^2} \quad (1.7)$$

### *Analysis of Variance*

To study the variance among the expression values of the 4 replicate spots, we assemble the matrix  $\mathbf{X}(92 \times 35 \times 4)$ , a 3-dimensional expression array, containing the

expression of the genes of **C** (shown in Figure 5a of the paper) in **E<sub>11</sub>**, **E<sub>12</sub>**, **E<sub>13</sub>**, and **E<sub>14</sub>**. Let  $fv$  be the coefficient of variance:

$$fv = \left| \frac{s}{\bar{x}} \right| \quad (1.8)$$

$s$  and  $\bar{x}$  refer to the standard deviation and mean, respectively. **X** generates **Y**(92X35), the matrix of coefficients of variance. Each element in **Y** (shown in Figure 6a of the paper) is the coefficient of variance corresponding to the element in **C** having the same matrix coordinates.

#### *Study of Genetic Expression in Gliomas*

The matrix **Z**(92X35X4) is generated by ‘zeroing’ the 4 replicate log2(expression ratios) of **X**(92X35X4) that were filtered by  $f_4$ . The data of each 2-dimensional gene matrix of size 4X35 of **Z** are compared between tumor samples 4-16 and 17-32 by a 2-sample t-test.

#### **Real Time RT PCR**

The primers and annealing temperatures are as follow: G3PDH; forward: 5'-CAAGGTCATCCCTGAGCTGAAC-3', reverse: 5'-TCGCTGTTGAAGTCAGAGGAGAC-3', temp: 60°C. CALM2;

forward: 5'-TGGCTGACCAACTGACTGAAGAG-3', reverse: 5'-

GTA ACTCTGCTTCTGTGGGATTCTG-3', temp: 60°C. TUBB5; forward: 5'-

ATGGGCACGTTGCTCATCAG-3', reverse: 5'-TCGTTGTCGATGCAGTAGGTCTC-

3', temp: 60°C.

## TABLE

The genes whose expression vectors are listed in Figure 7 are shown in order. The first column contains the row number of the matrix in Figure 7, the second includes the IMAGE Ids, and the third contains the clone names.

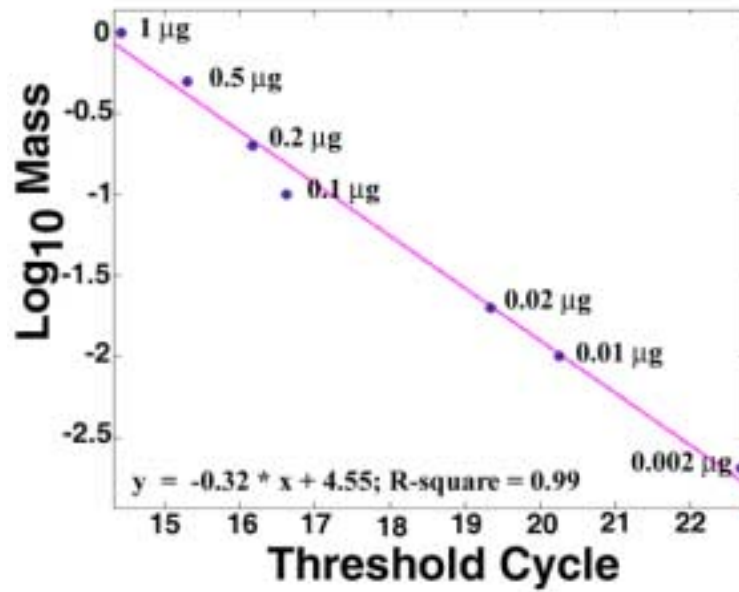
1.	'29830'	'EST'
2.	'26189'	'EST'
3.	'22355'	'RGS4'
4.	'150767'	'EST'
5.	'28347'	'NEFL'
6.	'40396'	'KIAA0319'
7.	'50080'	'EST'
8.	'135556'	'PEG3'
9.	'51067'	'CALM2'
10.	'38747'	'LOC51062'
11.	'114906'	'MAOB'
12.	'205072'	'EST'
13.	'51262'	'FLJ20041'
14.	'142573'	'FAF1'
15.	'26104'	'CHN1'
16.	'42377'	'TNRC15'
17.	'66495'	'CALM2'
18.	'23286'	'RTN1'
19.	'141299'	'RPS5P1'
20.	'41657'	'RTN3'
21.	'133070'	'CALM1'
22.	'33139'	'KIAA0844'
23.	'44590'	'CHN1'
24.	'30826'	'EST'
25.	'114898'	'EST'
26.	'491169'	'EST'
27.	'361174'	'SNAP25'
28.	'280244'	'EST'
29.	'486800'	'CALM2'
30.	'281298'	'PDP'
31.	'298836'	'MEF2C'
32.	'297267'	'EST'
33.	'503296'	'PEX1'
34.	'292459'	'D6S52E'
35.	'291724'	'EST'
36.	'502086'	'YWHAH'

37. '491670' 'ITPR1'  
38. '290290' 'SYT1'  
39. '33205' 'SLC12A7'  
40. '115784' 'FVT1'  
41. '136149' 'KRT7'  
42. '152802' 'PLA2G2A'  
43. '42090' 'KIAA1062'  
44. '29130' 'KIAA1548'  
45. '36118' 'EST'  
46. '162646' 'EST'  
47. '116856' 'FLJ10871'  
48. '126391' 'EST'  
49. '26280' 'SPARC'  
50. '115000' 'LOC51042'  
51. '200180' 'SPP1'  
52. '116431' 'EST'  
53. '34386' 'EST'  
54. '665126' 'FN1'  
55. '505564' 'PDGFRA'  
56. '262047' 'SPARC'  
57. '380396' 'CD74'  
58. '258999' 'COL3A1'  
59. '486617' 'PDGFRA'  
60. '503093' 'IGFBP7'  
61. '303100' 'FN1'  
62. '510179' 'HLA-B'  
63. '279619' 'MIC2'  
64. '300106' 'PAN2'  
65. '489707' 'GDI2'  
66. '240655' 'HLA-DRB5'  
67. '327550' 'HLA-DPB1'  
68. '504908' 'P5-1'  
69. '236270' 'IGL@'  
70. '234382' 'EST'  
71. '303146' 'KIAA0606'  
72. '381205' 'SPARC'  
73. '291409' 'MACS'  
74. '417148' 'HLA-DQA1'  
75. '376499' 'PDGFRA'  
76. '118081' 'FLJ20265'  
77. '72664' 'VIM'  
78. '61283' 'VIM'  
79. '343205' 'HLA-A'  
80. '302482' 'IGFBP7'  
81. '484707' 'HXB'  
82. '490090' 'KIAA1042'

83.	'489898'	'YB1'
84.	'239086'	'APBA3'
85.	'N/A1'	'N/A1'
86.	'365630'	'LAMA3'
87.	'365459'	'TPM1'
88.	'117727'	'KIAA0630'
89.	'488562'	'KIAA1494'
90.	'N/A1'	'N/A1'
91.	'41562'	'EST'
92.	'50043'	'MBP'
93.	'44862'	'PLP1'
94.	'184065'	'LAMR1'
95.	'27217'	'EST'
96.	'27108'	'EST'
97.	'186772'	'EST'
98.	'49713'	'PLP1'
99.	'36462'	'KIAA1189'
100.	'132628'	'C3ORF4'
101.	'27094'	'PTPRZ1'
102.	'153322'	'TU3A'
103.	'25705'	'EST'
104.	'301396'	'EST'
105.	'205130'	'INSL4'
106.	'148425'	'HBB'
107.	'201839'	'EST'
108.	'144221'	'HBB'
109.	'141698'	'KIAA0704'
110.	'200200'	'EST'
111.	'115460'	'PSMD1'
112.	'116445'	'EST'
113.	'136255'	'HBB'
114.	'241804'	'EST'
115.	'306759'	'ST3GALVI'
116.	'238991'	'EST'
117.	'247953'	'HBE1'
118.	'235958'	'SLC6A8'
119.	'230534'	'EST'
120.	'241976'	'HBA1'
121.	'188190'	'IGL@'
122.	'152922'	'IGL@'
123.	'186803'	'IGL@'
124.	'110243'	'IGL@'
125.	'162117'	'IGL@'
126.	'187684'	'IGKC'
127.	'187123'	'IGKC'
128.	'182531'	'IGL@'

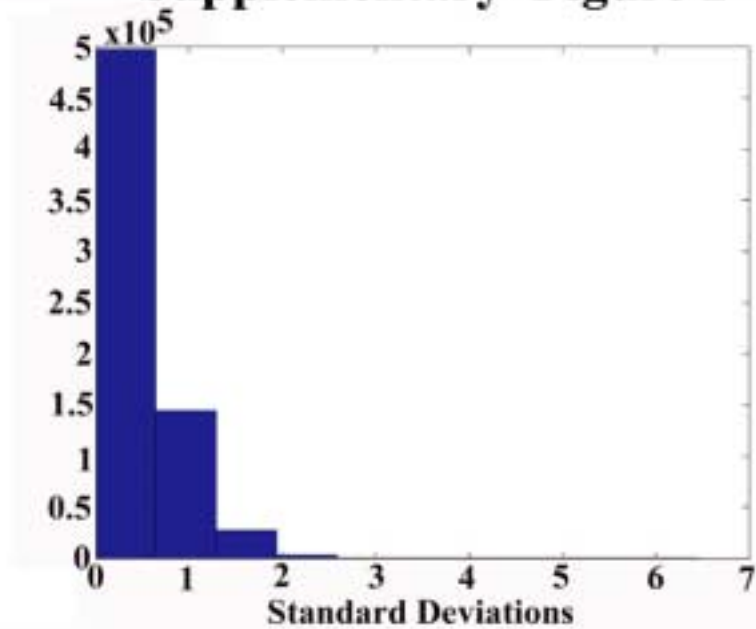
129. '35715' 'GPR51'  
130. '49250' 'EST'  
131. '182866' 'CRYAB'  
132. '162388' 'CRYAB'  
133. '44648' 'NEFL'  
134. '192325' 'PRO1489'  
135. '205633' 'SCYA4'  
136. '48938' 'EST'  
137. '66765' 'EST'  
138. '213575' 'EST'  
139. '41250' 'EST'  
140. '241681' 'EST'  
141. '44958' 'EST'  
142. '23684' 'AOC3'  
143. '51004' 'SLC1A2'  
144. '26249' 'SH3GL2'  
145. '42439' 'KNS2'  
146. '206429' 'CALM3'  
147. '205943' 'EST'  
148. '29649' 'EST'  
149. '299106' 'LAP18'  
150. '275422' 'FLJ12615'  
151. '49707' 'EST'  
152. '272204' 'KIAA0436'  
153. '289019' 'EST'  
154. '665278' 'EST'  
155. '277423' 'DCAMKL1'  
156. '171684' 'TUBB5'  
157. '172076' 'BNPI'  
158. '173674' 'NPTX1'  
159. '172534' 'SYN1'  
160. '357885' 'ATP2B1'  
161. '345809' 'DBI'  
162. '501678' 'EST'  
163. '41173' 'TUBG1'  
164. '289055' 'EST'  
165. '491100' 'EST'

# Supplementary Figure 1



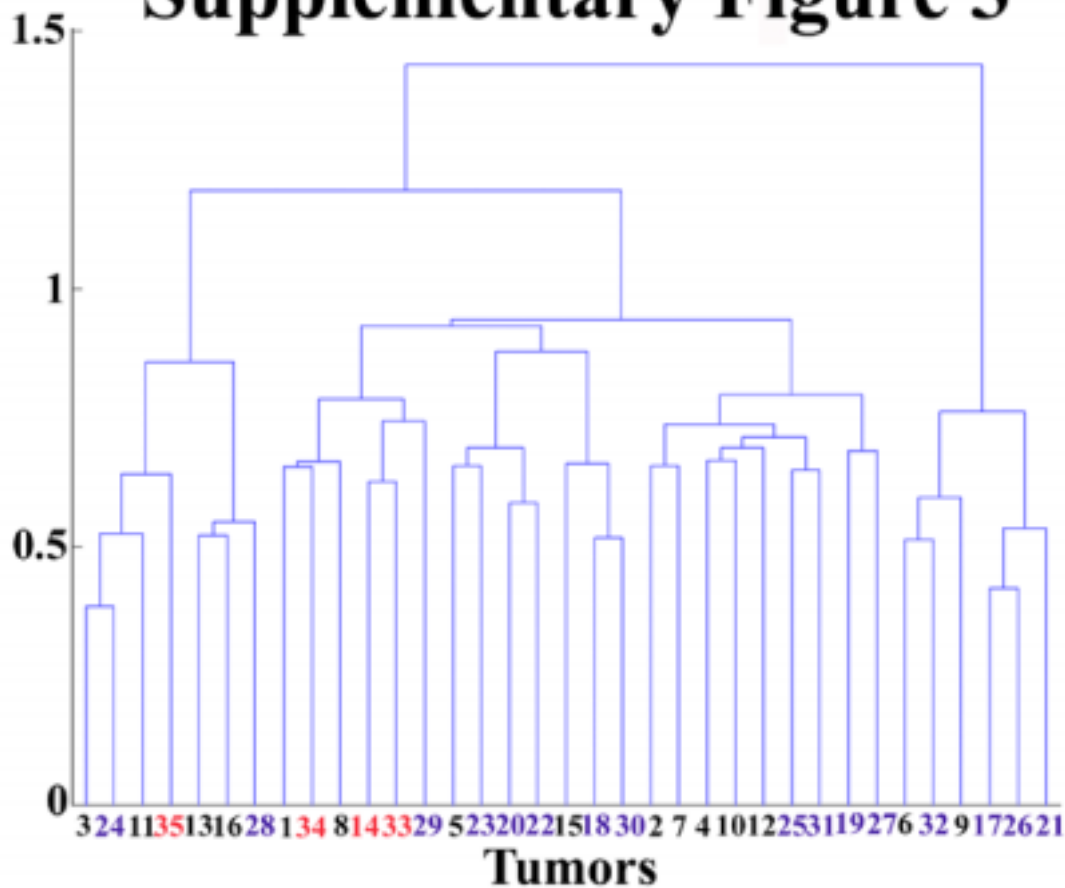
*Supplementary Figure 1:* Linear correlation between serial dilution of starting total RNA and the threshold cycles as measured by one-step hot-start real time RT-PCR for G3PDH using SYBR green. The value of a threshold cycle is computed as the maximum of the second differential of a growth curve. The melting points range from 84.38 to 84.78.

## Supplementary Figure 2



*Supplementary Figure 2:* Histogram showing a normal distribution of the standard deviations of the ‘prepared’ replicate log<sub>2</sub> ratio measurements. The overwhelming majority of the standard deviations are  $\leq 1$ .

## Supplementary Figure 3



*Supplementary Figure 3:* Clustering without the noise model does not replicate the pathological classification. The dendrogram shows the results of agglomerative hierarchical clustering of the mean values of the ‘prepared’ replicate measurements without application of the noise model using single linkage of Ward’s incremental sum of squares of the 1-Pearson product moment correlation matrix. Tumor numbers correspond to Figure 5a of the paper. Blue numbers refer to glioblastoma multiforme, black to lower-grade tumors, and red to samples showing radiation necrosis.