

# Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability

## Additional file 1

M.H. van Vliet<sup>\*1,2</sup>, F. Reyal<sup>3,5</sup>, H.M. Horlings<sup>3</sup>, M.J. van de Vijver<sup>3,4</sup>, M.J.T. Reinders<sup>1</sup>, L.F.A. Wessels<sup>1,2</sup>

<sup>1</sup>Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

<sup>2</sup>Bioinformatics and Statistics group, Department of Molecular Biology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

<sup>3</sup>Department of Pathology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

<sup>4</sup>Department of Pathology, Academic Medical Center, Meibergdreef 9, 1100 DD, Amsterdam, The Netherlands

<sup>5</sup>Department of Surgery, Institut Curie, 6 rue d'Ulm, 75005 Paris, France

Email: M.H. van Vliet\* - M.H.vanVliet@TUDelft.nl;

\*Corresponding author

### Supplementary text S1: Summation of SNRs

Hypothesis: SNR of pooled dataset  $C$  is the sum of the SNRs of the separate datasets  $A$  and  $B$  (given some assumptions)

$$SNR_C = \frac{1}{2}SNR_A + \frac{1}{2}SNR_B \quad (1)$$

$$\frac{\mu(C_1) - \mu(C_2)}{\sigma(C_1) + \sigma(C_2)} = \frac{1}{2} \frac{\mu(A_1) - \mu(A_2)}{\sigma(A_1) + \sigma(A_2)} + \frac{1}{2} \frac{\mu(B_1) - \mu(B_2)}{\sigma(B_1) + \sigma(B_2)} \quad (2)$$

Assumptions

- The two datasets  $A$  and  $B$  have the same number of samples  $N$ ; dataset  $C$  is the concatenation of  $A$  and  $B$ , and consequently has  $2N$  samples,
- The two classes in each dataset have equal priors, the first subscript indicates the class (1 or 2), the second subscript indicates a sample.

For the numerator the following holds:

$$\mu(C_1) - \mu(C_2) = \frac{1}{2N} \sum_{i=1}^{2N} C_{1i} - \frac{1}{2N} \sum_{i=1}^{2N} C_{2i} \quad (3)$$

$$= \frac{1}{2N} \sum_{i=1}^N A_{1i} + \frac{1}{2N} \sum_{i=1}^N B_{1i} - \frac{1}{2N} \sum_{i=1}^N A_{2i} - \frac{1}{2N} \sum_{i=1}^N B_{2i} \quad (4)$$

$$= \frac{1}{2}\mu(A_1) + \frac{1}{2}\mu(B_2) - \frac{1}{2}\mu(A_2) - \frac{1}{2}\mu(B_2) \quad (5)$$

For the denominator the following holds:

$$\sigma(C_1) + \sigma(C_2) = \sqrt{E(C_1^2) - E(C_1)^2} + \sqrt{E(C_2^2) - E(C_2)^2} \quad (6)$$

$$(7)$$

First simplicity for 1 class only:

$$\sigma(C_1)^2 = E(C_1^2) - E(C_1)^2 \quad (8)$$

$$= \frac{1}{2N} \sum_{i=1}^{2N} (C_{1i}^2) - \left(\frac{1}{2}\mu(A_1) + \frac{1}{2}\mu(B_1)\right)^2 \quad (9)$$

$$= \frac{1}{2N} \left( \sum_{i=1}^N (A_{1i}^2) + \sum_{i=1}^N (B_{1i}^2) \right) - \left(\frac{1}{2}\mu(A_1) + \frac{1}{2}\mu(B_1)\right)^2 \quad (10)$$

$$= \frac{1}{2N} \left( \sum_{i=1}^N (A_{1i}^2) + \sum_{i=1}^N (B_{1i}^2) \right) - \frac{1}{4}\mu(A_1)^2 - \frac{1}{4}\mu(B_1)^2 - \frac{1}{2}\mu(A_1)\mu(B_1) \quad (11)$$

$$= \frac{1}{2} \left( \frac{1}{N} \sum_{i=1}^N A_{1i}^2 + \frac{1}{N} \sum_{i=1}^N B_{1i}^2 - \frac{1}{2}\mu(A_1)^2 - \frac{1}{2}\mu(B_1)^2 - \mu(A_1)\mu(B_1) \right) \quad (12)$$

$$= \frac{1}{2} \left( \sigma(A_1)^2 + \sigma(B_1)^2 + \frac{1}{2}\mu(A_1)^2 + \frac{1}{2}\mu(B_1)^2 - \mu(A_1)\mu(B_1) \right) \quad (13)$$

$$= \frac{1}{2} \left( \sigma(A_1)^2 + \sigma(B_1)^2 + \frac{1}{2}(\mu(A_1) - \mu(B_1))^2 \right) \quad (14)$$

Substituting 5 and 14 in 2:

$$SNR_C = \tag{15}$$

$$\frac{\frac{1}{2}\mu(A_1) + \frac{1}{2}\mu(B_1) - \frac{1}{2}\mu(A_2) - \frac{1}{2}\mu(B_2)}{\sqrt{\frac{1}{2}(\sigma(A_1)^2 + \sigma(B_1)^2 + \frac{1}{2}(\mu(A_1) - \mu(B_1))^2) + \frac{1}{2}(\sigma(A_2)^2 + \sigma(B_2)^2 + \frac{1}{2}(\mu(A_2) - \mu(B_2))^2)}} \tag{16}$$

$$\frac{\mu(A_1) + \mu(B_1) - \mu(A_2) - \mu(B_2)}{\sqrt{2\sigma(A_1)^2 + 2\sigma(B_1)^2 + (\mu(A_1) - \mu(B_1))^2 + 2\sigma(A_2)^2 + 2\sigma(B_2)^2 + (\mu(A_2) - \mu(B_2))^2}} \tag{17}$$

If:

$$\mu(A_1) = \mu(B_1) \tag{18}$$

$$\mu(A_2) = \mu(B_2) \tag{19}$$

$$\sigma(A_1) = \sigma(B_1) \tag{20}$$

$$\sigma(A_2) = \sigma(B_2) \tag{21}$$

both 17 and the righthand side of 2 simplify to:

$$\frac{\mu(A_1) - \mu(A_2)}{\sigma(A_1) + \sigma(A_2)} \tag{22}$$