Supplementary Methods

A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes

Benjamin Haibe-Kains, Christine Desmedt, Sherene Loi, Aedin C. Culhane, Gianluca Bontempi, John Quackenbush, Christos Sotiriou

Contents

1 Subtype Classification Model (SCM)

This section recapitulates the design and fitting of the Subtype Classification Model (SCM) as described in details in Supplementary information of (1, 3, 11).

In order to identify the molecular subtypes of breast cancer tumors, we performed a clustering in a two-dimensional space. The dimensions were defined by the ESR1 and ERBB2 module scores (1, 11). To facilitate comparison between datasets, we applied a robust linear scaling to each of these scores such that quantiles 2.5% and 97.5% were set to -1 and +1 respectively. This procedure was particularly efficient in datasets where skewed population of patients (such as those with different proportions of ER-/+ or HER2-/+ tumors) since only a few extreme cases (5%) are needed to perform the robust scaling, while not relying on outliers.

We used a simple clustering model that is a mixture of Gaussians with equal variance and shape (2, 7). Since Kapp et al. showed that only three main breast cancer molecular subtypes (basal-like, HER2-enriched and luminal) can be identified in multiple datasets (5), we fitted a mixture of three Gaussians to identify the ER-/HER2- (alias basal-like), HER2+ (alias HER2-enriched), and ER+/HER2- (alias luminal) subtypes. Moreover, since we showed that the discrimination between luminal A and B can be performed using proliferation-related genes (6), we used a proliferation module scores (referred herein by AURKA) to identify the low and high proliferative ER+/HER2- tumors. However, we and others reported that such a proliferation signal is a continuum and do not exhibit natural cutoffs representing different proliferative stages (4, 5, 8, 10); therefore we assumed that, in a representative population of breast cancer patients, half of the ER+/HER2- tumors are lowly proliferative (alias luminal A) and the other half are highly proliferative (alias luminal B).

Note that an implementation of our method for breast cancer molecular subtype identification is available from the R package `genefu`[1] (see functions `subtype.cluster`, `subtype.cluster.predict` and objects `scmod1`, `scmod2`, and `scmgene`).

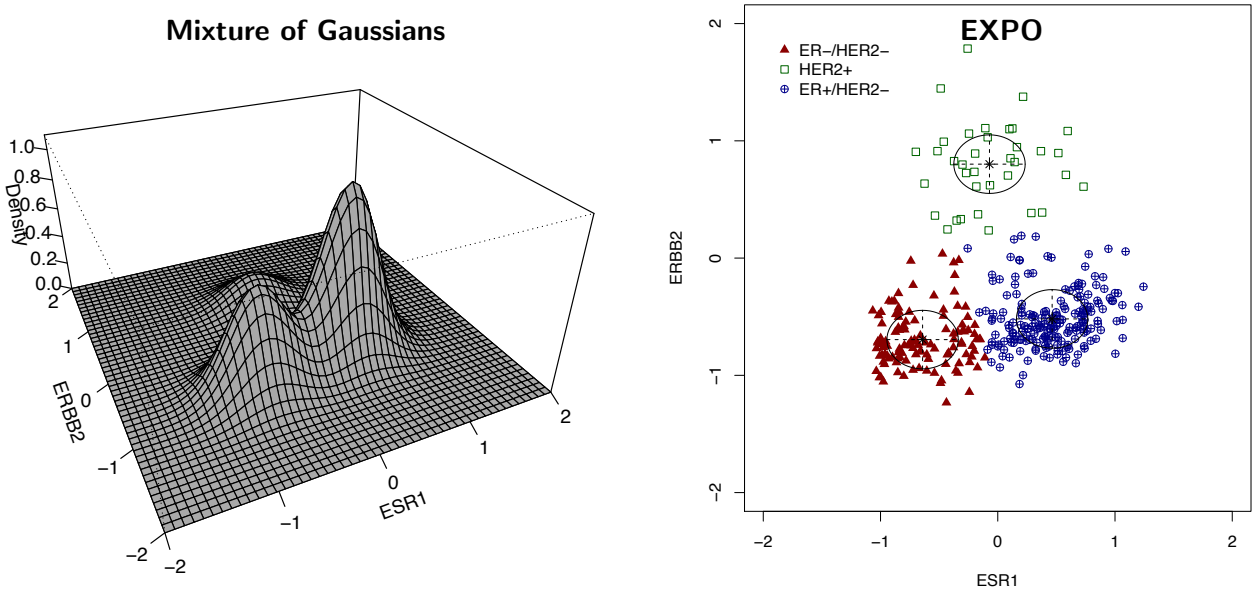
This subtype classification model (SCM), once fitted on the training set, returns a set of probabilities for a patient to belong to each molecular subtype.

SCMOD2

The first SCM, referred to as SCMOD2, has been published by Wirapati et al. where ER, HER2 signaling pathways and proliferation were quantified by averaging expressions of genes included in the modules ESR1, ERBB2 and AURKA respectively (11).

As can be seen in the figure below, the mixture of three Gaussians fitted on the training set (EXPO, see Table 1) enables the identification of the three main molecular subtypes: ER-/HER2- (alias basal-like), HER2+ (alias HER2-enriched), and ER+/HER2- (alias luminal).

[1] `http://www.bioconductor.org/packages/release/bioc/html/genefu.html`



The following table gives the parameters of the clustering model (mixture of three Gaussians with equal shape and variance) as fitted on the training set (EXPO, see Table 1).

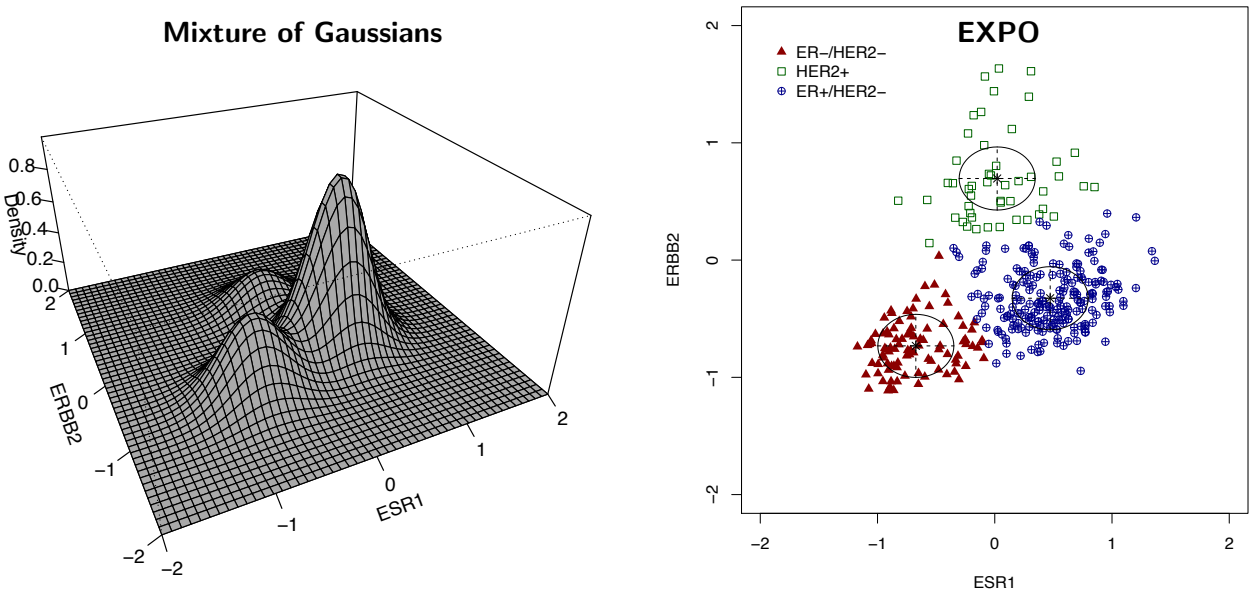
$\hat{\mu}$	ER-/HER2-	HER2+	ER+/HER2-
ESR1	-0.64	-0.07	0.46
ERBB2	-0.69	0.80	-0.52
$\hat{\Sigma} \times I$			
ESR1	0.09	0.09	0.09
ERBB2	0.06	0.06	0.06
$\hat{\pi}$	0.34	0.10	0.56

Finally, we estimated a cutoff value for the AURKA module in order to discriminate between low and high proliferative tumors. This cutoff was defined as the median and is equal to -0.27.

SCMOD1

Another version of the SCM, referred to as SCMOD1, has been published by Desmedt et al. where ER, HER2 signaling pathways and proliferation were quantified by averaging expressions of genes included in the modules ESR1, ERBB2 and AURKA respectively (1).

As can be seen in the figure below, the mixture of three Gaussians fitted on the training set (EXPO, see Table 1) enables the identification of the three main molecular subtypes: ER-/HER2- (alias basal-like), HER2+ (alias HER2-enriched), and ER+/HER2- (alias luminal).



The following table gives the parameters of the mixture of three Gaussians with equal shape and variance, as fitted on the training set (EXPO, see Table 1).

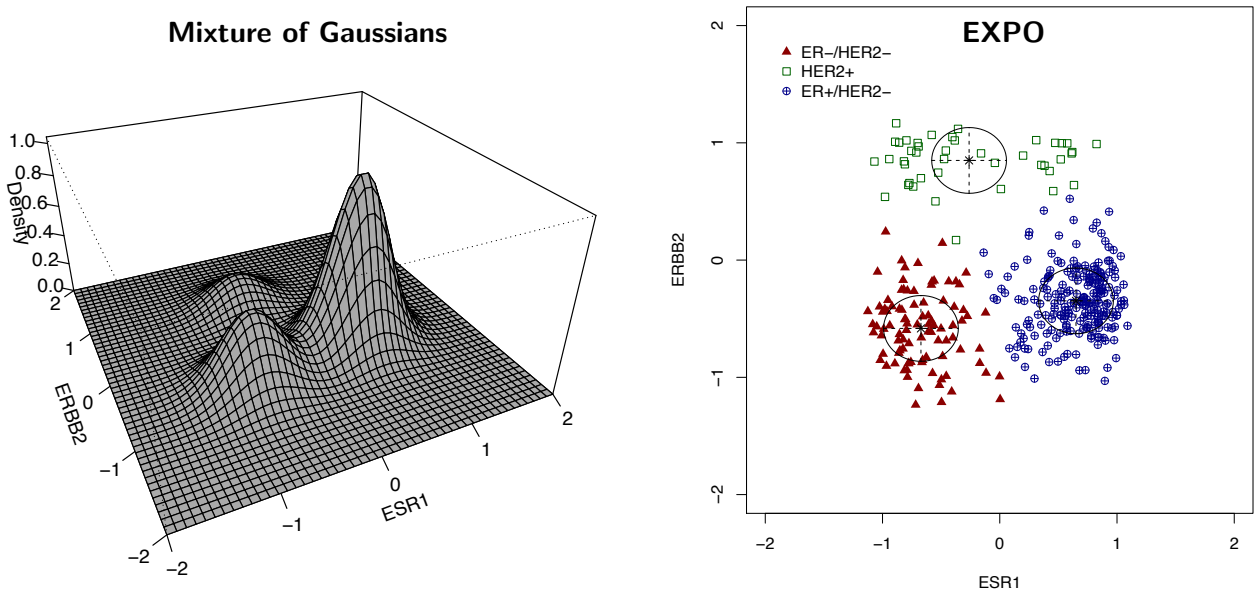
$\hat{\mu}$	ER-/HER2-	HER2+	ER+/HER2-
ESR1	-0.68	0.02	0.47
ERBB2	-0.73	0.70	-0.32
$\hat{\Sigma} \times I$			
ESR1	0.10	0.10	0.10
ERBB2	0.07	0.07	0.07
$\hat{\pi}$	0.27	0.14	0.59

Finally, we estimated a cutoff value for the AURKA module in order to discriminate between low and high proliferative tumors. This cutoff was defined as the median and is equal to -0.30.

SCMGENE

In the present work we developed a three-gene version of the SCM, referred to as SCMGENE, where ER, HER2 signaling pathways and proliferation were quantified by expression of single genes that are ESR1, ERBB2 and AURKA respectively (1).

As can be seen in the figure below, the mixture of three Gaussians fitted on the training set (EXPO, see Table 1) enables the identification of the three main molecular subtypes: ER-/HER2- (alias basal-like), HER2+ (alias HER2-enriched), and ER+/HER2- (alias luminal).



The following table gives the parameters of the mixture of three Gaussians with equal shape and variance, as fitted on the training set (EXPO, see Table 1).

$\hat{\mu}$	ER-/HER2-	HER2+	ER+/HER2-
ESR1	-0.67	-0.26	0.65
ERBB2	-0.58	0.84	-0.34
$\hat{\Sigma} \times I$			
ESR1	0.10	0.10	0.10
ERBB2	0.08	0.08	0.08
$\hat{\pi}$	0.27	0.12	0.61

Finally, we estimated a cutoff value for the AURKA module in order to discriminate between low and high proliferative tumors. This cutoff was defined as the median and is equal to -0.37.

2 R code for SCMGENE

In this section we describe in details how to fit SCMGENE using the EXPO dataset (see Table 1) by following two approaches: (i) a few R code lines relying on functions implemented in the *genefu* R/Biocinductor package[2]; (ii) a standalone code fitting the mixture of three Gaussians using ESR1, ERBB2 and AURKA gene expressions.

R code using genefu

Start an R session and download/install *genefu* version 1.3.4 available from the companion website of the publication[3].

Load the *genefu* library

```
> library(genefu)
```

Load the R workspace containing the dataset EXPO as provided in the companion website of the publication. R objects *data*, *annot* and *demo* contain the normalized gene expression values, probe annotations and patients' clinical information respectively.

```
> load("data/EXPO.RData")
```

Extract the probesets for ESR1, ERBB2 and AURKA genes used in SCMOD1 (1) and put them in a list called *modgene*.

```
> modgene <- lapply(scmod1$mod, function(x) { return(x[1, , drop=FALSE]) })
> print(modgene)
```

Construct the *scmgene.expo* object, which the SCMGENE model used in the present work, by using the function *subtype.cluster*. The files `scmgene_fit_EXPO.pdf` and `scmgene_model_EXPO.csv` are created in the current directory, they describe the fit of the mixture of the three Gaussians in terms of classification and parameters, respectively.

```
> pdf("scmgene_fit_EXPO.pdf", width=7, height=7)
> tt <- subtype.cluster(module.ESR1=modgene$ESR1, module.ERBB2=modgene$ERBB2,
module.AURKA=modgene$AURKA, data=data, annot=annot, do.mapping=FALSE, do.scale=TRUE,
rescale.q=0.05, plot=TRUE, filen=sprintf("scmgene_model_EXPO"))
> dev.off()
> scmgene.expo <- tt$model
```

The following function enables classification of the breast tumors into subtypes (function *subtype.cluster.predict*) using SCMGENE in an affymetrix HG-U133A or HG-U133PLUS2 microarray dataset (for example VDX, see Table 1). Note that, since SCMGENE has been trained on EXPO (Affymetrix HG-U133PLUS2), mapping is unnecessary (*do.mapping=FALSE*). A file `scmgene_classif_VDX.pdf` is created and illustrates the classification of the VDX dataset into the four molecular subtypes: ER-/HER2-,
HER2+, ER+/HER2- High Proliferation, and ER+/HER2- Low Proliferation which correspond to basal-like, HER2-enriched, luminal B and luminal A in the present work.

[2] `http://www.bioconductor.org/packages/devel/bioc/html/genefu.html`
[3] `http://compbio.dfci.harvard.edu/pubs/sbtpaper/`

```
> load(sprintf("data/VDX.RData"))
> pdf("scmgene_classif_VDX.pdf", width=7, height=7)
> sc.vdx <- subtype.cluster.predict(sbt.model=scmgene.expo, data=data, annot=annot,
do.mapping=FALSE, plot=TRUE, verbose=TRUE)
> dev.off()
```

The subtyping for each tumor can be accessed through *sc.vdx$subtype2*.

```
> print(table(sc.vdx$subtype2))
```

Since the function *subtype.cluster.predict* enables automatic mapping (*do.mapping=TRUE*), we can easily classify tumors into subtypes from a dataset using a non-Affymetrix microarray platform (for example NKI, see Table 1).

```
> load(sprintf("data/NKI.RData"))
> pdf("scmgene_classif_NKI.pdf", width=7, height=7)
> sc.nki <- subtype.cluster.predict(sbt.model=scmgene.expo, data=data, annot=annot,
do.mapping=TRUE, plot=TRUE, verbose=TRUE)
> dev.off()
```

Similarly the subtyping for each tumor can be accessed through *sc.nki$subtype2*.

```
> print(table(sc.nki$subtype2))
```

Subtyping of a single sample If the microarray platform and the single chip normalization procedure are standardized throughout an entire study (for example all tumor samples are profiled using Affymetrix HG-U133A and normalized with MAS5 or fRMA), then one could easily classify a single sample into its most likely subtype. Actually we just have to fit the model without scaling the gene expression data (*do.scale=FALSE* in function *subtype.cluster*).

Here is an example with the EXPO dataset where all samples have been profiled on Affymetrix HG-U133PLUS2 and normalized with fRMA: the first 352 samples are used for training SCMGENE (*data.training*) and the last sample is the single sample to classify (*single.test.sample*).

```
> library(genefu)
> load("data/EXPO.RData")
> modgene <- lapply(scmod1$mod, function(x) { return(x[1, , drop=FALSE]) })
> data.training <- data[1:(nrow(data)-1), , drop=FALSE]
> single.test.sample <- data[nrow(data), , drop=FALSE]
> scmgene.unscaled.expo <- subtype.cluster(module.ESR1=modgene$ESR1, module.ERBB2=modgene$
module.AURKA=modgene$AURKA, data=data.training, annot=annot, do.mapping=FALSE, do.scale=FA
> > sc <- subtype.cluster.predict(sbt.model=scmgene.unscaled.expo, data=single.test.sample
do.mapping=FALSE, plot=TRUE, verbose=TRUE)
```

The predicted subtype for the single tumor sample can be accessed through *sc$subtype2*.

```
> print(sc$subtype2)
```

Standalone R code

In this section we describe how to fit the mixture of three Gaussians used in SCMGENE and how to compute the maximum posterior probabilities. The code and its documentation are also available the the package *genefu*, function *subtype.cluster*.

Load the EXPO dataset.

```
> load("data/EXPO.RData")
```

Extract the probesets for ESR1, ERBB2 and AURKA genes used in SCMOD1 (1) and put them in a list called *modgene*.

```
> modgene <- lapply(scmod1$mod, function(x) { return(x[1, , drop=FALSE]) })
> print(modgene)
```

Extract the gene expression values for ESR1, ERBB2 and AURKA and rescale them. Note that rescale is unnecessary if you apply SCMGENE on datasets using the same microarray platform and normalization method (which is not the case in the present work).

```
> ge.expo <- lapply(modgene, function(x, y) {
xx <- y[ ,x[ ,1,drop=TRUE]]
qq <- quantile(xx, probs=c(0.025, 0.975))
return((((xx - qq[1]) / (qq[2] - qq[1])) - 0.5) * 2)
},y=data)
```

Let's now fit the mixture of three Gaussians using ESR1 and ERBB2 expressions (R package *mclust* available from CRAN[4]).The object *mg3* contains all the parameters of the mixture of three Gaussians.

```
> library(mclust)
> mg3 <- Mclust(data=cbind("ESR1"=ge.expo$ESR1,"ERBB2"=ge.expo$ERBB2), modelNames="EEI", G
```

The classification and corresponding posterior probabilities are stored in the *mg3* object. Note that in this case classes 1, 2 and 3 represent the HER2+, ER-/HER2-, and ER+/HER2- respectively.

```
> print(table(mg3$classification))
> print(mg3$z)
```

We can discriminate between luminal B and luminal A tumors by using AURKA gene expressions. Objects *subtype2* and *subtype.proba2* contain the subtyping and the corresponding posterior probabilities.

```
> nn <- c("ER-/HER2-", "HER2+", "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif")
> aurka.cutoff <- median(ge.expo$AURKA[!is.na(mg3$classification) &
mg3$classification == 3], na.rm=TRUE)
> subtype2 <- mg3$classification
> subtype2[!is.na(mg3$classification) & mg3$classification == 3 &
```

[4] http://cran.r-project.org/

```
ge.expo$AURKA < aurka.cutoff] <- 4
> subtype2 <- nn[c(2, 1, 3, 4)][subtype2]
> tt <- mg3$z[ ,3]
> tt2 <- rep(NA, length(ge.expo$AURKA))
> names(tt2) <- names(ge.expo$AURKA)
> iix <- ge.expo$AURKA < aurka.cutoff
> tt2[iix] <- (ge.expo$AURKA[iix] - min(ge.expo$AURKA[iix], na.rm=TRUE)) /
((max(ge.expo$AURKA[iix], na.rm=TRUE) - min(ge.expo$AURKA[iix], na.rm=TRUE)) * 2)
> tt2[!iix] <- 0.5 + (ge.expo$AURKA[!iix] - min(ge.expo$AURKA[!iix], na.rm=TRUE)) /
((max(ge.expo$AURKA[!iix], na.rm=TRUE) - min(ge.expo$AURKA[!iix], na.rm=TRUE)) * 2)
> tt <- cbind(tt * tt2, tt * (1 - tt2))
> subtype.proba2 <- cbind(mg3$z[ ,c(2, 1)], tt)
> colnames(subtype.proba2) <- nn
```

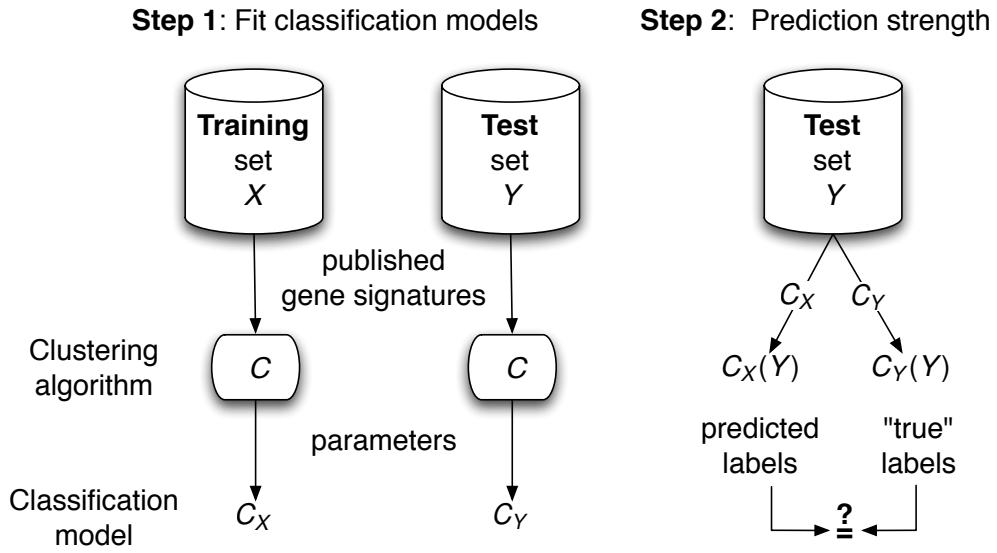
Using this mixture of Gaussians we can easily classify tumors present in the VDX dataset.

```
> load("data/VDX.RData")
> ge.vdx <- lapply(modgene, function(x, y) {
xx <- y[ ,x[ ,1,drop=TRUE]]
qq <- quantile(xx, probs=c(0.025, 0.975))
return((((xx - qq[1]) / (qq[2] - qq[1])) - 0.5) * 2)
},y=data)
> mgt <- estep(modelName="EEI", data=cbind("ESR1"=ge.vdx$ESR1,"ERBB2"=ge.vdx$ERBB2), param
> subtype2 <- map(mgt$z)
> subtype2[!is.na(map(mgt$z)) & map(mgt$z) == 3 &
ge.vdx$AURKA < aurka.cutoff] <- 4
> subtype2 <- nn[c(2, 1, 3, 4)][subtype2]
> tt <- mgt$z[ ,3]
> tt2 <- rep(NA, length(ge.vdx$AURKA))
> names(tt2) <- names(ge.vdx$AURKA)
> iix <- ge.vdx$AURKA < aurka.cutoff
> tt2[iix] <- (ge.vdx$AURKA[iix] - min(ge.vdx$AURKA[iix], na.rm=TRUE)) /
((max(ge.vdx$AURKA[iix], na.rm=TRUE) - min(ge.vdx$AURKA[iix], na.rm=TRUE)) * 2)
> tt2[!iix] <- 0.5 + (ge.vdx$AURKA[!iix] - min(ge.vdx$AURKA[!iix], na.rm=TRUE)) /
((max(ge.vdx$AURKA[!iix], na.rm=TRUE) - min(ge.vdx$AURKA[!iix], na.rm=TRUE)) * 2)
> tt <- cbind(tt * tt2, tt * (1 - tt2))
> subtype.proba2 <- cbind(mgt$z[ ,c(2, 1)], tt)
> colnames(subtype.proba2) <- nn
```

If one wants to classify tumors in a dataset generated with a different microarray platform, we suggest to use the function *geneid.map* in the *genefu* package.

3 Prediction strength

Illustration of the idea behind the prediction strength of a clustering/classification model C with one training set and one test set.



To compute the prediction strength statistic, one dataset was considered as training set and the remaining ones were considered as test sets. First the original gene lists (the intrinsic gene lists for SSPs and the gene modules for SCMs, see Figure 1c,d in the main manuscript) and algorithms were applied both on the training and test sets in order to tune the parameter that are the centroids for SSPs and the Gaussians for SCMs (see Figure 1a,b in the main manuscript); the resulting classifications were referred to as true labels. Note that for SSPs, the subtypes were identified as the main clusters which contain at least 5 clusters. Second, the classification model fitted on the training set was applied to all the test sets (predicted labels) and compared to the classification (true labels) obtained on the first step. The prediction strength was then used to quantify the similarity between both classifications in each dataset separately. Values range from 0 (low similarity) to 1 (high similarity), a prediction strength ≥ 0.8 being representative of a robust classification model (9).

Formally the prediction strength is defined as

$$ps = \min_{1 \leq j \leq u} \frac{1}{n_{k_j}(n_{k_j} - 1)} \sum_{i \neq i' \in k_j} D\left[C_X(Y)\right]_{ii'}$$

where

- $k \in K$ is a cluster of objects
- k_j's with $1 \leq j \leq u$, are the clusters defined by the clustering $C_Y(Y)$
- n_{k_j} is the number of objects in cluster k_j

- D is the co-membership matrix such that

$$D\left[C_X(Y)\right]_{ii'} = \begin{cases} 1 \text{ if } i, i' \in k \\ 0 \text{ otherwise} \end{cases}$$

The prediction strength can also be defined at the cluster and individual levels.
For cluster j;

$$ps_j = \frac{1}{n_{k_j}(n_{k_j}-1)} \sum_{i \neq i' \in k_j} D\left[C_X(Y)\right]_{ii'}$$

For individual i,

$$ps(i) = \frac{1}{\#A_k(i)} \sum_{i' \in A_k(i)} D\left[C_X(Y)\right]_{ii'}$$

where $A_k(i)$ are the objects indices $i' \in Y$ such that $i \neq i' \wedge D\left[C_Y(Y)\right]_{ii'} = 1$ and $\#A_k(i)$ is the number of indicies in $A_k(i)$.

References

[1] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart, and Christos Sotiriou. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. *Clin Cancer Res*, 14(16):5158–5165, 2008. doi: 10.1158/1078-0432.CCR-07-4756. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/14/16/5158.

[2] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of American Statistical Asscoiation*, 97(458):611–631, June 2002.

[3] Benjamin Haibe-Kains, Christine Desmedt, Francoise Rothe, Martine Piccart, Christos Sotiriou, and Gianluca Bontempi. A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome Biology*, 11(2):R18, 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-2-r18. URL http://genomebiology.com/2010/11/2/R18.

[4] Anna V Ivshina, Joshy George, Oleg Senko, Benjamin Mow, Thomas C Putti, Johanna Smeds, Thomas Lindahl, Yudi Pawitan, Per Hall, Hans Nordgren, John E L Wong, Edison T Liu, Jonas Bergh, Vladimir A Kuznetsov, and Lance D Miller. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*, 66(21):10292–301, Nov 2006. doi: 10.1158/0008-5472.CAN-05-4414.

[5] Amy Kapp, Stefanie Jeffrey, Anita Langerod, Anne-Lise Borresen-Dale, Wonshik Han, Dong-Young Noh, Ida Bukholm, Monica Nicolau, Patrick Brown, and Robert Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(1):231, 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-231. URL http://www.biomedcentral.com/1471-2164/7/231.

[6] Sherene Loi, Benjamin Haibe-Kains, Christine Desmedt, Francoise Lallemand, Andrew M. Tutt, Cheryl Gillet, Paul Ellis, Adrian Harris, Jonas Bergh, John A. Foekens, Jan G.M. Klijn, Denis Larsimont, Marc Buyse, Gianluca Bontempi, Mauro Delorenzi, Martine J. Piccart, and Christos Sotiriou. Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor-Positive Breast Carcinomas Through Genomic Grade. *J Clin Oncol*, 25(10):1239–1246, 2007. doi: 10.1200/JCO.2006.07.1522. URL http://jco.ascopubs.org/cgi/content/abstract/25/10/1239.

[7] G. McLachlan and D. Peel. *Finite Mixture Models*. Probability and statistics. Applied probability and statistics. J. Wiley and Sons, 2000.

[8] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, and Mauro Delorenzi. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *J. Natl. Cancer Inst.*, 98(4):262–272, 2006. doi: 10.1093/jnci/djj052. URL http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;98/4/262.

[9] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[10] Britta Weigelt, Alan Mackay, Roger A'hern, Rachael Natrajan, David S P Tan, Mitch Dowsett, Alan Ashworth, and Jorge S Reis-Filho. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol*, 11(4):339–49, Apr 2010. doi: 10.1016/S1470-2045(10)70008-5.

[11] Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frederic Schutz, Darlene Goldstein, Martine Piccart, and Mauro Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):R65, 2008. ISSN 1465-5411. doi: 10.1186/bcr2124. URL `http://breast-cancer-research.com/content/10/4/R65`.