Author's response to reviews

Title: An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer.

Authors:

Iker Sánchez Navarro (ikersana@gmail.com)
Angelo Gámez Pozo (angelogamez@gmail.com)
Álvaro Pinto (alvaro_pinto_marin@hotmail.com)
David Hardisson (dhardisson.hulp@salud.madrid.org)
Rosario Madero (madero.hulp@salud.madrid.org)
Rocío López (rlvacas@gmail.com)
Belén San José (bsjv@gmail.com)
Pilar Zamora (zamorapilar@terra.es)
Andrés Redondo (redondo12@gmail.com)
Jaime Feliu (jfeliu.hulp@salud.madrid.org)
Paloma Cejas (palomacejas@gmail.com)
Manuel González Barón (mgonzalezb.hulp@salud.madrid.org)
Juan Ángel Fresno Vara (jafresco@gmail.com)
Enrique Espinosa (eespinosa00@terra.es)

Version: 4 Date: 11 February 2010

Author's response to reviews: see over
Dear sir:

Thank you for your interest in the above referenced study. Please find below our point-by-point answer to reviewers’ comments. As you can see, the explanation about the mathematical model to find a gene profile is long and, for this reason, we have not included it in the new version of the manuscript. However, if you think that part of this text should be included, please let us know.

We have uploaded a new version of the manuscript, with changes outlined in blue color. There are two supplementary files with a new table and a new figure. The first page, the competing interest section and the references have been remodeled to fit the format required by BioMed Central medicine journals.
Reviewer's 1 report:
The authors analyze the expression of 83 genes from the 70-gene profile; the recurrence score and a 2-gene score in 153 paraffin-embedded breast tumor samples. The data for the 3 profiles has previously been compared and published by the authors. In this manuscript they develop a new 8-gene profile and compare it to the 3 other profiles. Although several prognostic and predictive profiles are published it is important to do more analyses to identify the best combination of genes in the profiles. Furthermore, it is an advantage to make tests with few genes that can be performed in local labs.

Discretionary Revisions
1) The filtering from 83 to 53 genes and use of fresh frozen samples should be explained in more detail.

RNA isolated from FFPE is considered a poor material for gene expression analysis owing to its extensive degradation. While microarray-based studies are highly sensitive to RNA degradation, qRT-PCR appears to be more robust and tolerates partial degradation of RNA (Endoh et al, J Clin Oncol 2004;22:811-9.). Several studies have demonstrated the feasibility of using FFPE tissues to perform gene expression profiling by qRT-PCR. Although RNA degradation leads to a loss of amplifiable templates, optimized normalization strategies could effectively compensate for this bias (Cronin et al, Am J Pathol 2004;164:35-42 // Antonov J et al, Lab Invest 2005;85:1040-50 // Paik S et al, N Engl J Med 2004;351:2817-26.).

We propose that the normalization factor should be able to control for the several levels of experimental variability in qRT-PCR, mainly the quantity of RNA and differences in enzymatic efficiencies, but also for differences in the quality of the starting material. We performed a previous work in which correlation in the quantitative, normalized, data between 30 matched FF and FFPE samples was determined per gene using the Pearson correlation coefficient. We excluded from subsequent analyses genes showing non-significant correlation. Statistical significance was set at p<0.005.

It is important to select the best set of genes when developing a predictive test to be used in the clinical setting. The number of genes that are actually informative is usually small compared to the number of genes that are noisy. Including too many noise genes can dilute the influence of the informative genes and reduce the accuracy of prediction, and makes interpretation and future use of the predictor more difficult. We consider that only the assays that are comparable between FF and FFPE tissue should be included in predictors, in order to guarantee reproducibility and reliability.

These results are explained in more detail in a manuscript called Comparison of Gene Expression Profiling by Quantitative Reverse Transcription PCR between Fresh Frozen and Formalin-Fixed, Paraffin Embedded Breast Cancer Tissues (I Sánchez-Navarro, A Gámez-Pozo, E Espinosa, A Pinto-Marín, D Hardisson, R López, R Madero, M Mendiola, M González-Barón and JA Fresno Vara), which is currently under consideration in the journal Biotechniques.
2) The feature selection method using p-values and correlation should also be explained in more detail

We evaluated the correlation between the expressions of each gene respect all other genes. Positively and significantly (p<0.05) correlated genes were included in the same correlation group. From each multigene correlation groups, only the two genes with the highest relation with DMFS were included in the 8-gene score. This method was previously described by Paik et al.

Major Compulsory Revisions
1) All samples in the data set are used to select the 8 genes. This introduces an information leakage in the leave-one-out cross validation (LOUCV). The bias is considerably smaller in this 83-gene (53 after filtering) data set compared to genome-wide analyses, but the manuscript would be improved by selection of genes in each LOUCV cycle, so that no information from the test set (the left out sample) was used in training. To build a final overall profile for validation in external data sets, the entire data set could be used.
2) The principal component method results in the "8-gene score". The cut-off point in this score is set to include 60% of patients in the low risk group. Why was the cut-off set to 60%? An alternative would be to select the optimal cut-off point in the single rounds of LOUCV.

These two comments relate to the prognostic model development, so we shall deal with them jointly.

As stated in the manuscript, we used the BRB Array Tools (developed by R. Simon and A. Peng) software to develop the gene expression-based predictor. All the reviewer requirements are implemented in BRB Array Tools (Analysis of DNA microarray expression data. Simon R. Best Pract Res Clin Haematol. 2009; 22: 271-82).

Regarding the cut-off point, we have added this text in M&M to clarify the issue: “The cut-off point was established prior to gene selection, validating the whole process using the LOOCV. We assessed cut-off points leaving from 10% to 90% of patients in the low risk group increasing a ten percent each time”. We have also rephrased this part of the text in Results: “We assessed cut-off points leaving from 10% to 90% of patients in the low risk group increasing a ten percent each time. Best performance was obtained when 60% of the patients were allocated to the low-risk group”.

The following text, extracted from BRB Array Tools Manual, shows the methodology we used:

The Survival Analysis Tool for finding genes whose expression is correlated with survival time fits proportional hazards models relating survival to each gene, one gene at a time and computes the p value for each gene for testing the hypothesis
that survival time is independent of the expression level for that gene. Gene lists are created based on these p values in the same way as in the Class Comparison tools. The p values can be used to construct gene lists using multivariate permutation tests for controlling the number or proportion of false discoveries. Or the gene list can simply consist of the genes with p values less than a specified threshold (0.001 is default). For more information regarding the multivariate permutation tests for controlling the number or proportion of false discoveries, please see the preceding section on Multivariate Permutation Tests for Controlling Number and Proportion of False Discoveries.

“The Survival Analysis Prediction Tool develops a gene expression based predictor of survival risk group. The number of risk groups and the risk percentiles for defining the groups are specified by the user. The survival risk groups are constructed using the supervised principal component method of E Bair and R Tibshirani (2004):


This method uses a Cox proportional hazards model to relate survival time to k “super-gene” expression levels, where k is selectable by the user (usually 1-3). The “supergene” expression levels are the first k principal component linear combinations of expression levels of the subset of genes that are univariately correlated with survival. The user specifies the threshold significance level (e.g. 0.001) for selecting the genes to be used in computing the principal components. The significance of each gene is measured based on a univariate Cox proportional hazards regression of survival time versus the log expression level for the gene. After selecting the genes, the principal components are computed, and the k-variable Cox proportional hazard regression analysis is performed. This provides a regression coefficient (weight) for each principal component.

Having developed a supervised principal component model as described above, to compute a prognostic index for a patient whose expression profile is described by a vector x of log expression levels, the following steps are performed. First the components of the vector x corresponding to the genes that were selected for use in computing the principal components are identified. Then the k principal components are computed. These are just linear combinations of the components of x, with the weights of each linear combination having been determined from the principal component analysis described above. Finally, the weighted average of these k principal component values is computed, using as weights the regression coefficients derived from the k-variable Cox regression described above. This computation provides a prognostic index for a patient with a log expression profile given by a vector x. A high value of the prognostic index corresponds to a high value of hazard of death, and consequently a relatively poor predicted survival.

In order to evaluate the predictive value of the method, Leave-One-Out-Cross-Validation is used. A single case is omitted and the entire procedure described
above is performed to create a prognostic index. This function is created from scratch on the training set with the one case removed, including determining the genes to be included in the calculation of the principal components. Having determined a prognostic index function for that training set, it is used to compute a prognostic index for the omitted observation. That value is compared to the prognostic index for the n-1 cases included in that training set (assuming that there are n distinct cases available in total). The prognostic index for the omitted patient is ranked relative to the prognostic index for the patients included in the cross-validated training set. The omitted patient is placed into a risk group based on his/her percentile ranking, the number of risk groups specified, and the cut-off percentiles specified for defining the risk groups. This analysis is repeated from scratch n times, leaving out a different case each time.

Having completed the computations described in the previous paragraph, we plot Kaplan-Meier survival curves for the cases predicted to have above average risk and the cases predicted to have below average risk. It is important to note that the risk group for each case was determined based on a predictor that did not use that case in its construction. Hence, the Kaplan-Meier curves are essentially unbiased and the separation between the curves gives a fair representation of the value of the expression profiles for predicting survival risk."

3) It would be helpful for the reader to have sensitivity and specificity for the 8-gene profile and the other 3 profiles in this data set. This would also help to compare the performance of 8-gene profile and the 70-gene profile in the NKI data set.

Michael J. Pencina and Ralph B. D’Agostino have investigated the properties of the overall concordance index introduced by Harrell as a natural extension of the ROC curve area to survival analysis (Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004;23:2109-23).

A concordance index (CI) is a widely applicable measure of predictive discrimination - one that applies to ordinary continuous outcomes, dichotomous diagnostic outcomes, ordinal outcomes, and censored time until event response variables. CI is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. CI measures predictive information derived from the set of predictor variables in a model. For predicting binary outcomes such as the presence of disease, concordance reduces to the proportion of all pairs of patients, one with and one without the disease, in which the patient having the disease had the higher predicted probability of disease. In this binary outcome case, concordance is essentially the Wilcoxon-Mann-Whitney statistic for comparing predictions in the two outcome groups, and it is identical to the area under a receiver operating characteristic (ROC) curve. CI estimates the probability of concordance between predicted and observed responses. A value of 0.5 indicates no predictive discrimination and a value of 1 indicates perfect separation of patients with different outcomes. CI can be used to quantify the predictive
discrimination of any quantitative predictive method, whether the response is continuous, ordinal, or binary.

We have included the information regarding the comparison of the 8-gene Score and other gene profiles using the NKI online database as supplementary table 4.

<table>
<thead>
<tr>
<th>Profile</th>
<th>Group</th>
<th>% good vs poor</th>
<th>DMFS at 5 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-gene Score</td>
<td>All patients</td>
<td>52 vs 48</td>
<td>85.7 ± 2.7 vs 54.8 ± 4.3</td>
</tr>
<tr>
<td></td>
<td>N-</td>
<td>53 vs 47</td>
<td>86.1 ± 3.9 vs 50.5 ± 6.2</td>
</tr>
<tr>
<td></td>
<td>N+</td>
<td>52 vs 48</td>
<td>89.0 ± 3.7 vs 59.1 ± 6.1</td>
</tr>
<tr>
<td></td>
<td>ER+</td>
<td>65 vs 35</td>
<td>88.8 ± 2.6 vs 54.2 ± 5.7</td>
</tr>
<tr>
<td>70-Gene Signature</td>
<td>All patients</td>
<td>39 vs 61</td>
<td>94.7 ± 2.1 vs 60.5 ± 3.8</td>
</tr>
<tr>
<td></td>
<td>N-</td>
<td>40 vs 60</td>
<td>93.4 ± 3.2 vs 56.2 ± 5.5</td>
</tr>
<tr>
<td></td>
<td>N+</td>
<td>38 vs 62</td>
<td>95.2 ± 2.6 vs 66.3 ± 5.2</td>
</tr>
<tr>
<td></td>
<td>ER+</td>
<td>50 vs 50</td>
<td>92.9 ± 2.4 vs 58.2 ± 4.7</td>
</tr>
<tr>
<td>Recurrence Score</td>
<td>All patients</td>
<td>35 vs 65</td>
<td>92.2 vs 58.5</td>
</tr>
<tr>
<td>Wound Response</td>
<td>All patients</td>
<td>23 vs 77</td>
<td>92.5 vs 63.6</td>
</tr>
</tbody>
</table>

4) The group of patients in the study is very diverse, resulting in a profile that is partially prognostic and partially predictive for a mixture of treatments. An sub-analysis would show the relevance of the 8-gene profile for decision of chemotherapy.

We have evaluated the performance of the 8-gene Score in patients receiving endocrine therapy and not chemotherapy. New Result and Discussion texts and a supplementary figure have been added to include this new information.
Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests
Reviewer’s 2 report:
This manuscript described an interesting study to define an 8-gene signature that can be used on FFPE samples using RT-PCR for hormonal receptor-positive breast cancer risk stratification. The advantage of this study is smaller number of genes used in the assay, FFPE samples used, RT-PCR assays with commercially available reagents.

- Minor Essential Revisions
1. One of the rationales for this study is to show that prognostic signatures can be performed in most laboratories using commercially available reagents. Good idea but difficult to implement. A diagnostic test is easier to implement in a central reference lab due to stringent quality control and quality assurance requirements. This is why both Mammaprint and Oncotype are offered in a central lab so that tests can be reproducibly performed. Point-of-care testing is only possible after stringent and regulated the production and validation of testing reagents and instrument. So the authors should make changes in their statement in the Introduction regarding "commercially available assays" into something like "commercially available assays with stringent diagnostics development".

The reviewer is right, so our statement about commercially available assays in the Introduction has been rephrased.

2. A major drawback of this study is there is no true independent validation of the 8-gene signature using the same RT-PCR assay on another FFPE cohort as acknowledged by authors in Discussion. Using datasets from other microarray expression studies can not be considered as true independent validation due to different assay format and sample types. Therefore, the performance shown in the study is likely an over-estimate of the true performance, and the comparison with other signatures can not be considered objective. The author should discuss this in Results and Discussion.

This issue is already commented in the body text in both Results” Although the use of other external databases does not constitute a formal validation, it may provide insight about the performance of the gene set." and Discussion “Furthermore, without an independent validation series, we cannot rule out overestimation of the prognostic value of the 8-gene Score. The application of our score to external data sets may temper this limitation, but it is not a formal validation. Without such validation, the score cannot facilitate treatment choices, as it neither distinguishes between therapy benefit nor identifies a group of patients that do not need any therapy”.

3. page 5 and 7, what is the poor correlation between fresh frozen and FFPE is used to discard genes?

RNA isolated from FFPE is considered a poor material for gene expression analysis owing to its extensive degradation. While microarray-based studies are highly sensitive to RNA degradation, qRT-PCR appears to be more robust and


We propose that the normalization factor should be able to control for the several levels of experimental variability in qRT-PCR, mainly the quantity of RNA and differences in enzymatic efficiencies, but also for differences in the quality of the starting material. We performed a previous work in which correlation in the quantitative, normalized data between 30 matched FF and FFPE samples was determined per gene using the Pearson correlation coefficient. We excluded from subsequent analyses genes showing non-significant correlation. Statistical significance was set at p<0.005.

When the study objective is to develop a predictive test to be implemented in the clinical setting, it is important to select the best set of genes. The number of genes that are actually informative is usually small compared to the number of genes that are noisy. Including too many noise genes can dilute the influence of the informative genes and reduce the accuracy of prediction, and makes interpretation and future use of the predictor more difficult. We consider that only the assays that are comparable between FF and FFPE tissue should be included in predictors, in order to guarantee reproducibility and reliability.

These results are explained in more detail in a manuscript called Comparison of Gene Expression Profiling by Quantitative Reverse Transcription PCR between Fresh Frozen and Formalin-Fixed, Paraffin Embedded Breast Cancer Tissues (I Sánchez-Navarro, A Gámez-Pozo, E Espinosa, A Pinto-Marín, D Hardisson, R López, R Madero, M Mendiola, M González-Barón and JA Fresno Vara), which is currently under consideration in Biotechniques.

4. page 7, what is the P value cutoff to select 17 genes from 53 genes?

We used a p<0.005. This information has been added in M&M.

5. The author mentioned leave-one-out in the Methods section for gene selection, is risk stratification (high- vs. low-risk) done in leave-one-out fashion? Are all the performance measures (log-rank P values, multivariate Cox analysis, Kaplan-Meier curves, etc) based on leave-one-out risk stratification or really just the plug-in performance (i.e. using the formula and cutoff point included in the Results section)? The author should clarify in the Results section. If latter, I would suggest they include the performance based on leave-one-out or other types of cross-validation.

As we stated in the manuscript, we used the BRB Array Tools (developed by R. Simon and A. Peng) software to develop the gene expression-based predictor. All

Regarding the cut-off point, we have included some text in M&M to clarify this issue: “The cut-off point was established prior to gene selection, validating the whole process using the LOOCV. We assessed cut-off points leaving from 10% to 90% of patients in the low risk group increasing a ten percent each time”. We have also rephrased this part of the text in Results: “We assessed cut-off points leaving from 10% to 90% of patients in the low risk group increasing a ten percent each time. Best performance was obtained when 60% of the patients were allocated to the low-risk group”.

The following text, extracted from BRB Array Tools Manual, shows the methodology we used:

“The Survival Analysis Tool for finding genes whose expression is correlated with survival time fits proportional hazards models relating survival to each gene, one gene at a time and computes the p value for each gene for testing the hypothesis that survival time is independent of the expression level for that gene. Gene lists are created based on these p values in the same way as in the Class Comparison tools. The p values can be used to construct gene lists using multivariate permutation tests for controlling the number or proportion of false discoveries. Or the gene list can simply consist of the genes with p values less than a specified threshold (0.001 is default). For more information regarding the multivariate permutation tests for controlling the number or proportion of false discoveries, please see the preceding section on Multivariate Permutation Tests for Controlling Number and Proportion of False Discoveries.

The Survival Analysis Prediction Tool develops a gene expression based predictor of survival risk group. The number of risk groups and the risk percentiles for defining the groups are specified by the user. The survival risk groups are constructed using the supervised principal component method of E Bair and R Tibshirani (2004):


This method uses a Cox proportional hazards model to relate survival time to k “super-gene” expression levels, where k is selectable by the user (usually 1-3). The “supergene” expression levels are the first k principal component linear combinations of expression levels of the subset of genes that are univariately correlated with survival. The user specifies the threshold significance level (e.g. 0.001) for selecting the genes to be used in computing the principal components. The significance of each gene is measured based on a univariate Cox proportional hazards regression of survival time versus the log expression level for the gene. After selecting the genes, the principal components are computed, and the k-
variable Cox proportional hazard regression analysis is performed. This provides a regression coefficient (weight) for each principal component.

Having developed a supervised principal component model as described above, to compute a prognostic index for a patient whose expression profile is described by a vector $x$ of log expression levels, the following steps are performed. First the components of the vector $x$ corresponding to the genes that were selected for use in computing the principal components are identified. Then the $k$ principal components are computed. These are just linear combinations of the components of $x$, with the weights of each linear combination having been determined from the principal component analysis described above. Finally, the weighted average of these $k$ principal component values is computed, using as weights the regression coefficients derived from the $k$-variable Cox regression described above. That computation provides a prognostic index for a patient with a log expression profile given by a vector $x$. A high value of the prognostic index corresponds to a high value of hazard of death, and consequently a relatively poor predicted survival.

In order to evaluate the predictive value of the method, Leave-One-Out-Cross-Validation is used. A single case is omitted and the entire procedure described above is performed to create a prognostic index. This function is created from scratch on the training set with the one case removed, including determining the genes to be included in the calculation of the principal components. Having determined a prognostic index function for that training set, it is used to compute a prognostic index for the omitted observation. That value is compared to the prognostic index for the $n$-1 cases included in that training set (assuming that there are $n$ distinct cases available in total). The prognostic index for the omitted patient is ranked relative to the prognostic index for the patients included in the cross-validated training set. The omitted patient is placed into a risk group based on his/her percentile ranking, the number of risk groups specified, and the cut-off percentiles specified for defining the risk groups. This analysis is repeated from scratch $n$ times, leaving out a different case each time.

Having completed the computations described in the previous paragraph, we plot Kaplan-Meier survival curves for the cases predicted to have above average risk and the cases predicted to have below average risk. It is important to note that the risk group for each case was determined based on a predictor that did not use that case in any way in its construction. Hence, the Kaplan-Meier curves are essentially unbiased and the separation between the curves gives a fair representation of the value of the expression profiles for predicting survival risk.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.