Author's response to reviews

Title: Molecular Risk Assessment of BIG 1-98 Participants by Expression Profiling using RNA from Archival Tissue

Authors:

Janine Antonov (janine.antonov@gmail.com)
Vlad Popovici (vlad.popovici@isb-sib.ch)
Mauro Delorenzi (Mauro.Delorenzi@isb-sib.ch)
Pratyaksha Wirapati (pratyaksha.wirapati@isb-sib.ch)
Anna Baltzer (anna.baltzer@dkf.unibe.ch)
Andrea Oberli (andrea.oberli@dkf.unibe.ch)
Beat Thürlimann (beat.thuerlimann@kssg.ch)
Anita Giobbie-Hurder (agiohur@jimmy.harvard.edu)
Giuseppe Viale (giuseppe.viale@ieo.it)
Hans Jörg Altermatt (altermatt@patholaenggasse.ch)
Stefan Aebi (stefan.aebi@insel.ch)
Rolf Jaggi (rolf.jaggi@dkf.unibe.ch)

Version: 2 Date: 30 November 2009

Author’s response to reviews: see over
Dear Dr. Le Good

Thank you for your reply and the four reviews regarding manuscript 1967450280304825 "Molecular Risk Assessment of BIG 1-98 Participants by Expression Profiling using RNA from Archival Tissue" by Antonov et al.

We thank the reviewers for their detailed and careful feedbacks and we discuss each comment point by point. The manuscript was revised according to their suggestions and we submit a new version of the manuscript. All the changes made in the manuscript are commented in our detailed feedback to the reviewers. This new version of the manuscript is clearer and we hope that it can be accepted for publication.

We would like to draw your attention to the specific criticism of Reviewer #3 (point 5): The postulated need for separate training and validation sets of tissue samples cannot be applied to our data as we used a tissue-independent method to develop the scores: the scores used in our study were developed prospectively from publicly available microarray data of breast cancer samples. The Big 1-98 samples were then used for validation only. This procedure is explained in the manuscript and we raise this important fact again in the discussion. Indeed, this was obvious to the other three reviewers who did not mention this issue in their reports.

We have explained the reasoning in great detail in our reply to the comments of Reviewer #3.

All the co-authors agree to this revised version of the manuscript and we look forward to hearing from you.

Kind regards

Rolf Jaggi
Reviewer 1

*The authors did not compare the two methods, but this should be done at least as a comment of the published results.*

We are well aware of the work done by the researchers in Pittsburgh and those in Genomic Health Inc. where the RNA analyses were done. A direct comparison cannot be done for two reasons:

**Tissue samples:** The data shown in the manuscript by Paik et al. (2006) is based on tissue samples of patients from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-20 trial. This material (and the same applies for other material that was analyzed by these groups) was not available to us and therefore, a direct comparison on the same material could not be performed. Instead, we received and analyzed tissues from patients of the International Breast Cancer Study Group (IBCSG) BIG 1-98 trial.

Both trials were done with tumor tissues from patients with primary, estrogen receptor positive breast cancer. B-20 patients had all node negative cancers (BIG 1-98 had both node negative and node positive cancers), B-20 patients received tamoxifen (Tam) or Tam plus either CMF or MF chemotherapy, whereas BIG 1-98 patients received Tam or letrozole (Let) with or without chemotherapy as a non-randomized option.

**Scores:** In our study we present data on estrogen receptor (ER), progesterone receptor (PGR), Her2, and PRO scores as well as the RISK score which is a combination of the four scores. RISK and Recurrence Score (RS) are similar as RS also represents a combination of the same categories of genes (and some additional genes). RS and RISK are similar but not identical as we mostly used other genes and a simple algorithm to combine expression values of several genes by assigning each gene in a subgroup the same weight. In RS the weight of several genes was different and the final RS was mathematically transformed to give a value between 0 and 100. We modified our manuscript to highlight some of the similarities: for example, we show that PRO_10 correlates to histological grades 1 to 3 (Fig. 2). This is similar to Paik et al. who compared RS to histological grade I-III (see Figures 1F and G in Paik et al., JCO 2006). Both, PRO and RISK scores were similar to the distribution of RS in Paik et al.

What is the reason to analyse more genes (than Paik et al)?

*Why did the authors create a new test using different genes, but belonging to the same group as the aforementioned kit?*

Several reports based on microarray data have revealed that prognosis and outcome of breast cancer patients can be estimated mainly from expression of genes related to proliferation, estrogen receptor and Her2. The genes used in our study were not selected based on their known biological function but instead from a large metaanalysis of microarray data by searching genes whose expression correlates with a prototype gene for proliferation (AURKA), estrogen receptor (ESR1 gene) and Her2 (ERBB2 gene). The genes on the microarrays were sorted and ranked according to their correlation to AURKA, ESR1 and ERBB2. We selected genes that were different from the genes in the RS because we wanted to document that the exact combination and also the number of genes is flexible. In addition, Genomic Health wants to protect their genes by a patent from being used by others for diagnosis and therefore we did not include genes from RS (with the exception of previously well known genes).

We also determined a PRO_5 score comprising the 5 proliferation genes present in RS. Significant scores were also measured with this smaller set of genes but PRO_5 had a lower prognostic power than PRO_10 (see page 9 and 12). Increasing the number of genes in each category appears to stabilize the resulting scores.

_Do they hope to analyse other populations of breast cancers, not only ER+ postmenopausal women?_ Recently, we started a feasibility test with patients who underwent surgery for breast cancer. Some of these patients were pre-menopausal and some were ER negative. Preliminary data suggest that the RISK_25 scores are generally high in ER negative patients as one would predict from the formula for the RISK score. However, it would be necessary to analyze more material from a clinical trial to make this observation reliable. As ER negative patients usually get a chemotherapy, molecular scoring would not alter treatment of these patients and therefore the score does not add essential information to guide therapeutic decisions.
Reviewer 2

The authors state in the paragraph titled ‘Conclusions’ that they have developed a novel method for isolation and demodification of RNA from FFPE tissues.

The formulation used in the ‘Conclusion’ section is indeed not appropriate, and it was re-written.

The ‘Methods’ section was also changed, it reads now:

The RNA was isolated from 4 sections (25 µm) of FF material and from 5-10 paraffin sections (10 µm thick) as described previously [26]. After demodification, the RNA was bound to silica-based columns, DNase I digested and eluted with water. The protocols and reagents for RNA isolation from FF and FFPE material were recently incorporated in commercial protocols (RNAready and FFPE RNAready, AmpTec, Hamburg, Germany).

• How much tissue was extracted for both FF and FFPE tissues?

  FF: 4 sections, 25 µm thick
  FFPE: 5-10 sections, 10 µm thick, the number of sections was adjusted to the size of the tissue in blocks

  The information was included in the Materials section (see previous point)

• How did the yields and quality from FF compare with FFPE?

  FF: RNA was extracted from frozen sections prepared from diagnostic tumor material. This material was shipped on ice to the pathology department where tissue was frozen and kryo sections prepared. On average 7 µg RNA were recovered depending on the size of the tumor material, the resulting RNA was of good although not excellent quality (RIN>6, as assessed on a Bioanalyzer).
  FFPE: Twelve µg RNA were isolated on average from FFPE material, again, the recovery depended on the size of the tumor in each block. The RNA from FFPE material was considerably degraded, the majority of RNA fragments was in the range of 200-500 nucleotides and bands of intact ribosomal RNA were usually not visible after separation on a Bioanalyzer. The RIN factor of this RNA was in the range of 2-3. The RIN factor does not seem to be very informative for this RNA and we used standardized qPCR reactions to estimate the quality of the RNAs. A detailed characterization between size of RNA, amplicon length and Ct values in qPCR was published previously (Oberli et al., 2008, ref [26]).

• Is there a reason why such a range of amplicon sizes (from 60 – 131bp) was chosen?

  The amplicons were selected from available Taq Man assays (Applied Biosystems). The size of the amplicons was one important parameter, but we also considered others. We have compared the impact of the amplicon size on the overall increase of Ct values in qPCR between RNA from FF and FFPE. Delta Ct values (Ct_{FFPE} – Ct_{FF} for same tumor) tends to increase with the size of the amplicons. The increase is mostly eliminated after normalization. This information may be of interest and we have included it as new supplementary Fig. 2.

• Can the authors confirm that there was no variation in reaction efficiency between FF and FFPE and disparate amplicon sizes?

  We defined quality standards for RNA from FFPE. They were based on a qualitative estimation of the electropherogram (Bioanalyzer) and on raw Ct values of three control genes in standardized qPCR reactions (5 ng RNA). The mean of the raw Ct values were determined for the three control genes (GUSB, RPLP0, UBB), RNAs for which the mean of the three Cts was >31 were excluded from further studies. The efficiency of each assay was tested on several intact RNAs from FF material and from several partially degraded RNAs from FFPE samples. The efficiency of reactions was calculated on serial dilutions of RNAs (using the One step RT-qPCR protocol). The efficiency was not affected by the quality of the RNA. The impact of amplicon sizes was discussed in the previous paragraph.

• Is the lack of a requirement for any pre-amplification step attributable to a large quantity of extracted total RNA or to an enhancement in harvesting due to some un-described modifications used for extraction?
As mentioned in the text (page 5) 7 RNAs were excluded from the analysis because the recovery from 10 sections was less than 1.5 µg. All the other samples revealed higher recoveries and therefore, pre-amplification was not necessary. Our experience with breast cancer samples was that RNA content is higher in cancer cells than in fat cells, the predominant type of tumor-surrounding cells. We tried to isolate RNA from fat cells using the same conditions but we failed to recover sufficient RNA for a RISK score analysis. Therefore, contaminating fat cells have little impact on RNA recovery from tumor tissue and on molecular scores. H&E stained sections were examined by an experience pathologist (H.J.A.) and 43 samples with less than 30% tumor cells were excluded from the analysis (9.8%). H&E sections from Big 1-98 were re-evaluated and about 30% of the samples contained 30% but less than 50% tumor cells and about 60% of samples contained 50-100% tumor cells.

- Were replicates included for each sample and if so how many?
We performed replicates (RNA isolated from independent sections of the same tissue block) from several FF and FFPE samples and the variability of scores was small as long as the tissue in the sections contained the same type of tumor. We also used tissue cores (regularly used for tissue microarrays) to compare histologically similar and histologically different areas of a single tumor (Schobesberger et al., 2008. BMC Cancer 8, 343). This reference was added (ref 46).

- How many technical replicates were included?
We performed replicates of qPCR reactions on TLDAs. The performance was very good and we observed Pearson correlation coefficients >0.95 for intact RNA and RNA from FFPE material. An example is shown below. The 342 samples of the Big 1-98 were not done in replicates. This is mentioned in the revised manuscript (page 5).

![TLDA replicates](image)

- Was a single control gene chosen for normalisation of all samples or were different ones used? If so please clarify which ones used which control gene.
We apologize for an error in the manuscript which says at one point that 5 control genes were used (page 5). The error was corrected. In fact we had included 5 control genes on the TLDA card but normalization was done with 3 (GUSB, RPLP0, UBB). We recapitulated the calculations with 5 control genes (GAPDH, GUSB, RPLP0, TFRC, UBB) which led to small changes but did not affect any of the results presented in this manuscript.

- Samples with a mean Ct of 31 were excluded from further analysis. Was this normalised against a single control gene or several? What was the range of Cts in such samples?
Samples with a mean Ct > 31 were excluded. qPCR was performed with 5 ng RNA for each of the 3 control genes. If the mean Ct of the 3 control genes (GUSB, RPLP0, UBB) was >31 (using the raw Ct values for the three genes) the RNA was excluded from further analyses. This was the case when RNA was of poor quality (e.g. strongly degraded or otherwise corrupted). This test was performed in regular qPCR reactions before applying the RNAs to TLDA plates to avoid failures on the cards. Thirty five RNAs were excluded from further analysis due to poor quality of RNA.
Although Supplemental Table 1 lists gene identifiers, it was difficult to determine the biological rationale for selecting the genes, rationale for determining the parameters of the algorithm and the rationale for selection of the algorithm itself.

Independent sets of genes were selected from a number of publicly available microarray data sets using a meta-analytical approach [27]. The genes either correlated with a prototype gene (AURKA for proliferation, ESR1 for ER-related genes, PGR for progesterone-related genes and ERBB2 for Her2-related genes) or they correlated with disease-free survival. Importantly, the genes were selected prior to the validation presented in this study. It was a logical consequence of the procedure that the genes also represented ER and PGR functions as well as Her2 and proliferation but the selection was not based on this biological information.

The authors have not adequately discussed the sources of bias in their study. Specifically, tumor size is known to relate to aggressiveness. Moreover, as shown in figure 2 the subdivisions of the histological grading, which are determined by the subjective scores by the pathologists, have a better resolution than the PRO_10 score. This raises the general question as to what value the molecular score has over the standard histological grading.

It is well known that clinical and pathological parameters including tumor size are related to prognosis, and we have analyzed and discussed them and summarized the clinical information in Table 2 (originally Table 1) and in Figure 2 of the manuscript. Further, we document that tumor size and histological grade are important parameters for patients outcome. However, multivariate proportional hazards regression analyses, so-called Cox models of disease-free survival, including the molecular parameters (PRO_10, PGR_5 and RISK_25) revealed that histological grade, tumor size and Ki-67 labeling were not independent prognosticators of disease-free survival and were indeed superseded by the molecular parameters (Table 3 of the revised manuscript = Table 2 of the first submission). These results suggest that molecular scores are more powerful predictors of prognosis than certain clinical parameters such as tumor diameter.

In addition, low and high RISK_25 scores allowed discriminating between good and poor prognosis in tumors with histological grade II. These tumors represent a heterogeneous group of cancers. This result may be clinically relevant as a reliable prediction of prognosis on the basis of clinical and/or pathological parameters is not currently possible for these tumors. The data presented in this manuscript document that molecular parameters outperformed clinical and pathological parameters.

The samples were not microdissected and were biased against tumors with a low percentage of cells because tumors with less than 30% tumor were removed from analysis (page 5). In routine pathologic settings, typical core needle biopsies contain less than 30% tumor. This means that an entire class of tumors was not available for scoring. The percentage of cells in the biopsy is entirely dependent on the biopsy sampling technique and not reflective of the tumor size. A biopsy sample could be obtained at the tumor perimeter, or center, resulting in samples with vastly different tumor cell percentages.

It is correct that our tissue samples were not microdissected, and we excluded all the samples containing less than 30% tumor cells. 43 samples had to be excluded from the analysis for this reason (Methods section on page 5). All the patients from BIG 1-98 were treated by mastectomy or by breast conserving surgery. Therefore, the tissues analyzed in this study originated from this surgical material and not from core needle biopsies. The blocks that were available were derived from representative tumor regions. In the Methods section we describe that we started with 437 tumors, and we give exact numbers and reasons why tumor samples were excluded. At the end, 342 tumors were analyzed on TLDA cards. Baseline characteristics of these tumors are summarized in Table 2 (originally Table 1). No sample was excluded later in the process, e.g. after qPCR data was generated.

The authors do not show that their gene score reveals any additional information that is not demonstrable through histopathological grading (Figure 2). The PRO_10 score only changes one unit across the spectrum of histopathological grading shown in Figure 2. The mean values fall within a very narrow range (14.0 – 15.0 units).
We respectfully disagree with the reviewer in this point. Multivariate proportional hazard regression analyses (cf. Table 3 of the revised manuscript = Table 2 of the original submission) revealed that the molecular scores PRO_10 and RISK_25 outperformed histopathological parameters (grading, tumor size, Ki-67 labeling) as described above. In addition, the scores allow classifying histological grade II tumors into low and high risk groups (Fig. 3C). The discrimination is statistically significant even in this fairly limited group of 195 grade II tumors (P=0.0024). In addition, we document that the lymph node status remains a very important prognostic factors in multivariate analyses. The recurrence score, RS, is scaled to a value between 0 and 100. This range results from a mathematical transformation of qPCR values but this does not improve the accuracy of the original data.

Taken together, we present evidence that molecular parameters are powerful tools which can improve the prediction of prognosis for patients with primary breast cancer. We performed a number of statistical tests to document the statistical power of these molecular parameters (Fig. 3, Table 3).

The most serious deficiency in the analysis is the lack of a test set to develop the molecular score and an independent sample set to validate the score. The models were not validated with an independent, blinded validation set. Bootstrapping will not overcome biases built into the tissue collection methods and cellular heterogeneity.

We respectfully disagree with the reviewer. While in the typical context of biomarker development the training set and the validation set must be independent to avoid incorrect overly optimistic conclusions about the accuracy of the prediction, this does not apply to the present report. We would like to emphasize that the genes used in this study were prospectively selected from publicly available microarray data and the scores were defined by giving equal weight to each gene in the four groups (proliferation, estrogen response, progesterone response, Her2 response). Thus, a training set was not used at all as the scores are based purely on “in silico” selection of genes. Therefore, the BIG 1-98 tissue samples reported here are an independent test set that can be used to validate our scores.

Obviously, if a complex algorithm had been developed by individualizing the weight of each gene in the scores such as used in OncotypeDx, a new and independent dataset would need to be available for validation. However, as outlined above, this does not apply to the present report.

Minor Essential Revisions

The table of the gene identifiers (Supplemental Table 1) should be included in the main body of the paper. This is essential information for the reader to understand which genes were used to develop the molecular scores.

We have moved the Supplementary Table 1 into the body of the manuscript and termed it Table 1. Original Table 1 and 2 became Table 2 and 3, respectively.

The paper suffers from a lack of focus. Initially the paper appears to focus on compatibility of archival FFPE tissue for generating molecular scores but the remainder of the paper discusses generation of a molecular score for predicting DFS. The authors may wish to clearly delineate the two studies.

Several papers have been published which used archival material for RNA expression studies. However, the data of several well-known and important studies were generated in collaboration with Genomic Health and the experimental part (RNA isolation, expression measurement) was carried out at Genomic Health. Here, we present a procedure which is independent from the methods used by Genomic Health. We strongly believe that it is important to provide the reader some original data which documents the quality of the raw and normalized data and the scores subsequently used for the samples of the clinical trial. In fact it may be of interest to document that no outliers were observed among the 82 pairs of samples.

Focus 1: By showing Figure 1 in the manuscript and suppl. Figures 2-4 we give the reader an opportunity to see some basic data. Focus 2 is closely related to focus 1 as it is a direct application of the technology to clinically relevant samples for which clinical data and survival data are available. We think that figure 1 contains important information without overloading the manuscript with technological aspects. The reader is free to consult the supplementary information or not.
There is a discrepancy on page 5 regarding the number of gene controls utilized (3 or 5). Please refer to the following sentences: “RNAs were tested by quantitative reverse transcription PCR (qRT-PCR) with five control genes leading to the further exclusion of 35 samples (8%)…” “For the remaining 342 RNAs… the expression of 34 genes (3 control genes and 31 genes for computing scores…” Which 5 control genes were used to exclude samples and why were 5 genes used to exclude data, yet 3 genes were used to develop the molecular score?

We thank the reviewer for this error which was corrected. Explanation: we included 5 control genes on the TLDAs. Analysis of the data revealed that 3 control genes were very similar to five genes. Therefore, all the data shown in the paper were based on 3 control genes (GUSB, RPLP0, UBB). In fact, we compared the results of Figure 1 (matched samples) when normalization was based on 3 or on 5 genes. Essentially the same results and correlations were obtained.

Figure 2 legend does not indicate what is shown in the box and whiskers plot, i.e. mean ±2SD or some other statistical data.
The text in the legend was revised

Figure 3 is difficult to interpret and would benefit from additional labeling for each Kaplan-Meier plot as well as additional explanations of each plot in the figure legend. Specifically, to each figure please add the molecular score category and the patient category or group used in each plot.
The panels were labeled and the legend was revised

Figure 4B – the reader needs to understand why the PGR_5 score has an inverse relation to the number of events, compared to the positive correlation for the PRO_10 and RISK_25 scores. Additional text describing the biology of the score associations would be very beneficial.
The results obtained with PGR_5 are discussed in relation to PgR as measured by IHC (Viale et al, ref. 29) and by microarray studies (Loi et al, ref. 52) We explicitly point to these results in the Discussion and we compare it to PRO_10 and RISK_25 scores.

In general the authors did not provide adequate rationale regarding choice of genes, algorithms, or relevance to breast cancer biology. The reader would greatly benefit from enhanced discussion of these topics.

Previous studies by others have shown that prognosis in breast cancer is dependent on the expression of genes involved in the regulation of proliferation. As described previously, the PRO_10 genes were selected through a bioinformatics approach. Biological aspects and pathways were not considered but obviously, many of the genes identified turned out to be related to cell cycle and proliferation. Similarly, genes related to estrogen receptor, progesterone receptor and Her2 were found as a consequence of the bioinformatics approach.

Additional discussion regarding effects of tissue heterogeneity should be discussed because these tumor samples were not microdissected. We can not know which cell population (tumor or stoma) contributed which percentage of signal to the molecular score.
This is an interesting aspect. We extended the "Discussion" by adding some text (page 12)

9. Page 11 – the authors compare their molecular score to OncotypeDX recurrence score and claim that their RISK_25 score can predict DFS. This statement is unwarranted unless the authors use their score on a blinded validation sample set, or on the original samples used to generate the OncotypeDX recurrence score.
This comment is related to the last major comment (see above). We are concerned about this objection but we are convinced that it is not appropriate and does not apply to the data and procedures applied in this manuscript. The samples from the Big 1-98 trial were not used for developing the scores but they were exclusively used to validate them. Therefore, the objection of the reviewer does not apply here. The original samples used by Paik et al. are derived of a clinical trial, they are not available and cannot be used.

Discretionary Revisions
1. Additional information regarding the percentage of tumor cells would be appreciated. For example, a list of the number of tumors with >90% tumor, number with 75% tumor, number with 50% tumor, etc.

Histological sections of each tumor sample were assessed by an experienced pathologist (H.J.A.) and samples containing less than 30% tumor cells were excluded from the study. Following the reviewer's suggestion, the remaining samples were re-assessed: about 30% of sections contained 30% but less than 50% tumor cells and about 60% of tumors contained 50-100% tumor cells. We also tried to isolate RNA from fatty tissue of tumor-free mammary gland material but we failed to get sufficient RNA to measure molecular scores. Therefore, fat cells in tissue lysates affect molecular scores only marginally.

2. The conclusions are very general and are not supported with specific data or confidence intervals or p values. These values are listed in the results section but the conclusions would appear stronger if they were supported statistically.

Adding numbers and p-values from the results in the "Discussion" and "Conclusion" may lead at least in part to a recapitulation of the "Results", or data in Figures and Tables. Instead, we tried to discuss some further aspects of the results, for example:

- we discuss some quality issues of our work which consisted of the comparison between intact RNA from FF material and partially degraded RNA from FFPE;
- we compare our results with similar data published in the literature in order to consolidate some of our conclusions: for example, the number of ER- or proliferation-related genes does not seem to be very critical when establishing a molecular score;
- we compare scores with immunohistochemical markers;
- we emphasize the potential role of molecular scores in personalized medicine.
Reviewer 4

…The authors should give the range in the mean Ct shifts for the individual assays and should explain whether any of the assays was too unreliable or showed a too high Ct shift in paraffin. All assays on the TLDA card could be measured in FF and FFPE material. Some assays were not used in the context of this manuscript. We observed a week positive correlation between amplicon size and delta Ct values (long amplicons tend to perform less well with RNA from FFPE than short amplicons). We include an additional Figure (suppl. Fig. 2A and 2B) to document this. After normalization, delta Ct values were close to 0 and the effect of amplicon size was no longer apparent (2B).