A prognostic signature based on three-genes expression in triple-negative breast tumours with residual disease

Joseph A Pinto, Jhajaira Araujo, Nadezhda K Cardenas, Zaida Morante, Franco Doimi, Tatiana Vidaurre, Justin M Balko and Henry L. Gomez

Residual disease after neoadjuvant chemotherapy (NAC) in triple-negative breast cancer (TNBC) is related with poor prognosis; however, the risk of recurrence after 3 years from surgery, becomes similar to other breast cancer subtypes indicating that TNBC is composed of tumours of different prognosis. To evaluate genes related to TNBC aggressiveness in the outcome of TNBC resistant to NAC, we profiled 82 samples of residual tumours whose expression for 449 genes was quantified with NanoString. The validation set (GSE25066) consisted of 113 TNBC cases with residual disease. The stepwise multivariate survival analysis performed by the Cox proportional hazards model selected CCL5, DDIT4 and POLR1C as independent prognostic factors for distant recurrence-free survival (DRFS). We developed a three-genes signature using the regression coefficients for each gene (−0.393 × CCL5 + 0.443 × DDIT4 + 0.490 × POLR1C). The median score in the discovery set (0.1494) identified two subgroups with different DRFS (P < 0.001). The median score in the validation set was 0.0024 and was able to discriminate patients with different DRFS (P = 0.002). In addition, the three-genes signature was a prognostic factor in TNBC patients regardless their response to NAC (data set GSE58812; P = 0.001) and in patients with oestrogen-receptor-negative tumours (data set GSE16446; P = 0.041). Here we describe a prognostic signature based on expression levels of CCL5, DDIT4 and POLR1C. The knowledge about the involvement of these genes in chemotherapy resistance could improve the therapeutic strategies in TNBC.

Nowadays, there are commercially available tests based on levels of gene expression; such as, the 21-genes recurrence score (OncotypeDx) and the 70-genes signature (MammoPrint). Despite the clinical utility, these tests have certain limitations. The 21-genes score (16 cancer-related and 5 control genes) calculates a recurrence score between 0 and 100, where the risk of recurrence increase with the score and estimate the likelihood of benefit with the adjuvant chemotherapy. This test is recommended by NCCN guidelines only in patients diagnosed with pT1–pT3 and pN0 and pN1mi ER+/HER2− breast tumours. On the other hand, the 70-genes signature calculates a score to assign a risk group (high or low risk) and estimate the probability of 10 years recurrence from diagnosis. The 70-genes score (although not recommended by NCCN guidelines) has FDA approval for luminal patients with stage I or II, with ≤ 3 lymph nodes involved and invasive tumours < 5 cm. Despite these molecular platforms have shown clinical utility, these genomic predictors are not useful in TNBC and new markers and predictors are needed in order to a improve the risk stratifications and therapeutic strategies.

The aim of our study was to evaluate genes related to TNBC aggressiveness to elaborate a gene signature associated with prognosis in terms of DRFS in TNBC resistant to NAC.

RESULTS
Selection of genes related with DRFS
Workflow is shown in Figure 1. Overall, median DRFS in the discovery set was 22.3 months. Univariate Cox regression of 449
A three-genes signature for triple-negative breast cancer

JA Pinto et al

Figure 1. Overview of research design and workflow.

Three-genes prognostic score
Stepwise multivariate survival analysis performed by the Cox proportional hazards model of the seven genes from the previous step, selected three genes as independent prognostic factors: CCL5 (Chemokine (CC motif) ligand 5), DDIT4 (DNA-damage-inducible genes transcript 4) and POLR1C (Polymerase (RNA I) polypeptide C, 30 Kd1a), with P-values of 0.002, 0.005 and 0.004, respectively. CCL5 had protective effect (hazard ratio (HR) < 1), whereas DDIT4 and POLR1C were associated to a worse outcome (HR > 1; Table 1).

The three-genes prognostic score results from the sum of multiplication of normalised expression levels of the genes by their respective regression coefficient as described in the following formula:

$$\text{Three-genes prognostic score} = (-0.393 \times \text{CCL5} + 0.443 \times \text{DDIT4} + 0.490 \times \text{POLR1C})$$

The median score in the discovery set was 0.1494. Using this median as cutoff, was possible to establish two risk groups.

Individual value of CCL5, DDIT4 and POLR1C in TNBC
We evaluated CCL5, DDIT4 and POLR1C in the Kaplan–Meier Plotter online platform (kmplot.org) in order to analyse their influence in the recurrence-free survival (RFS) of TNBC patients. When group of patients was split into two groups according to median of expression, CCL5 and DDIT4 were associated with RFS (P = 0.0012 and P = 0.00034, respectively). POLR1C was not statistically associated with RFS (P = 0.28); however, when systematically untreated patients were removed from the cohort, high expression of POLR1C was associated with a poor prognosis (P = 0.0059; Figure 2).

Prognostic value of the three-genes signature
Evaluated the score as a continuous variable in the CoxPH analysis, a HR = 2.72 for each unit change (P < 0.001; 95% CI: 1.72–4.28) was estimated. The risk groups established by the median cutoff shown statistically significant differences in the DRFS, with a median survival of 39.6 months for the low-risk group and 15.5 months for the high-risk group (P < 0.001; Figure 3a).

Validation cohort
Validation set (GSE25066). The median DRFS in this group of patients was 35.4 months. In this group, the median of the score was 0.0024 and was able to identify two subgroups with different outcomes where the median DRFS for the low-risk group was not reached and the median DRFS for the high-risk group was 22 months (P = 0.002) (Figure 3b).

Analysis of other data sets
Patients with TNBC regardless of their response to NAC (GSE58812). The end point evaluated in this data set was metastases-free survival (MFS). The median score was 0.1009. The median of MFS was not reached and 4-year MFS rates were 83.5% vs 57.3% for low-risk vs high-risk groups according the three-genes signature (P = 0.001; Figure 3c).

Patients with oestrogen negative tumours (GSE16446). The median DRFS was 61.1 months. The median score was 0.0137 and was able to classify into two subgroups with different prognosis, where the median DRFS for the low-risk group was not reached and the median DRFS for the high-risk group was 47.1 months (P = 0.041; Figure 3d).

Multivariate analysis
In the discovery set, in addition to three-genes prognostic signature, age, lymph node status were statistically related to the DRFS in the univariate analysis. In the multivariate analysis, only > 3 involved nodes (HR = 3.98, 95% CI: 1.73–9.12) and a poor three-genes prognostic signature (HR = 2.03, 95% CI: 1.02–4.05) were independent factors associated with shorter DRFS (Table 2).

In the validation set, in addition to the three-genes prognostic signature, the T-stage, nodal stage and clinical AJCC stage were statistically related to DRFS. In the multivariate analysis, T3 stage (HR = 1.878, 95% CI: 1.054–3.346), nodal stage 2–3 (HR = 2.78, 95% CI: 1.271–6.079) and the high-score group (HR = 2.358, 95% CI: 1.359–4.091) were independent poor prognostic factors (Table 3).

DISCUSSION
Triple-negative breast cancer is a classification obtained by exclusion criteria rather than representing a well-defined entity such as other breast cancer subtypes. The pCR after NAC is the

Table 1. Genes selected as independent prognostic factors in the stepwise multivariate survival analysis performed by the Cox proportional hazards model in the discovery set

| Gene    | Discovery set | P-value | Regression coefficient |
|---------|--------------|---------|------------------------|
| CCL5    | 0.68 (0.53–0.86) | 0.002 | −0.393 |
| DDIT4   | 1.56 (1.14–0.76) | 0.005 | 0.443 |
| POLR1C  | 1.63 (1.17–2.28) | 0.004 | 0.490 |

Abbreviations: CI, confidence interval; HR, hazard ratio. Values of hazard ratios and regression coefficients are shown.
main indicator for good prognosis. In Peru, a study by Neciosup et al.\textsuperscript{17} reported pCR rates of 9\%, lower frequency than those reported in other series for TNBC ($\approx$ 22\% and up to 36\% with addition of bevacizumab in the neoadjuvant setting).\textsuperscript{10,17,18}

In the last decade, there has been growing interest to unveil the molecular biology of TNBC leading to identification of six TNBC subtypes with distinct biology and prognosis, where our research group could identify loss of DUSP4 expression as mechanism of resistance to NAC and 90\% of TNBC have an actionable mutation candidates to be treated with targeted drugs.\textsuperscript{19,20}

In our study we found three genes independently related with the outcome in several TNBC data sets. In addition, these genes can be combined in a linear score. Unfortunately, there is unavailability of certain clinical data in public data sets, including time-to-event data, leading to an important limitation in its use. Lack of important information in public data sets despite journal requirements has been previously addressed.\textsuperscript{21,22} To overcome this issue, we selected patients with residual disease from data sets profiled with a different platform (affymetrix microarrays) and with samples taken prior NAC for our validation cohort. Although it is expected that gene expression patterns could change after NAC, our three-genes signature was able to identify groups with different prognosis evaluating samples profiled before NAC. On the other hand, CCLS, DDIT4 and POLR1C in samples profiled after treatment in large data sets of TNBC patients (evaluated in the online platform kmplot.org) shown a significant association of gene expression levels with the RFS (Figure 2).

The gene CCLS is located on 17q, and is part of a superfamily of secreted proteins involved in immunoregulatory and inflammatory processes. TCGA (The Cancer Genome Atlas) data show that 4\% of breast tumours have CCLS dysregulation (2\% in basal tumours), mainly related to genetic downregulation which has been previously associated with breast cancer progression.\textsuperscript{22} CCLS expressed by tumours recruits tumour infiltrant lymphocytes (TILs). TILs has an important role in the outcome of TNBC where $> 20\%$ of TILs are associated with an better prognosis, while lower proportion of TILs ($< 10\%$) is associated with genetic or transcriptomic alterations in Ras/MAPK pathway.\textsuperscript{23} On the other hand, CCLS was shown to be a mechanism of tumour escape in a

---

**Figure 2.** Individual impact of CCLS, DDIT4 and POLR1C genes recurrence-free survival (RFS) in triple-negative breast cancer patients (using the median of expression as cutoff) in the database of the online platform kmplot.org. (a) High expression of CCLS was associated with good prognosis ($P = 0.0012$). (b) High expression of DDIT4 confers poor prognosis ($P = 0.00034$). (c) Expression of POLR1C was not significantly associated with the RFS in all patients ($P = 0.28$); however, (d) excluding systematically untreated patients, a significant association is observed ($0.0059$).
mouse model of colorectal cancer, recruiting and improving the cytotoxic effects of T-regulatory cells against CD8+ T cells. The DDIT4 gene (DNA-damage-inducible transcript 4) encodes a protein related to adverse environmental conditions, whose action is the inhibition of mTOR. Despite the biological function of DDIT4, in our analysis this gene was related with tumour aggressiveness with an HR = 1.56 (\(P = 0.005\)) by each unit of change (Table 1). A recent report by Puissant et al. describe that the product of DDIT4 could be cleaved by caspase 3 modifying its cellular function exerting anti-proliferative activities. TCGA data show that dysregulation of this gene in 4% of breast tumours, mainly overexpression (altered in 12% of tumours basal subtype).

The POLR1C gene (30 kDa), encodes a subunit of the RNA polymerase enzymes I and III, responsible for RNA synthesis. The mutation in this gene causes Treacher Collins syndrome that affects the development of bones and facial tissues. POLR1C gene is recurrently amplified and overexpressed in gastric cancer. TCGA data indicate that 11% of breast cancer cases have dysregulations in this gene (overexpression and down-regulation). This gene is dysregulated in 21% of cases of basal tumours, where changes include overexpression. The molecular mechanisms of its involvement in tumour aggressiveness remain unclear.

In spite that this cohort of TNBC patients was evaluated previously and other genes were found to be related significantly with the outcome, such as a signature of MEK pathway activation and DUSP4 loss (corroborated in vivo and in vitro experiments), in the analysis we done in this work found other genes not previously reported. It is important to explore biomarkers under different strategies, because weak biological signals or interaction between markers could be ignored in some bioinformatics or statistical analysis, for example, when is used as multiple testing procedures.

Owing to the TNBC heterogeneity, some cases of this phenotype could correspond to luminal A, luminal B or HER2-enriched subtypes. In TNBC, the PAM50 test has a particular utility. PAM50 evaluates 50 genes required for determining intrinsic molecular subtypes of breast cancer. Because immunohistochemistry has limited ability to detect protein expression, PAM50 becomes a test that can identify more accurately the molecular subtypes within the TNBC phenotype, providing additional information for deciding a better therapeutic strategy.

Our prognostic signature, unlike other commercially available platforms, is based only in three-genes expression. This finding could lead to the development of a cheap molecular test based in reverse transcription PCR (RT-PCR) and suitable in the clinical routine. The different median values of the score obtained in each data set could be explained by the different prognostic factor between them and/or differences between microarray platforms. On the other hand, our three-genes prognostic score produced...
Table 2. Univariate and multivariate analysis of patient and tumour characteristics related to DRFS in the discovery set

| Discovery cohort | n (events) | Univariate analysis | Multivariate analysis |
|------------------|------------|---------------------|-----------------------|
|                   |            | HR (95% CI)         | P-value | HR (95% CI) | P-value |
| Age              |            |                     |         |             |         |
| ≤ 45             | 30 (21)    | 1                   | 0.010   | 0.424       |
| > 45             | 49 (23)    | 0.46 (0.25–0.83)    | 0.68 (0.26–1.75)     |
| Menopausal status|            |                     |         |             |         |
| Premenopausal    | 36 (24)    | 1                   | 0.044   | 0.544       |
| Postmenopausal   | 43 (20)    | 0.54 (0.30–0.98)    | 0.75 (0.30–1.90)     |
| Stage            |            |                     |         |             |         |
| IIA, IIB         | 6 (5)      | 1                   | 0.127   |             |         |
| IIIA, IIIB, IIIC | 73 (39)    | 0.48 (0.19–1.23)    |         |             |         |
| Involved lymph nodes |       |                     |         |             |         |
| 0                | 25 (9)     | 1                   | 0.003   | 0.026       |
| 1–3              | 31 (18)    | 1.66 (0.75–3.70)    | 0.213   | 1.37 (0.60–3.14) | 0.450 |
| > 3              | 22 (16)    | 3.98 (1.73–9.12)    | 0.001   | 3.98 (1.73–9.12) | 0.014 |
| Ki-67 score      |            |                     |         |             |         |
| ≤ 14             | 15 (10)    | 1                   | 0.192   |             |         |
| > 14             | 64 (34)    | 0.62 (0.31–1.27)    |         |             |         |
| Prognostic signature |       |                     |         |             |         |
| Low risk         | 40 (17)    | 1                   | 0.001   | 0.044       |
| High risk        | 39 (27)    | 2.86 (1.52–5.36)    | 2.03 (1.02–4.05)     |

Abbreviations: CI, confidence interval; HR, hazard ratio. Bold entries denote statistically significant values.

Table 3. Univariate and multivariate analysis of patient and tumour characteristics related to DRFS in the validation set 1 (GSE25066)

| GSE25066 validation set | n (events) | Univariate analysis | Multivariate analysis |
|--------------------------|------------|---------------------|-----------------------|
|                          |            | HR (95% CI)         | P-value | HR (95% CI) | P-value |
| Age                      |            |                     |         |             |         |
| ≤ 45                     | 44 (20)    | 1                   | 0.656   | 0.032       |
| > 45                     | 69 (35)    | 0.656 (0.508–1.532) |         |             |         |
| Clinical T stage         |            |                     |         |             |         |
| 0–2                      | 52 (18)    | 1                   | 0.006   | 0.032       |
| 3–4                      | 61 (37)    | 2.172 (1.232–3.831) | 1.878 (1.054–3.346) |
| Clinical nodal stage     |            |                     |         |             |         |
| 0                         | 26 (9)     | 1                   | 0.010   | 0.025       |
| 1                         | 52 (23)    | 1.706 (0.785–3.711) | 1.609 (0.740–3.499) |
| 2–3                      | 35 (23)    | 3.109 (1.429–6.767) | 0.004   | 2.779 (1.271–6.079) |
| Clinical AJCC stage      |            |                     |         |             |         |
| I–IIIB                   | 51 (18)    | 1                   | 0.004   |             |         |
| IIIA, IIIB, IIIC, Inflammatory | 62 (37) | 2.3 (1.306–4.048) |         |             |         |
| Histological grade       |            |                     |         |             |         |
| 2                         | 17 (9)     | 1                   | 0.542   |             |         |
| 3                         | 86 (38)    | 0.798 (0.386–1.651) |         |             |         |
| Prognostic signature     |            |                     |         |             |         |
| Low risk                 | 57 (20)    | 1                   | 0.002   | 0.003       |
| High risk                | 56 (35)    | 2.358 (1.359–4.091) | 2.308 (1.326–4.018) |

Abbreviations: CI, confidence interval; HR, hazard ratio. Bold entries denote statistically significant values.
robust results and could be evaluated in both biopsies and residual tumours after NAC. In conclusion, our analysis of 449 genes related to aggressive molecular signatures identified three genes (CCL5, DDIS4 and POLR1C) that were independent prognostic factors whose combination resulted in a predictor able to identify TNBC patients with different outcome. These data encourage the prospective clinical validation of these genes using the technology of real-time PCR. The products of CCL5, DDIS4 and POLR1C genes could be used for develop new therapeutic strategies in TNBC and basal breast cancer.

MATERIALS AND METHODS

Discovery set

We evaluated 114 patients with TNBC who had residual tumour after NAC (diagnosed and treated at the Instituto Nacional de Enfermedades Neoplásicas, Lima, Peru). Cases with insufficient tissue or cellularity (n = 5) or HER2-amplified (n = 7) were excluded. In total, 82 cases were evaluable for analysis, Formalin-fixed paraffin-embedded residual tumour from surgical specimens were serially cut into 3- to 5-μm thick, then cellularity evaluation and nucleic acid extraction was done. RNA was extracted and purified using the RNEasy FFPE kits (Qiagen GmbH, Hilden, Germany).

Genes selected for evaluation

Overall, evaluation of 449 transcripts were selected based on their inclusion in published gene expression signatures: the PAM50 genes, signatures associated with DUSP4 loss, with MEK activation, with the enrichment of TGFβ inducible genes after NAC or signatures based on their association with the post-NAC Ki-67 score.9,34,37,38

NanoString nCounter analysis

NanoString nCounter technique captures and counts individual messenger RNA transcripts using unique pairs of capture and reporter probes with a colour code generated by ordered fluorescent segments specific for each transcript. As enzymatic reactions are not used, there is no bias or decrease of sensitivity compared with other techniques, such as microarrays.9,89 RNA extracted from formalin-fixed paraffin-embedded residual tumour was run on the nCounter Analysis System (NanoString Technologies, Seattle, WA, USA), according to the manufacturer’s protocol. Code sets were synthesised targeting 499 genes.

Validation set

GSE525066. We selected 113 TNBC (determined by immunohistochemistry) with residual disease after NAC. Gene profiling was done with U133A Affymetrix microarray platform (Affymetrix U133A chip, Santa Clara, CA, USA) in biopsies before NAC. List microarray probes evaluated in this data set are shown in Supplementary Table S1.

Analysis of other data sets

GSE58812. Composed by 107 TNBC patients regardless their response to neoadjuvant treatment (these data were not included in the data set). U133A Affymetrix microarray platform in biopsies before NAC. List of microarray probes evaluated in this data set are shown in Supplementary Table S1.

GSE16446. All patients in this data set were ER negative. We selected 47 patients with HER2 not-amplified tumours. Gene profiling with the U133 Plus 2.0 Affymetrix array was done in biopsies before NAC. List of microarray probes evaluated in this data set are shown in Supplementary Table S2.

Gene expression data preprocessing

In the discovery set, expression levels were normalised with Spike-controls, log2 transformed and median centred. In the data sets GSE25066, GSE58812 and GSE16446, probes for the same gene were collapsed to the higher value, then the expression data were log2 transformed and median centred. Values of normalised data values are shown in Supplementary Tables S3–S6.

Prognostic signature

The expression of 449 genes was evaluated to select genes strongly associated to DRFS (P < 0.05) using Cox regression models. Genes significantly associated with DRFS were further tested using Cox proportional hazards regression with the stepwise method of selection, identifying those genes that were independent prognostic factors. We used a linear combination of the normalised values of gene expression levels multiplied by a weighting value for each gene (regression coefficients) to calculate a risk score for each patient. The proportional hazards assumption over time for the final Cox model for the dichotomised risk score was tested graphically using log–log survival functions, and the assumption of appropriateness in discovery and validation sets was confirmed (Supplementary Figure S1).

Survival analysis

DRFS was estimated by the Kaplan–Meier method. The log-rank test was used as the method of statistical inference. After calculating the risk score, the median was estimated. Using the median (specific to each group of patients) as a cutoff, two risk subgroups were established. Survival curves were compared using the log-rank test. A P value < 0.05 was considered statistically significant.

Ethical considerations

This study involves a reanalysis of gene expression and clinical data obtained in a previous study that was approved by the IRB from the Instituto Nacional de Enfermedades Neoplásicas (INEN 10–018).
