# Subtype specific metagene-based prediction of outcome after neoadjuvant and adjuvant treatment in breast cancer

Maurizio Callari, Vera Cappelletti, Francesca D'Aiuto, Valeria Musella, Antonio Lembo, Fabien Petel, Thomas Karn, Takayuki Iwamoto, Paolo Provero, Maria Grazia Daidone*, Luca Gianni, Giampaolo Bianchini

## Supplementary Information

## 1. Data collection and clinico-pathological features of employed case series

All downloaded datasets were checked to verify that only profiles from primary invasive breast cancer receiving no pre-surgery treatment were included. Samples not satisfying such criteria, cell line and normal tissue profiles were removed. Data were also checked for the presence of replicated CEL files uploaded with different GSM ID and redundancy was removed.

Datasets were grouped in four main collections (GENERIC, PROGNOSTIC, TAM and CHEMO) according with available information and treatment received by patients. All relevant clinico-pathological features were equalized and summarized in Supplementary Tables 1-3 for the last three collections.

## 2. Data processing using alternative CDFs

CEL files were imported using the affy Bioconductor package [1] and normalized using fRMA [2] with an alternative CDF as previously described [3]. Such method was specifically optimized to process GEPs from FFPE samples, but we have highlighted an improvement also in the downstream analysis of GEPs obtained from frozen tissues. Moreover, since both the GeneChip Human Genome U133 Plus 2.0 and U133A were used in this study, we defined here a new alternative CDF (named RefSeq_common) using only probes unambiguously mapping a RefSeq transcript and common to the two chip versions. A total of 11782 probesets (one probeset per gene) was generated.

To validate the consistency of signals from the same probesets measured with the two chip versions (i.e. U133A and U133 Plus 2.0), a dataset (GSE17700) containing 16 breast cancer samples profiled by two different labs and, in each lab, using both U133A and Plus 2.0 was evaluated. Probeset correlation across the samples was very high, with a median value of 0.901 (Supplementary Figure 1). Notably, the agreement between U133A and Plus 2.0 data was much higher using our normalization approach than when using MAS5 normalization together with the standard CDF (median correlation = 0.656) (Supplementary Figure 1).

## 3. Development of metagene-based predictors of ER and HER2 status

For the training of an ER status predictor, we selected two datasets from the GENERIC collection (GSE5460 and GSE19615, n=169) having an ER status accurately and homogeneously defined by IHC. Area under ROC curve (AUC) was computed for all genes and, starting from the best performing ones, we increased the number of genes to be included in metagene computation until the best AUC was obtained. Best performance was obtained using the top 7 genes (*C6orf97, ESR1, EVL, ABAT, SLC39A6, GATA3, SCUBE2*) giving an AUC=0.993. A threshold was defined looking at the bimodal distribution of the metagene as well as looking at the agreement with ER status defined by IHC. Using a cut-off of 7.5, a Cohen's Kappa value of 0.925 was obtained. The validity of the threshold was confirmed in the whole GENERIC collection (n=1,186) (Supplementary Figure 2).

To further validate the metagene as a sufficiently accurate predictor of ER status, it was applied to both the PROGNOSTIC and the TAM datasets. In the PROGNOSTIC dataset an AUC=0.927 and Cohen's Kappa of 0.755 were obtained, that can be judged as high considering that the reported ER status was obtained with either IHC or radio-ligand assay methods using different or not specified thresholds. The prognostic value of ER status defined by our metagene was compared with that of the available ER status using Kaplan-Meier curves and superimposable results were obtained (Supplementary Figure 3). In the TAM dataset, where all but 5 samples were labeled as ER positive, 43 samples were classified as negative according to our metagene. All the 5 samples originally labeled as ER negative, were classified as negative by our metagene, and the 43 patients (ER-negative according to our metagene classification) had a significantly worst outcome compared with all the others (Supplementary Figure 3), suggesting that at least most of them were actually ER negative and did not therefore benefit from hormone treatment.

An analogous approach was applied to identify HER2 positive tumors. Three datasets from the GENERIC collection (GSE5460, GSE16391, and GSE21653, n=417) were selected to develop the HER2 status predictor. In such datasets, HER2-positive samples were defined as those with IHC score=3+ or FISH positive. An AUC=0.933 was obtained when using the top 10 genes associated with HER2 status to derive a metagene (*ERBB2, PGAP3, STARD3, GRB7, PNMT, PSMD3, GSDMB, RPL19, FGFR4, CAP1*), eight of which were in the amplicon region. Looking at the metagene distribution and at the agreement with the available HER2 status, a threshold at 8.35 was defined giving a Cohen's Kappa=0.738 (Supplementary Figure 4).

Finally, by combining the ER and HER2 status prediction in the PROGNOSTIC collection we observed for the three groups (ER-HER2-, HER2+ and ER+HER2-) the expected pattern of relapse during the follow-up, that is an early relapse, mostly within 5 years, for ER-HER2- and HER2+ tumors and a quite constant risk of metastasis until 10 years in ER+HER2- cases (Supplementary Figure 5).

The defined cut-offs for ER and HER2 predictors were applied to all public datasets employed in our analyses.

## 4. Identification of clusters of consistently correlated genes

The GENERIC collection was used to identify clusters of stably correlated genes as outlined in Supplementary Figure 6.

The collection was randomly split in a discovery and confirmation subset of 593 samples each. In the discovery subset, genes were subjected to hierarchical clustering and, by cutting the dendrogram at a correlation value of 0.4, 45 clusters each containing at least 25 genes were derived. Each cluster was then re-clustered in the confirmation subset and, genes herein correlated less than 0.4, were removed. Clusters that globally retained at least 20 of the original genes were maintained. This approach defined robust clusters of correlated genes because the genes selected as belonging to each cluster were independently confirmed in two split sets. To allow validation of our findings also in FFPE-derived case series, a re-clustering was also performed in a case series of 44 FFPE samples derived in our lab (GSE38554). As before, genes correlating less than 0.4 were removed and clusters still containing at least 10 genes were retained. As expected, some of the probesets which were not correlated in FFPE derived GEPs had spurious level of correlations (also negative), that can be only explained by the technical failure of hybridization of the corresponding probes using fragmented FFPE-derived mRNA (data not shown). However, the biological function captured by the cluster will be still accurately measured by the selected well performing probesets. At the end of the two selection steps, 40 of the 45 clusters were confirmed. Guided by Gene Ontology analysis, we picked up among them a T-cell related immune metagene, a metagene related with the proliferation process and a metagene containing *ESR1* and *ESR1*-related genes.

## 5. Refinement of selected clusters and metagene computation

Despite all genes belonging to the identified clusters were correlated each other, by looking at univariable Cox analysis, we noted that they were not all equally associated with outcome. Therefore, it seemed reasonable that a selection of best performing genes could improve the metagene performance.

A method was developed to improve the metagene association with outcome, according with the scheme reported in Supplementary Figure 7.

In a 10-fold cross-validation setting, univariable Cox regression analysis was performed for all genes belonging to the cluster being refined in 9/10 of samples. The number of genes to include in the computation of the metagene ranged from all to 10 genes and, for each value, the three tertiles having low, intermediate and high expression of the metagene were defined and used to assign the class in the remaining 1/10. After completing the cross validation, the 5-year DMFS for the cross-validated groups was estimated and log-rank test p-value was computed. The cross-validation was then repeated 100 times obtaining average survival and p-values, together with their standard deviations. This way it was possible to identify the best number of genes with lower univariable Cox p-values to be included in the refined metagene.

In the PROGNOSTIC collection, 179 ER-HER2- samples were identified by applying our genomic predictors of ER and HER2 status. In this subset the refinement procedure was applied to the immune cluster varying the number of genes from all in the cluster (n=92) to 10. Cross-validated refinement suggested to use the top 25 genes able to predict an average 5-year distant metastasis-free survival (DMFS) of 78, 75 and 58% for patients in the high, intermediate or low tertile of metagene levels, respectively (p<1e-2) (Supplementary Table 4).

The same cluster was refined in the subgroup of 122 HER2+ tumors from the PROGNOSTIC collection (Supplementary Table 5). This time, best performance was obtained using 10 genes with an average 5-year DMFS of 87, 81 and 51% for patients in the high, intermediate or low tertile of metagene levels, respectively (p<1e-4).

Immune genes selected in the two subtypes were 25 in ER-HER2- and 10 in HER2+ with 6 of them in common (*CXCL13, PRF1, IRF1, IKZF1, GZMB, HLA-E*). The difference could be a simple consequence of the fact that refinement was carried out in independent set of samples or a sign of partially different underlying biological mechanisms in the two subtypes. Under the first hypothesis, the six genes in common could be even more robust because they were identified as being the best in two independent sets. Moreover, these two subtypes usually constitute a minority of the samples in most studies (10-20% each), therefore such consensus T cell-related metagene (CTM) could facilitate a pooled analysis for ER-HER2- and HER2+ samples. As a first indication, we tested whether the common metagene worked well in the two subtypes in the PROGNOSTIC collection itself, although we know that such result could be someway overfitted (Supplementary Figure 8A-B). We also tested the CTM after stratifying HER2+ samples by ER status (Supplementary Figure 8C-D)

The proliferation cluster was refined in the subgroup of 508 ER+HER2- tumors identified in the PROGNOSTIC collection. By gradually reducing the number of genes included in the metagene, we obtained the best performance using the top 10 genes (*NCAPG, BUB1B, PRC1, CCNB2, RAD51AP1, ORC6, FANCI, UBE2C, AURKA, KIF20A*) associated with an average 5-year DMFS of 64, 88 and 91% for patients in the high, intermediate or low tertile of metagene levels, respectively (p<1e-13) (Supplementary Table 6).

The association with prognosis for the refined proliferation metagene was also validated in the cohort of patients receiving hormone treatment (TAM dataset), both in the same contest of lymph node negative patients and out of context in the lymph node positive group, confirming its independence from lymph node status (Supplementary Figure 9).

Looking at the Kaplan-Meier curves in Supplementary Figure 9 above, it can be noted that the group of patients with low proliferation tumors has a very good outcome, consequently trying to predict sensitiveness to hormone therapy including such subgroup is disadvantageous from a statistical point of view due to the very low number of events that limit the possibility to identify significant predictive factors. In keeping with this consideration, our approach was to exclude low proliferation ER+HER2- tumors (lower tertile of the proliferation metagene) and to focus on the remaining 394 samples of the TAM collection with intermediate or high proliferation. Here the refinement procedure was applied to the ER-related cluster, and we chose to use the top 10 genes (*ABAT, CA12, MCCC2, SCUBE2, LRIG1, FAM63A, C14orf45, MYB, CACNA1D, GATA3*). The generated metagene was able to stratify patients based on average 5-year DMFS. Ninety-four, 85, and 75% of patients were disease free for more than 5 years in the high, intermediate and low tertile, respectively (p<1e-5) (Supplementary Table 7).

With exploratory purposes, we tested whether the refined ER-related metagene provide any prognostic/predictive information in the ER+HER2+ subgroup. As reported in Supplementary Figure 10, no significant association was found.

It is reasonable to hypothesize that in patients with ER+HER2- breast cancers receiving adjuvant hormone treatment, the outcome could be consequence of a combination of the tumor aggressiveness and its ability to respond/resist to the targeted treatment. This means that a combined evaluation of the prognostic proliferation metagene and the predictive ER-related metagene might better stratify the patients.

Following our strategy of deriving simple and weight-independent predictors, we searched for reasonable cut-offs in the distribution of the two metagenes. By plotting the levels of the two metagenes and highlighting the events within 5 years, a median cut point seemed to be appropriate. The group with low proliferation and high ER-related metagenes had the best prognosis. On the opposite, those having high proliferation and low ER-related metagene had the worst prognosis, while the other two groups had an intermediate risk (Supplementary Figure 11).

## 6. Immune metagene prognostic and predictive value in chemotherapy receiving ER-HER2- and HER2+ subtypes

Association with treatment response of our CTM in ER- or HER2+ cases receiving neoadjuvant CT was evaluated in two datasets of the CHEMO collection: i) the Horak dataset (GSE41998), where patients received neoadjuvant doxorubicin/cyclophosphamide, followed by 1:1 randomization to ixabepilone or paclitaxel; ii) the Desmedt dataset (GSE16446) where patients only received a single agent treatment (epirubicin). A total of 146 and 114 ER- or HER2+ samples were identified, respectively. High, intermediate and low metagene expression was defined separately in each dataset according with tertiles. In both datasets comparable odds ratios were obtained, with 33% of patients with higher expression of the CTM having a significantly higher rate of pCR (Supplementary Figure 12).

The CTM was not evaluated in ER-HER2- and HER2+ samples of the two Hatzis datasets (GSE25055 and GSE25065) because most of these samples were fine needle aspiration (FNA) that are enriched in tumor cells while stromal cells are poorly and variably represented, distorting the evaluation of immune specific genes.

We investigated the association with long-term outcome of our CTM in HER2+ tumors (CHEMO collection) after stratifying by ER status (Supplementary Figure 13)

Association of the CTM with patients' outcome was separately evaluated in: i) Petel dataset (E-MTAB-365) where patients received not homogeneous adjuvant CT and ii) Desmedt dataset (GSE16446) where patients only received a single agent treatment (epirubicin). As before, high, intermediate and low metagene expression was defined in each dataset according with tertiles.

In Petel dataset only a trend was found in ER- or HER2+ group but this was caused by a lack of association with outcome in HER2+ subgroups whilst the association was strong and significant in ER-HER2- subgroup (Supplementary Figure 14).

Similar results were obtained in the Desmedt dataset. The number of HER2+ samples was too small to plot meaningful Kaplan-Meier curves but, as can be noted in Supplementary Figure 15, the performance of the CTM in the combined group was poorer than in ER-HER2- group only.

Patients achieving a pCR are known to have an excellent outcome, in particular in ER-HER2- and HER2+ tumors. However the double role (prognostic and predictive) of the CTM allows a significant stratification of the long-term risk also in patients with a residual disease after neoadjuvant treatment (Supplementary Figure 16). As reported previously, the prognostic value seems to be stronger in ER-HER2- tumors.

Besides Kaplan-Meier analysis, the immune metagene was evaluated in the different datasets and in the different subgroups by univariable Cox regression analysis that traced the conclusions above (Supplementary Table 8).

Performances of our CTM (Figure 2A-B in the main manuscript) were compared in the same set of data with performances of two previously published immune signatures: LCK [4] and Tfh [5] metagenes. Association with pCR is reported in Supplementary Figure 17 while association with long-term outcome is reported in Supplementary Figure 18.

For a direct comparison of the prognostic performances of our CTM with LCK and Tfh, we performed a multivariable analysis including all the three metagenes. Only the immune metagene remained significant, suggesting that even if the biomarkers provide similar prognostic information (all were significant in univariable analysis) it was outperforming the other two immune-related biomarkers (Supplementary Table 9).

The genes included in the CTM (i.e. GZMB, PRF1, IKZF1 and CXCL13) suggested that the signal provided by this immune marker is strongly associated with cytotoxic T cells.  To confirm this association, we assessed its correlation with two different T cell related metagenes (Tfh and LCK) in ER- or HER2+ samples from the CHEMO cohort. We also evaluated the association of our CTM with cell types (Treg, Macrophages) and regulatory signaling (co-inhibitory molecules expressed on T cell or antigen-presenting cells) which are expected to be involved in immune tolerance and escape. For this aim, we derived the corresponding specific gene signatures from Rooney et al [6] (Supplementary Figure 19). A positive correlation of different degree was found for all the immune markers with our CTM. The strongest correlation was found with the two T cell metagens, confirming that this signature is likely to capture engaged and activated T cells. The positive and significant association with co-inhibitory immune molecules suggests that the immune system activation captured by our metagene is associated with inhibitory and dampening signals engaged as negative regulatory feedbacks. This observation was similarly reported by many other authors (including Rooney et al).

## 7. Predictive and prognostic value of combined proliferation and ER-related metagenes in chemo-endocrine receiving ER+HER2- patients

In ER+HER2- subtype, high, intermediate and low risk groups where defined as described in the main manuscript and section 5 of this supplementary information. Association with response to treatment, looking at pCR as surrogate, was evaluated by logistic regression analysis in the Hatzis (n=250) and Horak (n=107) datasets. High-risk group was significantly associated with higher pCR rate in Hatzis dataset but not in Horak dataset (Supplementary Figure 20).

Although high-risk group seems to benefit the most from CT, it still showed the worst outcome in the Hatzis dataset as well as in the Petel dataset. Only ER+HER2- patients that actually received endocrine therapy were included in this analysis (Supplementary Figure 21).

As shown in Supplementary Figure 22, patients with a residual disease after neoadjuvant CT can be significantly stratified by our risk groups defined by combining expression levels of the proliferation end ER-related metagene.

Besides Kaplan-Meier analysis, the combined proliferation and ER-related metagenes were evaluated in the different datasets by univariable Cox regression analysis that traced the conclusions above (Supplementary Table 10).

## 8. Prediction in HER2- treated patients

Risk groups identified separately in ER-HER2- and ER+HER2- subtypes were then combined. Overall, the low, intermediate and high-risk groups had 91%, 83% and 72% 5 year DMFS, respectively (p=1.5 E-06, Supplementary Figure 23).

# Supplementary References

1. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level** . *Bioinforma* 2004, **20** (1367-4803 (Print)):307–315.

2. McCall MN, Bolstad BM, Irizarry RA: **Frozen robust multiarray analysis (fRMA)** . *Biostat* 2010, **11** (1468-4357 (Electronic)):242–253.

3. Callari M, Lembo A, Bianchini G, Musella V, Cappelletti V, Gianni L, Daidone MG, Provero P: **Accurate Data Processing Improves the Reliability of Affymetrix Gene Expression Profiles from FFPE Samples**. *PLoS One* 2014, **9**(1932-6203 (Electronic)):e86511.

4. Rody A, Holtrich U, Pusztai L, Liedtke C, Gaetje R, Ruckhaeberle E, Solbach C, Hanker L, Ahr A, Metzler D, Engels K, Karn T, Kaufmann M: **T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers**. *Breast Cancer Res* 2009, **11**(1465-542X (Electronic)):R15.

5. Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, De WA, Ravoet M, Le BH, Sibille C, Manfouo-Foutsop G, Veys I, Haibe-Kains B, Singhal SK, Michiels S, Rothe F, Salgado R, Duvillier H, Ignatiadis M, Desmedt C, Bron D, Larsimont D, Piccart M, Sotiriou C, Willard-Gallo K: **CD4(+) follicular helper T cell infiltration predicts breast cancer survival**. *J Clin Invest* 2013, **123**(1558-8238 (Electronic)):2873–2892.

6. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N: **Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity**. *Cell* 2015, **160**:48–61.

# Supplementary Tables

**Supplementary Table 1 - Clinico-pathological features for samples in the PROGNOSTIC collection**

| | | | **Dataset** | | |
|---|---|---|---|---|---|
| | **GSE2034** | **GSE7390** | **GSE11121** | **GSE5327** | **GSE2990** |
| | **Wang** | **Transbig** | **Mainz** | **Minn** | **Sotiriou** |
| **Total samples** | 286 | 198 | 200 | 58 | 84 |
| **Age** | | | | | |
| Range (years) | 26 - 83 | 24 - 60 | 25 - 90 | - | 34 - 73 |
| Median (years) | 52 | 46 | 60 | - | 57 |
| ≤ 50 | 129 | 142 | 51 | 0 | 29 |
| > 50 | 157 | 56 | 149 | 0 | 55 |
| Unknown | 0 | 0 | 0 | 58 | 0 |
| **Size** | | | | | |
| Range (cm) | - | 0.6 - 5 | 0.1 - 6 | - | 0 - 6 |
| Median (cm) | - | 2 | 2 | - | 2 |
| ≤ 2 | 146 | 102 | 112 | 0 | 48 |
| > 2 | 140 | 96 | 88 | 0 | 36 |
| Unknown | 0 | 0 | 0 | 58 | 0 |
| **Grade** | | | | | |
| G1 | 7 | 30 | 29 | 0 | 21 |
| G2 | 42 | 83 | 136 | 0 | 31 |
| G3 | 148 | 83 | 35 | 0 | 19 |
| Unknown | 0 | 2 | 0 | 58 | 13 |
| **Limph node status** | | | | | |
| Pos | 0 | 0 | 0 | 0 | 0 |
| Neg | 286 | 198 | 200 | 58 | 84 |
| Unknown | 0 | 0 | 0 | 0 | 0 |
| **Histotype** | | | | | |
| CDI | 0 | 158 | 0 | 0 | 0 |
| CDI+CLI | 0 | 11 | 0 | 0 | 0 |
| CLI | 0 | 13 | 0 | 0 | 0 |
| Other | 0 | 12 | 0 | 0 | 0 |
| Unknown | 286 | 4 | 200 | 58 | 84 |
| **ER status** | | | | | |
| Pos | 209 | 134 | 162 | 0 | 61 |
| Neg | 77 | 64 | 38 | 58 | 15 |
| Unknown | 0 | 0 | 0 | 0 | 5 |
| **HER2 status*** | | | | | |
| Pos | 51 | 27 | 24 | 10 | 13 |
| Neg | 235 | 171 | 176 | 48 | 71 |
| Unknown | 0 | 0 | 0 | 0 | 0 |
| **Follow-up time**** | | | | | |
| Range (months) | 2 - 171 | 4 - 299 | 1 - 240 | 7 - 156 | 2 - 174 |
| Median (months) | 86 | 144 | 91 | 87 | 96 |
| **Systemic treatment** | | | | | |
| Chemotherapy | 0 | 0 | 0 | 0 | 0 |
| Endocrine terapy | 0 | 0 | 0 | 0 | 0 |

* Based on genomic HER2 status predictor
** Time to Distant Metastasis

**Supplementary Table 2 - Clinico-pathological features for samples in the TAM collection**

| | | | Dataset | | |
|---|---|---|---|---|---|
| | **GSE12093** | **GSE17705** | **GSE9195** | **GSE6532** | **GSE6532** |
| | **Zhang** | **Symmans** | **Loi** | **Loi (PLUS2)** | **Loi (U133A)** |
| **Total samples** | 136 | 195 | 77 | 87 | 190 |
| **Age** | | | | | |
| Range (years) | - | - | 42 - 82 | 43 - 86 | 40 - 88 |
| Median (years) | - | - | 65 | 82 | 65 |
| ≤ 50 | 0 | 0 | 6 | 5 | 16 |
| > 50 | 0 | 0 | 71 | 62 | 165 |
| Unknown | 136 | 195 | 0 | 0 | 9 |
| **Size** | | | | | |
| Range (cm) | - | - | 1.1 - 6 | 1.1 - 7.5 | 0 - 8.2 |
| Median (cm) | - | - | 2.1 | | 2.4 |
| ≤ 2 | 0 | 0 | 34 | 43 | 67 |
| > 2 | 0 | 0 | 43 | 44 | 114 |
| Unknown | 136 | 195 | 0 | 0 | 9 |
| **Grade** | | | | | |
| G1 | 0 | 0 | 14 | 17 | 33 |
| G2 | 0 | 0 | 20 | 37 | 94 |
| G3 | 0 | 0 | 24 | 16 | 31 |
| Unknown | 136 | 195 | 19 | 17 | 32 |
| **Limph node status** | | | | | |
| Pos | 0 | 80 | 36 | 58 | 85 |
| Neg | 0 | 110 | 41 | 29 | 87 |
| Unknown | 136 | 5 | 0 | 0 | 18 |
| **Histotype** | | | | | |
| CDI | 0 | 0 | 0 | 0 | 0 |
| CDI+CLI | 0 | 0 | 0 | 0 | 0 |
| CLI | 0 | 0 | 0 | 0 | 0 |
| Other | 0 | 0 | 0 | 0 | 0 |
| Unknown | 136 | 195 | 77 | 87 | 190 |
| **ER status** | | | | | |
| Pos | 136 | 195 | 77 | 87 | 185 |
| Neg | 0 | 0 | 0 | 0 | 5 |
| Unknown | 0 | 0 | 0 | 0 | 0 |
| **HER2 status*** | | | | | |
| Pos | 10 | 9 | 6 | 6 | 22 |
| Neg | 126 | 186 | 71 | 81 | 168 |
| Unknown | | | | | |
| **Follow-up time**** | | | | | |
| Range (months) | 8 - 193 | 6 - 195 | 15 -135 | 6 - 202 | 0 - 153 |
| Median (months) | 85 | 113 | 99 | 137 | 62 |
| **Systemic treatment** | | | | | |
| Chemotherapy | 0 | 0 | 0 | 0 | 0 |
| Tamoxifen (5 years) | 136 | 195 | 77 | 87 | 190 |

* Based on genomic HER2 status predictor

** Time to Distant Metastasis

## Supplementary Table 3 - Clinico-pathological features for samples in the CHEMO collection

| | | | Dataset | | |
|---|---|---|---|---|---|
| | **E-MTAB-365** | **GSE16446** | **GSE25055** | **GSE25065** | **GSE41998** |
| | **Petel** | **Desmedt** | **Hazis discovery** | **Hazis validation** | **Horak** |
| **Patients** | 243 | 120 | 310 | 198 | 279 |
| **Age** | | | | | |
| Range (years) | 29-71 | - | 26 - 75 | 24 - 72 | 25 - 79 |
| Median (years) | 50 | - | 49 | 48 | 48 |
| ≤ 50 | 123 | 0 | 168 | 109 | 166 |
| > 50 | 120 | 0 | 142 | 89 | 113 |
| Unknown | 0 | 120 | 0 | 0 | 0 |
| **Size** | | | | | |
| Range (cm) | - | - | - | - | - |
| Median (cm) | - | - | - | - | - |
| ≤ 2 | 0 | 0 | 0 | 0 | 0 |
| > 2 | 0 | 0 | 0 | 0 | 0 |
| Unknown | 243 | 120 | 310 | 198 | 279 |
| **Grade** | | | | | |
| G1 | 25 | 2 | 19 | 13 | - |
| G2 | 92 | 20 | 117 | 64 | - |
| G3 | 122 | 92 | 151 | 107 | - |
| Unknown | 4 | 6 | 23 | 14 | 279 |
| **Limph node status** | | | | | |
| Pos | 202 | 65 | 223 | 128 | 0 |
| Neg | 38 | 55 | 87 | 70 | 0 |
| Unknown | 3 | 0 | 0 | 0 | 279 |
| **Histotype** | | | | | |
| CDI | 209 | 0 | 0 | 0 | 0 |
| CDI+CLI | 0 | 0 | 0 | 0 | 0 |
| CLI | 17 | 0 | 0 | 0 | 0 |
| Other | 4 | 0 | 0 | 0 | 0 |
| Unknown | 13 | 120 | 310 | 198 | 279 |
| **ER status** | | | | | |
| Pos | 159 | 3 | 167 | 107 | 121 |
| Neg | 83 | 117 | 143 | 91 | 158 |
| Unknown | 1 | 0 | 0 | 0 | 0 |
| **HER2 status** | | | | | |
| Pos | 33 | 28 | 9 | 14 | 26 |
| Neg | 177 | 92 | 301 | 184 | 253 |
| Unknown | 33 | 0 | 0 | 0 | 0 |
| **Pathologic Response** | | | | | |
| pCR | - | 16 | 57 | 42 | 69 |
| RD | - | 98 | 249 | 140 | 184 |
| Unknown | - | 6 | 3 | 16 | 26 |
| **Residual Cancer Burden** | | | 1 | | |
| RCB-0/I | - | - | 86 | 32 | 86 |
| RCB-II/III | - | - | 215 | 84 | 167 |
| Unknown | - | - | 9 | 82 | 26 |
| **Follow-up time*** | | | | | |
| Range (months) | 0-193 | 2-71 | 0 - 89 | 2-88 | - |
| Median (months) | 76 | 33 | 29 | 38 | - |
| **Systemic treatment** | | | | | |
| ***Neoadjuvant chemotherapy*** | | | | | |
| Epirubicin monotherapy | 0 | 120 | 0 | 0 | 0 |
| AC → Ixabepilone | 0 | 0 | 0 | 0 | 138 |
| AC → Paclitaxel | 0 | 0 | 0 | 0 | 127 |
| AC | 0 | 0 | 0 | 0 | 14 |
| Taxane-anthracycline regimen | 0 | 0 | 310 | 183 | 0 |
| ***Adjuvant chemotherapy*** | | | | | |
| Taxane-anthracycline regimen | 47 | 0 | 0 | 15 | 0 |
| Anthracycline regimen | 155 | 0 | 0 | 0 | 0 |
| HER2-targeted agents | 2 | 0 | 0 | 0 | 0 |
| Not specified chemotherapy‡ | 41 | 0 | 0 | 0 | 0 |
| ***Adjuvant endocrine therapy*** | | | | | |
| Endocrine therapy† | 128 | 0 | 167 | 107 | 0 |

* Time to Distant Metastasis

‡ These grup included adjuvant chemotherapy regimens not completely specified or hetherogenous

† Endocrine treatment were not specified

AC (doxorubicin and cyclophosphamide ) followed by ixabepilone (40 mg/m2) AC (doxorubicin and cyclophosphamide ) followed by weekly paclitaxel

**Supplementary Table 4 - Output of immune cluster refinement in ER-HER2- samples.** Five years DMFS for high, intermediate and low expression groups and log-rank test p-values (average and standard deviation of 100 10-fold cross-validations) are reported as a function of the number of genes.

| Number of genes | average 5-years DMFS (%) after 100 cross-validations | | | average -log10(P) | standard deviation of 5-years DMFS (%) after 100 cross-validations | | | standard deviation of -log10(P) |
|---|---|---|---|---|---|---|---|---|
| | Low exp | Intermediate exp | High exp | | Low exp | Intermediate exp | High exp | |
| 92 | 57.23 | 78.50 | 75.48 | 2.29 | 0.84 | 1.10 | 0.48 | 0.28 |
| 80 | 56.67 | 78.43 | 76.08 | 2.46 | 1.02 | 1.59 | 0.68 | 0.38 |
| 70 | 57.06 | 77.89 | 76.11 | 2.31 | 0.94 | 1.33 | 0.71 | 0.34 |
| 60 | 57.72 | 76.87 | 76.28 | 2.08 | 1.35 | 1.76 | 0.71 | 0.43 |
| 50 | 58.94 | 75.18 | 76.60 | 1.74 | 1.35 | 1.66 | 0.83 | 0.35 |
| 45 | 59.81 | 74.03 | 76.94 | 1.57 | 1.50 | 1.83 | 0.84 | 0.33 |
| 40 | 59.93 | 73.61 | 77.26 | 1.58 | 1.50 | 2.00 | 1.00 | 0.31 |
| 35 | 58.92 | 74.56 | 77.36 | 1.79 | 1.45 | 1.90 | 0.96 | 0.32 |
| 30 | 58.23 | 74.89 | 77.65 | 1.95 | 1.43 | 1.84 | 0.81 | 0.33 |
| **25** | **57.92** | **75.27** | **77.79** | **2.05** | **1.59** | **1.87** | **0.83** | **0.36** |
| 20 | 59.13 | 74.15 | 77.81 | 1.79 | 1.86 | 1.95 | 0.92 | 0.40 |
| 15 | 60.09 | 73.16 | 77.83 | 1.60 | 1.73 | 2.01 | 0.99 | 0.35 |
| 10 | 59.98 | 72.74 | 78.24 | 1.65 | 1.77 | 1.98 | 0.95 | 0.33 |

**Supplementary Table 5 - Output of immune cluster refinement in HER2+ samples.** Five years DMFS for high, intermediate and low expression groups and log-rank test p-values (average and standard deviation of 100 10-fold cross-validations) are reported as a function of the number of genes.

| Number of genes | average 5-years DMFS (%) after 100 cross-validations | | | average -log10(P) | standard deviation of 5-years DMFS (%) after 100 cross-validations | | | standard deviation of -log10(P) |
|---|---|---|---|---|---|---|---|---|
| | Low exp | Intermediate exp | High exp | | Low exp | Intermediate exp | High exp | |
| 92 | 52.61 | 80.47 | 85.55 | 3.33 | 1.04 | 1.26 | 0.43 | 0.27 |
| 80 | 51.14 | 81.91 | 85.57 | 3.74 | 1.21 | 1.05 | 0.46 | 0.30 |
| 70 | 51.90 | 81.35 | 85.53 | 3.55 | 1.55 | 1.39 | 0.41 | 0.40 |
| 60 | 52.35 | 81.11 | 85.45 | 3.43 | 1.42 | 1.28 | 0.37 | 0.34 |
| 50 | 52.47 | 81.11 | 85.38 | 3.40 | 1.41 | 1.29 | 0.42 | 0.34 |
| 45 | 52.49 | 81.10 | 85.38 | 3.41 | 1.63 | 1.61 | 0.38 | 0.42 |
| 40 | 52.41 | 81.15 | 85.30 | 3.42 | 1.62 | 1.60 | 0.41 | 0.41 |
| 35 | 52.68 | 80.89 | 85.24 | 3.36 | 2.00 | 1.98 | 0.38 | 0.50 |
| 30 | 52.67 | 80.51 | 85.25 | 3.35 | 1.83 | 1.99 | 0.41 | 0.49 |
| 25 | 54.08 | 79.25 | 85.20 | 3.02 | 1.94 | 2.16 | 0.57 | 0.48 |
| 20 | 54.55 | 79.15 | 85.07 | 2.92 | 1.83 | 2.12 | 0.73 | 0.46 |
| 15 | 53.88 | 79.78 | 85.22 | 3.10 | 1.93 | 1.76 | 0.88 | 0.43 |
| **10** | **51.26** | **81.34** | **86.56** | **4.02** | **1.87** | **2.05** | **1.30** | **0.57** |

**Supplementary Table 6 - Output of proliferation cluster refinement in 508 ER+HER2- samples.** Five-years DMFS for high, intermediate and low expression groups and log-rank test p-values (average and standard deviation of 100 10-fold cross-validations) are reported as a function of the number of genes.

| Number of genes | average 5-years DMFS (%) after 100 cross-validations | | | average -log10(P) | standard deviation of 5-years DMFS (%) after 100 cross-validations | | | standard deviation of -log10(P) |
|---|---|---|---|---|---|---|---|---|
| | Low exp | Intermediate exp | High exp | | Low exp | Intermediate exp | High exp | |
| 102 | 93.30 | 83.00 | 66.83 | 10.03 | 0.25 | 0.47 | 0.37 | 0.35 |
| 90 | 93.79 | 82.25 | 67.11 | 9.98 | 0.32 | 0.37 | 0.31 | 0.28 |
| 80 | 94.09 | 82.52 | 66.52 | 10.67 | 0.39 | 0.50 | 0.38 | 0.45 |
| 70 | 93.90 | 83.95 | 65.33 | 11.84 | 0.51 | 0.69 | 0.48 | 0.51 |
| 60 | 93.49 | 85.03 | 64.73 | 12.37 | 0.45 | 0.54 | 0.40 | 0.45 |
| 50 | 93.04 | 85.30 | 64.91 | 12.01 | 0.54 | 0.70 | 0.46 | 0.57 |
| 45 | 93.14 | 85.54 | 64.58 | 12.39 | 0.51 | 0.72 | 0.49 | 0.58 |
| 40 | 92.51 | 86.36 | 64.44 | 12.41 | 0.40 | 0.55 | 0.43 | 0.50 |
| 35 | 92.21 | 86.60 | 64.44 | 12.33 | 0.37 | 0.53 | 0.43 | 0.52 |
| 30 | 91.87 | 87.09 | 64.32 | 12.40 | 0.33 | 0.61 | 0.49 | 0.59 |
| 25 | 91.70 | 87.58 | 64.08 | 12.70 | 0.43 | 0.68 | 0.54 | 0.67 |
| 20 | 91.50 | 87.78 | 64.14 | 12.60 | 0.45 | 0.70 | 0.53 | 0.66 |
| 15 | 91.55 | 87.89 | 64.04 | 12.76 | 0.45 | 0.68 | 0.64 | 0.78 |
| **10** | **91.47** | **88.34** | **63.61** | **13.28** | **0.53** | **0.81** | **0.60** | **0.75** |

**Supplementary Table 7 - Output of ER-related cluster refinement in 394 high proliferation ER+HER2- samples.** Five-years DMFS for high, intermediate and low expression groups and log-rank test p-values (average and standard deviation of 100 10-fold cross-validations) are reported as a function of the number of genes.

| Number of genes | average 5-years DMFS (%) after 100 cross-validations | | | average -log10(P) | standard deviation of 5-years DMFS (%) after 100 cross-validations | | | standard deviation of -log10(P) |
|---|---|---|---|---|---|---|---|---|
| | Low exp | Intermediate exp | High exp | | Low exp | Intermediate exp | High exp | |
| 43 | 75.01 | 85.32 | 93.99 | 4.87 | 0.57 | 0.72 | 0.48 | 0.38 |
| 35 | 75.78 | 83.47 | 95.07 | 5.03 | 0.76 | 0.92 | 0.50 | 0.37 |
| 30 | 76.74 | 83.25 | 94.40 | 4.37 | 0.76 | 1.18 | 0.80 | 0.43 |
| 25 | 77.07 | 83.75 | 93.58 | 3.89 | 0.85 | 1.15 | 0.92 | 0.49 |
| 20 | 77.19 | 83.91 | 93.31 | 3.78 | 1.10 | 1.38 | 1.00 | 0.58 |
| 15 | 76.43 | 84.30 | 93.63 | 4.26 | 1.22 | 1.50 | 1.00 | 0.64 |
| **10** | **75.24** | **85.11** | **93.99** | **5.04** | **1.14** | **1.29** | **0.82** | **0.73** |

**Supplementary Table 8 - Univariable Cox regression analysis for the CTM in the combined dataset or in each dataset for ER- or HER2+ cases or separately for ER-HER2- and HER2+ subgroups.**

| | HR | CI.low | CI.up | P |
|---|---|---|---|---|
| ER- or HER2+ | | | | |
| All (n=205) | 0.57 | 0.41 | 0.77 | 0.00036 |
| Desmedt (n=107) | 0.54 | 0.33 | 0.87 | 0.01100 |
| Petel (n=98) | 0.61 | 0.40 | 0.92 | 0.01824 |
| ER-HER2- | | | | |
| All (n=122) | 0.44 | 0.30 | 0.65 | 0.00003 |
| Desmedt (n=80) | 0.47 | 0.28 | 0.78 | 0.00371 |
| Petel (n=42) | 0.39 | 0.21 | 0.74 | 0.00411 |
| HER2+ | | | | |
| All (n=83) | 0.74 | 0.42 | 1.29 | 0.28321 |
| Desmedt (n=27) | 1.08 | 0.26 | 4.44 | 0.91367 |
| Petel (n=56) | 0.70 | 0.37 | 1.30 | 0.25341 |

**Supplementary Table 9 – Univariable and multivariable Cox regression analysis in 205 ER-HER2- and HER2+ samples from the CHEMO collection (all biomarkers were used as continous variables)**

| **Univariable Cox regression** | | | | |
|---|---|---|---|---|
| | | **HR** | **CI (95%)** | **P** |
| **ER- or HER2+** | | | | |
| | **LCK** | 0.63 | (0.46-0.87) | 0.0054 |
| | **Tfh** | 0.53 | (0.35-0.79) | 0.0020 |

| **Multivariable Cox regression** | | | | |
|---|---|---|---|---|
| | | **HR** | **CI (95%)** | **P** |
| **ER- or HER2+** | | | | |
| | **CTM** | 0.32 | (0.12-0.90) | 0.03110 |
| | **LCK** | 1.40 | (0.65-3.01) | 0.39590 |
| | **Tfh** | 1.45 | (0.43-4.83) | 0.54700 |

**Supplementary Table 10 - Univariable Cox regression analysis for the combined proliferation and ER-related metagenes in ER+HER2- cases from the combined dataset or from each dataset**

| | | HR | CI.low | CI.up | P |
|---|---|---|---|---|---|
| ALL (n=350) | | | | | |
| | High vs Low risk | 3.73 | 1.63 | 8.51 | 0.0018 |
| | Interm vs Low risk | 2.20 | 0.97 | 5.01 | 0.0594 |
| Hatzis (n=242) | | | | | |
| | High vs Low risk | 4.72 | 1.33 | 16.75 | 0.0163 |
| | Interm vs Low risk | 2.99 | 0.85 | 10.48 | 0.0877 |
| Petel (n=108) | | | | | |
| | High vs Low risk | 12.52 | 1.60 | 97.87 | 0.0160 |
| | Interm vs Low risk | 6.54 | 0.83 | 51.65 | 0.0748 |