Supplementary patients and methods

1. Patients and methods, Cohort 1

Brief note: Here, we state our cohort definition and describe the design of our nested case-control study. Whilst we did not use nested design in our analysis, we did use the individual patients from the case-control study and given that this cohort has not previously been described, it is necessary to provide all details for the sake of completeness and clarity. In addition, we clearly define the population analyzed for the present study in the section "Present study population".

Patient cohort definition

Our patient cohort was defined as women diagnosed with primary breast cancer from January 1st 1997 to December 31st 2005 in the Stockholm health care region who were identified using the population based Stockholm-Gotland Breast Cancer Registry in conjunction with the unique 12 digit personal identification number assigned to every individual living in Sweden.

Original nested case-control study design

From the population-based cohort patients were selected who developed distant metastatic disease within the study period (cases). Next, controls (defined as patients free from disseminated disease for the same absolute time corresponding to the matched case) were selected and randomly matched them to each case according to the following criteria; assigned adjuvant therapy (chemotherapy (CT), endocrine therapy (ET) or a combination), age at primary breast cancer diagnosis (<45 years of age, 45-54 years, and >55 years), and calendar time period (1997-2000 or 2001-2005).

The nested case-control study cohort was comprised of 768 female study subjects (621 individuals including two with bilateral breast cancer tumors) aged 75 or younger with

available fresh frozen tumor tissue, excluding patients with primary tumor stage IV (predefined) at the time of initial primary tumor diagnosis. Most cases had three controls each (88.4%); 15 cases (7.9%) had four controls and 7 cases (3.7%) had two controls.

Present study population

We do not employ the nested case-control design in the present study, but rather take individual patients (n=621) to form a straightforward cohort. In doing so, it is important to note that we have an over-representation of metastatic events relative to a typical breast cancer population.

Clinical Information

All patient medical records were examined for detailed clinical information resulting in more than 200 clinical variables that form the basis for this study. Primary breast tumor ER, progesterone receptor (PR), lymph node status and tumor size were manually collected from original pathology reports. Tumor grade was re-evaluated by a breast cancer pathologist according to the Elston Ellis (1) grading system. Human epidermal growth factor receptor 2 (HER2), evaluated by Chromogenic in Situ Hybridization (CISH) (2), and Ki67 expression (MIB-1 antibody, 1:100 dilution, DAKO, Glostrup, Denmark) were reassessed in TMAs (P.K.W). Two TMA core blocks were used with two sections scored from each utilizing the same tumor sample as for the isolation of RNA. The TMAs were created under the supervision of and scored by a breast cancer pathologist. An overview of the numbers and percentages for each of the aforementioned clinicopathological variables in this cohort is shown split by ER status in Table 1 of the main manuscript and BRISQ sample handling criteria for cohort 1 is shown here:

BRISQ criteria, Cohort 1	
Data Elements	**Information**
Biospecimen type	Primary tumor material
Anatomical site	Breast
Clinical characteristics of patients	Pre & Post-menopausal breast cancer patients
Vital state of patients	Alive at the time of tumor excision
Clinical diagnosis of patients	Primary breast cancer
Pathology diagnosis	Table 1
Collection mechanism	Whole tumor excision, Formalin fixation + freezing samples
Type of stabilization	Dry ice and on ice
Type of long-term preservation	Formalin fixation + freezing
Constitution of preservation	10% formalin
Storage temperature	-80 ºC (frozen samples), room temperature
Storage duration	16 years
Shipping temperature	-80 ºC on Dry ice
Composition assessment and selection	Patients were selected on the basis of development of metastatic disease, See Patients & Methods.

2. Array profiling and normalisation

Gene Expression Profiling by Microarray

In this study, surgically excised tumours were divided in two with half being used for assessment of routine pathological markers (e.g. ER, PR, Her2, Ki67, grade) and the other half being fresh frozen for gene expression array profiling. Of note, the intra-tumour correlation for gene signatures (i.e. running signatures on both pieces of a primary breast tumour divided in two) has been demonstrated to be very high with an intra-class correlation of 0.90 (3). In addition, gene expression profiles from FFPE have been shown to be comparable to those from fresh frozen tissue (4–6) and the GGI, 70-gene, recurrence score

and PAM50 signatures have all compared on both FFPE and fresh frozen tissues (7–10) with a high degree of concordance observed.

Extraction of total RNA from frozen tumors was carried out using the Qiagen RNeasy Mini Kit (Qiagen, Germany). All patient tumor samples were profiled using NuGEN amplification protocol and hybridized using the HRSTA-2.0 custom human Affymetrix array GPL10379.

Details of the custom array are available at NCBI GEO depository as GPL10379. The order of the patient tumor samples to be profiled was randomized and the samples were subsequently profiled according to this order. The gene expression microarrays have been deposited into the Gene Expression Omnibus under the accession number of GSE48091. It should be noted that a potential bias exists in the dataset when using the nested case-control design (owing to RNA being extracted for the cases first and then controls) and as such the authors recommend appropriate validation datasets if pursuing a cases vs. controls analysis, further information here (11). We do not utilize this design in the present study.

Preprocessing and Normalization of the Microarray Gene Expression Arrays

The microarray gene expression analysis was done in the open source software R using the aroma.affymetrix package. Each gene expression array was individually background corrected and normalized using robust multichip averaging (RMA).

Cell cycle score (CCS)

The cell cycle score (CCS) was developed in the Metabric dataset (12) in order to derive broadly applicable cut-offs for cell cycle activity. Specifically, we aggregated all genes central to, or associated with, the cell cycle from three different databases (KEGG, HGNC,

Cyclebase) (13–15) and extracted the expression values for these genes (463 in total, see Supplementary Table S1) from all tumours of Metabric (n= 1992). Next, we summed the expression of all genes within each tumour and split the resulting continuous score into tertiles of low, intermediate and high cell cycle activity. For classification of our cohorts the continuous cell cycle score was derived in an identical manner and then scaled to match the Metabric cell cycle score before applying the same tertiles (33% and 66%) in order to classify every tumour into low, intermediate or high cell cycle activity bins. This study makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium. Funding for the project was provided by Cancer Research UK and the British Columbia Cancer Agency Branch.

3. Choice of Ki67 cutoff

We have previously performed an extensive comparison of the prognostic information provided by Ki67 with a range of cutoffs and how this relates to the prognostic information provided by binary and multi-level gene expression signatures (16). These findings indicated that choosing a cutoff between 10 to 19% provided largely consistent hazard ratios. Importantly however, in the same study, we found that applying the St. Gallen cutoff (17) of 20% greatly reduced the prognostic capacity of Ki67 in our hands. We concluded that given differences in staining procedures across pathology labs at a worldwide level, a reasonable cutoff could/should be determined by choosing the median Ki67 staining value for an individual lab. Based on this and additionally on wanting to avoid an overfit in the data by choosing a cutoff that maximizes prognostic capacity, we selected the median Ki67 value as an unbiased cutoff in Cohort 1 (16%). In order to keep methods and cutoffs identical, we also selected 16% as the Ki67 cutoff in Cohort 2.

References

1. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 2002;41:154–61.

2. Tanner M, Gancberg D, Di Leo A, Larsimont D, Rouas G, Piccart MJ, et al. Chromogenic in situ hybridization: a practical alternative for fluorescence in situ hybridization to detect HER-2/neu oncogene amplification in archival breast cancer samples. Am J Pathol. 2000;157:1467–72.

3. Harrell JC, Prat A, Parker JS, Fan C, He X, Carey L, et al. Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse. Breast Cancer Res Treat. 2012;132:523–35.

4. Mittempergher L, de Ronde JJ, Nieuwland M, Kerkhoven RM, Simon I, Th. Rutgers EJ, et al. Gene Expression Profiles from Formalin Fixed Paraffin Embedded Breast Cancer Tissue Are Largely Comparable to Fresh Frozen Matched Tissue. PLoS ONE. 2011;6:e17163.

5. Iddawela M, Rueda OM, Klarqvist M, Graf S, Earl HM, Caldas C. Reliable gene expression profiling of formalin-fixed paraffin-embedded breast cancer tissue (FFPE) using cDNA-mediated annealing, extension, selection, and ligation whole-genome (DASL WG) assay. BMC Med Genomics. 2016;9:54.

6. Musella V, Callari M, Buduo ED, Scuro M, Dugo M, Miodini P, et al. Use of Formalin-Fixed Paraffin-Embedded Samples for Gene Expression Studies in Breast Cancer Patients. PLOS ONE. 2015;10:e0123194.

7. Toussaint J, Sieuwerts AM, Haibe-Kains B, Desmedt C, Rouas G, Harris AL, et al. Improvement of the clinical applicability of the Genomic Grade Index through a qRT-PCR test performed on frozen and formalin-fixed paraffin-embedded tissues. BMC Genomics. 2009;10:424.

8. Sapino A, Roepman P, Linn SC, Snel MHJ, Delahaye LJMJ, van den Akker J, et al. MammaPrint molecular diagnostics on formalin-fixed, paraffin-embedded tissue. J Mol Diagn JMD. 2014;16:190–7.

9. Baehner FL. The analytical validation of the Oncotype DX Recurrence Score assay. ecancermedicalscience [Internet]. 2016 [cited 2017 Aug 18];10. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5045300/

10. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics [Internet]. 2015 [cited 2017 Aug 18];8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4546262/

11. Cunha SI, Bocci M, Lövrot J, Eleftheriou N, Roswall P, Cordero E, et al. Endothelial ALK1 Is a Therapeutic Target to Block Metastatic Dissemination of Breast Cancer. Cancer Res. 2015;75:2445–56.

12. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.

13. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457-462.

14. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43:D1079-1085.

15. Santos A, Wernersson R, Jensen LJ. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res. 2015;43:D1140–4.

16. Tobin NP, Lindström LS, Carlson JW, Bjöhle J, Bergh J, Wennmalm K. Multi-level gene expression signatures, but not binary, outperform Ki67 for the long term prognostication of breast cancer patients. Mol Oncol. 2014;8:741–52.

17. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24:2206–23.