Gene expression modules in primary breast cancers as risk factors for organotropic patterns of first metastatic spread: a case-control study

Katherine Lawler[1,2], Efterpi Papouli[3], Cristina Naceur-Lombardelli[4], Anca Mera[4,5], Kayleigh Ougham[6], Andrew Tutt[7], Siker Kimbung[8,9], Ingrid Hedenfalk[8,9], Jun Zhan[10], Hongquan Zhang[10], Richard Buus[11], Mitch Dowsett[11], Tony Ng[1,7,12,13], Sarah E. Pinder[4], Peter Parker[1,14], Lars Holmberg[5,15], Cheryl E. Gillett[4], Anita Grigoriadis[1,4,6,7*†] and Arnie Purushotham[1,4*†]

[1]School of Cancer Studies, CRUK King's Health Partners Centre, King's College London, Guy's Campus, London SE1 1UL, UK.

[2]Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, Guy's Campus, London SE1 1UL, UK.

[3]NIHR Comprehensive Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London, London WC2R 2LS, UK.

[4]Research Oncology, King's College London, Faculty of Life Sciences and Medicine, Guy's Hospital, London SE1 9RT, UK.

[5]Cancer Epidemiology Unit, King's College London, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK.

[6]Cancer Bioinformatics, King's College London, Innovation Centre, Cancer Centre at Guy's Hospital, London SE1 9RT, UK.

[7]Breast Cancer Now Research Unit, Innovation Centre, Cancer Centre at Guy's Hospital, King's Health Partners AHSC, King's College London, Faculty of Life Sciences and Medicine, London SE1 9RT, UK.

[8]Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden.

[9]CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden.

[10]Key Laboratory of Carcinogenesis and Translational Research, Ministry of Education of Beijing, Beijing, People's Republic of China, Laboratory of Molecular Cell Biology and Tumor Biology, Department of Anatomy, Histology and Embryology, Peking University Health Science Center, Beijing, People's Republic of China.

[11]The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK.

[12]Richard Dimbleby Department of Cancer Research, Randall Division of Cell and Molecular Biophysics, King's College London, Guy's Campus, London SE1 1UL, UK.

[13]UCL Cancer Institute, Paul O'Gorman Building, University College London, London WC1E 6DD, UK.

[14]London Research Institute, Lincoln's Inn Fields, London WC2A 3LY, UK.

[15]Uppsala University, Department of Surgical Sciences, Uppsala University Hospital, 75185 Uppsala, Sweden.

* Correspondence: anita.grigoriadis@kcl.ac.uk; EA-Purushotham@kcl.ac.uk

†Equal contributors

ABSTRACT

Background

Metastases from primary breast cancers can involve single or multiple organs at metastatic disease diagnosis. Molecular risk factors for particular patterns of metastastic spread in a clinical population are limited.

Methods

A case-control design including 1,357 primary breast cancers was used to study three distinct clinical patterns of metastasis which occur within the first six months of metastatic disease: bone and visceral metasynchronous spread, bone-only, and visceral-only metastasis. Whole-genome expression profiles were obtained using WG-DASL assays from formalin-fixed paraffin-embedded (FFPE) samples. A systematic protocol was developed for handling FFPE samples together with stringent data quality controls to identify robust expression profiling data. A panel of published and novel gene sets were tested for association with these specific patterns of metastatic spread and odds ratios (ORs) were calculated.

Results

Metasynchronous metastasis to bone and viscera was found in all intrinsic breast cancer subtypes, while IHC receptor status and specific IntClust subgroups were risk factors for visceral- or bone-only first metastases. Among gene modules, those related to proliferation increased the risk of metasynchronous metastasis (OR (95% C.I.) = 2.3 (1.1-4.8)) and visceral-only first metastasis (OR (95% C.I.) = 2.5 (1.2-5.1)) but not bone-only metastasis (OR (95% C.I.) = 0.97 (0.56-1.7)). A 21-gene module (*BV*) was identified in ER-positive breast cancers with metasynchronous metastasis to bone and viscera (AUC = 0.77), and its expression increased the risk of bone and visceral metasynchronous spread in this population. *BV* was further orthogonally validated with Nanostring nCounter in primary breast cancers, and was reproducible in their matched lymph nodes metastases and an external cohort.

Conclusion

This case-control study of WG-DASL global expression profiles from FFPE tumour samples, after careful quality control and RNA selection, revealed that gene modules in the primary

tumour have differing risks for clinical patterns of metasynchronous first metastases. Moreover, a novel gene module was identified as a putative risk factor for metasynchronous bone and visceral first metastatic spread, with potential implications for disease monitoring and treatment planning.

Keywords: breast cancer; metasynchronous metastases; gene expression pattern.

BACKGROUND

Development of metastatic breast cancer is a complex multi-step process manifesting with diverse temporal patterns involving single or multiple organs [1]. Metastases to multiple bone or visceral sites may be recorded as synchronous (reported at the same time), metasynchronous (where reported metastases are separated by a short time period, typically months) or asynchronous events with a significant delay between distant recurrences [2-4]. The median survival of patients with metastatic breast cancer is 18 to 24 months, although the range in survival spans between a few months to many years and often depends on the pattern or burden of metastatic spread. Most clinicians recommend initial treatment with chemotherapy for rapidly progressive visceral disease or in women with severe symptoms related to metastatic breast cancer [5]. Patients with bone metastases are often treated with osteoclast inhibitors, as these agents have been shown to reduce the risk of skeletal related events such as fractures, the need for surgery or radiation to bone, spinal cord compression, and hypercalcemia of malignancy. However, patients at risk of metasynchronous metastatic spread to bone and viscera may benefit from an alternative treatment strategy at the time of first metastatic presentation. Clinicians may pursue more aggressive pursue more aggressive therapy immediately (e.g. chemotherapy instead of endocrine therapy) for patients with bone metastasis who were at high likelihood of imminent visceral metastasis, or similarly add bone-directed therapy for patients with visceral metastasis who are at high likelihood of imminent bone metastasis. Identification of these patients at an early stage after primary diagnosis or during early metastatic disease is not well established.

Various prognostic factors influence the overall survival of patients with metastatic breast cancer, including hormone receptor status and axillary lymph node status at diagnosis, previous adjuvant chemotherapy, and the number of involved organs [6, 7]. ER-positive disease has a predilection to metastasise to bone, whereas basal-like and claudin-low breast cancers are associated with brain and lung relapses as first site of metastasis and HER2-positive tumours show a predilection to cerebellar metastasis [8-14]. Transcriptional features present in the primary invasive breast carcinoma can be intrinsic to metastatic progression [15] and are currently tested in clinical trials for patient stratification for treatment regimens [16, 17]. IntClust subtypes can stratify patients by disease-specific survival [18] but without attribution to specific metastatic patterns. Recently small-scale studies suggested circulating tumour-derived exosomes to be predictive for metastasis to individual bone or visceral metastatic sites [19]. It remains unclear, however, to what extent the primary tumour at the

time of diagnosis confers risk factors for clinical patterns of disease progression, manifesting as diverse temporal patterns of metastatic spread to single or multiple different organ sites.

With the increasing use of small diagnostic biopsy procedures prior to systemic treatment for cancer patients, there are limited opportunities to also collect frozen tissue. The latter has been the prime resource for genomic analysis and subsequent publication of gene signatures for prognostic and predictive use. However, despite the potentially vast resource available within diagnostic FFPE archives, they have remained largely untapped for exploratory genome-scale biomarker studies. The quality of RNA from FFPE material has been the key limitation on its subsequent use. Whilst modifications to the extraction techniques continue to make slight improvements to the degraded RNA, there have been greater developments in array-based gene expression profiling assays and emerging technologies for transcriptome sequencing from FFPE samples [20]. The Illumina Whole-Genome DASL (WG-DASL) assay is one such assay [21]. Several technical studies reported that DASL assays can produce reliable expression profiles from FFPE tumour tissue samples given adequate RNA quality, design and preprocessing of the resulting data [21-28].

Here, we designed a whole-genome expression profiling study using a case-control design to include primary invasive breast cancers with three clinically observed patterns of metastatic spread: (i) metasynchronous bone and visceral metastases (within six months of first metastasis); (ii) bone only, with delayed or no visceral metastasis; and (iii) viscera only, with delayed or no bone metastasis. Given the comprehensive time and detailed organ-site information in our cohort from which cases and controls were selected, we aimed to identify intrinsic molecular features of primary breast carcinomas associated with the distinct patterns of first metastatic spread which are observed in the clinic, and in particular those with metasynchronous bone and visceral metastases.

MATERIALS AND METHODS

Study population, study design and patient selection

The study population comprised 5,061 patients diagnosed with invasive primary breast cancer without distant metastasis at the time of diagnosis between 1975 and 2005 from Guy's Hospital, London UK. All patients had given consent for analysis of their tumour tissues. Median follow-up was 11yrs (time between entry and exit dates in the case-control design).

Three metastatic populations were defined according to the site and specificity of recurrence: (i) first recurrence to bone only, with no other metastatic site within six months ("*bone only*"), (ii) first recurrence to viscera (other organs) only, with no other metastatic site within six months ("*visceral only*"), and (iii) recurrence to bone and viscera within a period of six months ("*bone & visceral*"). Bone metastasis was defined as distant metastasis to bone or bone marrow, spinal cord compression, pathological fracture or hypercalcaemia. Visceral metastasis included distant metastasis to lung, liver, brain or ascites. In the study population, 1,598/5,061 (32%) developed distant metastasis: 413/1,598 (26%) to bone as first site ("*bone-only*"), 747 (47%) to visceral as first site ("*visceral-only*"), and 438 (27%) to bone and visceral within a six month period ("*bone & visceral*").

For each of the three metastatic populations ("*visceral-only*", "*bone & visceral*", "*bone-only*"), individually matched controls were randomly sampled using a case-control incidence based approach (Supplementary Methods). Briefly, each calendar time, *T* (e.g. 15th January 1999) that a case is diagnosed, one or more controls are randomly selected from the other members of the cohort who, at the time *T*, are still at risk of developing the outcome (distant metastasis). The controls are therefore matched to the case by time of event. A patient who is a control at one time can later become a case and/or a control again, and each of the control series therefore includes a combination of patients still at risk which enables efficient estimation of risks in the clinical population (among 1,200 case-control sets, 75% of patients selected as controls at any calendar time did not metastasise at all, and 25% went on to have a distant metastatic event). 400 case-control sets were randomly selected for each of the three metastatic populations, giving a total of 1,200 selected case-control sets. Case-control sets were then selected for tissue assessment and RNA extraction from formalin-fixed, paraffin embedded (FFPE) tissue blocks. Extracted RNA was available for a total of 742 case-control (1:1) pairs: 246 case-control pairs for "*visceral only*" cases, 258 for "*bone & visceral*" cases, and 238 for "*bone only*" cases. An overview diagram of patient cases is

shown in Figure 1 and a detailed overview of case-control sampling, random selection and extracted case-control pairs is shown in Supplementary Figure S1A&B. Supplementary Table 1 tabulates the number of cases with extracted case:control pairs (1:1), and the number of matched and unmatched cases and controls with available gene expression data.

RNA extraction and gene expression profiling

Formalin-fixed paraffin embedded (FFPE) samples of breast carcinomas were microdissected following tissue review. A total of 1,575 FFPE tissue blocks were assessed (H&E; Supplementary Methods). Primary tumour blocks from 1,357 patients were taken forward to micro-dissection and RNA extraction, with a total of 1,370 RNA samples (Supplementary Figure 1A). In addition, RNA was extracted from 100 matched positive axillary lymph nodes and from FFPE samples of six breast cancer cell lines. RNA extraction was outsourced to Gen-Probe Life Sciences Ltd (Manchester, UK). RNA sample quality, quantity and integrity were assessed before proceeding to Illumina HT-12 v4 BeadChips WG-DASL microarray. A detailed description of tissue selection, microdissection, RNA sample selection, hybridisation design and microarray data processing is provided in Supplementary Methods. Two gene expression data sets were produced following rigorous quality assessment: *GWDb* (containing primary tumour samples from 527 patients) and *GWDa* (containing primary tumour samples from 124 patients, after removing patients also present in *GWDb*). Patient characteristics for *GWDb* are provided by case-control series in Table 2. An overview diagram of each data set is provided in Supplementary Figure 1A. Gene expression microarray data has been deposited to Array Express E-MTAB-4003.

Intrinsic subtype assignments and gene module scores

PAM50 intrinsic subtype was assigned in accordance with Weigelt *et al*. [29] using median-centred data and matching probes to centroid identifiers via gene symbol. The nearest centroid by Spearman correlation was assigned to each sample. IntClust subtypes were assigned using the iC10 package (v1.1.2) [30] for R/Bioconductor. Gene module scores for a panel of previously reported gene modules were estimated using the DART method [31] and further compared with weighted sum (weights (+1,-1) according to the direction of expression in the gene signature) (Supplementary Methods). Previously reported gene expression signatures were mapped to WG-DASL probes using Ensembl Gene ID, Entrez Gene ID or gene symbol, according to their original source (Supplementary Table S2). Where

multiple microarray probes mapped to a single Entrez Gene ID, the probe with the most variable gene expression across the datasets was used (based on standard deviation in the relevant data set).

Derivation and expression summary of gene module

To identify a candidate gene module for *'bone & visceral'* metastasis (*'BV'* gene module), *GWDb* was reduced to ER-positive case-control pairs, and top-ranked genes were identified using an exploratory differential expression analysis of *'bone & visceral'* and *'No metastasis'* groups, as follows. Top-ranked genes were identified by comparing *'bone & visceral'* vs. *'No metastasis'* (threshold FDR-adjusted $P < 0.2$; Mann-Whitney *U*-test). This procedure resulted in a list of 21 genes (19 up, 2 down) together with the direction of differential expression between the *'bone & visceral'* and *'No metastasis'* groups (up, down) which defines the *BV* gene module. To inspect the candidate *BV* module within expression data sets, the expression of the gene module was summarised using a weighted sum with weights (+1,-1) according to the original direction of differential expression in the gene module.

NanoString gene expression analysis

For a subset of 192 samples, expression was validated for 150 selected genes by analysing total RNA (200 ng) with the nCounter platform (NanoString Technologies). Expression data were normalised using the NanoStringNorm package in R [32]. Background correction was done by subtracting the negative control probes ('mean.2sd'). Expression values were normalised to the geometric mean of fifteen housekeeping genes. Expression values were $\log_2$ transformed and standardised within each sample (geometric mean). An expression score for the *BV* gene module was calculated among ER-positive samples using a weighted sum (weights (+1,-1) according to direction in the *BV* module) of mean-centred, standard deviation-scaled *BV* genes.

Statistical analysis

For each case-control series, conditional logistic regression models (modelling individually matched pairs) and logistic regression models (unconditional, disregarding the case-control matching) were used to estimate odds ratios (ORs) and 95% confidence intervals (CIs). For intrinsic molecular subtypes, ORs were estimated for each subtype compared with a baseline

subtype. IHC-derived subtypes were compared with ER-positive, HER2-negative tumours [33]. PAM50 subtypes were compared with the 'LuminalA' subtype (a good prognosis group [34]). IntClust subtypes were compared with the baseline IntClust3 cluster [30]. Gene module scores were scaled within each case-control series so that 95% of values lay within the range [-1,1] [35]. FDR/Benjamini-Hochberg multiple testing correction was applied to p-values across the panel of reported gene modules within each test [36]. A quartile analysis, in which cases were binned according to the quartile thresholds of the respective control series and conditional logistic regression models fitted for each quartile compared with the first quartile, showed similar trends in OR to the models which treat gene module scores as continuous variables (not shown). The Wilcoxon test for matched pairs and Mann-Whitney *U* test (unpaired) was used to test for differences in gene module scores between cases and controls in each series. All statistical analysis was conducted in the R environment (v3.1.2) (www.r-project.org). Conditional regression models were fitted using the function *clogistic()* in the package *Epi* (v2.0) [37] (*Case_Control ~ x + strata(pair.id)*). Logistic regression models were fitted using function *glm(family='binomial')* in the base package *stats*. A Sweave document is provided in Supplementary Methods.

RESULTS

Patient characteristics and sample processing

Clinico-pathological information for extracted RNA samples from 742 cases and their case-matched controls are summarised in Table 1. A rigorous inspection of extracted RNA and WG-DASL data was performed to ensure that expression profiles could be obtained across the span of storage times and inferior quality data were excluded from further analysis (Supplementary Figures S5-7). Primary tumour samples, which passed rigorous WG-DASL quality controls, were assigned to a discovery set (*GWDb*, 527 patients) or a smaller independent data set (*GWDa*, 124 patients) (Supplementary Figure S1A&B). Clinico-pathological information for the three case-control series in data set *GWDb* is shown in Table 2. Before proceeding to the analyses of the three case-control series, the clinico-pathological characteristics for each set were inspected. Patient characteristics of *GWDb* and *GWDa* retained the originally selected distribution of organ-specific metastatic spread (Supplementary Table S1A lists the number of cases for the case:control pairs (1:1), and the number of matched and unmatched cases and controls with available gene expression data). On inspection of *GWDb*, there were predominantly immunohistochemically (IHC)-defined ER-positive (~68%), 18% IHC HER2-positive and 32% IHC ER-negative breast cancers. Primary carcinomas were predominantly treatment naïve and invasive ductal carcinoma of no-special histological type. Patients with a '*bone only'* pattern of first metastatic spread were more likely to report a visceral metastasis beyond a 6 month period from first metastasis (37%), than the *'visceral only'* group to a later bone metastasis (16%) (Supplementary Table S1B). The *'visceral only'* case series had a greater proportion of grade 3 primary tumours and a smaller proportion receiving endocrine therapies than the other two case series in *GWDb*, while the *'bone only'* case series had the lowest proportion of patients treated with chemotherapy (Table 2). Supplementary Figure S2 provides an illustration and descriptive summary of the temporal patterns of the single and multiple sites of metasynchronous metastatic spread present amongst all carcinomas in *GWDb*.

Metastatic spread among breast cancer subtypes

Next we asked to what extent patients with particular molecular subtypes of breast cancer, as currently defined in the research setting, were at risk of metasynchronous bone and visceral, bone only, or visceral only patterns of first metastasis observed in the clinic.

Molecular subtypes in *GWDb* set were defined by the IHC status of ER/HER2 [33], assigned to the PAM50 [34] and the IntClust subtypes [18, 30]. Initially, PAM50 estimates for each tumour were compared with IHC-defined subtypes and overall a good accordance was observed: 'Luminal A' samples were 89% IHC ER-positive; 62% of 'basal-like' cases were of triple-negative phenotype (IHC ER-, PgR- and HER2-negative; TNBC); and 73% of 'HER2-enriched' samples were IHC HER2-positive among samples with available IHC status.

Second, the molecular subtypes were tested within each case-control series using conditional logistic regression (Table 3). IHC ER-positive patients showed increased risk of *'bone only'* and decreased risk of *'visceral only'* and *'bone & visceral'* metastatic spread. In *GWDb*, breast cancer patients of the 'HER2-enriched' PAM50 subtype showed increased risk of '*visceral only*', and 'Luminal B' of *'bone only'*, compared to the 'Luminal A' baseline. Patients with tumours classified as IntClust5 had an increased risk of *'visceral-only'* spread compared to IntClust3 baseline tumours. Unconditional logistic regression models had similar OR point estimates (Table 4). The IHC ER-negative/HER2-positive subtype showed an increased risk of '*visceral only*' and '*bone & visceral*' spread compared to ER-positive/HER2-negative carcinomas. Among IntClust classes with IntClust3 as the reference class, '*bone & visceral'* had similar risks to the *'visceral only'* group with the exception that IntClust2 and IntClust4 showed increased risk for *'visceral only'* and *'bone only'* but not *'bone & visceral'*. Subtypes found to be risk factors for *'bone & visceral*' spread were also risk factors for *'visceral only'* or *'bone only'* events from either the conditional or unconditional logistic regression models, indicating that molecular subtypes do not confer risks specifically for '*bone & visceral*' events in this study.

Third, the tumour molecular subtypes of all patients in *GWDb* were tabulated and compared to the metastatic pattern of every patient irrespective of the case-control design, with the aim of providing a descriptive overview [30] of all primary tumours in *GWDb* (Figure 2). As predicted, IHC ER-negative and HER2-positive tumours were enriched for the '*visceral only'* cases, and TNBC was decreased for *'bone only'* events (Figure 2, IHC). The breakdown of IHC subtypes in the *'bone & visceral'* group lies in between the *'bone only'* and *'visceral only'* groups and does not appear to be dominated by the IHC associations that would be expected for one case type or the other (*'bone only'* or *'visceral only'*). The differences were less clear across the PAM50 subgroups (Figure 2, PAM50). IntClust5 had the highest prevalence of '*visceral only*' events, whereas IntClust9 followed by IntClust6 had the most '*bone &*

visceral' events (Figure 2, IntClust). The *'bone only'* group were mainly ER-positive breast cancers and were predominantly assigned to IntClust3, 4 and 7 subtypes. In contrast, patients with no reported metastases were enriched for Luminal A and IntClust3 subtypes (IntClust in '*NoMetastasis'* group; χ^2, *P<1e-5*), and *'bone & visceral'* events were present in all IntClust subtypes (IntClust in '*bone & visceral*'; χ^2, *P:0.3*). In summary, primary breast cancers with *'bone & visceral'* metasynchronous metastatic pattern were found across multiple molecular breast cancer subtypes in *GWDb*, indicating that there was no evidence of increased risk specifically for '*bone & visceral*' events among these current breast cancer classifications (Tables 3&4).

Prognostic gene modules are indicative of organ-specific metastatic predilection

Several studies have reported gene expression modules indicative of particular organ-specific metastatic spread (reviewed in [38, 39]). We therefore asked whether some of those modules were also activated across our three metastatic groups and to what extent the activation of individual gene modules in the primary tumour is a risk for the clinically observed patterns of metasynchronous *'bone & visceral'*, '*bone only'*, or *'visceral only'* first metastasis. Primary tumour expression modules were selected if they were previously reported to be associated with: (i) features of proliferation, cell motility, presence of stem-cell-like cells and immune/lymphocytic infiltration, and (ii) organ-specific metastasis (Supplementary Table S2).

The gene modules were each tested as a risk for each pattern of metastatic spread within *GWDb* using conditional logistic regression (Figure 3A) and logistic regression on complete case-control pairs (Figure 3B). In addition, logistic regression models were fitted using all controls and all cases (to avoid discarding both samples from a pair due to missing data from *GWDb*), and using ER-stratified data with or without discarding samples from ER-mismatching pairs (Supplementary Figure S3). Pairwise correlation of gene modules confirms that proliferation signatures are highly correlated in this data set and there is also correlation between other modules previously reported to represent metastasis to individual sites and between immune-related signatures (Figure 3C), consistent with other studies [35, 40]. Expression of proliferation-associated genes have been repeatedly shown to be associated with the prognosis of ER-positive breast cancers [41]. In our study, gene modules

related to proliferation increased the risk of metasynchronous bone and visceral metastasis (for example, conditional logistic regression model of *PTEN* module, OR (95% C.I.) = 2.3 (1.1-4.8); Supplementary Table S3) and *'visceral only'* first metastasis (*PTEN*, OR (95% C.I.) = 2.5 (1.2-5.1); Supplementary Table S3) but not *'bone only'* metastasis (*PTEN*, OR (95% C.I.) = 0.97 (0.56-1.7); Supplementary Table S3). Risk associations were observed by logistic regression modelling within ER-positive or ER-negative tumours (Supplementary Table S3, Supplementary Figure 3A).

In addition, with the alternative aim of providing a descriptive summary of gene module activation among all breast carcinomas present in *GWDb*, two exploratory analyses were performed irrespective of the case-control design: tumours with each pattern of metastatic spread were compared with all tumours with no metastasis using a logistic regression model (Supplementary Figure S3B), and time-to-metastasis to individual sites (lung, liver, bone, brain) was modelled irrespective of the patterns of first metastatic spread or any other metastases at any time during follow-up (Supplementary Figure S3C&D). In agreement with other studies, neither the logistic regression with reference to tumours with no metastasis nor the time-to-event analyses can be interpreted in the standard epidemiological sense estimating associations between exposures and outcomes, due to the sample selection methods employed in this study. These models are presented here as exploratory hypothesis-generating tools only with no inference implied for the breast cancer population. Metastasis to any site was associated with proliferation signatures irrespective of ER status. (Supplementary Figure S3C). A number of gene modules indicated nominal significance but would not pass a multiple testing correction (Supplementary Fig 3). In IHC ER-negative breast cancers TGF-β response [42] and hypoxia response gene sets [43] were activated in carcinomas with '*visceral only*' metastases compared with those which did not metastasise (Supplementary Figure S3B), while TGF-β response and hypoxia response gene sets showed an association with time-to-lung metastasis (Supplementary Figure S3D). A stem cell module which is a strong indicator of short relapse in TNBC [44] was present in ER-negative breast cancers with '*visceral only'* metastases, and a module related to intermediate tissue burden and progression from stemness/basal-like cells [45] was associated with the '*visceral only'* and '*bone & visceral'* groups, while the low tissue-burden/basal-like module (derived from metastatic cells from tissues with low metastatic burden [45]) was underexpressed in the '*bone only*' group compared with cancers with no reported metastases. Taken together, while we were able to recapitulate previously reported associations between

gene signatures and metastasis to bone or visceral organs within our study, there were no specific gene module with distinctive risk factors for the clinically observed metasynchronous '*bone & visceral'* spread group.

A prognostic gene module for metasynchronous metastatic spread

As the question remains whether any molecular features could be identified in primary tumours at the time of diagnosis for metasynchronous *'bone & visceral'* metastases, we then aimed to extract specific gene expression patterns associated with the *'bone & visceral'* metastatic group. In order to control for interactions between ER status and metastatic group, the discovery set (*GWDb*) was stratified by ER status by removing those case-control pairs with differing IHC ER status (Supplementary Methods). We proceeded with a total of 175 ER-positive breast cancer patients. Exploratory differential expression analysis was performed between the *'bone & visceral*' group and tumours with no metastasis and the top-ranked genes (FDR-adjusted $p < 0.2$; see Methods) was taken forward for further exploratory analysis (Supplementary Table S4). ROC curve analysis revealed an AUC of 0.77 for a 21 gene set for the '*bone & visceral'* metastatic group (termed '*BV*'; Figure 4A, Supplementary Table S4), in comparison to an AUC of 0.66 and 0.56 for '*visceral only'* and *'bone only'*, respectively, while combining all three series indicated an overall AUC of 0.83 for any metastatic site (Supplementary Figure S4).

The risk of *'bone & visceral'* spread from *BV* gene module was next estimated using the *'bone & visceral'* case-control series within *GWDb* and *GWDa* (Supplementary Table S5). A significant risk of '*bone & visceral'* spread was observed within *GWDb* (OR (95% C.I.) = 6.0 (3.1-12.2). In the independent data set *GWDa*, OR estimate were also positive (OR (95% C.I.) = 1.9 (0.38-9.7), and a shift towards increased *BV* scores in the cases compared to the controls were found (Mann-Whitney *U,* p=0.3), however due to small sample size it was not significant (Supplementary Table S5). Together these results indicated that this *BV* gene module might confer an increased risk of the *'bone & visceral'* pattern of metastatic spread.

To further explore the relevance of our *BV* gene module, orthogonal validation of the discovery was obtained on three levels: (i) with NanoString, by testing the expression of these genes in a representative subset of 192 samples, in which *BV* scores were highly

concordant with WG-DASL values (Figure 4B; Pearson's correlation, cor=0.78, *P*<1e-4); (ii) the expression of the *BV* gene set was reproduced in matched lymph node metastases within our cohort (Figure 4C); and (iii) we investigated the gene expression of the *BV* module in an external data set of lymph node metastases from breast cancers with known metastatic disease [46]. The *BV* gene module exhibited increased expression in the lymph node metastases of those patients with a *'bone & visceral'* pattern of first metastatic spread compared with *'visceral only'* (Figure 4D; Mann-Whitney *U* test, *P:0.04*) and *'bone only'* (Figure 4D; Mann-Whitney *U* test, *P:0.1*, n.s.) groups. These results are in line with our exploratory analyses and together are the first demonstration towards developing an intrinsic risk factor in primary breast cancer for metasynchronous bone and visceral first distant recurrences. Inspection of the genes comprising the candidate *BV* module indicated an enrichment for association with condensing chromosomes and the kinetochore (Supplementary Figure S8).

DISCUSSION

Metastasis represents the major cause of death for breast cancer patients. Over the last few years, numerous molecular-based prognostic tests of varying specificity have emerged, indicating that primary breast carcinomas display expression profiles associated with organ-specific dissemination, however few studies have addressed synchronous and metasynchronous patterns of metastatic spread [47]. Treatment strategies and monitoring for patients could potentially be tailored if prediction of single or multi-organ metastasis could be estimated at an early stage. As a step towards this goal, this study estimates potential risks for particular patterns of metastatic spread associated with intrinsic subtypes and gene modules in the primary tumour. A gene expression module present in primary invasive carcinomas associated with concurrent or short-term delays between the development of bone and visceral metastasis was identified and validated in an independent series of lymph node metastases.

Predilection for metastatic spread for breast cancer has previously been associated with gene modules enriched in the primary tumour. A common feature of these are markers of cell proliferation, such as the GENE70, PTEN, the centrosomal kinase *AURKA* [48, 49] as well as multiple processes related to chromosomal instability including CIN70 [50]. In this study, we

found that a gene module containing components of the kinetochore (*CENPO, SPC25, CASC5, SKA3, CENPE*; GO CC term 'kinetochore') was associated with the occurrence of metasynchronous bone and visceral metastases within six months of the first metastasis. The regulation of genes encoding kinetochore components has been hypothesised to drive chromosomal instability [51], whereby the upregulation of kinetochore genes may reflect the activation of a cell division program [52]. We speculate that the association of this gene module with rapid multitropic bone and visceral spread after first metastasis points to a mechanism of chromosomal instability, enabling the development of subclones and selection of metastatic tumour cells for invasion and adaption at multiple bone and visceral sites. Gene modules related to proliferation or mammary stem cells might be expected to influence the synchronicity of multiple metastases, and were indeed found to be significant risks for multiple bone and visceral first metastases.

Limitations of this study include the imposition of a timeline which defines metasynchronous metastases: for example, in our datasets a change in the definition of metasynchronous from 6-month to 12-month would have led 10% of *'bone only'* and 5% of *'visceral only'* to have been considered metasynchronous (Supplementary Figure 2A), and other definitions of metasynchronicity could be imposed which may affect the estimated risks for each case type. The AUC for the *BV* signature was higher for all metastasis than for metasynchronous bone and visceral metastasis: from the point of view of clinical translation, further work would need to establish whether *BV* or other putative signatures for patterns of metastatic spread could add value over existing signatures such as Oncotype or Mammaprint [16, 17].

Metastasis is a complicated, multi-step process and our understanding of the multiple factors involved is still partial. In the last decade, genomic profiling has attempted to fill this knowledge gap, however these studies have primarily used fresh-frozen tissue, had restricted numbers of primary and metastatic cases, and incomplete information on the site and time to development of the metastatic spread. This has limited the utility and clinical applicability of these modules. There is evidence that some tumours have a predilection for colonising specific tissues in clinical populations (e.g. [10]), while animal model and recent next generation sequencing studies also support a role for subclonal adaptations to the metastatic niche (e.g. [53]). In this study, we focused on the tumour as one part of this complex metastatic cascade which is close to clinical diagnostic practise and patient management. We hypothesised that intrinsic subtypes and gene modules confer risk of particular patterns of

metastatic spread in some tumours. We addressed this question by designing a case-control study for particular patterns of metastatic spread.

Pre-clinical models have contributed to our understanding of metastastic spread but they might not capture many of the processes which are important in a clinical setting - including alterations to the immune system or incorporation of specific latency periods to study multi-organ metastatic spread in parallel - and many gene signatures originate from ER-negative cell line and PDX models. Multiple lines of evidence indicate that intrinsic subtypes and gene modules have different metastatic potential within clinical populations [18, 54, 55]. We sought to address whether gene modules could confer risk for specific temporal patterns of metastatic spread using a large tumour archive with detailed clinical follow up. An efficient case-control design is required given that multiple breast cancer subtypes and metastatic patterns are present in any clinical population. This study therefore focused on three specific patterns of metastatic spread based on epidemiological observations from the same clinical population [56].

As recently shown by Iddawela *et al.* [28] and others, the WG-DASL platform together with stringent quality control and data processing steps can produce reliable results from FFPE breast tumour tissue. Here, by starting with a well characterised cohort of 5,061 patients with long-term follow up of whom 1,598 developed metastasis, we have created a well-annotated and sufficiently large cohort to investigate molecular risk factors for single and multiple organotropic metasynchronous metastatic disease. While many samples were excluded during the quality control steps, the processed gene expression data sets enabled an investigation of gene expression modules as risk factors for specific patterns of first metastatic spread.

The use of an incidence-based case-control design ensured that non-metastasising primary tumour samples were also included from the same range of calendar dates of diagnosis, and enabled efficient estimation of effects in a standard clinical population. Cunha *et al.* [57] recently reported using a case-control design to estimate the effect of *ALK1* expression in frozen breast tumours as a molecular risk factor for metastatic spread. Here, our use of an incidence-based case-control design enabled the estimation of genome-wide molecular risk factors for three clinically observed patterns of first metastatic spread (bone only, visceral only, or bone and visceral within a six month time period). Further work on patterns of long-term disease progression may focus on a more defined population, such as

ER-positive/HER2-negative or high grade tumours, and make use of appropriate platforms for assaying fewer FFPE tumour samples from a stratified clinical population.

Patients at risk of metasynchronous bony and visceral metastases could potentially profit from closer disease monitoring and may benefit from a more radical bisphosphonate and chemotherapeutic combination treatment strategy up front in the metastatic setting. In ER-positive disease the benefits of hormonal therapy in managing visceral metastases from breast cancer are much lower than those offered by chemotherapy. Conversely, Perez *et al.* [58] suggest that, for patients with bone metastases, efforts should be made to select the least aggressive therapy to avoid excessive toxicity. Progress towards the optimisation of disease monitoring and treatment strategies fundamentally requires a better understanding of risks which could be estimated at primary diagnosis, together with an improved understanding of additional emerging risks as breast cancers progress to a metastatic setting (for example, establishment of a metastatic niche). Initiatives such as AURORA, a large multinational, collaborative metastatic breast cancer molecular screening programme [59] will further shed light on our knowledge of whether the gene expression patterns found in primary breast cancers is similar to those in metastatic material. This would add to our understanding of the metastatic process as well as guide treatment regimens for metastatic breast cancer. A liver-selective gene module was among the set of gene modules, where these genes were under-expressed in primary tumours from patients who subsequently developed liver metastases, but displayed high expression in the liver metastases [46]. We observed the low-level expression of these 17 genes in ER-positive cancers, which metastasised to the liver. Of note, this inverse correlation in the direction of transcript levels of some genes between primary tumours and metastases was not unique to this gene module, but was also seen in another gene module associated with infiltration of the blood-brain barrier [60] and has also been recently reported in ovarian cancers [61]. Due to the scarcity of available samples, a clear biological conclusion cannot be drawn about such inverse correlation and we hypothesise that the process of adaption to the new microenvironment or development of the pre-metastatic niche [62, 63] favours those primary tumour traits. Ultimately, larger cohorts with matched primary and metastatic lesions are required to elucidate the clinical relevance of these transcriptional changes after primary diagnosis.

CONCLUSIONS

In conclusion, by analysing gene expression profiles in a large cohort of well characterised primary breast carcinomas and lymph node metastases in patients with long-term follow-up and detailed metastatic spread information, we were able to identify patterns informative of multiple organotropic metasynchronous metastatic spread. Further investigations are necessary in order to tease out the contributing components, which could be relevant for tailoring systemic therapeutic regimens, monitoring response or resistance to therapy, and warranting close imaging/biomarker followup with the institution of early intervention strategies as required. These genomewide expression data across extensive case-control series will provide a useful resource facilitating further studies of the biology and clinical relevance of single and metasynchronous metastatic spread, and might enable rational personalised treatment strategies to be developed for patients at risk of bone or visceral metastasis with subsequent metasynchronous metastasis.

ABBREVIATIONS

FFPE: formalin-fixed, paraffin-embedded; ER: estrogen receptor; PgR: progesterone receptor; HER2: human epidermal growth factor receptor 2. IHC: Immunohistochemistry. TNBC: triple-negative breast cancer. BH: Benjamini-Hochberg method for multiple testing correction.

DECLARATIONS

Ethical Approval and Consent to Participate

Ethical approval was obtained from King's Health Partners Cancer Biobank, London, UK, a HTA Licensed (12121) and NHS REC approved Tissue Bank (12/EE/0493). All patient samples were pseudoanonymised by the Biobank.

Consent for publication

Not applicable

Availability of data and material

Gene expression microarray data has been deposited to Array Express E-MTAB-4003.

Competing interests

The authors declare no competing interests.

Funding

Patient samples and data were provided by King's Health Partners Cancer Biobank, which is supported by the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award and the Experimental Cancer Centre at King's College London. KL was supported by the CRUK and EPSRC Comprehensive Cancer Imaging Centre at KCL and UCL (C1519/A10331 and C1519/A16463) and European Union Framework Programme 7 HEALTH-2010 grant entitled 'Imagint' (grant number 259881). AG and AT were supported by the Breast Cancer Now Research Unit (former Breakthrough Breast Cancer Research) funding at King's College London and by the

National Institute for Health Research Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.

Authors' contributions

KL participated in data analysis, interpretation, and helped to draft the manuscript.
EP performed the WG-DASL assays and participated in data analysis.
CN participated in acquisition of RNA samples.
AM participated in acquisition of follow-up data.
KO participated in data analysis and archiving.
AT participated in study concept and design, and data interpretation.
SK participated in design and data acquisition for the independent cohort.
IH participated in design and data acquisition for the independent cohort.
JZ participated in interpretation of the data.
HZ participated in interpretation of the data.
RB participated in NanoString data acquisition.
MD participated in NanoString data acquisition and critical reading of the manuscript.
TN participated in study concept and design, and data interpretation.
SEP participated in study concept and design, and acquisition of RNA samples.
PP participated in study concept and design, and data interpretation.
LH designed the case-control series, and participated in the design of data analyses and interpretation.
CEG participated in study concept and design, acquisition of RNA samples, and helped to draft the manuscript.
AG participated in study concept and design, design of data analyses, and wrote the manuscript.
AP conceived the study, participated in study design and interpretation, and wrote the manuscript.
All authors read and approved the final manuscript.

Acknowledgements

The authors thank Dr. O. Agbaje for assistance with case-control selection and description of the protocol, Dr. S. Irshad for helpful comments on clinical perspectives, and anonymous reviewers for their constructive and insightful comments.

REFERENCES

1. Kimbung S, Loman N, Hedenfalk I: **Clinical and molecular complexity of breast cancer metastases**. *Semin Cancer Biol* 2015, **35**:85-95.
2. Plunkett TA, Smith P, Rubens RD: **Risk of complications from bone metastases in breast cancer: implications for management**. *Eur J Cancer* 2000, **36**(4):476-482.
3. Coleman RE: **Clinical features of metastatic bone disease and risk of skeletal morbidity**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2006, **12**(20 Pt 2):6243s-6249s.
4. Marlow R, Honeth G, Lombardi S, Cariati M, Hessey S, Pipili A, Mariotti V, Buchupalli B, Foster K, Bonnet D *et al*: **A novel model of dormancy for bone metastatic breast cancer cells**. *Cancer Res* 2013, **73**(23):6886-6899.
5. Wilcken N, Hornbuckle J, Ghersi D: **Chemotherapy alone versus endocrine therapy alone for metastatic breast cancer**. *Cochrane Database of Systematic Reviews* 2003.
6. Largillier R, Ferrero JM, Doyen J, Barriere J, Namer M, Mari V, Courdi A, Hannoun-Levi JM, Ettore F, Birtwisle-Peyrottes I *et al*: **Prognostic factors in 1,038 women with metastatic breast cancer**. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* 2008, **19**(12):2012-2019.
7. Alanko A, Heinonen E, Scheinin T, Tolppanen EM, Vihko R: **Significance of estrogen and progesterone receptors, disease-free interval, and site of first metastasis on survival of breast cancer patients**. *Cancer* 1985, **56**(7):1696-1700.
8. Lai R, Dang CT, Malkin MG, Abrey LE: **The risk of central nervous system metastases after trastuzumab therapy in patients with breast carcinoma**. *Cancer* 2004, **101**(4):810-816.
9. Harrell JC, Prat A, Parker JS, Fan C, He X, Carey L, Anders C, Ewend M, Perou CM: **Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse**. *Breast Cancer Research and Treatment* 2012, **132**:523-535.
10. Kennecke H, Yerushalmi R, Woods R, Cheang MC, Voduc D, Speers CH, Nielsen TO, Gelmon K: **Metastatic behavior of breast cancer subtypes**. *J Clin Oncol* 2010, **28**(20):3271-3277.
11. Metzger-Filho O, Sun Z, Viale G, Price KN, Crivellari D, Snyder RD, Gelber RD, Castiglione-Gertsch M, Coates AS, Goldhirsch A *et al*: **Patterns of Recurrence and outcome according to breast cancer subtypes in lymph node-negative disease: results from international breast cancer study group trials VIII and IX**. *J Clin Oncol* 2013, **31**(25):3083-3090.
12. Sihto H, Lundin J, Lundin M, Lehtimaki T, Ristimaki A, Holli K, Sailas L, Kataja V, Turpeenniemi-Hujanen T, Isola J *et al*: **Breast cancer biological subtypes and protein expression predict for the preferential distant metastasis sites: a nationwide cohort study**. *Breast cancer research : BCR* 2011, **13**(5):R87.

13. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JGM, Foekens Ja, Martens JWM: **Subtypes of breast cancer show preferential site of relapse**. *Cancer Research* 2008, **68**:3108-3114.
14. Vaz-Luis I, Ottesen RA, Hughes ME, Marcom PK, Moy B, Rugo HS, Theriault RL, Wilson J, Niland JC, Weeks JC *et al*: **Impact of hormone receptor status on patterns of recurrence and clinical outcomes among patients with human epidermal growth factor-2-positive breast cancer in the National Comprehensive Cancer Network: a prospective cohort study**. *Breast cancer research : BCR* 2012, **14**(5):R129.
15. Mittempergher L, Saghatchian M, Wolf DM, Michiels S, Canisius S, Dessen P, Delaloge S, Lazar V, Benz SC, Tursz T *et al*: **A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences**. *Molecular Oncology* 2013, **7**:987-999.
16. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, Geyer CE, Jr., Dees EC, Perez EA, Olson JA, Jr. *et al*: **Prospective Validation of a 21-Gene Expression Assay in Breast Cancer**. *N Engl J Med* 2015, **373**(21):2005-2014.
17. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, Pierga JY, Brain E, Causeret S, DeLorenzi M *et al*: **70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer**. *N Engl J Med* 2016, **375**(8):717-729.
18. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y *et al*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups**. *Nature* 2012:1-7.
19. Hoshino A, Costa-Silva B, Shen T-L, Rodrigues G, Hashimoto A, Tesic Mark M, Molina H, Kohsaka S, Di Giannatale A, Ceder S *et al*: **Tumour exosome integrins determine organotropic metastasis.** *Nature* 2015, **527**:329-335.
20. Sinicropi D, Qu K, Collin F, Crager M, Liu ML, Pelham RJ, Pho M, Dei Rossi A, Jeong J, Scott A *et al*: **Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue**. *PLoS ONE* 2012, **7**(7):e40092.
21. April C, Klotzle B, Royce T, Wickham-Garcia E, Boyaniwsky T, Izzo J, Cox D, Jones W, Rubio R, Holton K *et al*: **Whole-genome gene expression profiling of formalin-fixed, paraffin-embedded tissue samples**. *PLoS ONE* 2009, **4**(12):e8162.
22. Abramovitz M, Ordanic-Kodani M, Wang Y, Li Z, Catzavelos C, Bouzyk M, Sledge GW, Jr., Moreno CS, Leyland-Jones B: **Optimization of RNA extraction from FFPE tissues for expression profiling in the DASL assay**. *BioTechniques* 2008, **44**(3):417-423.
23. Ravo M, Mutarelli M, Ferraro L, Grober OM, Paris O, Tarallo R, Vigilante A, Cimino D, De Bortoli M, Nola E *et al*: **Quantitative expression profiling of highly degraded RNA from formalin-fixed, paraffin-embedded breast tumor biopsies by oligonucleotide microarrays**. *Lab Invest* 2008, **88**(4):430-440.
24. Reinholz MM, Eckel-Passow JE, Anderson SK, Asmann YW, Zschunke MA, Oberg AL, McCullough AE, Dueck AC, Chen B, April CS *et al*: **Expression profiling of formalin-fixed paraffin-embedded primary breast tumors using cancer-specific and whole genome gene panels on the DASL(R) platform**. *BMC Med Genomics* 2010, **3**:60.
25. Saleh A, Zain RB, Hussaini H, Ng F, Tanavde V, Hamid S, Chow AT, Lim GS, Abraham MT, Teo SH *et al*: **Transcriptional profiling of oral squamous cell carcinoma using formalin-fixed paraffin-embedded samples**. *Oral Oncol* 2010, **46**(5):379-386.
26. Waddell N, Cocciardi S, Johnson J, Healey S, Marsh A, Riley J, da Silva L, Vargas AC, Reid L, kConFab *et al*: **Gene expression profiling of formalin-fixed, paraffin-embedded familial breast tumours using the whole genome-DASL assay**. *J Pathol* 2010, **221**(4):452-461.
27. Mittempergher L, de Ronde JJ, Nieuwland M, Kerkhoven RM, Simon I, Rutgers EJT, Wessels LFa, Van't Veer LJ: **Gene expression profiles from formalin fixed paraffin embedded breast cancer tissue are largely comparable to fresh frozen matched tissue.** *PloS ONE* 2011, **6**:e17163.
28. 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**(1):54.
29. Weigelt B, Mackay A, A'Hern R, Natrajan R, Tan DS, Dowsett M, Ashworth A, Reis-Filho JS: **Breast cancer molecular profiling with single sample predictors: a retrospective analysis**. *The Lancet Oncology* 2010, **11**(4):339-349.
30. Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SA, Caldas C: **Genome-driven integrated classification of breast cancer validated in over 7,500 samples**. *Genome Biol* 2014, **15**(8):431.
31. Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Tutt A, Ng T, Teschendorff AE: **DART: Denoising Algorithm based on Relevance network Topology improves molecular pathway activity inference**. *BMC Bioinformatics* 2011, **12**:403.
32. Waggott DM: **NanoStringNorm: Normalize NanoString miRNA and mRNA Data**. In., vol. R package version 1.1.21. http://CRAN.R-project.org/package=NanoStringNorm; 2015.
33. Lips EH, Mulder L, De Ronde JJ, Mandjes IaM, Koolen BB, Wessels LFa, Rodenhuis S, Wesseling J: **Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response**. *Breast Cancer Research and Treatment* 2013, **140**:63-71.
34. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al*: **Supervised risk predictor of breast cancer based on intrinsic subtypes**. *J Clin Oncol* 2009, **27**(8):1160-1167.
35. Ignatiadis M, Singhal SK, Desmedt C, Haibe-Kains B, Criscitiello C, Andre F, Loi S, Piccart M, Michiels S, Sotiriou C: **Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis**. *J Clin Oncol* 2012, **30**(16):1996-2004.
36. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1995, **57**:289-300.
37. Carstensen B, Plummer M, Laara E, Hills M: **Epi: A package for Statistical Analysis in Epidemiology. R package version 2.0. http://CRAN.R-project.org/package=Epi**. 2016.
38. Nguyen DX, Bos PD, Massague J: **Metastasis: from dissemination to organ-specific colonization**. *Nat Rev Cancer* 2009, **9**(4):274-284.
39. Vanharanta S, Massague J: **Origins of metastatic traits**. *Cancer Cell* 2013, **24**(4):410-421.
40. Stover DG, Coloff JL, Barry WT, Brugge JS, Winer EP, Selfors LM: **The Role of Proliferation in Determining Response to Neoadjuvant Chemotherapy in Breast Cancer: A Gene Expression-Based Meta-Analysis**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2016, **22**(24):6039-6050.
41. Weigelt B, Pusztai L, Ashworth A, Reis-Filho JS: **Challenges translating breast cancer gene signatures into the clinic**. *Nat Rev Clin Oncol* 2012, **9**(1):58-64.
42. Shipitsin M, Campbell LL, Argani P, Weremowicz S, Bloushtain-Qimron N, Yao J, Nikolskaya T, Serebryiskaya T, Beroukhim R, Hu M *et al*: **Molecular definition of breast tumor heterogeneity.** *Cancer Cell* 2007, **11**:259-273.
43. Lu X, Yan CH, Yuan M, Wei Y, Hu G, Kang Y: **In vivo dynamics and distinct functions of hypoxia in primary tumor growth and organotropic metastasis of breast cancer**. *Cancer Res* 2010, **70**(10):3905-3914.
44. Soady KJ, Kendrick H, Gao Q, Tutt A, Zvelebil M, Ordonez LD, Quist J, Tan DW, Isacke CM, Grigoriadis A *et al*: **Mouse mammary stem cells express prognostic markers for triple-negative breast cancer**. *Breast cancer research : BCR* 2015, **17**:31.
45. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, Zhou A, Eyob H, Balakrishnan S, Wang CY *et al*: **Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells**. *Nature* 2015, **526**(7571):131-135.
46. Kimbung S, Johansson I, Danielsson A, Veerla S, Egyhazi Brage S, Frostvik Stolt M, Skoog L, Carlsson L, Einbeigi Z, Lidbrink E *et al*: **Transcriptional Profiling of Breast Cancer Metastases Identifies Liver Metastasis-Selective Genes Associated with Adverse Outcome in Luminal A Primary Breast Cancer**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2015:1078-0432.CCR-1015-0487-.

47. Ho VK, Gijtenbeek JM, Brandsma D, Beerepoot LV, Sonke GS, van der Heiden-van der Loo M: **Survival of breast cancer patients with synchronous or metachronous central nervous system metastases**. *Eur J Cancer* 2015.
48. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C: **Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2008, **14**(16):5158-5165.
49. Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, Sotiriou C: **A three-gene model to robustly identify breast cancer molecular subtypes**. *J Natl Cancer Inst* 2012, **104**(4):311-325.
50. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z: **A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers**. *Nat Genet* 2006, **38**:1043-1048.
51. Yuen KW, Montpetit B, Hieter P: **The kinetochore and cancer: what's the connection?** *Curr Opin Cell Biol* 2005, **17**(6):576-582.
52. Thiru P, Kern DM, McKinley KL, Monda JK, Rago F, Su KC, Tsinman T, Yarar D, Bell GW, Cheeseman IM: **Kinetochore genes are coordinately up-regulated in human tumors as part of a FoxM1-related cell division program**. *Mol Biol Cell* 2014, **25**(13):1983-1994.
53. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H *et al*: **Subclonal diversification of primary breast cancer revealed by multiregion sequencing**. *Nat Med* 2015, **21**(7):751-759.
54. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:10869-10874.
55. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B *et al*: **Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis**. *J Natl Cancer Inst* 2006, **98**(4):262-272.
56. Purushotham A, Shamil E, Cariati M, Agbaje O, Muhidin A, Gillett C, Mera A, Sivanadiyan K, Harries M, Sullivan R *et al*: **Age at diagnosis and distant metastasis in breast cancer--a surprising inverse relationship**. *Eur J Cancer* 2014, **50**(10):1697-1705.
57. Cunha SI, Bocci M, Lovrot J, Eleftheriou N, Roswall P, Cordero E, Lindstrom L, Bartoschek M, Haller BK, Pearsall RS *et al*: **Endothelial ALK1 Is a Therapeutic Target to Block Metastatic Dissemination of Breast Cancer**. *Cancer Res* 2015, **75**(12):2445-2456.
58. Perez JE, Machiavelli M, Leone BA, Romero A, Rabinovich MG, Vallejo CT, Bianco A, Rodriguez R, Cuevas MA, Alvarez LA: **Bone-only versus visceral-only metastatic pattern in breast cancer: analysis of 150 patients. A GOCS study. Grupo Oncologico Cooperativo del Sur**. *Am J Clin Oncol* 1990, **13**(4):294-298.
59. Zardavas D, Maetens M, Irrthum A, Goulioti T, Engelen K, Fumagalli D, Salgado R, Aftimos P, Saini KS, Sotiriou C *et al*: **The AURORA initiative for metastatic breast cancer**. *Br J Cancer* 2014, **111**(10):1881-1887.
60. Bos PD, Zhang XH, Nadal C, Shu W, Gomis RR, Nguyen DX, Minn AJ, van de Vijver MJ, Gerald WL, Foekens JA *et al*: **Genes that mediate breast cancer metastasis to the brain**. *Nature* 2009, **459**(7249):1005-1009.
61. Brodsky AS, Fischer A, Miller DH, Vang S, MacLaughlan S, Wu HT, Yu J, Steinhoff M, Collins C, Smith PJ *et al*: **Expression profiling of primary and metastatic ovarian tumors reveals differences indicative of aggressive disease**. *PLoS ONE* 2014, **9**(4):e94476.
62. Lu X, Kang Y: **Organotropism of breast cancer metastasis**. *Journal of Mammary Gland Biology and Neoplasia* 2007, **12**:153-162.
63. Zhang L, Zhang S, Yao J, Lowery FJ, Zhang Q, Huang W-C, Li P, Li M, Wang X, Zhang C *et al*: **Microenvironment-induced PTEN loss by exosomal microRNA primes brain metastasis outgrowth**. *Nature* 2015.

FIGURE LEGENDS

Figure 1. Overview diagram of cases and gene expression data sets. For each metastatic population ("*visceral only*", "*bone & visceral*", "*bone only*") 400 cases were sampled, and three possible controls were matched to each case by calendar time of event. A random sample of case-control sets was taken forward to tissue assessment for RNA extraction (Supplementary Methods). Extracted RNA was available for a total of 742 case-control (1:1) pairs, comprising a total of 1,277 individual patients. A detailed overview of patients and samples included in the design and in the gene expression data sets is shown in Supplementary Figure S1A, B.

Figure 2. Distribution of three metastatic groups across breast cancer subtypes within the dataset GWDb. Barplots illustrate the proportion of immunohistochemically defined (IHC), prediction analysis of microarray 50 (PAM50) and IntClust breast cancer subtypes present in each metastatic group (a) and the proportion of each metastatic group assigned to IHC, PAM50 and IntClust subtypes (b). The patient number for each group is shown on the top of each column. ER estrogen receptor, HER2 human epidermal growth factor receptor, TNBC triple-negative breast cancer

Figure 3. Forestplots of log(OR) estimated from univariate conditional logistic regression and logistic regression models of illustrative gene modules in *GWDb* for each metastatic group. A. Conditional logistic regression (matched pairs). **B.** Logistic regression using complete case-control pairs. Alternative grey and green bars indicate the broad categorisation of illustrative gene modules (Proliferation', 'Immune', *etc*.) as shown in the module labels on the left of the plot. **C.** Heatmap of pairwise correlation for a panel of gene modules (Pearson correlation, 'complete' clustering; irrespective of case type or case-control status). Gene modules are listed in Supplementary Table S2.

Figure 4. *BV* module for metasynchronous metastatic spread. A. Discovery of the *BV* module. *BV* module scores, shown as density plots for each metastatic group in the ER-positive case-control paired breast cancer cases from the discovery set (*GWDb*) together with the corresponding ROC curve for the *'bone & visceral'* group. **B.** Estimated *BV* expression score based on NanoString quantification compared with those obtained from the WG-DASL

platform. **C.** Correlation plot displaying the *BV* gene set in primary tumours and their matched lymph node metastases. Dots are colour coded according to the metastatic groups. **D.** Boxplot of *BV* module scores in lymph node metastases of the GSE46141 data, displayed according to their reported patterns of first metastatic spread.

TABLE LEGENDS

Tables 1 and 2. Patient characteristics for case-control series.

Table 1. Patients with available high-quality RNA sample.

Table 2. Patients present in the discovery set (*GWDb).*

Table 3. Estimated ORs for molecular subtype representations based on conditional logisitic regression. IHC-derived subtypes were compared with ER-positive, HER2-negative tumours. PAM50 subtypes were compared with the 'LuminalA' subtype. IntClust subtypes were compared with the baseline IntClust3 cluster. ORs were calculated from conditional logistic regression models.

Table 4. Estimated ORs for molecular subtype representations based on logisitic regression. IHC-derived subtypes were compared with ER-positive, HER2-negative tumours. PAM50 subtypes were compared with the 'LuminalA' subtype. IntClust subtypes were compared with the baseline IntClust3 cluster. ORs were calculated from logistic (unmatched) regression models, to avoid missing data where one sample of a pair is not available in *GWDb*.

ADDITIONAL DATA FILES

Supplementary Methods (PDF). Detailed description of tissue preparation including microtomy, de-waxing and staining, micro-dissection and preparation of materials, and gene expression processing and quality control.

Supplementary Figures (PDF). Contains Supplementary Figures S1-S8.

Supplementary Tables (XLSX). Contains Supplementary Tables S1-S2, S4-S5.

Supplementary Table S3 (XLSX). Contains Supplementary Table S3.

Table 1. Patient characteristics in the case-control series: patients with extracted RNA sample

		V cases	*V* controls	*BV* cases	*BV* controls	*B* cases	*B* controls
Number of patients		246	232	258	245	238	222
		Number (%)					
Age at hist. diagnosis	Median (years)	57.1	51.5	55.7	51.0	55.9	51.2
Grade	1	11 (4%)	33 (14%)	7 (3%)	36 (15%)	20 (8%)	20 (9%)
	2	81 (33%)	89 (38%)	113 (44%)	127 (52%)	114 (48%)	97 (44%)
	3	135 (55%)	87 (38%)	122 (47%)	62 (25%)	81 (34%)	88 (40%)
	Unknown	19 (8%)	23 (10%)	16 (6%)	20 (8%)	23 (10%)	17 (8%)
ER IHC status	Positive	147 (60%)	160 (69%)	162 (63%)	174 (71%)	188 (79%)	158 (71%)
	Negative	99 (40%)	72 (31%)	96 (37%)	71 (29%)	50 (21%)	64 (29%)
PR status	Positive	97 (39%)	121 (52%)	116 (45%)	131 (53%)	146 (61%)	113 (51%)
	Negative	149 (61%)	111 (48%)	142 (55%)	114 (47%)	92 (39%)	109 (49%)
HER2 status	Positive	58 (24%)	35 (15%)	55 (21%)	30 (12%)	35 (15%)	23 (10%)
	Negative	108 (44%)	102 (44%)	116 (45%)	98 (40%)	108 (45%)	107 (48%)
	Unknown	80 (33%)	95 (41%)	87 (34%)	117 (48%)	95 (40%)	92 (41%)
Tumour size	<= 2 cm	96 (40%)	113 (49%)	99 (38%)	136 (56%)	101 (42%)	117 (53%)
	>2 cm	146 (60%)	113 (49%)	155 (60%)	101 (41%)	133 (56%)	99 (45%)
	Unknown	– –	6 (3%)	4 (2%)	8 (3%)	4 (2%)	6 (3%)
Lymph nodes positive	0	51 (21%)	97 (42%)	61 (24%)	109 (44%)	61 (26%)	103 (46%)
	1–3	74 (30%)	81 (35%)	64 (25%)	80 (33%)	70 (29%)	64 (29%)
	4+	74 (30%)	28 (12%)	81 (31%)	28 (11%)	64 (27%)	29 (13%)
	Unknown	47 (19%)	26 (11%)	52 (20%)	28 (11%)	43 (18%)	26 (12%)
Invasive subtype	NOS/no special type	217 (88%)	189 (81%)	219 (85%)	206 (84%)	196 (82%)	189 (85%)
	Ductal - mucinous	1 (0%)	3 (1%)	4 (2%)	2 (1%)	1 (0%)	4 (2%)
	Ductal - tubular	– –	6 (3%)	1 (0%)	7 (3%)	3 (1%)	3 (1%)
	Ductal - other	6 (2%)	3 (1%)	1 (0%)	3 (1%)	1 (0%)	3 (1%)
	Lobular - classical	14 (6%)	23 (10%)	19 (7%)	22 (9%)	29 (12%)	13 (6%)
	Lobular - pleomorphic	3 (1%)	2 (1%)	6 (2%)	3 (1%)	3 (1%)	4 (2%)
	Lobular - other	2 (1%)	3 (1%)	3 (1%)	2 (1%)	4 (2%)	1 (0%)
	Other	3 (1%)	3 (1%)	5 (2%)	– –	1 (0%)	5 (2%)
Surgery type (any time)	Breast conserving	58 (24%)	80 (34%)	70 (27%)	82 (33%)	64 (27%)	72 (32%)
	Breast conserving + mastectomy	53 (22%)	28 (12%)	56 (22%)	32 (13%)	37 (16%)	27 (12%)
	Mastectomy	106 (43%)	110 (47%)	99 (38%)	120 (49%)	115 (48%)	114 (51%)
	Unknown	29 (12%)	14 (6%)	33 (13%)	11 (4%)	22 (9%)	9 (4%)
Radiotherapy (adj/neo)	Yes	168 (68%)	140 (60%)	197 (76%)	126 (51%)	194 (82%)	125 (56%)
	No	78 (32%)	92 (40%)	61 (24%)	119 (49%)	44 (18%)	97 (44%)
Hormone treatment (adj/neo)	Yes	204 (83%)	120 (52%)	230 (89%)	130 (53%)	218 (92%)	114 (51%)
	No	42 (17%)	112 (48%)	28 (11%)	115 (47%)	20 (8%)	108 (49%)
Chemotherapy	Neo-adj only	– –	3 (1%)	3 (1%)	67 (27%)	8 (3%)	3 (1%)
	Neo-adj and adj	6 (2%)	1 (0%)	9 (3%)	1 (0%)	1 (0%)	2 (1%)
	Adj only	146 (59%)	85 (37%)	150 (58%)	2 (1%)	101 (42%)	67 (30%)
	No	94 (38%)	143 (62%)	96 (37%)	175 (71%)	128 (54%)	150 (68%)

Hist.: histological assessment, ER: IHC estrogen receptor status determined by immunohistochemical assessment, PR progesterone receptor, HER2 human epidermal growth factor receptor 2, adj adjuvant, neo neoadjuvant

Table 2. Patient characteristics in the case-control series: patients in the discovery cohort GWDb

		V cases	*V* controls	*BV* cases	*BV* controls	*B* cases	*B* controls	Case series	
Number of patients		105	82	106	86	98	90	χ^2	*p*
Age at hist. diag.	Median (years)	55.4	52.0	52.5	50.3	55.9	48.8		
Grade	1	3 (3%)	12 (15%)	3 (3%)	13 (15%)	8 (8%)	7 (8%)		
	2	27 (26%)	32 (39%)	44 (42%)	47 (55%)	44 (45%)	42 (47%)		
	3	68 (65%)	34 (41%)	55 (52%)	19 (22%)	38 (39%)	37 (41%)		
	Unknown	7 (7%)	4 (5%)	4 (4%)	7 (8%)	8 (8%)	4 (4%)	18.2	0.01
ER IHC status	Positive	57 (54%)	52 (63%)	69 (65%)	62 (72%)	77 (79%)	65 (72%)		
	Negative	48 (46%)	30 (37%)	37 (35%)	24 (28%)	21 (21%)	25 (28%)	13.3	0.001
PR status	Positive	39 (37%)	40 (49%)	53 (50%)	56 (65%)	59 (60%)	51 (57%)		
	Negative	66 (63%)	42 (51%)	53 (50%)	30 (35%)	39 (40%)	39 (43%)	10.9	0.004
HER2 status	Positive	31 (30%)	13 (16%)	22 (21%)	13 (15%)	15 (15%)	11 (12%)		
	Negative	44 (42%)	38 (46%)	51 (48%)	41 (48%)	45 (46%)	42 (47%)		
	Unknown	30 (29%)	31 (38%)	33 (31%)	32 (37%)	38 (39%)	37 (41%)	7.0	0.1
Tumour size	<= 2 cm	41 (39%)	39 (48%)	39 (37%)	44 (51%)	38 (39%)	47 (52%)		
	>2 cm	64 (61%)	42 (51%)	65 (61%)	38 (44%)	60 (61%)	41 (46%)	0.1	1.0
	Unknown	– –	1 (1%)	2 (2%)	4 (5%)	– –	2 (2%)		
Lymph nodes positive	0	15 (14%)	33 (40%)	21 (20%)	32 (37%)	24 (24%)	39 (43%)		
	1–3	34 (32%)	29 (35%)	31 (29%)	28 (33%)	30 (31%)	30 (33%)		
	4+	37 (35%)	10 (12%)	40 (38%)	11 (13%)	26 (27%)	10 (11%)		
	Unknown	19 (18%)	10 (12%)	14 (13%)	15 (17%)	18 (18%)	11 (12%)	6.1	0.4
Invasive subtype	NOS/no special type	96 (91%)	70 (85%)	85 (82%)	73 (85%)	82 (84%)	81 (90%)		
	Ductal - mucinous	– –	2 (2%)	1 (1%)	– –	1 (1%)	1 (1%)		
	Ductal - tubular	– –	– –	1 (1%)	3 (3%)	2 (2%)	1 (1%)		
	Ductal - other	1 (1%)	1 (1%)	1 (1%)	– –	– –	3 (3%)		
	Lobular - classical	5 (5%)	5 (6%)	7 (7%)	9 (10%)	12 (12%)	2 (2%)		
	Lobular - pleomorphic	1 (1%)	1 (1%)	6 (6%)	– –	– –	1 (1%)		
	Lobular - other	– –	2 (2%)	2 (2%)	1 (1%)	1 (1%)	– –		
	Other	2 (2%)	1 (1%)	1 (1%)	– –	– –	1 (1%)	1.7	0.4 †
Surgery type (any time)	Breast conserving	30 (29%)	34 (41%)	35 (33%)	43 (50%)	31 (32%)	39 (43%)		
	Breast conserving + mastectomy	17 (16%)	11 (13%)	19 (18%)	7 (8%)	15 (15%)	13 (14%)		
	Mastectomy	47 (45%)	33 (40%)	41 (39%)	30 (35%)	43 (44%)	35 (39%)		
	Unknown	11 (10%)	4 (5%)	11 (10%)	6 (7%)	9 (9%)	3 (3%)	1.2	1.0
Radiotherapy (adj/neo)	Yes	79 (75%)	52 (63%)	84 (79%)	55 (64%)	82 (84%)	59 (66%)		
	No	26 (25%)	30 (37%)	22 (21%)	31 (36%)	16 (16%)	31 (34%)	2.2	0.3
Hormone treatment (adj/neo)	Yes	84 (80%)	38 (46%)	97 (92%)	57 (66%)	92 (94%)	51 (57%)		
	No	21 (20%)	44 (54%)	9 (8%)	29 (34%)	6 (6%)	39 (43%)	11.0	0.004
Chemotherapy	Neo-adj only	– –	2 (2%)	2 (2%)	1 (1%)	5 (5%)	3 (3%)		
	Neo-adj and adj	4 (4%)	– –	4 (4%)	1 (1%)	1 (1%)	– –		
	Adj only	65 (62%)	33 (40%)	67 (63%)	24 (28%)	43 (44%)	28 (31%)		
	No	36 (34%)	47 (57%)	33 (31%)	60 (70%)	49 (50%)	59 (66%)	8.7	0.01 ‡
Metastatic events in control series		No mets, 64 (78%) V = 8, BV = 8, B = 2		No mets, 66 (77%) V = 10, BV = 5, B = 5		No mets, 69 (77%) V = 4, BV = 10, B = 7			

†ductal versus lobular, ‡any versus no chemotherapy. Abbreviations as for Table 1

Table 3

		V SERIES		n, paired		n, informative				BV SERIES		n, paired		n, informative				B SERIES		n, paired		n, informative				ANY CASE TYPE		n, paired		n, informative		
		Cases	Controls	Cases	Controls	Cases	Controls	OR (95% CI)	p	Cases	Controls	Cases	Controls	Cases	Controls	OR (95% CI)	p	Cases	Controls	Cases	Controls	Cases	Controls	OR (95% CI)	p	Cases	Controls	Cases	Controls	Cases	Controls	OR (95% CI)	p
Extracted RNA																															
ER IHC	**ER-**	99	76	*99*	*76*	*69*	*46*	Ref.		96	74	*96*	*74*	*67*	*45*	Ref.		50	68	*50*	*68*	*32*	*50*	Ref.		245	218	*245*	*218*	*168*	*141*	Ref.	
	ER+	147	170	*147*	*170*	*46*	*69*	0.67 (0.46-0.97)	0.03	162	184	*162*	*184*	*45*	*67*	0.67 (0.46-0.98)	0.04	188	170	*188*	*170*	*50*	*32*	1.6 (1.0-2.4)	0.05	497	524	*497*	*524*	*141*	*168*	0.84 (0.67-1.0)	0.1
	Total	246	246	*246*	*246*	*115*	*115*			258	258	*258*	*258*	*112*	*112*			238	238	*238*	*238*	*50*	*50*			742	742	*742*	*742*	*309*	*309*		
IHC subtypes	**ER+HER2-**	68	70	*42*	*49*	*22*	*29*	Ref.		78	74	*48*	*55*	*15*	*22*	Ref.		86	84	*56*	*48*	*21*	*13*	Ref.		232	228	*146*	*152*	*58*	*64*	Ref.	
	ER+HER2+	23	25	*11*	*17*	*11*	*17*	0.8 (0.36-1.8)	0.6	25	17	*12*	*12*	*10*	*10*	1.3 (0.48-3.3)	0.7	22	15	*13*	*11*	*10*	*8*	1.1 (0.43-3.0)	0.8	70	57	*36*	*40*	*31*	*35*	0.96 (0.58-1.6)	0.9
	ER-HER2+	35	12	*16*	*8*	*15*	*7*	2.2 (0.88-5.7)	0.09	30	13	*17*	*9*	*16*	*8*	2.2 (0.87-5.3)	0.1	13	11	*7*	*8*	*4*	*5*	0.84 (0.22-3.2)	0.8	78	36	*40*	*25*	*35*	*20*	1.7 (0.97-3.0)	0.07
	TNBC	36	31	*24*	*19*	*20*	*15*	1.4 (0.69-2.9)	0.3	33	25	*18*	*19*	*12*	*13*	1.1 (0.48-2.7)	0.8	14	23	*5*	*14*	*3*	*12*	0.25 (0.07-0.89)	0.03	83	79	*47*	*52*	*35*	*40*	0.93 (0.58-1.5)	0.8
Not assigned (missing IHC data)		*84*	*108*							*92*	*129*							*103*	*105*							*279*	*342*						
	Total	246	246	*93*	*93*	*68*	*68*			258	258	*95*	*95*	*53*	*53*			238	238	*81*	*81*	*38*	*38*			742	742	*269*	*269*	*159*	*159*		
GWDb																															
PAM50	**Luminal A**	21	34	*10*	*20*	*7*	*17*	Ref.		33	45	*19*	*25*	*10*	*16*	Ref.		32	45	*16*	*25*	*9*	*18*	Ref.		86	124	*45*	*70*	*26*	*51*	Ref.	
	Luminal B	17	9	*10*	*9*	*6*	*5*	2.7 (0.62-11)	0.2	26	13	*13*	*5*	*12*	*4*	2.5 (0.79-8.1)	0.1	28	16	*15*	*6*	*14*	*5*	4.2 (1.3-14)	0.02	71	38	*38*	*20*	*32*	*14*	2.9 (1.5-5.9)	0.002
	Basal	28	17	*11*	*10*	*11*	*10*	1.7 (0.62-4.7)	0.3	21	13	*14*	*9*	*7*	*2*	3.0 (0.61-15)	0.2	9	17	*6*	*10*	*4*	*8*	0.66 (0.19-2.3)	0.5	58	47	*31*	*29*	*22*	*20*	1.4 (0.7-2.7)	0.3
	HER2	30	15	*15*	*6*	*11*	*2*	8.6 (1.6-46)	0.01	20	12	*4*	*9*	*2*	*7*	0.4 (0.081-2.1)	0.3	18	12	*6*	*4*	*5*	*3*	2.9 (0.57-15)	0.2	68	39	*25*	*19*	*18*	*12*	2.3 (1.0-5.0)	0.05
	Normal	9	11	*5*	*6*	*4*	*5*	1.6 (0.35-7.4)	0.5	6	7	*1*	*3*	*1*	*3*	0.5 (0.050-5.4)	0.6	11	8	*7*	*5*	*6*	*4*	2.5 (0.55-11)	0.2	26	26	*13*	*14*	*11*	*12*	1.4 (0.6-3.5)	0.4
Not assigned (missing from GWDb)		*141*	*160*							*152*	*168*							*140*	*140*							433	468						
	Total	246	246	*51*	*51*	*39*	*39*			258	258	*51*	*51*	*32*	*32*			238	238	*50*	*50*	*38*	*38*			742	742	*152*	*152*	*109*	*109*		
IntClust, grouped	**1,6,9**	9	21	*6*	*11*	*5*	*10*	Ref.		14	25	*9*	*14*	*8*	*13*	Ref.		19	25	*11*	*13*	*8*	*10*	Ref.		42	71	*26*	*38*	*21*	*33*	Ref.	
	2,4	24	18	*13*	*10*	*10*	*7*	2.6 (0.70-9.5)	0.2	14	18	*6*	*9*	*3*	*6*	0.86 (0.18-4.1)	0.9	26	14	*12*	*7*	*10*	*5*	2.1 (0.49-8.6)	0.3	64	50	*31*	*26*	*23*	*18*	1.9 (0.88-4.2)	0.1
	5	20	6	*9*	*1*	*9*	*1*	17 (1.6-188)	0.02	17	3	*4*	*1*	*4*	*1*	7.3 (0.66-81)	0.1	7	5	*3*	*1*	*3*	*1*	5.1 (0.43-61)	0.2	44	14	*16*	*3*	*16*	*3*	9.0 (2.3-35)	0.002
	7,8	12	22	*4*	*16*	*4*	*16*	0.66 (0.15-2.9)	0.6	22	23	*12*	*13*	*10*	*11*	1.4 (0.51-3.9)	0.5	24	27	*10*	*14*	*6*	*10*	0.89 (0.25-3.1)	0.9	58	72	*26*	*43*	*20*	*37*	0.95 (0.49-1.9)	0.9
	3	20	11	*8*	*7*	*8*	*7*	2.1 (0.47-9.3)	0.3	13	8	*9*	*7*	*5*	*3*	2.2 (0.43-10.7)	0.4	4	10	*3*	*7*	*3*	*7*	0.59 (0.13-2.7)	0.5	37	29	*20*	*21*	*16*	*17*	1.5 (0.64-3.3)	0.4
Not assigned (missing from GWDb)		*141*	*160*							*152*	*168*							*140*	*140*							*433*	*468*						
	Total	246	246	*51*	*51*	*46*	*46*			258	258	*51*	*51*	*40*	*40*			238	238	*50*	*50*	*41*	*41*			742	742	*152*	*152*	*127*	*127*		
IntClust	**3**	9	21	*6*	*11*	*5*	*10*			14	25	*9*	*14*	*8*	*13*			19	25	*11*	*13*	*8*	*10*			42	71	*26*	*38*	*21*	*33*	Ref.	
	1	12	0	*8*	*0*	*8*	*0*			9	6	*4*	*4*	*4*	*4*			5	13	*3*	*7*	*3*	*7*			26	19	*15*	*11*	*15*	*11*	2.3 (0.88-6.2)	0.09
	2	6	4	*4*	*4*	*4*	*4*			4	1	*2*	*0*	*2*	*0*			8	3	*4*	*2*	*4*	*2*			18	8	*10*	*6*	*10*	*6*	3.0 (0.90-10)	0.07
	4	18	14	*9*	*6*	*9*	*6*			10	17	*4*	*9*	*3*	*8*			18	11	*8*	*6*	*3*	*6*			46	42	*21*	*20*	*18*	*17*	1.5 (0.67-3.5)	0.3
	5	20	6	*9*	*1*	*9*	*1*			17	3	*4*	*1*	*4*	*1*			7	5	*3*	*1*	*3*	*1*			44	14	*16*	*3*	*16*	*3*	8.7 (2.2-34)	0.002
	6	4	1	*2*	*1*	*2*	*1*			7	2	*4*	*1*	*4*	*1*			5	2	*3*	*0*	*3*	*0*			16	5	*9*	*2*	*9*	*2*	6.0 (1.2-31)	0.03
	7	5	13	*1*	*10*	*1*	*10*			9	11	*5*	*5*	*4*	*4*			17	14	*4*	*8*	*3*	*7*			31	38	*10*	*23*	*8*	*21*	0.63 (0.25-1.6)	0.3
	8	7	9	*3*	*6*	*3*	*6*			13	12	*7*	*8*	*7*	*8*			7	13	*6*	*6*	*5*	*5*			27	34	*16*	*20*	*15*	*19*	1.3 (0.56-2.8)	0.6
	9	4	7	*1*	*5*	*1*	*5*			10	5	*3*	*2*	*2*	*1*			8	2	*5*	*1*	*5*	*1*			22	14	*9*	*8*	*8*	*7*	1.8 (0.56-5.7)	0.3
	10	20	11	*8*	*7*	*8*	*7*			13	8	*9*	*7*	*5*	*3*			4	10	*3*	*7*	*3*	*7*			37	29	*20*	*21*	*16*	*17*	1.4 (0.61-3.1)	0.4
Not assigned (missing from GWDb)		*141*	*160*							*152*	*168*							*140*	*140*							*433*	*468*						
	Total	246	246	*51*	*51*	*50*	*50*			258	258	*51*	*51*	*43*	*43*			238	238	*50*	*50*	*43*	*43*			742	742	*152*	*152*	*136*	*136*		

Footnotes: Conditional logistic regression.

Epi::clogistic, Case_Control ~ x + strata(Random_Selection)

Note: Counts of pairs (not missing) and informative entries in the conditional logistic model are shown for information.

Table 4

		V SERIES		n, uniq			BV SERIES		n, uniq			B SERIES		n, uniq			ANY CASE TYPE		n, uniq		
		Cases	Controls	Controls	OR (95% CI)	p	Cases	Controls	Controls	OR (95% CI)	p	Cases	Controls	Controls	OR (95% CI)	p	Cases	Controls	Controls	OR (95% CI)	p
Extracted RNA																					
ER IHC	**ER-**	99	76	72	Ref.		96	74	71	Ref.		50	68	64	Ref.		245	218	184	Ref.	
	ER+	147	170	160	0.67 (0.46-0.97)	0.04	162	184	174	0.69 (0.47-1.0)	0.05	188	170	158	1.5 (1.0-2.3)	0.05	497	524	436	0.86 (0.68-1.1)	0.2
	Total	**246**	**246**	**232**			**258**	**258**	**245**			**238**	**238**	**222**			**742**	**742**	**620**		
IHC subtypes	**ER+HER2-**	68	70	67	Ref.		78	74	72	Ref.		86	84	80	Ref.		232	228	196	Ref.	
	ER+HER2+	23	25	23	0.99 (0.50-1.9)	1.0	25	17	17	1.4 (0.68-2.8)	0.4	22	15	14	1.5 (0.71-3.1)	0.3	70	57	43	1.4 (0.90-2.1)	0.1
	ER-HER2+	35	12	12	2.9 (1.4-6.2)	0.005	30	13	13	2.1 (1.0-4.5)	0.04	13	11	9	1.3 (0.55-3.4)	0.5	78	36	31	2.1 (1.4-3.4)	0.001
	TNBC	36	31	30	1.2 (0.66-2.1)	0.6	33	25	23	1.3 (0.71-2.5)	0.4	14	23	23	0.57 (0.27-1.2)	0.1	83	79	68	1.0 (0.71-1.5)	0.9
	Total	**162**	**138**	**132**			**166**	**129**	**125**			**135**	**133**	**126**			**463**	**400**	**338**		
Not assigned (missing IHC data)		*84*	*108*	*100*			*92*	*129*	*120*			*103*	*105*	*96*			*279*	*342*	*282*		
Total (extracted RNA)		*246*	*246*	*232*			*258*	*258*	*245*			*238*	*238*	*222*			*742*	*742*	*620*		
GWDb																					
PAM50	**Luminal A**	21	34	33	Ref.		33	45	42	Ref.		32	45	42	Ref.		86	124	109	Ref.	
	Luminal B	17	9	9	3.0 (1.1-8.2)	0.03	26	13	12	2.7 (1.2-6.4)	0.02	28	16	15	2.5 (1.1-5.4)	0.02	71	38	35	2.6 (1.6-4.2)	0.0002
	Basal	28	17	15	2.9 (1.3-6.9)	0.01	21	13	13	2.1 (0.9-4.8)	0.09	9	17	16	0.74 (0.28-1.9)	0.5	58	47	40	1.8 (1.1-3.0)	0.02
	HER2	30	15	14	3.4 (1.5-8.0)	0.005	20	12	12	2.1 (0.92-5.1)	0.08	18	12	10	2.4 (0.98-6.0)	0.06	68	39	33	2.6 (1.6-4.4)	<1e-4
	Normal	9	11	11	1.3 (0.45-3.6)	0.6	6	7	7	1.1 (0.32-3.6)	0.9	11	8	7	2.1 (0.7-6.2)	0.2	26	26	23	1.4 (0.76-2.7)	0.3
	Total	**105**	**86**	**82**			**106**	**90**	**86**			**98**	**98**	**90**			**309**	**274**	**240**		
Not assigned (not in GWDb)		*141*	*160*	*150*			*152*	*168*	*159*			*140*	*140*	*132*			*433*	*468*	*380*		
Total (extracted RNA)		*246*	*246*	*232*			*258*	*258*	*245*			*238*	*238*	*222*			*742*	*742*	*620*		
IntClust, grouped	**3**	9	21	20	Ref.		14	25	25	Ref.		19	25	23	Ref.		42	71	66	Ref.	
	1,6,9	20	8	8	5.6 (1.8-18)	0.003	26	13	12	3.9 (1.5-10)	0.01	18	17	15	1.5 (0.45-1.5)	0.4	64	38	35	2.9 (1.6-5.1)	0.0003
	2,4	24	18	17	3.1 (1.2-8.9)	0.03	14	18	17	1.5 (0.6-3.9)	0.4	26	14	13	2.4 (1.0-6.1)	0.05	64	50	42	2.4 (1.4-4.2)	0.002
	5	20	6	5	8.9 (2.7-34)	0.001	17	3	3	10 (2.8-49)	0.001	7	5	4	2.1 (0.55-9.1)	0.3	44	14	11	6.3 (3.0-14)	<1e-4
	7,8	12	22	21	1.3 (0.44-3.7)	0.7	22	23	21	1.9 (0.78-4.6)	0.2	24	27	26	1.1 (0.49-2.6)	0.8	58	72	59	1.5 (0.91-2.6)	0.1
	10	20	11	11	4.0 (1.4-12)	0.01	13	8	8	2.9 (0.99-9.0)	0.06	4	10	9	0.54 (0.13-1.9)	0.4	37	29	27	2.2 (1.2-4.1)	0.02
	Total	**105**	**86**	**82**			**106**	**90**	**86**			**98**	**98**	**90**			**309**	**274**	**240**		
Not assigned (not in GWDb)		*141*	*160*	*150*			*152*	*168*	*159*			*140*	*140*	*132*			*433*	*468*	*380*		
Total (extracted RNA)		*246*	*246*	*232*			*258*	*258*	*245*			*238*	*238*	*222*			*742*	*742*	*620*		

Footnotes: Logistic regression model.

stats::glm, Case_Control ~ x, family='binomial'

Note: This is an unpaired test. A small number of patients were duplicated within a control series by design; here, patients were de-duplicated within each control series ("n, uniq"). There are no duplicates within each case series by design. Duplicates were permitted between cases and controls. Entries with missing values in the independent variable were excluded.

Figure 1

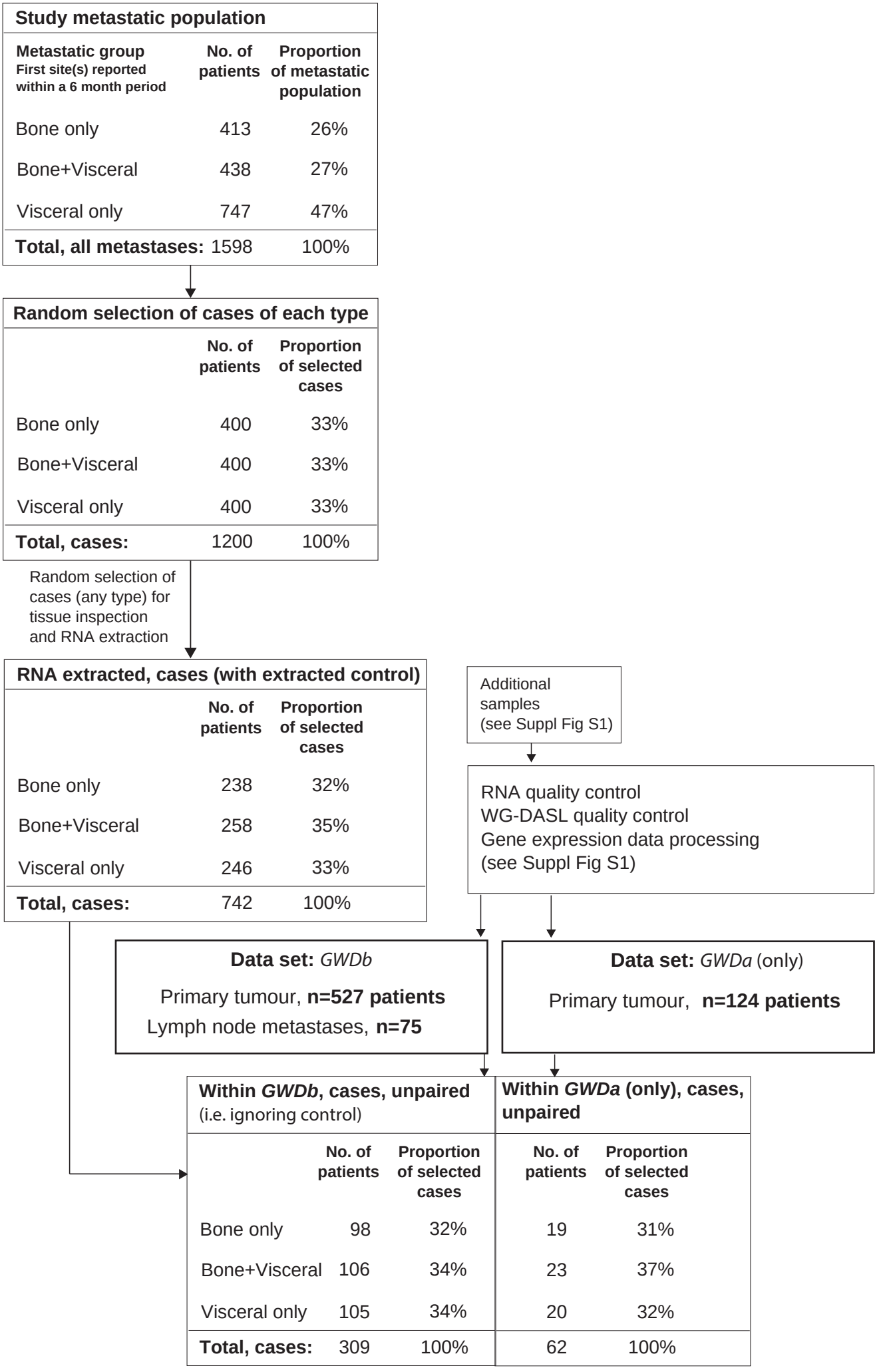

Study metastatic population
Metastatic group
First site(s) reported within a 6 month period
No. of patients
Proportion of metastatic population
Bone only 413 26%
Bone+Visceral 438 27%
Visceral only 747 47%
Total, all metastases: 1598 100%
Random selection of cases of each type
No. of patients
Proportion of selected cases
Bone only 400 33%
Bone+Visceral 400 33%
Visceral only 400 33%
Total, cases: 1200 100%
Random selection of cases (any type) for tissue inspection and RNA extraction
RNA extracted, cases (with extracted control)
No. of patients
Proportion of selected cases
Bone only 238 32%
Bone+Visceral 258 35%
Visceral only 246 33%
Total, cases: 742 100%
Additional samples (see Suppl Fig S1)
RNA quality control
WG-DASL quality control
Gene expression data processing
(see Suppl Fig S1)
Data set: GWDb
Primary tumour, n=527 patients
Lymph node metastases, n=75
Data set: GWDa (only)
Primary tumour, n=124 patients
Within GWDb, cases, unpaired (i.e. ignoring control)
No. of patients
Proportion of selected cases
Bone only 98 32%
Bone+Visceral 106 34%
Visceral only 105 34%
Total, cases: 309 100%
Within GWDa (only), cases, unpaired
No. of patients
Proportion of selected cases
19 31%
23 37%
20 32%
62 100%

Figure 2

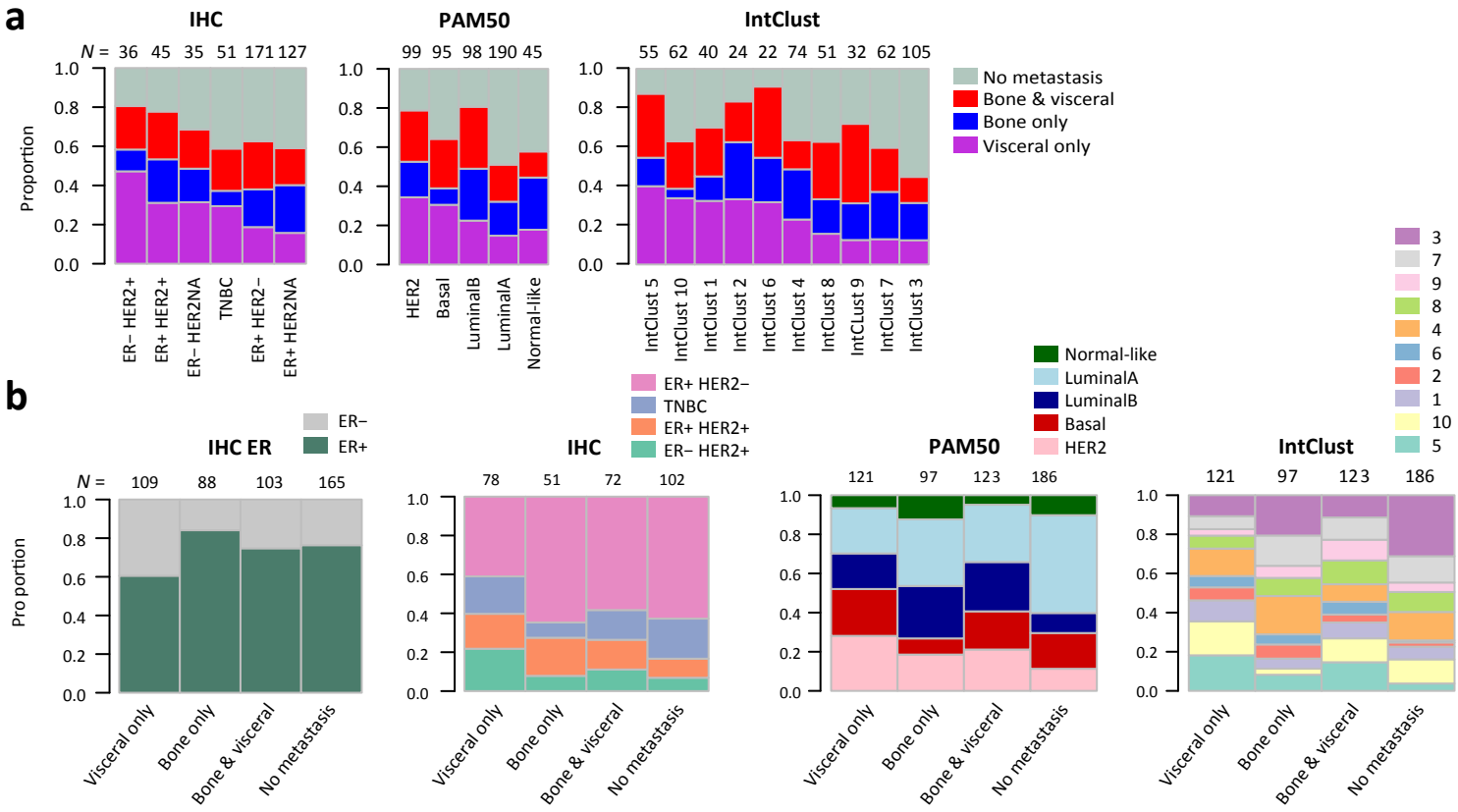

a
IHC
PAM50
IntClust
N = 36 45 35 51 171 127
99 95 98 190 45
55 62 40 24 22 74 51 32 62 105
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
ER− HER2+
ER+ HER2+
ER− HER2NA
TNBC
ER+ HER2−
ER+ HER2NA
HER2
Basal
LuminalB
LuminalA
Normal-like
IntClust 5
IntClust 10
IntClust 1
IntClust 2
IntClust 6
IntClust 4
IntClust 8
IntCLust 9
IntClust 7
IntClust 3
No metastasis
Bone & visceral
Bone only
Visceral only
b
IHC ER
ER−
ER+
IHC
ER+ HER2−
TNBC
ER+ HER2+
ER− HER2+
PAM50
Normal-like
LuminalA
LuminalB
Basal
HER2
IntClust
3
7
9
8
4
6
2
1
10
5
N = 109 88 103 165
78 51 72 102
121 97 123 186
121 97 123 186
Pro portion
Visceral only
Bone only
Bone & visceral
No metastasis

Figure 3

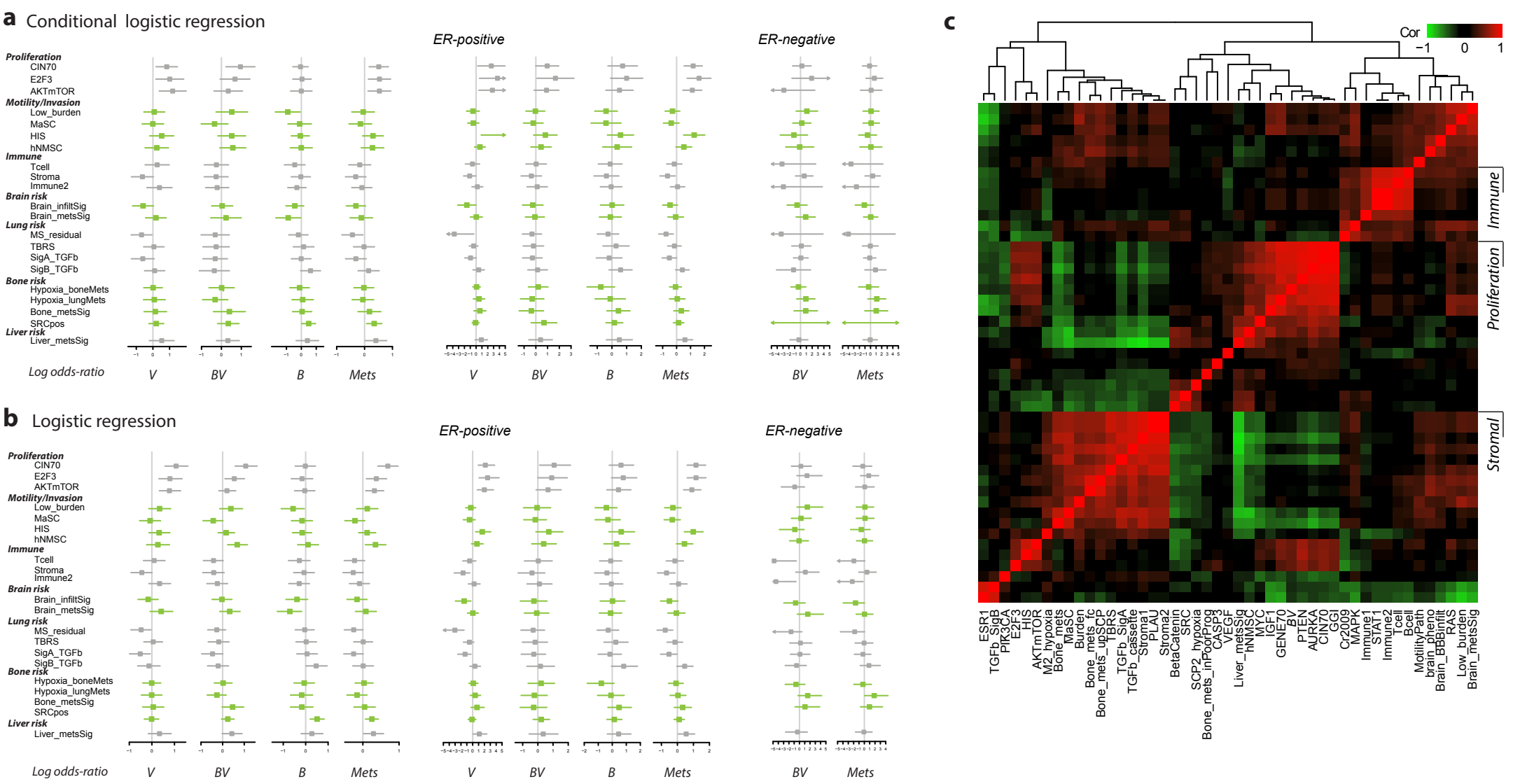
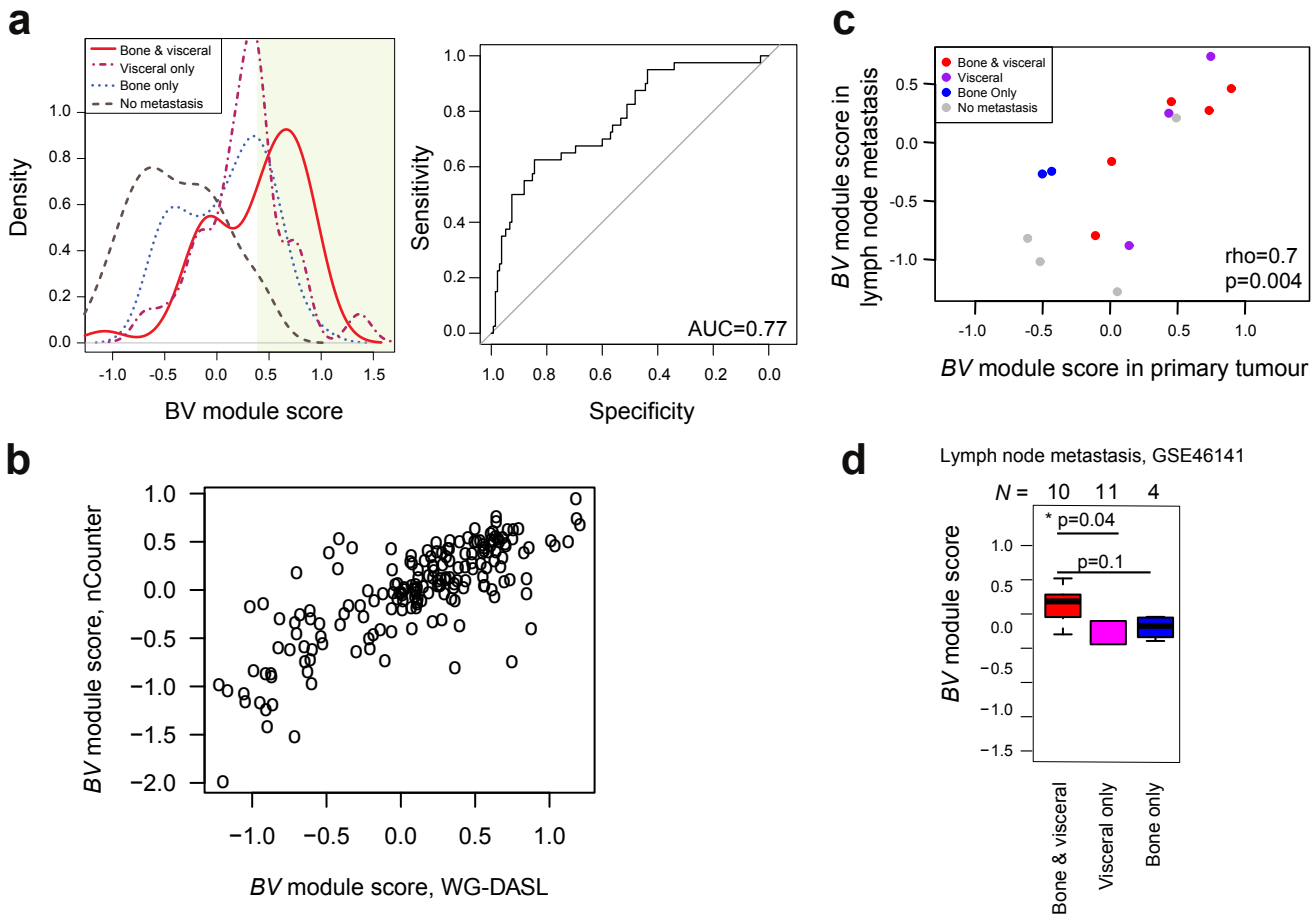

a Conditional logistic regression
ER-positive
ER-negative
Proliferation
CIN70
E2F3
AKTmTOR
Motility/Invasion
Low_burden
MaSC
HIS
hNMSC
Immune
Tcell
Stroma
Immune2
Brain risk
Brain_infiltSig
Brain_metsSig
Lung risk
MS_residual
TBRS
SigA_TGFb
SigB_TGFb
Bone risk
Hypoxia_boneMets
Hypoxia_lungMets
Bone_metsSig
SRCpos
Liver risk
Liver_metsSig
Log odds-ratio
V
BV
B
Mets
b Logistic regression
ER-positive
ER-negative
Proliferation
CIN70
E2F3
AKTmTOR
Motility/Invasion
Low_burden
MaSC
HIS
hNMSC
Immune
Tcell
Stroma
Immune2
Brain risk
Brain_infiltSig
Brain_metsSig
Lung risk
MS_residual
TBRS
SigA_TGFb
SigB_TGFb
Bone risk
Hypoxia_boneMets
Hypoxia_lungMets
Bone_metsSig
SRCpos
Liver risk
Liver_metsSig
Log odds-ratio
V
BV
B
Mets
c
Cor
−1
0
1
Immune
Proliferation
Stromal
ESR1
TGFb_SigB
PIK3CA
E2F3
HIS
AKTmTOR
M2_hypoxia
Bone_mets
MaSC
Burden
Bone_mets_fc
Bone_mets_upSCP
TBRS
TGFb_SigA
TGFb_cassette
Stroma1
PLAU
Stroma2
BetaCatenin
SRC
SCP2_hypoxia
Bone_mets_inPoorProg
CASP3
VEGF
Liver_metsSig
hNMSC
MYC
IGF1
GENE70
BV
PTEN
AURKA
CIN70
GGI
Cr2009
MAPK
Immune1
STAT1
Immune2
Tcell
Bcell
MotilityPath
brain_pheno
Brain_BBinfilt
RAS
Low_burden
Brain_metsSig

Figure 4


a
Bone & visceral
Visceral only
Bone only
No metastasis
Density
BV module score
Sensitivity
Specificity
AUC=0.77
c
Bone & visceral
Visceral
Bone Only
No metastasis
BV module score in lymph node metastasis
BV module score in primary tumour
rho=0.7
p=0.004
b
BV module score, nCounter
BV module score, WG-DASL
d
Lymph node metastasis, GSE46141
N = 10 11 4
* p=0.04
p=0.1
BV module score
Bone & visceral
Visceral only
Bone only