ARTICLE
Genetics and Genomics

A key genomic subtype associated with lymphovascular invasion in invasive breast cancer

Sasagu Kurozumi[1,2], Chitra Joseph [1], Sultan Sonbul[1], Sami Alsaeed[1], Yousif Kariri[1], Abrar Aljohani[1], Sara Raafat[1], Mansour Alsaleem[1], Angela Ogden[1], Simon J Johnston[1], Mohammed A Aleskandarany[1,3], Takaaki Fujii[2], Ken Shirabe[2], Carlos Caldas[4], Ibraheem Ashankyty[5], Leslie Dalton[6], Ian O Ellis[1], Christine Desmedt[7], Andrew R Green[1], Nigel P Mongan[8,9] and Emad A Rakha[1,3]

BACKGROUND: Lymphovascular invasion (LVI) is associated with the development of metastasis in invasive breast cancer (BC). However, the complex molecular mechanisms of LVI, which overlap with other oncogenic pathways, remain unclear. This study, using available large transcriptomic datasets, aims to identify genes associated with LVI in early-stage BC patients.
METHODS: Gene expression data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort ($n = 1565$) was used as a discovery dataset, and The Cancer Genome Atlas (TCGA; $n = 854$) cohort was used as a validation dataset. Key genes were identified on the basis of differential mRNA expression with respect to LVI status as characterised by histological review. The relationships among LVI-associated genomic subtype, clinicopathological features and patient outcomes were explored.
RESULTS: A 99-gene set was identified that demonstrated significantly different expression between LVI-positive and LVI-negative cases. Clustering analysis with this gene set further divided cases into two molecular subtypes (subtypes 1 and 2), which were significantly associated with pathology-determined LVI status in both cohorts. The 10-year overall survival of subtype 2 was significantly worse than that of subtype 1.
CONCLUSION: This study demonstrates that LVI in BC is associated with a specific transcriptomic profile with potential prognostic value.

BACKGROUND
Outcomes for early-stage breast cancer (BC) patients have improved over recent decades as a result of better diagnostic accuracy, targeted drug therapies, in addition to improvements in early diagnosis.[1] However, the ten-year mortality rates of BC patients remain ~20% which is attributable to the development of metastasis.[2] Several histopathological features have been studied as prognostic factors in BC, including tumour size, lymph node status and histological grade,[3–5] which are strongly associated with outcome. Lymphovascular invasion (LVI) is an early event in the development of metastasis and is a potent prognostic factor.[6] Although the molecular profiles associated with tumour differentiation in terms of histological type and grade and development of lymph node metastasis have been well characterised,[7–9] the molecular mechanisms of LVI and associated genes that may represent therapeutic targets or biomarkers remain to be identified. The main challenge in determining the molecular profiles associated with LVI status in BC stems from the lack of LVI status in the available large-scale

molecular studies in addition to the inherent subjectivity of morphological assessment of LVI status.

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[10] and The Cancer Genome Atlas (TCGA)[11] cohorts are currently the largest genomic and transcriptomic datasets of early-stage BC patients with clinical follow-up. In this study, using these large transcriptomic datasets combined with thorough histological assessment of LVI, we applied bioinformatic analysis to evaluate the genes associated with LVI and assessed the prognostic value of genomic subtype based on LVI status.

METHODS
The METABRIC cohort
In the METABRIC study,[10] mRNA was extracted from primary tumours of female patients, and mRNA expression was evaluated using the Illumina TotalPrep RNA Amplification Kit and Illumina Human HT-12 v3 Expression BeadChips (Ambion,

[1]Nottingham Breast Cancer Research Centre, Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham, UK; [2]Department of General Surgical Science, Gunma University Graduate School of Medicine, Gunma, Japan; [3]Faculty of Medicine, Menoufyia University, Shebin al Kawm, Egypt; [4]Cancer Research UK Cambridge Institute and Department of Oncology, University of Cambridge, Cambridge, UK; [5]Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia; [6]Department of Histopathology, St. David's South Austin Medical Center, Austin, TX, USA; [7]Laboratory for Translational Breast Cancer Research, Department of Oncology, KU Leuven, Leuven, Belgium; [8]Biology and Translational Research, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, UK and [9]Department of Pharmacology, Weill Cornell Medicine, New York, NY, USA
Correspondence: Emad A Rakha (Emad.Rakha@nottingham.ac.uk)

A key genomic subtype associated with lymphovascular invasion in invasive…
S Kurozumi et al.

1130

Warrington, UK). LVI status of 1565 patients within the METABRIC cohort, which were histologically assessed using haematoxylin and eosin (H&E) stained slides. For the Nottingham subset included in METABRIC ($n = 285/1565$), LVI status was additionally assessed by immunohistochemistry (IHC) utilising CD31, CD34 and D2-40,[12] and the final LVI status was confirmed using a combination of multiple H&E tumour sections and IHC. Considering the different methods of LVI assessment, cases were divided into two groups: (1) the Nottingham cases and (2) the remaining METABRIC cases ($n = 1280$). Gene transcript expression levels between LVI-positive and LVI-negative cases were compared for each group, as described in the 'Bioinformatics analysis' section.

The TCGA cohort
The data from the TCGA[11] cohort of female BC patients ($n = 854$) was extracted from the Genomic Data Commons Data Portal and cBioPortal website.[13,14] Briefly, the datasets of mRNA expression from RNASeqV2 were accessed along with de-identified clinical information for several clinicopathological factors and outcomes. Digital H&E-stained slides from the TCGA_BRCA cohort were accessed via the cBioPortal website, and LVI status was quantified by an expert breast pathologist (LD).

Bioinformatics analysis
Analysis of mRNA expression data from METABRIC has been previously described.[10] Differentially expressed genes (DEGs) between LVI-positive and LVI-negative cases were identified using the weighted average difference (WAD) method, and the DEGs were selected according to the WAD ranking.[15,16] Lists of the top 350 genes associated with LVI for the WAD assay in both (1) the Nottingham cases in the METABRIC cohort ($n = 285$) and (2) other METABRIC cases ($n = 1280$) are shown in Supplementary Tables 1 and 2. Overlapping DEGs between the two groups were included in the gene set associated with LVI.

The Cluster 3.0 package was used for clustering and heat map construction.[17] Clustering analysis was performed using METABRIC data as the discovery set and validated using TCGA data as the validation set. TCGA mRNA data were log2-transformed prior to clustering analysis.

For pathway analysis, the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) was used to calculate significantly enriched gene ontologies and pathways associated with these genes.[18,19] The false discovery rate was controlled using the Benjamini–Hochberg procedure in WebGestalt, with an adjusted-$p < 0.01$ considered statistically significant.

Statistical analysis
Statistical analyses were conducted using IBM SPSS Statistics for Windows, version 24.0 (IBM Corp., Armonk, NY, USA). The chi-squared test was used to assess differences among several clinicopathological factors, including LVI status, tumour size, lymph node status, histological grade, oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor 2 (HER2) and molecular subtypes, as stratified by the LVI-associated genomic subtype.

Kaplan–Meier survival curves of 10-year overall survival (OS) were plotted for the METABRIC and TCGA cohorts. The 10-year OS in this study was defined as the day of death within 10 years or the day of completing follow-up from the day of surgery. In univariate and multivariate analyses, 95% confidence intervals (CIs) were assessed using the Cox proportional hazards regression model to determine the associations between clinicopathological factors (LVI status, tumour size, lymph node status, histological grade, ER, PR and HER2), including the LVI-associated genomic subtype and prognosis.

RESULTS
Clinicopathological and prognostic significance of LVI status
In the METABRIC cohort, 635/1,565 (41%) were LVI-positive and 930 (59%) were LVI-negative. The LVI-positivity rate was 41.1% (117/285) in the Nottingham cases and 40.5% (518/1,280) in the remaining METABRIC cases. In the TCGA cohort, 295/854 (35%) patients were LVI-positive and 559 (65%) were LVI-negative. In both cohorts, LVI positivity was significantly associated with large tumour size (METABRIC: $p < 0.0001$; TCGA: $p = 0.00055$), positive nodal status (METABRIC and TCGA: both $p < 0.0001$) and high histological grade (METABRIC and TCGA: both $p < 0.0001$; Supplementary Table 3).

The survival of LVI-positive BC patients was significantly worse compared with LVI-negative patients in the METABRIC (hazard ratio [HR] 1.70, 95% CI 1.45–2.01, $p < 0.0001$; Fig. 1a) and TCGA cohorts (HR 2.2, 95% CI 1.46–3.38, $p = 0.00019$; Fig. 1b). Univariate and multivariate analyses of both METABRIC and TCGA datasets are summarised in Supplementary Table 4. Univariate analysis using the Cox proportional hazards regression model identified LVI-positive status, large tumour size (METABRIC: HR 1.82, 95% CI 1.49–2.21, $p < 0.0001$; TCGA: HR 1.81, 95% CI 1.08–3.04, $p = 0.025$), positive nodal status (METABRIC: HR 2.06, 95% CI 1.74–2.44, $p < 0.0001$; TCGA: HR 1.85, 95% CI 1.20–2.85, $p = 0.0056$), negative ER status (METABRIC: HR 1.66, 95% CI 1.38–1.99, $p < 0.0001$; TCGA: HR 1.89, 95% CI 1.19–2.98, $p = 0.0065$) and negative PR status (METABRIC: HR 1.67, 95% CI 1.42–1.98, $p < 0.0001$; TCGA: HR 1.68, 95% CI 1.08–2.61, $p = 0.020$) as poor prognostic factors in both cohorts. In addition, significant prognostic factors included high histological grade (HR 1.63, 95% CI 1.37–1.93, $p < 0.0001$) and positive HER2 status (HR 1.92, 95% CI 1.54–2.38, $p < 0.0001$) in the METABRIC cohort. LVI positivity was an independent poor prognostic factor in multivariate analysis (METABRIC: HR 1.29, 95% CI 1.07–1.56, $p = 0.0073$; TCGA: HR 2.19, 95% CI 1.32–3.62, $p = 0.0023$; Supplementary Table 4).

Genes associated with LVI
The overlapping DEGs between (1) the Nottingham cases in the METABRIC cohort ($n = 285$) and (2) remaining METABRIC cases ($n = 1280$) included 42 significantly overexpressed and 57 down-regulated genes (Table 1, Supplementary Tables 5 and 6).

The 99 genes in the LVI-related set were significantly associated with gene ontologies, including 'GO: 0005615 Extracellular space', 'GO: 0072562 Blood microparticle' and 'GO: 0031012 Extracellular matrix' (Table 2). All significant pathways existed in the category 'Cellular component' of gene ontology (Supplementary Fig. 1).
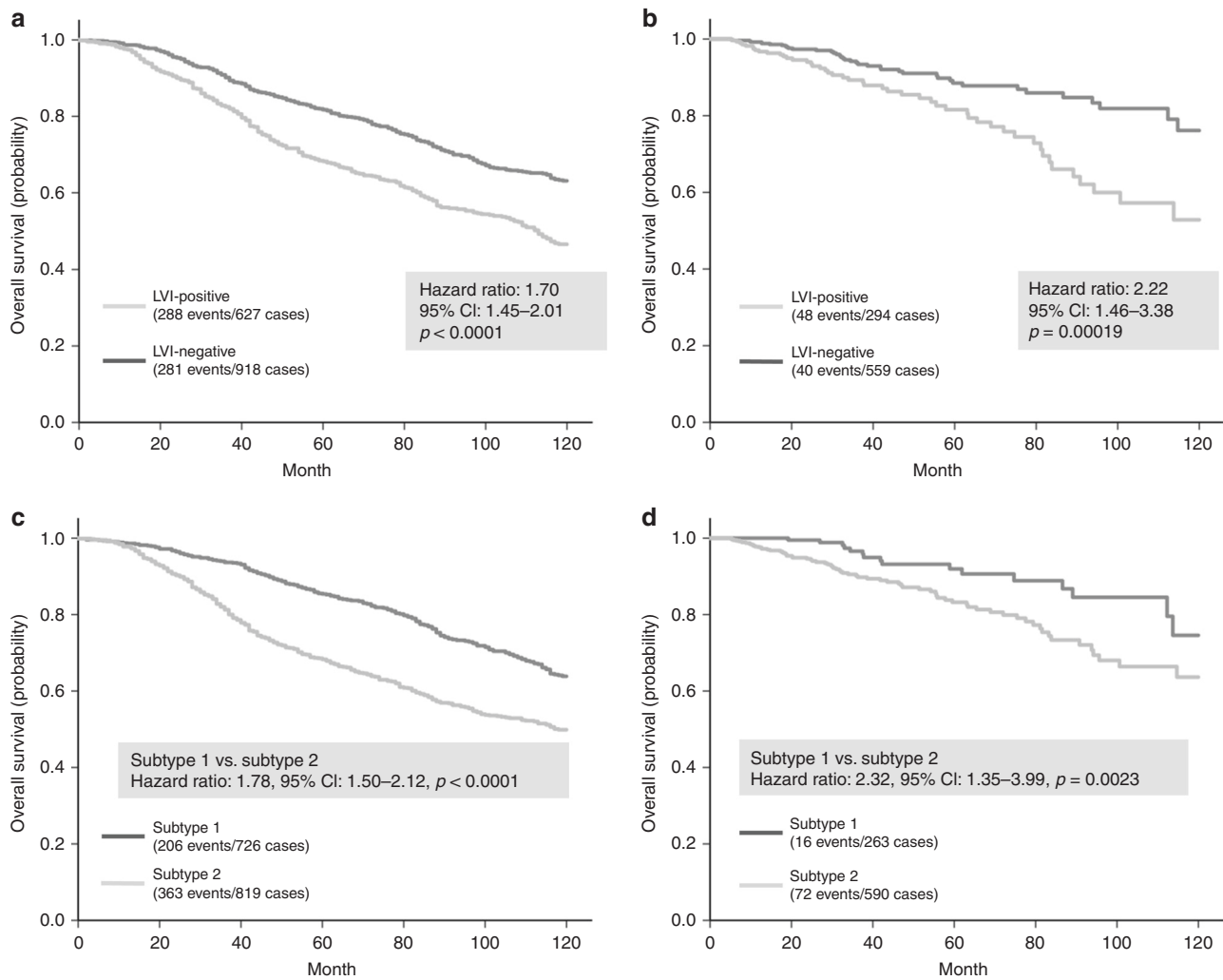
Hierarchical clustering was used to further analyse these 99 genes based on similarity in expression (Fig. 2a). Clustering in the discovery (METABRIC) cohort classified cases into two subtypes, namely, subtypes 1 ($n = 738$ cases; 45%) and 2 ($n = 827$; 55%) (Fig. 2b). The dendrogram of METABRIC cases, in which the pattern of the branches indicates the relationship for each case, is shown in Supplementary Fig. 2.

To validate these results, hierarchical clustering was conducted on the TCGA cohort using the same 99 genes. The dendrogram classifying these 854 cases is shown in Supplementary Fig. 3, again showing the cases split into two groups: subtypes 1 and 2, with 263 (31%) and 591 (69%) cases, respectively (Fig. 2c).

In both cohorts, LVI positivity was significantly more prevalent in subtype 2 tumours than those of subtype 1 (METABRIC and TCGA: $p < 0.0001$; Table 3).

Clinicopathological and prognostic significance of the LVI-related gene sets
In the METABRIC and TCGA cohorts, subtype 2 was significantly associated with large tumour size (both $p < 0.0001$), high histological grade (both $p < 0.0001$), ER negativity (both $p < 0.0001$), PR negativity (both $p < 0.0001$) and HER2 positivity (both $p < 0.0001$;

A key genomic subtype associated with lymphovascular invasion in invasive…
S Kurozumi et al.

1131

Fig. 1 Cumulative survival of BC patients stratified by LVI status. **a** Ten-year overall survival in the METABRIC cases was significantly worse in the LVI-positive group than in the LVI-negative group. **b** In TCGA cases, significant differences were noted in patient overall survival in the LVI-positive and LVI-negative groups. Cumulative survival of breast cancer patients stratified by LVI-related genomic subtypes. **c** Ten-year overall survival in breast cancer patients with LVI-related genomic subtypes. Subtype 2 was significantly worse compared with subtype 1 in the METABRIC cohort. **d** Classification of LVI-related genomic subtype was a significant prognostic factor in the TCGA cohort

Table 3). Interestingly, 69% of luminal B, 95% HER2-enriched and 90% basal-like BC were classified as subtype 2 in the METABRIC cohort.

Patients with LVI-related subtype 2 had a significantly worse prognosis compared with those presenting with subtype 1 tumours in both cohorts (METABRIC: HR 1.78, 95% CI 1.50–2.12, $p < 0.0001$; TCGA: HR 2.32, 95% CI 1.35–3.99, $p = 0.0023$; Fig. 1c, d). In multivariate survival analysis, the LVI-related genomic subtype was an independent poor prognostic factor in both cohorts (METABRIC: HR 1.32, 95% CI 1.07–1.63, $p = 0.0098$; TCGA: HR 2.76, 95% CI 1.19–6.38, $p = 0.018$; Fig. 3 and Supplementary Table 7).

DISCUSSION
In this study, we identified a 99-gene set significantly associated with LVI status in the METABRIC dataset. We validated this finding using the TCGA dataset. LVI is a biomarker for aggressive BC and is considered predictive for metastasis.[20] In other cancer types, gene sets associated with vascular invasion have been previously described, for example in hepatocellular carcinoma[21] and endometrial cancer.[22] Mannelqvist et al.[23] suggested that an 18-gene set associated with vascular invasion in endometrial cancer[22] was consistently associated with hormone receptor negativity, HER2 positivity, basal-like phenotype, reduced patient survival in BC patients. In line with these findings, the present study found that 69% of luminal B, 95% HER2-enriched and 90% basal-like BCs were subtype 2 in the METABRIC cohort. Subtype 2 was significantly associated with LVI positivity. However, of the 18 genes identified in Mannelqvist et al., only different isoforms of matrix metallopeptidase (MMP) and serpin family E member (SERPINE) were present in our 99-gene set.

The underlying molecular mechanisms driving LVI in BC, which are potential therapeutic targets, have yet to be identified. The 99 genes in the LVI-related gene signature from this study are significantly associated with extracellular pathways. In previous work, Klahan et al.[24] suggested their gene set associated with LVI was related to extracellular matrix components using microarray data from 108 BC patients. Epithelial–mesenchymal transition (EMT)-implicated genes in prostate cancer have also been associated with pathways relating to the extracellular space.[25]

A key genomic subtype associated with lymphovascular invasion in invasive…
S Kurozumi et al.

1132

Table 1. List of 99 genes significantly associated with lymphovascular invasion

Upregulated genes			Downregulated genes		
APOC1	KRT7	UCP2	ACTG2	FCGBP	S100A4
APOE	KRT8	YWHAZ	ANG	FGD3	SELENOM
CALML5	LAPTM4B		ANXA1	FOS	SERPINA3
CCNB2	LRRC26		C1S	FST	SERPINE2
CDCA5	LY6E		CDC42EP4	GAS1	SGCE
COX6C	MMP11		CEBPD	GSTP1	SLC40A1
DNAJA4	MX1		CFB	HBA2	SLC44A1
EEF1A2	NME1		CFD	HBB	SRPX
ELF3	NOP56		CLIC6	HLA-DQA1	STC2
ERBB2	PGAP3		CXCL12	IL17RB	SUSD3
GNAS	PITX1		CXCL14	MAOA	TNS3
HMGA1	PTTG1		CYBRD1	MFAP4	TPM2
HMGB3	S100P		CYP4X1	MGP	TXNIP
HSPB1	SCD		DCN	MT1E	UBD
IDH2	SLC52A2		DKK3	NDP	VIM
IFI27	SLC9A3R1		DPYSL2	NINJ1	VTCN1
ISG15	SPDEF		DUSP1	PDGFRL	ZBTB20
KRT18	TM7SF2		EEF1B2	PLGRKT	
KRT18P55	UBE2C		FBLN1	PYCARD	
KRT19	UBE2S		FCER1A	RPL3	

The extracellular matrix comprises a network of structural proteins, and reorganisation of this matrix is required for cancer to progress.[26] The EMT is thought to play an important role in the process of metastasis to distant sites, and certain EMT markers are related to LVI status in BC.[12] In the 99 gene LVI signature set, there are several genes associated with extracellular pathways that are implicated in BC prognosis. For example, heat shock protein 27 (HSPB1), is associated with BC aggressiveness and metastasis.[27] HSPB1 expression is upregulated in the early phase of cell differentiation, which implies that HSPB1 may play an important role in controlling the growth and migration of cancer stem-like cells.[28] Another example is apolipoprotein C1 (APOC1), which is considered as a prognostic biomarker for triple-negative BC.[29] APOC1 is thought to regulate the inflammatory response in cancer tissues,[30] which may be closely related to the elimination of proliferating cancer cells.[31] Upregulation of MMPs is also related to cancer cell proliferation, invasion and epithelial-to-mesenchymal transformation and is indicative of a poor prognosis for BC patients.[32] As an example, MMP-11, which belongs to the MMP family, promotes BC development by inhibiting apoptosis as well as enhancing the migration and invasion of BC cells.[33] Additional functional studies of these genes are necessary to explore the association of aberrant gene function and proteins related to LVI in BC.

Comparison of the METABRIC and TCGA cohorts was a limiting factor in this study, in terms of the different methods used to quantify and statistically analyse gene expression and in the approaches to LVI evaluation. We previously developed a method for the accurate detection of LVI using immunostaining for CD34 or D2-40.[12] In the Nottingham cases, we evaluated LVI status using strict criteria based on both morphology and immunohistochemistry. However, for the TCGA BRCA cohort, we evaluated LVI status using H&E-stained slides alone from the cBioPortal database. Although LVI evaluation using only one H&E slide is feasible, it may be difficult to clearly identify LVI negativity.[34] In present study, the LVI-positivity rates were

Table 2. Gene ontology pathways significantly associated with 99 genes related to lymphovascular invasion

Ontology	Name	Genes in Ontology	Observed	Expected	Enrichment	p-value	Genes
GO:0005615	Extracellular space	1385	23	6.52	3.53	<0.0001	SERPINA3, DCN, CFD, FBLN1, DKK3, ANG, GSTP1, ANXA1, HBB, HSPB1, APOC1, APOE, MFAP4, NDP, SERPINE2, S100A4, CFB, CXCL12, C1S, ACTG2, YWHAZ, STC2, CXCL14
GO:0072562	Blood microparticle	110	7	0.52	13.51	0.00043	SERPINA3, HBB, APOE, CFB, C1S, ACTG2, YWHAZ
GO:0031012	Extracellular matrix	503	11	2.37	4.64	0.0079	DCN, FBLN1, ANG, HSPB1, APOE, MFAP4, MGP, MMP11, NDP, SERPINE2, VIM

A key genomic subtype associated with lymphovascular invasion in invasive…
S Kurozumi et al.

1133
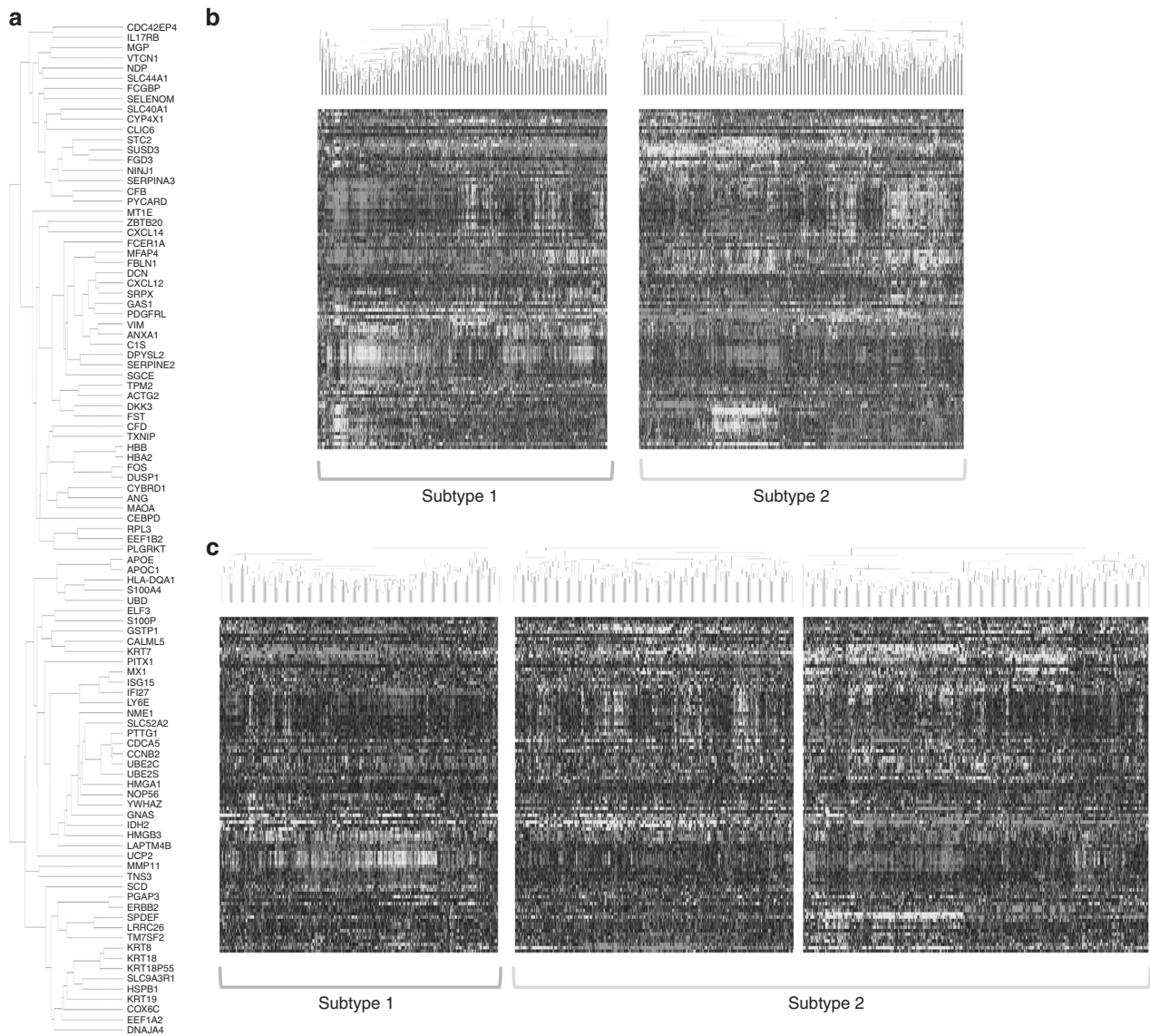
Fig. 2 Cluster analysis of the gene set associated with LVI. **a** The dendrogram of 99 LVI-related genes using METABRIC cohort, in which the pattern of the branches indicates the relationship for each gene. Heat maps in accordance with the LVI-related gene set for the **b** METEBRIC and **c** TCGA cohorts showed that all cases were clearly divided between subtypes 1 and 2 using cluster analysis

closely similar between the Nottingham cases, the remaining METABRIC cases and TCGA_BRCA cases using the different LVI-evaluations. Although our results might suggest the adequacy of LVI evaluation with only one H&E-stained slide, further analysis with the larger cohorts to assess the LVI status using both H&E and IHC slides is necessary to report accurately on LVI status.

Microarrays were used to evaluate mRNA expression in the METABRIC analysis. In contrast, RNA-seq using NGS was used in the TCGA analysis. Microarray platforms have been used and validated for nearly two decades, and this approach has been widely used for evaluating multi-gene expression. Conversely, the unbiased genome-wide RNA-seq method allows for the analysis of all annotated transcripts in addition to the identification of novel transcripts, splice junctions and noncoding RNAs. These techno-logical and methodological differences may underpin the known challenges of relating microarray and RNA sequencing data
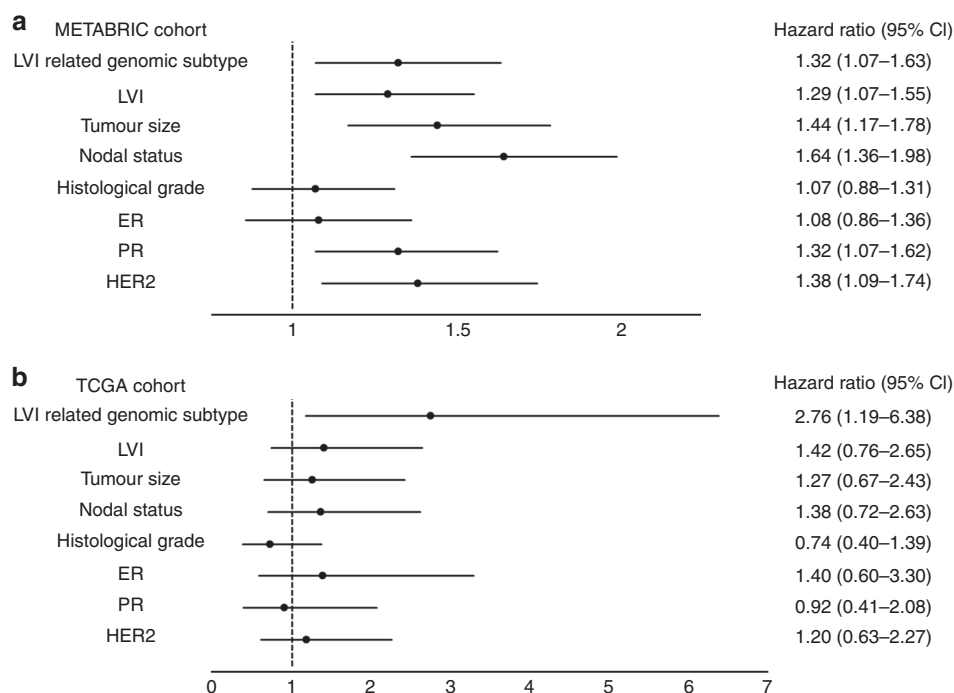
between studies.[35,36] For example, the different approaches can have different lower limits of detection or may encompass different genomic regions. Thus, we cannot assume that the methods are interchangeable, and doing so would require rigorous cross-assay comparisons.[37] Although there is statistical agreement across the different cohorts in the present study, further analysis using identical technologies (microarray and/or NGS assays) may provide clearer validation of the LVI gene signature.

In conclusion, we have confirmed the suitability and prog-nostic significance of our LVI-evaluation approach using the METABRIC and TCGA cohorts. We have determined genomic subtype associated with LVI status and patient outcome in BC, therefore, providing an experimental tool which may serve to unravel the complex gene networks associated with LVI with potential clinical relevance. Consistency between clinical cohorts stratified by LVI-gene signature may be further

A key genomic subtype associated with lymphovascular invasion in invasive…
S Kurozumi et al.

1134

Table 3. Clinicopathological significance of genomic subtypes related to lymphovascular invasion

METABRIC cohort

Factors		LVI-associated genomic subtypes			p-value
		Subtype 1	Subtype 2	Total	
LVI	Positive	262 (35.5%)	373 (45.1%)	635	<0.0001
	Negative	476 (64.5%)	454 (54.9%)	930	
Tumour size	≥2 cm	454 (61.9%)	613 (75.2%)	1067	<0.0001
	<2 cm	279 (38.1%)	202 (24.8%)	481	
Nodal status	Positive	307 (41.7%)	428 (51.9%)	735	<0.0001
	Negative	429 (58.3%)	396 (48.1%)	825	
Histological grade	Grade 3	187 (26.5%)	586 (72.8%)	773	<0.0001
	Grade 1, 2	519 (73.5%)	219 (27.2%)	738	
ER	Positive	707 (95.8%)	497 (60.1%)	1204	<0.0001
	Negative	31 (4.2%)	330 (39.9%)	361	
PR	Positive	533 (72.2%)	295 (35.7%)	828	<0.0001
	Negative	205 (27.8%)	532 (64.3%)	737	
HER2	Positive	20 (2.7%)	168 (20.3%)	188	<0.0001
	Negative	718 (97.3%)	659 (79.7%)	1377	
Molecular subtypes	Luminal A	467 (63.5%)	126 (15.3%)	593	<0.0001
	Luminal B	121 (16.5%)	272 (32.9%)	393	
	HER2-enriched	10 (1.4%)	171 (20.7%)	181	
	Basal-like	24 (3.3%)	222 (26.9%)	246	
	Normal-like	113 (15.4%)	35 (4.2%)	148	

TCGA cohort

Factors		LVI-associated genomic subtypes			p-value
		Subtype 1	Subtype 2	Total	
LVI	Positive	61 (23.2%)	234 (39.6%)	295	<0.0001
	Negative	202 (76.8%)	357 (60.4%)	559	
Tumour size	T 2–4	164 (62.4%)	451 (76.3%)	615	<0.0001
	T 1	99 (37.6%)	140 (23.7%)	239	
Nodal status	Positive	128 (48.9%)	295 (50.3%)	423	0.71
	Negative	134 (51.1%)	292 (49.7%)	426	
Histological grade	Grade 3	28 (11.3%)	324 (56.9%)	352	<0.0001
	Grade 1, 2	219 (88.7%)	245 (43.1%)	464	
ER	Positive	246 (97.6%)	393 (68.7%)	639	<0.0001
	Negative	6 (2.4%)	179 (31.3%)	185	
PR	Positive	235 (94.0%)	311 (54.8%)	546	<0.0001
	Negative	15 (6.0%)	257 (45.2%)	272	
HER2	Positive	20 (9.6%)	113 (23.0%)	133	<0.0001
	Negative	189 (90.4%)	378 (77.0%)	567	

ER oestrogen receptor, *PR* progesterone receptor, *LVI* Lymphovascular invasion

A key genomic subtype associated with lymphovascular invasion in invasive…
S Kurozumi et al.

1135

Fig. 3 Survival analysis based on clinicopathological characteristics including LVI-related genomic subtype. Forest plots showing the hazard ratios and 95% CI of the multivariate survival analyses in **a** the METABRIC cohort and **b** the TCGA cohort. The LVI-related genomic subtype was an independent prognostic factor in both cohorts

improved by using the same definitions and evaluation methods for LVI status.

AUTHOR CONTRIBUTIONS
S.K. participated in its design, experimentation, analysis, interpretation, and manuscript drafting. C.J., S.S., S.A., Y.K., A.A., M.A., M.A.A. and N.P.M. collected the genomic and clinical data and assisted in making the study design and evaluating the results obtained. L.D. mainly performed histopathological examinations. S.R., A.O., S.J., T.F., K.S., C.C., I.A., I.O.E., C.D., and A.R.G. contributed to theoretical organisation of the manuscript. E.A.R. conceived and supervised the study, participated in its design, interpretation, and analysis, including drafting. All authors contributed to drafting and reviewing the manuscript and approved the submitted and final version.

ADDITIONAL INFORMATION

Competing interests: Ibraheem Alshankyty is a consultant/advisory board in Molecular Diagnostics Lab, College of Applied Med. Sci., KAU. The remaining authors declare no competing interests.

Ethics approval and consent to participate: This study was approved by the Nottingham Research Ethics Committee 2 (Reference title: Development of a molecular genetic classification of breast cancer). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data availability: The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES
1. Marshall, D. C., Webb, T. E., Hall, R. A., Salciccioli, J. D., Ali, R., Maruthappu, M. Trends in UK regional cancer mortality 1991–2007. *Br. J. Cancer* **114**, 340–347 (2016).
2. Liedtke, C., Mazouni, C., Hess, K. R., André, F., Tordai, A., Mejia, J. A. et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J. Clin. Oncol.* **26**, 1275–1281 (2008).
3. Wo, J. Y., Chen, K., Neville, B. A., Lin, N. U. & Punglia, R. S. Effect of very small tumor size on cancer-specific mortality in node-positive breast cancer. *J. Clin. Oncol.* **29**, 2619–2627 (2011).
4. Hernandez-Aya, L. F., Chavez-Macgregor, M., Lei, X., Meric-Bernstam, F., Buchholz, T. A., Hsu, L. et al. Nodal status and clinical outcomes in a large cohort of patients with triple-negative breast cancer. *J. Clin. Oncol.* **29**, 2628–2634 (2011).
5. Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V. et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res.* **12**, 207 (2010).
6. Rakha, E. A., Martin, S., Lee, A. H., Morgan, D., Pharoah, P. D., Hodi, Z. et al. The prognostic significance of lymphovascular invasion in invasive breast carcinoma. *Cancer* **118**, 3670–3680 (2012).
7. Yates, L. R. & Desmedt, C. Translational genomics: practical applications of the genomic revolution in breast cancer. *Clin. Cancer Res.* **23**, 2630–2639 (2017).
8. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J. et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl Cancer Inst.* **98**, 262–272 (2006).
9. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).

10. Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

11. Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A. et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).

12. Mohammed, R. A., Martin, S. G., Mahmmod, A. M., Macmillan, R. D., Green, A. R., Paish, E. C. et al. Objective assessment of lymphatic and blood vascular invasion in lymph node-negative breast carcinoma: findings from a large case series with long-term follow-up. *J. Pathol.* **223**, 358–365 (2011).

13. Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

14. Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pl1 (2013).

15. Kadota, K., Nakai, Y. & Shimizu, K. A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms Mol. Biol.* **3**, 8 (2008).

16. Alexander-Dann, B., Pruteanu, L. L., Oerton, E., Sharma, N., Berindan-Neagoe, I., Módos, D. et al. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. *Mol. Omics* **14**, 218–236 (2018).

17. De Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).

18. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33**, W741–W748 (2005).

19. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* **45**, W130–W137 (2017).

20. Aleskandarany, M. A., Sonbul, S. N., Mukherjee, A. & Rakha, E. A. Molecular mechanisms underlying lymphovascular invasion in invasive breast cancer. *Pathobiology* **82**, 113–123 (2015).

21. Mínguez, B., Hoshida, Y., Villanueva, A., Toffanin, S., Cabellos, L., Thung, S. et al. Gene-expression signature of vascular invasion in hepatocellular carcinoma. *J. Hepatol.* **55**, 1325–1331 (2011).

22. Mannelqvist, M., Stefansson, I. M., Bredholt, G., Hellem Bø, T., Oyan, A. M., Jonassen, I. et al. Gene expression patterns related to vascular invasion and aggressive features in endometrial cancer. *Am. J. Pathol.* **178**, 861–871 (2011).

23. Mannelqvist, M., Wik, E., Stefansson, I. M. & Akslen, L. A. An 18-gene signature for vascular invasion is associated with aggressive features and reduced survival in breast cancer. *PLoS ONE* **9**, e98787 (2014).

24. Klahan, S., Wong, H. S., Tu, S. H., Chou, W. H., Zhang, Y. F., Ho, T. F. et al. Identification of genes and pathways related to lymphovascular invasion in breast cancer patients: a bioinformatics analysis of gene expression profiles. *Tumour Biol.* **39**, 1010428317705573 (2017).

25. Zhao, M., Liu, Y. & Qu, H. Expression of epithelial-mesenchymal transition-related genes increases with copy number in multiple cancer types. *Oncotarget* **7**, 24688–24699 (2016).

26. Jena, M. K. & Janjanam, J. Role of extracellular matrix in breast cancer development: a brief update. *F1000Res* **7**, 274 (2018).

27. Musiani, D., Konda, J. D., Pavan, S., Torchiaro, E., Sassi, F., Noghero, A. et al. Heat-shock protein 27 (HSP27, HSPB1) is up-regulated by MET kinase inhibitors and confers resistance to MET-targeted therapy. *FASEB J.* **28**, 4055–4067 (2014).

28. Wei, L., Liu, T. T., Wang, H. H., Hong, H. M., Yu, A. L., Feng, H. P. et al. Hsp27 participates in the maintenance of breast cancer stem cells through regulation of epithelial-mesenchymal transition and nuclear factor-κB. *Breast Cancer Res.* **13**, R101 (2011).

29. Song, D., Yue, L., Zhang, J., Ma, S., Zhao, W., Guo, F. et al. Diagnostic and prognostic significance of serum apolipoprotein C-I in triple-negative breast cancer based on mass spectrometry. *Cancer Biol. Ther.* **17**, 635–647 (2016).

30. Ko, H. L., Wang, Y. S., Fong, W. L., Chi, M. S., Chi, K. H., Kao, S. J. Apolipoprotein C1 (APOC1) as a novel diagnostic and prognostic biomarker for lung cancer: a marker phase I trial. *Thorac. Cancer* **5**, 500–508 (2014).

31. Kurozumi, S., Fujii, T., Matsumoto, H., Inoue, K., Kurosumi, M., Horiguchi, J. et al. Significance of evaluating tumor-infiltrating lymphocytes (TILs) and programmed cell death-ligand 1 (PD-L1) expression in breast cancer. *Med. Mol. Morphol.* **50**, 185–194 (2017).

32. Merdad, A., Karim, S., Schulten, H. J., Dallol, A., Buhmeida, A., Al-Thubaity, F. et al. Expression of matrix metalloproteinases (MMPs) in primary human breast cancer: MMP-9 as a potential biomarker for cancer invasion and metastasis. *Anticancer Res.* **34**, 1355–1366 (2014).

33. Zhang, X., Huang, S., Guo, J., Zhou, L., You, L., Zhang, T. et al. Insights into the distinct roles of MMP-11 in tumor biology and future therapeutics (Review). *Int J. Oncol.* **48**, 1783–1793 (2016).

34. Rakha, E. A., Abbas, A., Pinto Ahumada, P., ElSayed, M. E., Colman, D., Pinder, S. E. et al. Diagnostic concordance of reporting lymphovascular invasion in breast cancer. *J. Clin. Pathol.* **71**, 802–805 (2018).

35. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9**, e78644 (2014).

36. Wolff, A., Bayerlová, M., Gaedcke, J., Kube, D., Beißbarth, T. A comparative study of RNA-Seq and microarray data analysis on the two examples of rectal-cancer patients and Burkitt Lymphoma cells. *PLoS ONE* **13**, e0197162 (2018).

37. Merker, J. D., Oxnard, G. R., Compton, C., Diehn, M., Hurley, P., Lazar, A. J. et al. Circulating tumor DNA analysis in patients with cancer: American Society of Clinical Oncology and College of American Pathologists Joint Review. *J. Clin. Oncol.* **36**, 1631–1641 (2018).