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 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]

[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, Texas, 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
[9]Department of Pharmacology, Weill Cornell Medicine, New York, USA

Corresponding author:
Prof Emad Rakha
Department of Histopathology,
Division of Cancer and Stem Cells, School of Medicine,
The University of Nottingham and Nottingham University Hospitals NHS Trust,
Nottingham City Hospital, Nottingham, NG5 1PB, UK
Email: Emad.Rakha@nottingham.ac.uk

ABSTRACT

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 characterized 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.

KEY WORDS: invasive breast cancer, lymphovascular invasion, gene signature

INTRODUCTION

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.

MATERIALS AND 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, Warrington, UK). LVI status of 1,565 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 = 285/1,565), 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 = 1,280). 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, $p < 0.0001$; Figure 1-a) and TCGA cohorts (HR 2.2, 95% CI 1.46–3.38, $p = 0.00019$; Figure 1-b). 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 = 1,280) included 42 significantly overexpressed and 57 downregulated 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 Figure 1).

Hierarchical clustering was used to further analyse these 99 genes based on similarity in expression (Figure 2-a). Clustering in the discovery (METABRIC) cohort classified cases into two subtypes, namely, subtypes 1 (n = 738 cases; 45%) and 2 (n = 827; 55%) (Figure 2-b). The dendrogram of METABRIC cases, in which the pattern of the branches indicates the relationship for each case, is shown in Supplementary Figure 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 Figure 3, again showing the cases split into two groups: subtypes 1 and 2, with 263 (31%) and 591 (69%) cases, respectively (Figure 2-c).

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 $p < 0.0001$; 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, $p = 0.0023$; Figure 1-c, 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; Figure 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]. 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 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 technological 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 prognostic 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 improved by using the same definitions and evaluation methods for LVI status.

Additional Information

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.

Availability of data and material

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

Ibraheem Alshankyty is a consultant/advisory board in Molecular Diagnostics Lab, College of Applied Med. Sci., KAU. There were no competing interests for the other authors.

Funding

This work was supported by the University of Nottingham (Nottingham Life Cycle 6).

Authors' contributions

SK participated in its design, experimentation, analysis, interpretation, and manuscript drafting. CJ, SS, SA, YK, AA, MA, MAA and NPM collected the genomic and clinical data and assisted in making the study design and evaluating the results obtained. LD mainly performed histopathological examinations. SR, AO, SJ, TF, KS, CC, IA, IOE, CD and ARG contributed to theoretical organization of the manuscript. EAR 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.

Acknowledgements

We thank the Nottingham Health Science Biobank and Breast Cancer Now Tissue Bank for providing the tissue samples. We also thank the University of Nottingham (Nottingham Life Cycle 6) and the METABRIC group members.

Supplementary Information

Supplementary information is available at the British Journal of Cancer’s website

REFERENCES

1. Marshall DC, Webb TE, Hall RA, Salciccioli JD, Ali R, Maruthappu M. Trends in UK regional cancer mortality 1991-2007. Br J Cancer 2016, **114**, 340-347
2. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol 2008, **26**, 1275-1281
3. Wo JY, Chen K, Neville BA, Lin NU, Punglia RS. Effect of very small tumor size on cancer-specific mortality in node-positive breast cancer. J Clin Oncol 2011, **29**, 2619-2627.
4. Hernandez-Aya LF, Chavez-Macgregor M, Lei X, Meric-Bernstam F, Buchholz TA, Hsu L et al. Nodal status and clinical outcomes in a large cohort of patients with triple-negative breast cancer. J Clin Oncol 2011, **29**, 2628-2634
5. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res 2010, **12**, 207
6. Rakha EA, Martin S, Lee AH, Morgan D, Pharoah PD, Hodi Z et al. The prognostic significance of lymphovascular invasion in invasive breast carcinoma. Cancer 2012, **118**, 3670-3680
7. Yates LR, Desmedt C. Translational genomics: practical applications of the genomic revolution in breast cancer. Clin Cancer Res 2017, **23**, 2630-2639
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 2006, **98**, 262-272
9. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE et al. Predicting cancer outcomes from histology and

genomics using convolutional networks. Proc Natl Acad Sci U S A 2018, **115**, E2970-E2979

10. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012, **486**, 346-352
11. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015, **163**, 506-519
12. Mohammed RA, Martin SG, Mahmmod AM, Macmillan RD, Green AR, Paish EC 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 2011, **223**, 358-365
13. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012, **2**, 401-404
14. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013, **6**, pl1
15. Kadota K, Nakai Y, Shimizu K. A weighted average difference method for detecting differentially expressed genes from microarray data. Algorithms Mol Biol 2008, **3**, 8
16. Alexander-Dann B, Pruteanu LL, 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 2018, **14**, 218-236

17. De Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics 2004, **20**, 1453–1454
18. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005, **33**, W741-W748
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 2017, **45**, W130-W137
20. Aleskandarany MA, Sonbul SN, Mukherjee A, Rakha EA. Molecular mechanisms underlying lymphovascular invasion in invasive breast cancer. Pathobiology 2015, **82**, 113-123
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 2011, **55**, 1325-1331
22. Mannelqvist M, Stefansson IM, Bredholt G, Hellem Bø T, Oyan AM, Jonassen I et al. Gene expression patterns related to vascular invasion and aggressive features in endometrial cancer. Am J Pathol 2011, **178**, 861-871
23. Mannelqvist M, Wik E, Stefansson IM, Akslen LA. An 18-gene signature for vascular invasion is associated with aggressive features and reduced survival in breast cancer. PLoS One 2014, **9**, e98787
24. Klahan S, Wong HS, Tu SH, Chou WH, Zhang YF, Ho TF et al. Identification of genes and pathways related to lymphovascular invasion in breast cancer patients: A bioinformatics analysis of gene expression profiles. Tumour Biol 2017, **39**, 1010428317705573

25. Zhao M, Liu Y, Qu H. Expression of epithelial-mesenchymal transition-related genes increases with copy number in multiple cancer types. Oncotarget 2016, **7**, 24688-24699
26. Jena MK, Janjanam J. Role of extracellular matrix in breast cancer development: a brief update. F1000Res 2018, **7**, 274
27. Musiani D, Konda JD, 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 2014, **28**, 4055-4067
28. Wei L, Liu TT, Wang HH, Hong HM, Yu AL, Feng HP 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 2011, **13**, R101
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 2016, **17**, 635-647
30. Ko HL, Wang YS, Fong WL, Chi MS, Chi KH, Kao SJ. Apolipoprotein C1 (APOC1) as a novel diagnostic and prognostic biomarker for lung cancer: A marker phase I trial. Thorac Cancer 2014, **5**, 500-508
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 2017, **50**, 185-194
32. Merdad A, Karim S, Schulten HJ, 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 2014, **34**, 1355-1366

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 2016, **48**, 1783-1793
34. Rakha EA, Abbas A, Pinto Ahumada P, ElSayed ME, Colman D, Pinder SE et al. Diagnostic concordance of reporting lymphovascular invasion in breast cancer. J Clin Pathol 2018, **71**, 802-805
35. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 2014, **9**, e78644
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 2018, **13**, e0197162
37. Merker JD, Oxnard GR, Compton C, Diehn M, Hurley P, Lazar AJ 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 2018, **36**, 1631-1641

Figure legends

Figure 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 gene signatures. Subtype 2 was significantly worse compared with subtype 1 in the METABRIC cohort. (d) Classification of LVI-related gene signature was a significant prognostic factor in the TCGA cohort.

Figure 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.

Figure 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.

Supplementary file legends

Supplementary Table 1. List of top 350 genes significantly associated with lymphovascular invasion in the Nottingham cohort

Supplementary Table 2. List of top 350 genes significantly associated with lymphovascular invasion in the remaining METABRIC cases

Supplementary Table 3. Correlation between lymphovascular invasion and clinicopathological characteristics

Supplementary Table 4. Survival analysis based on clinicopathological characteristics including lymphovascular invasion

Supplementary Table 5. Full gene name list of the 99 genes significantly associated with lymphovascular invasion

Supplementary Table 6. Mean value, standard error of the mean (SEM), subtraction and weighted average difference (WAD) ranking in the 99 genes significantly associated with lymphovascular invasion

Supplementary Table 7. Survival analysis based on clinicopathological characteristics including LVI-related genomic subtype

Supplementary Figure 1. Significant pathways associated with LVI-related gene set

Supplementary Figure 2. The dendrogram of METABRIC cases for hierarchical clustering analysis

Supplementary Figure 3. The dendrogram of TCGA cases for hierarchical clustering analysis

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*	

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*

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

METABRIC cohort						TCGA cohort					
Factors		LVI-associated genomic subtypes			*p*-value	Factors		LVI associated genomic subtypes			*p*-value
		Subtype 1	Subtype 2	Total				Subtype 1	Subtype 2	Total	
LVI	**Positive**	262 (35.5%)	373 (45.1%)	635	<0.0001	**LVI**	**Positive**	61 (23.2%)	234 (39.6%)	295	<0.0001
	Negative	476 (64.5%)	454 (54.9%)	930			**Negative**	202 (76.8%)	357 (60.4%)	559	
Tumour size	**≥ 2cm**	454 (61.9%)	613 (75.2%)	1067	<0.0001	**Tumour size**	**T 2-4**	164 (62.4%)	451 (76.3%)	615	<0.0001
	< 2cm	279 (38.1%)	202 (24.8%)	481			**T 1**	99 (37.6%)	140 (23.7%)	239	
Nodal status	**Positive**	307 (41.7%)	428 (51.9%)	735	<0.0001	**Nodal status**	**Positive**	128 (48.9%)	295 (50.3%)	423	0.71
	Negative	429 (58.3%)	396 (48.1%)	825			**Negative**	134 (51.1%)	292 (49.7%)	426	
Histological grade	**Grade 3**	187 (26.5%)	586 (72.8%)	773	<0.0001	**Histological grade**	**Grade 3**	28 (11.3%)	324 (56.9%)	352	<0.0001
	Grade 1, 2	519 (73.5%)	219 (27.2%)	738			**Grade 1, 2**	219 (88.7%)	245 (43.1%)	464	
ER	**Positive**	707 (95.8%)	497 (60.1%)	1204	<0.0001	**ER**	**Positive**	246 (97.6%)	393 (68.7%)	185	<0.0001
	Negative	31 (4.2%)	330 (39.9%)	361			**Negative**	6 (2.4%)	179 (31.3%)	639	
PR	**Positive**	533 (72.2%)	295 (35.7%)	828	<0.0001	**PR**	**Positive**	235 (94.0%)	311 (54.8%)	546	<0.0001
	Negative	205 (27.8%)	532 (64.3%)	737			**Negative**	15 (6.0%)	257 (45.2%)	272	

HER2	**Positive**	20 (2.7%)	168 (20.3%)	188	<0.0001	**HER2**	**Positive**	20 (9.6%)	113 (23.0%)	133	<0.0001
	Negative	718 (97.3%)	659 (79.7%)	1377			**Negative**	189 (90.4%)	378 (77.0%)	567	
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							

Abbreviations: ER, Oestrogen receptor; PR, Progesterone receptor; LVI, Lymphovascular invasion.

Supplementary Table 7. Survival analysis based on clinicopathological characteristics including LVI-related genomic subtype

METABRIC cohort					TCGA cohort				
Factors		Multivariate analysis			Factors		Multivariate analysis		
		Hazard Ratio	95% CI	*p*-value			Hazard Ratio	95% CI	*p*-value
LVI related genomic subtype	Subtype 1	Reference			LVI related genomic subtype	Subtype 1	Reference		
	Subtype 2	1.32	1.07-1.63	0.0098		Subtype 2	2.76	1.19-6.38	0.018
LVI	Negative	Reference			LVI	Negative	Reference		
	Positive	1.29	1.07-1.55	0.0075		Positive	1.42	0.76-2.65	0.28
Tumour size	< 2cm	Reference			Tumour size	T1	Reference		
	≥ 2cm	1.44	1.17-1.78	0.00055		T2-4	1.27	0.67-2.43	0.47
Nodal status	Negative	Reference			Nodal status	Negative	Reference		
	Positive	1.64	1.36-1.98	<0.0001		Positive	1.38	0.72-2.63	0.33
Histological grade	Grade 1, 2	Reference			Histological grade	Grade 1, 2	Reference		
	Grade 3	1.07	0.88-1.31	0.49		Grade 3	0.74	0.40-1.39	0.35
ER	Positive	Reference			ER	Positive	Reference		
	Negative	1.08	0.86-1.36	0.51		Negative	1.40	0.60-3.30	0.44
PR	Positive	Reference			PR	Positive	Reference		
	Negative	1.32	1.07-1.62	0.0095		Negative	0.92	0.41-2.08	0.84
HER2	Negative	Reference			HER2	Negative	Reference		
	Positive	1.38	1.09-1.74	0.0074		Positive	1.20	0.63-2.27	0.58

Abbreviations: ER, Oestrogen receptor; PR, Progesterone receptor; LVI, Lymphovascular invasion.

Supplementary Table 6. Mean value, standard error of the mean (SEM), subtraction and weighted average difference (WAD) ranking in the 99 genes significantly associated with lymphovascular invasion

Upregulated genes												
	Nottingham cases						Remaining METABRIC cases					
LVI	Positive		Negative		Subtraction	WAD ranking	Positive		Negative		Subtraction	WAD ranking
Genes	Mean	SEM	Mean	SEM			Mean	SEM	Mean	SEM		
APOC1	10.28	0.91	10.00	0.98	0.28	25	9.98	0.99	9.84	1.04	0.14	64
APOE	11.73	0.67	11.53	0.74	0.20	31	11.64	0.82	11.54	0.82	0.10	81
CALML5	7.56	2.05	7.04	1.88	0.52	61	7.50	2.16	7.16	1.93	0.34	42
CCNB2	8.25	0.92	8.04	0.93	0.21	326	8.07	0.90	7.90	1.01	0.16	186
CDCA5	8.50	0.96	8.28	0.98	0.22	232	8.48	0.95	8.31	1.04	0.17	100
COX6C	12.98	0.63	12.89	0.65	0.09	263	12.89	0.68	12.83	0.67	0.05	277
DNAJA4	9.17	0.75	8.98	0.81	0.20	205	9.20	0.82	9.09	0.81	0.11	227
EEF1A2	9.08	2.00	8.46	1.92	0.63	3	9.05	2.09	8.90	2.04	0.15	105
ELF3	8.69	0.73	8.49	0.86	0.20	274	8.92	0.79	8.81	0.82	0.11	287
ERBB2	10.83	1.59	10.63	1.33	0.20	65	10.92	1.46	10.62	1.22	0.30	1
GNAS	12.75	0.45	12.62	0.41	0.13	101	12.93	0.53	12.87	0.48	0.05	262
HMGA1	8.48	0.64	8.28	0.77	0.20	303	8.50	0.75	8.38	0.77	0.12	298
HMGB3	7.72	0.89	7.38	0.83	0.34	166	7.64	0.88	7.48	0.91	0.16	327
HSPB1	12.26	0.73	12.07	0.74	0.19	32	12.21	0.79	12.11	0.84	0.10	53
IDH2	9.63	0.88	9.45	0.72	0.18	179	9.63	0.83	9.51	0.86	0.12	131
IFI27	11.95	1.30	11.73	1.15	0.21	24	11.63	1.40	11.57	1.35	0.06	334
ISG15	9.74	1.32	9.54	1.35	0.20	140	9.75	1.35	9.61	1.36	0.14	69
KRT18	11.70	0.96	11.51	1.07	0.20	36	11.84	1.05	11.75	1.07	0.09	79
KRT18P55	10.35	0.97	10.16	1.02	0.18	120	10.17	1.08	9.99	1.10	0.18	18

KRT19	**12.51**	**1.14**	**12.37**	**1.33**	**0.14**	**88**	**12.58**	**1.22**	**12.51**	**1.26**	**0.07**	**149**
KRT7	**9.45**	**1.44**	**9.27**	**1.35**	**0.18**	**223**	**9.36**	**1.53**	**9.25**	**1.46**	**0.11**	**187**
KRT8	**10.16**	**0.92**	**9.96**	**0.99**	**0.21**	**90**	**10.46**	**0.99**	**10.32**	**1.01**	**0.14**	**43**
LAPTM4B	**10.43**	**1.06**	**10.23**	**0.92**	**0.20**	**85**	**10.18**	**1.21**	**10.09**	**1.13**	**0.09**	**246**
LRRC26	**9.92**	**1.63**	**9.72**	**1.51**	**0.20**	**125**	**9.92**	**1.55**	**9.83**	**1.53**	**0.09**	**257**
LY6E	**10.45**	**1.00**	**10.25**	**0.95**	**0.20**	**72**	**10.62**	**1.03**	**10.45**	**1.00**	**0.17**	**14**
MMP11	**10.53**	**1.39**	**10.38**	**1.50**	**0.15**	**184**	**10.47**	**1.36**	**10.27**	**1.52**	**0.19**	**11**
MX1	**11.07**	**1.30**	**10.75**	**1.30**	**0.32**	**9**	**11.17**	**1.32**	**11.05**	**1.29**	**0.13**	**37**
NME1	**11.55**	**0.73**	**11.43**	**0.71**	**0.12**	**217**	**11.31**	**0.72**	**11.20**	**0.67**	**0.11**	**67**
NOP56	**9.75**	**0.47**	**9.59**	**0.48**	**0.16**	**242**	**9.95**	**0.56**	**9.87**	**0.58**	**0.09**	**286**
PGAP3	**8.91**	**1.39**	**8.69**	**1.15**	**0.21**	**199**	**8.81**	**1.26**	**8.56**	**0.95**	**0.25**	**20**
PITX1	**9.29**	**1.55**	**8.84**	**1.60**	**0.45**	**11**	**9.34**	**1.59**	**9.23**	**1.61**	**0.10**	**253**
PTTG1	**9.29**	**0.87**	**9.12**	**0.94**	**0.17**	**257**	**9.10**	**0.91**	**8.93**	**1.01**	**0.16**	**73**
S100P	**9.70**	**2.32**	**9.26**	**2.31**	**0.45**	**5**	**9.52**	**2.34**	**9.22**	**2.24**	**0.30**	**3**
SCD	**10.88**	**0.97**	**10.75**	**0.92**	**0.14**	**200**	**10.92**	**1.11**	**10.78**	**1.03**	**0.14**	**30**
SLC52A2	**9.18**	**0.61**	**9.02**	**0.63**	**0.16**	**309**	**9.29**	**0.72**	**9.13**	**0.67**	**0.16**	**76**
SLC9A3R1	**10.77**	**1.04**	**10.59**	**1.01**	**0.18**	**108**	**10.97**	**0.98**	**10.87**	**1.02**	**0.10**	**109**
SPDEF	**9.48**	**1.39**	**9.34**	**1.46**	**0.14**	**346**	**9.74**	**1.34**	**9.55**	**1.45**	**0.19**	**23**
TM7SF2	**8.70**	**0.93**	**8.44**	**0.88**	**0.26**	**132**	**8.70**	**0.95**	**8.56**	**0.91**	**0.14**	**170**
UBE2C	**9.27**	**1.09**	**9.03**	**1.16**	**0.24**	**106**	**9.25**	**1.17**	**8.99**	**1.31**	**0.25**	**10**
UBE2S	**9.29**	**0.71**	**9.02**	**0.73**	**0.27**	**70**	**9.35**	**0.84**	**9.21**	**0.85**	**0.14**	**94**
UCP2	**8.93**	**0.91**	**8.71**	**0.90**	**0.22**	**189**	**9.10**	**0.94**	**8.94**	**0.92**	**0.16**	**84**
YWHAZ	**12.00**	**0.59**	**11.84**	**0.58**	**0.15**	**79**	**12.08**	**0.63**	**11.95**	**0.62**	**0.13**	**17**
Downregulated genes												
	Nottingham cases						**Remaining METABRIC cases**					

LVI	Positive		Negative		Subtraction	WAD ranking	Positive		Negative		Subtraction	WAD ranking
Genes	Mean	SEM	Mean	SEM			Mean	SEM	Mean	SEM		
ACTG2	8.75	2.50	9.01	1.81	-0.26	100	8.48	1.62	8.78	1.63	-0.30	7
ANG	8.19	0.94	8.44	1.01	-0.25	186	8.14	1.07	8.29	1.14	-0.15	179
ANXA1	10.91	0.68	11.08	0.71	-0.17	95	10.45	0.91	10.58	0.99	-0.13	50
C1S	10.11	0.93	10.34	0.88	-0.24	53	9.67	1.02	9.77	1.10	-0.10	217
CDC42EP4	10.22	0.38	10.37	0.62	-0.15	197	10.40	0.66	10.48	0.65	-0.08	232
CEBPD	10.09	0.53	10.21	0.70	-0.13	316	10.12	0.84	10.20	0.81	-0.08	333
CFB	10.10	2.45	10.51	1.48	-0.41	4	10.42	1.70	10.54	1.64	-0.12	72
CFD	9.48	1.39	9.87	1.38	-0.40	10	9.24	1.29	9.42	1.36	-0.19	33
CLIC6	8.17	4.43	8.54	2.17	-0.37	50	8.15	2.21	8.43	2.24	-0.28	19
CXCL12	9.44	1.16	9.72	1.00	-0.28	45	9.05	1.10	9.23	1.20	-0.18	49
CXCL14	8.31	2.30	8.67	1.52	-0.36	49	8.18	1.57	8.39	1.61	-0.21	65
CYBRD1	9.73	1.03	9.91	1.03	-0.18	162	9.68	1.15	9.76	1.19	-0.09	318
CYP4X1	8.44	3.82	8.77	1.89	-0.32	64	8.65	1.90	8.88	1.94	-0.23	24
DCN	9.07	1.34	9.23	1.24	-0.16	325	8.46	1.33	8.64	1.43	-0.19	75
DKK3	9.45	0.94	9.72	0.88	-0.27	54	9.07	0.91	9.22	0.93	-0.15	90
DPYSL2	9.82	0.43	9.98	0.60	-0.16	214	9.73	0.68	9.85	0.76	-0.12	107
DUSP1	10.32	0.90	10.44	0.96	-0.12	348	9.89	1.40	10.04	1.45	-0.15	48
EEF1B2	11.20	0.34	11.33	0.52	-0.12	219	10.93	0.78	11.01	0.80	-0.08	159
FBLN1	10.59	1.04	10.86	0.93	-0.27	17	10.51	1.04	10.63	1.12	-0.11	86
FCER1A	7.41	1.20	7.76	1.27	-0.36	144	6.95	1.07	7.15	1.22	-0.21	293
FCGBP	8.72	2.50	9.11	1.64	-0.39	19	8.76	1.61	8.96	1.61	-0.20	38
FGD3	8.81	1.30	9.19	1.11	-0.38	21	9.19	1.20	9.30	1.20	-0.12	173
FOS	10.12	1.85	10.24	1.37	-0.12	349	9.53	1.66	9.74	1.69	-0.21	13

FST	**8.22**	**1.20**	**8.60**	**1.10**	**-0.38**	**41**	**8.04**	**1.02**	**8.24**	**1.03**	**-0.20**	**89**
GAS1	**8.92**	**1.05**	**9.11**	**0.92**	**-0.19**	**227**	**8.45**	**1.06**	**8.63**	**1.11**	**-0.18**	**80**
GSTP1	**10.76**	**1.21**	**10.99**	**0.93**	**-0.23**	**33**	**10.61**	**1.20**	**10.80**	**1.10**	**-0.19**	**8**
HBA2	**9.40**	**2.26**	**9.55**	**1.48**	**-0.15**	**308**	**9.03**	**1.52**	**9.26**	**1.57**	**-0.23**	**16**
HBB	**9.34**	**2.23**	**9.59**	**1.47**	**-0.24**	**73**	**8.62**	**1.62**	**8.91**	**1.69**	**-0.29**	**9**
HLA-DQA1	**10.37**	**1.03**	**10.52**	**0.99**	**-0.15**	**188**	**10.04**	**1.30**	**10.12**	**1.31**	**-0.08**	**347**
IL17RB	**7.50**	**1.18**	**7.75**	**1.17**	**-0.25**	**340**	**7.56**	**1.06**	**7.73**	**1.04**	**-0.17**	**224**
MAOA	**7.42**	**1.62**	**7.83**	**1.25**	**-0.41**	**84**	**7.48**	**1.33**	**7.67**	**1.37**	**-0.19**	**178**
MFAP4	**8.48**	**1.52**	**8.73**	**1.26**	**-0.25**	**141**	**8.31**	**1.17**	**8.50**	**1.31**	**-0.19**	**83**
MGP	**12.95**	**1.44**	**13.28**	**1.10**	**-0.33**	**2**	**12.73**	**1.34**	**12.87**	**1.43**	**-0.14**	**6**
MT1E	**9.78**	**1.26**	**10.10**	**1.10**	**-0.32**	**15**	**9.75**	**1.23**	**9.84**	**1.19**	**-0.09**	**225**
NDP	**6.81**	**2.19**	**7.26**	**1.53**	**-0.45**	**152**	**6.91**	**1.65**	**7.14**	**1.68**	**-0.23**	**229**
NINJ1	**10.21**	**0.29**	**10.33**	**0.53**	**-0.12**	**314**	**10.40**	**0.55**	**10.49**	**0.53**	**-0.09**	**151**
PDGFRL	**8.95**	**1.04**	**9.23**	**0.95**	**-0.28**	**67**	**8.51**	**1.01**	**8.65**	**1.09**	**-0.15**	**146**
PLGRKT	**9.88**	**0.38**	**10.05**	**0.63**	**-0.18**	**160**	**9.62**	**0.78**	**9.72**	**0.78**	**-0.10**	**182**
PYCARD	**9.90**	**0.83**	**10.10**	**0.88**	**-0.19**	**122**	**10.05**	**0.94**	**10.13**	**0.94**	**-0.08**	**323**
RPL3	**12.76**	**0.29**	**12.89**	**0.46**	**-0.13**	**112**	**12.70**	**0.53**	**12.76**	**0.53**	**-0.06**	**212**
S100A4	**10.87**	**0.71**	**11.00**	**0.76**	**-0.13**	**209**	**10.46**	**0.86**	**10.55**	**0.90**	**-0.10**	**134**
SELENOM	**10.09**	**0.49**	**10.33**	**0.69**	**-0.23**	**55**	**10.20**	**0.70**	**10.35**	**0.66**	**-0.15**	**34**
SERPINA3	**12.09**	**3.18**	**12.25**	**1.79**	**-0.16**	**63**	**12.05**	**1.79**	**12.27**	**1.68**	**-0.21**	**2**
SERPINE2	**9.97**	**0.89**	**10.28**	**0.96**	**-0.31**	**16**	**9.82**	**1.03**	**9.93**	**1.08**	**-0.11**	**123**
SGCE	**8.87**	**0.97**	**9.09**	**0.89**	**-0.21**	**176**	**8.49**	**1.10**	**8.63**	**1.14**	**-0.14**	**168**
SLC40A1	**9.83**	**1.18**	**10.07**	**1.29**	**-0.24**	**59**	**9.71**	**1.32**	**9.83**	**1.40**	**-0.12**	**113**
SLC44A1	**11.03**	**0.25**	**11.24**	**0.50**	**-0.21**	**46**	**10.93**	**0.57**	**11.01**	**0.55**	**-0.09**	**138**
SRPX	**8.24**	**1.02**	**8.43**	**0.93**	**-0.20**	**328**	**7.84**	**1.00**	**8.04**	**1.14**	**-0.19**	**121**

STC2	**9.26**	**3.41**	**9.70**	**1.94**	**-0.44**	**6**	**9.73**	**1.96**	**9.90**	**1.93**	**-0.17**	**28**
SUSD3	**8.46**	**2.30**	**8.99**	**1.55**	**-0.53**	**7**	**8.67**	**1.57**	**8.87**	**1.57**	**-0.20**	**45**
TNS3	**9.84**	**0.32**	**10.04**	**0.50**	**-0.19**	**129**	**9.98**	**0.62**	**10.07**	**0.58**	**-0.09**	**216**
TPM2	**10.48**	**0.74**	**10.61**	**0.75**	**-0.13**	**275**	**10.32**	**0.78**	**10.40**	**0.82**	**-0.07**	**348**
TXNIP	**10.16**	**0.35**	**10.29**	**0.63**	**-0.14**	**259**	**9.92**	**0.72**	**10.00**	**0.76**	**-0.09**	**269**
UBD	**8.16**	**2.50**	**8.55**	**1.66**	**-0.39**	**40**	**7.98**	**1.56**	**8.12**	**1.61**	**-0.14**	**263**
VIM	**12.25**	**0.39**	**12.41**	**0.58**	**-0.16**	**57**	**12.05**	**0.77**	**12.13**	**0.83**	**-0.08**	**103**
VTCN1	**9.12**	**3.46**	**9.34**	**2.00**	**-0.22**	**134**	**9.08**	**2.01**	**9.26**	**1.96**	**-0.19**	**39**
ZBTB20	**8.79**	**0.51**	**8.96**	**0.63**	**-0.17**	**300**	**8.92**	**0.68**	**9.02**	**0.72**	**-0.10**	**342**

Supplementary Table 5. Full gene name list of the 99 genes significantly associated with lymphovascular invasion

Gene symbol	Gene name
ACTG2	actin gamma 2
ANG	angiogenin
ANXA1	annexin A1
APOC1	apolipoprotein C1
APOE	apolipoprotein E
C1S	complement C1s
CALML5	calmodulin-like 5
CCNB2	cyclin B2
CDC42EP4	CDC42 effector protein 4
CDCA5	cell division cycle associated 5
CEBPD	CCAAT/enhancer-binding protein delta
CFB	complement factor B
CFD	complement factor D
CLIC6	chloride intracellular channel 6
COX6C	cytochrome c oxidase subunit 6C
CXCL12	C-X-C motif chemokine ligand 12
CXCL14	C-X-C motif chemokine ligand 14
CYBRD1	cytochrome b reductase 1
CYP4X1	cytochrome P450 family 4 subfamily X member 1
DCN	decorin
DKK3	dickkopf WNT signaling pathway inhibitor 3
DNAJA4	DnaJ heat shock protein family (Hsp40) member A4

DPYSL2	**dihydropyrimidinase-like 2**
DUSP1	**dual specificity phosphatase 1**
EEF1A2	**eukaryotic translation elongation factor 1 alpha 2**
EEF1B2	**eukaryotic translation elongation factor 1 beta 2**
ELF3	**E74 like ETS transcription factor 3**
ERBB2	**erb-b2 receptor tyrosine kinase 2**
FBLN1	**fibulin-1**
FCER1A	**Fc fragment of IgE receptor Ia**
FCGBP	**Fc fragment of IgG binding protein**
FGD3	**FYVE, RhoGEF and PH domain containing 3**
FOS	**Fos proto-oncogene, AP-1 transcription factor subunit**
FST	**follistatin**
GAS1	**growth arrest specific 1**
GNAS	**GNAS complex locus**
GSTP1	**glutathione S-transferase pi 1**
HBA2	**hemoglobin subunit alpha 2**
HBB	**hemoglobin subunit beta**
HLA-DQA1	**major histocompatibility complex, class II, DQ alpha 1**
HMGA1	**high mobility group AT-hook 1**
HMGB3	**high mobility group box 3**
HSPB1	**heat shock protein family B (small) member 1**
IDH2	**isocitrate dehydrogenase (NADP(+)) 2, mitochondrial**
IFI27	**interferon alpha inducible protein 27**
IL17RB	**interleukin 17 receptor B**
ISG15	**ISG15 ubiquitin-like modifier**

KRT18	**keratin 18**
KRT18P55	**keratin 18 pseudogene 55**
KRT19	**keratin 19**
KRT7	**keratin 7**
KRT8	**keratin 8**
LAPTM4B	**lysosomal protein transmembrane 4 beta**
LRRC26	**leucine rich repeat containing 26**
LY6E	**lymphocyte antigen 6 family member E**
MAOA	**monoamine oxidase A**
MFAP4	**microfibrillar-associated protein 4**
MGP	**matrix Gla protein**
MMP11	**matrix metallopeptidase 11**
MT1E	**metallothionein 1E**
MX1	**MX dynamin like GTPase 1**
NDP	**NDP, norrin cystine knot growth factor**
NINJ1	**ninjurin 1**
NME1	**NME/NM23 nucleoside diphosphate kinase 1**
NOP56	**NOP56 ribonucleoprotein**
PDGFRL	**platelet derived growth factor receptor like**
PGAP3	**post-GPI attachment to proteins 3**
PITX1	**paired like homeodomain 1**
PLGRKT	**plasminogen receptor with a C-terminal lysine**
PTTG1	**pituitary tumor-transforming 1**
PYCARD	**PYD and CARD domain containing**
RPL3	**ribosomal protein L3**

S100A4	**S100 calcium binding protein A4**
S100P	**S100 calcium binding protein P**
SCD	**stearoyl-CoA desaturase**
SELENOM	**selenoprotein M**
SERPINA3	**serpin family A member 3**
SERPINE2	**serpin family E member 2**
SGCE	**sarcoglycan epsilon**
SLC40A1	**solute carrier family 40 member 1**
SLC44A1	**solute carrier family 44 member 1**
SLC52A2	**solute carrier family 52 member 2**
SLC9A3R1	**SLC9A3 regulator 1**
SPDEF	**SAM pointed domain containing ETS transcription factor**
SRPX	**sushi repeat containing protein, X-linked**
STC2	**stanniocalcin 2**
SUSD3	**sushi domain containing 3**
TM7SF2	**transmembrane 7 superfamily member 2**
TNS3	**tensin 3**
TPM2	**tropomyosin 2 (beta)**
TXNIP	**thioredoxin interacting protein**
UBD	**ubiquitin D**
UBE2C	**ubiquitin conjugating enzyme E2 C**
UBE2S	**ubiquitin conjugating enzyme E2 S**
UCP2	**uncoupling protein 2**
VIM	**vimentin**
VTCN1	**V-set domain containing T-cell activation inhibitor 1**

YWHAZ	**tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta**
ZBTB20	**zinc finger and BTB domain containing 20**

Supplementary Table 4. Survival analysis based on clinicopathological characteristics including lymphovascular invasion

METABRIC cohort								TCGA cohort							
Factors		Univariate analysis			Multivariate analysis			Factors		Univariate analysis			Multivariate analysis		
		Hazard Ratio	95% CI	*p*-value	Hazard Ratio	95% CI	*p*-value			Hazard Ratio	95% CI	*p*-value	Hazard Ratio	95% CI	*p*-value
LVI	Negative	Reference			Reference			LVI	Negative	Reference			Reference		
	Positive	1.70	1.45-2.01	<0.0001	1.29	1.07-1.56	0.0073		Positive	2.22	1.46-3.38	0.00019	2.19	1.32-3.62	0.0023
Tumour size	< 2cm	Reference			Reference			Tumour size	T1	Reference			Reference		
	≥ 2cm	1.82	1.49-2.21	<0.0001	1.48	1.21-1.83	0.00018		T2-4	1.81	1.08-3.04	0.025	1.33	0.77-2.31	0.30
Nodal status	Negative	Reference			Reference			Nodal status	Negative	Reference			Reference		
	Positive	2.06	1.74-2.44	<0.0001	1.63	1.35-1.97	<0.0001		Positive	1.85	1.20-2.85	0.0056	1.13	0.67-1.92	0.65
Histological grade	Grade 1, 2	Reference			Reference			Histological grade	Grade 1, 2	Reference			Reference		
	Grade 3	1.63	1.37-1.93	<0.0001	1.16	0.96-1.40	0.13		Grade 3	1.46	0.94-2.25	0.092	-		
ER	Positive	Reference			Reference			ER	Positive	Reference			Reference		
	Negative	1.66	1.38-1.99	<0.0001	1.14	0.91-1.43	0.25		Negative	1.89	1.19-2.98	0.0065	1.70	0.82-3.50	0.15
PR	Positive	Reference			Reference			PR	Positive	Reference			Reference		
	Negative	1.67	1.42-1.98	<0.0001	1.38	1.13-1.69	0.0020		Negative	1.68	1.08-2.61	0.020	1.21	0.60-2.42	0.60
HER2	Negative	Reference			Reference			HER2	Negative	Reference			Reference		
	Positive	1.92	1.54-2.38	<0.0001	1.45	1.15-1.83	0.0019		Positive	1.51	0.83-2.77	0.18	-		

Abbreviations: ER, Oestrogen receptor; PR, Progesterone receptor; LVI, Lymphovascular invasion.

Supplementary Table 3. Correlation between lymphovascular invasion and clinicopathological characteristics

METABRIC cohort						TCGA cohort					
Factors		LVI status			*p*-value	Factors		LVI status			*p*-value
		Positive	Negative	Total				Positive	Negative	Total	
Tumour size	≥ 2cm	485 (76.5%)	582 (63.7%)	1067	<0.0001	Tumour size	T 2-4	234 (79.3%)	381 (68.2%)	615	0.00055
	< 2cm	149 (23.5%)	332 (36.3%)	481			T 1	61 (20.7%)	178 (31.8%)	239	
Nodal status	Positive	430 (67.8%)	305 (32.9%)	735	<0.0001	Nodal status	Positive	226 (77.1%)	197 (35.4%)	423	<0.0001
	Negative	204 (32.2%)	621 (67.1%)	825			Negative	67 (22.9%)	359 (64.6%)	426	
Histological grade	Grade 3	374 (60.7%)	399 (44.6%)	773	<0.0001	Histological grade	Grade 3	155 (55.0%)	197 (36.9%)	352	<0.0001
	Grade 1, 2	242 (39.3%)	496 (55.4%)	738			Grade 1, 2	127 (45.0%)	337 (63.1%)	464	
ER	Positive	473 (74.5%)	731 (78.6%)	1204	0.058	ER	Positive	219 (76.6%)	420 (78.1%)	639	0.63
	Negative	162 (25.5%)	199 (21.4%)	361			Negative	67(23.4 %)	118 (21.9%)	185	
PR	Positive	319 (50.2%)	509 (54.7%)	828	0.080	PR	Positive	194 (68.6%)	352 (65.8%)	546	0.43
	Negative	316 (49.8%)	421 (45.3%)	737			Negative	89 (31.4%)	183 (34.2%)	272	
HER2	Positive	105 (16.5%)	83 (8.9%)	188	<0.0001	HER2	Positive	50 (21.2%)	83 (17.9%)	133	0.29
	Negative	530 (83.5%)	847 (91.1%)	1377			Negative	186 (78.8%)	381 (82.1%)	567	

Molecular subtypes	**Luminal A**	225 (35.5%)	368 (39.7%)	593	0.0021	
	Luminal B	178 (28.1%)	215 (23.2%)	393		
	HER2-enriched	87 (13.7%)	94 (10.1%)	181		
	Basal-like	99 (15.6%)	147 (15.8%)	246		
	Normal-like	44 (7.0%)	104 (11.2%)	148		
Abbreviations: ER, Oestrogen receptor; PR, Progesterone receptor; LVI, Lymphovascular invasion.						

Supplementary Table 2. List of top 350 genes significantly associated with lymphovascular invasion in the remaining METABRIC cases

Genes	WAD value	WAD ranking
ERBB2	0.191	1
SERPINA3	−0.170	2
S100P	0.145	3
TFF3	0.141	4
PIP	−0.123	5
MGP	−0.120	6
ACTG2	−0.120	7
GSTP1	−0.120	8
HBB	−0.119	9
UBE2C	0.116	10
MMP11	0.115	11
ACTG1	−0.112	12
FOS	−0.107	13
LY6E	0.107	14
CNTNAP2	0.106	15
HBA2	−0.105	16
YWHAZ	0.104	17
FLJ40504	0.104	18
CLIC6	−0.102	19
PGAP3	0.101	20

SFRP1	**−0.099**	**21**
HLA-A	**−0.097**	**22**
SPDEF	**0.096**	**23**
CYP4X1	**−0.094**	**24**
STARD10	**0.093**	**25**
SORD	**0.093**	**26**
LTF	**−0.092**	**27**
STC2	**−0.092**	**28**
C19orf33	**0.092**	**29**
SCD	**0.091**	**30**
ATP5E	**0.089**	**31**
abParts	**−0.089**	**32**
CFD	**−0.089**	**33**
SELM	**−0.087**	**34**
X64709	**0.087**	**35**
TOP2A	**0.087**	**36**
MX1	**0.086**	**37**
FCGBP	**−0.086**	**38**
VTCN1	**−0.086**	**39**
KRT17	**−0.086**	**40**
GSTM2	**−0.086**	**41**
CALML5	**0.085**	**42**
KRT8	**0.084**	**43**

NQO1	**0.084**	**44**
SUSD3	**−0.084**	**45**
TMBIM6	**0.083**	**46**
EEF1G	**−0.083**	**47**
DUSP1	**−0.082**	**48**
CXCL12	**−0.082**	**49**
ANXA1	**−0.082**	**50**
NFIX	**−0.081**	**51**
BOLA2B	**0.080**	**52**
HSPB1	**0.080**	**53**
FOXC1	**−0.080**	**54**
FAM83H	**0.079**	**55**
C10orf116	**0.078**	**56**
STAT1	**0.078**	**57**
NUSAP1	**0.078**	**58**
MYH11	**−0.078**	**59**
S100A16	**0.077**	**60**
PSMB3	**0.077**	**61**
GINS2	**0.076**	**62**
COL4A5	**−0.076**	**63**
APOC1	**0.076**	**64**
CXCL14	**−0.076**	**65**
KIAA0101	**0.076**	**66**

NME1	**0.075**	**67**
GRB7	**0.075**	**68**
ISG15	**0.075**	**69**
AGR2	**0.074**	**70**
HIST1H2AC	**−0.073**	**71**
CFB	**−0.073**	**72**
PTTG1	**0.073**	**73**
FAM129A	**−0.073**	**74**
DCN	**−0.072**	**75**
GPR172A	**0.072**	**76**
ATP9A	**0.072**	**77**
CLEC3A	**0.072**	**78**
KRT18	**0.071**	**79**
GAS1	**−0.071**	**80**
APOE	**0.071**	**81**
TPM1	**−0.071**	**82**
MFAP4	**−0.070**	**83**
UCP2	**0.070**	**84**
SPP1	**−0.070**	**85**
FBLN1	**−0.070**	**86**
CDC20	**0.069**	**87**
C8orf55	**0.069**	**88**
FST	**−0.068**	**89**

DKK3	**−0.068**	**90**
PAM	**−0.068**	**91**
NME4	**0.068**	**92**
ZAK	**−0.068**	**93**
UBE2S	**0.067**	**94**
MFGE8	**−0.066**	**95**
PUF60	**0.066**	**96**
MT1X	**−0.066**	**97**
EGR1	**−0.066**	**98**
TUBA1B	**0.065**	**99**
CDCA5	**0.065**	**100**
NAT1	**−0.064**	**101**
SRP9	**−0.064**	**102**
VIM	**−0.064**	**103**
PDLIM1	**−0.064**	**104**
EEF1A2	**0.064**	**105**
SQLE	**0.064**	**106**
DPYSL2	**−0.064**	**107**
COL16A1	**−0.064**	**108**
SLC9A3R1	**0.063**	**109**
NAPRT1	**0.063**	**110**
RRM1	**0.063**	**111**
HIST1H4C	**0.063**	**112**

SLC40A1	**−0.063**	**113**
PPAP2B	**−0.063**	**114**
EZR	**0.062**	**115**
CYC1	**0.062**	**116**
BST2	**0.062**	**117**
WWP1	**0.062**	**118**
STC1	**−0.062**	**119**
JUN	**−0.062**	**120**
SRPX	**−0.062**	**121**
RPS26	**0.062**	**122**
SERPINE2	**−0.061**	**123**
TMEM97	**0.061**	**124**
PRC1	**0.061**	**125**
TNC	**−0.061**	**126**
CMTM7	**−0.061**	**127**
CITED4	**−0.061**	**128**
SEZ6L2	**−0.061**	**129**
TSC22D1	**−0.061**	**130**
IDH2	**0.061**	**131**
HNRNPA1L2	**−0.060**	**132**
38777	**0.060**	**133**
S100A4	**−0.060**	**134**
VPS28	**0.060**	**135**

ZFP36	−0.060	136
CCDC130	−0.059	137
SLC44A1	−0.059	138
CD24	−0.058	139
EIF3E	0.058	140
PRNP	−0.058	141
PLAT	−0.058	142
MAL2	0.058	143
DDIT4	0.058	144
CGNL1	−0.058	145
PDGFRL	−0.058	146
ITM2A	−0.058	147
ARL6IP1	0.058	148
KRT19	0.057	149
CRIP1	0.057	150
NINJ1	−0.057	151
TSPAN13	0.057	152
CPNE3	0.057	153
ECHDC2	−0.057	154
CSE1L	0.057	155
GLA	−0.057	156
SLC7A2	−0.057	157
CUEDC1	0.056	158

EEF1B2	**−0.056**	**159**
PRDX1	**0.056**	**160**
BCAS4	**0.056**	**161**
TUBB2B	**−0.056**	**162**
CSTB	**−0.056**	**163**
ATP5EP2	**0.055**	**164**
PGM1	**−0.055**	**165**
GSDMB	**0.055**	**166**
FSCN1	**−0.055**	**167**
SGCE	**−0.055**	**168**
STK3	**0.055**	**169**
TM7SF2	**0.055**	**170**
EIF2C2	**0.055**	**171**
PABPC1	**0.055**	**172**
FGD3	**−0.054**	**173**
RBBP8	**−0.054**	**174**
NCOA3	**0.054**	**175**
PBX3	**−0.054**	**176**
ORMDL3	**0.054**	**177**
MAOA	**−0.054**	**178**
ANG	**−0.053**	**179**
SERHL2	**0.053**	**180**
FBLN2	**−0.053**	**181**

C9orf46	**−0.053**	**182**
MMP7	**−0.053**	**183**
TMEM106C	**0.053**	**184**
ALCAM	**0.053**	**185**
CCNB2	**0.053**	**186**
KRT7	**0.053**	**187**
SGK223	**−0.053**	**188**
MFSD3	**0.053**	**189**
ALOX5	**−0.053**	**190**
ALOX5AP	**−0.053**	**191**
CA2	**−0.053**	**192**
ATP6V0B	**0.052**	**193**
FGFR3	**0.052**	**194**
APOD	**−0.052**	**195**
TGOLN2	**−0.052**	**196**
ZDHHC8	**−0.052**	**197**
ZFP36L2	**−0.052**	**198**
MYC	**−0.052**	**199**
NCRNA00152	**0.052**	**200**
PCOLCE	**−0.052**	**201**
RPL19	**0.052**	**202**
MCM4	**0.052**	**203**
NTN4	**−0.052**	**204**

FOSB	**−0.052**	**205**
LASS6	**0.052**	**206**
EIF3G	**−0.052**	**207**
CKMT1B	**0.052**	**208**
COL6A1	**−0.051**	**209**
TMEM14C	**−0.051**	**210**
CST3	**−0.051**	**211**
RPL3	**−0.051**	**212**
SLC38A1	**0.051**	**213**
FRMD6	**−0.051**	**214**
SLC5A6	**0.051**	**215**
TNS3	**−0.051**	**216**
C1S	**−0.051**	**217**
PLSCR3	**−0.051**	**218**
CANT1	**0.050**	**219**
PTPN1	**0.050**	**220**
SC5DL	**−0.050**	**221**
ITM2B	**−0.050**	**222**
MYL6	**0.050**	**223**
IL17RB	**−0.050**	**224**
MT1E	**−0.050**	**225**
CSDA	**−0.050**	**226**
DNAJA4	**0.050**	**227**

TNFSF10	**−0.050**	**228**
NDP	**−0.049**	**229**
C12orf44	**0.049**	**230**
SERF2	**0.049**	**231**
CDC42EP4	**−0.049**	**232**
CYP4Z1	**−0.049**	**233**
LOC389493	**−0.049**	**234**
ADM	**−0.049**	**235**
TMEM101	**−0.049**	**236**
HERPUD1	**−0.049**	**237**
DENND1B	**0.049**	**238**
IFI44L	**0.049**	**239**
MRPL27	**0.049**	**240**
ALPL	**−0.049**	**241**
WLS	**−0.049**	**242**
CXCL10	**0.049**	**243**
ARMCX1	**−0.049**	**244**
KRT15	**−0.049**	**245**
LAPTM4B	**0.049**	**246**
CLDN3	**0.049**	**247**
ZBTB20	**−0.049**	**248**
COPS5	**0.048**	**249**
DNAJC12	**−0.048**	**250**

ID3	**−0.048**	**251**
UBE2E3	**−0.048**	**252**
PITX1	**0.048**	**253**
GAPDH	**0.048**	**254**
HLA-B	**−0.048**	**255**
SDCBP	**−0.048**	**256**
LRRC26	**0.048**	**257**
TNFRSF14	**−0.048**	**258**
CRTAP	**−0.048**	**259**
C8orf4	**−0.048**	**260**
NOSTRIN	**−0.048**	**261**
GNAS	**0.047**	**262**
UBD	**−0.047**	**263**
FAM127A	**−0.047**	**264**
CHI3L2	**−0.047**	**265**
GATA3	**−0.047**	**266**
AURKA	**0.047**	**267**
SCPEP1	**−0.047**	**268**
TXNIP	**−0.047**	**269**
ZNF148	**0.047**	**270**
QPCT	**−0.047**	**271**
CD248	**−0.047**	**272**
PRDX2	**−0.047**	**273**

BOLA2	**0.047**	**274**
GRINA	**0.047**	**275**
hNp95	**0.047**	**276**
COX6C	**0.047**	**277**
RPL30	**0.047**	**278**
IGJ	**−0.047**	**279**
TGFBR2	**−0.047**	**280**
STIP1	**0.047**	**281**
TDG	**0.047**	**282**
KRT6B	**−0.047**	**283**
CLN3	**0.047**	**284**
PTGDS	**−0.046**	**285**
NOP56	**0.046**	**286**
ELF3	**0.046**	**287**
ASAP1	**0.046**	**288**
C8orf84	**−0.046**	**289**
SLC1A5	**0.046**	**290**
MLPH	**0.046**	**291**
KIAA0182	**0.046**	**292**
FCER1A	**−0.046**	**293**
BZW2	**0.046**	**294**
MAPT	**−0.046**	**295**
GSN	**−0.046**	**296**

TMED3	**0.046**	**297**
HMGA1	**0.046**	**298**
ATP5H	**0.046**	**299**
CSNK1E	**−0.046**	**300**
CAMK2N1	**0.046**	**301**
ERGIC1	**0.046**	**302**
CR613620	**0.045**	**303**
ENPP5	**−0.045**	**304**
GGCT	**0.045**	**305**
C17orf97	**−0.045**	**306**
CAPS	**0.045**	**307**
KIAA1598	**0.045**	**308**
SERPINA1	**−0.045**	**309**
RPS19	**0.045**	**310**
SLC39A11	**0.045**	**311**
SAT1	**−0.045**	**312**
ACTB	**−0.045**	**313**
NUCB1	**−0.045**	**314**
SEMA6A	**−0.045**	**315**
CRISPLD2	**−0.045**	**316**
TMEM62	**0.045**	**317**
CYBRD1	**−0.045**	**318**
MT1G	**−0.045**	**319**

PTTG3	**0.045**	**320**
GADD45A	**−0.045**	**321**
RNASE1	**−0.045**	**322**
PYCARD	**−0.045**	**323**
LPIN1	**−0.045**	**324**
PPIC	**−0.045**	**325**
DQ893812	**−0.045**	**326**
HMGB3	**0.045**	**327**
ZHX1	**0.044**	**328**
NUDT1	**0.044**	**329**
POLR3GL	**−0.044**	**330**
TP53INP1	**0.044**	**331**
TUFM	**0.044**	**332**
CEBPD	**−0.044**	**333**
IFI27	**0.044**	**334**
SOX18	**−0.044**	**335**
RACGAP1	**0.044**	**336**
ST3GAL1	**0.044**	**337**
H2AFX	**0.044**	**338**
PTK2	**0.044**	**339**
SNX3	**−0.044**	**340**
CCDC92	**−0.044**	**341**
AK001020	**−0.044**	**342**

FAM110A	**0.044**	**343**
SCGB1D2	**−0.044**	**344**
IFIT1	**0.044**	**345**
FKBP9L	**−0.044**	**346**
HLA-DQA1	**−0.044**	**347**
TPM2	**−0.043**	**348**
CDS1	**0.043**	**349**
CLIP3	**−0.043**	**350**

Supplementary Table 1. List of top 350 genes significantly associated with lymphovascular invasion in the Nottingham cohort

Gene symbols	WAD value	WAD ranking
C10orf116	−0.339	1
MGP	−0.291	2
EEF1A2	0.256	3
CFB	−0.236	4
S100P	0.217	5
STC2	−0.217	6
SUSD3	−0.215	7
CDH1	0.212	8
MX1	0.209	9
CFD	−0.201	10
PITX1	0.199	11
COMP	−0.194	12
FABP4	−0.190	13
C1orf64	−0.187	14
MT1E	−0.174	15
SERPINE2	−0.173	16
FBLN1	−0.170	17
PLIN4	−0.169	18
FCGBP	−0.167	19
CIDEC	−0.165	20

FGD3	**−0.164**	**21**
FASN	**0.164**	**22**
ESR1	**−0.162**	**23**
IFI27	**0.161**	**24**
APOC1	**0.157**	**25**
SCUBE2	**−0.153**	**26**
SLC7A5	**0.151**	**27**
TFF1	**−0.150**	**28**
LGALS1	**−0.150**	**29**
ALDOA	**0.149**	**30**
APOE	**0.147**	**31**
HSPB1	**0.146**	**32**
GSTP1	**−0.145**	**33**
ADH1A	**−0.145**	**34**
LGALS3BP	**0.144**	**35**
KRT18	**0.143**	**36**
TOMM7	**−0.143**	**37**
IFI6	**0.142**	**38**
GPX3	**−0.142**	**39**
UBD	**−0.142**	**40**
FST	**−0.142**	**41**
SFRP4	**−0.141**	**42**
SHISA2	**−0.141**	**43**

RPS13	**−0.141**	**44**
CXCL12	**−0.140**	**45**
SLC44A1	**−0.139**	**46**
AGR3	**−0.138**	**47**
abParts	**0.137**	**48**
CXCL14	**−0.136**	**49**
CLIC6	**−0.135**	**50**
PIP	**0.135**	**51**
TGFBR3	**−0.135**	**52**
C1S	**−0.134**	**53**
DKK3	**−0.134**	**54**
SELM	**−0.133**	**55**
DPYSL3	**−0.132**	**56**
VIM	**−0.131**	**57**
DBNDD1	**0.131**	**58**
SLC40A1	**−0.131**	**59**
MT1A	**−0.130**	**60**
CALML5	**0.129**	**61**
ACTG1	**0.128**	**62**
SERPINA3	**−0.127**	**63**
CYP4X1	**−0.127**	**64**
ERBB2	**0.125**	**65**
MDK	**0.124**	**66**

PDGFRL	**−0.123**	**67**
RPL26	**−0.123**	**68**
TACSTD2	**0.122**	**69**
UBE2S	**0.122**	**70**
KIAA0531	**−0.119**	**71**
LY6E	**0.118**	**72**
HBB	**−0.118**	**73**
RPS9	**−0.118**	**74**
RPS3	**−0.118**	**75**
S100A6	**−0.118**	**76**
RPS6	**−0.117**	**77**
EIF3E	**−0.116**	**78**
YWHAZ	**0.116**	**79**
CA12	**−0.116**	**80**
HLA-DPA1	**−0.116**	**81**
PHGDH	**0.115**	**82**
HSP90AA1	**0.115**	**83**
MAOA	**−0.115**	**84**
LAPTM4B	**0.115**	**85**
ACTB	**0.115**	**86**
SEZ6L2	**0.115**	**87**
KRT19	**0.115**	**88**
TFAP2A	**0.114**	**89**

KRT8	**0.114**	**90**
PRRX1	**−0.114**	**91**
PPP1R1B	**−0.113**	**92**
RPS20	**−0.113**	**93**
BEX1	**−0.113**	**94**
ANXA1	**−0.112**	**95**
RUSC1	**0.112**	**96**
IFITM1	**−0.111**	**97**
CAP2	**−0.111**	**98**
PLAC9	**−0.111**	**99**
ACTG2	**−0.111**	**100**
GNAS	**0.111**	**101**
RARRES2	**−0.111**	**102**
HLA-DRA	**−0.110**	**103**
RPL10A	**−0.110**	**104**
RPS18	**−0.110**	**105**
UBE2C	**0.110**	**106**
GLTSCR2	**−0.109**	**107**
SLC9A3R1	**0.109**	**108**
GFRA1	**−0.109**	**109**
CTSD	**0.109**	**110**
AKR1C2	**−0.108**	**111**
RPL3	**−0.108**	**112**

HSP90AB1	**0.107**	**113**
HSPA1A	**0.107**	**114**
ARHGEF6	**−0.107**	**115**
AKR1C3	**−0.107**	**116**
RPL13AP6	**−0.106**	**117**
LPAR1	**−0.106**	**118**
SOCS2	**−0.106**	**119**
FLJ40504	**0.105**	**120**
SAA1	**−0.105**	**121**
PYCARD	**−0.105**	**122**
HLA-DMA	**−0.105**	**123**
COL8A1	**−0.105**	**124**
LRRC26	**0.105**	**125**
MYB	**−0.105**	**126**
IRX3	**0.105**	**127**
RPS17	**−0.104**	**128**
TNS3	**−0.104**	**129**
MFAP5	**−0.104**	**130**
RPL24	**−0.103**	**131**
TM7SF2	**0.103**	**132**
TMSB10	**0.102**	**133**
VTCN1	**−0.102**	**134**
HIST1H2BK	**0.102**	**135**

AX746718	**−0.102**	**136**
MGST1	**0.102**	**137**
ADAM15	**0.101**	**138**
RPS14	**−0.100**	**139**
ISG15	**0.100**	**140**
MFAP4	**−0.099**	**141**
APOD	**0.099**	**142**
AZGP1	**0.099**	**143**
FCER1A	**−0.099**	**144**
MT2A	**−0.099**	**145**
CPB1	**0.099**	**146**
ATP6V1B1	**0.098**	**147**
S100A8	**0.098**	**148**
RPL21	**−0.098**	**149**
RPS26P11	**−0.098**	**150**
DPT	**−0.097**	**151**
NDP	**−0.097**	**152**
IL6ST	**−0.097**	**153**
SMARCA1	**−0.097**	**154**
PDK3	**−0.096**	**155**
TPST2	**−0.096**	**156**
GAS6	**−0.096**	**157**
SLC38A1	**−0.096**	**158**

RPL35A	**−0.095**	**159**
C9orf46	**−0.095**	**160**
PPP1R3C	**−0.095**	**161**
CYBRD1	**−0.095**	**162**
CNN3	**−0.095**	**163**
ITPRIPL2	**−0.095**	**164**
TUBA1C	**0.094**	**165**
HMGB3	**0.094**	**166**
ATP6AP1	**0.094**	**167**
HIST1H4C	**−0.093**	**168**
CSTB	**0.093**	**169**
RAI14	**−0.093**	**170**
TCEAL4	**−0.093**	**171**
CLDN7	**0.093**	**172**
PLS3	**−0.092**	**173**
CR610863	**0.092**	**174**
BTG2	**−0.092**	**175**
SGCE	**−0.092**	**176**
WBP5	**−0.092**	**177**
ALDH2	**−0.091**	**178**
IDH2	**0.091**	**179**
ACOX2	**−0.091**	**180**
SERPINF1	**−0.091**	**181**

CIDEA	**−0.091**	**182**
RPL27A	**−0.091**	**183**
MMP11	**0.090**	**184**
EFEMP1	**−0.090**	**185**
ANG	**−0.090**	**186**
CCL15	**−0.090**	**187**
HLA-DQA1	**−0.090**	**188**
UCP2	**0.090**	**189**
RPL36	**−0.089**	**190**
ECM2	**−0.089**	**191**
S100A9	**0.089**	**192**
BTG1	**−0.088**	**193**
C13orf15	**−0.088**	**194**
CITED2	**−0.088**	**195**
HOXB2	**0.088**	**196**
CDC42EP4	**−0.088**	**197**
CAV1	**−0.088**	**198**
PGAP3	**0.088**	**199**
SCD	**0.087**	**200**
FAU	**−0.087**	**201**
LRRC17	**−0.087**	**202**
PROM2	**0.087**	**203**
CCL5	**−0.087**	**204**

DNAJA4	**0.087**	**205**
IFITM2	**−0.087**	**206**
ARHGEF3	**−0.087**	**207**
HCST	**−0.087**	**208**
S100A4	**−0.087**	**209**
HIST1H4H	**0.086**	**210**
ALDH3A2	**−0.086**	**211**
RFTN1	**−0.086**	**212**
YWHAQ	**0.086**	**213**
DPYSL2	**−0.086**	**214**
RPL22	**−0.086**	**215**
PFKP	**0.085**	**216**
NME1	**0.085**	**217**
COMMD6	**−0.085**	**218**
EEF1B2	**−0.085**	**219**
NFKBIZ	**0.085**	**220**
VCAM1	**−0.084**	**221**
CALM1	**0.084**	**222**
KRT7	**0.084**	**223**
SLC25A5	**0.083**	**224**
MGC87042	**−0.083**	**225**
BCAP31	**0.083**	**226**
GAS1	**−0.083**	**227**

MMP9	**0.083**	**228**
FTL	**0.083**	**229**
MDH2	**0.082**	**230**
C8orf40	**−0.082**	**231**
CDCA5	**0.082**	**232**
GLYATL2	**0.082**	**233**
TPSAB1	**−0.082**	**234**
RNASE1	**0.082**	**235**
HLA-DRB6	**−0.081**	**236**
HLA-DQB1	**−0.081**	**237**
CRYAB	**−0.081**	**238**
CPA3	**−0.081**	**239**
C10orf10	**−0.081**	**240**
TUBB	**0.081**	**241**
NOP56	**0.081**	**242**
FERMT2	**−0.081**	**243**
PRKCDBP	**−0.080**	**244**
CD24	**0.080**	**245**
GRN	**0.080**	**246**
MXRA5	**−0.080**	**247**
LASP1	**0.080**	**248**
WISP2	**−0.080**	**249**
POLD2	**0.080**	**250**

POTEKP	**0.080**	**251**
ARL6IP5	**−0.080**	**252**
GBP2	**−0.080**	**253**
TSPYL5	**0.080**	**254**
FLNB	**−0.079**	**255**
H2AFY2	**0.079**	**256**
PTTG1	**0.079**	**257**
COX5A	**0.079**	**258**
TXNIP	**−0.078**	**259**
EIF3L	**−0.078**	**260**
ATHL1	**−0.078**	**261**
CHCHD2	**0.078**	**262**
COX6C	**0.078**	**263**
BCL2	**−0.078**	**264**
XBP1	**−0.078**	**265**
EPN1	**0.078**	**266**
FOXA1	**0.078**	**267**
CCND2	**−0.078**	**268**
LSM1	**−0.078**	**269**
SNAR-A3	**0.078**	**270**
ZNF217	**0.077**	**271**
RPSA	**−0.077**	**272**
CD36	**−0.077**	**273**

ELF3	**0.077**	**274**
TPM2	**−0.077**	**275**
SAPS2	**−0.077**	**276**
NFIB	**−0.077**	**277**
MBOAT7	**0.077**	**278**
ATP5B	**0.076**	**279**
C7orf41	**−0.076**	**280**
ABCB9	**0.076**	**281**
CDR2L	**0.076**	**282**
RPS28	**−0.076**	**283**
LMTK3	**0.076**	**284**
P4HB	**0.075**	**285**
ATP5C1	**0.075**	**286**
F13A1	**−0.075**	**287**
ULK1	**0.075**	**288**
KLHDC9	**−0.075**	**289**
ZG16B	**0.074**	**290**
TMED9	**0.074**	**291**
ZMIZ1	**0.074**	**292**
ATP2A2	**0.074**	**293**
RPL27	**−0.074**	**294**
GPI	**0.074**	**295**
WNK4	**−0.074**	**296**

RPL35	**−0.074**	**297**
RSL24D1	**−0.074**	**298**
CYB561	**0.074**	**299**
AK001020	**−0.073**	**300**
LUM	**−0.073**	**301**
ACP5	**0.073**	**302**
HMGA1	**0.073**	**303**
FBP1	**−0.073**	**304**
FTH1	**0.073**	**305**
MELK	**0.073**	**306**
FMOD	**−0.072**	**307**
HBA2	**−0.072**	**308**
GPR172A	**0.072**	**309**
TIGA1	**−0.072**	**310**
GSTM2	**0.072**	**311**
TSPAN9	**−0.072**	**312**
POLB	**−0.072**	**313**
NINJ1	**−0.072**	**314**
RPL5	**−0.071**	**315**
CEBPD	**−0.071**	**316**
ASNS	**0.071**	**317**
RBP1	**−0.071**	**318**
UBA1	**0.071**	**319**

AKT1	**0.071**	**320**
DARC	**−0.071**	**321**
RERG	**−0.071**	**322**
PALLD	**−0.071**	**323**
OMD	**−0.071**	**324**
DCN	**−0.071**	**325**
CCNB2	**0.071**	**326**
COL9A2	**−0.071**	**327**
SRPX	**−0.071**	**328**
CTDSPL	**−0.071**	**329**
ARHGEF2	**−0.070**	**330**
RPS27	**−0.070**	**331**
THBS2	**−0.070**	**332**
HSPD1	**0.070**	**333**
ARHGDIA	**0.070**	**334**
ANKRD30A	**−0.070**	**335**
PTRF	**−0.070**	**336**
FOXO3	**−0.070**	**337**
GIPC1	**0.070**	**338**
CHPT1	**−0.070**	**339**
IL17RB	**−0.070**	**340**
SYBU	**−0.070**	**341**
TUFT1	**0.070**	**342**

CCDC25	**−0.070**	**343**
ARID5B	**−0.070**	**344**
IGFBP4	**−0.070**	**345**
SPDEF	**0.070**	**346**
VDAC2	**0.070**	**347**
DUSP1	**−0.070**	**348**
FOS	**−0.069**	**349**
COPG	**0.069**	**350**

Fig. 1

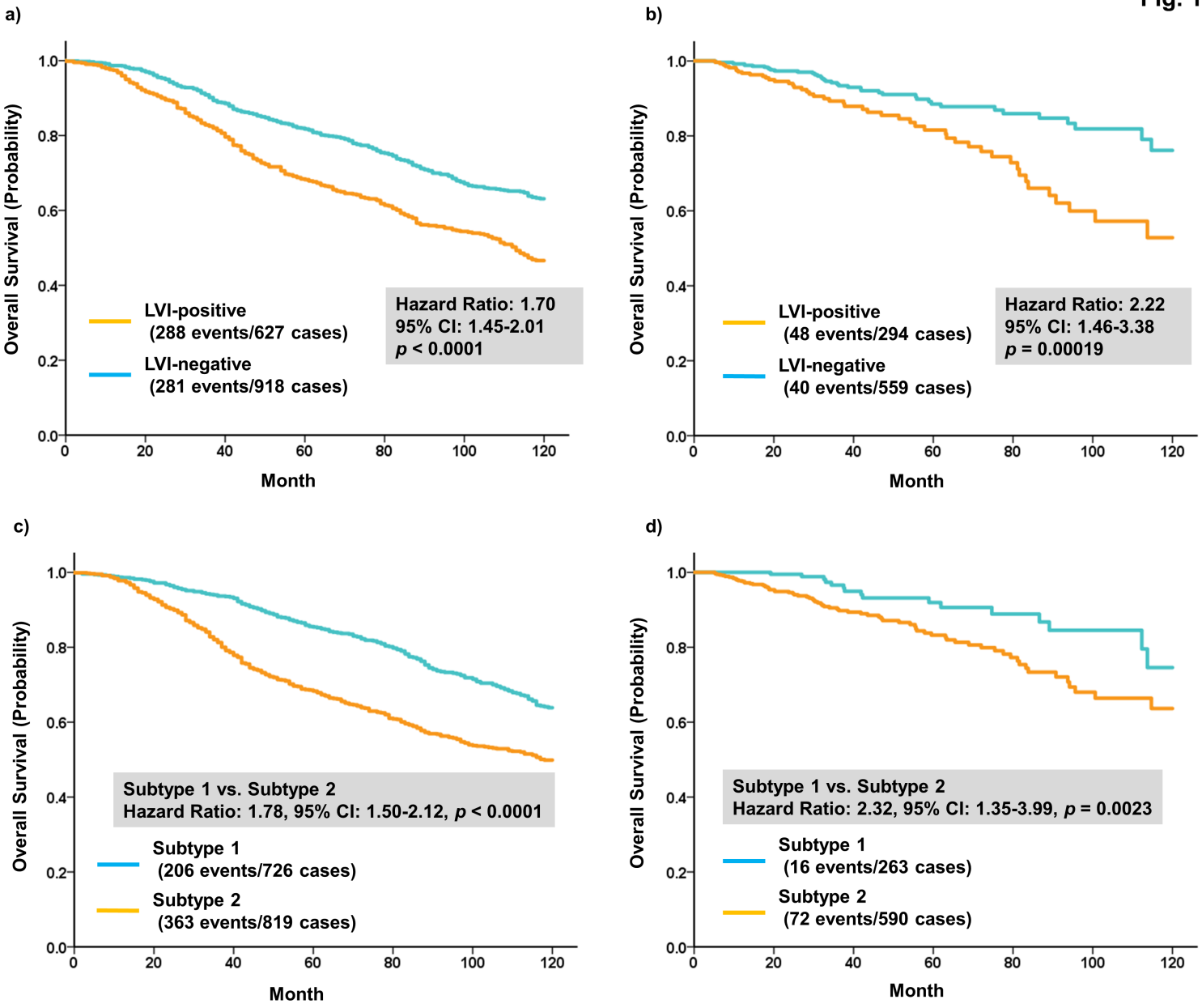
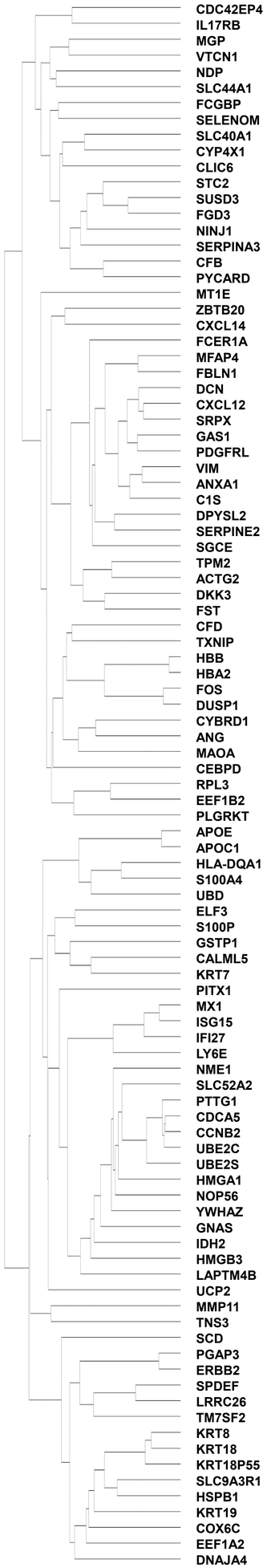
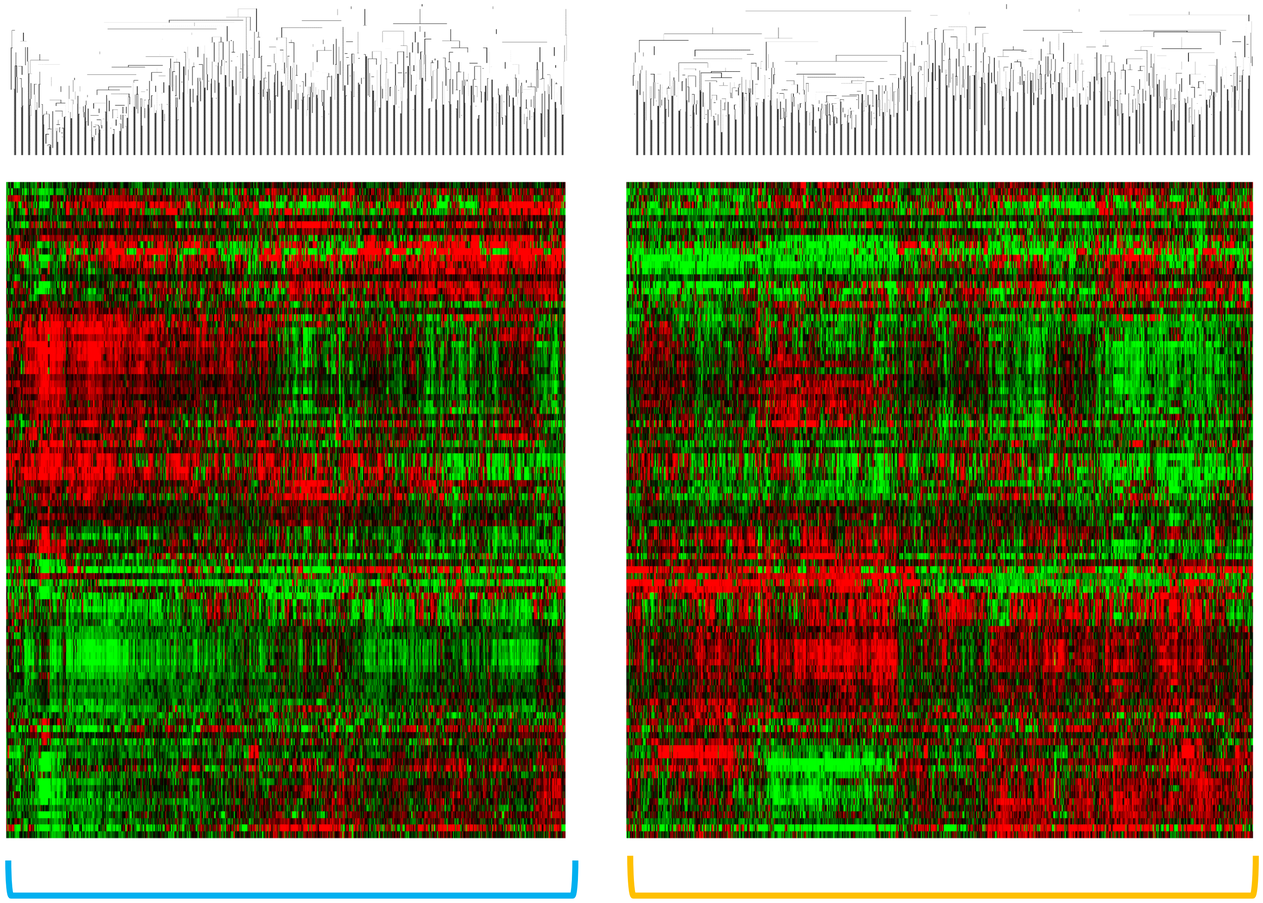

a)
Overall Survival (Probability)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Month
LVI-positive
(288 events/627 cases)
LVI-negative
(281 events/918 cases)
Hazard Ratio: 1.70
95% CI: 1.45-2.01
p < 0.0001
b)
Overall Survival (Probability)
Month
LVI-positive
(48 events/294 cases)
LVI-negative
(40 events/559 cases)
Hazard Ratio: 2.22
95% CI: 1.46-3.38
p = 0.00019
c)
Overall Survival (Probability)
Month
Subtype 1 vs. Subtype 2
Hazard Ratio: 1.78, 95% CI: 1.50-2.12, p < 0.0001
Subtype 1
(206 events/726 cases)
Subtype 2
(363 events/819 cases)
d)
Overall Survival (Probability)
Month
Subtype 1 vs. Subtype 2
Hazard Ratio: 2.32, 95% CI: 1.35-3.99, p = 0.0023
Subtype 1
(16 events/263 cases)
Subtype 2
(72 events/590 cases)

Fig. 2

a)


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

b)


Subtype 1
Subtype 2

c)

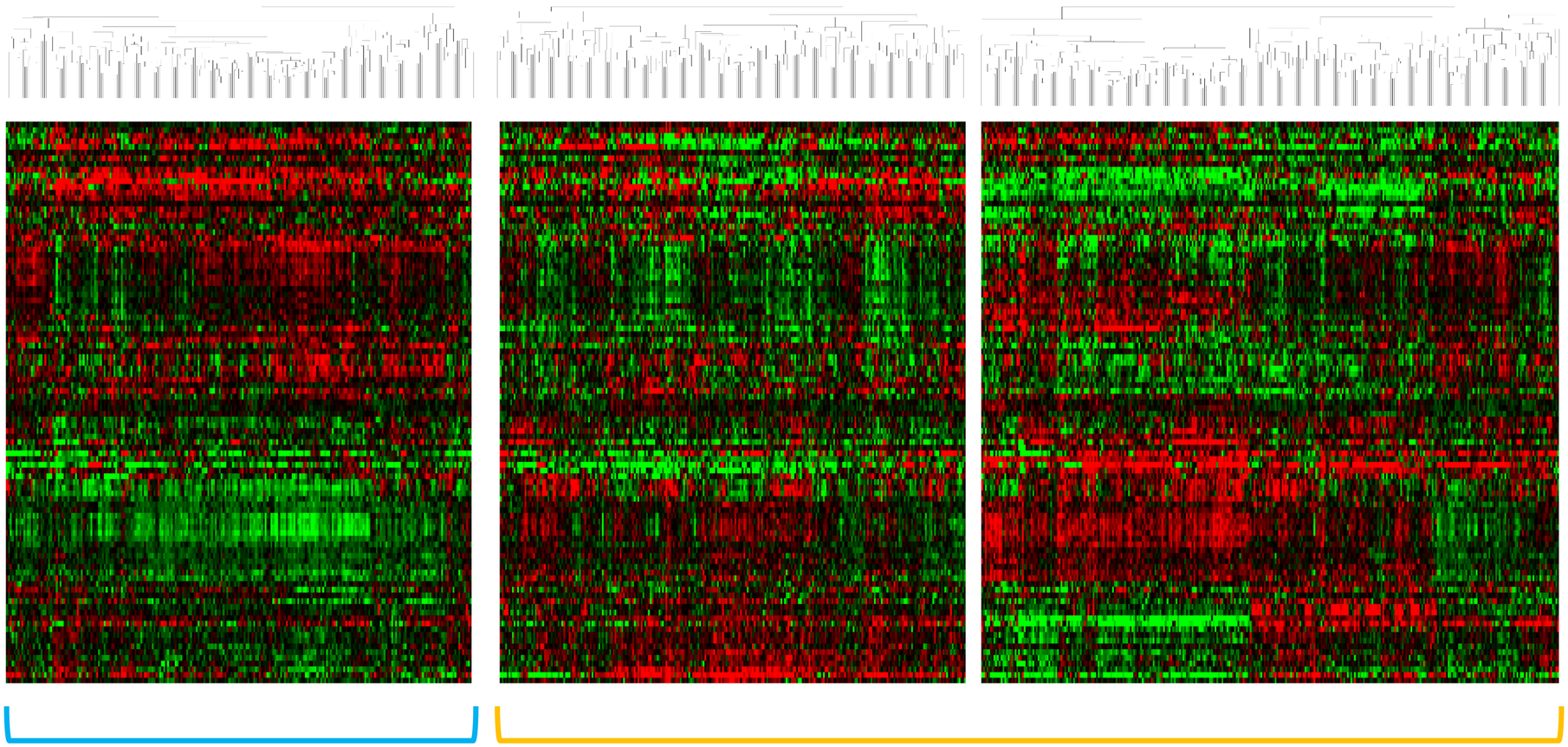
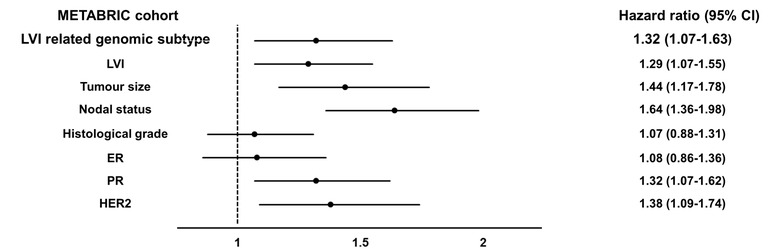
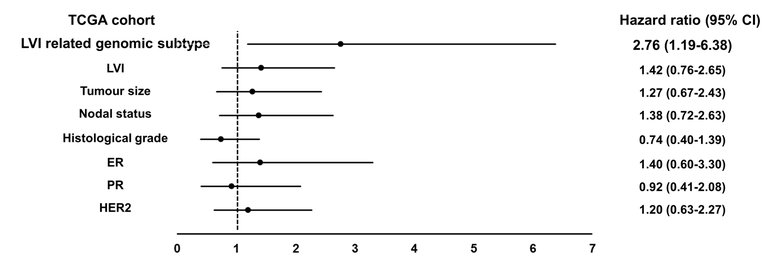

Subtype 1
Subtype 2

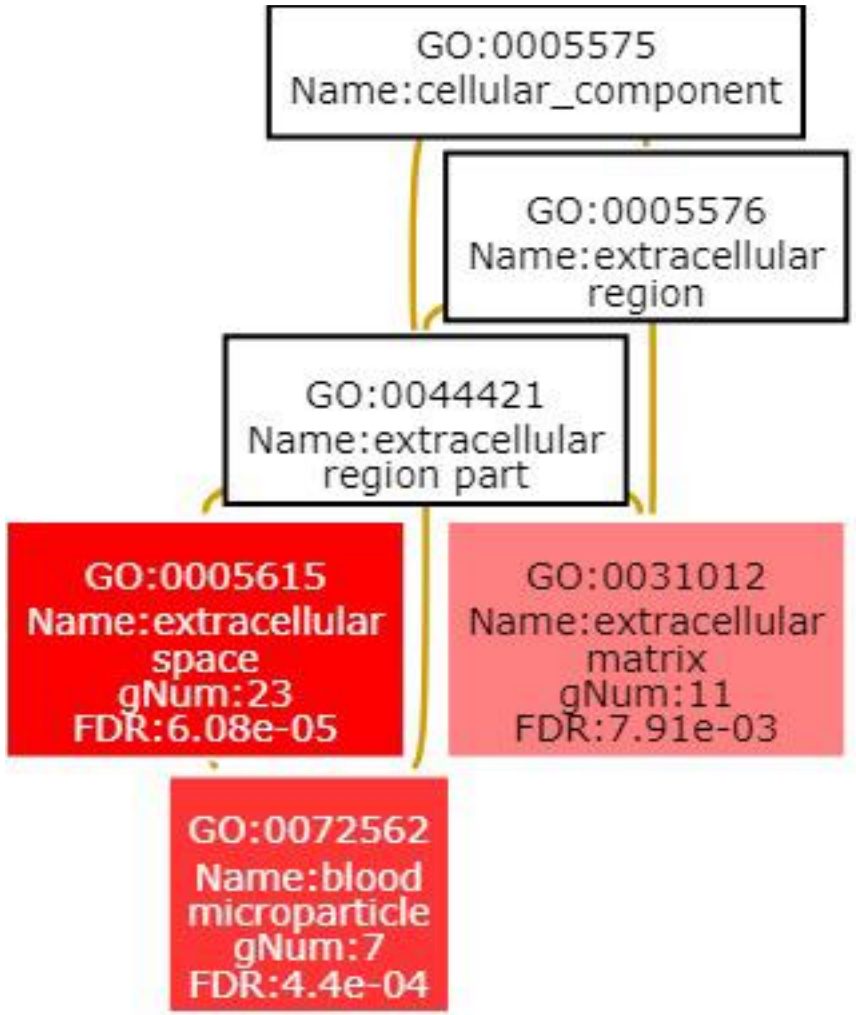
Fig. 3

a)

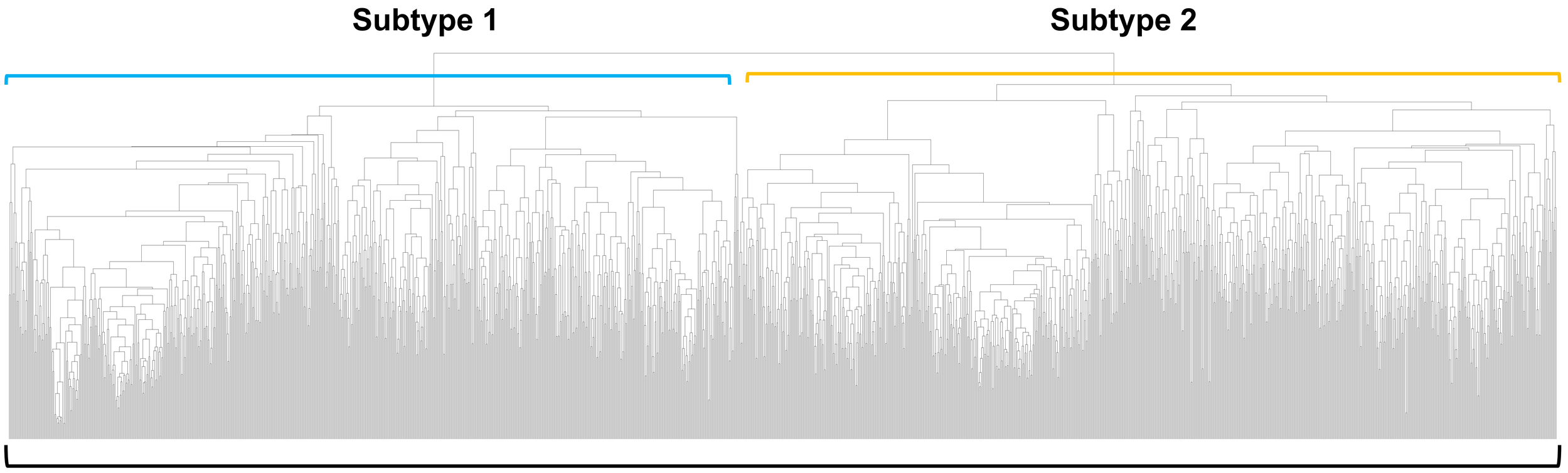

METABRIC cohort
Hazard ratio (95% CI)
LVI related genomic subtype
1.32 (1.07-1.63)
LVI
1.29 (1.07-1.55)
Tumour size
1.44 (1.17-1.78)
Nodal status
1.64 (1.36-1.98)
Histological grade
1.07 (0.88-1.31)
ER
1.08 (0.86-1.36)
PR
1.32 (1.07-1.62)
HER2
1.38 (1.09-1.74)
1
1.5
2

b)

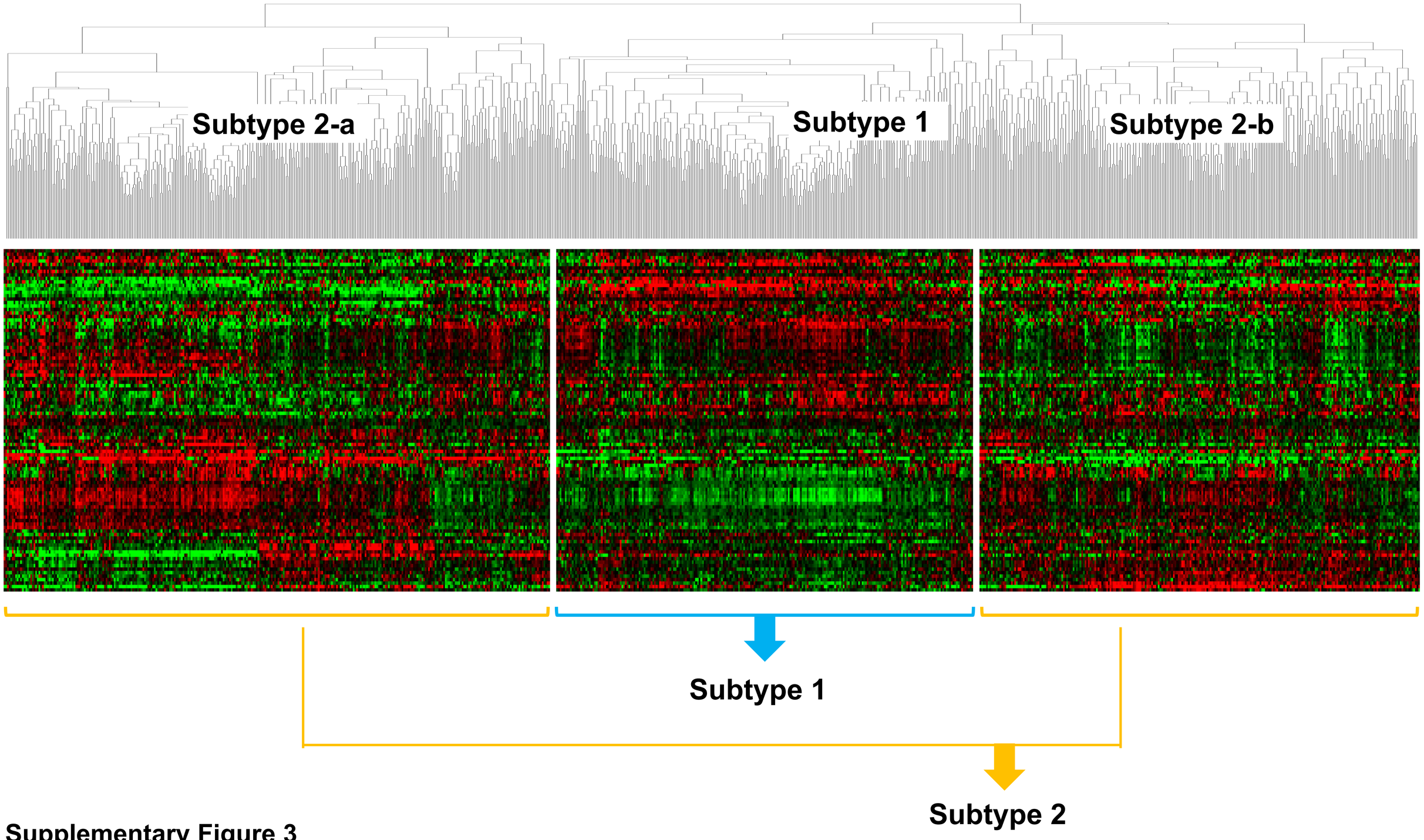

TCGA cohort
Hazard ratio (95% CI)
LVI related genomic subtype
2.76 (1.19-6.38)
LVI
1.42 (0.76-2.65)
Tumour size
1.27 (0.67-2.43)
Nodal status
1.38 (0.72-2.63)
Histological grade
0.74 (0.40-1.39)
ER
1.40 (0.60-3.30)
PR
0.92 (0.41-2.08)
HER2
1.20 (0.63-2.27)
0
1
2
3
4
5
6
7



Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3