Breast Cancer Research

RESEARCH **Open Access**

Clustering of HR + /HER2− breast cancer in an Asian cohort is driven by immune phenotypes

Jia-Wern Pan[1*], Mohana Ragu[1], Wei-Qin Chan[1], Siti Norhidayu Hasan[1], Tania Islam[2], Li-Ying Teoh[2], Suniza Jamaris[2], Mee-Hoong See[2], Cheng-Har Yip[3], Pathmanathan Rajadurai[3,4], Lai-Meng Looi[5], Nur Aishah Mohd Taib[2], Oscar M. Rueda[6], Carlos Caldas[7,8], Suet-Feung Chin[7], Joanna Lim[1] and Soo-Hwang Teo[1,9]

Abstract

Breast cancer exhibits significant heterogeneity, manifesting in various subtypes that are critical in guiding treatment decisions. This study aimed to investigate the existence of distinct subtypes of breast cancer within the Asian population, by analysing the transcriptomic profiles of 934 breast cancer patients from a Malaysian cohort. Our findings reveal that the HR + /HER2− breast cancer samples display a distinct clustering pattern based on immune phenotypes, rather than conforming to the conventional luminal A-luminal B paradigm previously reported in breast cancers from women of European descent. This suggests that the activation of the immune system may play a more important role in Asian HR + /HER2− breast cancer than has been previously recognized. Analysis of somatic mutations by whole exome sequencing showed that counter-intuitively, the cluster of HR + /HER2− samples exhibiting higher immune scores was associated with lower tumour mutational burden, lower homologous recombination deficiency scores, and fewer copy number aberrations, implicating the involvement of non-canonical tumour immune pathways. Further investigations are warranted to determine the underlying mechanisms of these pathways, with the potential to develop innovative immunotherapeutic approaches tailored to this specific patient population.

*Correspondence:
Jia-Wern Pan
jiawern.pan@cancerresearch.my
Full list of author information is available at the end of the article

Introduction

Breast cancer is a highly heterogeneous disease, characterized by diverse subtypes that have important implications in guiding treatment [1–4]. These distinct subtypes encompass variations in clinical presentation, molecular profiles, and response to treatment, highlighting the complex nature of breast cancer. The classification of breast cancer subtypes has provided valuable insights into disease prognosis and treatment selection, enabling more personalized and targeted therapeutic approaches.

Currently, there are two generally accepted methodologies to classify breast cancer subtypes: PAM50 [1, 2, 5] and Integrative Clusters [3, 4, 6]. Under the PAM50 classification scheme, breast cancer is classified into four main subtypes according to gene expression-based clusters: luminal A, luminal B, HER2− enriched, and basal-like, as well as an additional small group usually labelled as “normal-like”. These categories roughly correspond to the biomarker- and treatment-based clinical subtypes of HR+/HER2−, HR+/HER2+, HR−/HER2+, and triple negative breast cancer (TNBC). The Integrative Cluster classification scheme, on the other hand, integrates both transcriptomic and copy number data to classify breast cancer into 11 subtypes with different driver mutations. These subtypes include the HER2+-associated IntClust 5, the TNBC-associated IntClust 10, as well as subtypes with low copy number variation (IntClust 4+and 4−).

However, it is important to recognize that the existing subtypes were predominantly derived from studies conducted in populations of European descent. Given the clinical utility of molecular subtyping, there is a growing need to investigate breast cancer subtypes in different ethnic populations to account for potential variations in genetic, environmental, and lifestyle factors. For example, the various breast cancer subtypes are known to have different associations with reproductive risk factors [7, 8] and BMI [7], and these lifestyle risk factors are likely to differ substantially between different populations. Additionally, different breast cancer subtypes are associated with different breast cancer risk genetic loci [9–11], which are also distributed differently in different populations [12]. Importantly, previous studies have found differences in the prevalence of certain subtypes in non-European populations, with the African population having higher numbers of TNBC-associated samples [13], and the Asian population having a higher prevalence of HER2− enriched [14] or luminal B [15] samples.

The heterogeneity in clinicopathological features across different populations suggests the possibility that intrinsic molecular subtypes differ across different populations as well. For example, breast cancers from the Nigerian population have a high prevalence of homologous recombination deficiency (HRD), *TP53* mutations, and structural variation indicative of a more aggressive biology compared to tumours from Western populations [13]. It has also been identified previously that Asian breast cancers exhibit higher immune scores compared to the Western population [14, 15]. These findings suggest potential differences in the underlying biology of breast cancer in different ethnic groups and highlight the importance of population-specific studies across diverse populations.

Investigating breast cancer subtypes in Asians holds significant promise for improving our understanding of the disease and optimizing treatment strategies for this specific population. Additionally, it may provide insights into the underlying genetic and molecular mechanisms that contribute to breast cancer pathogenesis, allowing for the development of more tailored and effective therapeutic interventions. Existing studies which investigated the molecular subtypes underlying breast cancer in Asian populations have reported that molecular subtypes are largely conserved between Asian and Western populations [15–17]. However, the cohort sizes in these studies are relatively small, and some are only focused on a subpopulation, such as young breast cancer, and may not be representative of the entire Asian population. In this study, we explored the possibility of unique subtypes of breast cancer within the Asian population by examining the transcriptomic profiles of breast cancer patients from a relatively large cohort. Our analyses found the HR+/HER2− breast cancer samples in our cohort display a distinct clustering pattern based on immune phenotypes instead of the luminal A-luminal B paradigm. This suggests that the activation of the immune system may play a more important role in Asian HR+/HER2− breast cancer than has been previously recognized, which may have important clinical implications.

Materials and methods

Study cohort

Our study cohort consists of 934 Malaysian women with breast cancer. Patients were recruited to the MyBrCa Genetics study [18] from the Subang Jaya Medical Centre and the University Malaya Medical Centre between 2012 and 2018. Peripheral blood samples and breast tumour tissues were acquired from each patient. For breast tumour tissues, representative fresh tumour tissues were obtained and frozen during surgical resection of the tumour. These tumour samples were then stored in liquid nitrogen. Tumour samples were then sectioned for DNA and RNA extraction. The top and bottom sections were stained with haematoxylin and eosin and reviewed for tumour content. Tumour samples with an average tumour content of $<30\%$ ($n=50$) and/or insufficient DNA ($n=8$) were excluded from the study. Patients with bilateral breast cancer were also excluded ($n=14$). Patient enrolment and genetic analyses were approved by the Ethics Committee of Subang Jaya Medical Centre (Reference no: 201208.1) and the Medical Ethics Committee of the University Malaya Medical Centre (Reference no: 842.9) and written informed consent was provided by each patient.

Nucleic acid extraction and sequencing of tumour and matched normal specimens

For samples which were part of our original MyBrCa tumour cohort publication, DNA and RNA extraction and sequencing were performed as previously published [14]. This study also included an additional 448 samples. For these samples, DNA extraction from blood samples was performed using the Maxwell 16 Blood DNA Purification Kit with a Maxwell 16 Instrument, following standard protocol. For tumour samples, DNA was extracted using the QIAGEN DNeasy Blood and Tissue Kit following standard protocol and quantified using the Qubit HS DNA Assay kit and Qubit 2.0 fluorometer (Life Technologies Inc).

RNA was extracted from tumour samples using the QIAGEN miRNeasy Mini Kit with a QIAcube, according to standard protocol. Quantification of total RNA was performed using a Nanodrop 2000 Spectrophotometer and RNA integrity was established through an Agilent 2100 Bioanalyzer. For both DNA and RNA sequencing (RNA-seq), samples with a concentration above 20.0 ng/µL were selected. Additionally, an RNA integrity number above 7 was required for samples to be selected for RNA-seq.

DNA libraries were produced from 50 ng of genomic DNA using the Nextera Rapid Capture Exome kit (Illumina, San Diego, USA) as per manufacturer's instructions. Exome capture was achieved in pools of 3 and subjected to paired end 75 sequencing on a NovaSeq platform (Illumina, San Diego, USA) at 40 × depth for blood samples and 80 × depth for tumour samples. Prior to exome capture, 4 nM pools of DNA libraries from tumour samples were also selected for single end 50 shallow whole-genome sequencing at 0.1 × depth.

RNA libraries were prepared from 550 ng of total RNA from tumour samples using the TruSeq Stranded Total RNA HT kit with Ribo-Zero Gold (Illumina, San Diego, USA) as per manufacturer's instructions, subjected to paired end 75 sequencing on a NovaSeq platform (Illumina, San Diego, USA) at 40 × depth.

Gene expression analysis

RNA-seq reads were aligned to the hs37d5 human genome and the ENSEMBLE GrCh37 release version 87 human transcriptome via the STAR aligner (v.2.5.3a) [19]. Gene-level counts were calculated with featureCounts (v. 1.5.3) [20]. Gene-level count matrices for the cohort were transformed into normalized log2 counts-per-million (logCPM) using the voom function from the limma (v. 3.34.9) R package. The transformed matrices were then subtyped according to PAM50 and SCMgene designations using the Genefu package in R (v. 2.14.0).

Clustering and classification analysis

To identify unique subtypes, unsupervised k-means clustering was performed on the gene-level count matrices for the MyBrCa cohort. We also evaluated the use of different numbers of genes as our feature set, by ranking each gene according to the median absolute deviation of each gene within the cohort, and using either the top 1000 genes with the highest median absolute deviation, the top 5000, or all genes. To ensure robustness, an extensively implemented consensus clustering method [21], with 1000 iterations and 0.9 subsampling ratio, was used to assess clustering stability. Consensus clustering was implemented by the ConsensusClusterPlus function of the R package ConsensusClusterPlus with k-means clustering algorithm using Pearson correlation distance. Our final clustering model used $k=12$ with the top 5000 genes.

Hierarchical clustering analysis

Hierarchical clustering was conducted on the gene-level count matrices for the MyBrCa cohort using the hclust package from R with default parameters.

Shallow whole genome sequencing alignment and copy number aberration (CNA) assessment

Sequenced reads were mapped to the hg19 reference genome using bwa-mem, sorted using samtools and dedupped using picard (http://broadinstitute.github.io/picard). Mapped reads were analysed using QDNAseq [22] to obtain 100 kb segmented copy number profiles using standard protocol and default parameters. CNAs were called using CGHcall (v 2.40) as implemented in the QDNASeq R package, which calls each segment as normal, copy number gain, copy number loss, amplification or deletion using a mixture model. ENSEMBL hg19 genes with HUGO names were mapped to the segmented copy number calls by their start positions to determine the copy number status for each gene in each sample.

Profiling the tumour immune microenvironment

Overall immune cell infiltration in the bulk tumour samples was assessed from RNA-seq TPM gene expression scores using ESTIMATE (v. 1.0.13) [23], gene set variation analysis (GSVA) (v. 1.26) using combined immune cell gene sets from Bindea et al. [24] and immune scores from Thorsson et al. [25]. For each sample, immune features predictive of checkpoint inhibitor immunotherapy was also scored. This was done using IMPRES scores (only 14 out of 15 of the predictive features were available in our datasets) as well as GSVA using the Expanded IFN-gamma gene set [26, 27]. The relative abundance of specific immune cell populations was estimated from RNA-seq TPM scores with the CIBERSORT [28] web tool, as well as through GSVA with individual immune gene sets from Bindea et al. [24].

Mutational signatures

The weights of previously reported breast cancer mutational signatures using COSMIC matrices (Single Base Substitutions (SBS) Signatures 1, 2, 3 and 13) were established using deconstructSigs [29]. The proportion of variants associated with each mutational signature was determined only for samples with at least 15 detected single nucleotide variants (SNVs).

HRD scores

The following measures of HRD were determined as described previously: (1) loss of heterozygosity (LOH), (2) large-scale state transitions (LST), and (3) telomeric allelic imbalance (TAI) [30, 31]. Allele-specific copy number (ASCN) profiles on paired normal-tumour BAM files were classified via Sequenza [32] and utilised to analyse the individual measure scores and HRD-sum scores via scarHRD R package [33].

Differential gene expression and functional enrichment analysis

Gene expression was analysed with the DEseq2 package, an R-based open-source software designed to analyse transcriptomic data for differential expression, as previously described [34]. Gene set enrichment analyses (GSEA) was performed to compare clusters using the Hallmark pathways from the Molecular Signatures Database [35] as well as KEGG pathways. These analyses were performed with default parameter settings using 1,000 permutations and an FDR cutoff of 0.05. Single-sample gene set enrichment analyses (ssGSEA) and gene set variation analyses (GSVA) were also performed for each individual sample for specific Hallmark and KEGG pathway gene sets, including the Hallmark complement, hypoxia, IL6-JAK-STAT3 signalling, inflammatory response, and interferon gamma response pathways, as well as the KEGG cGAS-STING, T-cell receptor signalling, antigen processing and presentation, and TGF-β signalling pathways, using the GSVA package in R.

Survival analysis

Survival data of patients were obtained from the Malaysian National Registry record of deaths. Survival length was defined as the period between the date of diagnosis of patients until the date of death for deceased patients, or the date when the Malaysian National Registry was last queried for patients assumed to be still alive. Cox proportional hazard models were built using the coxph function from the survival package and plotted using ggadjustedcurves and ggforest functions from the survminer package in R (v. 4.3.1). For comparison of overall survival between MyBrCa clusters, the cluster with relatively large sample sizes and the best survival (Clusters 2) was selected as the reference group, and the p-value from comparison with Cluster 1, which had the worst survival among the larger clusters, was reported. For comparison between HR + /HER2− clusters, Cluster 7 (Group 1) was used as the reference group as comparison with the Group 2 clusters and the p value of the cluster with best survival was reported.

Statistical analysis

The Wilcoxon test and the Chi-square test were executed for comparisons of variables between categories. Unpaired t-tests were used to compare continuous variables between two groups. All tests were two-tailed and a significance level of $p = 0.05$ was used. Statistical analyses were performed using R v4.0. All box and whiskers plots in the main and supplementary figures were constructed with boxes indicating the 25th percentile, the median and the 75th percentile, whiskers showing the maximum and minimum values within 1.5 times the inter-quartile range, and outliers were not shown.

Results

Clinical characteristics of the MyBrCa cohort

The MyBrCa tumour cohort comprises of 934 female breast cancer patients of self-reported Malaysian nationality who were sequentially recruited from two Malaysian hospitals, Subang Jaya Medical Centre and Universiti Malaya Medical Centre. This cohort consists of a mix of Chinese, Malay, or Indian ancestry. The clinical characteristics of these patients are shown in Table 1.

Identification of Clusters Associated with Subtypes of MyBrCa patients

After sample processing, sequencing, and data processing, we were able to obtain transcriptomic profiles for 934 samples, which we used to conduct a cluster analysis. Using k-means clustering, we iteratively analysed the clustering of these samples across a range of *k* values

Table 1 Clinico-pathological characteristics of the study cohort. Group 1 and Group 2 are the two groups of HR+/HER2− patients described below. Statistical significance was determined using Student's t-test or Pearson's chi-squared test

	Overall	**Group 1**	**Group 2**	**Statistical Significance**
Subjects (n)	934	61	421	
Patient age (Mean ± SD)	53.78 ± 11.65	50.56 ± 10.08	54.16 ± 11.89	0.0247
Clinical subtypes (n(%))				
HR−/HER2+	138 (14.78)	5 (8.20)	5 (1.19)	0.0004
HR+/HER2−	432 (46.25)	39 (63.93)	312 (74.11)	
HR+/HER2+	126 (13.49)	10 (16.39)	60 (14.25)	
TNBC	159 (17.02)	4 (6.56)	7 (1.66)	
N/A	79 (8.46)	3 (4.92)	37 (8.79)	
TNM Stage (n(%))				
0	23 (2.46)	2 (3.28)	8 (1.90)	0.7430
I	146 (15.63)	8 (13.11)	67 (15.91)	
II	428 (45.82)	28 (45.90)	198 (47.03)	
III	270 (28.91)	20 (32.79)	114 (27.08)	
IV	40 (4.28)	2 (3.28)	24 (5.70)	
N/A	27 (2.89)	1 (1.64)	10 (2.38)	
Grade (n(%))				
1	27 (2.87)	3 (4.92)	22 (5.23)	0.3310
2	385 (40.45)	31 (50.82)	248 (58.91)	
3	420 (44.46)	22 (36.07)	113 (26.84)	
N/A	102 (12.22)	5 (8.20)	38 (9.03)	
Histology (n(%))				
Ductal carcinoma	800 (85.65)	48 (76.19)	361 (82.23)	< 0.0001
Lobular carcinoma	30 (3.21)	11 (17.46)	11 (2.51)	
Other	1 (0.11)	0 (0.00)	0 (0.00)	
N/A	103 (11.03)	2 (3.17)	49 (11.16)	

and number of features (number of genes included—see "Methods"). The optimum *k* value and feature set was selected by comparing our clustering results with the PAM50 clustering. Given that Her-2 enriched breast cancer is a well validated, biologically distinct subtype with copy number amplification of the ERBB2 gene, we selected a clustering result which grouped HER2− enriched samples consistently into a single cluster to be used for all subsequent analyses. This clustering result had the *k* value of 12 and a feature set of the top 5000 genes with the highest median absolute deviation within the cohort (Fig. 1).

Comparisons with PAM50, IntClust, and clinical subtypes indicated that our clustering results largely support the notion of distinct basal-like/IntClust10/TNBC and HER2− enriched/IntClust5/HER2 + groups of samples (Fig. 1). However, our clustering results were very different compared to previous clustering methods for the subtypes of HR + samples (Fig. 1a), suggesting that the grouping of HR + samples in our cohort may be driven by a different phenotypic paradigm relative to other populations.

Comparison of pathway expression across MyBrCa clusters

Next, to investigate the differences between our clusters, we conducted differential gene expression and pathway analyses. We compared each cluster to the other clusters with similar clinical subtypes, with a particular focus on the HR + clusters. Using GSEA, we found that the HR + clusters 2, 4, 5, and 7 differed from each other primarily in immune-related pathways such as the complement, inflammatory response, and interferon gamma (IFN-γ) response pathways (Fig. 2c, Additional file 2: Table S1). Pathway enrichment analyses returned similar results, with an over-representation of genes involved in immune-related pathways when comparing between the four HR + clusters. To follow up on these results, we calculated the scores for several gene expression-based immune-scoring methods, including ESTIMATE, IMPRES, the Ayers expanded IFN-γ gene set, and the

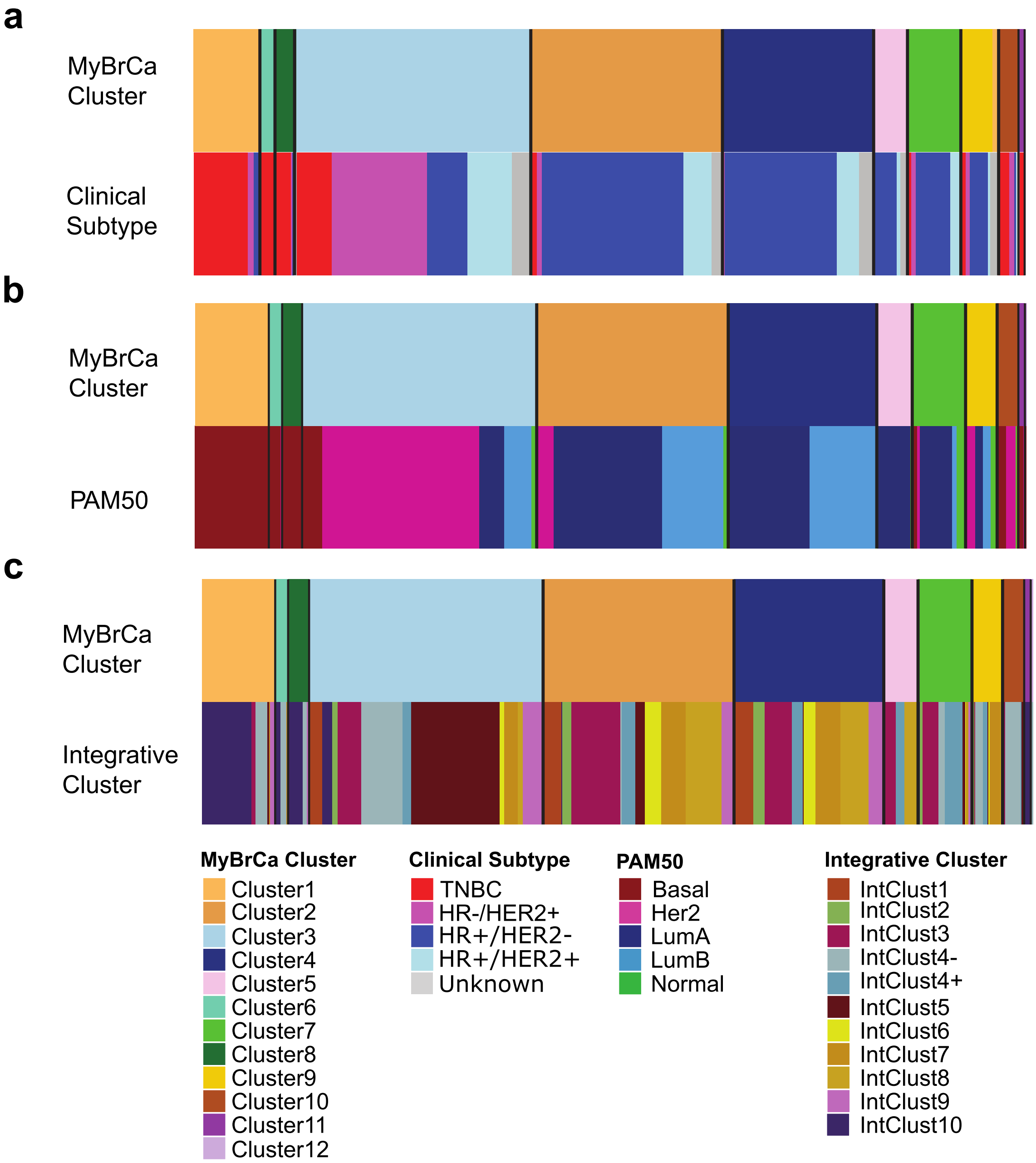

Fig. 1 Comparison of MyBrCa clusters with **a** clinical subtypes, **b** PAM50 and **c** Integrative Clusters

immune scores from Thorsson et al. (2018) for each sample (Fig. 2b). We found that there was a marked difference for many of these scores between our HR+clusters, consistent with the notion that the clustering of HR+samples in our cohort is driven by differences in immune-related phenotypes. We identified a group of clusters with consistently low or intermediate immune scores (Group 2), comprising of Clusters 2, 4 and 5, while Cluster 7 (Group 1) has consistently high immune scores.

We also investigated several KEGG pathways known to mediate IFN-γ, including cGAS-STING, T-cell receptor signalling, antigen processing and presentation, and TGF-β signalling pathways. We compared gene expression for genes in these pathways between our Group 1 and Group 2 samples using GSVA and ssGSEA, and the results indicated that all of these pathways were upregulated in Group 1 as compared to Group 2 (Additional file 1: Fig. S1).

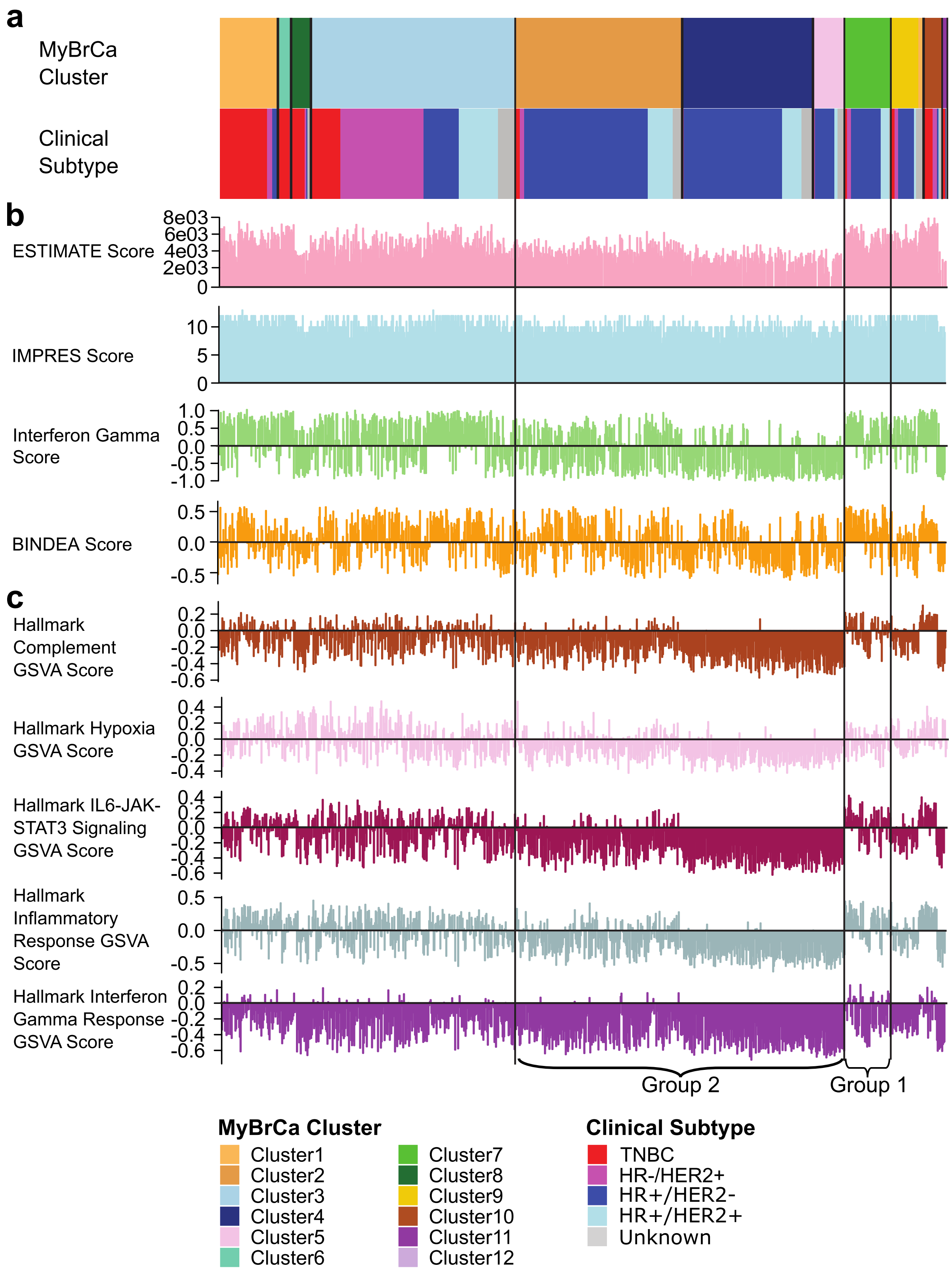

Fig. 2 Comparison of MyBrCa clusters with **a** clinical subtypes, **b** immune scores and **c** Gene Set Variation Analysis (GSVA) scores. HR+ clusters with high immune scores (Group 1) and low immune scores (Group 2) are identified

We validated our results by conducting unsupervised hierarchical clustering of transcriptomic profiles of the MyBrCa cohort. Hierarchical clustering of our samples gave similar results to our k-means clustering analyses in that our HR + /HER2− samples were again clustered into two main groups, and the segregation of the two groups was primarily driven by immune-related pathways according to pathway analyses (Additional file 1: Fig. S2, Additional file 3: Table S2).

In addition, we also evaluated the association between ER status and clustering of the HR + /HER2− samples in our cohort. We found that there is a significant difference in the prevalence of ER positivity between the 4 clusters that are primarily HR + /HER2− ($p = 1.3e-5$; Additional file 4: Table S3a), with Cluster 7/Group 1 having a higher prevalence of ER-negative samples. Similarly, we also found a significant difference in PAM50 subtype distribution between the four HR + /HER2− clusters ($p < 1.0e−5$). Cluster 5 contained no luminal B samples, while Cluster 7 had more basal-like, HER2− enriched, and normal-like samples compared to the other clusters (Additional file 4: Table S3b).

Genomic profiles of HR + /HER2− clusters

Given the growing importance of immunotherapy in cancer treatment, biomarkers to identify a subset of HR + /HER2− breast cancer patients that have an active immune microenvironment may be useful in both clinical and research settings. Since both whole-exome and shallow whole-genome sequencing data are available for the majority of samples in our study cohort, we next compared the mutational and copy number profiles of samples in the high immune scoring HR + /HER2− cluster (Group 1) to those in the intermediate and low immune-scoring HR + /HER2− clusters (Group 2), in order to identify molecular features that may be associated with an active immune microenvironment in a HR + /HER2− breast cancer background. First, we compared the prevalence of known somatic driver mutations in both groups and found that somatic *TP53* mutations were more common in Group 1, but no other driver mutations were significantly different (Fig. 3a). Next, we compared the prevalence of mutational signatures and found that the aging-associated mutational signature SBS1 was more prevalent in Group 2 (Fig. 3b). Contrary to expectations, samples in Group 1 had on average fewer somatic mutations (Fig. 3c) and CNAs (Additional file 1: Fig. S3) compared to Group 2, as well as lower scores for HRD-associated LOH and LST (Fig. 3d).

Immune profiles of HR + /HER2− clusters

Next, we asked if there was a difference in the composition of immune cells in the tumour microenvironment between the two groups that could be associated with immune activation. We first confirmed that there was a significant difference between the two groups by comparing several general immune score markers and found that Group 1 indeed had significantly higher scores for the combined Bindea gene set, Ayers expanded IFN-gamma gene set, IMPRES score, as well as scores for cytotoxic cells (Fig. 4a). Following that, we used CIBERSORT transcriptomic deconvolution to measure the relative abundances of 22 different immune cell types in the tumour microenvironment and compared the abundance of each cell type between the two groups. We found significant differences between the two groups across all 22 immune cell types (Fig. 4b). Interestingly, immune cell types associated with the adaptive immune system such as CD4 + memory T-cells, CD8 + T-cells, and B-cells had a higher relative abundance in Group 1, whereas Group 2 had a higher relative abundance of cells typically associated with the innate immune system, such as M2 macrophages and mast cells (Fig. 4b). Moreover, Group 1 samples were found to have higher abundance of lymphoid lineage cells, including B cells, CD8 + T cells and CD4 + T cells, whereas immune cells which were more abundant in Group 2 were mostly from the myeloid lineage, such as neutrophils and mast cells. There was also a notable difference in the prevalence of macrophage phenotypes between Group 1 and Group 2. Pro-inflammatory M1 macrophages were found to be more abundant in Group 1, while non-activated M0 macrophages and anti-inflammatory M2 macrophages are more abundant in Group 2 (Fig. 4b).

Survival analyses of MyBrCa clusters

Following that, we looked for prognostic differences between our clusters by conducting survival analyses. Survival analyses were performed using Cox proportional hazard models for overall survival, adjusting for tumour stage and grade. In general, HR + clusters, (Clusters 2, 4, 5 and 7) appeared to have better survival than the rest of the cohort (Fig. 5d, Additional file 1: Fig. S4a). As expected, clusters comprising of mainly TNBC subtypes, namely Clusters 1, 8 and 11, were associated with the worst survival, followed by Cluster 3 which comprises of a mix of clinical subtypes which are mainly HER2 + or TNBCs.

We also asked if survival differs between the HR + /HER2− clusters with distinct immune phenotypes (Group 1 versus Group 2). Group 1 was associated with

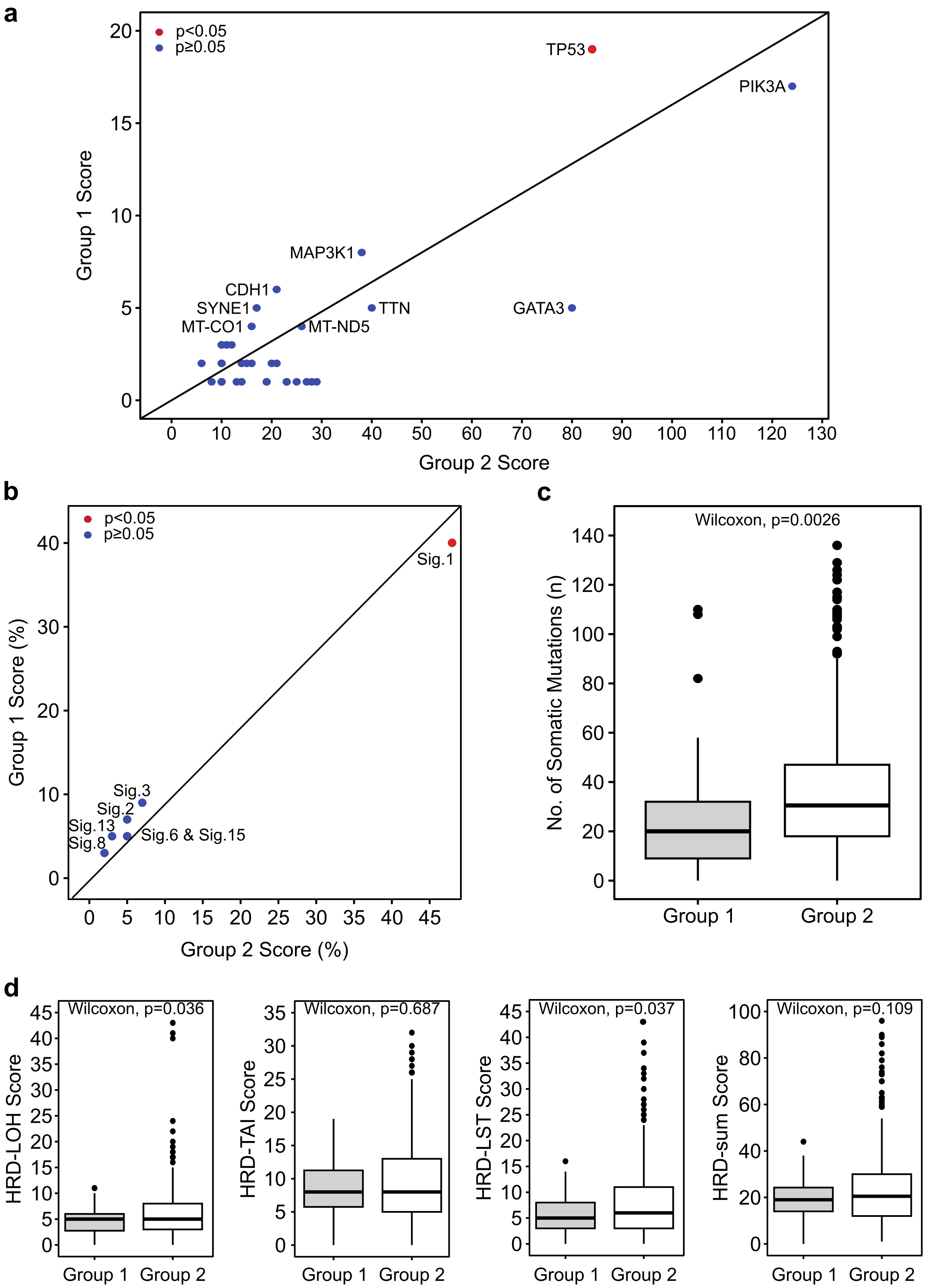

Fig. 3 Comparison of characteristics between HR+/HER2− Group 1 and Group 2. **a** Somatic driver mutations in Group 1 and Group 2. **b** Mutational signatures of Group 1 and Group 2. **c** Number of somatic mutations in Group 1 and Group 2. **d** Comparison of homologous recombination deficiency (HRD) scores between Group 1 and Group 2. LOH = loss of heterozygosity, LST = large-scale state transitions, TAI = telomeric allelic imbalance. A significance level of p = 0.05 was used

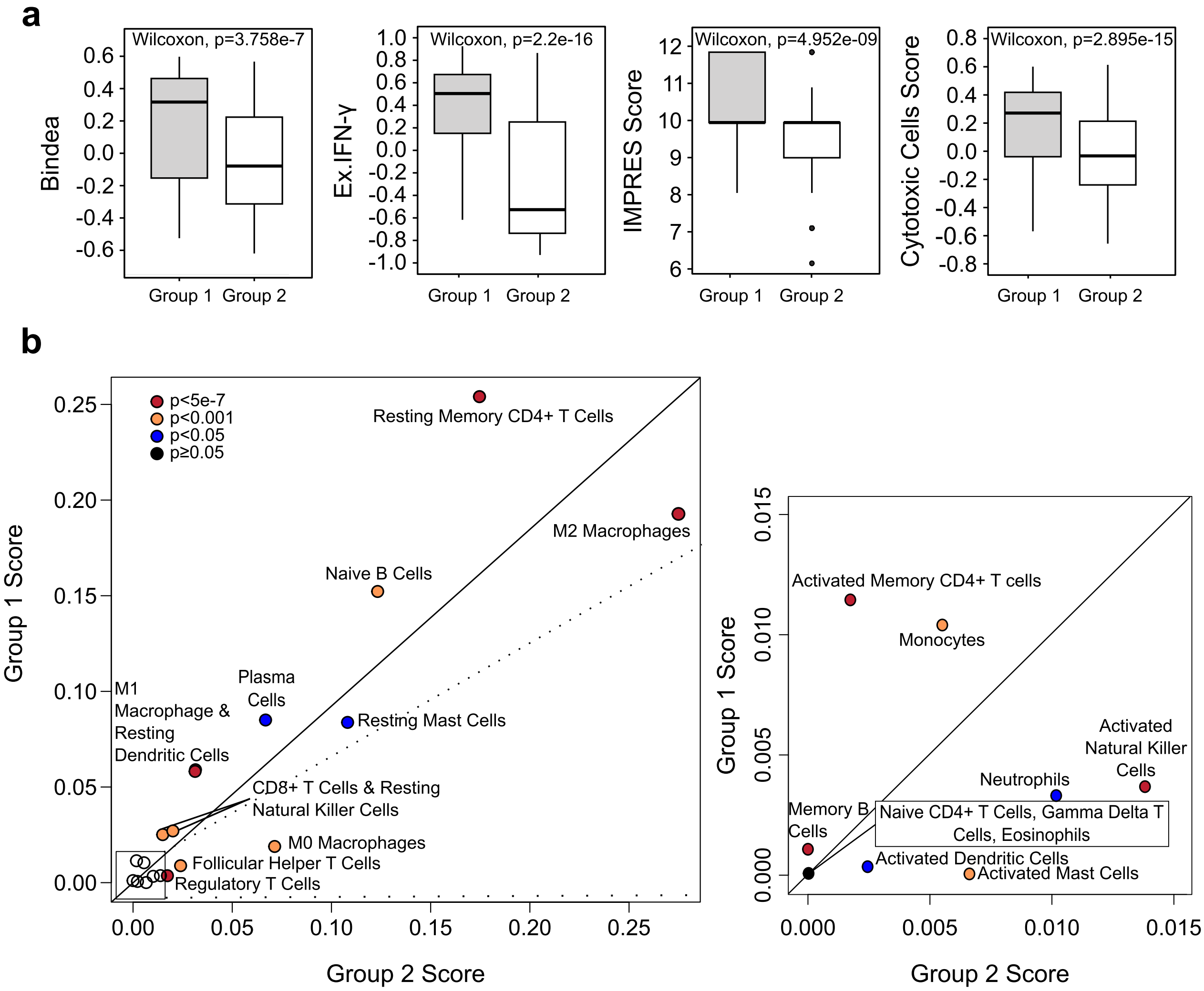

Fig. 4 Comparison of immune profiles of Group 1 and Group 2. **a** Scores of Group 1 and Group 2 for the Bindea combined gene set, Ayers expanded IFN-gamma gene set, IMPRES score and scores for the Bindea cytotoxic cells gene set. **b** Comparison of abundance of immune cell types in the tumour microenvironment in Group 1 and Group 2

a slightly poorer survival than Group 2, but the difference was not statistically significant (Fig. 5e, Additional file 1: Fig. S4b). Relative to Cluster 7 (Group 1), both Clusters 2 and 5 (Group 2) had better survival, although the difference was insignificant, while Cluster 4 (Group 2) had similar survival as Cluster 7 (Fig. 5f, Additional file 1: Fig. S4c).

(See figure on next page.)

Fig. 5 Prognostic analyses between different clusters **a** Overall survival by stage, adjusted for grade and clinical subtypes. **b** Overall survival by grade, adjusted for stage and clinical subtypes. **c** Overall survival by clinical subtypes, adjusted for stage and grade. **d** Overall survival by MyBrCa clusters, adjusted for stage and grade. In the right panel, small clusters with n < 20 were removed. **e** Overall survival by immune group, adjusted for stage and grade. **f** Overall survival by MyBrCa clusters adjusted for stage and grade, showing only clusters in Group 1 or Group 2. Adjustments were made using a Cox proportional hazard model. Sample sizes are reported in brackets. For **d** and **f**, Cox proportional hazard ratio p-values between the group with the best survival and the group with poorest survival are reported. For **d**, the *p*-value between Clusters 1 and 2 was reported. For **f**, the *p*-value between Cluster 7 and Cluster 5 was reported

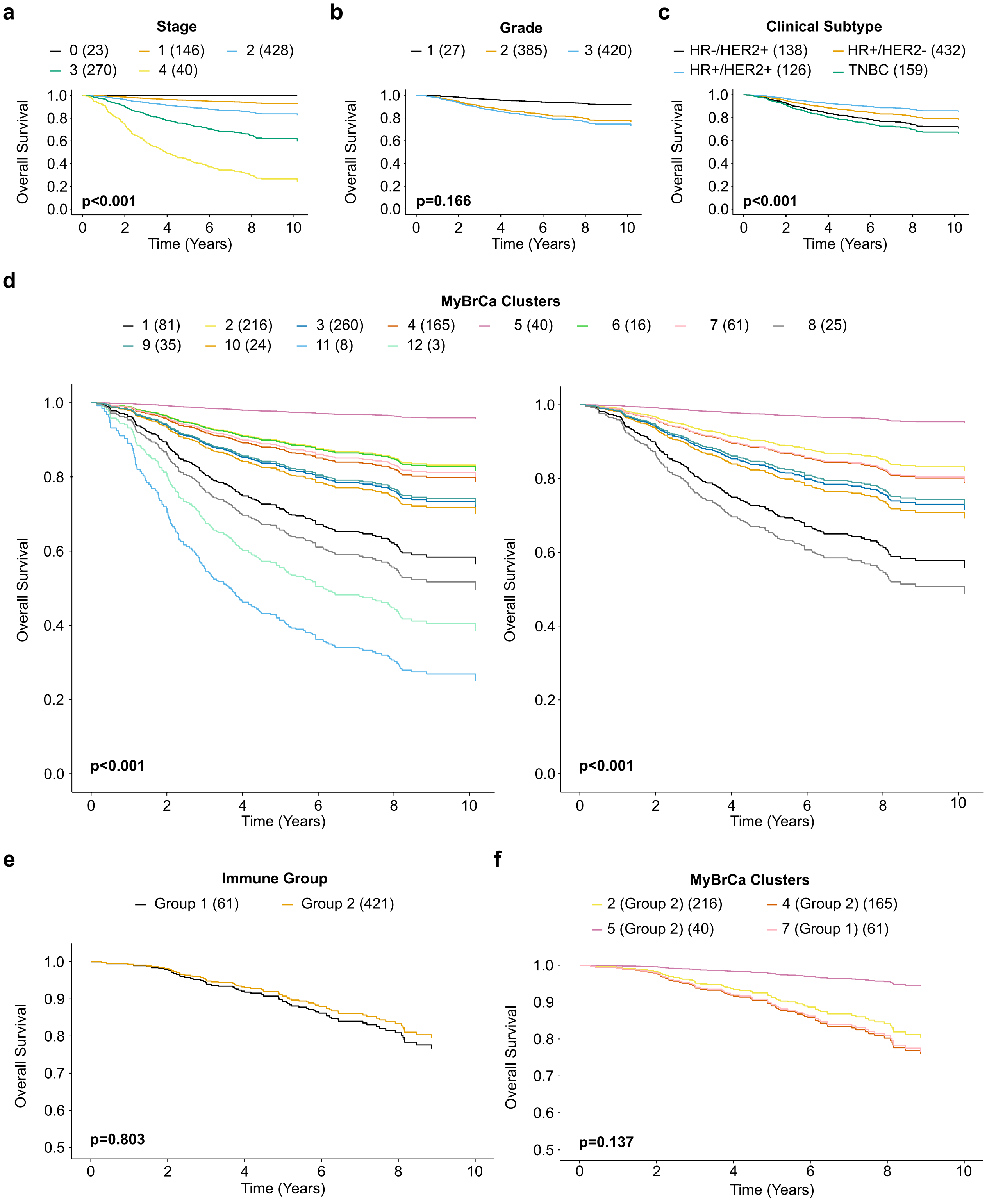

Fig. 5 (See legend on previous page.)

Discussion

The aim of this study was to look for novel subtypes of breast cancer from cluster analyses using transcriptomic data from a large Malaysian cohort of 934 breast tumours. Our classifier grouped these samples into 12 clusters which were highly associated with clinical subtypes. Comparisons of our clustering results to the PAM50 and IntClust classification schemes in the MyBrCa cohort suggest that intrinsic subtypes of breast cancer are largely conserved between Asian and Western cohorts for HER2+and TNBC subtypes. Survival analyses of our k-means clusters were also consistent with what might be expected based on the clinical subtypes which each cluster comprise of and also consistent with previous findings. On the other hand, our results from both k-means clustering and hierarchical clustering demonstrated a unique clustering pattern of Asian HR+/HER2− breast cancer samples. Instead of conforming to the conventional luminal A-luminal B paradigm, differences in immune-related pathways, immune scores as well as immune profiles suggest that clustering of these samples were driven by immune phenotypes.

This observation suggests that in Asian populations with HR+/HER2− breast cancer, the activation of the immune system may play a more crucial role compared to other populations. These findings are consistent with previous studies which reported a more immune-active tumour microenvironment in the Asian population, supporting the notion that immune responses could potentially have a heightened importance in Asian HR+/HER2− breast cancer patients [15, 36, 37]. These findings are also consistent with previous studies reporting that luminal breast cancers in Asians can be further stratified based on immune profiles [38, 39]. Additionally, comparison of the clustering results of our study to Integrative Clusters suggests that the majority of our HR+/HER2− immune high group (Group 1) also classified as IntClust 4, which was previously identified as having strong immune signatures in the METABRIC cohort [3], further supporting our findings.

Previous studies have suggested that the immune activity in HR+/HER2− breast tumours is associated with different outcomes compared to other breast cancer subtypes [40]. For example, the abundance of tumour-infiltrating lymphocytes (TILs) is associated with worse outcomes in HR+/HER2− breast cancer, whereas it is associated with better outcomes in other subtypes [41]. This difference may be due to a different balance of immune cell subsets within the tumour microenvironment—such a higher abundance of FOXP3+T-cells [42], or due to differences in the tumour immune microenvironment such as CTLA-4 expression [43]. In addition to an association with poorer prognosis, the presence of TILs in HR+/HER2− breast cancer is also thought to be an independent predictor of response to neoadjuvant chemotherapy [44].

The underlying factors driving the increased activation of the immune system in Asian breast cancer remains to be elucidated, however we suggest that this may be due to a combination of genetic and environmental factors. Genetic variants that lower the threshold for immune activation during breast cancer may be more common in Asian populations—for example, germline deletion of the *APOBEC3B* gene, a cytidine deaminase that has roles in both cancer mutagenesis and innate immunity, is much more common in Asian and Oceanic populations than in Western populations [45, 46]. Additionally, lifestyle factors such as parity and BMI are known to influence cancer immunity [47, 48] as well as risk for developing specific breast cancer subtypes [7, 8], and these lifestyle factors also differ between Asian and Western populations, with Asian populations generally having higher (albeit decreasing) parity [49, 50], and lower BMI [51] compared to their Western counterparts. Additionally, immunity generally declines with increasing age, and Asian breast cancer studies so far have indicated that the average age at diagnosis is lower than in Western populations [52, 53], although this may be due to a birth cohort effect [52, 54]. As a whole, these studies suggest plausible mechanisms or combinations of mechanisms by which the immune system in Asian breast cancer is more active.

Our study also revealed associations between the cluster of Asian HR+/HER2− breast cancers exhibiting high immune scores and several molecular characteristics. Specifically, this cluster was composed primarily of HR+/HER2− samples, but contained a higher number of ER-negative and basal-like/normal-like samples compared to other HR+/HER2− clusters, suggesting that this cluster straddles the continuum between hormonally-driven breast cancers and other TNBC-associated breast cancer subtypes. Additionally, we observed a higher prevalence of *TP53* somatic mutations within this cluster, suggesting that reduced functionality of this well-known tumour suppressor gene may contribute to immune activation in Asian HR+/HER2− breast cancer. Interestingly, despite the strong immune activation, these Asian HR+/HER2− breast cancers demonstrated lower overall numbers of somatic SNVs and indels, as well as fewer CNAs and lower HRD scores. This is opposite to what we might expect given the current paradigm for immune activation in cancer, where high tumour mutational burden generates a high neoantigen load, leading to immune activation via tumour antigen presentation [55, 56], suggesting that immune activation in this specific subset of Asian HR+/HER2− breast cancers may be governed by non-canonical mechanisms.

In our study, the cluster of Asian HR+/HER2− breast cancers characterized by high immune scores also displayed a notable association with specific immune cell types linked to adaptive immunity. This cluster exhibited a higher prevalence of immune cell populations associated with antigenic presentation and recognition pathways, suggesting a potential mechanism of immune activation in these tumours. Furthermore, we observed a lower number of immuno-suppressive M2 macrophages within this cluster, further supporting the notion of an immunologically active tumour microenvironment. These findings are consistent with the involvement of adaptive immune responses in promoting antitumor immunity in Asian HR+/HER2− breast cancer.

In conclusion, our classifier subtyped Asian breast cancer into biologically distinct clusters which revealed that the clustering of Asian HR+/HER2− breast cancer is driven by immune phenotypes. These findings may be important because HR+breast cancer is traditionally associated with a less immune active tumour microenvironment and thus less responsive to immunotherapy, but the findings from this study support other recent studies suggesting that a subset of HR+/HER2− breast cancer may be more likely to respond to immunotherapy.

Abbreviations

MyBrCa	The Malaysian Breast Cancer cohort
HRD	Homologous recombination deficiency
TNBC	Triple negative breast cancer
RNA-seq	RNA sequencing
CNA	Copy number aberration
GSVA	Gene set variation analysis
SBS	Single base substitutions
SNV	Single nucleotide variation
GSEA	Gene set enrichment analysis
IFN-γ	Interferon gamma

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-024-01826-5.

Additional file 1. Supplementary Figures S1–S4.

Additional file 2. Supplementary Table S1.

Additional file 3. Supplementary Table S2.

Additional file 4. Supplementary Table S3.

Acknowledgements

Cancer Research Malaysia receives charitable funding from the Scientex Foundation, Estée Lauder Companies, Vistage Malaysia, Yayasan PETRONAS, and Yayasan Sime Darby which contributed to the funding of this study. Funding was also provided by a research grant from the Newton-Ungku Omar Fund (MRC Ref: MR/P012442/1) to SFC and SHT. OMR, CC, and SFC also receive funding from Cancer Research UK. All genomics work was undertaken by the Genomics Core Facility CRUK Cambridge Institute.

Author contributions

JWP, MR and SHT contributed to the experimental design, data analysis, supervised experiments, wrote the manuscript and generated figures. WQC contributed to data analysis, writing of manuscript and generation of figures. SNH, TI, LYT, SJ, MHS, CHY, PR, LML, and AMT contributed to sample collection and processing and data collection, while OMR and SFC generated and processed sequencing data. PR and LML provided histopathology expertise, and collected clinical data together with MHS, TI, SJ, LYT, CHY and AMT. CC, SFC and SHT contributed to obtaining funding for the project. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

Availability of data and materials

The whole exome sequencing and RNA-seq data included in this study are available in the European Genome-phenome Archive under Accession nos. EGAS00001006518 and EGAS00001004518. Access to controlled patient data will require the approval of the Data Access Committee. Further information is available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

Patient recruitment and sample collection was reviewed and approved by the Independent Ethics Committee, Ramsay Sime Darby Health Care (Reference nos: 201109.4 and 201208.1), as well as the Medical Ethics Committee of the University Malaya Medical Centre (Reference no: 842.9). Written informed consent to participation in research was given by each individual patient.

Competing interests

The authors declare no conflict of interest.

Author details

[1] Cancer Research Malaysia, No. 1, Jalan SS12/1A, 47500 Subang Jaya, Malaysia. [2] Department of Surgery, Faculty of Medicine, University Malaya, 50603 Kuala Lumpur, Malaysia. [3] Subang Jaya Medical Centre, No. 1, Jalan SS12/1A, 47500 Subang Jaya, Malaysia. [4] Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Malaysia. [5] Department of Pathology, Faculty of Medicine, University Malaya, 50603 Kuala Lumpur, Malaysia. [6] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. [7] Department of Oncology, Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [8] NIHR Cambridge Biomedical Research Centre and Cambridge Experimental Cancer Medicine Centre, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK. [9] Faculty of Medicine, University Malaya Cancer Research Institute, University Malaya, 50603 Kuala Lumpur, Malaysia.

Received: 16 January 2024 Accepted: 15 April 2024
Published online: 22 April 2024

References

1. Bernard PS, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
2. Network CGA, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
3. Curtis C, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.
4. Pereira B, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.
5. Perou CM, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.
6. Rueda OM, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019;567:399–404.
7. Yang XR, et al. Associations of Breast cancer risk factors with tumor subtypes: a pooled analysis from the breast cancer association consortium studies. JNCI J Natl Cancer Inst. 2011;103:250.
8. Mao X, et al. Association of reproductive risk factors and breast cancer molecular subtypes: a systematic review and meta-analysis. BMC Cancer. 2023. https://doi.org/10.1186/s12885-023-11049-0.
9. Michailidou K, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017. https://doi.org/10.1038/nature24284.

10. Milne RL, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet. 2017. https://doi.org/10.1038/ng.3785.
11. Dorling L, et al. Breast Cancer Risk Genes: Association analysis in more than 113,000 women. N Engl J Med. 2021. https://doi.org/10.1056/nejmoa1913948.
12. Ho WK, et al. Polygenic risk scores for prediction of breast cancer risk in Asian populations. Genet Med. 2022. https://doi.org/10.1016/j.gim.2021.11.008.
13. Pitt JJ, et al. Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features. Nat Commun. 2018;9:4181.
14. Pan JW, et al. The molecular landscape of Asian breast cancers reveals clinically relevant population-specific differences. Nat Commun. 2020;11:1–12.
15. Kan Z, et al. Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. Nat Commun. 2018;9:1725.
16. Huang X, et al. Molecular portrait of breast cancer in China reveals comprehensive transcriptomic likeness to Caucasian breast cancer and low prevalence of luminal A subtype. Cancer Med. 2015;4:1016–30.
17. Yu, K., Lee, C. H., Tan, P. H. & Tan, P. Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations. http://clincancerres.aacrjournals.org (2004).
18. Tan M-M, et al. A case-control study of breast cancer risk factors in 7,663 women in Malaysia. PLoS ONE. 2018;13:e0203469.
19. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
20. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
21. Qi L, et al. Multi-omics data fusion for cancer molecular subtyping using sparse canonical correlation analysis. Front Genet. 2021;12:607817.
22. Scheinin I, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res. 2014;24(12):2022–32.
23. Yoshihara K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
24. Bindea G, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013;39:782–95.
25. Thorsson V, et al. The immune landscape of cancer. Immunity. 2018;48:812–30.
26. Auslander N, et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat Med. 2018;24:1545–9.
27. Ayers M, et al. IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest. 2017;127:2930–40.
28. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
29. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.
30. Birkbak NJ, et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discov. 2012;2:366–75.
31. Telli ML, et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple negative breast cancer. Clin Cancer Res. 2016;22:3764–73.
32. Favero F, et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015;26:64–70.
33. Sztupinszki Z, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. npj Breast Cancer. 2018;4:16.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
35. Liberzon A, et al. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1:417–25.
36. Chen C-H, et al. Disparity in tumor immune microenvironment of breast cancer and prognostic impact: Asian versus Western populations. Oncologist. 2020;25:e16–23.
37. Jin X, et al. Molecular classification of hormone receptor-positive HER2− negative breast cancer. Nat Genet. 2023. https://doi.org/10.1038/s41588-023-01507-7.
38. Park C, et al. Integrative molecular profiling identifies a novel cluster of estrogen receptor-positive breast cancer in very young women. Cancer Sci. 2019;110:1760–70.
39. Zhu B, et al. Immune gene expression profiling reveals heterogeneity in luminal breast tumors. Breast Cancer Res. 2019;21:1–11.
40. Pellegrino B, et al. Luminal breast cancer: risk of recurrence and tumor-associated immune suppression. Mol Diagn Ther. 2021. https://doi.org/10.1007/s40291-021-00525-7.
41. Denkert C, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol. 2018;19:40–50.
42. Liu S, et al. Prognostic significance of FOXP3+ tumor-infiltrating lymphocytes in breast cancer depends on estrogen receptor and human epidermal growth factor receptor-2 expression status and concurrent cytotoxic T-cell infiltration. Breast Cancer Res. 2014. https://doi.org/10.1186/s13058-014-0432-8.
43. Ostapchuk YO, et al. Functional heterogeneity of circulating T regulatory cell subsets in breast cancer patients. Breast Cancer. 2018. https://doi.org/10.1007/s12282-018-0874-4.
44. Denkert C, et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. 2010. https://doi.org/10.1200/JCO.2009.23.7370.
45. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. Population stratification of a common APOBEC gene deletion polymorphism. PLoS Genet. 2007;3:0584–92.
46. Pan JW, et al. Germline *APOBEC3B* deletion increases somatic hypermutation in Asian breast cancer that is associated with Her2 subtype, *PIK3CA* mutations and immune activation. Int J Cancer. 2021;148(10):2489–501. https://doi.org/10.1002/ijc.33463.
47. Swaby A, Atallah A, Varol O, Cristea A, Quail DF. Lifestyle and host determinants of antitumor immunity and cancer health disparities. Trends in Cancer. 2023. https://doi.org/10.1016/j.trecan.2023.08.007.
48. Krause AL, et al. Parity improves anti-tumor immunity in breast cancer patients. Oncotarget. 2017. https://doi.org/10.18632/oncotarget.20756.
49. Sim X, et al. Ethnic differences in the time trend of female breast cancer incidence: Singapore, 1968–2002. BMC Cancer. 2006. https://doi.org/10.1186/1471-2407-6-261.
50. Perry CS, Otero JC, Palmer JL, Gross AS. Risk factors for breast cancer in East Asian women relative to women in the West. Asia-Pacific J Clin Oncol. 2009. https://doi.org/10.1111/j.1743-7563.2009.01242.x.
51. Deurenberg P, Deurenberg-Yap M, Guricci S. Asians are different from Caucasians and from each other in their body mass index/body fat per cent relationship. Obes Rev. 2002. https://doi.org/10.1046/j.1467-789X.2002.00065.x.
52. Sung H, et al. Female breast cancer incidence among Asian and Western populations: more similar than expected. J Natl Cancer Inst. 2015. https://doi.org/10.1093/jnci/djv107.
53. Lin CH, et al. The emerging epidemic of estrogen-related cancers in young women in a developing Asian country. Int J Cancer. 2012. https://doi.org/10.1002/ijc.26249.
54. Shen YC, et al. Significant difference in the trends of female breast cancer incidence between Taiwanese and Caucasian Americans: implications from age-period-cohort analysis. Cancer Epidemiol Biomark Prev. 2005. https://doi.org/10.1158/1055-9965.EPI-04-0932.
55. Germano G, et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. Nature. 2017;552:116–20.
56. Yarchoan M, Johnson BA, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. Nat Rev Cancer. 2017;17:209–22. https://doi.org/10.1038/nrc.2016.154.

Publisher's Note