







# The molecular landscape of Asian breast cancers reveals clinically relevant population-specific differences

Jia-Wern Pan<sup>1,7</sup>, Muhammad Mamduh Ahmad Zabidi<sup>1,7</sup>, Pei-Sze Ng<sup>1,2</sup>, Mei-Yee Meng<sup>1</sup>, Siti Norhidayu Hasan<sup>1</sup>, Bethan Sandey <sup>3</sup>, Stephen-John Sammut<sup>3</sup>, Cheng-Har Yip <sup>2,4</sup>, Pathmanathan Rajadurai<sup>4</sup>, Oscar M. Rueda <sup>3</sup>, Carlos Caldas <sup>3,5,6</sup>, Suet-Feung Chin <sup>3,8</sup>✉ & Soo-Hwang Teo <sup>1,2,8</sup>✉

Molecular profiling of breast cancer has enabled the development of more robust molecular prognostic signatures and therapeutic options for breast cancer patients. However, non-Caucasian populations remain understudied. Here, we present the mutational, transcriptional, and copy number profiles of 560 Malaysian breast tumours and a comparative analysis of breast cancers arising in Asian and Caucasian women. Compared to breast tumours in Caucasian women, we show an increased prevalence of HER2-enriched molecular subtypes and higher prevalence of *TP53* somatic mutations in ER+ Asian breast tumours. We also observe elevated immune scores in Asian breast tumours, suggesting potential clinical response to immune checkpoint inhibitors. Whilst HER2-subtype and enriched immune score are associated with improved survival, presence of *TP53* somatic mutations is associated with poorer survival in ER+ tumours. Taken together, these population differences unveil opportunities to improve the understanding of this disease and lay the foundation for precision medicine in different populations.

<sup>1</sup>Cancer Research Malaysia, No. 1, Jalan SS12/1A, 47500 Subang Jaya, Malaysia. <sup>2</sup>University Malaya Cancer Research Institute, Faculty of Medicine, University Malaya, Kuala Lumpur, Malaysia. <sup>3</sup>Cancer Research UK, Cambridge Institute & Department of Oncology, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. <sup>4</sup>Subang Jaya Medical Centre, No. 1, Jalan SS12/1A, 47500 Subang Jaya, Malaysia. <sup>5</sup>Cambridge Breast Cancer Research Unit, CRUK Cambridge Cancer Centre, Cambridge, UK. <sup>6</sup>NIHR Cambridge Biomedical Research Centre and Cambridge Experimental Cancer Medicine Centre, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK. <sup>7</sup>These authors contributed equally: Jia-Wern Pan, Muhammad Mamduh Ahmad Zabidi. <sup>8</sup>These authors jointly supervised: Suet-Feung Chin, Soo-Hwang Teo. ✉email: [suet-feung.chin@cruk.cam.ac.uk](mailto:suet-feung.chin@cruk.cam.ac.uk); [soohwang.teo@cancerresearch.my](mailto:soohwang.teo@cancerresearch.my)

**B**reast cancer is a heterogenous disease, where histopathological and clinico-demographic features guide treatment choices—for example, use of endocrine treatment in oestrogen receptor-positive (ER+) tumours, use of HER2-targeted treatment in Her2-positive tumours and use of chemotherapy in women with poor prognostic features, such as high grade. The development of molecular classifiers based on gene expression<sup>1,2</sup> has led to a better understanding of the molecular subtypes of breast cancer, and the development of molecular-based prognostic tools such as OncoTypeDx has led to improved clinical decision-making—for example, in determining the benefit of chemotherapy in node-negative, ER+ disease<sup>3</sup>. Recognising that breast cancer is a copy number driven disease<sup>4</sup>, we have improved the stratification of breast cancers by integrating copy number alterations (CNAs) and gene expression in the classification of 2000 primary tumours from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)<sup>5</sup>. The subtypes identified, designated as Integrative Clusters (IntClust), have very distinct molecular features, drivers and clinical courses<sup>6,7</sup>.

The molecular characterisation of breast tumours has also enabled the development of new therapies, and the selection of patients for the appropriate therapies. Gene expression<sup>8,9</sup>, targeted sequencing<sup>7</sup>, whole-exome sequencing (WES)<sup>10–12</sup> and whole-genome sequencing (WGS) have revealed the genomics drivers of breast cancer, uncovering hitherto unrecognised therapeutic possibilities. Furthermore, mutational analysis has identified tumour mutational burden<sup>13</sup> and homologous recombination (HR) deficiency<sup>14</sup> as potential biomarkers for response to immunotherapy and poly (ADP-ribose) polymerase (PARP) inhibitors<sup>15,16</sup>, respectively.

Notably, the extent to which findings from these studies, conducted predominantly in women of European descent, can be applied to other populations where there are differences in genetic, lifestyle and environmental factors remains largely understudied. Triple-negative breast cancer is more common in African-American women and a recent genomic analysis of 194 tumours show an increased HR deficiency signature, pervasive *TP53* mutations and greater structural variations, indicating a more-aggressive biology<sup>17</sup>. Breast cancer in Asians tend to occur at a younger age, with a higher proportion of pre-menopausal, oestrogen receptor (ER) negative and human epidermal growth factor receptor 2 (HER2) receptor-positive disease (reviewed in Yap et al.<sup>18</sup>). A recent genomic analysis of 187 early-onset Asian breast cancers show a higher prevalence of *TP53* mutations and enrichment in immune signatures<sup>19</sup>. In addition, analysis of 465 triple-negative breast cancers in Chinese women demonstrated broad similarities in tumours of the same subtype arising in Asian and Caucasian women, although Chinese patients had a significantly higher proportion of the luminal androgen receptor (LAR) subtype of TNBCs relative to Caucasian or African-American patients<sup>9</sup>.

Given the potential impact of tumour subtypes, candidate drivers, mutational signatures and immune profiles on treatment options for breast cancer patients, and the hitherto lack of detailed information in Asian breast cancer patients, we have performed WES, shallow WGS (sWGS) and RNA-sequencing (RNA-seq) of 560 breast tumours from a cohort of Asian patients in Malaysia and report the impact of population differences on the molecular profiles of breast cancer. Relative to Caucasian women, we found a higher prevalence of HER2-enriched molecular subtypes, as well as a higher prevalence of *TP53* somatic mutations in ER+ Asian breast tumours. Importantly, we also found that Asian breast tumours have elevated immune scores. HER2 subtype and elevated immune scores were found to be associated with improved survival, whereas presence of *TP53* somatic mutations was associated with poorer survival in ER+

tumours. Taken together, our findings set the stage for improving precision medicine efforts for breast cancer in the Asian populations.

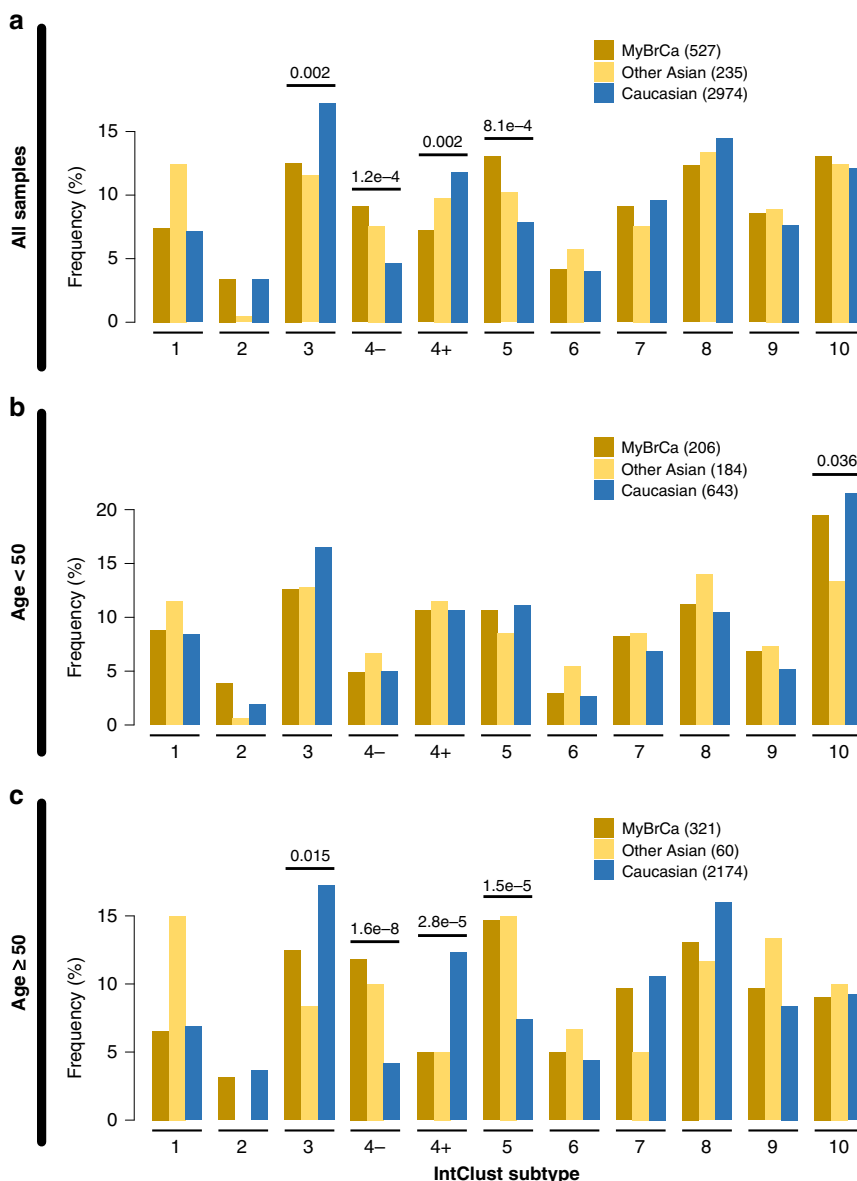
## Results

**Study population and clinicopathological characteristics.** Primary tumour tissue and blood samples (including three matched bilateral and one matched primary-recurrence samples) were obtained from 560 female patients with breast cancer treated at the Subang Jaya Medical Centre (SJMC), Malaysia between 2012 and 2017 who were recruited to the Malaysian Breast Cancer (MyBrCa) study<sup>20</sup>. The patients were recruited sequentially as seen in the clinic from a year to year period. After excluding samples that failed quality checks, data were generated for RNA-seq ( $n = 527$ ), WES ( $n = 546$ ) and sWGS ( $n = 533$ ). Compared with patients in The Cancer Genome Atlas (TCGA)<sup>10</sup>, our patients were younger, presented at later stages and had higher proportions of ductal carcinoma and Her2 positivity (28.9% MyBrCa versus 19.3% TCGA when NAs are excluded; Supplementary Table 1). The MyBrCa tumour cohort includes Malaysians from a number of different ethnicities, primarily Chinese (89%), Malay (4.6%) and Indian (3.4%) (Supplementary Table 1), and is as a whole genetically similar to other East/Southeast Asian populations according to genotyping analysis<sup>21</sup>. A principal component analysis of our RNA-seq data did not reveal any significant differences across different ethnicities (Supplementary Fig. 1).

**Asian breast tumours exhibit higher prevalence of Her2-positive disease.** Clustering of the tumour profiles of the MyBrCa samples according to IntClust<sup>5</sup> and PAM50<sup>2</sup> (Fig. 1 and Supplementary Fig. 2, respectively) showed a higher prevalence of the Her2-enriched molecular subtypes, particularly IntClust 5 (13.1% in MyBrCa versus 7.9% in Caucasians; Fig. 1a) and the PAM50 Her2-enriched subtype (23.3% in MyBrCa versus 9.9% in Caucasians; Supplementary Fig. 2a). We also observed a higher prevalence of ER-negative Integrative Cluster 4 (IntClust 4-; 9.1% in MyBrCa, 7.6% in other Asians versus 4.6% in Caucasians; Fig. 1a).

Given that the median age of diagnosis of breast cancer was 49 years in Asians and 61 years in Caucasians, we examined the distribution of subtypes in women above and below the age of 50. In women below 50 years of age ( $n = 206$ ), there were no statistically significant differences between Asians and Caucasians in IntClust subtypes (Fig. 1b), but there was an increased prevalence in the PAM50 Luminal B subtype (Supplementary Fig. 2b). In women above 50 years of age ( $n = 321$ ; Fig. 1c, Supplementary Fig. 2c), we see the significant differences noted above in the whole population analysis, i.e., there was a markedly higher prevalence of Her2-positive disease: (a) IntClust 4- (11.8% in MyBrCa, 10% in other Asians, 4.1% in Caucasians), (b) IntClust 5 (14.6% in MyBrCa, 15% in other Asians, 7.4% in Caucasians) and (c) PAM50 Her2-enriched subtype (27.4% in MyBrCa, 30% in other Asians, 10.6% in Caucasians), suggesting that it is in this age group where population-specific differences are mainly observed.

In addition, we also conducted a naive combined cluster analysis using gene expression data from the MyBrCa and TCGA Caucasian cohorts in order to determine whether there were any molecular subtypes that may be unique to Asians or Caucasians. However, unsupervised k-means consensus clustering did not reveal any exclusive clusters, but did support a higher prevalence of Her2-enriched subtypes and a lower prevalence of luminal subtypes in Asians (Supplementary Fig. 3).



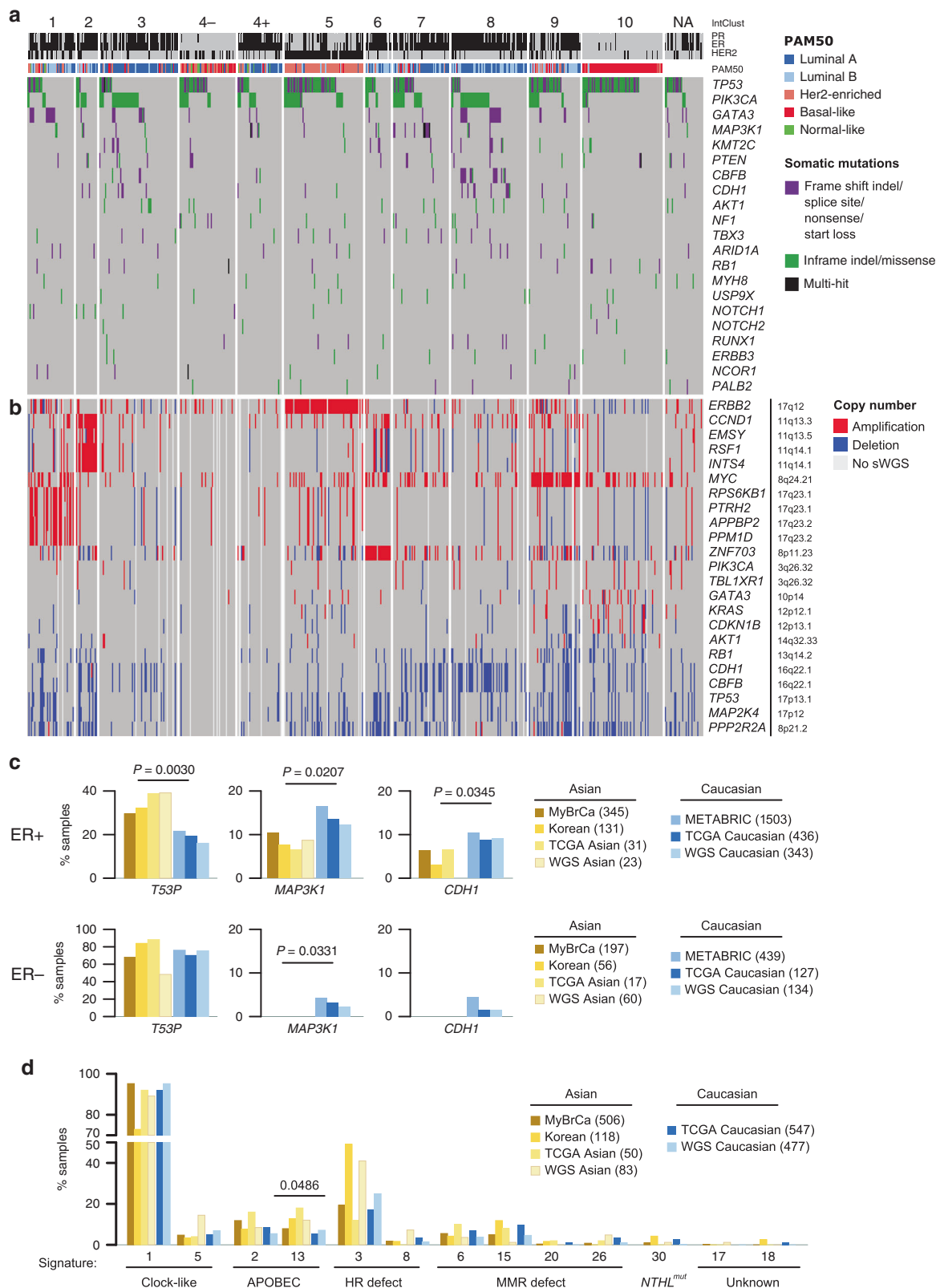
**Fig. 1 Molecular subtypes of Malaysian breast cancer.** Comparison of Integrative Cluster molecular subtype distribution across Malaysian (MyBrCa), other Asian (Korean, TCGA Asian) and Caucasian (TCGA Caucasian, METABRIC, Nik-Zainal 2016<sup>23</sup>) cohorts. Comparisons were done using the full cohorts **a** as well as with only patients below **b** or above **c** the age of 50. Numbers above the bars are *p* values denoting significant differences between Asians and Caucasians for that subtype, as determined by Pearson’s chi-square test. Numbers in the figure legend indicate sample size.

Finally, we classified the TNBC samples of the MyBrCa cohort into LAR, mesenchymal-like (MES), immunomodulatory (IM) and basal-like immune-suppressed (BLIS) TNBC subtypes using hierarchical clustering with the 80-gene signature from Burstein et al.<sup>22</sup> (Supplementary Fig. 4a). We found that the prevalence of each subtype was comparable with previous studies, with a relatively high prevalence of the LAR subtype (35%) that is consistent with reports from other Asian cohorts<sup>9</sup> (Supplementary Fig. 4b).

**Asian ER-positive breast tumours exhibit elevated TP53 mutations.** To understand the mutational architecture of MyBrCa tumours, we performed WES at a median coverage of 75× (range 5–123×) and 40× (range 5–77×) for 546 paired tumours and matched normal blood, respectively, and determined their somatic single-nucleotide variant and short insertion or deletion (indel) mutations. We identified 39,372 SNVs (18,508

nonsynonymous) and 1262 indels, with a median somatic mutation count of 45 SNVs and 2 indels per sample. We found similar subtype-specific nonsynonymous sSNVs and indel mutations as previously reported (Fig. 2a)<sup>7,23</sup>. The most frequently mutated genes are *TP53* (42.9%), *PIK3CA* (27.8%), *GATA3* (9.7%), *MAP3K1* (5.5%), *KMT2C* (4.2%), *PTEN* (4.2%), *CBFB* (4.0%), *CDH1* (4.0%), *AKT1* (3.3%) and *NF1* (3.1%). Within the molecular subtypes, we observed that *TP53* mutation rates were high in IntClusts 5 and 10, and low in IntClusts 3 and 8. In contrast, *PIK3CA* mutation rates were high in IntClusts 3 and 8, with 96.4% of the *PIK3CA* variants being missense variants and concentrated at the known hotspot positions (Supplementary Fig. 5). The majority of *GATA3* mutations were protein truncating (98.1%; Supplementary Fig. 5), with high frequencies in IntClusts 1 and 8. *MAP3K1* mutation frequencies were high in IntClust 7, whereas mutations of *CBFB* and *CDH1* were high in IntClust 8.

Interestingly, *TP53* mutations were more common in Asians compared with Caucasians (42.9% compared with 30–35%;



**Fig. 2** Mutational landscape of MyBrCa tumours. **a** Somatic nonsynonymous SNV and indel mutations of top mutated genes. Genes are sorted by the total mutation rates of MyBrCa cohort. **b** Copy number aberration of breast cancer-related genes. Genes are sorted according to difference in number of samples carrying amplified over deleted copy of the genes, and genes within the same locus are grouped together for clarity. Samples are ordered as in **a**. NA: IntClust classification could not be determined owing to unavailability of RNA-seq. **c** Comparison of mutational prevalence of *TP53*, *MAP3K1* and *CDH1* in Asian (MyBrCa, Kan et al., 2018<sup>19</sup> “Korean”, Asian TCGA, Korean samples from Nik-Zainal et al., 2016<sup>23</sup> “WGS Asian”) and Caucasian (METABRIC, Caucasian TCGA, Nik-Zainal et al., 2016<sup>23</sup> “WGS Caucasian”) breast tumours, separated by ER status. **d** Frequencies of samples in Asian or Caucasian datasets carrying breast cancer-related mutational signatures. *P* value from two-sided student’s *t* test.

Supplementary Fig. 6), and this difference was observed only in ER+ tumours and IntClust 8 (Fig. 2c, Supplementary Figs. 6–8;  $p = 0.003$  and  $p = 0.035$ , respectively). However, there was no significant difference in the location of the mutations, with 83% and 84.8% of mutations occurring in the DNA binding domain of *TP53* for ER+ Asian and Caucasian samples, respectively (86.6% vs 83% for ER- samples; Supplementary Fig. 9). *MAP3K1* mutation rates were lower in Asians irrespective of ER status (Fig. 2c;  $p = 0.02$  and  $p = 0.03$  in ER+ and ER-, respectively), whereas *CDH1* mutation rates were lower in ER+ Asian tumours ( $p = 0.03$ ), although there was no difference in *CDH1* mutation rates after accounting for histological subtype (Supplementary Fig. 10).

We performed driver gene analyses<sup>24</sup> (Supplementary Fig. 11, Supplementary Tables 2 and 3) and identified well-established breast cancer oncogenes and tumour suppressors. For example, regardless of ER status or ethnicity, *PIK3CA* and *AKT1* have high oncogene (ONC) but low tumour suppressor gene (TSG) scores, consistent with their established roles as known breast cancer oncogenes, while known tumour suppressors *TP53*, *RBI* and *PTEN* have elevated TSG scores. We also observed high TSG scores for *MAP3K1* and high ONC for *CDKN1B* in ER+ as previously noted<sup>7</sup>. We also observed preferentially higher *SF3B1* mutations in ER+ samples, with the recurrent K700E mutation enriched in Caucasians ER+ tumours<sup>25</sup> (Caucasian ER+ ONC = 58.6 vs Asian ER+ ONC = 12.5). We additionally examined pairwise association of somatic mutations (Supplementary Fig. 12) in Asian or Caucasian breast tumours. We observed mutual exclusivity between *TP53* and each of *CDH1* and *GATA3*, between *PIK3CA* and each of *CDH1* and *MAP3K1*, and between *GATA3* and *CBFB* as previously noted<sup>7</sup> in Asians and Caucasians tumours (Supplementary Fig. 12). Finally, an analysis of tumours with germline pathogenic variants in six high to moderate breast cancer susceptibility genes did not reveal any significant differences between our cohort and TCGA<sup>26</sup>, except for a marginally significant higher prevalence in germline *PALB2* mutations (Supplementary Table 4).

Analysis of CNA using whole-genome sequencing at 0.1× coverage revealed frequently altered regions in chromosomes 1, 8, 16 and 17 (Supplementary Fig. 13), similar to what has been found in previous large genomic studies (TCGA, METABRIC). High-level amplifications were observed in known oncogenes such as *ERBB2* and *MYC*, while deletions were observed in loci encompassing *TP53* and *MAP2K4*. We observed IntClust-specific CNAs (Fig. 2b and Supplementary Fig. 14): for example, amplification of *ERBB2* at 17q12 in IntClust5; amplification of 17q23 encompassing *RPS6KB1*, *PTRH2*, *APPBP2* and *PPM1D* in IntClust1, amplification of two distinct amplicons around 11q13 (*CCND1* and *EMSY*, *RSF1*, and *INTS4*) in IntClust2, and amplification of *ZNF703* at 8p12 in IntClust6.

We also used copy number data to estimate the fraction of cancer cells that harbour mutations for nine common driver genes, and found that most genes had median mutational cancer cell fractions (CCF) of 1, suggesting that mutations in these genes are clonal and tend to occur early during tumour development (Supplementary Fig. 15).

Together, WES and sWGS analyses suggest that the subtype-specific somatic SNVs, indels, CNA and CCF patterns in our cohort are similar to that previously described in other Asian or European populations, with the notable exception that *TP53* mutations are more common in ER+ Asian breast cancers.

**Asian and Caucasian breast tumours exhibit similar mutational signatures.** We compared the mutational signatures previously detected in breast tumours<sup>23,27,28</sup>: Signature 1 and 5

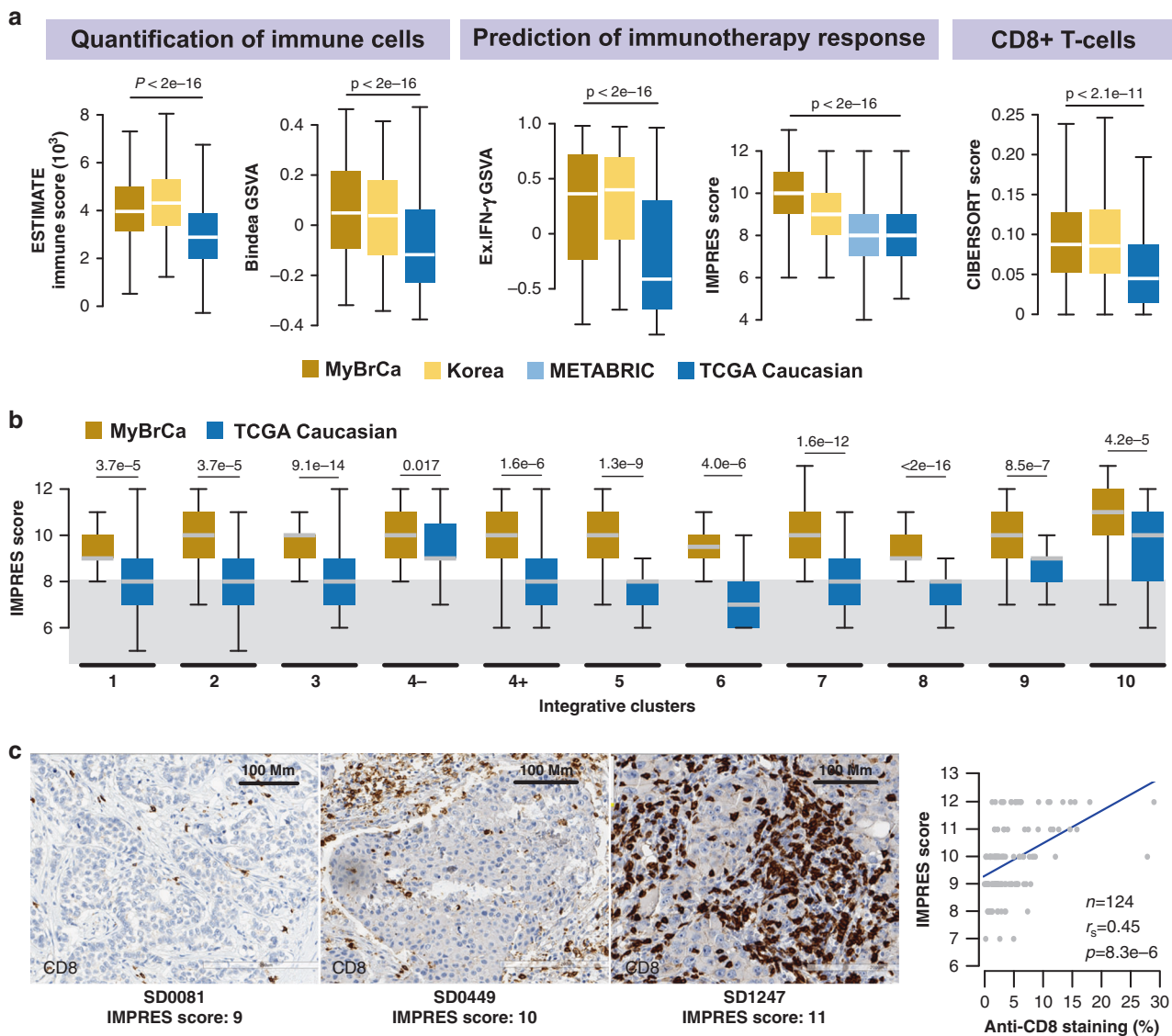
(clock-like), Signatures 2 and 13 (APOBEC enzymatic activity), Signatures 3 and 8 (DNA repair deficiency), Signatures 6, 15, 20 and 26 (mismatch repair (MMR) deficiency), and Signature 30 (base-excision repair protein Nth Like DNA Glycosylase 1 deficiency). We also included Signatures 17 and 18 that have been previously detected in breast tumours, yet of unknown aetiology.

We found only minor differences in the mutational signatures between Asian and Caucasian breast tumours (Fig. 2d, Supplementary Fig. 16). Whilst we observed a higher percentage of Korean samples (Korean and WGS Asian) with the HR deficiency Signature 3 as previously reported<sup>19</sup>, this was not observed in other Asian datasets (MyBrCa and TCGA Asian; Fig. 2d). Significantly, we found a higher percentage of Asian breast tumours with Signature 13, consistent with the higher prevalence of the *APOBEC3B* germline deletion polymorphism among the population ( $p = 0.0486$ ). No differences for other mutational signatures were observed (Fig. 2d).

We extended the analysis to molecular subtypes and found no significant difference in the distribution of mutational signatures based on ethnicity. IntClusts 3 and 8 samples primarily have high Signature 1, and moderate number of samples with Signatures 2, 13 and 3, and consistent with this, ER+ luminal A tumours by PAM50 classification show an elevated percentage of samples with Signature 1 as well (Supplementary Fig. 17, 18). IntClust 5 samples have high Signature 2 and 13, and consistent with this, the Her2-enriched tumours by PAM50 classification show an elevated percentage of samples with Signatures 2 and 13, suggesting increased APOBEC enzymatic activity<sup>29</sup> within the Her2-enriched cohort. By contrast, IntClust 10 samples have a higher percentage of samples with the HR deficiency Signature 3, and consistent with this, the basal-like samples by PAM50 classification have a high percentage of samples with Signature 3. Together, these results suggest that there is no significant difference in the carcinogenic processes driving the development of each subtype of breast tumours in both Asian and Caucasian women.

**Asian breast cancers exhibit enriched immune scores.** Given that population differences in the genetic, environmental and lifestyle exposures that drive carcinogenesis could shape the tumour microenvironment and lead to different outcomes, we performed pathway gene set enrichment analyses (GSEA) to determine pathways that are differentially activated between our cohort and TCGA. Interestingly, we found that 5 out of the top 15 most differentially expressed pathways were related to the immune system (Supplementary Table 5). We determined the immune scores for MyBrCa, Korean, METABRIC and TCGA Caucasian cohorts according to five different scoring methods: ESTIMATE<sup>30</sup>, gene set variation analysis (GSVA) using gene sets for immune cells<sup>31</sup>, GSVA using the expanded interferon-gamma gene set<sup>32</sup>, the IMPRES method<sup>33</sup> and CD8+ T cells scores from CIBERSORT<sup>34</sup>. Remarkably, all five methods showed significantly elevated immune scores in both Asian cohorts relative to the Caucasian cohorts (Fig. 3a). Furthermore, a multiple regression analysis of IMPRES scores for the MyBrCa and TCGA cohorts demonstrated that the difference in IMPRES scores between the two cohorts remained significant even after controlling for age, stage, histological subtype, menopausal status, tumour content, and HR/HER2 positivity (Supplementary Table 6).

As immune profiles are subtype-dependent, we compared IMPRES scores between MyBrCa and the TCGA Caucasian cohort across breast cancer subtypes. We observed the highest immune enrichment was in IntClust 10, which comprises mainly of TNBCs, in concordance with published studies<sup>35,36</sup>. More interestingly, we found that the IMPRES scores were significantly



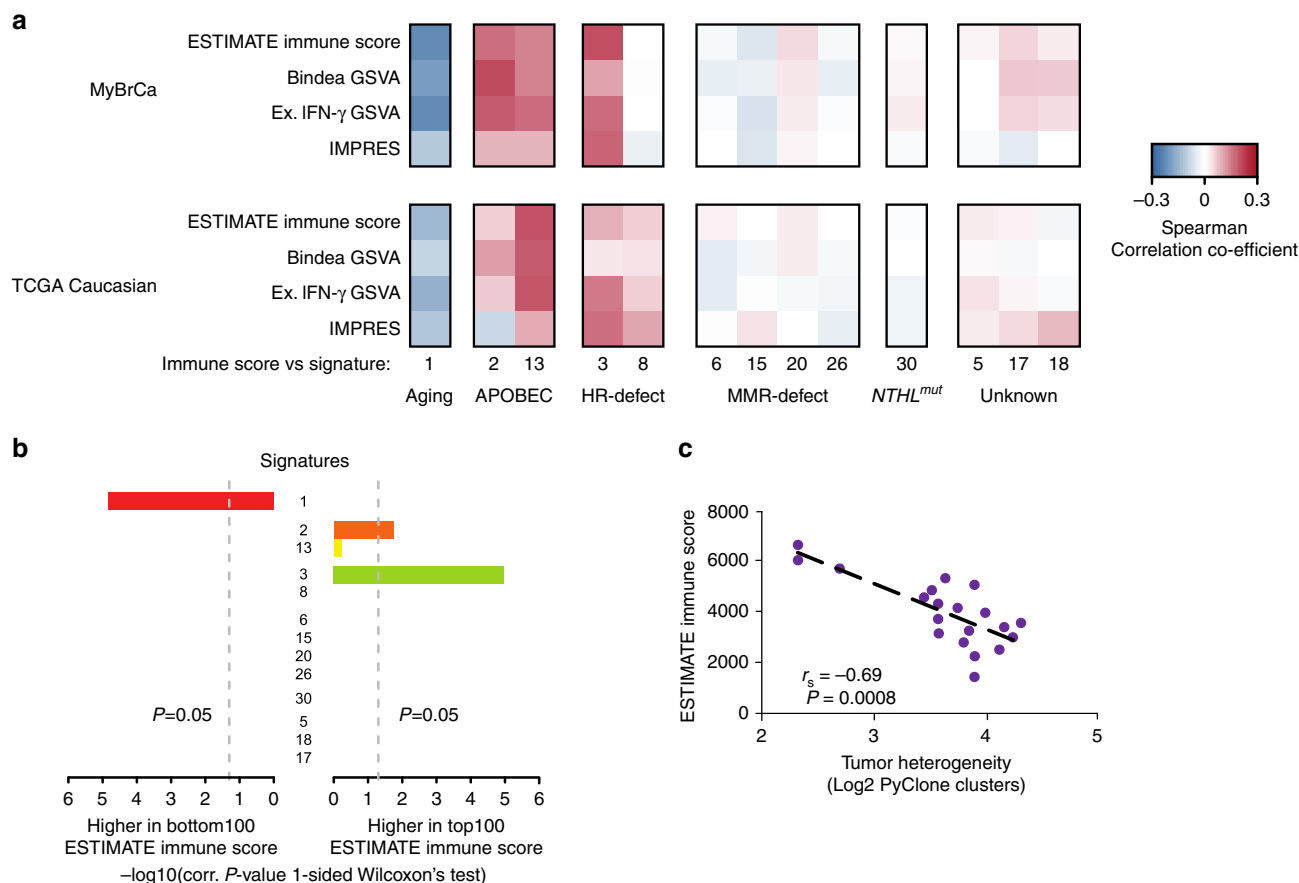
**Fig. 3** The tumour immune microenvironment of Malaysian breast cancer. **a** Comparison of the tumour immune microenvironment in breast tumours from the Malaysian (MyBrCa,  $n = 527$ ), Korean ( $n = 168$ ), METABRIC (MB,  $n = 997$ ) and Caucasian TCGA ( $n = 638$ ) cohorts across five quantification methods—from left to right: ESTIMATE immune scores, GSVA using Bindea et al.'s (2013)<sup>31</sup> combined immune gene set, GSVA using an expanded IFN- $\gamma$  gene set (Ayers et al. 2017<sup>32</sup>), IMPRES (Auslander et al. 2018<sup>33</sup>) and CD8+ T-cell scores from CIBERSORT. Outliers not shown.  $P$  values indicated are for one-way ANOVA. **b** Breakdown of IMPRES scores across MyBrCa and TCGA cohorts by IntClust subtype. Grey area indicates the threshold (score = 8) used by Auslander et al.<sup>33</sup> to separate high versus low scores. Outliers not shown.  $P$  values indicated are for two-sided student's  $t$  tests. **a–b** The boxes in box plots indicate 25th percentile, median and 75th percentile, whereas whiskers show the maximum and minimum values within 1.5 times the interquartile range from the edge of the box. **c** Validation of tumour immune scores in the Malaysian cohort using IHC; representative images shown with staining for CD8 in brown. IHC experiments were repeated in 124 samples across a range of IMPRES scores, with the results summarised in the figure on the right. Right: correlation between percentage of cell stained with anti-CD8 by IHC versus IMPRES scores.  $P$  value shown is for Spearman's correlation coefficient.

higher for MyBrCa cohorts compared to TCGA Caucasian samples independent of subtype classification (Fig. 3b). Similar patterns were found using the other immune scores (Supplementary Fig. 19) and classification methods (Supplementary Fig. 20), suggesting a systemic enrichment of immune scores in Asian breast cancers that occurs across the majority of molecular subtypes.

To identify immune cell types that drive the enrichment of immune scores in Asians, we used CIBERSORT on RNA-seq gene expression profiles to quantify the relative abundance of 22 different immune cell types in the tumour immune microenvironment. We found CD8+ T cells and macrophages are enriched

in Asian breast cancers (Fig. 3a, Supplementary Fig. 21), and this was consistent with TIMER and GSVA with Bindea gene sets (Supplementary Figs. 22, 23), suggesting an overall enrichment of cytotoxic NK and T cells in Asian tumours.

In a subset of 124 patients our cohort, we compared each patient's IMPRES score with anti-CD8 staining of tumour-infiltrating lymphocytes (TILs) in archival formalin-fixed paraffin-embedded (FFPE) blocks and found a high correlation ( $r_s = 0.45$ ; Fig. 3c). Similarly, IMPRES scores were highly correlated with CD3 IHC scores ( $r_s = 0.52$ ; Supplementary Fig. 24), and ESTIMATE immune scores were highly correlated with CD3 and CD8 IHC scores (Supplementary Fig. 24). Together, these data suggest that



**Fig. 4 Factors associated with immune scores. a** Heatmap depicting Spearman correlation of coefficient between mutation signatures in MyBrCa or TCGA Caucasian samples versus immune scores. **b** Bar plots depicting  $P$  values from one-sided Wilcoxon's test of the top versus bottom quartile MyBrCa samples, ranked according to ESTIMATE scores. **c** Comparison of ESTIMATE immune scores to the tumour heterogeneity of each MyBrCa sample, quantified as the log<sub>2</sub> of the number of clusters predicted by PyClone analysis. Samples were stratified into 20 groups according to ESTIMATE immune score.  $P$  value shown is for Spearman's correlation coefficient.

differences in immune scores derived from RNA-seq data are at least in part due to differences in the TILs within each tumour.

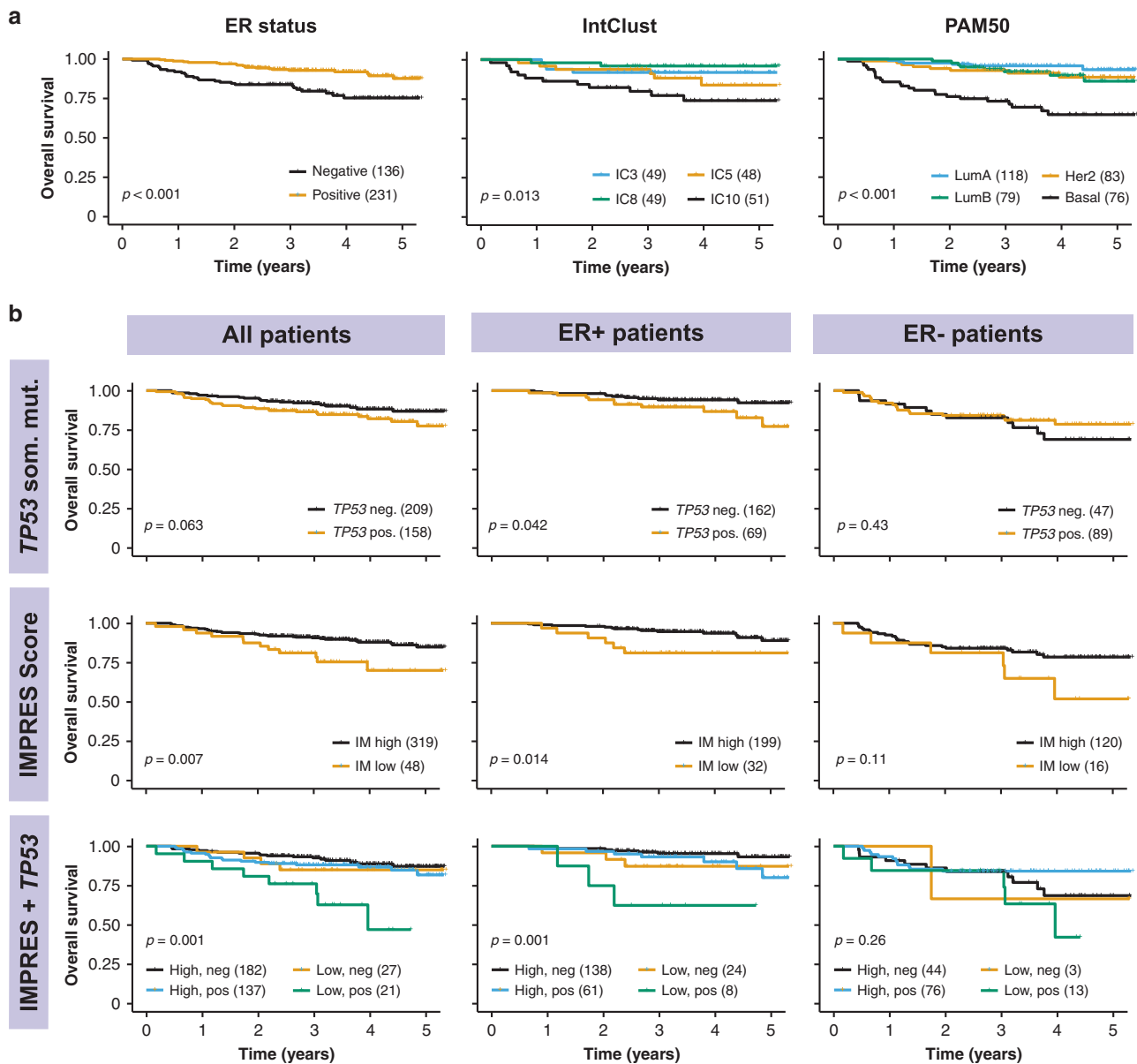
**Factors contributing to enriched immune scores.** To explore the factors contributing to enriched immune scores in Asians, we first explored whether the generation of neoantigens through different mutational processes is associated with immune scores. Indeed, we found that Signatures 2 and 13 (APOBEC) and Signature 3 (HR deficiency) are positively correlated with immune scores, whereas Signature 1 (aging) is negatively correlated (Fig. 4a, b).

We also asked if the immune signatures were associated with the underlying tumour heterogeneity of each sample. To quantify tumour heterogeneity, we used PyClone<sup>37</sup> to estimate the number of subclonal clusters in each sample. A comparison of ESTIMATE immune scores with the number of PyClone clusters revealed a strong negative association ( $r_s = -0.69$ ,  $p = 0.0008$ ; Fig. 4c), suggesting that tumours in the MyBrCa cohort with high immune scores have correspondingly low tumour heterogeneity. Furthermore, we also found that the MyBrCa samples have significantly lower tumour heterogeneity relative to TCGA samples across subtypes (Supplementary Fig. 25). These data are consistent with a model in which the stronger immune response in Asian breast cancer patients leads to higher selective pressure on tumour cells and, ultimately, more homogenous tumours in those patients<sup>38,39</sup>.

Next, in order to identify clinical or demographic factors that could contribute to the elevated immune scores observed in Asian

breast cancer samples, we conducted linear regression analysis of immune scores against all available clinical and demographic data, including age, tumour size, tumour stage, tumour grade, and tumour content. These analyses showed that elevated IMPRES scores were associated with molecular subtype and age, but not with other clinical or demographic factors (Supplementary Table 7). Following up on this result, multiple linear regression analysis of immune scores against all available molecular data, adjusting for age and molecular subtype, showed that mutational signature 3, tumour heterogeneity (PyClone clusters) and predicted neoantigen binding to be independently associated with IMPRES scores (Supplementary Table 8), although the overall predictive ability of the model was weak (adjusted  $R^2 = 0.13$ ).

Finally, to determine the biological pathways that are associated with higher immune scores, we conducted GSEA comparing tumours in the top quartile of IMPRES scores to tumours in the bottom quartile. We found that the systemic lupus erythematosus (SLE) pathway was the most enriched pathway for tumours in the top IMPRES score quartile (Supplementary Table 9). In addition, we also found enrichment in the cGAS-STING cytosolic DNA-sensing pathway, and decreased expression of the TGF- $\beta$  signalling pathway in the top quartile, which is concordant with previous studies on the tumour immune microenvironment (Supplementary Table 9)<sup>40-42</sup>. We confirmed these results by comparing IMPRES scores for our cohort to GSVA scores using the KEGG gene sets, and found a strong



**Fig. 5** Survival analyses of MyBrCa patients. **a** 5-year overall survival of MyBrCa patients stratified according to ER status (left), the four most common Integrative Clusters (middle), or PAM50 subtype (right). **b** 5-year overall survival of MyBrCa patients stratified according to presence (*TP53* Pos.) or absence (*TP53* Neg.) of *TP53* somatic mutations (top), tumour IMPRES score (“IM High” = 9 or more, “IM Low” = 8 or less; middle), or a combined analysis of IMPRES score and *TP53* somatic mutations (bottom). Results are shown for all patients (left), as well as for only ER+ patients (middle), or ER– patients (right). For all figures, only patients with more than 2 years of follow-up data were included. Brackets in the figure legends indicate sample size for each group, and  $p$  values shown are for unadjusted log-rank tests. Crosses indicate patient dropouts.

correlation between the two scores for all three pathways (Supplementary Fig. 26).

**Clinical impact of genomic profiles of Asian breast cancer.** To determine the potential clinical impact of the differences in molecular profiles of breast cancers in Asian women, we determined the impact of the observations on overall survival. As expected, we found that ER– patients had poorer survival compared with ER+ patients, patients with IntClust 10 had poorer survival relative to other IntClusts, and patients with basal tumours by PAM50 had poorer survival relative to other molecular subtypes (Fig. 5a). In ER+ patients alone, *TP53* mutations were associated with poorer survival (Fig. 5b). Patients with high IMPRES scores had significantly better survival in both unadjusted (Fig. 5b) and adjusted (Supplementary Fig. 27) models, and

the effect was stronger in ER+ patients (Fig. 5b). Interestingly, ER+ patients with both a low IMPRES score and *TP53* mutation appear to have markedly poorer survival than other patients (Fig. 5b), although the sample size was small ( $n = 8$ ). Overall, these data suggest that the differences in molecular profiles of Asian breast cancer patients could have important clinical implications, particularly in patients with ER+ tumours.

## Discussion

To our knowledge, this is the largest and most extensive molecular profiling study on breast tumours arising in Asian women. Our study of Malaysian breast tumours complements previous genomics studies of Asian breast tumours from China and South Korea<sup>9,19,23</sup>, and more importantly, enables more-comprehensive comparisons of Asian breast tumours with breast tumours arising



in women of European descent<sup>5,10,23</sup>. Comparisons conducted in this study revealed a higher prevalence of the Her2 molecular subtypes, higher prevalence of *TP53* mutations in ER+ disease and higher overall immune signatures in Asian breast cancer, all of which could impact clinical outcomes.

Despite the lower mean age of breast cancer in our cohort compared with TCGA, and the association of younger age with ER- breast cancer, we did not find an increased proportion of triple-negative breast cancer. In contrast to other Asian studies, which selected for individuals with young onset breast cancer<sup>19</sup>, we also did not find a large difference in the frequency of germline carriers for *BRCA1*, *BRCA2*, *PALB2*, *ATM* or *CHEK2* relative to TCGA. The prevalence of germline BRCA mutations that we report is slightly lower than that reported by studies that have looked at germline BRCA mutations in the Asian population<sup>43–45</sup>, which is likely owing to a higher mean age of diagnosis and our cohort being unselected for age and family history of breast cancer.

The higher prevalence of the Her2 subtype in our cohort, the South Korean and the TCGA Asian cohorts is consistent with immunohistochemistry-based epidemiological studies among Asians and Asian Americans<sup>46–49</sup>. It is becoming increasingly apparent that the lifestyle and genetic risk factors may have overlapping but sometimes distinct effects in different subtypes of breast cancer. In fact, a large study of 28,095 breast cancer patients has suggested that parity may be associated with early-onset Her2-positive breast cancers<sup>50</sup>. Incidentally, the MyBrCa cohort<sup>20</sup> has a higher median parity relative to the Caucasian cohort sequenced by Nik Zainal et al.<sup>23</sup>. Further work on the basis for population differences in risk is warranted to understand implications for prevention.

The higher prevalence of *TP53* mutations in ER+ breast cancer in our cohort, the South Korean and the TCGA Asian cohorts is consistent with previous findings<sup>19</sup>. However, unlike liver cancer where *TP53* mutations may be linked to population-specific risk factors<sup>51</sup>, we found no difference in the type or location of mutations nor in mutational signatures in Asian and Caucasian breast tumours, suggesting that the source of mutations may be similar in both populations. Regardless of the cause of the mutations, *TP53* mutations have been suggested to be prognostic, but with opposite effects in different subtypes of breast cancer<sup>52,53</sup>, and in ER+ /HER2- disease, *TP53* mutations appear to be associated with higher risk to recurrence and poorer prognosis<sup>54–56</sup>. Notably, expression of *TP53* mutations may have different effects in different populations<sup>57</sup>, highlighting the need for further clinical studies in diverse populations.

Finally, we observed an elevated immune score, driven by higher cytotoxic TILs, in both Malaysian and Korean breast tumours relative to the Caucasian tumours. In our cohort, elevated immune score is also associated with better overall survival. Our results are consistent with recent microarray-based analyses in Asian breast cancer patients<sup>58</sup>. Intriguingly, ethnic-specific differences in the tumour microenvironment, such as hypoxia have been reported and appeared to be associated with higher *TP53* and lower *PIK3CA* mutations in Asians<sup>59</sup>. We found that the elevated immune scores are associated with the SLE, cGAS-STING cytosolic DNA-sensing and TGF- $\beta$  signalling pathways, HR deficiency, neoantigen prediction and tumour heterogeneity. However, these associations only account for a small proportion of the population variance in immune scores, highlighting the need for further studies in this area.

In summary, this study may have important implications on our understanding of breast cancers arising in different populations. It has been reported that Asian-American breast cancer patients have better survival than other racial/ethnic groups<sup>60,61</sup>, and the molecular basis of breast cancers in Asian women

reported in this study highlights some of the possible explanations for these population-specific differences in outcome.

## Methods

**Biospecimen collection and pathological assessment.** In all, 661 tissue samples were collected from breast cancer patients recruited as part of the MyBrCa study at the Subang Jaya Medical Centre, Subang Jaya, Malaysia. The patients were recruited sequentially as seen in the clinic from a year to year period. The project was reviewed and approved by the Independent Ethics Committee, Ramsay Sime Darby Health Care (reference no: 201109.4 and 201208.1), and written informed consent was given by each individual patient. Matching blood samples were obtained prior to surgery, whereas tumour samples were sectioned from the primary tumour during surgery and were immediately frozen and stored in liquid nitrogen.

Patients were excluded from this study for the following criteria: no corresponding tumour samples ( $n = 5$ ), no corresponding germline samples ( $n = 5$ ) and those who withdrew consent ( $n = 12$ ). Tumour samples were further excluded after clinicopathological review if they were found to be from rare histological subtypes and other breast diseases ( $n = 5$ ). Tumour samples were then sectioned for DNA and RNA extraction, and 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 those with insufficient DNA ( $n = 8$ ) were excluded from the study.

A small minority of patients that were included in this study ( $n = 26$ ) received neoadjuvant chemotherapy prior to tissue collection. Tissue samples also included four cases of bilateral breast cancer and 1 case of recurrence.

**Sample selection and genomic material extraction.** DNA from blood samples was extracted using the Maxwell 16 Blood DNA Purification Kit with the Maxwell 16 Instrument, according to standard protocol. DNA from tumour samples was extracted with the QIAGEN DNeasy Blood and Tissue Kit according to standard protocol, followed by quantitation using the Qubit HS DNA Assay kit and Qubit 2.0 fluorometer (Life Technologies Inc). RNA from tumour samples was extracted using the QIAGEN miRNeasy Mini Kit with a QIAcube, according to standard protocol. Total RNA was quantitated using Nanodrop 2000 Spectrophotometer and RNA integrity was confirmed using Agilent 2100 Bioanalyzer. For DNA samples, only samples with a concentration above 20.0 ng/ $\mu$ L were included for sequencing. For RNA samples, only samples with a concentration of at least 20.0 ng/ $\mu$ L and a RIN number above 7 were included for sequencing.

**DNA-sequencing libraries.** DNA libraries were generated from 50 ng of genomic DNA using the Nextera Rapid Capture Exome kit (Illumina, San Diego, USA) as per manufacturer's instructions. Prior to exome capture, 4 nM pools of DNA libraries ( $n = 48$ ) was subjected to single end 50 shallow WGS. Exome capture was performed in pools of 3 and subjected to paired end 75 sequencing on a HiSEQ4000 platform (Illumina, San Diego, USA).

**RNA-sequencing libraries.** RNA libraries were prepared from 550 ng of total RNA using the TruSeq Stranded Total RNA HT kit with Ribo-Zero Gold (Illumina, San Diego, USA) as per manufacturer's instructions and subjected to paired end 75 sequencing on a HiSEQ4000 platform (Illumina, San Diego, USA).

**RNA-sequencing alignment and quality assessment.** RNA-seq reads were mapped to the hs37d5 human genome and the ENSEMBL GRCh37 release 87 human transcriptome using the STAR aligner (v. 2.5.3a)<sup>62</sup>. Samples with <15 million mapped fragments were excluded from further RNA-seq analyses. Gene-level aligned fragment counts were generated using featureCounts (v. 1.5.3), while gene-level expression in transcripts-per-million (TPM) was calculated using RSEM (v1.2.31)<sup>63</sup>. Genes with an average count of less than 10 or an average TPM score of <0.1 were filtered out and excluded from further analyses. Variant calling from RNA-seq data for sample ID reconfirmation and downstream analyses was also conducted using the GATK Best Practices workflow for RNA-seq—in brief, STAR-mapped reads were processed to mark duplicates with Picard, followed by exon splitting & trimming and base recalibration using GATK, and lastly variants were called using HaplotypeCaller. Sample identities were verified as described below in the WES section.

**Molecular subtyping.** Gene-level count matrices for the SD and TCGA cohort were transformed into  $\log_2$  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). In addition, each matrix was quantile-normalised and subtyped according to integrative clusters using the iC10 R package (v. 1.5). For the METABRIC cohort, the normalised microarray expression matrix from the discovery set was used for GeneFu and iC10 subtyping without modification. For IntClust 4, we designated each sample as being ER+ or ER- by fitting a two-component mixture model to the distribution of *ESR1* expression in TPM using an expectation-maximisation algorithm as implemented in the mixtools (v. 1.1) package in R.

**Unsupervised cluster analysis.** K-means consensus clustering was conducted using gene-median centred TPM gene expression scores for the MyBrCa and TCGA Caucasian cohorts, using the ConsensusClusterPlus (v.1.46) package in R.

**Classification of TNBC subtypes.** TPM gene expression scores for MyBrCa TNBC samples were first normalised using gene-median centreing. Then, the 80 genes from the TNBC classifier in Burnstein et al.<sup>22</sup> was used to conduct hierarchical clustering, as implemented in the heatmap.2 function in R using default options. Each cluster was assigned to the four subtypes according to pathway expression: high expression of hormone-related pathways for the LAR subtype, high expression of immune pathways and developmental growth genes as the IM, low expression of immune pathways but high expression of developmental growth genes as the BLIS subtype, and finally the remaining cluster with moderate expression of epithelial–mesenchymal transition-related pathways as the MES subtype.

**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)<sup>30</sup>, as well as with GSVA (v. 1.26) using the combined immune cell gene sets from Bindea et al.<sup>31</sup>. For each sample, we also scored the immune features predictive of checkpoint inhibitor immunotherapy using IMPRES scores<sup>33</sup> (only 14 out of 15 of the predictive features were available in our datasets) as well as with GSVA using the Expanded IFN-gamma gene set from Ayers et al.<sup>32</sup>. The relative abundance of specific immune cell populations was estimated from RNA-seq TPM scores with the CIBERSORT<sup>34</sup> and TIMER<sup>64</sup> web tools, as well as through GSVA with individual immune gene sets from Bindea et al.<sup>31</sup>. All box plots shown are constructed in the same way—the boxes indicate 25th percentile, median, and 75th percentile, whereas whiskers show the maximum and minimum values within 1.5 times the inter-quartile range from the edge of the box.

**Regression analyses of clinical and molecular features.** To assess the associations between various clinical and molecular features, we performed multiple regression analyses using the base stats package in R (v. 3.5.1). For continuous variables such as IMPRES score, regular linear regression was applied using the “lm” function to estimate regression coefficients and statistical significance.

**Shallow WGS alignment and CNA assessment.** The sequenced reads were mapped to the hg19 reference genome using bwa-mem, sorted using samtools and deduplicated using picard (<http://broadinstitute.github.io/picard>). Mapped reads were analysed using QDNaseq<sup>65</sup> to obtain 100 kb segmented copy number profiles using standard protocol and default parameters. Copy number aberrations 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.

**WES analysis alignment and quality assessment.** The sequenced reads were trimmed for Illumina adapters using trim\_galore! (<https://github.com/FelixKrueger/TrimGalore>) and mapped to the human genome b37 plus decoy genome using bwa-mem version 0.7.12. The mapped reads were merged and sorted using samtools, and de-duplicated using picard (<http://broadinstitute.github.io/picard>). Realignment and base quality score recalibration were performed using GATK3<sup>66</sup>. Realignment around indels was done concurrently on the tumour and normal pairs. For bilateral cases, the two tumours and normal were realigned simultaneously. Sample identities were verified by determining the concordance of the genotypes using GATK3 HaplotypeCaller<sup>66</sup>. In brief, the haplotypes at 2000–5000 dbSNP positions were called for all samples (RNA-seq and WES experiments) in an all-versus-all manner. True match will typically have upwards of 95% concordance between samples from the sample individuals. Samples with ambiguous identities were excluded from further analyses.

**Data download.** We downloaded the TCGA<sup>10</sup> and METABRIC<sup>5</sup> data via the Genomic Data Commons Data Portal or cBioPortal<sup>67</sup>. We additionally downloaded data for Nik-Zainal et al.<sup>23</sup> (from <ftp://ftp.sanger.ac.uk/pub/cancer/Nik-ZainalEtAl-560BreastGenomes>) and Zhengyan Kan et al.<sup>19</sup> (from <https://www.nature.com/articles/s41467-018-04129-4>). For Nik-Zainal WGS somatic mutation analyses, we delimited our analysis to only the Nextera Rapid Capture regions with additional flanking 100 bp.

**Short-nucleotide variants calling.** We first constructed panel-of-normals by artefact detection mode of GATK3 Mutect2<sup>66</sup>, retaining only sites that are present in two or more normal samples. In all steps, we only concentrated on the Nextera Rapid Capture regions with additional flanking 100 bp. For SNVs, we used positions that are called by Mutect2, and applied following filters: minimum 10 reads in tumour and 5 reads in normal samples, OxoG metric <0.8, variant allele frequency (VAF) 0.075 or more, *p* value for Fisher’s exact test on the strandedness of the reads 0.05 or more, and  $S_{AF} > 0.75$ . For positions that are present in five samples or more, we removed two positions that were not in COSMIC and in single tandem

repeats. We also removed variants that have VAF at least 0.01 in gnomAD, and considered only variants that are supported by at least four alternate reads, with at least two reads per strand. For indels, we also required the positions to be called by Strelka2<sup>68</sup>. We manually salvaged an indel in PALB2 gene in sample SD0028 which we have identified using PCR. We annotated the variants using Oncotator version 1.9.9.0. We plotted mutation positions as lollipop plots using MutationMapper<sup>67</sup>.

**Cancer cell fractions.** CCFs were calculated for nine common breast cancer driver genes using VAFs, copy number and tumour purity estimates obtained from WES, following the methodology described in Pereira et al.<sup>7</sup>. Copy number and tumour purity were generated for this analysis using the ASCAT R package (v. 2.5.2) on allele counts generated from WES bam files using alleleCounter (v. 4.0.1).

**Driver genes analysis.** We considered samples from WGS and WES experiments (MyBrCa, TCGA Asian, Korean, WGS Asian, TCGA Caucasian, WGS Caucasian) for driver gene analyses<sup>24</sup>. We considered genes that are mutated in at least 1% of samples in Asian ER–, Asian ER+, Caucasian ER+ or Caucasian ER–. We calculated ONC score as the percentage of missense mutation counts at the most recurrent amino-acid positions over all mutations, and considered only genes with at least five recurrent mutations. We calculated TS (score) as the percentage of inactivating mutation counts over all mutations, and considered only genes with at least five inactivating mutations. We selected genes with either ONC or TSG scores of >20, and queried the Integrative Onco Genomics database ([www.intogen.org](http://www.intogen.org)) for known cancer driver genes (Supplementary Tables 2 and 3). We sorted 40 genes according to the enrichment in Asian ER+ over Caucasian ER+ and plotted the prevalence and the ONC/TSG scores (Supplementary Fig. 11).

**Association patterns of driver genes.** We considered the top 15 mutated genes from driver genes analysis. We calculated odds ratio and determined the likelihood of co-occurrence or mutual exclusivity of missense or inactivating mutations using Fisher’s exact test. We displayed associations with *p* value of <0.05 after Bonferroni correction and plotted the log odds ratio.

**Mutational signatures.** We used deconstructSigs<sup>69</sup> to determine the weights of previously reported breast cancer mutational signatures (Signatures 1, 2, 13, 3, 8, 6, 15, 20, 26, 5, 17, 18 and 30) from COSMIC matrices (version 2, March 2015) in samples with at least 15 sSNVs. We determined the proportions of mutational signatures in the different data sets by combining the variants from all samples, and used the function plotSignatures to plot the mutational spectra. To determine the difference in mutational signature weights between the top and bottom immune quartile of MyBrCa samples, we ranked the samples by their ESTIMATE scores, and performed one-sided rank sum Wilcoxon’s test on the two categories.

**Germline mutation analysis.** Carriers of deleterious germline mutations in *BRCA1*, *BRCA2*, *TP53*, *PALB2*, *ATM* and *CHEK2* were identified from targeted sequencing of the MyBrCa cohort (BRIDGES study, in review), which were confirmed with Sanger sequencing. Comparable data from TCGA was taken from Huang et al.<sup>26</sup>.

**Correlation between immune scores and mutational signatures.** We calculated Spearman’s correlation coefficient between the mutational signatures and immune scores and plotted the correlations as heatmaps.

**Differential expression and pathway analyses.** Gene-level count matrices were normalised using the “Trimmed Mean of *M*-values” method implemented in the edgeR (v. 3.20.9) R package. The count matrices were then transformed into logCPM using the voom function from the limma package in R. Differentially expressed genes were determined by empirical Bayes moderation of the standard errors towards a common value from a linear model fit of the transformed count matrices as implemented in the limma package, with the threshold for differential expression set as false discovery rate (FDR) < 0.001 and absolute log fold-change >0.2. For pathway analysis, DE genes were further filtered for FDR < 1e-50, and ranked according to absolute log fold-change. The top 1000 ranked genes were queried in Reactome ([www.reactome.org](http://www.reactome.org)) to determine the top enriched pathways. Further pathway analyses were also conducted using GSEA (v. 3.0) on quantile-normalised gene-level count data. GSEA was run for 1000 permutations with the v. 6.2 Hallmark and KEGG gene sets from MSigDB.

**Validation of immune scores with immunohistochemistry.** FFPE blocks for 207 patients with sequencing data were sectioned and stained for anti-CD3 (clone 2GV6, predilute; Ventana Medical Systems), anti-CD4 (clone SP35, predilute; Ventana Medical Systems), anti-CD8 (clone SP57, predilute; Ventana Medical Systems) and anti-PD-1 (clone SP263, predilute; Ventana Medical Systems) using an automated immunostainer (Ventana BenchMark ULTRA; Ventana Medical Systems, Tucson, AZ) without any dilution. Stained slides were digitised using an Aperio AT2 whole slide scanner. CD3, CD4 and CD8 staining was quantified using the Aperio Positive Pixel digital pathology tool and PDL-1 expression was

determined using the Combined Positive Score system. The data were further verified by a pathologist (PR).

**Quantification of tumour heterogeneity.** Tumour heterogeneity was determined using PyClone (v 0.13.1)<sup>37</sup> with default options to estimate the number of sub-clonal clusters within each tumour sample. Allele counts used for the PyClone input were extracted from the GATK output MAF files, whereas copy number input data were generated by ASCAT (v. 2.5.2) from WES allele counts generated by alleleCounter (v 4.0.1). Tumour heterogeneity was also separately quantified for each sample in the MyBrCa and TCGA cohorts using the MATH method described in Pereira et al.<sup>7</sup> from MAF files from WES.

**Neoantigen analysis.** Sample HLAs were determined using Polysolver from tumour and normal DNA WES data. Only HLA alleles that were concordant in tumour and normal WES data were considered. Somatic mutations were annotated using VEP. All possible neoantigen peptides (9- to 11-mers) encompassing the nonsynonymous mutations were predicted using a combination of NetMHCpan and NetMHC on the pVAC-seq platform<sup>70</sup>. Only neoantigens with predicted binding of less than 500 nM were considered.

**Survival analysis.** Overall survival data were obtained for each patient by querying their names and identity card numbers against the Malaysian National Registry records of deaths. Patients that did not return any matches against the database were assumed to still be alive, and vice versa. Length of survival was defined as the period of time from the date when patients were recruited into the study until the date of death as recorded by the Malaysian National Registry for patients who have passed away, or until the date when the Malaysian National Registry was last queried for patients assumed to still be alive. For all survival analyses in this study, only patients with at least two years of survival data were included ( $n = 367$ ). Unadjusted Kaplan–Meier analyses and log-rank tests were conducted using the survival package in R (v. 2.44) and plotted using the “ggsurvplot” function from the survminer R package (v. 0.4.4). Cox proportional hazard models were built using the “coxph” function from the survival package and plotted using the “ggforest” function from survminer.

**Box and whiskers plots.** All box and whiskers plots in the main and supplemental figures are constructed with boxes indicating 25th percentile, median and 75th percentile, and whiskers showing the maximum and minimum values within 1.5 times the inter-quartile range from the edge of the box, with outliers not shown.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Sequencing data for this study (WES, RNA-seq and sWGS bam files) are available on the European Genome-phenome Archive under the study accession number [EGAS00001004518](https://www.ebi.ac.uk/ena/browser/view/EGAS00001004518). Access to controlled patient data will require the approval of the MyBrCa Tumour Genomics Data Access Committee upon request to Soo-Hwang Teo at [genetics@cancerresearch.my](mailto:genetics@cancerresearch.my). Publicly available data sets that were included in this study are accessible as follows: TCGA<sup>10</sup> and METABRIC<sup>5</sup> data are available via the Genomic Data Commons Data Portal (<https://portal.gdc.cancer.gov/>) and cBioportal (<https://www.cbioportal.org/>)<sup>67</sup>, whereas data from Nik-Zainal et al.<sup>23</sup> are available from <ftp://ftp.sanger.ac.uk/pub/cancer/Nik-ZainalEtAl-560BreastGenomes> and data from Zhengyan Kan et al.<sup>19</sup> are available from <https://www.nature.com/articles/s41467-018-04129-4>. Other public databases that were accessed include Reactome ([www.reactome.org](http://www.reactome.org)), the Integrative Onco Genomics database ([www.intogen.org](http://www.intogen.org)) and gnomAD (gnomad.broadinstitute.org). The remaining data are available within the Article, Supplementary Information or from the authors upon request.

## Code availability

Code that was used to create the main figures in this study is available at [https://github.com/panjw0/MyBrCa\\_Genomics](https://github.com/panjw0/MyBrCa_Genomics).

Received: 8 April 2020; Accepted: 17 November 2020;

Published online: 22 December 2020

## References

1. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874 (2001).
2. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
3. Sparano, J. A. et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).
4. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
5. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
6. Rueda, O. M. et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399–404 (2019).
7. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
8. Lehmann, B. D. et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).
9. Jiang, Y. Z. et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell* **35**, 428–440.e5 (2019).
10. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
11. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
12. Banerji, S. et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
13. Rizvi, N. A. et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
14. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
15. Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
16. Bryant, H. E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913–917 (2005).
17. Pitt, J. J. et al. Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features. *Nat. Commun.* **9**, 4181 (2018).
18. Yap, Y. S. et al. Insights into breast cancer in the east vs the west: a review. *JAMA Oncol.* <https://doi.org/10.1001/jamaoncol.2019.0620> (2019).
19. Kan, Z. et al. Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. *Nat. Commun.* **9**, 1725 (2018).
20. Tan, M.-M. et al. A case-control study of breast cancer risk factors in 7,663 women in Malaysia. *PLoS ONE* **13**, e0203469 (2018).
21. Ho, W.-K. et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nat. Commun.* **11**, 3833 (2020).
22. Burstein, M. D. et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res.* **21**, 1688–1698 (2015).
23. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
24. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
25. Maguire, S. L. et al. SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J. Pathol.* **235**, 571–580 (2015).
26. Huang, K. & Lin et al. Pathogenic germline variants in 10,389 adult cancers. *Cell* **173**, 355–370.e14 (2018).
27. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
28. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
29. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
30. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
31. Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).
32. Ayers, M. et al. IFN- $\gamma$ -related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).
33. Auslander, N. et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat. Med.* **24**, 1545–1549 (2018).
34. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
35. Desmedt, C. et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.* **14**, 5158–5165 (2008).
36. Stanton, S. E. & Disis, M. L. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J. Immunother. Cancer* **4**, 59 (2016).
37. Roth, A. et al. PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
38. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).

39. Wolf, Y. et al. UVB-induced tumor heterogeneity diminishes immune response in melanoma. *Cell* **179**, 219–235.e21 (2019).
40. Letterio, J. J. & Roberts, A. B. Regulation of immune responses by TGF- $\beta$ . *Annu. Rev. Immunol.* **16**, 137–161 (1998).
41. Woo, S.-R. et al. STING-dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. *Immunity* **41**, 830–842 (2014).
42. Teschendorff, A. E. et al. Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC Cancer* **10**, 604 (2010).
43. Momozawa, Y. et al. Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. *Nat. Commun.* **9**, 1–7 (2018).
44. Sun, J. et al. Germline mutations in cancer susceptibility genes in a large series of unselected breast cancer patients. *Clin. Cancer Res.* **23**, 6113–6119 (2017).
45. Deng, M. et al. Prevalence and clinical outcomes of germline mutations in *BRCA1/2* and *PALB2* genes in 2769 unselected breast cancer patients in China. *Int. J. Cancer* **145**, 1517–1528 (2019).
46. Abubakar, M. et al. Breast cancer risk factors, survival and recurrence, and tumor molecular subtype: analysis of 3012 women from an indigenous Asian population. *Breast Cancer Res.* **20**, 114 (2018).
47. Gomez, S. L. et al. Breast cancer in Asian Americans in California, 1988–2013: increasing incidence trends and recent data on breast cancer subtypes. *Breast Cancer Res. Treat.* **164**, 139–147 (2017).
48. Telli, M. L. et al. Asian ethnicity and breast cancer subtypes: a study from the California Cancer Registry. *Breast Cancer Res. Treat.* **127**, 471–478 (2011).
49. Zhang, L. et al. Comparison of breast cancer risk factors among molecular subtypes: a case-only study. *Cancer Med.* **8**, 1882–1892 (2019).
50. Brouckaert, O. et al. Reproductive profiles and risk of breast cancer subtypes: a multi-center case-only study. *Breast Cancer Res.* **19**, 112–119 (2017).
51. Schulze, K. et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
52. Silwal-Pandit, L. et al. TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance. *Clin. Cancer Res.* **20**, 3569–3580 (2014).
53. Griffith, O. L. et al. The prognostic effects of somatic mutations in ER-positive breast cancer. *Nat. Commun.* **9**, 3416–3476 (2018).
54. Luen, S. J. et al. Association of somatic driver alterations with prognosis in postmenopausal, hormone receptor-positive, HER2-negative early breast cancer: a secondary analysis of the BIG 1-98 randomized clinical trial. *JAMA Oncol.* **4**, 1335–1343 (2018).
55. Basho, R. K. et al. Clinical outcomes based on multigene profiling in metastatic breast cancer patients. *Oncotarget* **7**, 76362–76373 (2016).
56. Meric-Bernstam, F. et al. Survival outcomes by TP53 mutation status in metastatic breast cancer. *JCO Precis. Oncol.* **2018**, PO.17.00245 (2018).
57. Loo, L. W. M. et al. MDM2, MDM2-C, and mutant p53 expression influence breast cancer survival in a multiethnic population. *Breast Cancer Res. Treat.* **174**, 257–269 (2019).
58. Chen, C.-H. et al. Disparity in tumor immune microenvironment of breast cancer and prognostic impact: Asian versus Western populations. *Oncologist* (2019) <https://doi.org/10.1634/theoncologist.2019-0123>.
59. Bhandari, V. et al. Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).
60. Ellis, L. et al. Racial and ethnic disparities in cancer survival: the contribution of tumor, sociodemographic, institutional, and neighborhood characteristics. *J. Clin. Oncol.* **36**, 25–33 (2018).
61. Trinh, Q.-D. et al. Cancer-specific mortality of Asian Americans diagnosed with cancer: a nationwide population-based assessment. *J. Natl. Cancer Inst.* **107**, 54 (2015).
62. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
63. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
64. Li, B. et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
65. 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.* **24**, gr.175141.114–2032 (2014).
66. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
67. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
68. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
69. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
70. Hundal, J. et al. pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).

## Acknowledgements

This project was funded by a research grant from the Newton-Ungku Omar Fund (MRC Ref: MR/P012442/1). Cancer Research Malaysia also receives charitable funding from the Scientex Foundation, Estée Lauder Companies, Yayasan Petronas, and Yayasan Sime Darby, which contributed to the funding of this study. This work (SFC, OMR, and CC) is also partly funded by Cancer Research UK (A16942). The authors would thank Dr. Tan Min Min for help with data curation, nurses, and staff who helped with sample collection, as well as Dr Saira Bahnu Mohamed Yousoof and all staff at the Subang Jaya Medical Centre Tissue Diagnostics laboratory for assistance with histopathological sample retrieval and processing. We also acknowledge the contribution of the Core Genomics Facility at the CRUK Cambridge Institute, where the sequencing work was conducted.

## Author contributions

J.W.P. and M.M.A.Z. co-led the data analysis and wrote the manuscript. P.S.N., M.Y.M., S.N.H., and C.H.Y. contributed to sample collection and processing and data collection, whereas B.S., O.M.R. and S.F.C. generated and collected data. S.J.S. interpreted results and guided the data analysis. P.R. provided histopathology expertise, and together with C.H.Y. collected clinical data. O.M.R., C.C., S.F.C. and S.H.T. designed experiments, interpreted results, and drafted the manuscript. The project was directed and co-supervised by C.C., S.F.C. and S.H.T., and were responsible for final editing.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-20173-5>.

**Correspondence** and requests for materials should be addressed to S.-F.C. or S.-H.T.

**Peer review information** *Nature Communications* thanks Yeon Hee Park and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020