Differential whole-genome doubling and homologous recombination deficiencies across breast cancer subtypes from the Taiwanese population

Chia-Hsin Wu1,16, Chia-Shan Hsieh2,16, Yo-Cheng Chang3,16, Chi-Cheng Huang4, Hsien-Tang Yeh5, Ming-Feng Hou6, Yuan-Chiang Chung7, Shih-Hsin Tu8, King-Jen Chang9, Amrita Chattopadhyay10, Liang-Chuan Lai11, Tzu-Pin Lu12, Yung-Hua Li1, Mong-Hsun Tsai13 & Eric Y. Chuang1,10,14,15

Whole-genome doubling (WGD) is an early macro-evolutionary event in tumorigenesis, involving the doubling of an entire chromosome complement. However, its impact on breast cancer subtypes remains unclear. Here, we performed a comprehensive and quantitative analysis of WGD and its influence on breast cancer subtypes in patients from Taiwan and consequently highlight the genomic association between WGD and homologous recombination deficiency (HRD). A higher manifestation of WGD was reported in triple-negative breast cancer, conferring high chromosomal instability (CIN), while HER2+ tumors exhibited early WGD events, with widely varied CIN levels, compared to luminal-type tumors. An association of higher activity of de novo indel signature 2 with WGD and HRD in Taiwanese breast cancer patients was reported. A control test between WGD and pseudo non-WGD samples was further employed to support this finding. The study provides a better comprehension of tumorigenesis in breast cancer subtypes, thus assisting in personalized treatment.
Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer-related death worldwide. In recent years, there has been a sharp rise in breast cancer incidence in the Asia-Pacific region, which has brought increased visibility of Asians as a distinct breast cancer patient population comprising 22.8% of all cases globally. Notably, the incidence of breast cancer in the Asia-Pacific population exhibits remarkable variation (range of age-standardized rate: 32.8–59.8 per 100,000 population). Taken together, thorough genomic characterization of SNVs and CNAs allows us to hypothesize that large-scale chromosomal abnormalities, which dominate the genomic landscape of cancer. In this study, the previous biological finding of breast cancer subtypes caused by different processes revealed indel signatures to be a better and more reliable predictor of WGD-linked HRD phenotype than the substitution signatures. This study aspires to shed light on the understanding of tumorigenesis by revealing the molecular basis for breast cancer subtypes, with the goal of enhancing personalized treatment strategies.

Results

Patient characteristics. A total of 116 patients were diagnosed between 27 and 85 years (median 53.0 years) of age, frequently at early stages (80% at stage I-II) and with an intermediate histologic grade (53%). The majority of BCTW patients (60%) were classified with luminal A subtype (defined as HR+/HER2-), while the remaining patients were classified evenly among HER2+ (15%), TNBC (13%) and luminal B (HR+/HER2+; 12%) subtypes. Table 1 summarizes the clinical characteristics of all patients included in this study.

Mutational landscape of BCTW samples. A total of 13,174 somatic SNVs (median 53.5), including 5375 nonsynonymous variants (median 26) and 1374 somatic small insertions/deletions (indels; median 5), were identified from the patients in this study. The median tumor mutation burden (TMB) was 1 mutation/Mb; however, a few breast cancer patients demonstrated hyper-mutation with a TMB of more than 10 mutations per Mb. (Fig. 1a). Interestingly, TMB was reported to be positively associated with the mutational status of gene PIK3CA (P = 0.034) using a one-tailed Mann–Whitney U test.

To detect the significantly mutated genes (SMGs) in BCTWs, four robust approaches (Supplementary Fig. 1) were applied to genes with expression in human breast tissue (genes with no expression were excluded). The results revealed the following: (1) the frequency-based approach (dNdScv) identified the gene COMP as a BCTW-specific SMG that has not been included previously in the Cancer Gene Census-v86 (Supplementary Note 1). Another four genes, TP53, PIK3CA, GATA3, and AKTI
(previously reported), were also reported as SMGs under positive selection, with a higher prevalence of positively accumulated mutations than expected\(^\text{14}\). Somatic mutations in \textit{PIK3CA} were predominantly missense, while \textit{TP53} exhibited a variety of alterations (Fig. 1a). (2) To complement the aforementioned frequency-based approach, the 20/20 rule-based approach\(^\text{15}\), which evaluates the proportion of missense mutations and loss-of-function mutations in the gene of interest, identified 10 additional SMGs, including \textit{SF3B1}, \textit{NF1}, \textit{MAP3K1}, \textit{ARID1A}, \textit{NUMA1}, \textit{CDH1}, \textit{RBI}, \textit{KMT2A}, \textit{PTEN}, and a BCTW-specific SMG, \textit{ERN1}. (3) Assessment of mutations with high functional impact (OncodriveFML, a domain-based approach)\(^\text{16}\) revealed...
Fig. 1 Mutational landscape of somatic alteration and single base substitution mutational signatures (SBSs) in 116 breast cancer samples. a Rows represent significantly mutated genes (SMGs), and columns represent individual tumors. Samples are arranged to emphasize mutural exclusivity among alterations. SMGs are ordered according to the frequency of nonconsensus single-nucleotide variations/indels. The stacked bar plot depicts the tumor mutation burden (TMB; mutations/covered bases; y-axis) for individual tumors (x-axis). Key clinical features are annotated for each tumor. Clinical characteristics and mutation types are indicated with color. b The bar plot with the 95% confidence interval indicates mutational frequencies of BCTWs, compared with those from The Cancer Genome Atlas benchmark cohorts (right; BCTW, Taiwanese; CCSN, Caucasian; AFRAM, African American; ASAM, Asian American). c Heatmap of the cosine similarity results for the three de novo SBSs of the Taiwanese population (y-axis), coded by color. In the scale bar, the cosine similarity (range 0–1) represents the extent of similarity to a particular signature of COSMIC (x-axis). Among the 30 COSMIC SBSs, the APOBEC- and age-related signatures were the most similar mutational signatures detected in the Taiwanese population (dark red), while one signature was dissimilar to any COSMIC mutational signatures and thus is considered unknown.

the well-known SMGs GATA3 and TP53. (4) Parsing locally clustered 'hotspot mutations' (Onco driveCLUST, the feature-based approach) identified 1 BCTW-specific SMG, PIGT, occurring in 2.6% of BCTW samples.

The identified SMGs were then compared across breast cancer subtypes in BCTWs. TP53 remained largely unaltered in the HR +/HER- subtype (14.1%), but TP53 mutations were enriched in both HR +/HER2 + and HER2 + cancers (23.1%, 33.3%) (Fig. 1a), with the highest frequency of TP53 mutations (50.0%) in TNBC tumors. The prevalence of PIK3CA mutations was similar between the HR +/HER2- (17.2%) and TNBC (21.4%) subtypes and lower in HER2 + tumors (6.7%) than in HR+/HER2 + tumors (15.4%).

Comparative analysis across populations was briefly performed using BCTWs and The Cancer Genome Atlas (TCGA) benchmark cohorts (Fig. 1b). The prevalence of GATA3 mutation in BCTWs was the same as that in Asian Americans. Genes KMT2A and RB1 displayed similar mutation prevalence for both the BCTW and TCGA cohorts, except for Caucasians. Interestingly, two of the most commonly mutated genes in BCTWs, TP53 and PIK3CA (20.7 and 15.5%, respectively), were present at lower frequencies in BCTWs than in TCGA cohorts (26.8 and 35.4% in Caucasians; 40.4 and 18.4% in African Americans; 50.0 and 37.5% in Asian Americans). Higher mutation frequency in CDH1 and MAP3K1 were also observed in TCGA cohorts as compared to BCTWs. In contrast, the BCTW cohort had higher mutation rates in AKT1, NFI, and SF3B1. Our analysis also identified somatic mutations in genes that turned out to be only locally prevalent (ERN1, PIGT, and COMP) (Fig. 1b).

To characterize the mutational processes that contributed to the BCTW cohort, a mutational signature analysis was performed. Among the 30 identified pan-cancer single base substitution (SBS) mutational signatures released in the Catalogue of Somatic Mutations in Cancer (COSMIC), there are 13 SBS signatures related to breast cancer. A cosine similarity analysis, to compare the de novo SBS mutational signatures (de novo SBS) derived from BCTW samples with those from COSMIC, identified two signatures that were highly similar to COSMIC SBS signatures (COSMIC SBS; Fig. 1c). De novo SBS1 was highly similar to an age-dependent COSMIC SBS (cosine similarity 0.933), while de novo SBS2 displayed high similarity to COSMIC SBS2 and COSMIC SBS13 (cosine similarity 0.806 and 0.834, respectively), which might be due to overactivity of the APOBEC family of cytidine deaminases. However, de novo SBS3 did not show much similarity to any of the COSMIC SBS signatures (all cosine similarity <0.55).

WGD and CIN in breast cancer subtypes. In total, 19.0% of BCTW patients were found with tumors that underwent WGD. The rate of WGD exhibited remarkable variation across breast cancer subtypes, affecting 50.0% of TNBC samples versus only 14.1% of HR+/HER2- tumors, 23.1% of HR+/HER2 + and 20.0% of HER2 + tumors (Fig. 2a). The TMB of WGD-positive tumors was found to be higher than those of WGD-negative tumors (median TMB of 2.018 and 0.835, respectively; \( P = 1.484 \times 10^{-6} \); one-tailed Mann–Whitney U test). Next, compared to non-WGD samples, the relatively frequent alterations of cancer genes were found in WGD samples, where genes MYC, EIF4EBP1, and FGFR1 displayed amplifications; DUSP4, LEPROT1, NRG1, and WRN deletions; and MUC16 demonstrated mutations (\( P < 0.05 \); odds ratio >1; logistic regression). HR+/HER2 + tumors, further, exhibited a greater degree of CIN than HR+/HER- tumors, while the highest level of CIN was observed in TNBC samples, consistent with their higher WGD frequency (Fig. 2b). Additionally, WGD was found to occur early, before subclonal diversification but after the acquisition of driver alterations that induce CIN. The timing of WGD across subtypes was estimated, based on the fraction of mutations occurring post-WGD compared to pre- and post-WGD combined. On average, TNBC samples exhibited complex clonal WGD events due to a wider variance compared to HR+/HER- and HR+/HER2 + tumors (Fig. 2c), whereas HER2 + tumors revealed relatively early WGD events within subtypes.

Clonality and timing of driver events within subtypes. Alterations in well-known cancer genes or SMGs were frequently clonal and occurred before WGD, which suggests their involvement in tumorigenesis. In the TNBC subtype, these alterations included mutations in CDK6; amplification of MYC, CCND1, and KIT; and deletion of DUSP4, LEPROT1, NRG1, WRN, MXI1, PIK3R1, RB1, and STK11 (Fig. 3a, b). Other genes that were subject to subclonal involvement, suggesting late alterations in tumor progression, included ATM and 22q arm (deletion). In the HER2 + subtype, the mutation in APC and amplification of CCNE1, MDM2, ERBB2, AKT1, EGF, HSF3B, and CDK12 were early clonal alterations (Supplementary Fig. 2a, b); while deletion of CDKN2A and CDKN2B, along with 9p deletion, were late events. In the HR+/HER2 + subtype, the mutation in TGFBR2, ACVR1B, and KIT, as well as amplification of EGF, FGFR1, ERBB2, EIF4EBP1, and CDK12, occurred as clonal alterations before WGD (Supplementary Fig. 3a, b). Notably, deletion of DUSP4, LEPROT1, NRG1, and WRN within this subtype were primarily subclonal and occurred late. In the HR+/HER- subtype, the early amplification in ATM, AKT1/2, PIK3CA, and STK11; amplification of MYC, CCND1, CDK6, MDM2, AKT2/3, EIF4EBP1, FGFR1, and HRAS; and deletion of RB1, as well as pre-WGD deletion of 17p, 18q, 19p, and 22q, were also clonal (Supplementary Fig. 4a, b). Post-WGD mutations were found to occur in APC and NOTCH1. Interestingly, driver mutations in TP53 were predominantly clonal, occurring before WGD in the HR+/HER2 + subtype, but after WGD in the HER2 + subtype. Similarly, other driver events in SMGs, including mutation in PIK3CA in the HER2 + subtype, while predominantly subclonal, occurred after WGD, which might result in reduced treatment efficacy.
The mutational processes that shaped tumor progression were further explored for breast cancer subtypes. In the TNBC subtype, a large decrease in the age-related signature from the pre-WGD to the post-WGD period/stage was accompanied by an increase in the APOBEC cytidine deaminase activity (Fig. 3c). A predominance of the rate of clonal and subclonal mutations in the APOBEC signature is indicative of a long period of mutagenesis latency in the HER2\(^+\) subtype (Supplementary Fig. 2c). A similar phenomenon was observed for the APOBEC family in the HR\(^+\)/HER2\(-\) subtype, responsible for shaping the developmental trajectory of these tumors (Supplementary Fig. 3c). HR\(^+\)/HER2\(-\) samples exhibited dynamic equilibrium in age and APOBEC signatures, suggesting that alternative mutational processes might dominate tumor behavior later in the tumor lifetime (Supplementary Fig. 4c).

**WGD with HRD.** To quantitatively evaluate the extent of HRD in the WGD samples, the HRD score was first calculated by combining three genomic scar scores\(^{20}\). An association test found patients with WGD to be characterized by relatively high HRD scores (\(P = 4.8 \times 10^{-8}\); one-tailed Mann–Whitney \(U\) test; Fig. 4a). In each subtype, WGD-positive tumors had higher HRD scores compared to those of the WGD-negative tumors, but no difference was observed for WGD-positive tumors among subtypes (Fig. 4b). Given that 13.2% of BCTW patients exhibited HRD (score \(\geq 42\)), the association of prevalence of the HRD phenotype with WGD status was further assessed in the subtypes. In the TNBC cohort, 35.7% of patients frequently harbored HRD that was simultaneously accompanied with WGD in 28.6% of cases (Fig. 4c). Similarly, for the HR\(^+\)/HER2\(-\) and HER2\(^+\) cohort, there were more HRD cases with WGD than without WGD, in contrast to HR\(^+\)/HER2\(+\) tumors, where no difference was observed.

**WGD with alternative DSB repair processes.** To investigate whether the alternative DSB repair processes that operate in HRD cancers are associated with resulting mutational patterns, we applied the same SBS signature extraction approach (non-negative matrix factorization) to categorize indel mutational signatures\(^{21}\). De novo indel signature 2 (de novo ID2) was highly
similar to COSMIC indel mutational signatures six and eight (cosine similarity 0.85 and 0.83, respectively; Fig. 5a, b). Deletion patterns in these two COSMIC signatures have been associated with the characterization of DSB repair by two distinct forms of nonhomologous end-joining activity21. The COSMIC indel mutational signature six displayed longer stretches of overlapping microhomology at deletion boundaries that stem from HRD, whereas the COSMIC indel mutational signature 8 revealed relatively shorter microhomology at deletion boundaries. Interestingly, the pattern of the de novo ID2 was also similar to the mixtures of COSMIC indel mutational signatures six and eight (Fig. 5b). Also, patients with HRD were linked to a high contribution22 of de novo ID2 ($P = 1.633 \times 10^{-4}$; one-tailed Mann–Whitney U test; Fig. 5c). Furthermore, we found that the contribution of de novo ID2 in WGD-positive tumors was much higher than in those without WGD ($P = 2.16 \times 10^{-2}$; one-tailed Mann–Whitney U test; Fig. 5d). We further performed a control test to rule out the possibility that a higher contribution of de novo ID2 is caused by disproportionate numbers of indels in the WGD samples. Our results displayed no statistical significance in 978 out of 1000 times implying that the high proportion of indels in the WGD samples had no role to play in the contribution of de novo ID2 in WGD.

**Discussion**

WGD potentially aggravates CIN and accelerates cancer genome evolution, which has previously been shown to be associated with poor prognosis and drug resistance3. In this study, it can be hypothesized that WGD might explain distinct genomic
complexity across breast cancer subtypes. The prevalence of WGD displayed remarkable variability among subtypes, with its effects mostly observed in TNBC tumors. Moreover, eight somatic alterations frequently altered in the WGD samples were found across breast cancer subtypes. On average, TNBC exhibited more complex clonal WGD events when compared to other subtypes, as did HER2+ tumors with relatively early WGD events. Furthermore, TNBC samples had high levels of CIN, and CIN levels varied widely in HER2+ tumors. These findings could be the initial indication that WGD primarily affects TNBC and HER2+ tumor progression, which might change the practice of clinical care, and provide insight into breast cancer biology.

Investigation of early somatic events might assist in assessing their evolutionary impact on breast cancer in terms of both the WGD and mutational processes. The timing of cancer driver events determines their involvement in tumor initiation or progression. The clonality of drivers may facilitate therapeutic decision making, as the accumulation of subclonal alterations in a proportion of cells may result in reduced drug efficacy. However, the underlying history of deletions is complex and uncertain, hence not identifiable except with additional strong assumptions about the CN paths that are allowable. Therefore, we assume that the current copy number paths from sequencing data can be used to uncover and determine the timing of mutation and copy number alterations, based on a prior study.23. FGFR1 amplification, an early clonal somatic event in hormone receptor-positive tumors, conferred antiestrogen resistance to ER+ breast cancer24. Treatment regimens by prioritizing on drug targets with combinations of ER and FGFR antagonists might improve the robust and uniform treatment response by inhibiting their binding to DNA in early ER+ tumor progression. Similarly, the amplification of EIF4EBP1, a downstream effector of mTOR, has been
Fig. 5 Indel mutational signatures (IDs) with homologous recombination deficiency (HRD) and whole-genome doubling (WGD). 

a Heatmap of the cosine similarity results for the de novo IDs (y-axis), coded by color. In the scale bar, the cosine similarity (range 0–1) represents the extent of similarity to a particular signature of COSMIC (x-axis). Among the 17 COSMIC IDs, the signatures associated with the double-strand repair by nonhomologous end-joining were the most similar mutational signatures detected in the Taiwanese population (dark red), while other signatures are of unknown cause in COSMIC or dissimilar.

b The mutational spectrum of de novo signature ID2 and COSMIC signatures ID6 and ID8. 

c Box plots indicate the contribution of de novo signature ID2 (y-axis) across HRD and non-HRD tumors (x-axis). 

d Box plots indicate the contribution of de novo signature ID2 (y-axis) across WGD and non-WGD tumors (x-axis). One-tailed Mann–Whitney U tests were performed to calculate the p-values for the comparison of medians in c and d. Overall, 116 breast cancer samples were used to perform analyses in a–d.
reported to be associated with a poor response to endocrine treatment in the hormone receptor-positive cohort\(^9\). A combination of ER and mTOR signaling targeted therapies might suppress hormone receptor signaling for breast tumorigenesis. In TNBC, deletions of **DUSP4**, **LEPROTL1**, **NRG1**, and **WRN** occurred before WGD, in contrast to **HR** +/+ **HER2** +/+, in which they were post-WGD. It is suggested that WGD might lead to unique therapeutic decisions and serve as a precursor of late events. These alterations could inform the design of adjuvant trials in patients with WGD and further clinical and functional investigation. Furthermore, most SMGs were altered before WGD and were predominantly clonal across subtypes but were observed post-WGD in **HER2** +/+. Notably, **CDK12** exhibited similar gene amplification-driven proteogenomic patterns to **ERBB2**\(^2-6\), both were enriched in **HER2** +/+ and **HR** +/+ **HER2** +/+ as clonal alterations before WGD, which should potentially be therapeutically exploited. In the mutational processes of TNBC, a large decrease in the proportion of samples with a common age-related signature was accompanied by an increase in APOBEC-mediated signatures during tumor development. Targeting the activity of APOBEC might help limit subclone diversification. These findings may have important implications for understanding the specific molecular pathogenesis and therapeutic control in breast cancer.

Although our cohort was not enriched in bi-allelic BRCA1/2 alterations, the HRD scoring measurement has the ability to identify additional patients, beyond BRCA deficiency, with high sensitivity to platinum agents or PARP inhibitors\(^27\). We quantitatively evaluated the extent of HRD in the WGD samples. Our observation showed that WGD-positive tumors were associated with HRD for all subtypes, and thus all might benefit from HRD-related treatment. In TNBC subtypes, patients frequently harboring WGD displayed the mutational characterization of the HRD phenotype. Therefore, future clinical trials could incorporate the WGD status for baseline treatment stratification, as several studies have suggested that some (wild-type BRCA) tumors also respond to platinum agents or PARP inhibitors, even in sporadic TNBCs\(^28-31\). It is hoped that such effort will help to diminish therapeutic vulnerabilities in breast cancer outcomes in the future.

This study provides accumulating evidence that the presence of a mutational signature of HRD can also be employed for optimal treatment decisions of cancer patients\(^32\). As COSMIC substitution signature three does not provide a precise assessment of HRD\(^33,34\), we focused on the contribution of deletion patterns attributed to HRD, which exhibit the characteristics of non-homologous end-joining\(^6,7\). Given that de novo signature ID2 was identified as a mixture of COSMIC indel mutational signatures six and eight, it suggests that two distinct forms of non-homologous end-joining activity are operative in patients with HRD\(^21\). Furthermore, we observed that higher contribution of de novo signature ID2 is associated with WGD in BCTWs. However, the higher contribution of de novo ID2 could originate from the disproportionate numbers of indels in the WGD samples. To rule out this possibility, a control test between the WGD and pseudo non-WGD samples was employed and supported our finding. It was further observed that signature ID2 is a predictor of the interplay between HRD and WGD. Taken together, this finding might provide avenues of assessment of both WGD and HRD phenotypes for the rendering of therapeutic decisions for individual cancer patients.

In this study, we found that WGD might exhibit distinct genomic complexity across breast cancer subtypes, and highlights the association between HRD and WGD through an extensive analysis using WES data from BCTW patients. The impact of WGD and the timing of genome abnormalities across subtypes of breast cancer, which have not been characterized in the Asian cohort before, are explored in this study. In the BCTW cohort, we observed the subtype heterogeneity of WGD in Asian populations. This can be further validated in other populations, in the future. Our work has outlined the subtype specificity of WGD and **CIN** in BCTW patients and provided insights into the genomic basis of de novo ID2 linked to WGD and HRD in breast cancer. Future studies with suitable cell lines, which contain both the diploid and polyploid subclones with HRD phenotype from human breast cancers in four subtypes, would be needed to support our hypothesis. These suitable cell lines would be isolated at different passages and sequenced to validate the findings in this study. However, future cell line studies would be out of the scope of the current study. Such molecular-level findings may shed light on the understanding of tumorigenesis in breast cancer, and help develop personalized treatment strategies in an important step toward the complete cure of the disease.

**Methods**

**Study population and specimens.** Subjects recruited for this study included a subset of clinically diagnosed breast cancer patients in Taiwan who underwent surgical resection at four different hospitals in Taiwan (Lotung Poh-Ai Hospital, Cathay General Hospital, Kaohsiung Medical University Hospital, and Cheng Ching Hospital). The study protocols were reviewed and approved by the Institutional Review Boards of these hospitals, and informed consent was obtained from all patients for conducting WES and corresponding analyses. Fresh-frozen samples were collected from tumor and matched adjacent normal tissues of 104 patients with breast cancer. Additional paired normal-tumor samples that were formalin-fixed and paraffin-embedded were available from a set of 12 breast cancer patients. Breast cancer subtypes were identified using immunohistochemistry and fluorescence in situ hybridization. To obtain BCTW-specific genetic features, the WES data was compared with sequences obtained from the TCGA database. TCGA cohorts without mutation or race information were excluded from the downstream comparisons via postfiltering strategies based on the TCGA MC3 project, leading to a total of 723 cases including Caucasians (526 cases, 72.8%), African Americans (141 cases, 19.5%), and Asian Americans (56 cases, 7.7%); linked clinical data for the TCGA cohorts was recorded in the eBioPortal for Cancer Genomics\(^35,36\).

**Exome capture, library construction, and sequencing.** For the generation of standard exome capture libraries, we used the Agilent SureSelect XT Reagent Kit protocol for an Illumina HiSeq paired-end sequencing library (catalog#G9061A). In all cases, the SureSelect XT Human All Exon Version 6 (60 Mb) probe set was used. We used 1000 ng genomic DNA to construct each library. Each adapter-ligated sample was purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) and analyzed on a Bioanalyzer DNA1000 chip. A total of 750 ng of the hybridized sample was prepared for hybridization with the capture baits, and the sample was hybridized for 90 min at 65°C, captured with Dynabeads MyOne Streptavidin T1 (Invitrogen), washed, and amplified using Agencourt beads. We used the Agilent protocol to add index tags by post-hybridization amplification. Finally, all samples were sequenced on an Illumina HiSeq4000 instrument using the 150PE protocol.

**Sequencing data processing.** To compare with TCGA cohorts, the analytic pipelines were run with similar parameter values as that of TCGA, based on the GDC data user guide organized by its institutional research network. The processed read pairs were mapped to the human reference genome (hg19) using BWA-MEM (v.0.7.15)\(^37\). The Picard module (v.2.6.0) was utilized to sort BAM files containing the sequence alignment data in binary format. Duplicate reads were marked for exclusion in a subsequent analysis using the MarkDuplicates tool in Picard. The resulting BAM files were processed with the Genome Analysis Toolkit Best Practices workflow (v.3.7.0) to correct mapping and sequencing errors\(^38\). First, indel realignment was performed around the Mills and 1000 Genomes gold standard INDELS to improve the alignment accuracy. Then base quality score recalibration was conducted to assign an accurate confidence score to each base using known variants in dbsNP138 and the Mills and 1000 Genomes gold standard INDELS\(^39\). To improve downstream variant detection, we added 100 bp-interval padding to confirm all reads within and outside the targeted region.

**Variant calling.** Following the TCGA DNA-Seq analysis pipeline, somatic SNV and indel calling was conducted using MuTect (v.3.7.0) from both the tumor and matched normal whole-exome samples in targeted exons\(^40\), utilizing the COSMIC and dbSNP138 as reference sites of known somatic and germline mutations\(^4-5\). To confirm all calls within the targeted region, the initial SNV indel calls were extracted using SelectVariants. The standard filter settings were the same as the TCGA MC3 project\(^30\). To reduce the false-positive calls, we constructed a Panel-of-Normals as filters of
contamination and miscalled germline mutations using MuTect2 on all the normal sample genomes. We removed those mutations from our Panel-of-Normals if they were present in more than one normal sample. Furthermore, the likely 8-oxoguanine contamination and miscalled germline mutations using MuTect2 on all the normal genomes were excluded to reduce the artefactual sensitivity of the algorithm. To establish the associations between CNAs and somatic processes, such as ultraviolet light exposure, carcinogens, and aging22. To characterize the SBS mutational signatures originating from the accumulation of histonic mutagenic activity in BCTW, we utilized non-negative matrix factorization to extract de novo SMGs. The mutational signatures present in these tumors were utilized to calculate the HRD score using the R package scarHRD61. A predefined HRD threshold of 42 was used to identify high-copy-number regions (defined as regions that encompass a wide range of chromosome-level abnormalities. CIN is defined as the percentage of the genome in length affected by CNAs45 and is given by:

\[
\text{CIN} = \frac{\sum_{i=1}^{L} \alpha_i}{L}
\]

where \( L \) is the total length of the genome and \( u_i \) represents the altered length in \( i \).

Mutational signature analysis. The final portrait of mutations was determined by the duration of exposure to each mutational process in patients. Mutational signatures were deciphered from the substitution context defined by 96 mutation types, each represented by 12 mutation rates as triplets of frequencies. A combination of mutation rates for each signature was determined to create a consensus list of signatures (1) \( d\)NSV (0.0.0.9), a maximum-likelihood \( dN/dS \) method, quantifies selection of missense, nonsense, and splice mutations, at the level of individual genes, groups of genes, or at the whole-genome level in cancer and somatic evolution15. \( d\)NSV was run using default parameters. Genes with a \( q \) value <0.05 were considered as SMGs. (2) OncodriveCLUST (v.1.0.0), a method used to identify oncogenes with a bias towards mutation clustering for genes with nonsynonymous mutation more than expected of synonymous mutations within the protein sequence47. OncodriveCLUST was run using default parameters. Genes with a \( q \) value <0.05 were considered as SMGs. (3) Read depth-based SNV- and indel-mutation classification was considered as a late event after WGD. The timing of WGD was estimated using methodology adapted from the previous work23,26,27. Clonal mutations in regions of overlapping copy number events were timed as early or late with respect to WGD. Mutations were classified as early if the mutation copy number was >1 and the MCN ≥ 2. Late mutations were called with the mutation copy number ≥ 1 and MCN ≥ 2. Clonal mutations that could not be timed were considered as untimed, while any subclonal mutations were considered as late events after WGD. The timing of WGD was estimated based on the fraction of mutations occurring post-WGD compared to pre- and post-WGD combined23.

Clonality of somatic alterations and timing of WGD. To infer the clonality, the estimates of purity and allele-specific copy number from FACETS were used as input for the Absolute module to obtain the cancer cell fraction for all somatic mutations with corresponding variant allele frequencies in all tumor samples. Somatic alterations were further timed relative to WGD using methodology adapted from the previous work23,26,27. Clonal mutations in regions of overlapping copy number events were timed as early or late with respect to WGD. Mutations were classified as early if the mutation copy number was >1 and the MCN ≥ 2. Late mutations were called with the mutation copy number ≥ 1 and MCN ≥ 2. Clonal mutations that could not be timed were considered as untimed, while any subclonal mutations were considered as late events after WGD. The timing of WGD was estimated based on the fraction of mutations occurring post-WGD compared to pre- and post-WGD combined23.

FACETS was also utilized to estimate the cancer cell fraction to distinguish the clonal and subclonal CNAs. Clonal arm-level amplification was timed as early and late with respect to WGD by an average ratio of mutation copy numbers according to the following over Amplification analysis using the R package scarHRD61. A predefined HRD threshold of 42 was used to identify high-copy-number regions (defined as regions that encompass a wide range of chromosome-level abnormalities. CIN is defined as the percentage of the genome in length affected by CNAs45 and is given by:

\[
\text{CIN} = \frac{\sum_{i=1}^{L} \alpha_i}{L}
\]

where \( L \) is the total length of the genome and \( u_i \) represents the altered length in \( i \).
status of SMGs and mutation burden. Logistic regression was performed to identify relatively frequent cancer gene alterations in the WGD samples. All statistical analyses were performed using R version 3.5.1. Sample sizes are included in each figure legend. The allelic ratio plots and sequencing quality metric for assessing WGD analysis are shown in Supplementary Figs. 5–26 and Supplementary Data 2. The bootstrapping data for 95% confidence interval of mutational frequencies in each population (Fig. 1b) are shown in Supplementary Data 3.

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

Data availability
Raw whole-exome sequencing data used in this study can be found in the NCBI database under the BioProject accession PRJNA729775. The molecular and clinical data of breast cancer patients from the TCGA cohort are available in the following repositories: TCGA BRCA: https://portal.gdc.cancer.gov/ (MiTecl2 Mutation Annotation Format file) and cBioPortal: https://www.cbioportal.org/ (clinical data).

Code availability
All of the tools used in this study are publicly available. Statistical and other analyses were performed using software R (version 3.5.1) and described in the method section.

Received: 23 August 2020; Accepted: 5 August 2021;
Published online: 09 September 2021

References
1. Cancer Today (powered by GLOBOCAN 2018). Vol. 2019 (Available at: https://gco.iarc.fr/today/home (Accessed: 17th January 2019).
2. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 490, 61–70 (2012).
3. Dewhurst, S. M. et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov. 4, 175–185 (2014).
4. Bielack, S. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat. Genet. 50, 1189–1195 (2018).
5. Davoli, T. & de Lange, T. The causes and consequences of polyplody in normal development and cancer. Annu Rev. Cell Dev. Biol. 27, 585–610 (2011).
6. Ceccaldi, R., Rondinelli, B. & D’Andrea, A. D. Repair pathway choices and consequences at the double-strand break. Trends Cell Biol. 26, 52–64 (2016).
7. Sfeir, A. & Symington, L. S. Microhomology-mediated end joining: a back-up repair pathway for genome maintenance. Nature. 474, 338–345 (2011).
8. Lord, C. J. & Ashworth, A. PARP inhibitors: Synthetic lethality in the clinic. Cell. 144, 27–40 (2011).
9. Riaz, N. et al. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. Nat. Commun. 8, 857 (2017).
10. Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. Science. 357, eaan2507 (2017).
11. Futreal, P. A. et al. A census of human cancer genes. Nat. Rev. Cancer. 4, 177–183 (2004).
12. Lord, C. J. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer. 18, 696–705 (2018).
13. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. Cell 171, 1029–1041.e21 (2017).
14. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. Proc. Natl Acad. Sci. USA 113, 14330–14335 (2016).
15. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 17, 128 (2016).
16. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 29, 2238–2243 (2013).
17. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45, D777–D783 (2017).
18. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature. 500, 415–421 (2013).
19. Telli, M. L. et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with favorable prognosis. Cancer Res. 76, 3764–3773 (2016).
20. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature. 578, 94–101 (2020).
21. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. Cell 173, 246–259 (2018).
22. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. N. Engl. J. Med. 376, 2109–2121 (2017).
23. Formisano, L. et al. Association of FGFR1 with ERalpha Maintains Ligand-Independent ER Transcription and Mediates Resistance to Estrogen Deprivation in ER(+) Breast Cancer. Clin. Cancer Res. 23, 6138–6150 (2017).
24. Velteborg, M. A. et al. Genomic patterns resembling BRCA1+ and BRCA2-mutated breast cancers predict benefit of intensified carboplatin-based chemotherapy. Breast Cancer Res. 16, R47 (2014).
25. Alli, E., Sharma, V. B., Sunderesakumar, P. & Ford, J. M. Defective repair of oxidative DNA damage in triple-negative breast cancer confers sensitivity to inhibition of poly(ADP-ribose) polymerase. Cancer Res. 69, 3589–3596 (2009).
26. Hastak, K., Alli, E. & Ford, J. M. Synergistic chemosensitivity of triple-negative breast cancer cell lines to poly(ADP-Ribose) polymerase inhibition, gemcitabine, and cisplatin. Cancer Res. 70, 7970–7980 (2010).
27. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. 23, 517–525 (2017).
28. Patch, A. et al. Whole-genome characterization of chemoresistant ovarian cancer. Nature. 521, 489–494 (2015).
29. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 534, 47–54 (2016).
30. Gao, J. et al. Integrative analysis of complex cancer genomes and clinical profiles using the cBioPortal. Sci. Signal. 6, p1 (2013).
31. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 6, 271–281.e7 (2018).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25, 1754–1760 (2009).
33. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
34. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 30, 308–311 (2002).
35. Gbulsksis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
36. Costello, M. et al. Discovery and characterization of artificial mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 41, e67 (2013).
37. McLaren, W. et al. The ensemble variant effect predictor. Genome Biol. 17, 12 (2016).
38. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 536, 285–291 (2016).
39. Lin, J. C., Fan, C. T., Liao, C. C. & Chen, Y. S. Taiwan Biobank: making cross-database convergence possible in the Big Data era. Gigascience. 7, 1–4 (2018).
40. Vogelstein, B. et al. Cancer genome landscapes. Science. 339, 1546–1558 (2013).
41. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 173, 371–385.e18 (2018).
42. Ghoujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 11, 367 (2010).
43. Blozik, F., Janssens, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 10, 33 (2018).
44. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics. 20, 685 (2019).
45. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 44, e131 (2016).
46. Fattmeel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41 (2011).
52. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. Nat. Methods. 15, 531–534 (2018).
53. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 30, 413–421 (2012).
54. Amin-Mansour, A. et al. Genomic evolutionary patterns of leiomyosarcoma and liposarcoma. Clin. Cancer Res. 25, 5135–5142 (2019).
55. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat. Commun. 7, 11479 (2016).
56. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci. Transl. Med. 7, 283ra54 (2015).
57. Jolly, C. & Loo, Van P. Timing somatic events in the evolution of cancer. Genome Biol. 19, 95 (2018).
58. Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. Br. J. Cancer. 107, 1776–1782 (2012).
59. Birkbak, N. J. et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discov. 2, 366–375 (2012).
60. Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res. 72, 5454–5462 (2012).
61. Sztupinszki, Z. et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. NPJ Breast Cancer. 4, 16 (2018).

Acknowledgements
We would like to thank TCGA, COSMIC, Exome Aggregation Consortium, Taiwan Biobank, and The Human Protein Atlas for making their data publicly available. Finally, we thank all the participating individuals for their contribution to this study. This work was supported by the Center of Biotechnology, National Taiwan University, Taiwan [grant number GTZ300].

Author contributions
Y.-C.C. (Yo-Cheng Chang), M.-H.T. and E.Y.C. designed the study. Y.-H.L., M.-H.T. and L.-C.L. performed the experiments. Collection of specimens was coordinated by C.-C.H., H.-T.Y., M.-F.H., Y.-C.C. (Yuan-Chiang Chung), S.-H.T. and K.-J.C. Clinical data was organized by C.-C.H., H.-T.Y., Y.-H.L., T.-P.L. and C.-S.H. Data analyses were conducted by C.-H.W. and C.-S.H. C.-H.W. performed statistical analyses. C.-H.W., C.-S.H., A.C., M.-H.T. and E.Y.C. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02597-x.

Correspondence and requests for materials should be addressed to Mong-Hsun Tsai or Eric Y. Chuang.

Peer review information Communications Biology thanks Ruping Sun and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Brooke LaFlamme and Eve Rogers. Peer reviewer reports are available.

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.

© The Author(s) 2021