

Article

# Comprehensive Cohort Analysis of Mutational Spectrum in Early Onset Breast Cancer Patients

Mohit K. Midha <sup>1,2</sup>, Yu-Feng Huang <sup>1</sup>, Hsiao-Hsiang Yang <sup>3</sup>, Tan-Chi Fan <sup>4</sup>,  
Nai-Chuan Chang <sup>4</sup>, Tzu-Han Chen <sup>1</sup>, Yu-Tai Wang <sup>5</sup>, Wen-Hung Kuo <sup>6</sup>, King-Jen Chang <sup>6</sup>,  
Chen-Yang Shen <sup>7</sup>, Alice L. Yu <sup>4,8</sup>, Kuo-Ping Chiu <sup>1,2,9,\*</sup> and Chien-Jen Chen <sup>1</sup>

<sup>1</sup> Genomics Research Center, Academia Sinica, Taipei 11529, Taiwan; mohitkmidha@sinica.edu.tw (M.K.M.); yfh1202@sinica.edu.tw (Y.-F.H.); chentzuhan@gmail.com (T.-H.C.); chencj@gate.sinica.edu.tw (C.-J.C.)

<sup>2</sup> Institute of Biochemistry and Molecular Biology, National Yang-Ming University, Taipei 112, Taiwan

<sup>3</sup> Department of Medical Research, Hsinchu Mackay Memorial Hospital, Hsinchu 300, Taiwan; vul4790127@gmail.com

<sup>4</sup> Institute of Stem Cell and Translational Cancer Research, Chang Gung Memorial Hospital at Linkou and Chang Gung University, No. 5, Fu-Shin St., Kuei Shang, Taoyuan 333, Taiwan; tcfan@cgmh.org.tw (T.-C.F.); logome@gmail.com (N.-C.C.); a1yu@ucsd.edu (A.L.Y.)

<sup>5</sup> National Center for High-Performance Computing, Hsinchu Science Park, Hsinchu 300, Taiwan; yutaiwang@narlabs.org.tw

<sup>6</sup> Department of Surgery, National Taiwan University Hospital, Taipei 100, Taiwan; brcancer@gmail.com (W.-H.K.); kingjen@ntu.edu.tw (K.-J.C.)

<sup>7</sup> Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan; bmcys@ibms.sinica.edu.tw

<sup>8</sup> Department of Pediatrics, University of California in San Diego, San Diego, CA 92161, USA

<sup>9</sup> Department of Life Sciences, College of Life Sciences, National Taiwan University, Taipei 10617, Taiwan

\* Correspondence: chiukp@gate.sinica.edu.tw; Tel.: +886-2-2787-1257; Fax: +886-2-2789-9924

Received: 16 June 2020; Accepted: 24 July 2020; Published: 28 July 2020



**Abstract:** Early onset breast cancer (EOBC), diagnosed at age ~40 or younger, is associated with a poorer prognosis and higher mortality rate compared to breast cancer diagnosed at age 50 or older. EOBC poses a serious threat to public health and requires in-depth investigation. We studied a cohort comprising 90 Taiwanese female patients, aiming to unravel the underlying mechanisms of EOBC etiopathogenesis. Sequence data generated by whole-exome sequencing (WES) and whole-genome sequencing (WGS) from white blood cell (WBC)–tumor pairs were analyzed to identify somatic missense mutations, copy number variations (CNVs) and germline missense mutations. Similar to regular breast cancer, the key somatic mutation-susceptibility genes of EOBC include *TP53* (40% prevalence), *PIK3CA* (37%), *GATA3* (17%) and *KMT2C* (17%), which are frequently reported in breast cancer; however, the structural protein-coding genes *MUC17* (19%), *FLG* (16%) and *NEBL* (11%) show a significantly higher prevalence in EOBC. Furthermore, the top 2 genes harboring EOBC germline mutations, *MUC16* (19%) and *KRT18* (19%), encode structural proteins. Compared to conventional breast cancer, an unexpectedly higher number of EOBC susceptibility genes encode structural proteins. We suspect that mutations in structural proteins may increase physical permeability to environmental hormones and carcinogens and cause breast cancer to occur at a young age.

**Keywords:** early onset breast cancer (EOBC); missense mutations; nonsynonymous mutations; somatic mutations; germline mutations

## 1. Introduction

Early onset breast cancer (EOBC) diagnosed at age ~40 or younger is associated with a poorer prognosis, higher recurrence rate and higher mortality rate than breast cancer (BC) diagnosed at later

ages [1–3]. Overall, there are over one million (e.g., 1.38 million in 2008 alone) new breast cancer cases each year worldwide [4,5]. Among these cases, approximately 4%–14% are EOBC: ~6.6% in American women [6], ~12% in Taiwanese women [7] and over 13% in Singaporean women [8]. While late-onset breast cancer (LOBC) is becoming more manageable, EOBC has continued to increase steadily in many countries over the past few decades and demands intensive investigation [3,6,9–11]. Breast cancer encompasses a broad spectrum of heterogeneous malignancies resulting from polygenic susceptibility. Based on the pattern of gene expression and histopathologic staining [12–14], previous studies have classified breast cancer into several clinically relevant subtypes, each of which is more or less associated with epidemiological risk factors and clinical responses [11,15,16].

A broad spectrum of risk factors has been associated with breast cancer. These risk factors include genetic predisposition (e.g., familial history and physiological history), lifestyle (e.g., dietary, living and physical activity habits) and environmental factors (e.g., environmental carcinogens and hormones). These risk factors may vary significantly between ethnic groups and racial origins [4,11,17], making the attempt to elucidate the underlying etiopathologic causes a great challenge; thus, a broad range of analyses targeting as many potential risk factors as possible are strongly desired.

This study aims to obtain comprehensive insight into EOBC. We surveyed somatic mutations, copy number variations (CNVs) and germline mutations in 90 patients to reveal the etiopathologic factors involved in the tumorigenesis of EOBC. Through the comparison of missense mutations between white blood cell (WBC)–tumor pairs, we examined the association of these molecular alterations with biologic pathways and clinical subtypes. The results indicate that a significantly higher fraction of EOBC susceptibility genes encode extracellular structural proteins compared to breast cancer susceptibility genes as a whole.

## 2. Results

### 2.1. Structure of the EOBC Cohort

The EOBC cohort included all five subtypes (Table 1). The median age of the patients was 37, and the tumor stages ranged from Ia–IVb, with the largest number of patients at stages IIa and IIb. Eleven patients had a family history of breast cancer, including 2 sisters (IDs 7768 and 7942), who will be further discussed below.

**Table 1.** Taiwanese early onset breast cancer (EOBC) cohort structure and associated sequencing methods.

Category	Description	Number of Patients	WES	WGS
Subtype	Her2+	9	8	1
	Luminal A	34	34	0
	Luminal B/Her2+	20	18	2
	Luminal B/Her2-	14	13	1
	Triple negative	13	13	0
Age group	< 37 (median)	39	37	2
	≥ 37	51	49	2
Stage group	Ia, Ib	15	13	2
	IIa, IIb	42	41	1
	IIIa, IIIb, IIIc	24	23	1
	IVa, IVb	5	5	0
	Unknown	4	4	0
Family history	No family history	79	75	4
	With family history	11	11	0

WES—whole-exome sequencing; WGS—whole-genome sequencing.

The base coverage of the WGS libraries ranged between 30.6–44.2-fold, while that of the WES libraries ranged between 40–167-fold (Supplementary Table S1).

## 2.2. Somatic Mutation Analysis

A total of 17,175 somatic mutations distributed in 7475 protein-coding genes were identified. These mutations belonged to many categories, among which missense mutations constituted the majority (91.7%) (Supplementary Table S2).

To facilitate further discussion, a table in which patient clinical data were integrated with somatic mutation-associated information, especially for *TP53*, *BRCA1* and *BRCA2*, was generated (Supplementary Table S3) and the issues discussed in the remainder of this section are deduced from the information provided in table.

The *TP53* and *PIK3CA* together with three other randomly selected genes were validated by Sanger sequencing using patient BC0145 as an example (Supplementary Figure S1). The mutations in genomic locations found in tumor tissues were confirmed for all genes tested.

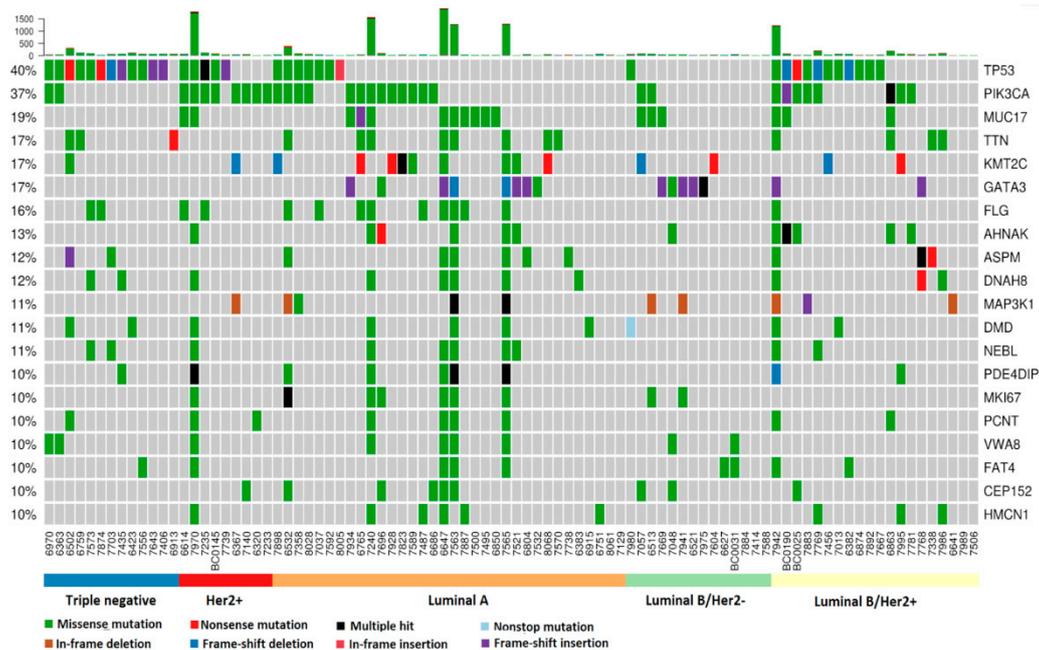
As frequently observed in breast cancer, a small portion of patients may exhibit distinctly high number of genes with somatic mutations. Here, among the 90 patients, 6 patients were found to have somatic mutations in more than one thousand genes (ranging from 1237–1926 genes, avg. 1520 per patient); these were categorized as high-mutation (HM) patients (Supplementary Figure S2). The remaining patients exhibited somatic mutations in 7–402 (avg. 63) genes and were categorized as low-mutation (LM) patients.

A striking characteristic of the HM patients was that all of them had somatic mutations in the *BRCA1* gene, and half of them also had somatic mutation(s) in the *TP53* and/or *BRCA2* gene (Supplementary Table S3). All three genes encode tumor suppressors involved in the DNA repair checkpoint and are among the most common susceptibility genes in breast cancer. The influence of these genes in determining the HM phenotype was further suggested by sisters 7768 and 7924 (both luminal B/Her2+). The elder sister, 7768, did not exhibit any mutation in any of these genes and belonged to the LM group, while the younger sister, 7924, had mutations in all three genes and belonged to the HM group. (More information regarding these sisters is shown in later section entitled “Case study of sisters in the cohort”.) Notably, the HM phenotype did not seem to correlate with clinical stage or with any particular subtype.

The genes that were most commonly affected by somatic mutations were similar to those previously reported in breast cancer (Figure 1). *TP53* and *PIK3CA* were distinctly ranked as first and second, with prevalences of 40% and 37%, respectively, followed by extracellular structural protein-coding genes *MUC17* (19%), *TTN* (17%) and *FLG* (16%), together with the transcription factors *KMT2C/MLL3* (17%) and *GATA3* (17%).

Notably, *TP53* somatic mutations showed a 92% prevalence among triple-negative patients (followed by a 55% prevalence in luminal B/Her2+ and 56% in Her2+), while *PIK3CA* somatic mutations showed an 89% prevalence among Her2+ patients (Supplementary Table S4). Interestingly, *GATA3* mutations were mainly associated with the luminal B/Her2- (36%) and luminal A (24%) subtypes.

*MUC17* encodes mucin 17, which is a highly expressed membrane-bound glycoprotein in intestinal epithelial cells that provides protection for the cells where it is expressed. The *TTN* gene encodes titin or connectin, which functions as a molecular spring by providing elasticity to muscle. Mutations in titin proteins result in the weakening of muscle and are responsible for heart disease and muscular dystrophy. The *FLG* gene encodes the filaggrin protein, which undergoes proteolysis to generate filaggrin monomers. By integrating into the cellular membrane, filaggrin monomers strengthen the skin barrier.



**Figure 1.** Top genes with somatic mutations with at least 10% prevalence among EOBC patients. The number on the top of each bar represents the number of mutations in a patient. High-mutation (HM) patients (6 total) can be distinctly identified. In the central panel, patients’ mutated genes are labeled with various colors, each representing a particular type of somatic mutation. The names of the mutated genes are listed on the right side of the panel, while the corresponding prevalence is shown on the left side of the panel.

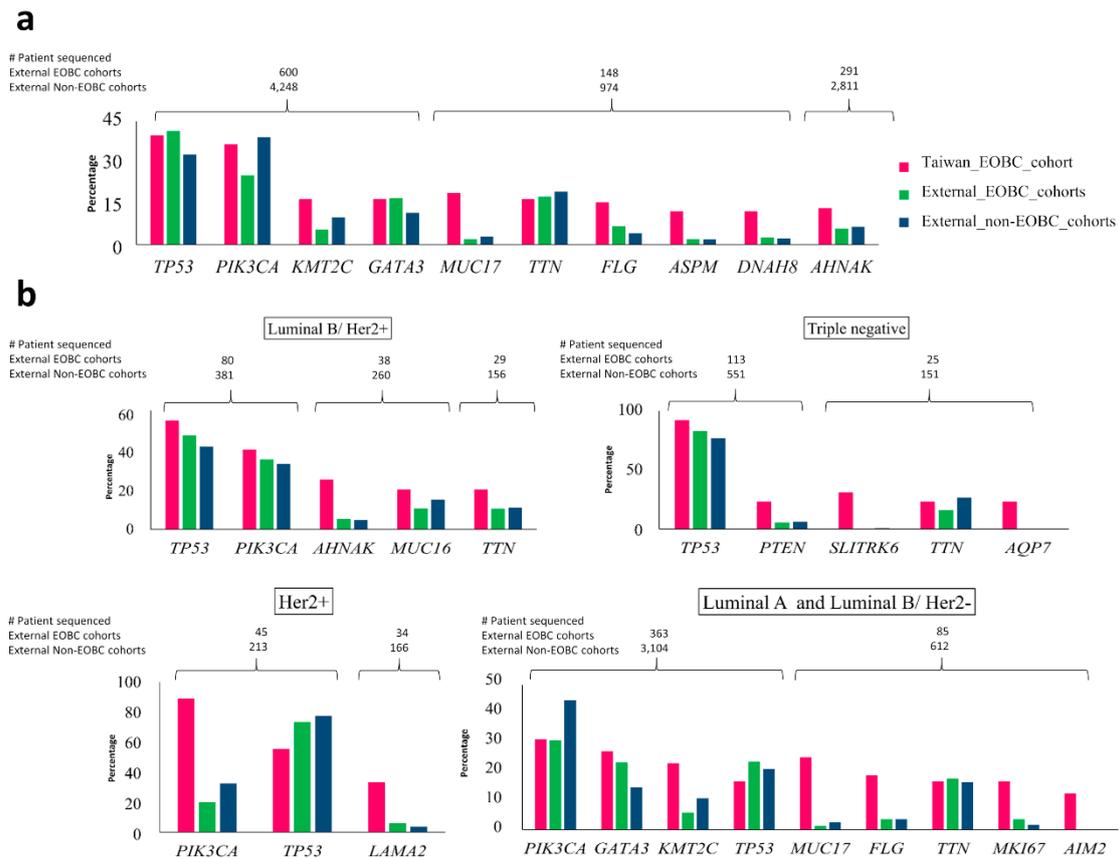
### 2.3. Comparison of Top Genes with Somatic Mutations in Taiwanese EOBC Cohort to Top Genes in EOBC and Non-EOBC Groups from Other Breast Cancer Cohorts

To gain more insight into the common and unique genic features of Taiwanese EOBC patients, we compared most prevalent genes in our EOBC cohort to their counterparts in EOBC and non-EOBC groups generated by combining 6 non-Taiwanese (external) cohorts.

Among the six non-Taiwanese cohorts, most patients were non-EOBC cases (50.93%–92.78%), while EOBC cases ranged only between 7.22%–49.07% (Supplementary Table S5). Overall, 690 (13.97%) and 4248 (86.03%) patient’s mutation and clinical data (including Taiwanese EOBC cohort) were analyzed for characterization of EOBC and non-EOBC, respectively.

For the publicly available breast cancer datasets, most of the frequently mutated BC susceptible genes were observed in both EOBC and non-EOBC groups, although the mutation frequencies were significantly different for certain genes (Supplementary Table S6). Mutations in *TP53* and *PIK3CA* genes were predominant among both EOBC and non-EOBC patients in the public datasets. Mutation frequency of *TP53*, *GATA3* and *ESR1* were higher in EOBC cases than non-EOBC cases. Mutations in *PIK3CA*, *CDH1*, *KMT2C* and *MAP3K1* were more frequently observed in non-EOBC cases.

We compared mutation frequencies of the top 10 genes with somatic mutations in the Taiwanese EOBC cohort to mutation frequencies of these genes in the pooled EOBC and non-EOBC patients (Figure 2, panel a).



**Figure 2.** Comparison of top genes with somatic mutations identified in Taiwanese EOBC cohort to the pooled EOBC and pooled non-EOBC groups in other cohorts. (a) Patients from external cohorts were divided into EOBC and non-EOBC patients and then pooled together to form EOBC and non-EOBC groups and compared with Taiwanese EOBC counterparts; (b) subtype-based comparison of top genes with somatic mutations in Taiwanese EOBC-to-EOBC and non-EOBC counterparts in external cohorts. For the comparison, EOBC and non-EOBC groups were further divided into subtypes. The most prevalent genes in each subtype of Taiwanese EOBC were compared to their counterpart genes in each subtype of EOBC and non-EOBC groups from external cohorts. For external cohorts luminal A and luminal B/Her2- subtypes were indistinguishable due to lack of information on ki67. These two subtypes from the Taiwanese EOBC cases were combined and compared with external cohorts. Percentage of patients with mutation in each gene is shown by vertical bars and the number of patients sequenced for each gene in external EOBC and non-EOBC cohorts is shown at the top of that gene.

The mutation frequency of *TP53* was similar in Taiwanese EOBC and pooled EOBC (40% and 41.5%, respectively), but higher than that of pooled non-EOBC (32.9%). A similar pattern was observed for *GATA3* (Taiwan EOBC, 16.7%; pooled EOBC, 17%; pooled non-EOBC, 11.6%). *PIK3CA* was mutated in 25.3% of pooled EOBC patients, 36.7% of Taiwanese EOBC patients and 39.3% in pooled non-EOBC patients. The mutation frequency of *KMT2C* was found to be highest in Taiwanese EOBC patients (16.7%), followed by pooled non-EOBC (10%) and pooled EOBC patients (5.5%). Mutation frequencies of *MUC17* (19%), *FLG* (15.6%), *AHNAK* (13.3%), *ASPM* (12.2%) and *DNAH8* (12.2%) were distinctly higher in Taiwanese EOBC patients than that of pooled EOBC and non-EOBC groups from other cohorts.

We then divided all pooled EOBC group and non-EOBC group into subtypes and compared the prevalences of the top genes in each subtype (Figure 2, panel b). For the Taiwanese EOBC cohort, luminal B/Her2+ subtype had *TP53* (55%), *PIK3CA* (40%), *AHNAK* (25%), *MUC16* (20%) and *TTN* (20%) as the most prevalently mutated genes; the triple-negative subtype had *TP53* (92%), *SLITRK6*

(31%), *PTEN* (23%), *TTN* (23%) and *AQP7* (23%) as the most prevalently mutated genes; and the Her2+ subtype had *PIK3CA* (88.9%), *TP53* (55.6%) and *LAMA2* (33.3%) as the most prevalently mutated genes.

Through the comparison, we observed that in the luminal B/Her2+ subtype, *TP53*, *PIK3CA*, *AHNAK*, *MUC16* and *TTN* have similar mutation frequencies among pooled EOBC and pooled non-EOBC patients from external cohorts and that their mutation frequencies in the Taiwanese EOBC cohort are slightly higher. We observed similar mutation frequencies for *TP53*, *PTEN* and *TTN* in the triple-negative subtype of pooled EOBC and pooled non-EOBC groups from external cohorts, but their mutation frequencies were higher in the Taiwanese EOBC cohort. In the Her2+ subtype of Taiwanese EOBC patients, mutation frequencies of *PIK3CA* and *LAMA2* (88.9% and 33.3%, respectively) were higher than that of pooled EOBC (20% and 5.9%, respectively) and pooled non-EOBC groups (32.4% and 3.6%, respectively) of other cohorts, whereas the *TP53* mutation frequency (55.6%) was lower than that of pooled EOBC (73.3%) and pooled non-EOBC (77.5%) groups. In the Her2+ subtype of external cohorts, *TP53* was the most prevalently mutated in both pooled EOBC and pooled non-EOBC patients, whereas in the Her2+ subtype of the Taiwanese cohort, *PIK3CA* was most prevalently mutated. In “luminal A and luminal B/Her2-” subtypes with ER+/PR+/Her2- status, mutation frequency of *PIK3CA* was lower in all EOBC patients than that of non-EOBC patients, while *GATA3* was higher.

Mutation frequencies of the top 10 genes in the Taiwanese EOBC cohort were compared with EOBC and non-EOBC groups of each external cohort individually. For *TP53*, variations in mutation frequencies among individual EOBC cohorts were higher than non-EOBC cohorts (Supplementary Figure S3). Variations in mutation frequencies of *PIK3CA* and *TTN* were moderately high in EOBC compared to non-EOBC patients across all cohorts, whereas variations in mutation frequencies of the other seven genes among EOBC and non-EOBC groups were similar across all cohorts.

We also compared the survival and disease/progression free survival along with the molecular characteristics of pooled EOBC and pooled non-EOBC groups. The survival of EOBC and non-EOBC was analyzed over a 420-month period. Patients from the non-EOBC group exhibited poor survival probabilities. Age, being an important risk factor for non-EOBC patients, corresponds to poor survival. Disease/progression-free survival time of patients from the pooled non-EOBC group was significantly higher than that of the pooled EOBC group over a 220-month period (Supplementary Figure S4). The hazard ratio of EOBC was 1.96 (confidence interval 1.05–3.51, *p*-value 0.0321) indicating an almost two times higher risk of cancer progression among the EOBC patients than among the non-EOBC patients.

#### 2.4. Association of Somatic Mutations with Family History of Breast Cancer

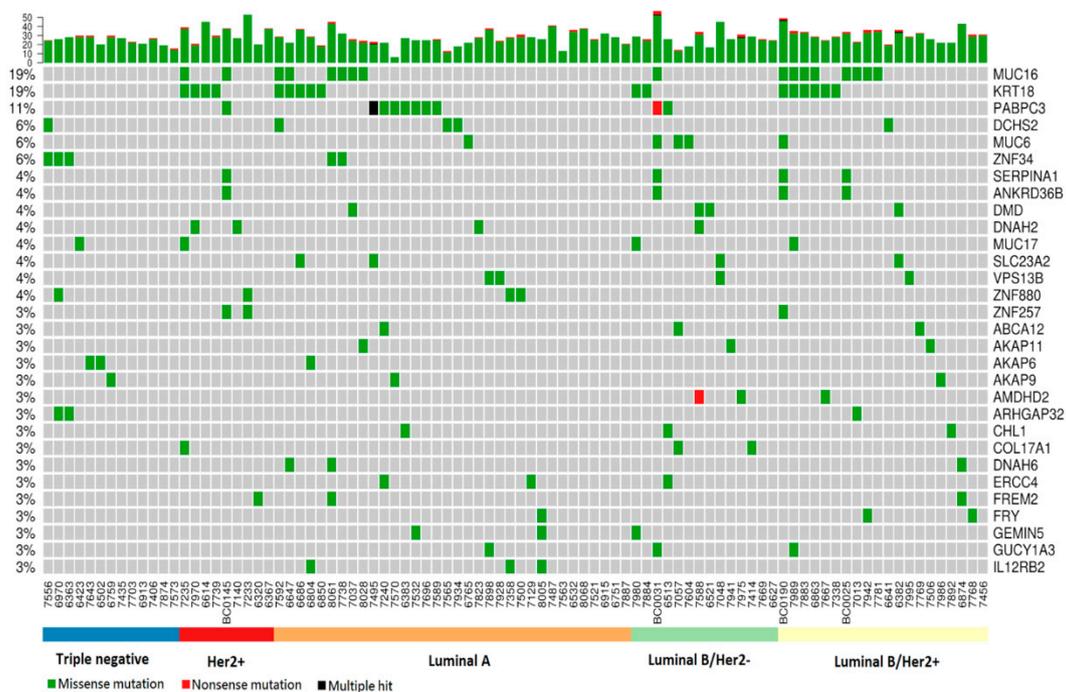
There were 11 patients in the cohort with a family history of breast cancer. Patients with and without a family history of breast cancer seemed to differ in the preferentially acquired somatic mutations. Patients with a family history were more likely to have somatic mutations in *TP53* (56%), *TTN* (36%) and *GATA3* (27%), while patients without a familiar history were likely to have somatic mutations in *PIK3CA* (39%), *TP53* (38%) and *MUC17* (20%). However, the data may not be representative because of the small population size for patients with a family history.

#### 2.5. Copy Number Variation Is Associated with Subtypes

CNV analysis revealed an association of CNVs with subtypes, each of which was in turn associated with CNVs of particular proto-oncogenes and tumor suppressor genes in the affected regions (Supplementary Table S7). Most evidently, the Her2+ subtype was associated with copy number increases in the proto-oncogenes *MLLT6*, *TBC1D3* and *TAF15*, but not in any tumor suppressor, while the luminal A subtype was associated with copy number increases in the proto-oncogenes *GNAS*, *TRIM27*, *MDM2*, *MCF2 L* and *RARA*, but in no tumor suppressor genes.

## 2.6. Germline Mutations

A total of 2690 germline mutations, affecting 2170 genes, were identified. Among these mutations, 2595 were missense mutations, while 95 were nonsense mutations. The genes with the most germline mutations included *MUC16* (19% prevalence), *KRT18* (19%), *PABPC3* (11%), *DCHS2* (6%), *MUC6* (6%) and *ZNF34* (6%) (Figure 3). Many of these genes encode extracellular and/or structural proteins.



**Figure 3.** Genes with germline mutations among the 90-patient EOBC cohort. (Top) Each bar represents the number of mutations for a patient; (Central panel) Patient mutated genes are labeled with various colors, each representing a particular type of germline mutation. The names of the mutated genes are listed on the right, with the corresponding prevalence shown on the left side.

*MUC16* encodes mucin 16, an o-glycosylated protein that forms a protective mucous layer on the surface of epithelia. *KRT18* (keratin 18) encodes keratin 18, a type I intermediate cytoskeleton filament. *PABPC3* encodes poly(A)-binding protein, cytoplasmic 3, involved in mRNA metabolism and translational initiation. *DCHS2* encodes dachsous cadherin-related-2, a large protein containing a large number of cadherin domains.

## 2.7. Pathway Analysis

Combinatorial pathway analyses of subtypes by integrating somatic mutations and germline mutations revealed a number of interesting pathways (Table 2). Genes in small cell lung cancer and focal adhesion pathways were commonly altered across all five subtypes. Mutations identified in Her2+ patients were enriched in osteoclast differentiation, sphingolipid-signaling and thyroid hormone-signaling pathways. ECM receptor interactions and PI3K-Akt-signaling pathways were altered in luminal A and both of the luminal B subtypes. Interestingly, in both luminal B/Her+ and luminal B/Her2- subtypes, ABC (ATP-binding cassette) transporter pathway was altered exclusively. Three pathways (namely small cell lung cancer, focal adhesion and endometrial cancer pathways) were altered in triple-negative subtype patients, which were common with other subtypes. The luminal B/Her2- subtype exhibited the most diverse affected pathways. Overall, the ABC transporter pathway, followed by ECM-receptor interaction and the focal adhesion and PI3K-Akt pathways, were the most significantly affected pathways among the patients in the Taiwanese EOBC cohort as a whole.

**Table 2.** Subtype-based pathway analysis.

Pathway ID	Pathway	Her2+	Luminal_A	Luminal B Her2+	Luminal B Her 2-	Triple Negative	Whole Cohort
* hsa05222	Small cell lung cancer	0.001	0.009	0.032	0.006	0.020	0.014
hsa04380	Osteoclast differentiation	0.006	–	–	–	–	–
* hsa05146	Amoebiasis	0.021	0.018	–	–	–	–
* hsa05200	Pathways in cancer	0.022	–	–	0.012	–	–
hsa04919	Thyroid hormone-signaling pathway	0.026	–	–	–	–	–
hsa04510	Focal adhesion	0.028	0.013	0.009	0.0001	0.051	0.004
hsa04071	Sphingolipid-signaling pathway	0.030	–	–	–	–	–
hsa04512	ECM–receptor interaction	–	0.001	0.008	0.007	–	0.001
* hsa05016	Huntington’s disease	–	0.010	–	–	–	0.031
hsa04151	PI3K–Akt-signaling pathway	–	0.016	0.009	0.001	–	0.032
* hsa05213	Endometrial cancer	–	–	0.006	0.021	0.048	–
hsa02010	ABC transporters	–	–	0.024	0.030	–	0.00004
hsa03460	Fanconi anemia pathway	–	–	0.039	–	–	–
hsa04015	Rap1-signaling pathway	–	–	–	0.0004	–	–
hsa05230	Central carbon metabolism in cancer	–	–	–	0.002	–	–
* hsa05218	Melanoma	–	–	–	0.014	–	–
* hsa05215	Prostate cancer	–	–	–	0.020	–	–
hsa04060	Cytokine–cytokine receptor interaction	–	–	–	0.024	–	–
hsa04923	Regulation of lipolysis in adipocytes	–	–	–	0.030	–	–
* hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	–	–	–	0.037	–	–
hsa04520	Adherens junction	–	–	–	0.037	–	–
hsa05205	Proteoglycans in cancer	–	–	–	0.043	–	–
* hsa04930	Type II diabetes mellitus	–	–	–	0.043	–	–
hsa04611	Platelet activation	–	–	–	0.048	–	–
* hsa05210	Colorectal cancer	–	–	–	0.049	–	–
hsa04630	Jak–STAT-signaling pathway	–	–	–	0.050	–	–
hsa04530	Tight junction	–	–	–	–	–	0.008
hsa04974	Protein digestion and absorption	–	–	–	–	–	0.017

\* indicates pathway related to diseases that may not be breast cancer.

### 2.8. Case Study of Sisters in the Cohort

To better understand whether family history plays a pathologic role that is predictive of EOBC tumorigenesis, we paid special attention to the sister pair with a focus on the common and different features between their germline backgrounds. These sisters (IDs 7768 and 7942) were diagnosed with breast cancer at ages 29 and 28, respectively and both belonged to the luminal B/Her2+ subtype and presented very similar CNV patterns (Supplementary Table S8).

In terms of germline mutations, patient 7768, belonging to the LM group, exhibited 31 mutated sites distributed in 31 genes, while patient 7942, belonging to the HM group, exhibited 37 mutated sites distributed in 36 genes. Nineteen sites distributed in 18 genes were shared by these two women, including a site in the *FRY* gene that showed a 3% prevalence among all EOBC patients (Supplementary Figure S5). However, they exhibited very different somatic mutations and mutations in *TP53*, *BRCA1* and *BRCA2* found in patient 7942 seemed to have been acquired after birth.

### 3. Discussion

We surveyed germline and somatic missense mutations as well as copy number variations in the EOBC cohort of Taiwanese women. Comparison of our EOBC with external EOBC and non-EOBC cohorts showed that *TP53* had the highest mutation frequency in EOBC across all cohorts, indicating that the *TP53*-mutation rate was significantly higher in EOBC patients. On the other hand, *PIK3CA* had the highest rate of mutation in non-EOBC patients. This phenomenon was consistent across all cohorts analyzed [18]. In addition, the Taiwanese population has the highest prevalence of *TP53* among the Asian populations that have been investigated. In a South Korean cohort of 229 patients with ER+ and Her2- breast cancer, *TP53* was found mutated in 10% of the patients [19], compared to 16.7% in 'luminal A and luminal B/Her2- combined' in our cohort. In a study of 116 triple-negative patients from Thailand [20], 76% were found to carry *TP53* mutations, comparing to 92% in the same subtype in our cohort. Recapitulating, the high prevalence of mutations in *TP53* implies a distinct molecular characteristic of EOBC and associated subtypes.

Family history was known to be associated with higher risk of EOBC than LOBC [21]. In American women, patients with family history of breast or ovarian cancer in first degree relatives had higher risk for EOBC (4.9% to 10.0%) and lower risk for LOBC (2.0% to 2.1%) [21]. Inherited genetic risk factors such as *BRCA1* and *BRCA2* mutations are observed more often among breast cancer patients, but are not the sole causal factors [22–24]. In a study of 1987 Dutch women with or without breast cancer, family history of breast cancer in first degree relatives was found to be one of the prominent risk factors for EOBC [25]. In our study, 11 EOBC patients had family history of breast cancer up to second degree of relatives. Although *BRCA1/2* or *TP53* mutations among the family members have been reported as a risk factors of EOBC, our study was limited due to the lack of background genetic information of family members for comparison. Several other studies related to breast cancers did not solely focus on EOBC in particular [18,26–29], while studies with a focus on EOBC either analyzed a particular subtype or preselected genes, and the cohort may be too small and thus not representative [19,20,23,30–34]. Our study included 5 major subtypes of breast cancer in young women (less than 40) and to the best of our knowledge, this is so far the only comprehensive study of EOBC in Taiwanese women. Similar to the previous reports by other groups, somatic mutations in genes such as *TP53*, *PIK3CA* and *GATA3* are associated with EOBC [19,20,30,33], while high prevalences in somatic mutations were observed in certain genes such as *KMT2C*, *FLG*, *MUC17* and *ASPM* in our cohort.

It was astonishing to observe an extremely large number of mutations in structural protein-coding genes in the studied group. This phenomenon was shown by multiple lines of evidence. First, the top 2 genes with germline mutations were *MUC16* and *KRT18*, both of which exhibited a 19% prevalence and encoded structural proteins. In addition, more mucin-coding genes, including *MUC6* and *MUC17*, were highly mutated. Second, compared to conventional breast cancer, somatic mutations in *MUC17*, *FLG* and *NEBL* were of much higher prevalence among EOBC patients. Third, these observations were further strengthened by the pathway analysis, which showed that the focal adhesion and ECM–receptor interaction pathways were the most prevalent affected pathways.

The *MUC16* gene contains multiple occurrences of the SEA domain (Sea urchin sperm protein, enterokinase, agrin) and its protein is thought to play a role in forming a protective layer for epithelial cells from pathogens and foreign particles/ligands [35]. In our analysis, we reported germline variations in this gene at regions where copies of the SEA domain are located. These variations are nonsynonymous in nature and may change the structural integrity of the barrier. Several studies have reported on the role of *MUC16* in different types of cancers including breast cancer [36–40]. As these alterations are germline variations, we hypothesize that, in breast cells they may have allowed the uptake/release of cell-signaling molecules leading to initiation/blockage of signaling pathways, subsequently triggering tumor development and progression at an early age in these women. Apart from *MUC16* which has germline variation present in 19% of the EOBC patients, *KRT18* and *PAPBC3* genes were having germline variations in 19% and 11% of patients, respectively. Interestingly, the latter two variations were mutually exclusive. *KRT18* is a known triple negative breast cancer marker and has also been

reported to play a role as oncogene in colorectal cancer [41,42]. *PABPC3* was found to have germline mutation in a breast cancer cohort of Tunisian females [43].

Extracellular and membrane-bound structural proteins form an occlusive barrier to protect an organism from various physical and chemical impacts from the environment. Based on these observations, we hypothesize that mutations in these structural proteins may result in a leaky epidermis and, thus, weaken the physical barrier, which would become more permeable to environmental toxins such as environmental carcinogens or hormones. The enhanced accumulation of toxins within breast cells will increase somatic mutations and initiate tumorigenesis at an early age. In fact, this possibility has been well documented for the filaggrin protein, encoded by the *FLG* gene [44], and functional defects resulting from loss-of-function mutations of *FLG* have been found to be a major risk factor for eczema [45,46]. Thus, besides driver mutations and passenger mutations acting in the downstream of tumorigenesis, mutations in extracellular protein-coding genes may act in the upstream and failure in gatekeeping is responsible for tumorigenesis developing at a young age. The role of gatekeeper may be even more important for the triple-negative subtype, in which ER, PR, Her2 receptors are absent or expressed at extremely low levels.

It is also noteworthy that some of the highly mutated structural genes, such as *TTN* and *MUC17*, encode very large proteins, and one may suspect that high mutation rates in these genes may simply be due to size effect. In fact, both *TTN* and *MUC17* have been directly or indirectly recognized as tumor-associated genes. For example, although most commonly known as a molecular spring by providing elasticity to muscle, *TTN* was found to associate with chromosomes to provide elasticity [47]. With close association with chromosomes, it is reasonable to speculate that mutations in *TTN* may cause mutations in the genome. Direct evidence was provided by Tan and colleagues. In their comprehensive survey of 23 cancers in COSMIC data, *TTN* was ranked second highest in terms of numbers of somatic mutations [48]. Furthermore, the large gene size of *TTN* was noticed in a report by Greenman and colleagues. In their study, *TTN* was ranked as the single most prevalent gene among 518 protein kinase genes in 210 different human cancers associated with an extremely high density of cancer driver mutations [49]. On the other hand, although mucins are essential components of the mucous layer in extracellular space, mucins were also known to be involved in inflammation and cancer [50], while Mucin 17 was further found to inhibit the progression of human gastric cancer [51].

Since tumor tissue availability is limited and varies with surgical requirements for the patient or size of tumor, it is hard to acquire samples for NGS and validation of mutations. In our analysis, we were able to validate mutations in the tumor of one of the patients. Our analysis, although limited by sample size, still correlates with external cohorts and provides important insights into EOBC in Taiwanese women.

## 4. Materials and Methods

### 4.1. WGS: Tissue Sample Collection, DNA Preparation and Whole-Genome Sequencing

Genomic DNA samples from WBC-tumor pairs from 4 patients (BC0145, BC0190, BC0025 and BC0031) ages of < 40 years were prepared at the Institute of Stem Cell and Translational Cancer Research (ISCTCR) at Chang Gung Memorial Hospital, Taiwan and stored at  $-80^{\circ}\text{C}$  according to routine laboratory practice. All genomic DNA samples were extracted by using the Gentra Puregene Blood kit (Qiagen, cat# 158445) and sequenced using the HiSeq X Ten platform (Macrogen, Inc., South Korea) with  $150 \times 150$  bp paired-end (PE) reads.

### 4.2. WES: Patient Recruitment, DNA Preparation and Whole-Exome Sequencing

A total of 86 female EOBC patients were recruited for the study between 2005 and 2010. All patients were diagnosed at age 41 or younger at National Taiwan University Hospital (NTUH). Eleven of them had a family history of breast cancer (mother or sister). Paired peripheral blood samples (buffy

coats) and tumor tissues were collected during surgery, and tumor tissue samples were examined by pathologists at NTUH.

Genomic DNA from the WBC samples was isolated immediately after blood collection and stored at  $-20\text{ }^{\circ}\text{C}$ , while the tumor tissues were stored at  $-20\text{ }^{\circ}\text{C}$  until DNA isolation. Genomic DNA from the tumor tissues was extracted using the DNeasy blood and tissue kit (Qiagen) and quantified by using the Qubit dsDNA BR assay kit (Invitrogen). Library preparation and exome enrichment were performed using the SureSelect human all exon system (Agilent Technologies) following the manufacturer's instructions for Illumina PE sequencing. Briefly, the DNA sample was sheared into 150–200 bp fragments. By using the paired-end sample preparation kit, the 5'-ends of the fragments were blunt ended and 'A' bases were added to the 3' ends. Index-specific PE adaptors were then ligated to both ends of the fragments. After removing the unligated adaptors, the ligated fragments were PCR-amplified using an Illumina InPE1.0 primer and a SureSelect indexing precapture primer. The quality and quantity of the purified samples were assessed by using an Agilent 2100 Bioanalyzer. For hybridization, the samples were processed with the Agilent SureSelect automated library prep and capture system. PE sequencing was conducted by using the Illumina genome analyzer IIX (GAIIIX) system (Illumina) at  $75 \times 75$  bp.

#### 4.3. Subtype Classification

Patients were classified based on the expression of the ER (estrogen receptor), PR (progesterone receptor), Her2 (human epidermal growth factor receptor 2) and Ki67 proteins [52,53]. Protein expression data were provided by the NTUH pathology lab. Immunohistochemical staining of proteins was performed on formalin-fixed, paraffin-embedded tissue sections and double-read by two pathologists. For the assessment of ER and PR expression, samples with  $>10\%$  positively stained nuclei were defined as positive. Her2 expression was scored as 0, 1+, 2+ or 3+ based on the percentage and intensity of staining. A score of 0 or 1+ was classified as negative and a score of 3+ was classified as positive. Samples with a score of 2+ were examined by fluorescence in situ hybridization (FISH), and those showing positive staining were classified as Her2-positive (Her2+). The cutoff value for the Ki67 protein was staining of 14% of cells. According to the immunohistochemical profiles, these tumors were classified into five subtypes: (1) luminal A (ER+ and/or PR-positive (PR+), Her2-negative (Her2-), Ki67  $< 14\%$ ); (2) luminal B/Her2- (ER+ and/or PR+, Her2-, Ki67  $\geq 14\%$ ); (3) luminal B/Her2+ (ER+ and/or PR+, Her2+, any Ki67); (4) Her2+ - non-luminal (ER-, PR-negative (PR-) and Her2+); and (5) Triple-negative (ER-, PR- and Her2-).

#### 4.4. Sequence Data Analysis Workflow

Raw reads from WGS were filtered to obtain quality reads. The procedure included (1) mapping all reads to the PhiX genome using Bowtie2 and then selecting unmappable reads using seqtk; (2) removing Illumina sequencing adapters using Cutadapt; (3) removing low-quality bases from the 5'- and 3'-ends using PRINseq with QV30 as the cutoff; (4) removing reads with at least two ambiguous bases using PRINseq; (5) selecting reads with  $\geq 30$  bases using PRINseq; (6) selecting quality reads using the NGS QC toolkit [54], (QV20, 70%); and (7) obtaining paired-end and single-end (SE) reads for further analysis. Quality reads were then mapped against human genome assembly hg19 using the Burrows wheeler aligner (BWA) version 0.7.5a [55]. Duplicates were marked using the MarkDuplicates module of PICARD tools (v1.98). The standard GATK pipeline [56], which includes RealignTargetCreator, IndelRealigner, BaseRecalibrator and PrintReads, was used to process bam files before variant calling.

#### 4.5. Somatic Mutation Analysis

Paired WBC and tumor sequence data were compared by using MuTect2 (GATK v4beta) with the default parameters to identify somatic mutations, followed by region- and feature-based annotations to identify somatic variants using ANNOVAR v2016Feb01 [57]. The Maftools program in R was used for further analyses, such as plotting mutation status.

#### 4.6. Validation of Mutations

To validate mutations, specific PCR primers flanking the mutation sites of five genes were designed and used to generate PCR fragments containing the mutation sites in patient BC0145. PCR fragments were cloned into the pZBack vector (Tools Biotechnology) and the constructs were used to transform *E. coli* DH5-Alpha competent cells, which were then cultured in selective agar plates. Colonies were picked and cultured in aqueous media. Plasmids were isolated with mini-preps and sequenced with Sanger sequencing.

#### 4.7. Cross-Comparison of Top Genes with Somatic Mutations between the EOBC Cohort and Other Breast Cancer Studies

To assess the molecular differences in EOBC and non-EOBC individuals, we collected six publicly available breast cancer datasets, which contain mutation and clinical information, from cBioPortal [58,59] and analyzed these datasets in a systematic manner. These datasets were obtained from 5975 samples including (1) 1918 samples from the breast cancer database [26]; (2) 2509 samples from METABRIC [27,28]; (3) 103 samples from the breast invasive carcinoma database [5]; (4) 100 samples from the Sanger Institute [29]; (5) 1108 samples from the TCGA provisional database; and (6) 237 samples from the metastatic breast cancer project. Patients without gender and/or age information, together with male patients, were excluded. Female patients were dichotomized into two groups on the basis of age at diagnosis, and patients  $\leq 41$  years were categorized as EOBC and patients above 41 years as non-EOBC.

For the comparison of somatic mutation patterns between EOBC and non-EOBC patients, we first analyzed somatic mutations for all these patients to identify EOBC- or non-EOBC-specific genes and then compared the prevalences of the 10 genes most susceptible to somatic mutations in the Taiwanese EOBC cohort to that in all other cohorts, either on individual or pooled basis for EOBC and non-EOBC separately. Disease/progression free survival time was defined as the time from diagnosis until first occurrence of relapse, progressive disease, secondary cancer or death or, if none occurred, until last contact. Patients from external cohorts were dichotomized into EOBC and non-EOBC groups based on age ( $< 41$  years) and survival probability was estimated by Kaplan–Meier survival curves. The relative risk of disease/progression free survival of EOBC was estimated by univariate Cox-proportional hazard regression analysis.

We further divided patients from all datasets into subtypes, based on the criteria of the St. Gallen subtypes classification mentioned previously in the “subtype classification” section. Since most cohorts lacked Ki67 information, classification of luminal A and luminal B/Her2- was not feasible, so we combined these two subtypes into “luminal A and luminal B/Her2-, with receptor status ER+/PR+/Her2-” and analyzed them together with luminal B/Her2+, Her2+ and triple-negative subtypes based on ER, PR and Her2 status. These subtypes were further compared for EOBC and non-EOBC patients combined.

#### 4.8. Germline Mutation Analysis

GATK HaplotypeCaller was used to identify germline mutations in WBC samples and annotations were added using ANNOVAR. To select novel germline mutations that may potentially be associated with EOBC, we compared our data with public databases. First, we checked for these mutations in the Taiwan Biobank (database for SNP information from 1000 normal Taiwanese people, <https://taiwanview.twbiobank.org.tw/index>). Allele frequencies (AFs) were calculated for each germline missense mutation in 90 patients and compared with the AF at the same location in the Taiwan Biobank using the chi-squared test. Genomic locations with significant differences ( $p$ -value  $\leq 0.05$ ) between the sample and Biobank populations are reported in the results. Germline mutations were also checked with the dbSNP and COSMIC databases. Those mutations not present in dbSNP were retained for further analysis.

#### 4.9. Copy Number Variation Analysis

Bam files for WBCs and tumors were further processed with GATK modules (CalculateTargetCoverage, CombineReadCounts, CreatePanelOfNormals, NormalizeSomaticReadCounts, PerformSegmentation and CallSegments) to obtain coverage information. We analyzed the CNV profiles of 90 EOBC patients by grouping them according to molecular subtypes to obtain subtype-specific alterations. GISTIC2.0 was used to call targets with somatic copy number alterations in each subtype [60].

#### 4.10. Pathway Analysis for Germline and Somatic Mutations

Six lists of genes were prepared for pathway analysis, one for each of the five subtypes and one for the whole cohort. Each list comprised genes with germline mutations and genes with somatic mutations. The following criteria were followed to select genes for pathway analysis: (1) genes with germline mutations in  $\geq 2$  patients were selected, except that all genes in the luminal B/Her2- subtype were included; (2) genes with somatic mutations in at least 2 patients or in  $\geq 10\%$  of the patients (whichever is higher), except that all genes in the luminal B/Her2- subtype were included; and (3) for the whole cohort pathway analysis, the list of genes with germline mutations in  $\geq 2$  patients and the list of genes with somatic mutations in at least 5% of the total patients were combined to be used as the input. The DAVID 6.8 webserver was used for pathway analysis and *Homo sapiens* was selected as the background organism. KEGG pathways with *p*-values of less than 0.05 were selected as enriched pathways in EOBC patients.

#### 4.11. Ethics Approval and Consent to Participate

All human materials used in the study for either WGS or WES analysis were reviewed and approved by the Human Subject Research Ethics Committee/ Institution Review Board (IRB) of the Academia Sinica (Approval number: AS-IRB02-99065), NTUH and the ISCTCR, Chang Gung Memorial Hospital. All participating patients provided their clinical information and informed consent for the research. All regulatory guidelines were strictly complied with throughout the study.

#### 4.12. Availability of Data and Material

Key genomic aberrations and clinical data are provided in supplementary files. The BAM files generated and/or analyzed during current study are available in NCBI with BioProject ID: PRJNA574486. Somatic mutation and germline mutation data are available in our website at <https://topsciencebiotech.com/main/eobc/>.

## 5. Conclusions

We performed comprehensive analysis of the EOBC mutation spectrum in Taiwanese women. Our study highlighted distinguishable molecular features of EOBC and non-EOBC patients by incorporating other breast cancer datasets. Further, we identified diverse mutation profiles among molecular subtypes. Our results indicated that mutations in the genes of ABC transporters and ECM–receptor interaction, focal adhesion and PI3K–Akt-signaling pathways could be potential therapeutic targets. Additionally, somatic mutations alone may not be sufficient to characterize EOBC disease, while germline mutations in structural proteins may also act as a causal factor to initiate EOBC tumorigenesis. Functional characterization of key mutations may provide further insights into the mechanism of cancer development to improve clinical outcomes.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2072-6694/12/8/2089/s1>, Supplementary Figure S1: Validation of somatic mutations by Sanger sequencing, Supplementary Figure S2: Number of genes affected by somatic mutations among patients in Taiwanese EOBC cohort, Supplementary Figure S3: Comparison of top 10 genes with somatic mutations in Taiwanese EOBC cohort to the individual EOBC (A) and individual non-EOBC (B) groups in each external cohort, Supplementary Figure S4: Kaplan Meier disease/progression-free survival plot for pooled EOBC and pooled non-EOBC patients from external cohorts, Supplementary Figure S5: Common and specific germline mutations in sisters with a family history of breast

cancer, Supplementary Table S1: Library statistics, Supplementary Table S2: Characterization of somatic mutations, Supplementary Table S3: Clinical information and somatic mutations of all Taiwanese EOBC patients (As of May, 2017), Supplementary Table S4: Subtype-based top genes with somatic mutations, Supplementary Table S5: Description of non-Taiwanese breast cancer cohorts used for comparison, Supplementary Table S6: Differentially mutated genes in pooled EOBC and pooled non-EOBC of external cohorts, Supplementary Table S7: Copy number variations in different cytobands by subtype, Supplementary Table S8: Overlap in CNV regions between sisters 7768 and 7942.

**Author Contributions:** Conceptualization, C.-J.C., C.-Y.S., A.L.Y., K.-P.C. and M.K.M.; methodology, C.-J.C., A.L.Y., C.-Y.S., K.-P.C., T.-H.C. and M.K.M.; validation, M.K.M., K.-P.C., A.L.Y., T.-C.F. and T.-H.C.; formal analysis, K.-P.C., M.K.M., Y.-F.H., H.-H.Y. and C.-Y.S.; investigation, K.-P.C., M.K.M., Y.-F.H., H.-H.Y., A.L.Y. and C.-Y.S.; resources, C.-J.C., C.-Y.S., A.L.Y., K.-P.C., T.-C.F., W.-H.K., Y.-T.W., K.-J.C. and N.-C.C.; data curation, K.-P.C., M.K.M., Y.-F.H., H.-H.Y. and C.-Y.S.; writing—original draft preparation, K.-P.C., M.K.M., C.-Y.S. and A.L.Y.; writing—review & editing, K.-P.C., C.-J.C., M.K.M., C.-Y.S. and A.L.Y.; supervision, C.-J.C., C.-Y.S., A.L.Y. and K.-P.C.; project administration, C.-J.C., C.-Y.S., A.L.Y. and K.-P.C.; funding acquisition, C.-J.C., A.L.Y. and K.-P.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by a Summit Project grant for the study of early onset breast cancer and a grant for the Biosignature Project in collaboration with Chang Gung University Hospital. Both were provided by the Academia Sinica.

**Acknowledgments:** We appreciate all patients and research staff participating in the study. We thank Yung Ming Jeng (Department of Pathology, National Taiwan University Hospital) for the contributions of the IHC and FISH analyses. We are grateful to the National Center for High-performance Computing (NCHC) for computer time and facilities.

**Conflicts of Interest:** No potential conflicts of interest were declared.

## References

- Bonnier, P.; Romain, S.; Charpin, C.; Lejeune, C.; Tubiana, N.; Martin, P.M.; Piana, L. Age as a prognostic factor in breast cancer: Relationship to pathologic and biologic features. *Int. J. Cancer* **1995**, *62*, 138–144. [[CrossRef](#)] [[PubMed](#)]
- Anderson, W.F.; Pfeiffer, R.M.; Dores, G.M.; Sherman, M.E. Comparison of age distribution patterns for different histopathologic types of breast carcinoma. *Cancer Epidemiol. Biomark. Prev.* **2006**, *15*, 1899–1905. [[CrossRef](#)] [[PubMed](#)]
- Yankaskas, B.C. Epidemiology of breast cancer in young women. *Breast Dis.* **2005**, *23*, 3–8. [[CrossRef](#)] [[PubMed](#)]
- Bray, F.; McCarron, P.; Parkin, D.M. The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Res.* **2004**, *6*, 229–239. [[CrossRef](#)]
- Banerji, S.; Cibulskis, K.; Rangel-Escareno, C.; Brown, K.K.; Carter, S.L.; Frederick, A.M.; Lawrence, M.S.; Sivachenko, A.Y.; Sougnez, C.; Zou, L.; et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **2012**, *486*, 405–409. [[CrossRef](#)] [[PubMed](#)]
- Anders, C.K.; Johnson, R.; Litton, J.; Phillips, M.; Bleyer, A. Breast cancer before age 40 years. *Semin. Oncol.* **2009**, *36*, 237–249. [[CrossRef](#)]
- Shen, Y.C.; Chang, C.J.; Hsu, C.; Cheng, C.C.; Chiu, C.F.; Cheng, A.L. Significant difference in the trends of female breast cancer incidence between Taiwanese and Caucasian Americans: Implications from age-period-cohort analysis. *Cancer Epidemiol. Biomark. Prev.* **2005**, *14*, 1986–1990. [[CrossRef](#)]
- Foo, C.S.; Su, D.; Chong, C.K.; Chng, H.C.; Tay, K.H.; Low, S.C.; Tan, S.M. Breast cancer in young Asian women: Study on survival. *ANZ J. Surg.* **2005**, *75*, 566–572. [[CrossRef](#)]
- Narod, S.A. Breast cancer in young women. *Nat. Rev. Clin. Oncol.* **2012**, *9*, 460–470. [[CrossRef](#)]
- Chang, L.-Y.; Yang, Y.-L.; Shyu, M.-K.; Hwa, H.-L.; Hsieh, F.-J. Strategy for Breast Cancer Screening in Taiwan: Obstetrician-Gynecologists Should Actively Participate in Breast Cancer Screening. *J. Med. Ultrasound* **2011**, *20*, 1–7. [[CrossRef](#)]
- Brenner, D.R.; Brockton, N.T.; Kotsopoulos, J.; Cotterchio, M.; Boucher, B.A.; Courneya, K.S.; Knight, J.A.; Olivotto, I.A.; Quan, M.L.; Friedenreich, C.M. Breast cancer survival among young women: A review of the role of modifiable lifestyle factors. *Cancer Causes Control* **2016**, *27*, 459–472. [[CrossRef](#)] [[PubMed](#)]
- Sorlie, T.; Tibshirani, R.; Parker, J.; Hastie, T.; Marron, J.S.; Nobel, A.; Deng, S.; Johnsen, H.; Pesich, R.; Geisler, S.; et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 8418–8423. [[CrossRef](#)] [[PubMed](#)]

13. Perou, C.M.; Sorlie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; et al. Molecular portraits of human breast tumours. *Nature* **2000**, *406*, 747–752. [[CrossRef](#)] [[PubMed](#)]
14. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490*, 61–70. [[CrossRef](#)]
15. Brewster, A.M.; Chavez-MacGregor, M.; Brown, P. Epidemiology, biology, and treatment of triple-negative breast cancer in women of African ancestry. *Lancet Oncol.* **2014**, *15*, 625–634. [[CrossRef](#)]
16. Aebi, S.; Sun, Z.; Braun, D.; Price, K.N.; Castiglione-Gertsch, M.; Rabaglio, M.; Gelber, R.D.; Crivellari, D.; Lindtner, J.; Snyder, R.; et al. Differential efficacy of three cycles of CMF followed by tamoxifen in patients with ER-positive and ER-negative tumors: Long-term follow up on IBCSG Trial IX. *Ann. Oncol.* **2011**, *22*, 1981–1987. [[CrossRef](#)]
17. Pharoah, P.D.; Antoniou, A.; Bobrow, M.; Zimmern, R.L.; Easton, D.F.; Ponder, B.A. Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **2002**, *31*, 33–36. [[CrossRef](#)] [[PubMed](#)]
18. Nik-Zainal, S.; Davies, H.; Staaf, J.; Ramakrishna, M.; Glodzik, D.; Zou, X.; Martincorena, I.; Alexandrov, L.B.; Martin, S.; Wedge, D.C.; et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **2016**, *534*, 47–54. [[CrossRef](#)]
19. Ahn, S.G.; Yoon, C.I.; Lee, J.H.; Lee, H.S.; Park, S.E.; Cha, Y.J.; Cha, C.; Bae, S.J.; Lee, K.A.; Jeong, J. Low PR in ER(+)/HER2(-) breast cancer: High rates of TP53 mutation and high SUV. *Endocr. Relat. Cancer* **2018**. [[CrossRef](#)]
20. Niyomnaitham, S.; Parinyanitikul, N.; Roothumnong, E.; Jinda, W.; Samarnthai, N.; Atikankul, T.; Suktitipat, B.; Thongnoppakhun, W.; Limwongse, C.; Pithukpakorn, M. Tumor mutational profile of triple negative breast cancer patients in Thailand revealed distinctive genetic alteration in chromatin remodeling gene. *PeerJ* **2019**, *7*. [[CrossRef](#)]
21. Dite, G.S.; Jenkins, M.A.; Southey, M.C.; Hocking, J.S.; Giles, G.G.; McCredie, M.R.; Venter, D.J.; Hopper, J.L. Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations. *J. Natl. Cancer Inst.* **2003**, *95*, 448–457. [[CrossRef](#)]
22. Cui, J.; Hopper, J.L. Why are the majority of hereditary cases of early-onset breast cancer sporadic? A simulation study. *Cancer Epidemiol. Biomark. Prev.* **2000**, *9*, 805–812. [[PubMed](#)]
23. Loizidou, M.; Marcou, Y.; Anastasiadou, V.; Newbold, R.; Hadjisavvas, A.; Kyriacou, K. Contribution of BRCA1 and BRCA2 germline mutations to the incidence of early-onset breast cancer in Cyprus. *Clin. Genet.* **2007**, *71*, 165–170. [[CrossRef](#)]
24. Walsh, P.C. Re: Increased cancer risks for relatives of very early-onset breast cancer cases with and without BRCA1 and BRCA2 mutations. *J. Urol.* **2011**, *185*. [[CrossRef](#)]
25. De Bock, G.H.; Jacobi, C.E.; Seynaeve, C.; Krol-Warmerdam, E.M.; Blom, J.; van Asperen, C.J.; Cornelisse, C.J.; Klijn, J.G.; Devilee, P.; Tollenaar, R.A.; et al. A family history of breast cancer will not predict female early onset breast cancer in a population-based setting. *BMC Cancer* **2008**, *8*. [[CrossRef](#)] [[PubMed](#)]
26. Razavi, P.; Chang, M.T.; Xu, G.; Bandlamudi, C.; Ross, D.S.; Vasan, N.; Cai, Y.; Bielski, C.M.; Donoghue, M.T.A.; Jonsson, P.; et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* **2018**, *34*, 427–438. [[CrossRef](#)] [[PubMed](#)]
27. Curtis, C.; Shah, S.P.; Chin, S.F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **2012**, *486*, 346–352. [[CrossRef](#)]
28. Pereira, B.; Chin, S.F.; Rueda, O.M.; Vollan, H.K.; Provenzano, E.; Bardwell, H.A.; Pugh, M.; Jones, L.; Russell, R.; Sammut, S.J.; et al. The somatic mutation profiles of 2433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **2016**, *7*. [[CrossRef](#)]
29. Stephens, P.J.; Tarpey, P.S.; Davies, H.; Van Loo, P.; Greenman, C.; Wedge, D.C.; Nik-Zainal, S.; Martin, S.; Varela, I.; Bignell, G.R.; et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **2012**, *486*, 400–404. [[CrossRef](#)]
30. Encinas, G.; Sabelnykova, V.Y.; de Lyra, E.C.; Hirata Katayama, M.L.; Maistro, S.; de Vasconcellos Valle, P.W.M.; de Lima Pereira, G.F.; Rodrigues, L.M.; de Menezes Pacheco Serio, P.A.; de Gouvea, A.; et al. Somatic mutations in early onset luminal breast cancer. *Oncotarget* **2018**, *9*, 22460–22479. [[CrossRef](#)]

31. Kadalayil, L.; Khan, S.; Nevanlinna, H.; Fasching, P.A.; Couch, F.J.; Hopper, J.L.; Liu, J.; Maishman, T.; Durcan, L.; Gerty, S.; et al. Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. *Nat. Commun.* **2017**, *8*. [[CrossRef](#)] [[PubMed](#)]
32. Shen, M.; Yang, L.; Lei, T.; Xiao, L.; Li, L.; Zhang, P.; Feng, W.; Ye, F.; Bu, H. BRCA1/2 mutation spectrum in Chinese early-onset breast cancer. *Transl. Cancer Res.* **2019**, *8*, 483–490. [[CrossRef](#)]
33. Zhang, Y.; Cai, Q.; Shu, X.O.; Gao, Y.T.; Li, C.; Zheng, W.; Long, J. Whole-Exome Sequencing Identifies Novel Somatic Mutations in Chinese Breast Cancer Patients. *J. Mol. Genet. Med.* **2015**, *9*. [[CrossRef](#)] [[PubMed](#)]
34. Ahsan, H.; Halpern, J.; Kibriya, M.G.; Pierce, B.L.; Tong, L.; Gamazon, E.; McGuire, V.; Felberg, A.; Shi, J.; Jasmine, F.; et al. A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. *Cancer Epidemiol. Biomark. Prev.* **2014**, *23*, 658–669. [[CrossRef](#)] [[PubMed](#)]
35. Marcos-Silva, L.; Narimatsu, Y.; Halim, A.; Campos, D.; Yang, Z.; Tarp, M.A.; Pereira, P.J.; Mandel, U.; Bennett, E.P.; Vakhrushev, S.Y.; et al. Characterization of binding epitopes of CA125 monoclonal antibodies. *J. Proteome Res.* **2014**, *13*, 3349–3359. [[CrossRef](#)]
36. Kanwal, M.; Ding, X.J.; Song, X.; Zhou, G.B.; Cao, Y. MUC16 overexpression induced by gene mutations promotes lung cancer cell growth and invasion. *Oncotarget* **2018**, *9*, 12226–12239. [[CrossRef](#)]
37. Li, X.; Pasche, B.; Zhang, W.; Chen, K. Association of MUC16 mutation with tumor mutation load and outcomes in patients with gastric cancer. *JAMA Oncol.* **2018**, *4*, 1691–1698. [[CrossRef](#)]
38. Patel, J.S.; Callahan, B.M.; Chobrutskiy, B.I.; Blanck, G. Matrix-Metalloprotease resistant mucin-16 (muc16) peptide mutants represent a worse lung adenocarcinoma outcome. *proteomics Clin. Appl.* **2019**, *13*. [[CrossRef](#)]
39. Taniguchi, T.; Woodward, A.M.; Magnelli, P.; McColgan, N.M.; Lehoux, S.; Jacobo, S.M.P.; Mauris, J.; Argueso, P. N-Glycosylation affects the stability and barrier function of the MUC16 mucin. *J. Biol. Chem.* **2017**, *292*, 11079–11090. [[CrossRef](#)]
40. Williams, K.A.; Terry, K.L.; Tworoger, S.S.; Vitonis, A.F.; Titus, L.J.; Cramer, D.W. Polymorphisms of MUC16 (CA125) and MUC1 (CA15.3) in relation to ovarian cancer risk and survival. *PLoS ONE* **2014**, *9*, e88334. [[CrossRef](#)]
41. Zhang, J.; Hu, S.; Li, Y. KRT18 is correlated with the malignant status and acts as an oncogene in colorectal cancer. *Biosci. Rep.* **2019**, *39*. [[CrossRef](#)] [[PubMed](#)]
42. Sporikova, Z.; Koudelakova, V.; Trojanec, R.; Hajduch, M. Genetic Markers in Triple-Negative Breast Cancer. *Clin. Breast Cancer* **2018**, *18*. [[CrossRef](#)] [[PubMed](#)]
43. Hamdi, Y.; Boujemaa, M.; Ben Rekaya, M.; Ben Hamda, C.; Mighri, N.; El Benna, H.; Mejri, N.; Labidi, S.; Daoud, N.; Naouali, C.; et al. Family specific genetic predisposition to breast cancer: Results from Tunisian whole exome sequenced breast cancer cases. *J. Transl. Med.* **2018**, *16*. [[CrossRef](#)] [[PubMed](#)]
44. O'Regan, G.M.; Sandilands, A.; McLean, W.H.I.; Irvine, A.D. Filaggrin in atopic dermatitis. *J. Allergy Clin. Immunol.* **2008**, *122*, 689–693. [[CrossRef](#)]
45. Bisgaard, H.; Simpson, A.; Palmer, C.N.; Bonnelykke, K.; McLean, I.; Mukhopadhyay, S.; Pipper, C.B.; Halkjaer, L.B.; Lipworth, B.; Hankinson, J.; et al. Gene-environment interaction in the onset of eczema in infancy: Filaggrin loss-of-function mutations enhanced by neonatal cat exposure. *PLoS Med.* **2008**, *5*. [[CrossRef](#)]
46. Henderson, J.; Northstone, K.; Lee, S.P.; Liao, H.; Zhao, Y.; Pembrey, M.; Mukhopadhyay, S.; Smith, G.D.; Palmer, C.N.; McLean, W.H.; et al. The burden of disease associated with filaggrin mutations: A population-based, longitudinal birth cohort study. *J. Allergy Clin. Immunol.* **2008**, *121*, 872–877. [[CrossRef](#)]
47. Machado, C.; Sunkel, C.E.; Andrew, D.J. Human autoantibodies reveal titin as a chromosomal protein. *J. Cell Biol.* **1998**, *141*, 321–333. [[CrossRef](#)]
48. Tan, H.; Bao, J.; Zhou, X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Sci. Rep.* **2015**, *5*. [[CrossRef](#)]
49. Greenman, C.; Stephens, P.; Smith, R.; Dalgliesh, G.L.; Hunter, C.; Bignell, G.; Davies, H.; Teague, J.; Butler, A.; Stevens, C.; et al. Patterns of somatic mutation in human cancer genomes. *Nature* **2007**, *446*, 153–158. [[CrossRef](#)]
50. Kufe, D.W. Mucins in cancer: Function, prognosis and therapy. *Nat. Rev. Cancer* **2009**, *9*, 874–885. [[CrossRef](#)]
51. Yang, B.; Wu, A.; Hu, Y.; Tao, C.; Wang, J.M.; Lu, Y.; Xing, R. Mucin 17 inhibits the progression of human gastric cancer by limiting inflammatory responses through a MYH9-p53-RhoA regulatory feedback loop. *J. Exp. Clin. Cancer Res.* **2019**, *38*. [[CrossRef](#)] [[PubMed](#)]

52. Vasconcelos, I.; Hussainzada, A.; Berger, S.; Fietze, E.; Linke, J.; Siedentopf, F.; Schoenegg, W. The St. Gallen surrogate classification for breast cancer subtypes successfully predicts tumor presenting features, nodal involvement, recurrence patterns and disease free survival. *Breast* **2016**, *29*, 181–185. [[CrossRef](#)] [[PubMed](#)]
53. Kondov, B.; Milenković, Z.; Kondov, G.; Petrushevska, G.; Basheska, N.; Bogdanovska-Todorovska, M.; Tolevska, N.; Ivković, L. Presentation of the molecular subtypes of breast cancer detected by immunohistochemistry in surgically treated patients. *Open Access Maced. J. Med. Sci.* **2018**, *6*, 961–967. [[CrossRef](#)] [[PubMed](#)]
54. Patel, R.K.; Jain, M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE* **2012**, *7*, e30619. [[CrossRef](#)]
55. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)] [[PubMed](#)]
56. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [[CrossRef](#)]
57. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*. [[CrossRef](#)]
58. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2*, 401–404. [[CrossRef](#)]
59. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **2013**, *6*. [[CrossRef](#)]
60. Mermel, C.H.; Schumacher, S.E.; Hill, B.; Meyerson, M.L.; Beroukhi, R.; Getz, G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **2011**, *12*. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).