cancers

MDPI

Article

Tumor-Infiltrating CD8 T Cells Predict Clinical Breast Cancer Outcomes in Young Women

Yong Won Jin [1,2] and Pingzhao Hu [1,2,3,4,*]

[1] Department of Biochemistry & Medical Genetics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB R3E0J9, Canada; jiny2@myumanitoba.ca

[2] Research Institute in Oncology and Hematology, Cancer Care Manitoba, Winnipeg, MB R3E 0V9, Canada

[3] Department of Electrical and Computer Engineering, Faculty of Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada

[4] Department of Computer Science, Faculty of Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

* Correspondence: pingzhao.hu@umanitoba.ca; Tel.: +1-204-789-3229

Received: 31 March 2020; Accepted: 24 April 2020; Published: 26 April 2020

Abstract: Young women with breast cancer have disproportionately poor clinical outcomes compared to their older counterparts. The underlying biological differences behind this age-dependent disparity are still unknown and warrant investigation. Recently, the tumor immune landscape has received much attention for its prognostic value and therapeutic targets. The differential tumor immune landscape between age groups in breast cancer has not yet been characterized, and may contribute to the age-related differences in clinical outcomes. Computational deconvolution was used to quantify abundance of immune cell types from bulk transcriptome profiles of breast cancer patients from two independent datasets. No significant differences in immune cell composition that were consistent in the two cohorts were found between the young and old age groups. Regardless of absence of significant differences, the higher tumor infiltration of several immune cell types, such as CD8+ T and CD4+ T cells, was associated with better clinical outcomes in the young but not in the old age group. Mutational signatures analysis showed signatures previously not found in breast cancer to be associated with tumor-infiltrating lymphocyte (TIL) levels in the young age group, whereas in the old group, all significant signatures were those previously found in breast cancer. Pathway analysis revealed different gene sets associated with TIL levels for each age group from the two cohorts. Overall, our results show trends towards better clinical outcomes for high TIL levels, especially CD8+ T cells, but only in the young age group. Furthermore, our work suggests that the underlying biological differences may involve multiple levels of tumor physiology.

Keywords: early-onset breast cancer; deconvolution; tumor infiltration lymphocytes; mutational signatures; pathway analysis

1. Introduction

Breast cancer is the most prevalent type of cancer in women worldwide [1]. It is a highly heterogeneous disease with multiple subtypes and classifications [2]. There are disparities in pathological features and disease outcomes between younger (age at diagnosis <40) versus older (age at diagnosis ≥40) breast cancer patients [1,3,4]. Young women with breast cancer are more commonly diagnosed with aggressive, invasive types of breast cancer that are difficult to treat. Studies found survival to be inversely associated with age at diagnosis [4–6]. The underlying biological cause for this age-dependent disparity in survival outcome is still unknown [1].

Recent studies have provided accumulating evidence that the presence of tumor-infiltrating lymphocytes (TILs) and their composition show significant associations with prognosis and response

to cancer treatments [7,8]. Specific TILs such as CD8+ cytotoxic T lymphocytes, T helper 1 cells, M1 macrophages, natural killer (NK) cells, and T-follicular helper (Tfh) cells have been reported as exhibiting anti-tumor activities, whereas T-regulatory (Treg) cells and M2 macrophages are known for their immune-inhibitory and thus pro-tumor activities [7]. Furthermore, measures of TILs have been shown to be markers for pathological complete remission, chemosensitivity, and improved recurrence-free survival, especially in non-luminal, receptor-negative breast cancers [9–11].

Today, there is an abundance of bulk transcriptomic data available publicly online [12–14]. These bulk transcriptomes often represent the average gene expression across a heterogeneous mixture of cells. If cellular components and their proportions can be identified from bulk transcriptomic data by computational methods, such in-silico methods can be used to characterize and quantify immune infiltrates in a cost-, time-, and labor-effective manner.

Deconvolution is a computational problem of simplifying a complex mixture into its individual constituents. In brief, most deconvolution algorithms see bulk transcriptome as a mixture where one gene of the mixture is a linear combination of that gene expressed across different cell types, weighted by the proportions of those cell types [15]. Deconvolution algorithms were previously applied to a large breast cancer transcriptomic dataset by others to characterize immune infiltrates across multiple samples [16,17]. After stratification by estrogen receptor (ER) status, Ali et al. found CD8+ T cells and activated memory T cells to be associated with favorable clinical outcomes in ER-negative tumors, which supported similar findings in the literature [7,16,18].

The tumor immune landscape, as predicted by computational deconvolution, has the potential to provide prognostic information as well as providing insight into immune functions within solid tumors. If age group differences in tumor immune landscape exist, these differences may be able to explain the age-dependent disparities in clinical outcomes in breast cancer.

2. Results

2.1. Estimates of Immune Cell Abundance by TIMER

Computational deconvolution of bulk gene expression data by the tumor immune estimation resource (TIMER) method, which used the constrained linear least-squares regression approach, allowed for abundance quantification of six immune cell types in each sample [19]. Differences in estimated immune cell abundance between age groups could not be distinguished by visual inspection of heatmaps for either cohort (Figure 1). The absolute differences in median between age groups in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort ranged only from 0.0065 for CD8+ T cells to 0.0715 for dendritic cells.

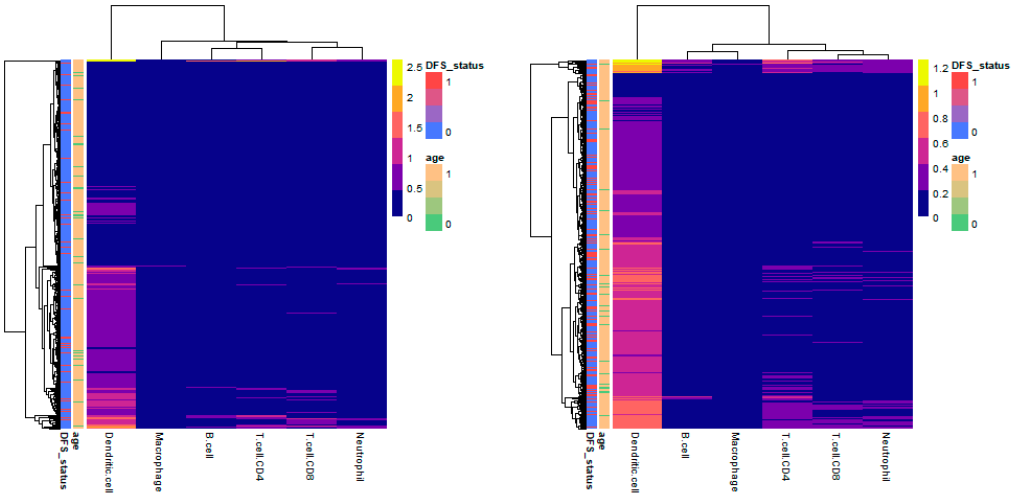

Figure 1. Heatmaps of abundance estimates for immune subsets predicted by computation deconvolution using the tumor immune estimation resource (TIMER) algorithm for the two cohorts: (**a**) The Cancer Genome Atlas (TCGA)-Breast Cancer (BRCA); and (**b**) Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). Row and column dendrograms show clustering of cases and cell types, respectively, according to Euclidean distance.

2.2. Immune Cell Type Abundance Significantly Associated with Disease-Free Survival

The Kaplan–Meier (KM) survival curve was used to associate abundance of each of the six immune cell types estimated by TIMER with disease-free survival (DFS) time and status separately for each age groups. Figure 2 shows that samples with high estimated CD8+ T cell abundance in The Cancer Genome Atlas (TCGA)-Breast Cancer (BRCA) cohort had significantly better prognosis (log-rank $p = 0.019$), which was also replicated in the METABRIC cohort (log-rank $p = 0.04$). Similar trends were also visible in the old age group; however, the differences in survival were less substantial between samples with high and low levels of CD8+ T cells. Figures S1 and S2 show results for all immune cell types from the TCGA-BRCA and METABRIC cohorts, respectively. For the other immune cell types, results were discordant between the two cohorts; however, macrophages seemed to consistently demonstrate little to no significant associations with DFS in the young age group in both the TCGA-BRCA (log-rank $p = 0.94$) and METABRIC (log-rank $p = 0.28$) cohorts.

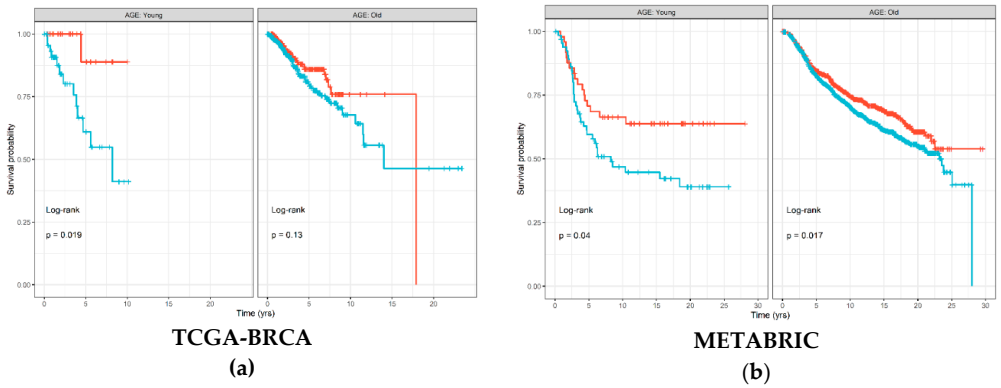

Figure 2. Disease-free survival Kaplan–Meier (KM) curve for: (**a**) TCGA-BRCA; and (**b**) METABRIC cohorts, grouped by age groups and stratified by high (red) and low (blue) CD8+ T cell levels estimated by TIMER and binarized by the maximally selected ranked statistics algorithm. Depicted *p*-values are from log-rank tests. p: *p-value*.

To quantify associations with survival, a Cox proportional hazards regression model was fit for each immune cell type to estimate a hazard ratio (HR), visualized as a forest plot for each age group and cohort in Figure 3. The trend across both cohorts for the young age group was that higher estimates of immune cell types resulted in lower HR, albeit with little significance, with the exception of macrophages. This trend, however, was not observed in samples from the old age groups. In particular, CD8+ T cells were consistently associated with better clinical outcome in the young age group in both TCGA-BRCA (HR 0.69; $p = 0.150$) and METABRIC (HR 0.81; $p = 0.110$) cohorts, as compared to the old age group. High levels of CD4+ T cells were also associated with lower HR for the young age group in both TCGA-BRCA (HR 0.58; $p = 0.150$) and METABRIC (HR 0.83; $p = 0.130$) cohorts but not in the old age group. Abundance of macrophages was consistently shown as having little relationship with survival in all samples.

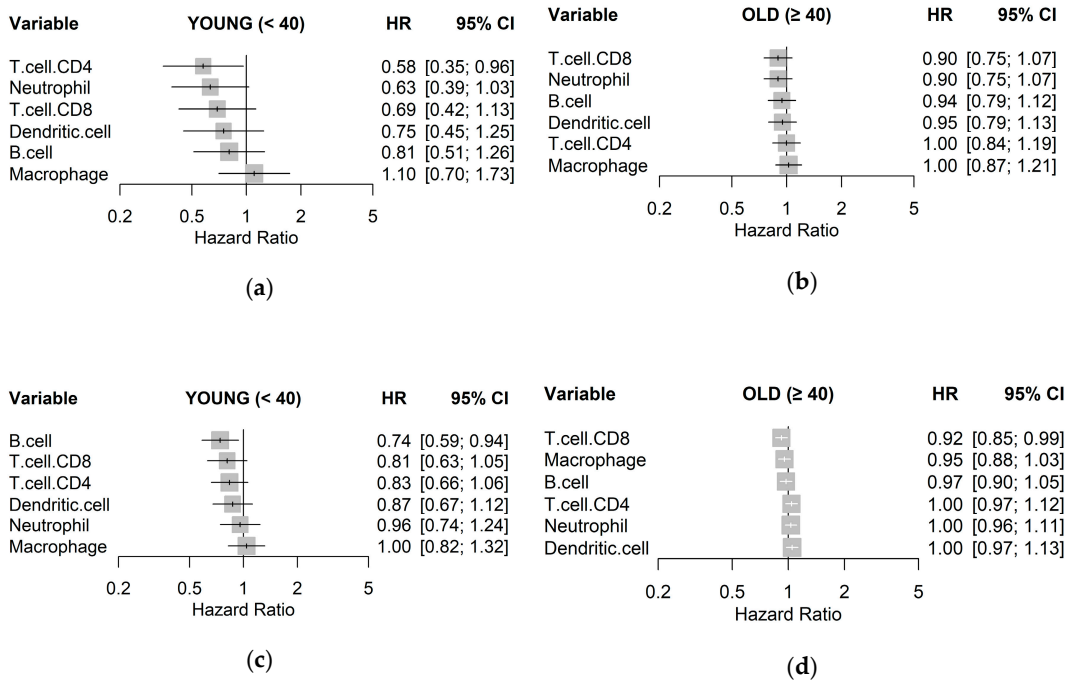

Figure 3. Unadjusted hazard ratios of each immune cell type quantified by TIMER, as individually estimated by univariable Cox regression models on disease-free survival for: (**a**) TCGA-BRCA young, (**b**) TCGA-BRCA old; (**c**) METABRIC young; and (**d**) METABRIC old cohorts with 95% confidence intervals (CIs).

2.3. The Mutational Signature Characteristic of High TILs Differs across Age Groups

We investigated whether mutational burdens from particular mutational signatures were significantly different between the age groups in the TCGA-BRCA cohort by analyzing the associated whole-exome sequencing (WES) data. Table S1 shows that contributions from signature 12 were significantly higher in the young age group (Mann–Whitney U, MWU $p = 0.00145$), with a median fold change of 1.48, as well as signature 14 (MWU $p = 0.0425$), with a median fold change of 2.18. None of the signatures were found to be significantly higher in the old age group compared to the young. Mutation data associated with the METABRIC cohort from targeted sequencing of 172 genes were also analyzed; however, results were largely discordant with TCGA-BRCA data. Reconstruction accuracy between estimated mutational burden and actual represented by mean cosine similarity was lower in the METABRIC (0.538) compared to TCGA-BRCA (0.775) cohorts, likely due to differences in the total number of mutations captured from sequencing.

The presence and abundance of TILs, especially CD8+ T cells, have been frequently and consistently indicated as an important factor to consider for prognosis and treatment of breast cancer [8,11,18]. Therefore, we referred to the abundance of CD8+ T cells estimated by TIMER as a measure of TILs for subsequent analyses. For each age group independently, we investigated whether contributions from particular mutational signatures were significantly different between high and low CD8+ T cell groups as estimated by TIMER in the TCGA-BRCA cohort. None of the mutational signatures were associated with TIL levels in the young age group. However, in the old age group, mutational contributions from signature 1 (MWU $p = 0.00504$; fold change 1.23), signature 2 (MWU $p = 0.0395$; fold change 0.77), signature 17 (MWU $p = 0.00319$; fold change 0.99), signature 26 (MWU $p = 0.0326$; fold change 2.02), and signature 30 (MWU $p = 0.0455$; fold change 1.36) were significantly associated with TIL levels estimated by TIMER.

2.4. Gene Set Enrichment for High TILs

Following our results from survival analyses which demonstrated this pattern in the young age group but not in the old, we sought to validate TIMER estimates by examining which gene sets from the Gene Ontology project were enriched in samples with high TIL independently for each age group. Table 1 shows the number of gene sets enriched for genes positively associated with TIL at various significance levels. At false discovery rate (FDR) $q < 0.05$, there was one overlapping positively enriched gene set between TCGA-BRCA and METABRIC cohorts for each age group: "cotranslational protein targeting to membrane" in the young age group, and "cilium movement" in the old age group. Gene set enrichment analysis (GSEA) results from the young TCGA-BRCA cohort showed enrichment of gene sets related to the mitochondria and cellular respiration, among several others (Table S2). In contrast, in the young age group from the METABRIC cohort, many of the enriched gene sets were related to the adaptive immune response, most notably T cell proliferation, selection, and regulation of cytotoxicity (Table S3). In the old age group from both cohorts, many of the top positively enriched gene sets were related to cilium assembly, movement, and ciliary transport (Tables S4 and S5). For enrichment maps of enriched gene sets for the young patient group in each cohort, refer to Figures 4 and 5.

Table 1. Number of positively enriched gene sets at different significance levels for each age group in each cohort from preranked gene set enrichment analysis (GSEA).

Age Group	Dataset	Number of Features	Number of Positive Gene Sets	Nominal $p < 0.01$	FDR $q < 0.25$	FDR $q < 0.05$	Number of Overlaps at FDR $q < 0.05$
Young	TCGA-BRCA	19,879	1370	126	135	33	1
Young	METABRIC	24,360	3765	246	204	30	
Old	TCGA-BRCA	20,201	2801	42	3	0	1
Old	METABRIC	24,360	3793	95	8	1	

[1] "Number of features" denotes the number of genes in the ranked gene list used as input for the analyses.

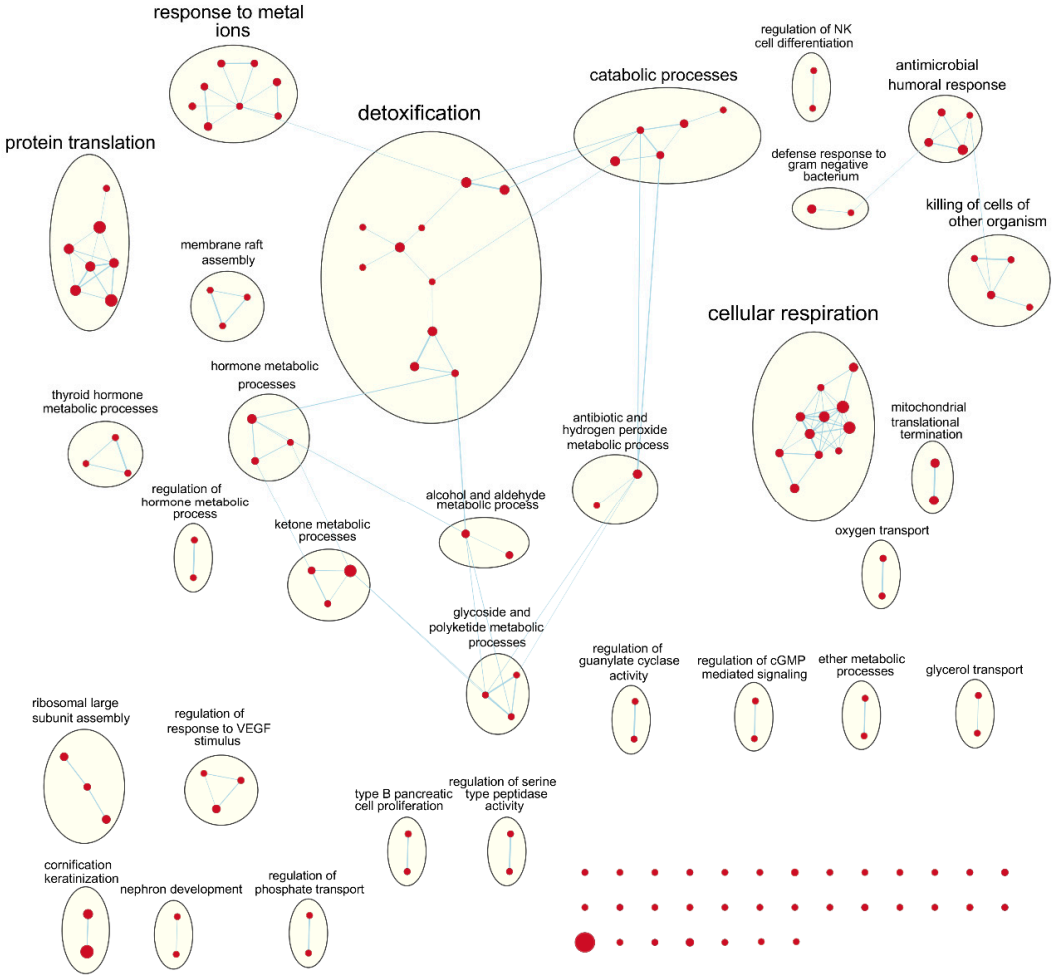

Figure 4. Enrichment map of results from preranked GSEA on the ranked gene list from the young TCGA-BRCA cohort. Nodes represent gene sets significant at FDR q-value < 0.25 and edges are drawn between nodes with similarity coefficient > 0.5. NK: natural killer; cGMP: cyclic guanosine monophosphate; VEGF: vascular endothelial growth factor.

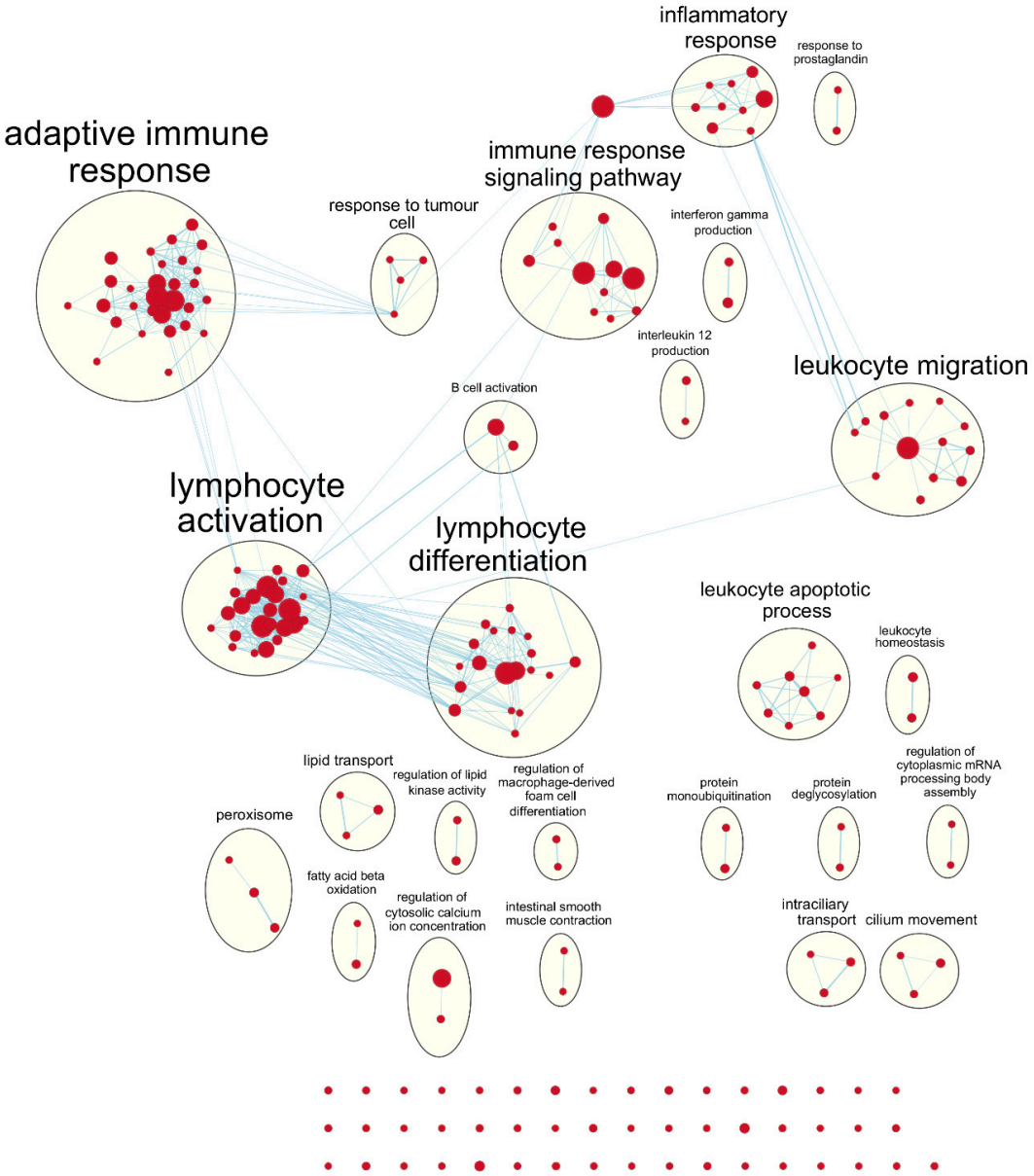

Figure 5. Enrichment map of results from preranked GSEA on the ranked gene list from the young METABRIC cohort. Nodes represent gene sets significant at FDR q-value < 0.25 and edges are drawn between nodes with similarity coefficient > 0.5. mRNA: messenger ribonucleic acid.

3. Discussion

Despite several reports in various parts of the world echoing the conclusion that young age at diagnosis is an indication for poor prognosis of breast cancer [4,6,20,21], researchers have been unsuccessful in identifying significant biological differences between young and old breast cancer patients. For example, Anders et al. found a number of genes that were differentially expressed between younger and older breast cancer patients, but none remained significant after correcting for subtype and other clinicopathological features [3,5]. Our results are consistent with previous findings in that we were also unable to uncover significant differences in immune cell compositions estimated by gene expression deconvolution between young and old patients.

Various immune subsets should decline with age in a phenomenon known as immunosenescence [22]. The older patient population was also expected to have greater myeloid potential and lesser lymphocyte potential due to involution of the thymus [22]. These observations

were not apparent in our analyses of differential immune profiles between age groups, which may be due to tumor-associated immune responses masking the global age-related changes in the immune system. Immunosenescence in women can also be attributed to deprivation of estrogen, which has immune-enhancing activities, during menopause [23]. In the datasets we used, 95.8% and 100% of the young patients (age at diagnosis <40) were pre-menopausal, whereas in the older patients (age at diagnosis ≥40) 20.3% and 16.5% were pre-menopausal in the TCGA-BRCA and METABRIC cohorts, respectively. Hence, menopausal state was a potential confounding variable in our study. Therefore, we fit a multiple linear regression model to further explore broader associations between estimated TIL, age at diagnosis (continuous variable), and menopausal state (categorical variable) (refer to Table S6). We found that TIL levels estimated by TIMER were significantly lower with older age at diagnosis (p-value: 0.0258 and 0.000282 for TCGA-BRCA and METABRIC cohorts, respectively). This finding is consistent with a recent meta-analysis exploring tumor-infiltrating lymphocytes and prognosis of early-stage triple-negative breast cancers [24]. However, menopausal state was not associated with TIL levels in both cohorts.

We found that specific immune subsets, in particular the CD8+ T cells, were significantly associated with disease-free survival in the young breast cancer patients but not in their older counterparts. This is in line with past studies that showed that higher infiltration of cytotoxic immune cells is indicative of better prognosis and greater odds of response to treatment [8,11]. Notably, Ali et al. showed that CD4+ and CD8+ T cells were more closely associated with favorable outcomes in ER-negative tumors than in ER-positive tumors [16]. Considering the fact that receptor status-negative tumors are much more common in the younger subset of breast cancer patients, our results are consistent with previous findings [4,7].

GSEA results showed that gene sets related to adaptive immune response and cytotoxic T cell processes were most obviously enriched in the young age group from the METABRIC cohort. On the other hand, gene sets related to the mitochondria and oxidative phosphorylation were positively enriched in the young age group from the TCGA-BRCA cohort. It may not be obvious, but mitochondria are important for T cell activation, proliferation, and differentiation because they are involved in processes such as immune synapse functions, production of reactive oxygen species, and various metabolic processes [25]. In the old age group from both cohorts, gene sets related to cilium assembly, movement, and ciliary transport were positively enriched for the high TIL phenotype, although with lower significance. Ciliary processes have recently been proposed to be involved in the formation of the immunological synapse between T cells and antigen-presenting cells, which is necessary for T cell activation and downstream functions [26,27]. This suggests that the TIMER algorithm may focus on different subset of genes for different datasets to provide estimates of immune cell composition.

Mutational signatures analysis showed that signatures 12 and 14 were significantly enriched in young compared to old breast cancer patients; however, the two signatures have not been previously associated with breast cancer [28,29]. Signature 12 was originally identified in liver cancer and is characterized by T > C substitutions with a transcriptional-strand-bias, which is indicative of being associated with transcription-coupled nucleotide excision repair [28]. Signature 14 was first identified in uterine cancer and is characterized by C > A and C > T substitutions, with a recent study suggesting that it is associated with microsatellite instability due to defects in mismatch repair [30]. However, our current knowledge on mutational signatures specific to breast cancer is likely reflective of the majority of breast cancer cases that are diagnosed in those over the age 40 (~93%). The landscape of mutational signatures specific to early onset breast cancer (~7%) has not been characterized previously. Further research is necessary to validate if the signatures previously not associated with breast cancer or novel signatures are responsible for mutations in the young subset of breast cancer patients. Within the young age group, none of the 30 single base substitution (SBS) signatures were significantly associated with TIL levels estimated by TIMER. However, within the old age group, mutational burdens from signatures 1, 26, and 30 were significantly higher in samples with high TIL levels, whereas signatures 2 and 17 were significantly associated with low TIL levels, although to varying degrees. Moreover,

mutational signatures that were significantly associated with TIL levels in the old age group were all previously reported in breast cancer patients, which once again suggests that TIL levels estimated by TIMER may also be indicative of tumor purity within samples.

TIMER has been frequently used on RNA-seq data from the TCGA database, since it was originally developed on TCGA data to work with bulk RNA-seq data [19,31,32]. However, it must be noted that the reference gene sets used in TIMER were curated from microarray gene expression data. To our knowledge, we are the first to employ the method for analyzing microarray data from the METABRIC cohort. Microarray gene expression data suffers from artifacts, such as limited profiling of genes with very low expression levels or saturation at very high expression levels, which poses potential challenges for data analysis compared to bulk RNA-seq data [19,33]. However, authors of TIMER method have stated that both RNA-seq and microarray data may be used as input to estimate the abundance of immune cell types in the tumor microenvironment [19,33]. They showed that TIMER is capable of producing highly concordant estimates between RNA-seq and microarray gene expression data generated from the same tumor samples [19]. This demonstrated that the algorithm used for deconvolution, constrained least-square regression (Methods: Equation (1)), is robust regardless of input data type [19,33].

TIMER only provides estimates for six broad immune cell types: B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells [19]. This can be advantageous because limiting number of cell types interrogated to those that are linearly separable prevents instability of estimates due to statistical co-linearity between cell types with very similar gene expression [19,32]. However, this is at the expense of resolution of the output. Even though they may originate from the same progenitor cell, different immune cell types and states can play vastly different roles within the tumor tissue; for example, M0 and M2 macrophages have traditionally been associated with pro-tumor responses whereas M1 macrophages have been associated with anti-tumor responses [7,18]. Our results from TIMER that found macrophages consistently insignificant to clinical outcomes in both cohorts may stem from the fact that TIMER cannot distinguish between the different activation states of macrophages. Despite its limitations, we were able to use TIMER as a tool to estimate the composition of immune cells within breast tumor samples and find specific immune subsets that showed significant trends with regards to disease-free survival in the young age group of breast cancer patients but not in the old.

4. Materials and Methods

4.1. Data

From The Cancer Genome Atlas (TCGA) database, gene-level RNA-seq expression data of patients with primary breast cancer (BRCA) were downloaded from Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) through TCGA-Assembler 2 [12,34]. The gene expression data had previously been processed by the RNAseqV2 pipeline, providing estimated counts and scaled estimate values from RNA-seq by Expectation Maximization (RSEM) [35]. Scaled estimates were converted to transcripts per million (TPM) values by multiplying by one million, resulting in a bulk RNA-seq dataset of 20,501 genes from 1095 breast tumor samples. We used RSEM-processed TPM measure because it was used previously to develop the TIMER method [19]. Mutations data from whole-exome sequencing (WES) and clinical data associated with the TCGA-BRCA cohort were downloaded from the cBioPortal for Cancer Genomics (https://www.cbioportal.org/) [36]. Samples from male patients (n = 12) were excluded from analyses, as well as any samples without associated age or survival data, resulting in final sample size of 989. There are inconsistencies in age thresholds used to define early onset breast cancer [1,4–6]. Here, we used a relatively conservative, but commonly used threshold (age at diagnosis = 40) to define it. Hence, of the 989 samples, 70 were identified as young/early-onset (age <40), and 919 were identified as old (age ≥40).

From the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort, gene-level microarray expression data and associated clinical data were downloaded from cBioPortal [13,36]. The gene expression data had previously been processed to log2 intensity values as described in the original publication [13]. Samples without associated age or survival data were removed, resulting in bulk microarray dataset of 24,360 genes from 1903 breast tumor samples, of which 116 were young and 1787 were old.

4.2. Immune Cell Type Deconvolution

Tumor immune estimation resource (TIMER) was used to deconvolute bulk gene expression data to estimate immune cell abundance [19]. In brief, most deconvolution algorithms see bulk transcriptome as a mixture where one gene of the mixture is a linear combination of that gene expressed across different cell types, weighted by the proportions of those cell types [15]. This can be represented by a simple Equation of a linear model:

$$y_i = \hat{\beta}_{i1}x_{i1} + \hat{\beta}_{i2}x_{i2} + \hat{\beta}_{i3}x_{i3} + \cdots + \hat{\beta}_{ik}x_{ik}\ , \tag{1}$$

where y_i is the gene expression level of a single gene i in the mixture data of n number of genes, x_1, x_2, $\ldots$, x_k are gene expressions of the one gene across k cell types, and $\hat{\beta}_1$, $\hat{\beta}_2$, $\ldots$, $\hat{\beta}_k$ are relative fractions of those cell types [15,19]. Given a signature expression matrix X, consisting of n rows for number of genes and k columns for number of cell types, the deconvolution algorithm produces estimates of cell fractions ($\hat{\beta}$). Specifically, TIMER utilizes constrained linear regression with non-negativity constraint on cell fractions to estimate the abundance of six immune cell types: B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells [19]. For settings, the sample tumor type was set to "BRCA" for breast cancer, and the default reference data and gene set were used for both cohorts.

The TIMER method was available for quick implementation through the immunedeconv package [15,19]. Expression data with rows as genes and columns as samples was provided to the algorithm as input, with which the algorithm estimated abundance of six immune cell types. Results were visualized as heatmaps and violin plots using the pheatmap [37] and ggpubr [38] packages, respectively.

4.3. Survival Analyses

Kaplan–Meier (KM) survival curves were visualized using survival [39] and survminer packages [40]. Maximally selected rank statistics (maxstat) as implemented in the survminer package was used to determine the optimal cut-off to binarize the immune cell abundance estimates from TIMER into "high" and "low" groups which produces the maximum log-rank statistic on disease-free survival (DFS) [41].

Cox regression was used to compute hazard ratios for each of the six immune cell types for which abundance was estimated by TIMER. Analyses were conducted separately for each age group to consider the potential violation of the Cox proportional-hazards assumption [16]. Similar to previous approach by Ali et al., estimates of immune cell abundance were converted to quartiles from 1–4 [16]. Results were visualized using the meta package [42].

4.4. Single Base Substitution Mutational Signatures

From the TCGA-BRCA cohort, 981 samples had mutation data available. Number of mutations for each of 96 possible single base substitution (SBS) types were summed using deconstructSigs package [43]. With the resulting matrix of SBS counts by samples, contributions of each of the 30 SBS signatures from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [44] were estimated for each sample using the R package MutationalPatterns [45]. Instead of excluding SBS signatures previously not extracted from breast cancer cohorts, all 30 signatures were incorporated in our analysis because the young subset of breast cancer patients is largely underrepresented in many cohorts that

have been used to extract those signatures [27]. Samples with pairwise cosine similarity between the reconstructed mutation counts matrix and the original input less than 0.5 were excluded (n = 23). Samples without gene expression or relevant clinical data were also excluded, resulting in 858 samples, of which 57 were young and 801 were old. Estimated contribution from each signature was compared by Mann–Whitney U (MWU) test between young and old age groups [27].

4.5. Gene Set Enrichment Analysis

To find genes positively associated with CD8+ T cell abundance estimated by TIMER in each cohort, simple linear regression models were fit for expression of each gene separately, with expression values treated as the continuous predictor variable and TIMER estimates as the continuous outcome variable. Models were fit separately for each age group, and the t-statistic value for each of the resulting regression coefficients were calculated as a measure of association to the CD8+ T cell abundance estimates. The list of genes ordered by decreasing t-statistic values for each age group was used as input for pre-ranked gene set enrichment analysis (GSEA) on the C5:BP (Gene Ontology biological processes) gene sets downloaded from the Molecular Signatures Database (MSigDB; http://software.broadinstitute.org/gsea/msigdb/index.jsp) [46,47]. Gene sets of size < 2 and > 500 were excluded, resulting in total of 7094 gene sets evaluated in the analysis. Results from GSEA were visualized by Cytoscape version 3.7.2 [48] using the Enrichment Map plugin [49]. Enriched gene sets with FDR q value < 0.25 were represented as nodes and any overlaps > 0.3 between nodes were represented as edges in the resulting network diagram. Nodes were grouped together and labelled by the clusterMaker2 [50], AutoAnnotate [51], and WordCloud [52] plugins, and resulting annotations were manually corrected.

4.6. Data Analysis Software

All analyses and visualizations were performed in R Project for Statistical Computing version 3.6.1 [53] and RStudio version 1.2.1335(Boston, MA, United States) [54]. Unless otherwise noted, all statistical analyses were conducted using the stats package [53], and two-sided $p < 0.05$ was considered significant.

5. Conclusions

In summary, we determined that a particular immune cell type, the cytotoxic CD8+ T cell, was significantly associated with disease-free survival in young breast cancer patients under the age 40, but not in their older counterparts. Furthermore, our analyses showed that single base substitution mutational signatures 12 and 14 were significantly enriched in the young patient group compared to the old, that were not previously associated with breast cancer. We also highlighted some potential limitations of our data and methods, especially with TIMER and its lack of resolution. Nonetheless, our work suggests that the underlying biological differences may stem from more abstract relationships involving multiple levels of tumor physiology and not age alone.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/12/5/1076/s1. Figure S1: 10-year disease-free survival Kaplan–Meier curves for TCGA-BRCA cohort; Figure S2: 10-year disease-free survival Kaplan–Meier curves for the METABRIC cohort; Table S1: Differences in mutational burden from 30 SBS signatures from COSMIC between young and old age groups in TCGA-BRCA cohort; Table S2: Preranked GSEA results for the gene list ranked by correlation with TIL levels in the young age group from TCGA-BRCA cohort; Table S3: Preranked GSEA results for the gene list ranked by correlation with TIL levels in the young age group from the METABRIC cohort; Table S4: Preranked GSEA results for the gene list ranked by correlation with TIL levels in the old age group from the TCGA-BRCA cohort; Table S5: Preranked GSEA results for the gene list ranked by correlation with TIL levels in the old age group from the METABRIC cohort; Table S6: Coefficients and statistical results of multiple linear regression between the response variable: TIL level estimated by TIMER, and predictor variables: age at diagnosis, and menopausal state.

Author Contributions: Conceptualization, P.H.; methodology, P.H. and Y.W.J.; software, Y.W.J.; validation, P.H. and Y.W.J.; formal analysis, Y.W.J.; investigation, Y.W.J.; resources, P.H.; data curation, Y.W.J.; writing—original draft preparation, Y.W.J.; writing—review and editing, Y.W.J. and P.H.; visualization, Y.W.J.; supervision, P.H.;

project administration, P.H.; funding acquisition, P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Canadian Breast Cancer Foundation, Natural Sciences and Engineering Research Council of Canada, and the University of Manitoba.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Johnson, R.H.; Anders, C.K.; Litton, J.K.; Ruddy, K.J.; Bleyer, A. Breast cancer in adolescents and young adults. *Pediatr. Blood Cancer* **2018**, *65*, e27397. [CrossRef] [PubMed]
2. Dai, X.; Li, T.; Bai, Z.; Yang, Y.; Liu, X.; Zhan, J.; Shi, B. Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.* **2015**, *5*, 2929–2943. [PubMed]
3. Anders, C.K.; Fan, C.; Parker, J.S.; Carey, L.A.; Blackwell, K.L.; Klauber-DeMore, N.; Perou, C.M. Breast carcinomas arising at a young age: Unique biology or a surrogate for aggressive intrinsic subtypes? *J. Clin. Oncol.* **2011**, *29*, 18–20. [CrossRef]
4. Anders, C.K.; Hsu, D.S.; Broadwater, G.; Acharya, C.R.; Foekens, J.A.; Zhang, Y.; Wang, Y.; Marcom, P.K.; Marks, J.R.; Nevins, J.R.; et al. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J. Clin. Oncol.* **2008**, *26*, 3324–3330. [CrossRef] [PubMed]
5. Johnson, R.H.; Hu, P.; Fan, C.; Anders, C.K. Gene expression in 'young adult type' breast cancer: A retrospective analysis. *Oncotarget* **2015**, *6*, 13688–13702. [CrossRef]
6. Martínez, M.T.; Oltra, S.S.; Peña-Chilet, M.; Alonso, E.; Hernando, C.; Burgues, O.; Chirivella, I.; Bermejo, B.; Lluch, A.; Ribas, G. Breast Cancer in Very Young Patients in a Spanish Cohort: Age as an Independent Bad Prognostic Indicator. *Breast Cancer Basic Clin. Res.* **2019**, *13*, 1178223419828766. [CrossRef]
7. Pruneri, G.; Vingiani, A.; Denkert, C. Tumor infiltrating lymphocytes in early breast cancer. *Breast* **2018**, *37*, 207–214. [CrossRef]
8. Ziai, J.; Gilbert, H.N.; Foreman, O.; Eastham-Anderson, J.; Chu, F.; Huseni, M.; Kim, J.M. CD8+ T cell infiltration in breast and colon cancer: A histologic and statistical analysis. *PLoS Med.* **2018**, *13*, e0190158.
9. Wang, K.; Shen, T.; Siegal, G.P.; Wei, S. The CD4/CD8 ratio of tumor-infiltrating lymphocytes at the tumor-host interface has prognostic value in triple-negative breast cancer. *Hum. Pathol.* **2017**, *69*, 110–117. [CrossRef]
10. Yang, X.; Ren, H.; Sun, Y.; Shao, Y.; Zhang, L.; Li, H.; Zhang, X.; Yang, X.; Yu, W.; Fu, J. Prognostic significance of CD4/CD8 ratio in patients with breast cancer. *Int. J. Clin. Exp. Pathol.* **2017**, *10*, 4787–4793.
11. Ali, H.R.; Provenzano, E.; Dawson, S.J.; Blows, F.M.; Liu, B.; Shah, M.; Bowden, S.J. Association between CD8+ T-cell infiltration and breast cancer survival in 12,439 patients. *Ann. Oncol.* **2014**, *25*, 1536–1543. [CrossRef]
12. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490*, 61–70. [CrossRef]
13. Curtis, C.; Shah, S.P.; Chin, S.F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Yuan, Y.; Gräf, S.; et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **2012**, *486*, 346–352. [CrossRef] [PubMed]
14. Clare, S.E.; Shaw, P.L. "Big Data" for breast cancer: Where to look and what you will find. *NPJ Breast Cancer* **2016**, *2*, 16031. [CrossRef]
15. Sturm, G.; Finotello, F.; Petitprez, F.; Zhang, J.D.; Baumbach, J.; Fridman, W.H.; List, M.; Aneichyk, T. Comprehensive evaluation of computational cell-type quantification methods for immuno-oncology. *Bioinformatics* **2019**. [CrossRef]
16. Ali, H.R.; Chlon, L.; Pharoah, P.D.P.; Markowetz, F.; Caldas, C. Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study. *PLoS Med.* **2016**, *13*, e1002194. [CrossRef]
17. O'Meara, T.; Safonov, A.; Casadevall, D.; Qing, T.; Silber, A.; Killelea, B.; Hitzis, C.; Pusztai, L. Immune microenvironment of triple-negative breast cancer in African-American and Caucasian women. *Breast Cancer Res. Treat.* **2019**. [CrossRef]

18. Nagarajan, D.; Mcardle, S.E.B. Immune Landscape of Breast Cancers. *Biomedicines* **2018**, *6*, 20. [CrossRef]
19. Li, B.; Severson, E.; Pignon, J.C.; Zhao, H.; Li, T.; Novak, J.; Jiang, P.; Shen, H.; Aster, J.C.; Signoretti, S.; et al. Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **2016**, *17*, 174. [CrossRef]
20. Anders, C.K.; Johnson, R.; Litton, J.; Phillips, M.; Bleyer, A. Breast Cancer before Age 40 Years. *Semin. Oncol.* **2009**, *36*, 237–249. [CrossRef]
21. Zeeshan, S.; Ali, B.; Ahmad, K.; Chagpar, A.B.; Sattar, A.K. Clinicopathological Features of Young Versus Older Patients With Breast Cancer at a Single Pakistani Institution and a Comparison With a National US Database. *J. Glob. Oncol.* **2019**, *5*, 1–6. [CrossRef] [PubMed]
22. Montecino-rodriguez, E.; Berent-maoz, B.; Dorshkind, K. Causes, consequences, and reversal of immune system aging. *J. Clin. Investig.* **2013**, *123*, 958–965. [CrossRef]
23. Gameiro, C.M.; Romao, F.; Castelo-Branco, C. Menopause and aging: Changes in the immune system—A review. *Maturitas* **2010**, *67*, 316–320. [CrossRef]
24. Loi, S.; Drubay, D.; Adams, S.; Pruneri, G.; Francis, P.A.; Lacroix-Triki, M.; Joensuu, H.; Badve, S.; Demaria, S.; Gray, R.; et al. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. *J. Clin. Oncol.* **2019**, *37*, 559–569. [CrossRef] [PubMed]
25. Desdín-Micó, G.; Soto-Heredero, G.; Mittelbrunn, M. Mitochondrial activity in T cells. *Mitochondrion* **2018**, *41*, 51–57. [CrossRef] [PubMed]
26. Le Borgne, M.; Shaw, A.S. Do T cells have a cilium? *Science* **2013**, *342*, 1177–1178. [PubMed]
27. Cassioli, C.; Baldari, C.T. A Ciliary View of the Immunological Synapse. *Cells* **2019**, *8*, 789. [CrossRef]
28. Nik-Zainal, S.; Davies, H.; Staaf, J.; Ramakrishna, M.; Glodzik, D.; Zou, X.; Martin, S.; Wedge, D.C.; Smid, M.; Van Loo, P.; et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **2016**, *534*, 47–54. [CrossRef]
29. Smid, M.; Rodríguez-González, F.G.; Sieuwerts, A.M.; Salgado, R.; Prager-Van der Smissen, W.J.; van Der Vlugt-Daane, M.; van Galen, A.; Staaf, J.; Brinkman, A.B.; Van de Vijver, M.J.; et al. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nat. Commun.* **2016**, *7*, 12910. [CrossRef]
30. Alexandrov, L.B.; Kim, J.; Haradhvala, N.J.; Huang, M.N.; Ng, A.W.; Wu, Y.; Islam, S.A. The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101. [CrossRef]
31. Kim, I.S.; Gao, Y.; Welte, T.; Wang, H.; Liu, J.; Janghorban, M.; Sheng, K.; Niu, Y.; Zhao, N.; Bado, I.; et al. Immuno-subtyping of breast cancer reveals distinct myeloid cell profiles and immunotherapy resistance mechanisms. *Nat. Cell Biol.* **2019**, *21*, 1113–1126. [CrossRef] [PubMed]
32. Wang, K.; Chen, R.; Feng, Z.; Zhu, Y.; Sun, X.; Huang, W.; Chen, Z. Identification of differentially expressed genes in non-small cell lung cancer. *Aging* **2019**, *11*, 11170–11185. [CrossRef]
33. Li, B.; Li, T.; Liu, J.S.; Liu, X.S. Computational deconvolution of tumor-infiltrating immune components with bulk tumor gene expression data. In *Bioinformatics for Cancer Immunotherapy, Methods in Molecular Biology*; Boegel, S., Ed.; Humana: New York, NY, USA, 2020; Volume 2120, pp. 249–262. [CrossRef]
34. Wei, L.; Jin, Z.; Yang, S.; Xu, Y.; Zhu, Y.; Ji, Y. TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics* **2018**, *34*, 1615–1617. [CrossRef]
35. Gao, F.G.; Parker, S.J.; Reynolds, M.S.; Silva, C.T.; Wang, B.L.; Zhou, W.; Akbani, R.; Bailey, M.; Balu, S.; Brooks, D.; et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst.* **2019**, *9*, 24–34.e10. [CrossRef]
36. Gao, J.; Aksoy, A.B.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, O.S.; Sun, Y.; Jacobson, A.; Sinha, R.; Cerami, E.; et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **2013**, *6*, pl1. [CrossRef]
37. Kolde, R. Pheatmap: Pretty Heatmaps. 2019. R Package Version 1.0.12. Available online: https://CRAN.R-project.org/package=pheatmap (accessed on 20 April 2020).
38. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. 2019. R Package Version 0.2.5. Available online: https://CRAN.R-project.org/package=ggpubr (accessed on 20 April 2020).
39. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model*; Springer: New York, NY, USA, 2000; ISBN 0-387-98784-3.

40. Kassambara, A.; Kosinski, M.; Biecek, P. Survminer: Drawing Survival Curves Using 'ggplot2'. 2019. R Package Version 0.4.6. Available online: https://CRAN.R-project.org/package=survminer (accessed on 20 April 2020).
41. Lausen, B.; Schumacher, M. Maximally Selected Rank Statistics. *Biometrics* **1992**, *48*, 73–85. [CrossRef]
42. Schwarzer, G. Meta: An R package for meta-analysis. *R. News* **2007**, *7*, 40–45.
43. Rosenthal, R.; McGranahan, N.; Herrero, J.; Taylor, B.S.; Swanton, C. Deconstructsigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **2016**, *17*, 31. [CrossRef] [PubMed]
44. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Fish, P. COSMIC: The Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **2019**, *47*, D941–D947. [CrossRef] [PubMed]
45. Blokzijl, F.; Janssen, R.; van Boxtel, R.; Cuppen, E. MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med.* **2018**. [CrossRef]
46. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Mesirov, J.P.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
47. Reimand, J.; Isserlin, R.; Voisin, V.; Kucera, M.; Tannus-Lopes, C.; Rostamianfar, A.; Wadi, L.; Meyer, M.; Wong, J.; Merico, D.; et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **2019**, *14*, 482–517. [CrossRef]
48. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
49. Merico, D.; Isserlin, R.; Stueker, O.; Emili, A.; Bader, G.D. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **2010**, *5*. [CrossRef] [PubMed]
50. Morris, J.H.; Apeltsin, L.; Newman, A.M.; Baumbach, J.; Wittkop, T.; Su, G.; Bader, G.D.; Ferrin, T.E. clusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinforma.* **2011**, *12*, 436. [CrossRef] [PubMed]
51. Kucera, M.; Isserlin, R.; Arkhangorodsky, A.; Bader, G.D. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations [version 1; peer review: 2 approved]. *F1000Research* **2016**, *5*. [CrossRef]
52. Oesper, L.; Merico, D.; Isserlin, R.; Bader, G.D. WordCloud: A Cytoscape plugin to create a visual semantic summary of networks. *Source Code Biol. Med.* **2011**, *6*, 7. [CrossRef]
53. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2019.
54. RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio Team: Boston, MA, USA, 2018.