OPEN

Breast tumor subgroups reveal diverse clinical prognostic power

Zhaoqi Liu, Xiang-Sun Zhang & Shihua Zhang

National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

Predicting the outcome of cancer therapies using molecular features and clinical observations is a key goal of cancer biology, which has been addressed comprehensively using whole patient datasets without considering the effect of tumor heterogeneity. We hypothesized that molecular features and clinical observations have different prognostic abilities for different cancer subtypes, and made a systematic study using both clinical observations and gene expression data. This analysis revealed that (1) gene expression profiles and clinical features show different prognostic power for the five breast cancer subtypes; (2) gene expression data of the normal-like subgroup contains more valuable prognostic information and survival associated contexts than the other subtypes, and the patient survival time of the normal-like subtype is more predictable based on the gene expression profiles; and (3) the prognostic power of many previously reported breast cancer gene signatures increased in the normal-like subtype and reduced in the other subtypes compared with that in the whole sample set.

C ancer genome aberrations observed through clinical and basic research have been used to categorize patients in an effort to improve clinical decision-making and develop more effective treatments. Although such grouping methods have improved treatment efficacy of many different cancers, overcoming heterogeneity within these populations is a major challenge. With the advent of high-throughput genomic technologies, many molecular-based diagnostics have been developed, and several have recently gained regulatory approval[1,2]. Many of these diagnostics are applicable to breast cancer, and suggest that individual molecular diagnostics for therapeutic strategies may provide objective, precise, and systematic prediction of clinical outcomes.

Breast cancer is no longer viewed as a single disease; rather, it is heterogeneous consisting of different subtypes on the molecular, histopathological, and clinical level with different prognostic and therapeutic implications[3–6]. Gene expression profiling has classified breast cancer into five biologically distinct intrinsic subtypes: luminal A, luminal B, HER2-enriched (HER2+), basal-like, and normal-like[3–5]. The luminal A and B subtypes are ER-positive, and luminal B is associated with a relatively worse outcome. Both HER2+ and basal-like breast cancers have poor outcomes. Parker et al.[6] developed an efficient classifier, called PAM50, to distinguish these five intrinsic subtypes using the expression of 50 "classifier genes". In a more recent study, a large breast cancer patient cohort (n ∼ 2000) was clustered into 10 molecularly defined subgroups with apparently distinct biology and disease-specific survival characteristics[7]. In addition, different breast cancer subtypes have different treatment responses[8,9]. For example, the basal-like and HER2+ subtypes are more sensitive to paclitaxel- and doxorubicin-containing preoperative chemotherapy than the luminal and normal-like cancers[8]. Another study suggested that the different molecular subtypes of breast cancer could be characterized by distinct response rates to neoadjuvant chemotherapy using a taxane and anthracycline-containing regimen[9].

The molecular heterogeneity among breast tumors suggests that respective stratified therapy and clinical prediction of prognosis would be beneficial. Here, patients within particular subtypes would be handled with special subtype-specific treatments[10]. Among the five intrinsic subtypes, basal-like breast cancer is of particular clinical interest because of its high frequency, poor prognosis, and its tendency to affect younger women[11]. Moreover, because this subtype lacks expression of estrogen receptor (ER), progesterone receptor (PR) and HER2, the basal-like breast cancers do not benefit from anti-estrogen hormonal therapies or trastuzumab. Although this subtype does benefit from chemotherapy, less toxic and more targeted treatment options are necessary[12]. Several molecular-based studies have focused on basal-like or triple negative breast cancers[11–16]. For example, Hassall et al.[14] identified a 14-gene signature to distinguish the basal-like subtype into two subgroups. They argued that this categorization would guide aggressive therapeutic regimens to the poor prognosis subgroup, and conversely avoid such therapy in low risk patients. In contrast to the tough basal-like subtype,
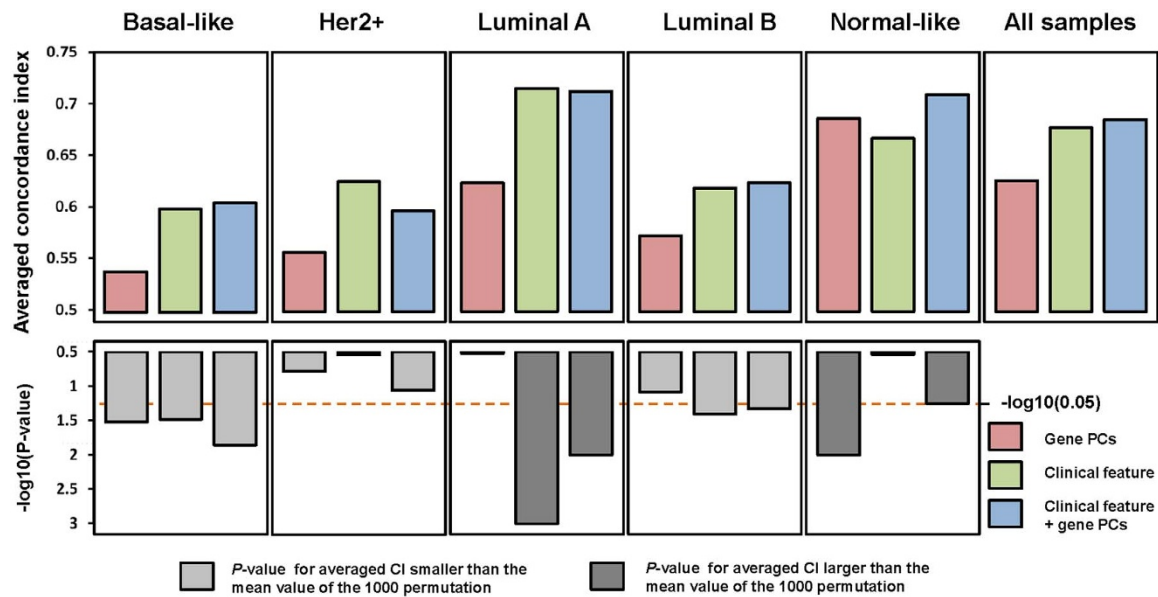
Figure 1 | Breast cancer subtypes show different prediction performance. Bar graphs show averaged CIs of the three cases of multivariate Cox PH model on the five breast cancer subtypes and the METABRIC whole dataset. Red and green bars represent the gene PCs case and clinical feature case respectively and blue bars represent these two variables together. P-values are from the permutation test. The red dashed line marks the significance level. P-values for corresponding averaged CIs larger and smaller than the mean value of the 1000 permutation results are assigned with different colors.

researchers have gained clinical success on HER2+ breast cancers because of effective therapeutic targeting of HER2. The presence of amplification of the HER2 gene confers sensitivity to the targeted chemotherapeutic agent herceptin (trastuzumab)[17].

In an effort to guide the selection of the most appropriate therapy for individual patients, numerous prognostic gene expression signatures have been reported[1,2,18–21]. One of the early studied signatures, called MammaPrint[1], is a commercially available microarray-based diagnostic, which evaluates the expression of 70 genes. More recently, OncotypeDX[2], a 21-gene quantitative RT-qPCR assay, was developed and predicts the risk of distant recurrence in tamoxifen-treated, node-negative breast cancers and their responsiveness to CMF chemotherapy[18]. Whereas many multi-gene signatures exist, Venet et al.[21] found that the prognostic abilities of many published breast cancer gene signatures are derived from their strong correlation to expression of genes associated with proliferation. Thus, this group developed a 131-gene proliferation-related signature called meta-PCNA[21]. In addition, Wu and Stein[22] recently proposed a network module-based method for identifying cancer prognostic signatures and discovered a novel 31-gene signature, which outperformed 48 published breast cancer gene signatures.

Given the limited patient number for many of these studies, the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) provided an unprecedented resource[7] which contains a large breast cancer patient cohort of ~2000 samples with detailed clinical measurements and genome-wide molecular profiles including gene expression and copy number variation data of the molecular patterns inside tumors. In line with this, Sage Bionetworks launched a competition called DREAM Breast Cancer Prognosis Challenge[23]. The goal of this competition is to assess the accuracy of computational models, like METABRIC, that use comprehensive molecular profiling data and clinical information to predict patient survival. However, we found that the molecular features such as gene expression only moderately improve the clinical prediction with regards to the whole cancer cohort.

Taking into account the heterogeneity of breast cancers and the subtype specific molecular signatures, we hypothesized that making clinical predictions for five subtypes separately would provide better prediction performance. In this study, we adopted PAM50 to identify the five heterogeneous breast cancer subtypes, and then systematically evaluated patient survival prediction performance on these subgroups using both clinical observations and the gene expression profiles of the METABRIC dataset. Then, we applied a network module-based cancer prognostic signature identification method on each subtype to search for network biomarkers to further demonstrate the differences of prediction performance.

Results

Given the heterogeneity among breast cancer patients, we sought to determine whether significant differences exist in different subgroups and whole cancer cohort for predicting survival time based on molecular and clinical data. To this end, we adopted the five PAM50 tumor groups defined in the METABRIC dataset for analysis. We then applied the Cox model to gene expression covariates, clinical feature covariates or the combination of these two for each of the five intrinsic subtypes and the METABRIC whole dataset respectively (see Materials and Methods, and Supplementary Information).

Breast cancer subtypes show different prognostic performance. As mentioned above, after defining the subgroups, we applied the multivariate Cox proportional hazards (multivariate Cox PH) model on different breast cancer subtypes. This analysis revealed significant differences in prognostic performance (Figure 1 and Supplementary Table S1). We also observed similar differences using a random survival forest model in the same manner (Supplementary Table S2). First, consistent with the DREAM Breast Cancer Prognosis Challenge, the multivariate Cox PH model using clinical feature covariates alone demonstrated performance comparable to the combination of clinical feature and gene expression data based on the whole population. Generally, the clinical feature covariates were more informative for predicting patient survival time than the gene-expression covariates on the five breast cancer subtype datasets and the whole dataset. The exception, however, was the normal-like tumor subtype. In addition, the predictive power of using both covariates together versus clinical features was increased slightly except for HER2+ (0.597 vs.0.625) and luminal A (0.712 vs.0.715)

tumors. These data suggests that the inclusion of gene expression only improves the prediction performance very limitedly.

Both the basal-like and HER2+ subgroups had poor survival prediction performance, while the luminal A and normal-like breast cancer subgroups demonstrated better performance. The clinical outcome of luminal A and normal-like subtypes, based on genetic and clinical covariates, was more predictable than the other three subgroups. Particularly, all three Cox models underperformed when applied to basal-like breast cancer. This observation is consistent with previous studies that the basal-like tumor subtype, referred to as triple-negative breast cancer in some literature, is associated with a particularly poor prognosis[4,5]. Compared with the basal-like subgroup, the HER2+ subgroup had a relatively weak prognosis without significance for all three models. In addition, the concordance index (CI) dropped when including the three gene principal components (PCs) to the clinical covariates on the Cox model. It is possible that the three PCs are not representative of the survival-related information hidden in the expression matrix for this subgroup.

Compared with luminal A tumors, luminal B breast cancer was associated with a worse outcome, which is consistent with previously published results[4,5]. The ER-positive and histological low-grade luminal A tumors have the highest significant averaged CIs using either clinical covariates alone (0.715, p-value = 0.001) or combined with gene expression covariates (0.712, p-value = 0.01). This result suggests that the clinical features of luminal A breast cancer are the most relevant to patient survival time among the five subgroups, and have the best prognostic power. The selected clinical covariates with the smallest p-values (Wald test) in the Cox model tested on the luminal A subtype were patient age at diagnosis, lymph node assessment, HER2 SNP6 state, and treatment received. The attendance of diagnosis age was probably due to our use of overall survival time as the patient survival outcome.

The normal-like breast cancer subtype (n = 200) was of particular interest because the high expression-based model score (0.686, p-value = 0.01) compared to the other four subtypes and the METABRIC whole dataset. In addition, unlike the other subgroups, gene expression covariate analysis of normal-like breast cancer was more predictive of patient survival than clinical features covariate analysis (0.686 vs. 0.667).

Taken together, these results demonstrate that 1) the predictive ability of the same method on different cancer subtypes was diverse, 2) gene expression data can improve the predictive performance to different degrees, and 3) the predictive power of gene expression covariates on the normal-like subtype was more informative than that on other subgroups, suggesting that gene expression data of normal-like subgroup contains more prognostic value for this subgroup than the other breast cancer subtypes. Further research should be conducted on this promising subtype-specific molecular dataset with the purpose of providing effective and reliable support for clinical diagnosis and personalized medicine treatment.

Given that gene expression data improves prediction performance to varying degrees, we further examined the potential prognostic power of gene expression covariates for the different breast cancer subtypes. To this end, the CI between the expression value and the survival time of the patients for each gene probe was calculated and used as a measure of its prognostic power. Then, all 49576 probes were ranked in terms of their CIs calculated for each subtype as well as the METABRIC whole dataset. This analysis also revealed that the prediction performance using gene expression profiles was diverse for each breast cancer subgroup (Figure 2A). Notably, the right tail of the CI distribution of the normal-like tumors extends beyond those of other subgroups. This demonstrates that gene expression profiles of normal-like tumors contain more prognostic probes. This is supported by the fact that 1299 gene probes have CIs larger than 0.6 for normal-like tumors, while only a few probes have CI values this high in other subgroups. In addition, the highest CI observed in

normal-like tumors was 0.688, whereas the highest score in other subgroups was 0.618. Thus, many prognostic genes show strong relevance between the expression value and patient survival in the normal-like tumors subgroup. This analysis confirms our observation using multivariate Cox PH that normal-like breast cancer benefits more from gene-expression data based predictions than other breast cancer subtypes.

We next explored the overlaps of the prognostic probes obtained on each subgroup with those obtained on the METABRIC and OsloVal datasets respectively (Figure 2B, C, D). Using this comparison, we found that normal-like breast cancers share 30 of the top 100 prognostic probes of the METABRIC whole dataset. The luminal B subtype was the next closest with only 12 common prognostic probes (Figure 2B). We also found that the prognostic probes of the five subtypes show a significant different performance on the METABRIC and OsloVal datasets respectively (Figure 2C, D). The top 100 prognostic probes obtained from the normal-like subtype had significantly higher CIs using both datasets compared those of the four other subtypes (Kruskal-Wallis test p-value < 0.0001 on METABRIC dataset and p-value < 0.05 on OsloVal dataset). Therefore, the prognostic gene probes in the normal-like subtype were the most consistent with those on the whole dataset. In addition, prognostic genes from the METABRIC whole dataset tend to be informative for prediction on the normal-like subtype, but not for other subtypes. Together, these observations suggest that the prognostic ability of a gene predictor on the whole breast cancer dataset may be due to its relative higher prognostic power on a special subtype.

Identifying subgroup-specific prognostic network gene modules. The analysis above revealed that gene expression data in different tumor subgroups demonstrate diverse prognostic power. Thus, we hypothesized that the functional network of biomarkers defined for breast cancer are only biologically meaningful for a set of tumor subgroups. To verify this hypothesis, we applied a recently published method[22] to the current five PAM50 subtypes and the METABRIC whole dataset to identify network biomarkers or gene modules that were significantly correlated with patient survival (see Materials and Methods). We found that six network modules obtained from the normal-like subgroup had CIs greater than 0.6, while only one in the basal-like tumors had a CI of 0.605. In addition, no modules satisfy this for all other subtypes and the whole tumor dataset (Figure 3 and Table 1). We named each module in terms of the name of the gene having the highest CI in the module. The BIRC5 module consisting of 25 genes generated from the normal-like subtype achieved the highest CI of 0.650. The second prognostic network module (MCM10 module) from normal-like subtype was comprised of 18 genes and had a CI of 0.640.

Based on the module overlap analysis, we extracted module biomarkers demonstrating significant overlaps with the BRIC5 module among those obtained from datasets of other subtypes. The NSUN2 module consisting of 12 genes from luminal A had the largest overlap with 11 genes also present in the BIRC5 module. However, it only gets a CI of 0.547 on the luminal A tumor. Other modules overlapping with more than one of the 25 genes had CIs lower than 0.55. Therefore, the BRIC5 module was only detected in the normal-like tumors, and it is highly correlated with its overall patient survival. Moreover, a permutation test revealed a p-value of 0.0025, suggesting that the BIRC5 module from the normal-like tumor data was not found by chance (Materials and Methods).

The BIRC5 module contained 25 genes and 9 of them had CIs larger than 0.60 (Figure 4A and B). The gene BIRC5 had the highest CI (0.667) among the 25 genes. The prognostic relevance of the BIRC5 module with normal-like tumors and all breast cancer tumors (using the METABRIC whole dataset and the OsloVal dataset respectively) was demonstrated by Kaplan-Meier survival curves.
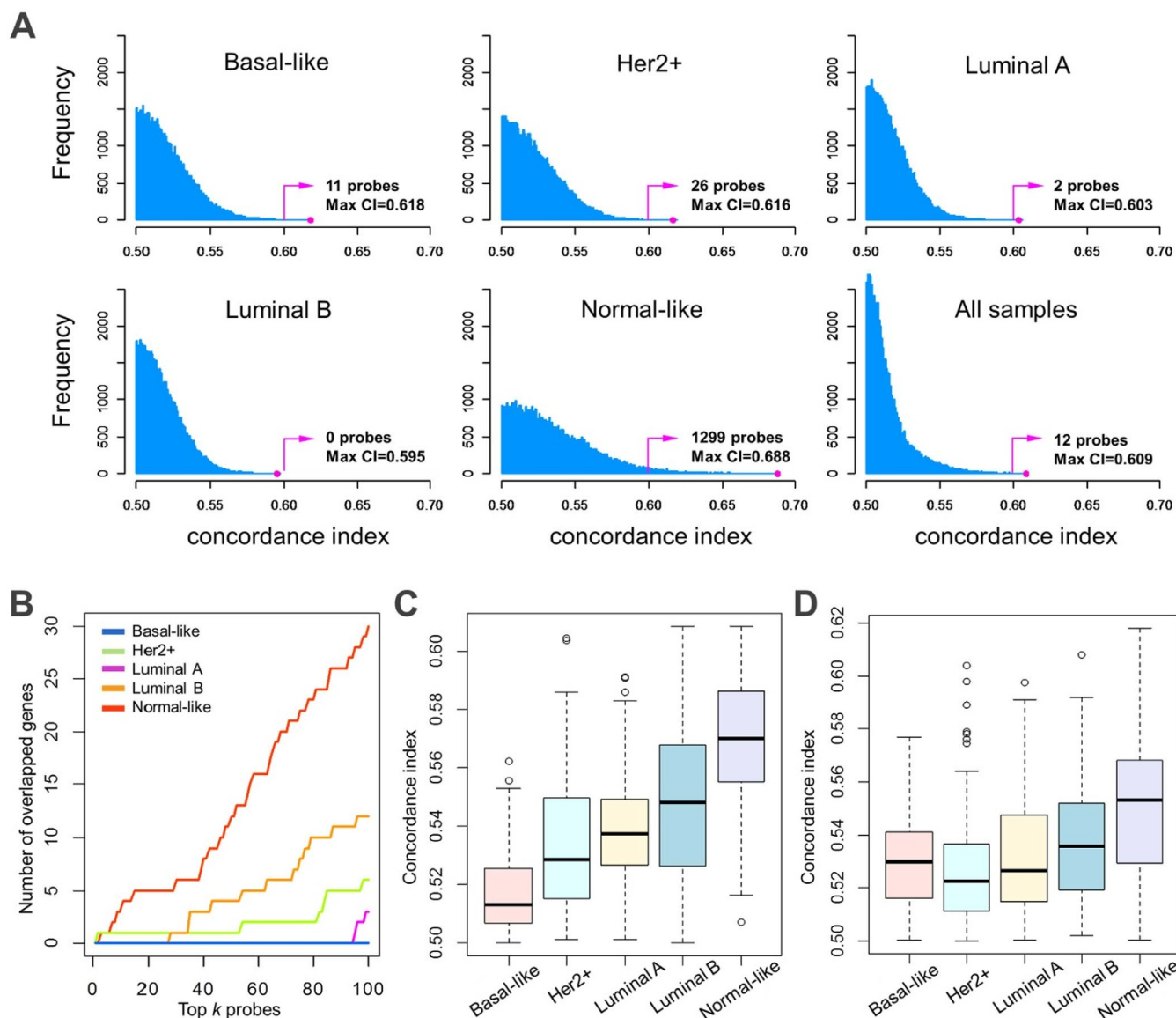
Figure 2 | **Gene-expression profiles of five breast cancer subtypes reveal different prognostic power.** (A) The distribution of CIs of all gene probes in the five breast cancer subtypes and METABRIC whole dataset. The x-axis denotes the CI and y-axis denotes frequency. On each dataset, gene probes are ranked according to their CIs. The number of probes with a CI higher than 0.60 and the maximal CI of the probes are also plotted for each dataset. (B) Comparison of the similarity of the prognostic probes for the five breast cancer subtypes with the METABRIC whole dataset. The points on the curve represent the size (y-value) of the overlap of the top x prognostic probes of a certain subtype and the METABRIC whole dataset. The rank of the probes is derived from (A) correspondingly. (C) Box-plot for the top 100 prognostic probes of each of the five subtypes with their CIs tested on the METABRIC whole dataset. The y-axis denotes the CI value. (D) Box-plot for the top 100 prognostic probes of each five subtypes with their CIs tested on the OsloVal dataset. The y-axis denotes the CI tested on the OsloVal dataset.
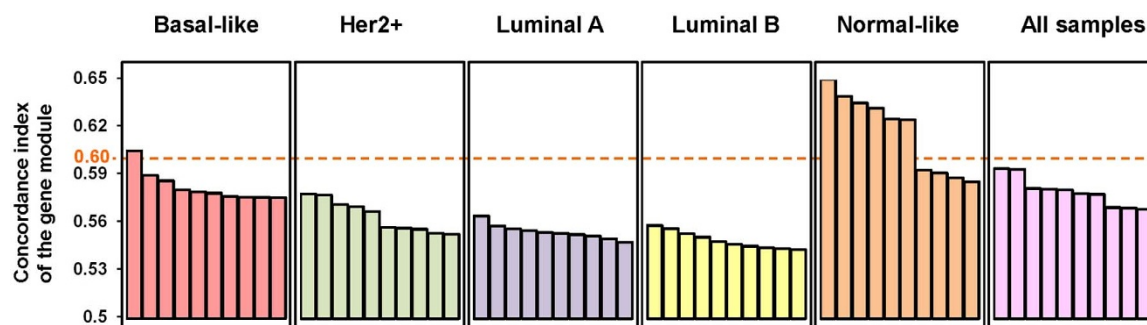

Figure 3 | **The CIs of the top 10 prognostic network modules from five distinct subtypes and the METABRIC whole dataset.** Modules generated from each dataset were ranked according to their CIs, which were calculated based on the averaged gene expression value of each module and survival time of corresponding patient cohorts. The dashed red line denotes CI = 0.60.

Table 1 | Network modules obtained from different datasets including five breast cancer subtypes and the METABRIC whole dataset

	Basal-like	HER2+	Luminal A	Luminal B	Normal-like	METABRIC
#modules	62	87	59	43	82	60
#genes	612	750	542	426	827	621
mSize	36	24	29	29	29	58
#(CI > 0.60)	1	0	0	0	6	0
hmCI	0.605	0.578	0.564	0.558	0.650	0.594
hmSize	12	5	7	5	25	5

#modules: number of modules; #genes: total number of genes contained in all modules; mSize: the largest size of modules; #(CI > 0.6): the number of modules having a CI larger than 0.60; hmCI: the highest CI of the modules; hmSize: the number of genes contained in the module with the highest CI.

The log-rank p-values for normal-like tumors and the two whole datasets were <0.01 (Figure 4C). However, the log-rank p-values for the other four subtypes were all larger than 0.01 (Supplementary Figure S1). Generally, patients with high expression of BIRC5 module genes were associated with poor overall survival. The BIRC5 gene is a member of the inhibitor of apoptosis (IAP) gene family, which encodes regulatory proteins that prevent apoptotic cell death. By doing functional enrichment analysis, we found that the BIRC5 modules are enriched in the following pathways: mitotic M-M/G1 phases, aurora A and B signaling and chromosome maintenance (Table 2 and Supplementary Table S4). Previous studies have shown that increased metrics of mitotic activity are relevant to worse survival outcome, which confirms our result that high expression of the BRIC5 module is associated with poorer survival (Figure 4C). The mitotic serine-threonine kinases have been shown to play roles in the regulation of cell cycle progression, the p53 pathway, and the checkpoint-response pathways[24]. Moreover, it has been shown that the expression of genes in the aurora A and B signaling pathway is
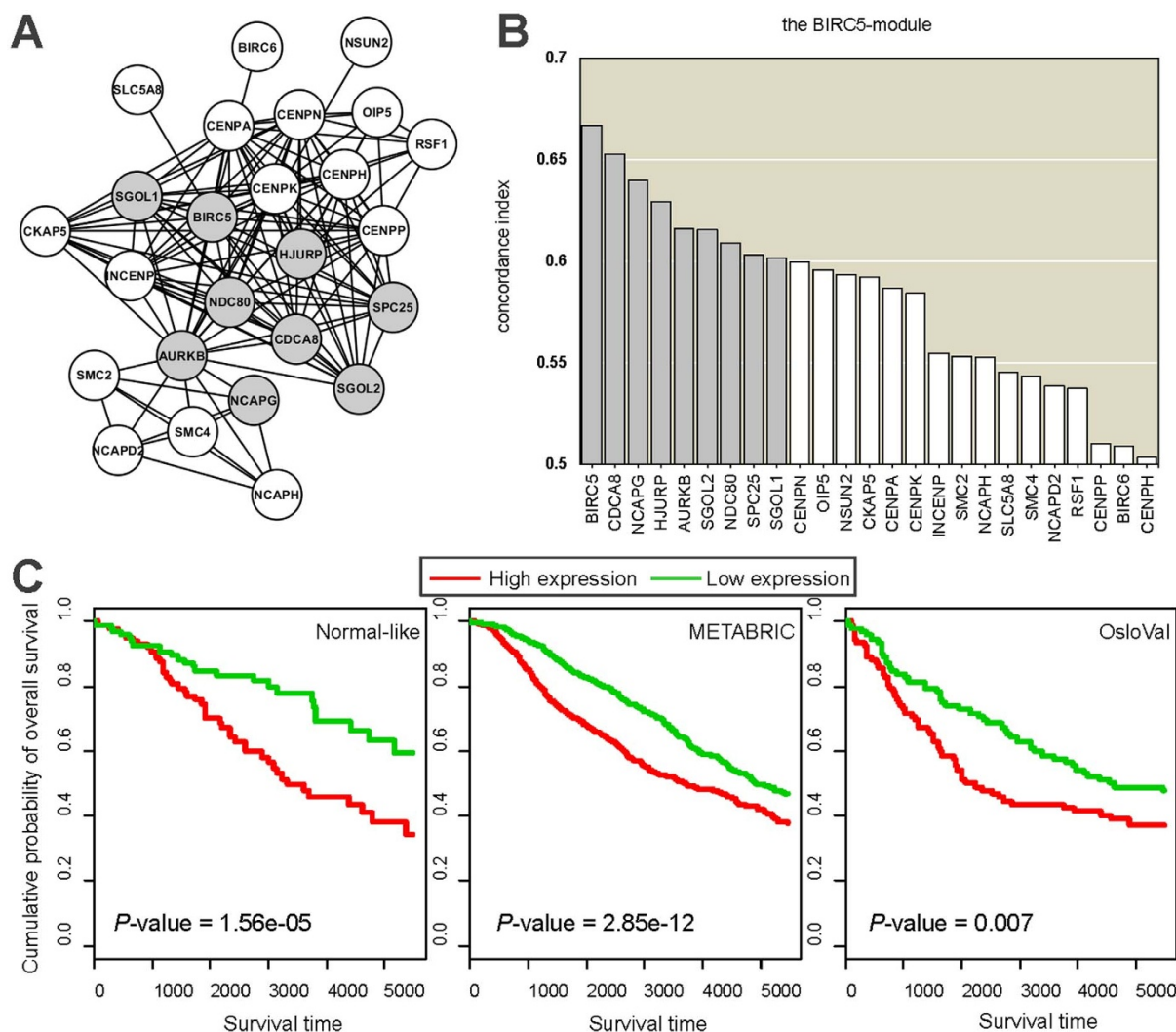

Figure 4 | The BIRC5 prognostic module. (A) The subnetwork of the 25-gene BIRC5 module. The nine genes in gray have CIs larger than 0.60. (B) The 25 genes are ranked according to their CIs. The nine genes with CIs larger than 0.60 are marked in gray (C) Kaplan-Meier cumulative survival curves of two breast cancer groups based on the expression of the BIRC5 module over a 15-year period on the normal-like tumor dataset, the METABRIC whole dataset and the OsloVal dataset respectively. The two patient groups in each plot were defined by partitioning the patients into two equal-sized sets using the median value of the averaged gene expression profile of the BIRC5 module.

Table 2 | Functional annotations of BIRC5 module based on pathway enrichment analysis

Gene set	Size	Overlap	FDR	BIRC5 module gene names in Gene set
Mitotic M-M/G1 phases	222	16	P < 5.0e-4	AURKB, SPC25, CDCA8, CENPA, INCENP, CENPN, CKAP5, SGOL2, SGOL1, CENPP, BIRC5, NDC80, CENPK, SMC2, SMC4, CENPH
Aurora B signaling	40	13	p < 5.0e-4	AURKB, NCAPH, CDCA8, CENPA, NCAPG, INCENP, NSUN2, SGOL1, BIRC5, NDC80, SMC2, SMC4, NCAPD2
Chromosome Maintenance	105	8	p < 3.3e-4	RSF1, HJURP, CENPA, OIP5, CENPN, CENPP, CENPK, CENPH
Aurora A signaling	31	4	p < 2.5e-4	AURKB,CENPA,CKAP5,BIRC5

cell-cycle related[25–29]. The chromosome maintenance was shown to be critical for stable chromosome function in mammalian and other eukaryotic cells[30]. These functional analyses reveal that the BRIC5 module is related to cell cycle and proliferation, which is likely relevant to cancer processes. Kaplan-Meier survival curves and functional enrichment analysis of the other six prognostic modules (CI > 0.6) were provided in Supplementary Information (Supplementary Table S3–S4 and Figure S2–S7).

Previously reported gene signatures show similar subtype-specificity with the BRIC5 module. Recently, Wu and Stein[22] discovered a 31-gene signature using five independent breast cancer datasets. In their research, the signature was compared with 48 published breast cancer gene signatures with regard to overall patient survival. We found that this 31-gene signature had a strong overlap with the 25-gene BIRC5 module obtained from normal-like tumors (Figure 5A). These two modules had 11 genes in common, among which, 6 were in the top 9 BIRC5 module genes with CIs larger than 0.60. Moreover, the BIRC5 gene was also found to be the most prognostic in this signature.

The winners of the Breast Cancer Prognosis Challenge developed a model based on three universal signatures, called attractor meta-genes, defined previously through a multicancer analysis of gene expression data[31,32]. We found that the mitotic CIN attractor meta-gene, which is the most prognostic of the METABRIC and OsloVal datasets, also has significant overlap with the BRIC5 module (Figure 5B). Here, 14 of the 25 genes in the BRIC5 module were in the top 100 genes of the CIN attractor.

In another study, Venet et al.[21] observed that many published breast cancer gene signatures have strong correlation to a cell proliferation-related gene set called meta-PCNA. This gene set contains 131 genes whose expression levels were correlated most positively with the proliferation marker PCNA. In our analysis, we found that the meta-PCNA signature significantly overlapped with the BIRC5 module and the MCM10 module, which are the top two prognostic modules obtained from the normal-like tumors (Figure 5C). In this comparison, 7 of the 25 BIRC5 module genes, and 10 of the 18 MCM10 module genes were among the meta-PCNA signature respectively. Notably, all three of these gene signatures were defined
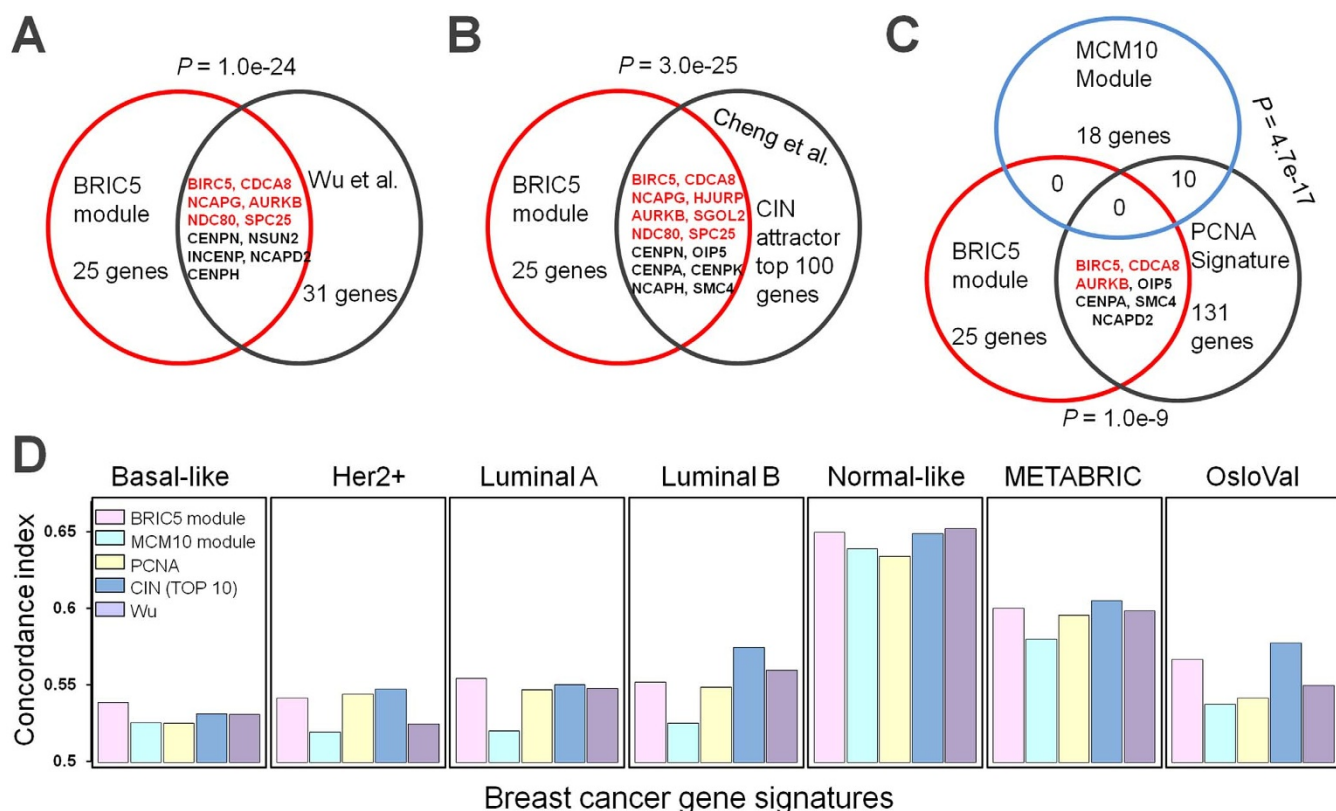

Figure 5 | Comparison of breast cancer prognostic gene signatures. Overlap between the BIRC5 module and three other signatures are shown. The three gene signatures include the 31-gene signature from Wu and Stein[22] (A), the top 100 genes of the CIN attractor metagene from Cheng et al.[32] (B), and the PCNA signature from Venet et al.[21] (C). P-values were calculated using the hypergeometric test. Overlapping genes among the 9 most prognostic BIRC5 module genes are labeled in red. (D) CIs of the five breast cancer signatures on the five subtypes, the METABRIC dataset and the OsloVal dataset.

based on whole breast cancer or multicancer datasets without considering tumor heterogeneity. Interestingly, they tend to share a significant number of genes with the modules found to be the most prognostic for the normal-like tumor subgroup.

Given this correlation, we deduced that these previously defined gene signatures were also tumor subtype specific, and have significantly high prognostic performance for the normal-like tumor subgroup. To confirm this, we calculated their corresponding CIs for each of the five subgroups, the METABRIC whole dataset and the OsloVal dataset by the similar strategy applied for our identified gene modules (Figure 5D). To be consistent with the original study, we used the top 10 genes of the CIN attractor metagene signature for calculation. This analysis revealed that all these gene signatures had the highest CIs for the normal-like breast cancer subgroup, and CIs on METABRIC dataset were also larger than those on other four subgroups. In addition to the BIRC5 module, CIN attractor and Wu's signature had comparable performances with subtle differences on the normal-like subtype and METABRIC dataset. We also adopted a multivariate Cox PH regression model to confirm our observations (Supplementary Table S5).

Recently, Fredlund et al.[33] discovered a breast cancer subtype-dependent network module containing fibroblast and stroma-related genes that are associated with less malignant tumors for luminal subtypes and aggressive lymph node positive disease among basal-like tumors. We investigated the prognosis ability of this stroma module consisting of 32 genes in the METABRIC dataset. However, unlike the five gene signatures, the stroma module doesn't show significant prognostic power (Supplementary Table S6) and it shared no overlap with the five gene signatures.

Distribution of tumors within the molecular subtypes doesn't affect the outcome of the analysis. Normal-like and luminal A breast cancers tend to have good prognosis based on gene expression and clinical feature predictors. However, compared to the other three subtypes, more patients with the typical "low risk" characteristics (i.e., ER positive, lymph node negative and low grade) exist in the luminal A and normal-like subgroups. The well-known OncotypeDx and Mammaprint signatures also suggest that expression data may provide information for this type of "low risk" patients that potentially differentiate patients at higher risk for distant metastasis or recurrent disease[1,2]. Thus, it is quite possible that the prediction ability of gene profiles is higher in the clinical "low risk" samples than in other clinically defined groups regardless of PAM50 subtypes. In other words, the relative better performance on normal-like and luminal A subtypes is a result of a significantly more "low risk" tumors present in these subgroups. Actually, The percent composition of each subtypes confirmed that normal-like and luminal A subtypes were indeed made up of more "low risk" patients than other three (Supplementary Figure S8). We then tested whether former mentioned gene signature have better performance on the clinical defined "low risk" patients (Supplementary Table S7). However, the results demonstrated that the clinical "low risk" subgroup cannot get special benefits from the well-developed gene signature, which means the observed diverse gene-based prognostic performance on the intrinsic breast cancer subtypes cannot be interpreted simply by these traditional clinical features.

Besides the risk levels, we also investigated the potential effect of tumor cellularity among different subtypes and found that the distribution of the tumor cellularity also has no much effect on the prognostic prediction of the normal-like tumors (Supplementary Figure S9 and Table S8). Interestingly, compared to other subtypes, the normal-like tumors possess a more balanced distribution of the three different cellularity types (low, moderate, high), which implying it is more heterogeneous in some extent. Furthermore, we studied the distribution of tumor histological type and observed that the normal-like tumors have a significant different histological type

distribution compared with other four subtypes (Chi-squared test, p-value < 1e-05, Supplementary Figure S10).

Discussion

Human cancers frequently display substantial tumor heterogeneity in all distinguishable phenotypic features, and this can have profound implications both for tumor development and therapeutic outcomes. In this study, we explored the diverse prognostic performance of standard prediction tools by applying genetic and clinical characteristics to breast cancer subtypes individually. We observed that prediction tools show distinct diversity of survival prediction ability when applied to breast cancer subtypes individually compared to whole breast cancer data. Unlike the other three subtypes, the luminal A and normal-like tumors are associated with good prognosis. Applying clinical feature covariates revealed strong prediction power for patient survival on luminal A breast cancer, which was not increased by adding gene expression covariates. However, gene expression covariates demonstrated significant prognostic power when applied to the normal-like subgroup.

Given the additional prognostic power associated with gene expression data for normal-like breast cancer, we examined this subtype with other gene signatures. We found that prognostic network biomarkers for the normal-like subtype had significant overlap with cancer signatures previously defined for whole breast cancer. The finding that previously reported gene signatures show similar subtype-specificity with the BRIC5 module suggests that the high prognostic ability of these gene signatures on the whole breast cancer samples likely results from their relative higher prognostic power on the normal-like subtype. They also support the finding that normal-like subtype patient survival is more predictable than other subtypes based on gene expression data. In conclusion, the gene-based signatures derived from all breast tumors have an extremely diverse clinical predictive ability when applied to intrinsic subtypes, and breast cancer heterogeneity greatly affects clinical prediction tasks.

However, unlike the other four breast cancer subtypes with well recognizable molecule characteristics, the significance of normal-like subtype is still largely undefined. Several studies suggest that the normal-like subtype is mainly an artifact resulting from a high percentage of normal cells in the tumor specimen[6,34]. Another study considered this subtype as a potential new one referred as claudin-low tumors[35]. These examples highlight the importance of considering tumor heterogeneity when predicting patient survival. In addition, the significant prognostic power of gene expression covariates on the normal-like breast tumors observed in this study support the theory that cancer biomarkers or signatures defined for the whole cancer cohort may be biased by the tumor heterogeneity. Thus, more detailed histologic, immunohistochemical, and gene expression analyses are needed to resolve this issue.

Developing computational methods based on genomic profiling to improve clinical diagnosis and survival prediction is an increasingly important issue in computational biology. In this study, we systematically evaluated the patient survival prediction performance of genomic and clinical data on the five intrinsic breast cancer subtypes. Our results revealed that molecular gene profiles and clinical features have different prognostic power when applied to the breast cancer subtypes individually. Specifically, gene expression profile of normal-like breast cancer contained more prognostic value than it did for the other four subgroups. In addition, we performed a network-based method on the five breast cancer cohorts to identify prognostic gene modules. The results of this analysis validated previous observations and uncovered a 25-gene module related to cell cycle and proliferation that was highly correlated with patient survival in the normal-like tumors. Thus, this study supports the notion of considering tumor heterogeneity when using gene expression data to predict patient survival, and opens new avenues for this type of analysis.

Methods

Materials. *METABRIC dataset.* The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset contains detailed clinical annotations, patient overall survival time, expression profiles, CNV profiles, and SNP genotypes derived from 1981 breast tumors collected from participants of the METABRIC trial[7]. Nearly all oestrogen receptor (ER)-positive and lymph node (LN)-negative patients did not receive chemotherapy, whereas ER-negative and LN-positive patients did. None of the HER2+ patients in this trial received trastuzumab. This dataset was accessed through Synapse (synapse.sagebase.org) and was used as the training data in the Sage Breast Cancer Challenge (BCC) competition. The expression profiles contain 49576 probe sets, performed on the Illumina HT 12v3 platform, re-normalized at Sage Bionetworks by the BCC Support Team. The clinical feature profiles contain 25 clinical covariates including size, grade, age at diagnosis, lymph nodes status, and the PAM50 subtype annotations. The dataset was composed of 328 basal-like tumors, 238 HER2+ tumors, 719 luminal A, 490 luminal B, 200 normal-like tumors and 6 samples with unclear category. More detailed description on METABRIC data is available at the Breast Cancer Challenge support page (https://sagebionetworks.jira.com/wiki/display/BCC/).

OsloVal dataset. The OsloVal cohort consists of 184 breast cancer patients collected from 1981 to 1999 at the Norwegian Radium Hospital. This relative small dataset was used as a validation set for the BCC, and as a result, has the same data structure and pre-processing as the METABRIC dataset. It contains 11 clinical features but no PAM50 subtype annotations. This dataset was accessed through Synapse (synapse.sagebase.org).

Methods. *Multivariate Cox PH regression model.* We adopted a multivariate Cox PH model to examine the association between covariates and survival time, and to predict clinical survival time. For gene expression data, we performed a supervised principal component analysis with the superpc package developed by Tibshirani et al.[36]. The first three principal components (PCs) were used as the gene expression input variables. Ten types of clinical features were selected from the 25 clinical covariates due to their level of completeness and availability in the OsloVal dataset. These ten clinical features/input covariates include: age at diagnosis, tumor size (cm), lymph node assessment, grade of disease, estrogen receptor (ER) immunohistochemistry status, HER2 SNP6 state, treatment received, HER2 expression, ER expression, and progesterone receptor (PR) expression (Supplementary Table S9). We used overall survival time as the survival outcome, and applied three-fold cross validation (CV) on the testing data. At each validation, two folds of the partition were used for performing the superpc analysis, training the multivariate Cox model, and selecting the clinical feature variables with Akaike information criterion. Afterwards, the trained model was tested on the third fold to conduct survival prediction. The concordance index (CI) was used to estimate the prognostic effect of the model. To eliminate the randomness of sample partition in the three-fold CV, each model was repeated 100 times. The final assessment of prognostic effect was measured by the average of the 300 CIs on the test folds. We applied this calculation process on gene expression covariates, clinical feature covariates or the combination of these two for each of the five intrinsic subtypes and the METABRIC whole dataset respectively. The p-values which indicated the significance of the results were calculated using a permutation test, in which samples were randomly divided into five classes keeping the subtype size and repeated 1000 times.

The concordance index. The concordance index is one of the most commonly used performance measures for assessing predicting models in survival analysis[37]. It is the probability of concordance between the predicted and the real survival, and calculated as follows:

$$CI = \frac{1}{N} \sum_{T_i \text{ uncensored}} \sum_{T_i < T_j} 1_{p_i < p_j}$$

The indicator function = 1 if $a < b$, otherwise it is 0. For patient i, p_i is the predicted survival time by the model p while T_i denotes the real survival time. A usable patient pair for calculating the CI is that the patient with the shorter survival time must be uncensored, or in other words, have event. N denotes the number of all usable patient pairs. Therefore, CI with a range of $0 \sim 1$ can be interpreted as the fraction of all patient pairs whose predicted survival times are correctly ordered among all patients that can actually be ordered. CI = 1 indicates perfect prediction accuracy and CI = 0.5 indicates a random guess. We use CI as a measurement of the prognostic power of individual genes and gene modules, and to assess the prediction performance of the multivariate Cox PH model. In this work, CIs lower than 0.5 were adjusted as follows: adj-CI = 1 − CI.

Identifying cancer prognostic signatures based on Reactome functional interaction (FI) network. A network biology method developed by Wu and Stein[22] was adopted for identifying cancer prognostic signatures. Using this method (see the Reactome FI cytoscape plug-in at http://wiki.reactome.org/index.php/Reactome_FI_Cytoscape_Plugin), we first discovered the modules using the Markov Clustering (MCL) procedure from a weighted gene functional interaction (FI) network consisting of 10,956 proteins and 209,988 interactions[38]. Then, an averaged expression profile for each module, calculated from the member gene expression profiles, was used for the downstream survival analysis. We applied this method onto the five PAM50 subtypes

and the METABRIC whole dataset to search for network biomarkers that were significantly correlated with patient survival for each cohort respectively.

To run the plug-in, probes were mapped to the gene symbol from HGNC (http://www.genenames.org/), and the expression profiles of multiple probes that were mapped to the same gene were averaged as the gene expression profile of this gene. Probes that either mapped to multiple genes or could not be mapped were removed. The mapping resulted in 15696 genes. For the MCL clustering, we used 5.0 as the inflation coefficient. Only those modules with a size of 5 or larger, and an average PCC equals to or larger than 0.25 were kept for further analysis. An expression profile was assigned for each module by averaging the gene expression profiles of all genes in the module, and using it to calculate the CI for each module on the corresponding patient group. Pathway and GO term enrichment analysis of the genes within a certain module were also conducted with this Cytoscape Plug-in.

To determine the statistical significance of the detectability of a network biomarker (e.g., a module generated from the normal-like breast cancer cohort), we performed a permutation test on the given patient group (e.g., the normal-like breast cancer cohort). 10000 simulated gene modules with the same size were randomly generated from the network, among which n modules get the CIs larger than that of the given module, leading to the permutation p-value of n/10000.

R package. We used R package BCC and predictiveModeling downloaded from Synapse (synapse.sagebase.org) to access the METABRIC and OsloVal datasets, build the models, and perform the cross validation. R package missForest was used to impute missing values in the METABRIC dataset. We used R package superpc to perform the principal component analysis (http://www-stat.stanford.edu/~tibs/superpc/). R package survival was used for Cox proportional hazards model, Kaplan-Meier survival analysis, Wald test, log-rank test and concordance index calculation. R package randSurvivalForest was applied to run the rand survival forest model.

1. van 't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
2. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817–2826 (2004).
3. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
4. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **9819**, 10869–10874 (2001).
5. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* **100**, 8418–8423 (2003).
6. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).
7. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
8. Rouzier, R. et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* **11**, 5678–5685 (2005).
9. Rody, A. et al. The erbB2+ cluster of the intrinsic gene set predicts tumor response of breast cancer patients receiving neoadjuvant chemotherapy with docetaxel, doxorubicin and cyclophosphamide within the GEPARTRIO trial. *Breast* **16**, 235–240 (2007).
10. Carey, L. A. et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* **295**, 2492–2502 (2006).
11. Prat, A. et al. Molecular Characterization of Basal-Like and Non-Basal-Like Triple-Negative Breast Cancer. *Oncologist* **18**, 123–133 (2013).
12. Perou, C. M. Molecular stratification of triple-negative breast cancers. *Oncologist* **16**, 61–70 (2011).
13. Rody, A. et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. *Breast Cancer Res* **13**, R97 (2011).
14. Hallett, R. M. et al. A gene signature for predicting outcome in patients with basal-like breast cancer. *Sci Rep* **2**, 227 (2012).
15. Moestue, S. A. et al. Metabolic biomarkers for response to PI3K inhibition in basal-like breast cancer. *Breast Cancer Res* **15**, R16 (2013).
16. Cheang, M. C. et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin Cancer Res* **14**, 1368–1376 (2008).
17. Kauraniemi, P. et al. Effects of Herceptin treatment on global gene expression patterns in HER2-amplified and nonamplified breast cancer cell lines. *Oncogene* **23**, 1010–1013 (2003).
18. Paik, S. et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *J Clin Oncol* **24**, 3726–3734 (2006).
19. Naderi, A. et al. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene* **26**, 1507–1516 (2006).
20. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).
21. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computat Biol* **7**, e1002240 (2011).
22. Wu, G. & Stein, L. A network module-based method for identifying cancer prognostic signatures. *Genome Biol* **13**, R112 (2012).

23. Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* **5**, 181re1–181re1 (2013).

24. Katayama, H. *et al.* Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53. *Nat Genet.* **36**, 55–62 (2003).

25. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell.* **144**, 646–674 (2011).

26. Bischoff, J. R. *et al.* A homologue of Drosophila aurora kinase is oncogenic and amplified in human colorectal cancers. *EMBO J.* **17**, 3052–3065 (1998).

27. Walter, A. O., Seghezzi, W., Korver, W., Sheung, J. & Lees, E. The mitotic serine/threonine kinase Aurora2/AIK is regulated by phosphorylation and degradation. *Oncogene* **19**, 4906–4916 (2000).

28. Wheatley, S. P., Carvalho, A., Vagnarelli, P. & Earnshaw, W. C. INCENP is required for proper targeting of Survivin to the centromeres and the anaphase spindle during mitosis. *Curr Biol.* **11**, 886–890 (2001).

29. Zeitlin, S. G., Shelby, R. D. & Sullivan, K. F. CENP-A is phosphorylated by Aurora B kinase and plays an unexpected role in completion of cytokinesis. *J Cell Biol.* **155**, 1147–1158 (2001).

30. De Lange, T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.* **19**, 2100–2110 (2005).

31. Cheng, W. Y., Ou Yang, T. H. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci Transl Med* **5**, 181ra50–181ra50 (2013).

32. Cheng, W. Y., Ou Yang, T. H. & Anastassiou, D. Biomolecular events in cancer revealed by attractor metagenes. *PLoS Computat Biol* **9**(2), e1002920 (2013).

33. Fredlund, E. *et al.* The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. *Breast Cancer Res.* **14**, R113 (2012).

34. Peddi, P. F., Ellis, M. J. & Ma, C. Molecular basis of triple negative breast cancer and implications for therapy. *Int J Breast Cancer* **2012**, 217185 (2012).

35. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* **8**, R76 (2007).

36. Bair, E. & Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**, e108 (2004).

37. Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P. & Raykar, V. C. On ranking in survival analysis: Bounds on the concordance index. *In Advances in Neural Information Processing Systems* 2007. pp. 1209–1216.

38. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* **11**, R53 (2010).

Acknowledgments

Author contributions

S.Z. and Z.L. conceived and designed this study; Z.L. conducted the test; Z.L., X.S.Z. and S.Z. analyzed the data and wrote the paper. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Liu, Z., Zhang, X.-S. & Zhang, S. Breast tumor subgroups reveal diverse clinical prognostic power. *Sci. Rep.* **4**, 4002; DOI:10.1038/srep04002 (2014).