Genomic Characteristics of Triple-Negative Breast Cancer Nominate Molecular Subtypes That Predict Chemotherapy Response

Jihyun Kim1, Doyeong Yu1, Youngmee Kwon2, Keun Seok Lee2, Sung Hoon Sim2,3, Sun-Young Kong3,4, Eun Sook Lee2,4, In Hae Park2,3, and Charny Park1

ABSTRACT

The heterogeneity of triple-negative breast cancer (TNBC) poses difficulties for suitable treatment and leads to poor outcome. This study aimed to define a consensus molecular subtype (CMS) of TNBC and thus elucidate genomic characteristics and relevant therapy. We integrated the expression profiles of 957 TNBC samples from published datasets. We identified genomic characteristics of subtype by exploring the pathway activity, microenvironment, and clinical relevance. In addition, drug response (DR) scores (n = 181) were computationally investigated using chemical perturbation gene signatures and validated in our own patient with TNBC (n = 38) who received chemotherapy and organoid biobank data (n = 64). Subsequently, cooperative functions with drugs were also explored. Finally, we classified TNBC into four CMSs: stem-like; mesenchymal-like; immunomodulatory; luminal-androgen receptor. CMSs also elucidated distinct tumor-associated microenvironment and pathway activities. Furthermore, we discovered metastasis-promoting genes, such as secreted phosphoprotein 1 by comparing with primary. Computational DR scores associated with CMS revealed drug candidates (n = 18), and it was successfully evaluated in cisplatin response of both patients and organoids. Our CMS recapitulated in-depth functional and cellular heterogeneity encompassing primary and metastatic TNBC. We suggest DR scores to predict CMS-specific DRs and to be successfully validated. Finally, our approach systemically proposes a relevant therapeutic prediction model as well as prognostic markers for TNBC.

Implications: We delineated the genomic characteristic and computational DR prediction for TNBC CMS from gene expression profile. Our systematic approach provides diagnostic markers for subtype and metastasis verified by machine-learning and novel therapeutic candidates for patients with TNBC.

Introduction

Triple-negative breast cancer (TNBC) is known to have poor prognosis and the tendency of early metastasis (1). Although TNBC shows a higher response rate to standard chemotherapy than hormone receptor–positive breast cancer, it has poor prognosis due to early development of multidrug resistance (2). Recently, targeted therapy based on genomic characteristics was introduced with promising results. For example, DNA-damaging agents like platinum agent or PARP1 inhibitor showed excellent efficacy in BRCA1/2 mutation carriers among heterogeneous TNBC groups (3). In addition, immune-modulating agent, anti–programmed death-ligand 1 (PD-L1) antibody, and atezolizumab demonstrated outstanding results in patients with metastatic TNBC whose tumors had higher PD-L1 expression (4). To find the appropriate drug for each patient with TNBC, a thorough understanding of the heterogeneity of TNBC and robust subtype identification are prerequisites.

Previous TNBC subtype identification studies have proposed various classification and subtypes of size 3 to 6. Lehmann’s classification has advanced to specify the heterogeneity and clinically evaluated several independent cohorts (5–7). Further studies proposed different-size subtypes of 3 or 4 based on next-generation sequencing using a more robust algorithm (8, 9). However, owing to restricted sample size (n < 200) originating from a single platform or insufficient algorithm optimization to refer diagnostic gene set, a robust TNBC subtype classification has not been established in the absence of multiple validations. In addition, previous classification was restricted to only primary tumors despite poor outcomes due to metastatic tumor development.

As an additional important factor of tumor heterogeneity, the tumor-associated microenvironment (TME) confers genetic diversity to tumor bulk cells. Cellular characteristics are decided by tumorigenic origin (10). As proven in a single-cell platform, the TME cell population cooperates with the chemotherapy response (11, 12). Cell abundance investigation to induce transitional diversity is unrestricted to a single-cell platform. Distinct cell abundance identification could be estimated from bulk tumor sample gene expression profiles using cell deconvolution technologies (13, 14). Expanded cellular investigation in terms of TME could reveal the origin of functional heterogeneity and specify therapeutic targets.

Due to the complex interactions between heterogeneous tumor cells and TME, there have been a few successes in attempts to search novel...
drug targets computationally in TNBC. To reveal the mode of action and to identify novel drug–target relationships, pharmacogenomics approaches utilizing gene expression or mutation profiles have been suggested (15, 16). However, most computational methods were performed using genetic profiles of cell lines, and validation was insufficient to follow actual patients’ responses. A systematic approach interconnected with molecular subtype diagnosis and therapeutic strategy is required to find effective treatment.

In the current study, utilizing gene expression profiles, we established various classification and prediction models. Machine-learning approaches referring related gene expression signature classified two types: a consensus molecular subtype (CMS) by support vector machine (SVM) and primary–metastasis by random-forest model. In-depth investigation using known gene signatures recapitulated genomic characteristics and clinical relevance of CMS and metastasis. We also carried out TME and mammary gland cells by TNBC bulk cell deconvolution and revealed the cellular heterogeneity of CMS. In additional study, we established drug response (DR) prediction model to compute DR score based on chemical perturbation gene signatures. We identified subtype-related DRs and pathway interventions. The DR predictions were validated in actual patient’s cisplatin response and matched gene expression profile as well as TNBC organoid (17).

Materials and Methods

CMS classification and functional investigation using gene expression profile

We collected 1,525 gene expression profiles from 14 high-throughput gene expression datasets related to metastasis and TNBC from the Gene Expression Omnibus (GEO; Supplementary Table S1). To eliminate the batch effect from the datasets, meta-analysis of the throughput gene expression datasets related to metastasis and TNBC Supplementary Fig. S1A). We removed samples considered as positive for ER, PR, and HER2 demonstrating enrichment scoring by referring to Hallmark gene sets from MSigDB or additional gene signatures (19–21). Key pathways inducing distinct CMS characteristics were chosen by limma test (P < 1e-05). To verify functionality for each CMS, we collected seven representative gene signatures to extract subtype characteristics, namely epithelial-to-mesenchymal transition (EMT), stromal, immune, TME, stemness, chromosomal instability (CIN), and hormone. Next, enrichment scores were estimated using GSVA. The EMT score was defined from normalized average expression values using a previously studied EMT gene set (22). The abundance of stromal and immune cells was estimated by the ESTIMATE tool (23). To perform cross-validation of immune infiltration, the TME score was calculated by xCell (14). The stemness score was determined from The Cancer Genome Atlas (TCGA) pan-cancer stemness index (24). In addition, we referred to embryonic stem (ES) cell-like gene expression signatures of Nanog homeobox (NANOG), octamer-binding transcription factor 4 (OCT4), SRY-related HMGB2 (SOX2), and MYC, which are known to be overexpressed in poorly differentiated tumors (25). The CIN score was derived from the CIN70 gene expression signature (26). Hormone response was summarized from the average score of the ER and androgen receptor (AR) response pathway scores. We compared the seven representative gene signature scores of our CMS with the Lehmann and Burstein classification. The difference in scores among subtypes was tested by the Mann–Whitney U test.

To reveal the TME heterogeneity, we estimated cell abundance, known as the deconvolution method, of nonimmune stromal and immune cells. The abundance of mammary gland cells was computed using MCPCounter using cell marker gene sets revealed by a single-cell study (27, 28). In addition, the proportion of immune cells was estimated using CIBERSORT (13).

To evaluate association between BRCAness and stem-like (SL) type, we classified TCGA BRCA TNBC into our CMSs and called germline mutations from TCGA breast cancer TNBC samples (total n = 1,101, TNBC n = 120) and mutation calling pipeline followed by GATK v3.3 germline short variant discovery pipeline (29). First, the mutations were called using HaploTypeCaller and filtered out low-quality variants by VariantFiltration. To extract pathogenic and functional germline mutations, we chose BRCA1/2 variants registered as “Pathogenic” or “Likely pathogenic” in ClinVar, INDEL, or truncated mutations (30).

To infer the genetic characteristics of metastatic transition in our aggregated data, we analyzed differentially expressed genes (DEG) between primary tumor and metastasis samples using limma (adjusted P < 0.001; Benjamini–Hochberg correction). Next, significant pathways were inferred from the DEGs by gene set enrichment analysis (GSEA) using ReactomFI (31). Finally, the DEGs based on the GSEA test (FDR < 0.05) were used to develop a gene–gene interaction network, visualized using Cytoscape. In addition, we generated a metastasis prediction model based on these prognostic markers. First, we tested for MFS and concordance index from DEGs previously comparing primary with metastasis samples. Tests were performed using the R package survcomp, and the identified gene set was defined as the metastasis prediction gene set (32). Next, to test the metastasis prediction availability of gene set, we established a random-forest machine-learning model to classify primary and metastasis samples using the metastasis prediction gene set identified from MFS test (P < 0.05). Finally, the random-forest model was evaluated by AUC metric using a 4-fold cross-validation procedure. Pathways of metastasis prediction gene set were also investigated using GSEA.
Computational DR prediction applying chemical perturbation gene signature and cross-validation

To explore the relevant DR status for each patient, we computed the DR scores in both aggregated data and biobank expression profiles (17). DR scores were calculated by the GSVA method using the MSigDB chemical and genetic perturbation signature gene set (17). Next, we selected drug signatures to examine response differences among CMSs using ANOVA. To select reliable DRs, we investigated drug names registered for clinical trials (NIH ClinicalTrials.gov database), considering synonyms provided in Drugbank (33). Meanwhile, paired Pearson correlation coefficients between pathway activities and DR scores were calculated to explore the functional association for each drug. To evaluate DR scores of aggregated data, we referred biobank expression profiles. First, organoid samples were classified into four subtypes using our CMS classifier, and DR scores were also equally calculated. Subtype specificity of biobank DR scores was also tested by ANOVA.

ROADMAP data analysis related to cisplatin response

Patients and sample preparation

In 2016, we introduced the ROADMAP program to implement genomic sequencing of tumors and blood coupled with clinical data in the management of advanced refractory breast cancer. As of June 2018, 38 female patients with metastatic TNBC were enrolled, and their genomic and clinical data were obtained after receiving written consent. Among the 38 patients, 27 patients had metastatic breast cancer, 6 patients had metastatic bone cancer, and 5 patients had metastatic brain cancer. The thickness of tumor samples was not uniform. Therefore, we used FFPE samples and aggregated data. Of the ROADMAP-enrolled patients, 31 patients had undergone cisplatin therapy, and the response log was available. Biobank cell models derived from patients with TNBC were divided into cisplatin response and nonresponse according to their drug high-throughput screening (HTS) IC₅₀ values. Cisplatin DR scores of ROADMAP patients and biobank cells were tested using the Wilcoxon rank-sum test. We also explored genes involved in cisplatin response by analyzing DEGs between cisplatin-response and -resistance samples using the Wilcoxon rank-sum test ($P < 0.05$) in the ROADMAP dataset. To avoid the artifact effect from the FFPE status, we eliminated genes that showed differences in expression between FFPE and fresh-frozen samples (Wilcoxon rank-sum test, $P < 0.1$). Finally, pathways involved in cisplatin response were investigated using ReactomeFI GSEA (31). In addition, survival analysis was performed to identify drug-related prognosis markers based on DEGs. The combined CI and HR were computed to get an overall score from independent datasets including aggregated data and the ROADMAP dataset of prognostic marker genes.

Results

CMS classification

The summary of our aggregated data composed of primary ($n = 601$) and metastatic ($n = 356$) TNBC is shown in Table 1. We confirmed to eliminate the batch effect via PCA plot (Supplementary Fig. S1A) and eliminated samples that highly expressed ER, PR, and HER2 using a bimodal filter on the gene expression distribution. The Lehmann subtype supports the most stringent ER, PR, and HER2 filtering to eliminate 15.3% of our aggregated data samples. On the other hand, 56% of Burstein samples to pass IHC were eliminated based on mRNA expression.

Table 1. Summary of gene expression profiles from the aggregated data.

<table>
<thead>
<tr>
<th>ID</th>
<th>Total</th>
<th>Primary</th>
<th>Metastasis</th>
<th>Tissue type</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE103091</td>
<td>107</td>
<td>53</td>
<td>21</td>
<td>74 Breast</td>
<td>GPL570</td>
</tr>
<tr>
<td>GSE10706</td>
<td>23</td>
<td>14</td>
<td>0</td>
<td>14 Breast</td>
<td>GPL570</td>
</tr>
<tr>
<td>GSE2034</td>
<td>286</td>
<td>223</td>
<td>0</td>
<td>225 Breast</td>
<td>GPL96</td>
</tr>
<tr>
<td>GSE46141</td>
<td>91</td>
<td>0</td>
<td>52</td>
<td>4 Breast, 8 liver, 24 LN, 14 skin, 1 ascites, 1 bone</td>
<td>GPL10379</td>
</tr>
<tr>
<td>GSE12726</td>
<td>204</td>
<td>16</td>
<td>129</td>
<td>145 Breast</td>
<td>GPL570</td>
</tr>
<tr>
<td>GSE4107</td>
<td>29</td>
<td>0</td>
<td>16</td>
<td>6 Bone, 8 brain, 2 lung</td>
<td>GPL570</td>
</tr>
<tr>
<td>GSE14018</td>
<td>36</td>
<td>0</td>
<td>27</td>
<td>5 Bone, 5 brain, 3 liver, 14 lung</td>
<td>GPL96</td>
</tr>
<tr>
<td>GSE46563</td>
<td>94</td>
<td>14</td>
<td>0</td>
<td>14 Breast</td>
<td>GPL6884</td>
</tr>
<tr>
<td>GSE46928</td>
<td>52</td>
<td>20</td>
<td>6</td>
<td>26 Breast</td>
<td>GPL96</td>
</tr>
<tr>
<td>GSE5327</td>
<td>58</td>
<td>29</td>
<td>9</td>
<td>38 Breast</td>
<td>GPL96</td>
</tr>
<tr>
<td>GSE54323</td>
<td>29</td>
<td>0</td>
<td>21</td>
<td>12 Bone, 6 liver, 3 LN</td>
<td>GPL570</td>
</tr>
<tr>
<td>GSE64935</td>
<td>120</td>
<td>0</td>
<td>74</td>
<td>11 Breast, 13 liver, 29 LN, 20 skin, 1 UNK</td>
<td>GPL10379</td>
</tr>
<tr>
<td>GSE7390</td>
<td>198</td>
<td>141</td>
<td>0</td>
<td>141 Breast</td>
<td>GPL96</td>
</tr>
<tr>
<td>GSE76124</td>
<td>198</td>
<td>91</td>
<td>1</td>
<td>92 Breast</td>
<td>GPL570</td>
</tr>
</tbody>
</table>

Total 1,525 601 356

Note: Matched clinical information is based on GEO. Sample counts after filtering out ER-, PR-, and HER2-positive samples are listed under primary and metastasis. A part of the samples include long-distance metastasis described in tissue type. In total, 601 primary and 356 metastasis samples were chosen for functional analysis.
In Lehmann’s classification, the mesenchymal SL type showed the highest stromal infiltration, but the discriminative power using the EMT score was better in our CMS (t test, metaCMS \( P = 2.86 \times 10^{-69} \), Lehmann \( P = 7.69 \times 10^{-31} \); Supplementary Fig. S1E). The genomic characteristics of mesenchymal type to associate with stemness were ambiguous (Supplementary Fig. S1E). Stemness was diffused across the basal-like (BL) 1, BL2, and mesenchymal subtypes. Consistent with a previous study, the average silhouette width of BL1 showed the worst result (5).

Burstein’s classification resembles with our result rather than Lehmann’s six subtypes, except for the BL immune-activation type (\( P < 2.2 \times 10^{-16}, \chi^2 \) test). When investigating Burstein’s dataset (\( n = 198 \)), the BL immune-activation type considerably included stemness samples (18%; Supplementary Fig. S1D and S1E). In immune and TME signatures (Supplementary Fig. S1E), our CMS was more discriminative rather than Burstein’s. Therefore, we finalized the TNBC CMS into the SL (32.1%), mesenchymal-like (MSL, 24.6%), IM (12.9%), and luminal AR (LAR, 30.5%) subtypes (Fig. 1A).

**Genomic characteristics and microenvironment of CMSs**

Key pathways activated in each subtype were identified from computational scores (\( P < 0.1e-06 \)). Interestingly, IM- and LAR-
associated pathways were also upregulated in the MSL subtype (Fig. 1A). EMT, TGFβ, and TP53 signals were constitutively upregulated restricted in the MSL. Our LAR showed the highest ER and AR response, particularly involved in adipogenesis and fatty acid metabolism. IFNα, IFNγ response, IL6–JAK–STAT3, IL2–STAT5, and inflammatory responses were activated in IM (Fig. 1A). In addition, both representative immune cell infiltration and TME signature show the highest score in IM (Fig. 1B). This indicates that IL–JAK–STAT signaling is implicated in immune infiltration, as identified in previous studies on TNBC (37). TNFα and apoptosis were activated in both IM and MSL. The characteristics of ES cells included an SL subtype, in which cell cycle, proliferation, and ES-associated pathways were activated, such as mitosis, E2F, MYC, WNT–β catenin, and Notch signaling (Fig. 1A and B). As previously known (26, 36), cell-cycle deficiency in the SL subtype induced the highest CIN (Fig. 1B; \( P = 1.35e-120 \)). Meanwhile, DNA repair and glycolysis were enhanced across SL and LAR.

The TME for each CMS presented distinct features in both stromal and immune cells. First, the IM was strongly immune-infiltrated and harbored memory B cells, CD8 T cells, activated CD4 T cells, gamma delta T cells, and activated natural killer cells (Supplementary Fig. S2A; \( P < 1.0e-07 \)), whereas MSL showed a high incidence of M2 macrophages, which was consistent with another study (9). Next, we explored the prognostic effect by immune cells to determine MFS or OS. The results showed that activation of CD8 T cells was associated with good prognosis (Supplementary Fig. S2B; MFS, \( P = 0.005 \); OS, \( P = 0.02 \)). Among the known immunotherapy targets such as CD20, CD52, PD-L1, CTLA4, CD52, and CTLA4 were significantly enhanced in IM (Supplementary Fig. S2C and S2D; ANOVA, \( P < 0.001 \) and Tukey test). Another bulk cell deconvolution revealed distinct characteristics of seven mammary gland cells depending on CMS (Fig. 1B; \( P < 3.66e-17 \)). Myoepithelial cells were enriched in the SL, consistent with a previous finding (39), and we additionally discovered that luminal progenitor cells were enhanced in SL. In MSL, adipocytes and matrix metalloproteinases categorizing as stromal cells were highly enriched, followed by luminal epithelial and myoepithelial cells. Interestingly, key pathways of adipogenesis and fatty acid metabolism were activated across the MSL and LAR (Fig. 1A), but adipocytes were enriched only in the MSL type. In addition, we confirmed that tumor-associated stromal cells were infiltrated in MSL from upregulation of nine markers (Supplementary Fig. S3; refs. 40, 41). Luminal epithelial cells showed a strong presence in the LAR, and blood cells undergoing the mitotic cell cycle process were mainly activated in the IM (Fig. 1B). In further investigation of SL characteristics for the SL subtype, we referred to previously known target genes of ES cell transcription factors and polycomb-regulated genes (25). As already identified in a previous study, target genes of transcription factors NANOG, OCT4, SOX2, and MYC were clearly activated in SL. In addition, we confirmed that known targets of polycomb repressive complex 2 (PRC2) downregulated in SL, PRC2 subunit 2 (SUZ12), PRC subunit (EED), and methylation at histone 3 lysine 27, are upregulated in the MSL subtype (Fig. 1B).

To investigate which subtype is dysregulated by BRCA1/2 germline mutations associated with BRCAaness, we classified TCGA BRCA TNBC samples into our CMS and called germline mutations. Fifty-four cases in TCGA TNBC samples (\( n = 120 \)) belonged to SL type. Finally, 8 SL-type samples in total 16 pathogenic BRCA1/2 germline mutations were identified (Fisher test, \( P < 1.89e-07 \)). Key pathways, G2–M checkpoint, MYC, mitotic spindle, and NOTCH activation in BRCA1/2 mutation samples resemble those shown in Fig. 1A (Supplementary Fig. S4; \( P < 2.13e-04 \)).

To investigate prognosis, we examined 10-year survival rates based on OS and MFS rates among CMSs. Survival test details are described in Supplementary Table S3. SL clearly presented poor MFS (median MFS = 5.9 years) and showed 2.53 times the HR (95% confidence interval (CI), 1.69–3.77; Fisher test, \( P = 5.40e-06 \)) compared with MSL, which had good prognosis (Fig. 1C). Moreover, high-grade patients were enriched in the SL subtype (44% of grade 3, \( \chi^2 \) test \( P = 5.10e-05 \)). The LAR subtype showed poor MFS as well as OS, consistent with the primary TNBC of the Curtis 2012 dataset METABRIC project (Fig. 1C), and the patients with LAR subtype also rapidly relapsed in line with the SL subtype (median MFS = 4.9 years) and had 1.5 times the ratio of risk than the MSL group (95% CI, 1.03–2.37; Fisher test \( P = 0.03 \); ref. 42). The MSL and IM groups presented relatively good prognosis in terms of both MFS and OS.

**Comparison between primary and metastatic TNBC**

We investigated the unique features between primary and metastatic TNBC. Metastasis samples (37%; \( n = 356 \)) were originated from primary TNBC and derived from several tissues, including skin, liver, lung, and brain (Table 1). When clustering all samples, key pathways shared with primary TNBC also discriminated metastasis into four subtypes (Fig. 1A). Proportions of SL and LAR subtypes were increased in metastasis (SL 35%, LAR 35%) compared with primary TNBC (SL 30%, LAR 28%), and the remaining two subtypes in metastasis were relatively decreased. Meanwhile, tissue types of metastatic samples were independent with CMS confirmed by the \( \chi^2 \) test.

To characterize the functional difference between primary cases and metastasis, we first found 734 DEGs (adjusted \( P < 0.001 \)) and revealed the biological pathways that were enriched in the DEGs using ReactomeFI (Materials and Methods; Fig. 2). These genes were associated with complement and coagulation, IL6-mediated signaling, Toll-like receptors, TNF signaling, and JAK–STAT signaling pathways (\( P < 0.001 \); Fig. 2A). Serpin family E member 1 (SERPINE1) and fibroblast growth factor 1, implicated in metastasis (43, 44), were found to be involved in metastasis via EGFR and FGF signaling pathways and angiogenesis (Fig. 2A and C).

We established metastasis prediction models based on identified prognostic markers (Supplementary Fig. S5A). Metastasis prediction gene set was chosen using MFS test (\( P < 0.05 \); \( n = 84 \)) from 734 DEGs comparing primary with metastasis samples, and prediction model was established using random-forest machine-learning method (Supplementary Table S4). The model was validated by a fourfold cross-validation (AUC 0.8; Supplementary Fig. S5B). Importantly, we report concordance index from these metastasis prediction gene sets (Fig. 2B). Among metastasis prediction genes, CIs were greater than 0.4 in FGF, JAK–STAT, and integrin signaling regulated by SHC adaptor protein 1 (SHC1), integrin subunit alpha M (ITGRAM), vascular cell adhesion molecule 1 (VCAM1), integrin subunit alpha 1 (ITGA1), and jagged canonical Notch ligand 2 (JAG2; Supplementary Fig. S3C and SSD). SHC1 involved in FGF signaling and cell migration is used as a prognostic marker for cancer treatment (45). We show high concordance (\( c = 0.55; 95\% CI, 0.47–0.62 \)) of this gene, which was activated in metastasis (Fig. 2C). ITGRAM (\( c = 0.48; 95\% CI, 0.44–0.53 \)) and ITGA1 (\( c = 0.52; 95\% CI, 0.48–0.58 \)) seem to enhance distance metastasis via integrin signal and coagulation cascading (Fig. 2A; Supplementary Table S5). JAG2 (\( c = 0.46; 95\% CI, 0.43–0.52 \)) was identified as a member of NOTCH signaling that promotes metastasis (Fig. 2C).

**Pharmacogenomics landscape investigation by CMS**

We established a pharmacogenomics landscape depending on CMS. Computational DR scores were calculated from 181 response or...
resistance drug signatures from MSigDB. In total, 48 drugs of the 181 signatures extracted to overlap with breast cancer clinical trials. DR scores were estimated in both our aggregated data and biobank. We extracted final 18 DR scores of 17 drugs to indicate CMS specificity (ANOVA FDR < 0.05). Next, DR scores and correlated pathways (γ > 0.65) were extracted as network edges for each subtype (Fig. 3). DR score differences among our CMSs remarkably resemble with biobank (Supplementary Fig. S6A). Chosen 18 DRs to shown difference among CMSs acquired high CI and partially associated with clinical outcomes (Supplementary Fig. S6A). Adaphostin, doxorubicin, and bexarotene assigned SL type were ranked in top 3 concordance index (c) and shown significant clinical outcomes (MFS P value = 3.31E-05). Trabectedin response of SL and ribavirin response and imatinib resistance of MSL were closely relevant to MFS. Tamoxifen resistance was significantly detected in LAR.

Dasatinib, imatinib, and cisplatin were estimated to have the highest resistance scores in the MSL subtype (Fig. 3A), and the DR scores for these three drugs correlated strongly with the EMT and KRAS pathways (Fig. 3B). In particular, EMT activation led to cisplatin resistance by functionally shared genes such as cellular communication network factor 2 (CCN2), vimentin (VIM), cysteine-rich angiogenic inducer 61 (CYR61), and 5’-nucleotidase (NT5E; Fig. 3B). The genes involved in EMT activation are known to be associated with the therapeutic effect of cisplatin (46, 47). Ribavirin, an eIF4E inhibitor, has been suggested as a response drug in MSL subtype. The target gene is known to promote EMT and metastasis (Fig. 3A; ref. 48). Interestingly, our MSL drug-pathway network demonstrates that ribavirin inhibits TNFα and apoptosis pathways activated in MSL (Figs. 1A and 3B).

The LAR subtype appeared strongly high resistant score against tamoxifen and gefitinib, though androgen blockade was considered as a suitable therapy (Fig. 3A and B). As already known, AR promotes tamoxifen agonist activity. Our prediction shows a strong intervention in ER early response (Fig. 3B). In LAR network, the drug with the

Figure 2. Functional investigation and gene interactions of DEGs between primary and metastatic TNBC. A, Significantly enriched pathways derived from the DEG set between primary and metastatic TNBC. B, Concordance index plot for metastasis prediction markers. Cox regression analysis was performed between binary groups according to gene expression. C, Gene interaction network of DEGs after GSEA. Red nodes indicate activated genes in metastatic samples, and blue nodes indicate activated genes in primary samples. Node sizes are -log(P values), and P values were calculated by t test with primary and metastatic samples.
Figure 3.
Overall status of functional interactions between DR and key pathway within each CMS. A, Boxplots of DR scores estimated from aggregated data, MSL, LAR, and SL, respectively. B, Interaction networks with drugs and key pathway to highly correlate ($r > 0.65$) for each CMS. The subtype of each network is equally ordered within the boxplots. Square and circle nodes represent key pathways and drugs, respectively. Node colors of pathways indicate pathway activity scores. Red drug node indicates resistance signature and blue the response. The width of edges represents gene degree overlap in drug and pathway gene sets. Numbers on edge represent the greatest gene degree for each subtype. C, DR score heatmap estimated from the gene expression profiles of ROADMAP patients. D, PFS plot of ROADMAP patients that underwent cisplatin therapy. E, Two waterfall plots enumerating cisplatin resistance DR score estimated from the ROADMAP expression profiles colored based on CMS (left plot) and patients’ cisplatin response profiles (right plot). F, Cisplatin IC$_{50}$ boxplot of 64 biobank TNBC cells for nonresponders (NR) and responders (R). Right boxplot is cisplatin resistance DR scores of NR and R from the biobank.
highest resistance, gefitinib, inhibits insulin-like growth factor 1 receptor (IGF1R) signaling. Consequently, the LAR subtype shows significant fold change in IGF1R expression (fold change = 8.38). On the other hand, triluglitol was identified as the only response drug that inhibits fatty acid oxidation (49). Fatty acid metabolism and adipogenesis pathways were activated across the MSL and LAR subtypes (Fig. 1A), but the response to triluglitol was more closely correlated to the LAR subtype.

The SL has shown resistance in doxorubicin, dasatinib, and fluorouracil, but adaphostin, cisplatin, and triluglitol were responsible (Fig. 3A). E2F, G2M, and MYC targets 1 signal were positively correlated with doxorubicin resistance (Fig. 3B; Supplementary Fig. S6A). Cyclin B2 (CCNB2), extra spindle pole bodies like 1, separate (ESPL1), marker of proliferation Ki-67 (MKI67), and DNA topoisomerase II alpha (TOP2A) genes, involved in E2F and G2M pathways, induced drug resistance and showed high expression in SL. Generally, ESPL1 and TOP2A encode for topoisomerase or its binding partner, and the genes' topoisomerase activity is known to trigger doxorubicin resistance (50).

In addition, to examine cisplatin efficacy in genomic unstable samples, we calculated CIN scores for each of these cases. Genomic unstable cases of high CIN scores were classified as cisplatin treatment (Supplementary Fig. S7; response \( r = 0.41, P = 0.06 \)). Glycogen synthase kinase 3 (GSK3) inhibitor, activated in IM type, regulates innate and adaptive immune response mediated by NK-x8 and STATs (51). Interestingly, biobank data show that the GSK3 inhibitor SB216763 responds only to IM, but also to partial LAR and SL types (Supplementary Fig. S5; see Discussion for more details).

**ROADMAP data analysis of cisplatin treatment patients**

We validated the response of patients to cisplatin with our computational result based on the pharmacogenomics approach. Of 38 patients with TNBC, 31 patients received cisplatin-containing chemotherapy at the metastatic setting (Table 2; Supplementary Table S6). Our SVM model classified all patients into SL (55.3%), MSL (10.5%), LAR (31.6%), and IM (2.6%; Fig. 3C; Supplementary Fig. S7A). Key pathways inferred from aggregated data were also investigated in ROADMAP data. EMT, ER response, and MYC pathways clearly resembled subtype-dependent pathway activities of the aggregated data, except for the IM subtype, which was due to sample size limitation (Supplementary Fig. S8A and S8B). Approximately 52% of the 31 patients responded to cisplatin treatment (R group), whereas 48% were nonresponders (NR group). In progression-free survival (PFS) analysis, the NR group showed poor PFS (\( P = 0.002; \) Fig. 3D).

Computational cisplatin DR scores of ROADMAP remarkably associated with actual patients' DR, and the computational scores showed subtype-dependency (\( P < 0.1; \) Fig. 3E). Three patients with MSL except one belonged to the NR. In contrast, 6 of 7 patients with LAR showed response to cisplatin treatment. In other words, patients with MSL-type appeared to be cisplatin-resistant compared with other subtypes, whereas LAR type was most sensitive for this treatment (\( P = 0.1; \) Wilcoxon test). Our method predicted that ribavirin and thiazolidinedione (TZD) response scores resembled cisplatin resistance DR scores in ROADMAP patients (Fig. 3C, top rectangle). Cisplatin was suitable for the LAR and SL subtypes (Fig. 3C, middle rectangle). LAR showed both tamoxifen resistance and androgen target treatment response (Fig. 3C, bottom rectangle). In addition, investigation of biobank drug HTS and expression profiles indicated that cisplatin efficacy (IC\(_{50}\) value) was remarkably consistent with our DR score (\( P = 6.0e-04; \) Fig. 3F).

**Table 2. Molecular subtype and clinical summary of ROADMAP patients.**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>MSL</th>
<th>LAR</th>
<th>IM</th>
<th>SL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
<td>4</td>
<td>12</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>Age</td>
<td>( &lt; 50 ) y</td>
<td>3 (75)</td>
<td>5 (41.7)</td>
<td>–</td>
</tr>
<tr>
<td>( \geq 50 ) y</td>
<td>1 (25)</td>
<td>7 (58.3)</td>
<td>1 (100)</td>
<td>11 (52.2)</td>
</tr>
<tr>
<td>Sample type</td>
<td>Fresh frozen</td>
<td>2 (50)</td>
<td>2 (16.7)</td>
<td>–</td>
</tr>
<tr>
<td>FFPE</td>
<td>2 (50)</td>
<td>10 (83.3)</td>
<td>1 (100)</td>
<td>15 (71.4)</td>
</tr>
<tr>
<td>Metastases</td>
<td>No metastases</td>
<td>–</td>
<td>4 (33.3)</td>
<td>–</td>
</tr>
<tr>
<td>Metastases</td>
<td>4 (100)</td>
<td>8 (66.7)</td>
<td>1 (100)</td>
<td>12 (57.1)</td>
</tr>
<tr>
<td>No of chemo before biopsy</td>
<td>0</td>
<td>–</td>
<td>3 (25)</td>
<td>–</td>
</tr>
<tr>
<td>1</td>
<td>3 (75)</td>
<td>4 (33.3)</td>
<td>1 (100)</td>
<td>9 (42.9)</td>
</tr>
<tr>
<td>( \geq 2 )</td>
<td>1 (25)</td>
<td>5 (41.7)</td>
<td>–</td>
<td>7 (35.3)</td>
</tr>
<tr>
<td>Cisplatin response</td>
<td>Response</td>
<td>1 (25)</td>
<td>7 (58.4)</td>
<td>1 (100)</td>
</tr>
<tr>
<td>Nonresponse</td>
<td>3 (75)</td>
<td>1 (8.3)</td>
<td>–</td>
<td>11 (52.4)</td>
</tr>
<tr>
<td>Unknown</td>
<td>–</td>
<td>4 (33.3)</td>
<td>–</td>
<td>3 (14.3)</td>
</tr>
<tr>
<td>Survival status</td>
<td>Alive</td>
<td>–</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Dead</td>
<td>3</td>
<td>5</td>
<td>–</td>
<td>13</td>
</tr>
<tr>
<td>Unknown</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>1</td>
</tr>
</tbody>
</table>

Genes involved in cisplatin resistance were identified in ROADMAP using the DEG set and GSEA analysis (Supplementary Table S7). We found SL- and MSL-related pathways such as mitosis pathways, DNA damage bypass, Wnt signaling pathway, and Beta integrin cell surface pathway to be enriched among the 847 DR-related DEGs (\( P < 0.0018; \) Supplementary Table S8). We next explored cisplatin response markers for the prognosis of PFS based on the expression of genes, such as E2F transcription factor 1 (E2F1), metastasis associated 1 (MTA1), plakophilin 1 (PKP1), and microtubule nucleation factor (TPX2). Upregulation of these four genes, characterized by shorter response times, indicated poor prognosis (\( P < 0.04; \) Supplementary Fig. S8C) and was also consistent with our aggregated data (Supplementary Fig. S8D).

Our DR scores strongly suggest that patients with cisplatin nonresponse are mainly enriched in the MSL subtype, which could be treated with ribavirin and TZD (Fig. 3C). Ribavirin also inhibits histone methyltransferase (EZH2) and treats leukemia cells (52). TZD belongs to a class of drugs used to improve glucose metabolism in type-2 diabetes, activates PPAR-\( \gamma \) to crosstalk in mesenchymal stem cells, and inhibits adipocyte differentiation (53). In addition, these drugs interfere in MSL-associated pathways that activate PRC2, H3K27, EED, SUZ12 (related with methyltransferase activity), and TZD-related PPAR-\( \gamma \) inhibition of adipogenesis (Fig. 1B).

**Discussion**

Through this study, we proposed four TNBC subtypes using an integrative expression profile. The robustness of the classification was validated using several gene signatures and a machine-learning method (Supplementary Fig. S1). In our classification, the absence of DNA variant evidence could be a limitation, but tumorigenic functions by variants could be inferred from transcriptional consensus change (21, 26). For example, BRCA1/2...
mutation could be related to the SL subtype activated in pathways associated with cell cycle (54). Our pathogenic BRCA1/2 enrichment in SL was consistent with Lehmann’s identification in BL1 and BL2 (5). Meanwhile, our results also summarized key pathways specific to subtypes encompassing primary and metastatic TNBC. These consensus key pathways are in accordance with previous investigations on the ES signature of aggressive breast cancer from glioma and bladder tumor (25).

We investigated precise stenness using various computational approaches like machine-learning (Fig. 1C, left) or ES cell signature scoring regulated by transcription factor target gene set (Fig. 1C, right). ES cell signatures showed more enriched results in the SL subtype, indicating that transcription factors SOX2, OCT4, and NANOG play a critical role in proliferation (25). However, MYC target activation was unrestricted in the SL type differently than in a previous report (25). MYC-amplified samples except the SL type existed in the IM subtype. In another study, MYC reduction induced the downregulation of CD47 and PD-L1 in mouse and human cells, which was in line with our results (55). Therefore, we infer that MYC upregulation is involved across the SL and IM types. Meanwhile, we suggested another regulator, PRC2, which was known to be exclusively activated in nonstemness samples (25). We specifically revealed mutual exclusiveness of PRC2 targets between SL and MSL, and this finding narrows down a PRC2 complex inhibitor target (56).

Cellular heterogeneity and tumor progenitors of TNBC could be elucidated from various cells in the breast tissue microenvironment. Based on our cell abundance results, the proliferative characteristics of SL seem to be originated from luminal progenitor cells (Fig. 1C). In a previous study, BRCA1/2 defective tumor samples originated from luminal progenitors (10). As already described, the SL is the closest to BRCA1/2 deficiency. Therefore, our hypothesis is consistent with a previous study that reported that BRCA1/2 mutation frequently recurred in TNBC that originated from luminal progenitor and not from basal stem cells. Another study has shown that adipocytes induce EMT in breast cancer cells (57), which was relevant to our results that adipocytes were enriched in the MSL subtype (Fig. 1C). Thus, our cell abundance analysis indicates the specific progenitor or interaction with other cells in the microenvironment, which could be involved in tumorigenic biological functions.

We successfully established metastasis prediction machine-learning models using 84 DEG gene set to compare between primary and metastasis, and the genes could be considered as prognostic markers. The metastasis prediction gene set was enriched in coagulation, cell migration, and cell-cell interaction required for enhancing metastasis (45). Especially, one of the identified genes, SHC1, regulates apoptosis and drug resistance in mammalian cells and is essential for tumor progression and metastatic spread in breast cancer mouse models (54).

Our computational DR estimated from the expression profile successfully recapitulated the comprehensive characteristics of functional intervention and drug target. Cisplatin response markers of SL subtype, E2F1, MTA1, PKP1, and TPX2 are discriminative in PFS and MFS (Supplementary Fig. S8). We suggest that these four genes could emerge as important drug resistance predictors. Despite the limited patient expression profiles in ROADMAP data, we successfully predicted high cisplatin resistance score in MSL, whereas the LAR subtype was associated with good outcome (Fig. 3E). Further, we identified several candidate drugs that could intervene with key pathways activated in each subtype: aplatidin, TZDs, and ribavirin for the MSL subtype by inhibiting EMT and metastasis (58, 59), cisplatin and trabectedin for SL-type cases with defective DNA-repair machinery (60). Such approach is useful to perform preclinical prediction. Additional evaluation using biobank data further confirmed our data on cisplatin resistance and ribavirin response in MSL (Supplementary Fig. S6). Interestingly, the proportion of IM subtype (4.2%) in biobank is lesser than the CMS analysis dataset (12.9%), and it overlaps with some LAR and SL samples with respect to GSK3 inhibitor SB216763 response (Supplementary Fig. S6). It suggests that partial IM samples are predicted to LAR and SL. We infer that this limitation originates from artificial reconstruction of TME for organoid culture system. In spite of TME change in organoid culture, partial patients show consistent DR in GSK3 inhibitor consistently with fresh-frozen samples of our aggregated dataset.

In summary, we defined TNBC CMS encompassing primary and metastasis tumors using large-scale expression profiles. Our CMS recapitulated in-depth genomic function and microenvironment heterogeneity. Particularly, we propose the abundance of mammary gland cells as a marker for tumorigenic origin and prognosis. Further, our computational approach to estimate DR was remarkably consistent with patients’ responses to cisplatin and drug screening of biobank organoid models. Therapeutic candidates for each CMS were suggested with pathway intervention. Thus, our study identifies molecular subtypes and provides a basis for relevant diagnostic and prognostic markers. Finally, our systematic computational approach suggests a potential to complement TNBC and DR study.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: I.H. Park, C. Park
Development of methodology: D. Yu
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Kim, Y. Kwon, K.S. Lee, S.H. Sim, S.Y. Kong, E.S. Lee, I.H. Park
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J. Kim, D. Yu, K.S. Lee, S.H. Sim, C. Park
Writing, review, and/or revision of the manuscript: J. Kim, Y. Kwon, K.S. Lee, I.H. Park
Study supervision: I.H. Park, C. Park

Acknowledgments
This research was supported by the National Cancer Center Grant (NCC-1611150, NCC-1910100) and the National Research Foundation of Korea (NRF, NRF-2018R1C1B6001475). We would like to acknowledge the patients with TNBC and the clinical help obtained for the ROADMAP project. TNBC gene expression meta-data were deposited in GEO and ROADMAP next-generation sequencing files have been deposited in NCBI SRA accession SRP156081. Our TNBC classification approach is provided as an R package in Bioconductor (https://bioconductor.org/packages/devel/bioc/html/TNBC.CMS.html).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 30, 2019; revised July 15, 2019; accepted November 4, 2019, published first November 8, 2019.
Kim et al.

References


TNBC Subtypes and Chemotherapy Response
Molecular Cancer Research

Genomic Characteristics of Triple-Negative Breast Cancer Nominate Molecular Subtypes That Predict Chemotherapy Response

Jihyun Kim, Doyeong Yu, Youngmee Kwon, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1541-7786.MCR-19-0453

Supplementary Material
Access the most recent supplemental material at:
http://mcr.aacrjournals.org/content/suppl/2019/11/08/1541-7786.MCR-19-0453.DC1

Cited articles
This article cites 60 articles, 10 of which you can access for free at:
http://mcr.aacrjournals.org/content/18/2/253.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link http://mcr.aacrjournals.org/content/18/2/253.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.