Directed Bayesian Networks established functional differences between breast cancer subtypes

Lucía Trilla-Fuertes[1¶], Andrea Zapater-Moros[2¶], Angelo Gámez-Pozo[1,2], Jorge M Arevalillo[3], Guillermo Prado-Vázquez[1], Mariana Díaz-Almirón[4], María Ferrer-Gómez[2], Rocío López-Vacas[2], Hilario Navarro[3], Enrique Espinosa[5,6], Paloma Maín[7], and Juan Ángel Fresno Vara[1,2,7*]

[1]Biomedica Molecular Medicine SL, Madrid, Spain

[2]Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Madrid, Spain

[3]Operational Research and Numerical Analysis, National Distance Education University (UNED), Spain

[4]Biostatistics Unit, La Paz University Hospital-IdiPAZ, Madrid, Spain

[5]Medical Oncology Service, La Paz University Hospital-IdiPAZ, Madrid, Spain

[6] Biomedical Research Networking Center on Oncology-CIBERONC, ISCIII, Madrid, Spain

[7]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain

¶ These authors contributed equally to this work.

*Corresponding Author:

E-mail: juanangel.fresno@salud.madrid.org (JAFV)

Short title: Directed networks established functional differences in breast cancer

Abstract

Breast cancer is a heterogeneous disease. In clinical practice, tumors are classified as hormonal receptor positive, Her2 positive and triple negative tumors. In previous works, our group defined a new hormonal receptor positive subgroup, the TN-like subtype, which has a prognosis and a molecular profile more similar to triple negative tumors. In this study, proteomics and directed Bayesian networks were used to characterize protein relationships in 106 breast tumor samples. Components obtained by these methods had a clear functional structure. The analysis of these components suggested differences in processes such as metastasis or proliferation between breast cancer subtypes, including our new subtype TN-like. In addition, one of the components, mainly related with metastasis, had prognostic value in this cohort. Functional approaches allow to build hypotheses about regulatory mechanisms and to establish new relationships among proteins in the breast cancer context.

Author Summary:

Breast cancer classification in the clinical practice is defined by three biomarkers (estrogen receptor, progesterone receptor and HER2) into hormone receptor positive, HER2+ and triple negative breast cancer (TNBC). Our group recently described a new ER+ subtype with molecular characteristics and prognosis similar to TNBC. In this study we propose a mathematical method, the directed Bayesian networks, as a useful tool to study protein interactions and differential biological processes in breast cancer subtypes, characterizing differences in relevant processes such as proliferation or metastasis and associated them with patient prognosis.

Introduction

Breast cancer is one of the most prevalent cancers in the world [1]. In clinical practice, breast cancer is classified according to the expression of hormonal receptors (estrogen or progesterone) and Her2, into positive hormonal receptor (ER+), HER2+ and triple negative (TNBC). In previous studies, our group defined a new ER+ molecular subgroup, named TN-like, with a molecular profile and a prognosis more similar to TNBC tumors [2]. The remaining ER+ tumors were considered as ER-true. We also found significant molecular differences among breast cancer subtypes. For instance, differences related with metabolism of glucose were described between ER-true, TN-like and TNBC tumors [2, 3].

Proteomics provides useful information about biological process effectors and may quantify thousands of proteins. Undirected probabilistic graphical models (PGM), based on a Bayesian approach, allow characterizing differences between tumor samples at functional level [2-5]. In this study we explored the utility of directed Bayesian networks in the molecular characterization of breast cancer. The main feature of targeted Bayesian networks is that they provide a hierarchical structure and targeted relationships between proteins.

In this work, we used proteomics and directed Bayesian networks to characterize protein relationships in a cohort of breast cancer tumor samples. These networks maintain a functional structure and it is possible to use this information to build prognostic signatures. This approach also reflected previously described interactions and it could be used to propose new hypotheses and mechanisms of regulation of these proteins.

Results

Patient characteristics

Clinical characteristics of this patient cohort have been previously described [2, 3, 6]. Briefly, one hundred and six patients were enrolled into the study. They all had node positive disease, Her2 negative and all had received adjuvant chemotherapy and hormonal therapy in the case of ER+ tumors. Among ER+ tumors, 50 patients were characterized as ER-true and 21 were defined as TN-like (Sup Table 1) [2].

MS analysis

Proteomics analyses from these samples has been previously described [3]. In summary, one hundred and two FFPE samples had enough protein to perform the MS analyses. After MS workflow, 96 samples provided useful protein expression data. After quality criteria, 1,095 proteins presented at least two unique peptides and detectable expression in at least 75% of the samples in at least one type of sample (either ER+ or TNBC).

Directed networks

Using proteomics data, directed acyclic graphs (DAG) were performed. Altogether, it was possible to establish 536 edges of which 414 are guided and 122 are undetermined. These edges formed 377 components formed by different number of nodes or proteins. An overview of the number of nodes (proteins) included in each component is provided in table 1.

Number of nodes	1	2	3	4	5	6	7	8	9	10	11	13	14	19	21	26	27	34	36
Number of components	229	70	24	12	15	2	6	3	4	2	1	2	1	1	1	1	1	1	1

Table 1: Characteristics of the components obtained from DAG. Number of nodes = number of proteins contained in each component, Number of components = directed components obtained.

84 We characterized components from DAG analysis. Components including less than 9 nodes
85 were dismissed because they were little informative. All components were named with the
86 number of nodes included by the DAG analysis, and the information was completed with
87 protein-protein interactions (PPI) based on experimental evidence obtained from Genes2FANS
88 (G2F) [7]. For example, component 14 includes 19 nodes, 14 nodes defined from our directed
89 Bayesian analysis and 5 extra nodes added by G2F.

90 Afterwards, components were interrogated for biological function. For components with less
91 than 27 nodes provided by DAG analysis, this functional analysis was performed by
92 bibliographical review. Thus, the main biological function of component 14 is metabolism.
93 Regarding components which had more than 27 nodes provided by the DAG analysis, gene
94 ontology analyses were done once they were completed by G2F information. Characteristics
95 about all components are supplied in table 2 and Supplementary File 1.

Component	Main function	Total number of nodes
Component 36	Extracellular exosome	65 nodes
Component 34	RNA processing and mTOR pathway	61 nodes
Component 19	Proliferation	26 nodes
Component 21	Proliferation & metastasis	34 nodes
Component 27	Metastasis	34 nodes
Component 14	Metabolism	19 nodes
Component 13	Proteasome	19 nodes
Component 11	None main function assigned	14 nodes
Component 10b	Proliferation	19 nodes
Component 9a	Growth	15 nodes
Component 9b	Proliferation & metastasis	22 nodes
Component 9c	Glycolysis	12 nodes
Component 9d	Transcription & ribosomes	20 nodes

96 **Table 2: Features of components obtained by DAG and G2F analysis.**

97 Component activity measurements

98 Component activities were calculated for each node. There were significant differences
99 between ER-true, TN-like and TNBC tumors in the component activity for component 34:
100 mRNA processing and mTOR, component 36: extracellular exosome, component 21:
101 proliferation and metastasis, component 27: metastasis, component 9a: growth, component

102 9b: proliferation and metastasis, component 9c: glycolysis, component 10b: metastasis,
103 component 13: proteasome and component 9d: transcription (Fig 1).

104 **Fig 1: Component activity measurements for ER-true, TN-like and TNBC respectively.**

105 Component 10b: metastasis

106 Component 10b activity showed prognostic value in our series, splitting our population into a
107 high and a low risk group and it can be used to make a distant metastasis-free survival (DMFS)
108 predictor (p=0.0068, HR= 0.2854, cut-off 40:60) (Fig 2). Additionally, dividing by molecular
109 subtype, this prognostic signature also split the population into low and high risk groups
110 although it is not statistically significant (Fig 3).

111 **Fig 2: Component 10b activity prognostic value in the whole cohort.**

112 **Fig 3: Component 10b activity prognostic value by subtype.**

113 Component 10b contains 10 proteins. Four out of nine edges (DYNLRB1-HSPH1; HSPH1-CFL1;
114 HBB-UCHL5; and UCHL5-CFL1) have been previously described in G2F database [7]. Then, G2F
115 added two more proteins (HSPH1 and UCHL5) to this component. Most of the proteins
116 included in this component are related with metastasis processes [8-12], whereas PBDC1
117 doesn´t have any associated function. On the other hand, HIST2H2BF is a histone and
118 HIST2H3PS2 is a histone pseudogene showing a directed relationship (Fig 4).

119 **Fig 4: Component 10b merged with PPI information provided by Genes2FANS.** Red arrows
120 indicate relationships from DAG; green lines indicate relationships from Genes2FANS database.
121 Orange nodes are common nodes between these two approaches, purple nodes are nodes
122 which come only from DAG and blue nodes are only from Genes2FANS.

123

124 # Discussion

125 In this study, we used proteomics and DAG to characterize relationships between proteins in
126 breast cancer tumor samples. Unlike other approaches, such as G2F [7], our DAG method
127 supplies directed relationships between proteins and a hierarchical structure. Traditionally, PPI
128 networks are based in relationships described in the literature. However, we built a directed
129 network, i.e. a graph formed by edges with a direction, using protein expression data without
130 other *a priori* information, so it was possible to propose new hypotheses about protein
131 interactions. We used probabilistic graphical models (PGM) because they offer a way to relate
132 many random variables with a complex dependency structure.

133 Arrows in directed networks indicate causality between two proteins, i.e. if protein A and B are
134 connected and protein A changes its expression value, protein B changes its expression value
135 as well. This approach allows making hypotheses about causal relationships between proteins
136 and proposes a hierarchical structure. G2F relationships supplies additional information to our
137 directed networks, so it served as a validation of the network coherence, i.e., in some cases, an
138 experimental relationship between two proteins connected in the directed network had been
139 previously described. In component 10b, for instance, CFL1 and DYNRBL1 had a described
140 common nexus in HSPH1 [7]. It is remarkable that, in component 27, DAG analysis established
141 a relationship between PTRF and PRDXCDBP which is widely described [13].

142 We demonstrated in previous works that non-directed Bayesian graphs provided functional
143 information [2-5]. Interestingly, a functional structure also appeared in this type of networks.
144 Component activities suggested differences in functions such as metastasis, proliferation,
145 proteasome or glycolysis. We have previously described differences in proteasome and
146 glycolysis between ER-true and TN-like subtypes using non-directed Bayesian networks [2].

147 On the other hand, it is widely known the role that actin-cytoskeleton plays and its regulation
148 in directional migration and metastasis, Therefore, it is not surprising that the proteins that
149 make up component 10b had some relation to this biological function and had a prognostic
150 value. Using component 10b activity, based on these proteins related with metastasis, it is
151 possible to split our population into a low and a high risk of relapse groups and, interestingly,
152 this prognostic value is also showed in the analysis by subtype. In previous studies we have
153 used functional node activities from non-directed Bayesian network to develop prognostic
154 predictors [4, 5]. Now, this approach is also validated in directed networks.
155 Component 10b presents three proteins showing influence over others: HIST2H2BF, DYNLRB1
156 and RHOG. Dynein light chain roadblock-type 1 (DYNRBL1), also known as km23-1, is a
157 component of the cytoplasmic dynein 1 complex. In colon cells, its depletion can block events
158 known to be involved in cell invasion and tumor metastasis, and it is also an actin cytoskeletal
159 linker critical for the dynamic regulation of cell motility and invasion since it induces a highly
160 organized actin stress fiber network [14]. DYNRBL1 regulates RhoA and motility-associated
161 actin modulating proteins, such as cofilin1 (CFL1) and coronin [15], suggesting that DYNRBL1
162 may represent a novel target for anti-metastatic therapy [9].
163 On the other hand, Ras-homolog family member G (RhoG) encodes a member of the Rho
164 family of small GTPases. Constitutively active RhoG induces morphological and cytoskeletal
165 changes similar to those induced by the combination of active Rac and Cdc42 working
166 upstream of them. Rac is activated at the leading edge of motile cells and induces the
167 formation of actin-rich lamellipodia protrusions, which serves as a major driving force of
168 cancer cell. The major downstream proteins for Rac are the WAVE family proteins, the
169 activators of the Arp2/3 complex. In addition, RhoG induces translocation of Dock4-ELMO
170 complex to the plasma membrane and enhances Rac1 activation which promotes migration
171 [16]. Moreover, RNA interference-mediated knockdown of RhoG in HeLa cells reduced cell
172 migration in Transwell and scratch-wound migration assays [11].

173 HIST2H2BF has been proposed as a pancreatic ductal adenocarcinoma biomarker[17], but
174 there is no available information about its role in metastasis or breast cancer.
175 In component 10b, both DYNLRB1 and RHOG expression modulates cofilin1 (CFL1). CFL1 is a
176 widely distributed intracellular actin-modulating protein that binds and depolymerizes
177 filamentous F-actin and inhibits the polymerization of monomeric G-actin in a pH-dependent
178 manner. Increased phosphorylation of this protein by LIM kinase aids in Rho-induced
179 reorganization of the actin cytoskeleton [18]. CFL1 is an actin-severing protein that creates free
180 barbed ends in the actin-severing process. Arp2/3 binding to these barbed ends allows the
181 elongation of new actin filaments. This synergy between CFL1 severing activity and Arp2/3-
182 generated dendritic nucleation resulting in the formation of stable invadopods and a
183 directional migration, linking CFL1and ARP2/3, through RhoG, to tumor cell invasion [15, 19].
184 CFL1 pathway was previously related with metastases processes in breast cancer [18].
185 Furthermore, DYNRBL1 is required for TGFb1 secretion. It seems to be TGFb pathway plays an
186 important role in cell migration in this sense, because TGFbRII signaling regulates Rho GTPase
187 degradation and actin dynamics. Moreover, TGFb pathway induces expression of the GEF NET1
188 which accumulates and promotes actin polymerization and on the other hand, it induces
189 expression of tropomyosins that regulate assembly of a contractile apparatus that serves the
190 motility of the responding cell during the EMT process [15] Additionally, TGFß production by
191 macrophages or dendritic cells (DCs) that have engulfed apoptotic cells can promote the
192 generation of inducible regulatory T (Tregs) that play a known protumoral role [16].
193 Therefore, the relationships between CFL1, RHOG and DYNLRB1 are well-established.
194 Interestingly, DAG graph adds the decorin (DCN) to these edges. DCN is a small leucine-rich
195 proteoglycan that promotes matrix organization by decreasing collagen uptake/degradation,
196 which constitutes a physical barrier against migration/motility of cancerous cells. The
197 alteration of the matrix stiffness leads to differential integrin activation and changes in
198 cytoskeletal organization by Rac, affecting cell motility and invasiveness [20]. Moreover, DCN

199 acts as a matrikine whose interaction with CXC chemokine receptor 4 (CXCR4) impairs its bind
200 with stromal cell-derived factor-1a (SDF-1a), preventing directional migration [20]. In breast
201 cancer tumors, high DCN expression in stroma correlated with lower tumor grade, low Ki67
202 levels and ER positivity. On the other hand, high expression of DCN in the malignant epithelium
203 was correlated with LN positivity, higher number of positive lymph nodes and HER2-positive
204 status compared to patients with low DCN expression [21].
205 By last, lysyl-tRNA synthetase (KARS), also known by KRS, was recently shown to induce cancer
206 cell migration through its interaction with the 67-kDa laminin receptor (67LR) who binds
207 laminin on ECM. The interaction of KRS with 67LR enhances the membrane stability of 67LR,
208 which in turn results in an increase laminin-dependent cell migration in metastasis [18]. KARS,
209 causes incomplete epithelial-mesenchymal transition and ineffective cell-extracellular matrix
210 adhesion for migration [22].
211 We used a mathematical method as DAG analysis and applied them to proteomics data of
212 breast cancer tumors in order to infer causal relationships between these proteins. This
213 method supplied some known relationships but also proposed new ones. Additionally, it
214 associated proteins with a similar function. Therefore, it seems that it is a good approach to
215 propose new hypotheses about mechanisms of action. Moreover, it was possible to associate
216 the results obtained by DAG analysis with prognosis and built a prognostic signature. As far we
217 know, this is the first time that this type of analysis is applied to clinical data and is associated
218 with clinical outcome.
219 Our study has some limitations. Proteomics provides complementary information to other
220 techniques such as genomics. However, an improvement in the number of detected proteins is
221 still necessary. Validation at the cellular level is also needed.

222 To sum up, in this study, we used proteomics and directed Bayesian networks to characterize
223 relationships between proteins in breast cancer tumors. This approach reflected some

224 previously described interactions and it could be used to propose new hypotheses and

225 mechanisms of action.

226

227 # Material and methods

228 ## Samples

229 One hundred and six FFPE samples from patients with breast cancer were recovered from I+12
230 Biobank and from IdiPAZ Biobank, both integrated in the Spanish Hospital Biobank Network.
231 The histopathological characteristics were reviewed by a pathologist to confirm tumor
232 content. Samples had to include no less than half of tumor cells. The endorsement of the
233 study was obtained by Hospital Doce de Octubre and Hospital Universitario La Paz Ethical
234 Committees. These samples were utilized in previous studies [2, 3, 6].

235 ## Protein preparation

236 Proteins were extracted from FFPE samples as previously described [23]. Briefly, FFPE sections
237 were deparaffinized in xylene and washed twice with absolute ethanol. Protein extracts from
238 FFPE samples were set up in 2% SDS buffer using a protocol based on heat-induced antigen
239 retrieval. Protein concentration was quantified using the MicroBCA Protein Assay Kit (Pierce-
240 Thermo Scientific). Protein extracts (10 µg) were processed with trypsin (1:50) and SDS was
241 removed from digested lysates using Detergent Removal Spin Columns (Pierce). Peptide
242 samples were additionally desalted using ZipTips (Millipore), dried, and resolubilized in 15 µL
243 of a 0.1% formic acid and 3% acetonitrile solution before MS experiments.

244 ## Label-free proteomics

245 Samples were analyzed on a LTQ-Orbitrap Velos hybrid mass spectrometer (Thermo Fischer
246 Scientific, Bremen, Germany) coupled to NanoLC-Ultra system (Eksigent Technologies, Dublin,
247 CA, USA) as described previously [2, 3]. Briefly, after separation, peptides were eluted with a
248 gradient of 5 to 30% acetonitrile in 95 minutes. The mass spectrometer was operated in data-
249 dependent mode (DDA), followed by CID (collision-induced dissociation) fragmentation on the
250 twenty most intense signals per cycle. The acquired raw MS data were processed by MaxQuant

251 (version 1.2.7.4) [24], followed by protein identification using the integrated Andromeda
252 search engine [25]. Briefly, spectra were searched against a forward UniProtKB/Swiss-Prot
253 database for human, concatenated to a reversed decoyed fasta database (NCBI taxonomy ID
254 9606, release date 2011-12-13). The maximum false discovery rate (FDR) was set to 0.01 for
255 peptides and 0.05 for proteins. Label free quantification was calculated on the basis of the
256 normalized intensities (LFQ intensity). Quantifiable proteins were defined as those detected in
257 at least 75% of samples in at least one type of sample (either ER+ or TNBC samples) showing
258 two or more unique peptides. Only quantifiable proteins were considered for subsequent
259 analyses. Protein expression data were log2 transformed and missing values were replaced
260 using data imputation for label-free data, as explained in [26], using default values. Finally,
261 protein expression values were z-score transformed. Batch effects were estimated and
262 corrected using ComBat [27]. All the mass spectrometry raw data files acquired in this study
263 may be downloaded from Chorus (http://chorusproject.org) under the project name Breast
264 Cancer Proteomics.

265 Network construction

266 PGM are graph-based representations of joint probability distributions where nodes represent
267 random variables and edges (directed or undirected) represent stochastic dependencies
268 among the variables. In particular, we have used a type of PGM called Bayesian networks (BN)
269 [28]. With these models, the dependences between the variables in our data are specified by a
270 directed acyclic graph (DAG).

271 Firstly, we find the BN that best explains our data [29]. There are different algorithms to learn
272 a DAG from data but we have selected the well-known PC algorithm, a constraint-based
273 structure learning algorithm [30] based on conditional independence tests. The PC algorithm
274 was shown to be consistent in some high-dimensional settings [31] and have some other
275 statistical properties which make it very useful. In this way, our data are represented by a large

276 graph that can be partitioned into several connected components. Then, we focused on
277 finding suitable subgraphs that give us a much more clear understanding of the interrelations
278 therein. All these procedures are implemented in R [32] within packages *pcalg* [31] and *graph*
279 [33]. We used protein expression data without other *a priori* information.

280 With the aim of compare and complete the information provided by BN, the tool Genes2FANS
281 (G2F), developed by Ma'ayan 's group was used [7]. This software provides information about
282 protein-protein interactions (PPI) based on experimental evidence. A PPI network using G2F
283 was built for each component and finally, both networks, the DAG and the PPI network, were
284 merged.

285 Gene Ontology Analyses

286 Protein to Gene Symbol conversion was performed using Uniprot (www.uniprot.org) and
287 DAVID (www.david.ncifcrf.gov) [34]. Gene Ontology Analysis was also done in DAVID selecting
288 only "*Homo sapiens*" background and GOTERM-FAT, Biocarta and KEGG databases. As
289 concerns small connected components, a search in the literature was done so as to establish
290 component main functions.

291 Component activity measurements

292 Component activities were calculated as previously described [3, 4]. Briefly, activity
293 measurement was calculated by the mean expression of all the proteins of each component
294 related with the established major component function.

295 Statistical analyses

296 Network analyses were performed using Cytoscape software [35]. Statistical analyses were
297 done in GraphPad Prism v6. Prognostic signatures were developed using R v3.2.4 and BRB
298 Array Tools, developed by Richard Simon and BRB Array Tools Development Team [36]. P-
299 values under 0.05 were considered statistically significant.

Funding statement

This study was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" (PI15/01310). LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). GP-V is supported by Conserjería de Educación, Juventud y Deporte of Comunidad de Madrid (IND2017/BMD7783). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All the authors have directly participated in the preparation of this manuscript and have approved the final version submitted. RL-V prepared the proteomics samples. JMA, MD-A, HN and PM contributed the directed graphical models. LT-F, AZ-M, AG-P, G-PV and MF-G performed the statistical analyses, the directed graphical model interpretation and the review of the literature. LT-F, AG-P, JAFV, and EE conceived of the study and participated in its design and interpretation. LT-F and AZ-M drafted the manuscript. AG-P, JAFV, and EE supported the manuscript drafting. AG-P and JAFV coordinated the study.

Competing interests

JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F and GP-V are employees of Biomedica Molecular Medicine SL. The other authors declare no competing interests.

Data Availability Statement

All the mass spectrometry raw data files acquired in this study may be downloaded from Chorus (http://chorusproject.org) under the project name Breast Cancer Proteomics.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359-86. Epub 2014/10/09. doi: 10.1002/ijc.29210. PubMed PMID: 25220842.
2. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Scientific Reports. 2017;7(1):10100. doi: 10.1038/s41598-017-10493-w.
3. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res; 2015. p. 2243-53.
4. de Velasco G, Trilla-Fuertes L, Gamez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepúlveda JM, et al. Urothelial cancer proteomics provides both prognostic and functional information. Sci Rep. 2017;7(1):15819. Epub 2017/11/17. doi: 10.1038/s41598-017-15920-6. PubMed PMID: 29150671; PubMed Central PMCID: PMCPMC5694001.
5. Trilla-Fuertes L, Gamez-Pozo A, M Arevalillo J, Diaz-Almiron M, Prado-Vazquez G, Zapater-Moros A, et al. Molecular characterization of breast cancer cell response to metabolic drugs. Oncotarget2018.
6. Gámez-Pozo A, Trilla-Fuertes L, Prado-Vázquez G, Chiva C, López-Vacas R, Nanni P, et al. Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics. PLoS One. 2017;12(6):e0178296. Epub 2017/06/08. doi: 10.1371/journal.pone.0178296. PubMed PMID: 28594844; PubMed Central PMCID: PMCPMC5464546.
7. Dannenfelser R, Clark NR, Ma'ayan A. Genes2FANs: connecting genes through functional association networks. BMC Bioinformatics. 2012;13:156. doi: 10.1186/1471-2105-13-156. PubMed PMID: 22748121; PubMed Central PMCID: PMCPMC3472228.
8. Kim DG, Choi JW, Lee JY, Kim H, Oh YS, Lee JW, et al. Interaction of two translational components, lysyl-tRNA synthetase and p40/37LRP, in plasma membrane promotes laminin-dependent cell migration. FASEB J. 2012;26(10):4142-59. Epub 2012/07/02. doi: 10.1096/fj.12-207639. PubMed PMID: 22751010.
9. Jin Q, Pulipati NR, Zhou W, Staub CM, Liotta LA, Mulder KM. Role of km23-1 in RhoA/actin-based cell migration. Biochem Biophys Res Commun. 2012;428(3):333-8. Epub 2012/10/15. doi: 10.1016/j.bbrc.2012.10.047. PubMed PMID: 23079622; PubMed Central PMCID: PMCPMC3513371.
10. Madak-Erdogan Z, Ventrella R, Petry L, Katzenellenbogen BS. Novel roles for ERK5 and cofilin as critical mediators linking ERα-driven transcription, actin reorganization, and invasiveness in breast cancer. Mol Cancer Res. 2014;12(5):714-27. Epub 2014/02/06. doi: 10.1158/1541-7786.MCR-13-0588. PubMed PMID: 24505128; PubMed Central PMCID: PMCPMC4020978.
11. Katoh H, Hiramoto K, Negishi M. Activation of Rac1 by RhoG regulates cell migration. J Cell Sci. 2006;119(Pt 1):56-65. Epub 2005/12/08. doi: 10.1242/jcs.02720. PubMed PMID: 16339170.
12. Järvinen TA, Prince S. Decorin: A Growth Factor Antagonist for Tumor Growth Inhibition. Biomed Res Int. 2015;2015:654765. Epub 2015/11/30. doi: 10.1155/2015/654765. PubMed PMID: 26697491; PubMed Central PMCID: PMCPMC4677162.
13. Mohan J, Morén B, Larsson E, Holst MR, Lundmark R. Cavin3 interacts with cavin1 and caveolin1 to increase surface dynamics of caveolae. J Cell Sci. 2015;128(5):979-91. Epub 2015/01/14. doi: 10.1242/jcs.161463. PubMed PMID: 25588833.

14. Jin Q, Ding W, Mulder KM. The TGFβ receptor-interacting protein km23-1/DYNLRB1 plays an adaptor role in TGFβ1 autoinduction via its association with Ras. Journal of Biological Chemistry. 2012;287(31):26453-63.
15. Moustakas A, Heldin C-H. Dynamic control of TGF-β signaling and its links to the cytoskeleton. FEBS letters. 2008;582(14):2051-65.
16. Green DR, Ferguson T, Zitvogel L, Kroemer G. Immunogenic and tolerogenic cell death. Nature reviews Immunology. 2009;9(5):353.
17. Castillo J, Bernard V, San Lucas FA, Allenson K, Capello M, Kim DU, et al. Surfaceome profiling enables isolation of cancer-specific exosomal cargo in liquid biopsies from pancreatic cancer patients. Ann Oncol. 2017. Epub 2017/09/25. doi: 10.1093/annonc/mdx542. PubMed PMID: 29045505.
18. Cho HY, Ul Mushtaq A, Lee JY, Kim DG, Seok MS, Jang M, et al. Characterization of the interaction between lysyl-tRNA synthetase and laminin receptor by NMR. FEBS letters. 2014;588(17):2851-8.
19. Hiramoto K, Negishi M, Katoh H. Dock4 is regulated by RhoG and promotes Rac-dependent cell migration. Experimental cell research. 2006;312(20):4205-16.
20. Feugaing DDS, Götte M, Viola M. More than matrix: the multifaceted role of decorin in cancer. European journal of cell biology. 2013;92(1):1-11.
21. Cawthorn TR, Moreno JC, Dharsee M, Tran-Thanh D, Ackloo S, Zhu PH, et al. Proteomic analyses reveal high expression of decorin and endoplasmin (HSP90B1) are associated with breast cancer metastasis and decreased survival. PloS one. 2012;7(2):e30992.
22. Nam SH, Kang M, Ryu J, Kim HJ, Kim D, Kim DG, et al. Suppression of lysyl-tRNA synthetase, KRS, causes incomplete epithelial-mesenchymal transition and ineffective cell‐extracellular matrix adhesion for migration. Int J Oncol. 2016;48(4):1553-60. Epub 2016/02/08. doi: 10.3892/ijo.2016.3381. PubMed PMID: 26891990.
23. Gámez-Pozo A, Ferrer NI, Ciruelos E, López-Vacas R, Martínez FG, Espinosa E, et al. Shotgun proteomics of archival triple-negative breast cancer samples. Proteomics Clin Appl. 2013;7(3-4):283-91. doi: 10.1002/prca.201200048. PubMed PMID: 23436753.
24. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26(12):1367-72. Epub 2008/11/26. doi: 10.1038/nbt.1511. PubMed PMID: 19029910.
25. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794-805. Epub 2011/01/25. doi: 10.1021/pr101065j. PubMed PMID: 21254760.
26. Deeb SJ, D'Souza RC, Cox J, Schmidt-Supprian M, Mann M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Mol Cell Proteomics. 2012;11(5):77-89. Epub 2012/03/21. doi: 10.1074/mcp.M111.015362. PubMed PMID: 22442255; PubMed Central PMCID: PMCPMC3418848.
27. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27. doi: 10.1093/biostatistics/kxj037. PubMed PMID: 16632515.
28. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann2014.
29. Neapolitan RE. Learning Bayesian Networks . Series in Artificial Intelligence. Prentice Hall2004.
30. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Adaptive Computation and Machine Learning. 2nd ed. The MIT Press2000.
31. Kalisch M, Maechler M, Colombo D, Maathius MH, Bluehlmann P. Causal Inference Using Graphical Models with the R Package pcalg. Journal of Statistical Software2012. p. 1-26.

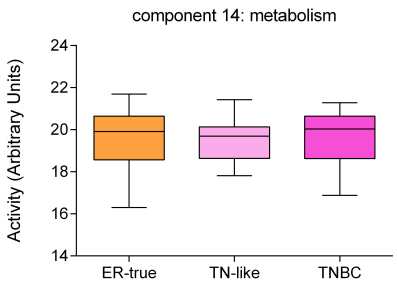
419 32. R Core Team. R: A language and environment for statistical computing. Vienna,Austria.
420 R Foundation for Stattistical Computing, 2013.
421 33. Gentleman R, Whalen E, Huber W, Falcon S. graph: A package to handle graph data
422 structures. R package version 1.54.0.

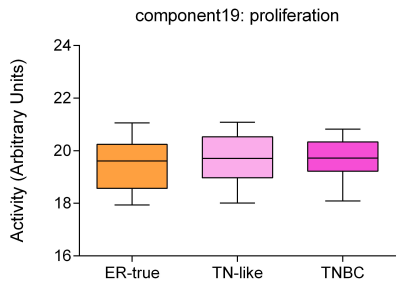
423 34. Huang dW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene
424 lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57. doi:
425 10.1038/nprot.2008.211. PubMed PMID: 19131956.
426 35. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a
427 software environment for integrated models of biomolecular interaction networks. Genome
428 Res. 2003;13(11):2498-504. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed
429 Central PMCID: PMCPMC403769.
430 36. Simon R. Roadmap for developing and validating therapeutically relevant genomic
431 classifiers. J Clin Oncol. 2005;23(29):7332-41. Epub 2005/09/06. doi:
432 10.1200/JCO.2005.02.8712. PubMed PMID: 16145063.

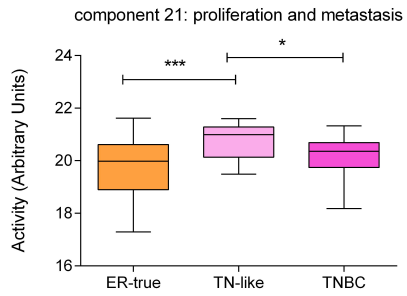
433 Supporting information captions

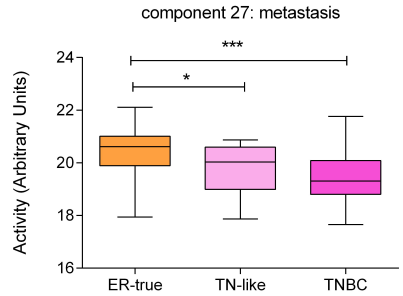
434 **Sup Table 1: Clinical characteristics of patient cohort.**

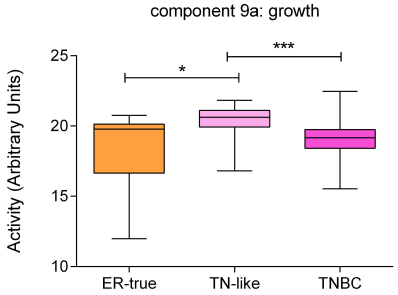
435 **Sup File 1: Components defined by DAG analysis.**

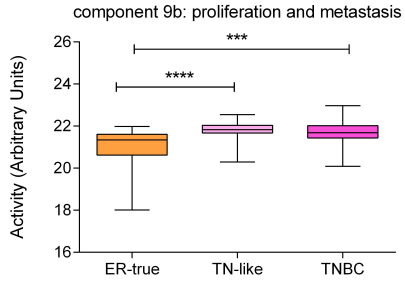

component 14: metabolism
Activity (Arbitrary Units)
24
22
20
18
16
14
ER-true
TN-like
TNBC

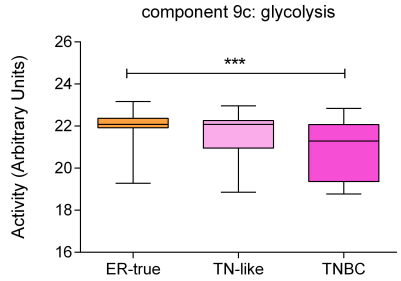

component19: proliferation
Activity (Arbitrary Units)
24
22
20
18
16
ER-true
TN-like
TNBC

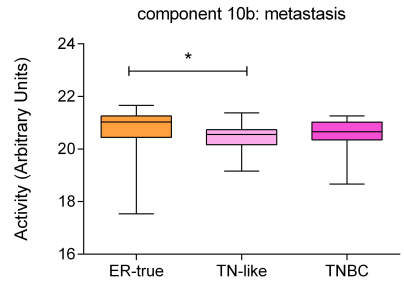

component 21: proliferation and metastasis
Activity (Arbitrary Units)
***
*
24
22
20
18
16
ER-true
TN-like
TNBC

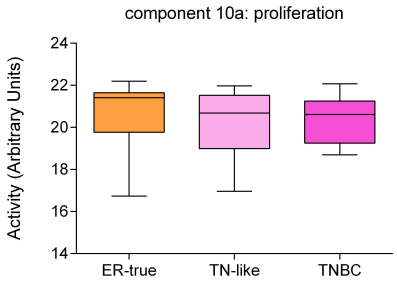

component 27: metastasis
Activity (Arbitrary Units)
***
*
24
22
20
18
16
ER-true
TN-like
TNBC

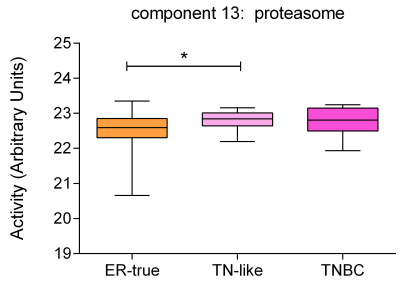

component 9a: growth
Activity (Arbitrary Units)
*
***
25
20
15
10
ER-true
TN-like
TNBC

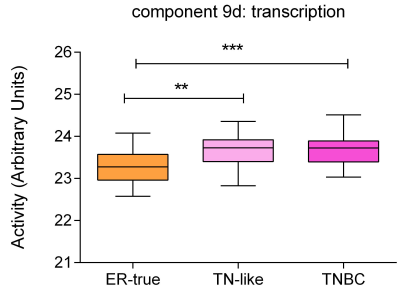

component 9b: proliferation and metastasis
Activity (Arbitrary Units)
***
****
26
24
22
20
18
16
ER-true
TN-like
TNBC

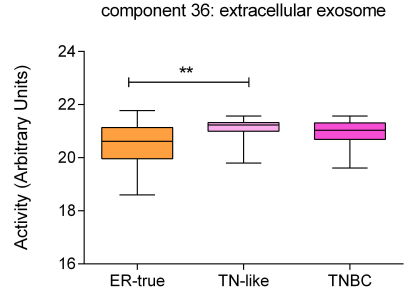

component 9c: glycolysis
Activity (Arbitrary Units)
***
26
24
22
20
18
16
ER-true
TN-like
TNBC

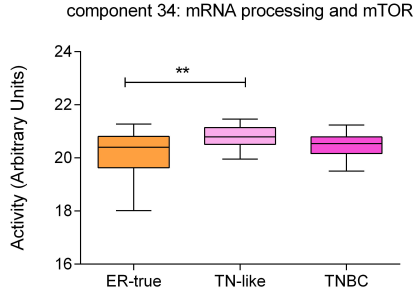

component 10b: metastasis
Activity (Arbitrary Units)
*
24
22
20
18
16
ER-true
TN-like
TNBC

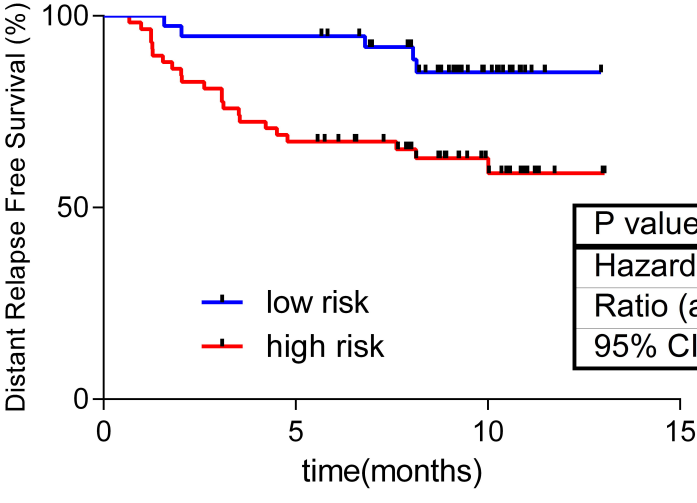

component 10a: proliferation
Activity (Arbitrary Units)
24
22
20
18
16
14
ER-true
TN-like
TNBC


component 13: proteasome
Activity (Arbitrary Units)
*
25
24
23
22
21
20
19
ER-true
TN-like
TNBC

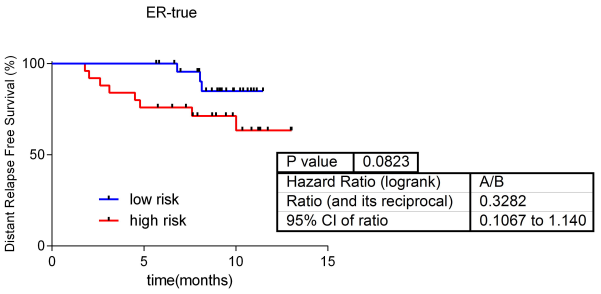

component 9d: transcription
Activity (Arbitrary Units)
***
**
26
25
24
23
22
21
ER-true
TN-like
TNBC

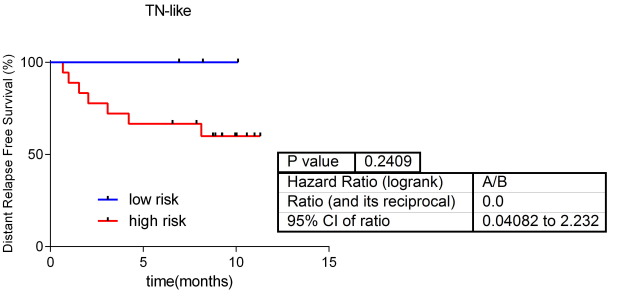

component 36: extracellular exosome
Activity (Arbitrary Units)
**
24
22
20
18
16
ER-true
TN-like
TNBC

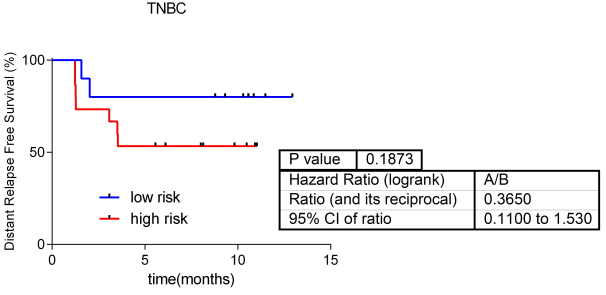

component 34: mRNA processing and mTOR
Activity (Arbitrary Units)
**
24
22
20
18
16
ER-true
TN-like
TNBC

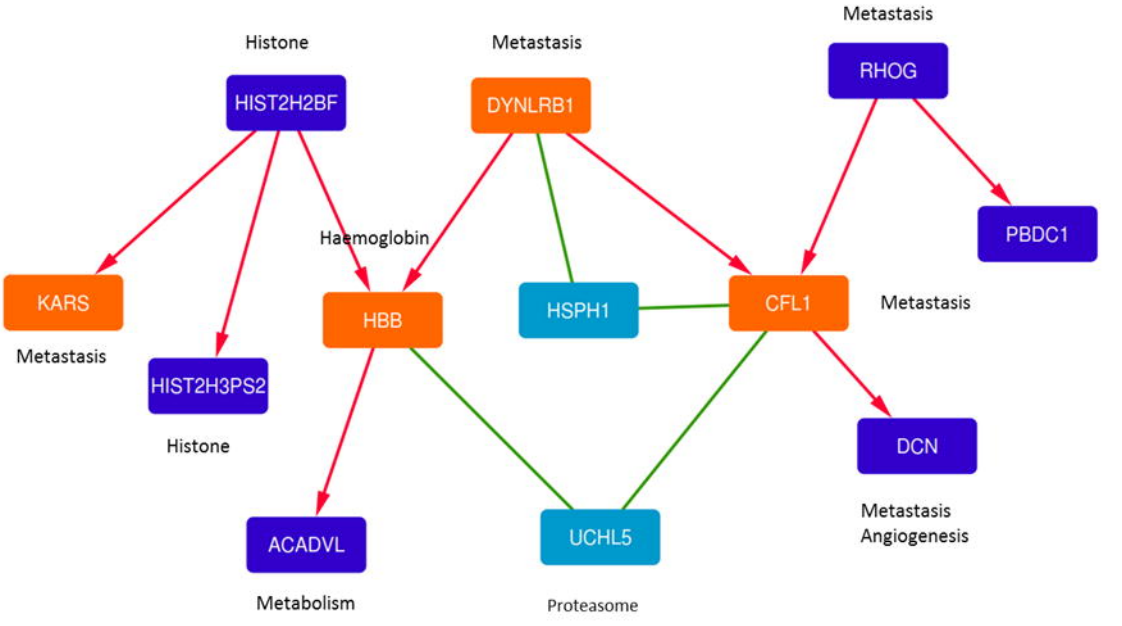


P value	0.0068
Hazard Ratio (logrank)	A/B
Ratio (and its reciprocal)	0.2854
95% CI of ratio	0.1635 to 0.7463


ER-true
Distant Relapse Free Survival (%)
100
50
0
0
5
10
15
time(months)
low risk
high risk
P value 0.0823
Hazard Ratio (logrank) A/B
Ratio (and its reciprocal) 0.3282
95% CI of ratio 0.1067 to 1.140


TN-like
Distant Relapse Free Survival (%)
100
50
0
0
5
10
15
time(months)
low risk
high risk
P value 0.2409
Hazard Ratio (logrank) A/B
Ratio (and its reciprocal) 0.0
95% CI of ratio 0.04082 to 2.232


TNBC
Distant Relapse Free Survival (%)
100
50
0
0
5
10
15
time(months)
low risk
high risk
P value 0.1873
Hazard Ratio (logrank) A/B
Ratio (and its reciprocal) 0.3650
95% CI of ratio 0.1100 to 1.530


Histone
HIST2H2BF
Metastasis
DYNLRB1
Metastasis
RHOG
PBDC1
Haemoglobin
KARS
Metastasis
HBB
HSPH1
CFL1
Metastasis
HIST2H3PS2
Histone
DCN
Metastasis
Angiogenesis
ACADVL
Metabolism
UCHL5
Proteasome