Clinical & Experimental Metastasis (2020) 37:657–674
https://doi.org/10.1007/s10585-020-10060-0

RESEARCH PAPER

Causal Bayesian gene networks associated with bone, brain and lung metastasis of breast cancer

Sung Bae Park[1] · Ki-Tae Hwang[2] · Chun Kee Chung[3,4] · Deodutta Roy[5] · Changwon Yoo[6]

Received: 17 July 2020 / Accepted: 14 October 2020 / Published online: 20 October 2020

Abstract

Using a machine learning method, this study aimed to identify unique causal networks of genes associated with bone, brain, and lung metastasis of breast cancer. Bayesian network analysis identified differentially expressed genes in primary breast cancer tissues, in bone, brain, and lung breast cancer metastatic tissues, and the clinicopathological features of patients obtained from the Gene Expression Omnibus microarray datasets. We evaluated the causal Bayesian networks of breast metastasis to distant sites (bone, brain, or lung) by (i) measuring how well the structures of each specific type of breast cancer metastasis fit the data, (ii) comparing the structures with known experimental evidence, and (iii) reporting predictive capabilities of the structures. We report for the first time that the molecular gene signatures are specific to the different types of breast cancer metastasis. Several genes, including CHPF, ARC, ANGPTL4, NR2E1, SH2D1A, CTSW, POLR2J4, SPTLC1, ILK, ALDH3B1, PDE6A, SCTR, ADM, HEY1, KCNF1, and UVRAG, were found to be predictors of the risk for site-specific metastasis of breast cancer. Expression of POLR2JA, SPTLC1, ILK, ALDH3B1, and the estrogen receptor was significantly associated with breast cancer bone metastasis. Expression of PDE6A and NR2E1 was causally linked to breast cancer brain metastasis. Expression of HEY1, KCNF1, UVRAG, and the estrogen and progesterone receptors was strongly associated with breast cancer lung metastasis. The causal Bayesian network structures of these genes identify potential interactions among the genes in distant metastases of breast cancer, including to the bone, brain, and lung, and may serve as target candidates for treatment of breast cancer metastasis.

Keywords Causal Bayesian networks · Breast neoplasms · Neoplasm metastasis · Bone · Brain · Lung

Introduction

Breast cancer is the most common cancer and the second leading cause of cancer-related death in women worldwide [1, 2]. Significant improvements in the diagnosis, treatment, and prevention of breast cancer have led to decreased mortality and localized breast cancer without distant metastasis is

Deodutta Roy and Changwon Yoo contributed equally to this work.

✉ Deodutta Roy
droy@fiu.edu

✉ Changwon Yoo
cyoo@fiu.edu

[1] Department of Neurosurgery, Seoul National University Boramae Medical Center, Seoul, Korea

[2] Department of Surgery, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea

[3] Department of Neurosurgery, Seoul National University College of Medicine, Seoul, Korea

[4] Department of Neurosurgery, Clinical Research Institute, Seoul National University Hospital, Seoul, Korea

[5] Department of Environmental Health Sciences, Stempel College of Public Health and Social Work, Florida International University, Miami, FL, USA

[6] Department of Biostatistics, Robert Stempel College of Public Health and Social Work, Florida International University, 11200 SW 8th Street AHC5, Miami, FL 33199, USA

now considered to be a manageable disease [2, 3]. Although more than 90% of patients with breast cancer do not have metastasis at the time of diagnosis, about 6% of patients with breast cancer are diagnosed with metastasis [4]. The bone, followed by the lung, brain, and liver, are the most common organs associated with breast cancer metastasis [4, 5]. The mortality and morbidity rates for breast cancer metastasis are 70% to 90% [6, 7]. The median survival of patients with non-metastatic breast cancer is > 85 months (~ 7 years) regardless of the breast cancer subtype [8]. However, the median survival of patients with metastatic breast cancer is 4 to 5 years, and patients with the triple-negative subtype have a median survival of only 10 to 13 months [9, 10].

Because surgery for metastatic lesions is not for oncological control but for palliative control of pain and preservation of function in the organ with metastases, the role of systemic therapy for metastatic breast cancer is important. The expression of HER2, the estrogen receptor (ER), and progesterone receptor (PR) is associated with progression and metastasis of breast cancer [11]. Depending on the breast cancer subtype (i.e., hormone receptor+/HER2−, HER2+, or triple-negative subtype), hormone therapy, targeted therapy, and chemotherapy are used as single agents, and combination therapy is used as the initial or later line of therapy for metastatic breast cancer [4, 12]. Although mutations in BRACA1, BRACA2, ERBB2, and ESR1 can be targeted with clinical efficacy, the clinical application of genomics plays a limited role at present [4, 13, 14]. The next-generation treatment of metastatic cancer requires a comprehensive understanding of both the pathological subtype and genomic and clinical profiles.

Several studies of metastatic breast cancer have reported the related gene signatures in an attempt to predict metastasis and recurrence [1, 15–17]. Previous studies have used applications such as Cytoscape to provide evidence of the presence and strength of the associations between breast cancer metastasis and the related pathways of selected genes [15, 18]. These applications allow for a holistic examination of the interactions of genes, the environment, and clinicopathological factors (e.g., age, sex, phenotype) that are associated with breast cancer metastasis. Understanding the causal relationships between breast cancer metastasis and effectors after metastasis may lead to opportunities to use these associated factors as candidates for therapy and prevention of breast cancer metastasis.

In the clinical and research fields, machine learning methods have been used to study statistical relationships in disease progression through the creation of causal networks from large and complicated health data [19, 20]. Statistical machine learning methods can help to identify the causes as key upstream regulators from a causal network inferred from genomic, clinical, and environmental data related to breast cancer metastasis. Causal Bayesian networks (CBNs) are used to learn the causal networks inferred from genomic data [20, 21]. A CBN is a directed acyclic graph comprising nodes, which represent the random variables being modeled, and intervening arrows, which represent the relationships between the random variables [19]. CBNs have been used to identify the role of osteoblasts in the formation of breast cancer bone metastasis [21].

As mentioned above, the direct causal networks between genes, clinical information, and pathological findings related to breast cancer with metastasis are not well understood. Therefore, we performed a CBN analysis of microarray and clinical and pathological data from the Gene Expression Omnibus (GEO) to obtain a causal network for bone, brain, and lung metastases of breast cancer.

Materials and methods

Data collection from GEO and data mining

Microarray datasets were retrieved from the GEO database of the National Center for Biotechnology Information (NCBI) of the US National Institutes of Health (https://www.ncbi.nlm.nih.gov/geo/; GEO and NCBI websites accessed in August 2018) [22]. The criteria for the enrolled datasets were as follows: (i) datasets with the GEO series (GSE) of human breast cancer alone or with bone, brain, or lung metastases; (ii) studies measuring gene expression in humans (*Homo sapiens*); (iii) studies that collected tissues extracted from breast cancer; and (iv) studies with clinical and pathological information, including age, pathology, expression of the ER, PR, and HER2 receptors, and adjuvant hormone therapy or chemotherapy.

Initially, four studies of the metastases of breast cancer were identified. One study (GSE 5327) did not provide other information about bone and brain metastases, and was excluded, leaving three studies (GSE 2603, 12276, and 2034) identified for the present study. These three studies included 365 samples from patients who were listed by sample accession numbers (GSM) in the GEO database (Tables 1 and 2). The clinical and pathological characteristics are summarized in Table 3. As shown in Table 3, the metastasis of breast cancer was significantly related to the expression of the PR. Expression of the ER and PR, and adjuvant hormone therapy were associated with a higher risk of lung metastasis alone, whereas expression of the ER was associated with brain metastasis of breast cancer alone.

We transformed the normalized gene expression levels into z-scores per gene and discretized the z-values into the categories less than -1 ($z < -1$), between -1 and 1 ($-1 \le z \le 1$), and more than 1 ($z > 1$) to represent low expression, no change in expression, and high expression of a gene, respectively [21]. To learn the CBNs, we used Bayesian

Table 1 Information of enrolled Gene Expression Omnibus series experiments and Gene Expression Omnibus Format in Text format sample files

	References	Study titles	GSE number	Number of subjects having breast cancer without metastasis/with metastasis (bone, brain or lung)	Gene information related with metastases
1	Minn et al. [2]	Genes that mediate breast cancer metastasis to lung	2603	55/27 (14 bone metastasis, 5 brain metastasis, 14 lung metastasis)	Twelve genes are significantly associated with lung-metastasis-free survival, including MMP1, CXCL1 and PTGS2
2	Bos et al. [3]	Genes that mediate breast cancer metastasis to the brain	12,276	19/185 (111 bone metastasis, 16 brain metastasis, 45 lung metastasis)	ST6GALNAC5 specifically mediates brain metastasis EGFR ligands and COX2 were linked to breast cancer infiltration of the lungs, but not the bones or liver
3	Wang et al. [4]	Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer	2034	191/95 (69 bone metastasis, 10 brain metastasis, 25 lung metastasis)	The study revealed a 76-gene signature that accurately predicts distant tumour recurrence
4	Minn et al. [5]	Lung metastasis genes couple breast tumor size and metastatic spread	5327	47/11 (7 lung metastasis)	Lung metastasis gene-expression signature (LMS)

Network Inference with Java Objects (BANJO), which is a computational modeling tool based on a data-driven method that uses Bayesian network frameworks to obtain directed inference networks [21, 23]. Computationally, learning CBN becomes exponentially expensive in search time as the number of parents in the CBN increases. In learning CBN, the usage of memory and required computing time by allowing six or more parents in a CBN structure is too expensive, it does not justify the mere gain of likelihood in the CBN structure with six or more parents. Because of this, to have parsimonious models, all CBNs learned from the data included in this study were obtained by limiting the possible number of parents to five.

Selection of candidate genes

Because the gene expression data in the three studies were collected using different microarray platforms comprising different numbers of gene probes, we first selected common genes that were present in all GSEs, and three GSEs from four studies and 13,229 genes were selected. After collection of all data in the three GSEs, we prepared a dataset (denoted as D°) comprising variables that represent the gene expression levels (low, no change, or high) of 13,229 genes and the following nine clinical variables for 365 patients: age, bone metastasis, brain metastasis, lung metastasis, ER expression, PR expression, HER2 receptor expression, adjuvant chemotherapy, and hormone therapy (Table 2). Using D°, we searched for additional relevant candidate genes among the 13,229 genes and selected the top 10% correlated genes associated with the presence of bone, brain, or lung metastasis (1323 genes) in D°. We also selected 159 signature genes associated with distant metastasis of breast cancer based on two published studies [10, 11]. Adding the 159 signature genes to the 1323 associated genes produced 1467 unique genes (15 genes were common between the signature and correlated genes). Finally, we created a dataset with 1476 variables (denoted as D^{1}) by extracting data for the expression of 1467 genes from D^{0} and adding the nine clinical variables for the 365 patients.

Overall analysis: learning CBN structure

For CBN structure learning, we performed independent runs with four different length of time, i.e., 3 h, 6 h, 12 h, and 24 h. For each of the 3 h, 6 h, 12 h, and 24 h run, we performed three independent CBN structure learning. Therefore, a total of 12 runs of independent CBN structure learning with total of $3\times3\text{ h}+3\times6\text{ h}+3\times12\text{ h}+3\times24\text{ h}=135$ h of runs were performed. Using the dataset D^{1}, we output 12 best log-likelihood CBN structures for each run: three best structures from each 3, 6, 12, and 24 h runs. From the 12 best log-likelihood structures reported for each run, we selected the network with the highest log-likelihood score and denoted this as S^{1} (note that S^{1} includes 1476 variables, each representing a gene expression or clinical information). The first-degree Markov blanket (MB) of variable X in the CBN (denoted as MB [X]) was defined as the set of variables that represents the direct causes (parents) of X, direct effects (children) of X, and direct causes (parents) of the direct effects (children) of X. (X itself was excluded from MB [X]). The second-degree MB of X, third-degree MB

Table 2 Enrolled three GSEs and 365 GSMs

GSE	GSM
GSE2603 (70 GSMs)	GSM50034, GSM50035, GSM50036, GSM50038, GSM50039, GSM50040, GSM50043, GSM50044,GSM50046, GSM50048, GSM50049, GSM50051, GSM50052, GSM50053, GSM50054, GSM50059, GSM50060,GSM50061, GSM50062, GSM50063, GSM50064, GSM50065, GSM50066, GSM50067, GSM50068, GSM50069, GSM50070, GSM50071, GSM50072, GSM50073, GSM50074, GSM50075, GSM50078, GSM50079, GSM50080, GSM50082, GSM50083, GSM50084, GSM50085, GSM50086, GSM50087, GSM50092, GSM50093, GSM50094, GSM50095, GSM50096, GSM50097, GSM50098, GSM50100, GSM50101, GSM50102, GSM50103, GSM50104, GSM50106, GSM50107, GSM50110, GSM50111, GSM50112, GSM50114, GSM50115, GSM50116, GSM50118, GSM50119, GSM50120, GSM50121, GSM50122, GSM50123, GSM50128, GSM50130, GSM50131
GSE12276 (53 GSMs)	GSM308256, GSM308257, GSM308258, GSM308259, GSM308260, GSM308261, GSM308262, GSM308263, GSM308264, GSM308265, GSM308266, GSM308267, GSM308268, GSM308269, GSM308270, GSM308271, GSM308272, GSM308273, GSM308274, GSM308275, GSM308276, GSM308277, GSM308278, GSM308279, GSM308280, GSM308281, GSM308282, GSM308284, GSM308285, GSM308286, GSM308287, GSM308289, GSM308290, GSM308291, GSM308292, GSM308293, GSM308296, GSM308297, GSM308298, GSM308299, GSM308300, GSM308301, GSM308302, GSM308303, GSM308304, GSM308305, GSM308306, GSM308307, GSM308308, GSM308309, GSM308310, GSM308311, GSM308312
GSE2034 (242 GSMs)	GSM36777, GSM36778, GSM36779, GSM36780, GSM36781, GSM36783, GSM36784, GSM36785, GSM36786, GSM36787, GSM36788, GSM36789, GSM36793, GSM36795, GSM36796, GSM36797, GSM36798, GSM36799, GSM36800, GSM36801, GSM36802, GSM36803, GSM36804, GSM36806, GSM36809, GSM36810, GSM36811, GSM36813, GSM36815, GSM36816, GSM36817, GSM36818, GSM36819, GSM36820, GSM36822, GSM36823, GSM36824, GSM36825, GSM36826, GSM36827, GSM36828, GSM36829, GSM36830, GSM36831, GSM36832, GSM36833, GSM36834, GSM36835, GSM36836, GSM36837, GSM36838, GSM36839, GSM36840, GSM36841, GSM36842, GSM36843, GSM36845, GSM36846, GSM36849, GSM36850, GSM36851, GSM36852, GSM36853, GSM36855, GSM36856, GSM36857, GSM36858, GSM36859, GSM36860, GSM36861, GSM36862, GSM36863, GSM36864, GSM36865, GSM36866, GSM36867, GSM36868, GSM36869, GSM36870, GSM36872, GSM36873, GSM36874, GSM36875, GSM36876, GSM36877, GSM36878, GSM36879, GSM36880, GSM36881, GSM36882, GSM36883, GSM36884, GSM36885, GSM36886, GSM36887, GSM36888, GSM36889, GSM36891, GSM36892, GSM36893, GSM36894, GSM36895, GSM36896, GSM36897, GSM36899, GSM36900, GSM36901, GSM36903, GSM36904, GSM36906, GSM36908, GSM36909, GSM36910, GSM36911, GSM36912, GSM36913, GSM36914, GSM36915, GSM36918, GSM36919, GSM36920, GSM36921, GSM36922, GSM36923, GSM36924, GSM36925, GSM36926, GSM36927, GSM36928, GSM36929, GSM36930, GSM36931, GSM36932, GSM36934, GSM36935, GSM36936, GSM36937, GSM36938, GSM36939, GSM36940, GSM36941, GSM36942, GSM36943, GSM36944, GSM36945, GSM36946, GSM36947, GSM36948, GSM36949, GSM36950, GSM36951, GSM36953, GSM36954, GSM36955, GSM36956, GSM36957, GSM36958, GSM36959, GSM36960, GSM36961, GSM36963, GSM36964, GSM36965, GSM36966, GSM36967, GSM36968, GSM36969, GSM36970, GSM36971, GSM36972, GSM36973, GSM36974, GSM36975, GSM36976, GSM36977, GSM36979, GSM36980, GSM36981, GSM36982, GSM36984, GSM36986, GSM36987, GSM36988, GSM36989, GSM36990, GSM36991, GSM36992, GSM36994, GSM36995, GSM36996, GSM36997, GSM36998, GSM36999, GSM37000, GSM37001, GSM37002, GSM37003, GSM37004, GSM37008,GSM37010, GSM37012, GSM37014, GSM37015, GSM37017, GSM37018, GSM37019, GSM37021, GSM37022, GSM37023, GSM37024, GSM37025, GSM37026 GSM37028, GSM37029, GSM37030, GSM37031, GSM37032, GSM37033, GSM37034, GSM37035, GSM37036,GSM37037, GSM37038, GSM37039, GSM37040, GSM37042, GSM37044, GSM37045,GSM37047, GSM37048, GSM37049, GSM37050, GSM37051, GSM37052, GSM37053,GSM37054, GSM37055, GSM37057, GSM37058, GSM37059, GSM37060, GSM37062

of X, etc., were defined as MB (MB [X]), MB (MB [MB (X)]), etc., respectively. Genes were sparsely connected in the best CBN networks that we learned from the datasets. This resulted with too few genes from the first-degree MB of the disease of interest. To include important genes that are closely connected to the disease, in the subsequent learning of the CBN structure, we selected the variables from either the second-degree MB or the third-degree MB of a disease. From S^1, we identified 139 variables comprising 93 signature genes having direct or indirect edges with bone, brain, or lung metastases nodes, 37 genes derived from the third-degree MB of bone, brain, or lung metastases nodes, and nine clinical variables. From dataset D^1, we created a new dataset with these 139 variables of the 365 patients (denoted as D^2). We again learned CBNs using D^2 [21]. This follow-up CBN learning using D^2 was performed the same as for CBN using D^1. From the 12 best log-likelihood structures reported by each independent run, we selected the network with the highest log-likelihood score, which we denoted as S^2 (note that S^2 includes 139 variables, each representing a gene expression and clinical information).

Table 3 Clinopatholgoical information of enrolled 365 patients with breast cancer

Clinopathology		Metastasis		Crude OR [95% confidence interval (CI)]	P	Bone		Crude OR [95% confidence interval (CI)]	P
		Yes	No			Yes	No		
Age									
< 40 years	38	17	21	Reff		11	27	Reff	
40–50 years	206	94	112	1.037 (0.517–2.079)	0.919	60	146	1.009 (0.470–2.163)	0.982
≥ 60 years	121	43	78	0.681 (0.325–1.427)	0.307	26	95	0.672 (0.295–1.532)	0.342
Estrogen receptor									
Positive	240	96	144	0.770 (0.498–1.191)	0.24	69	171	1.398 (0.844–2.316)	0.192
Negative	125	58	67	Reff		28	97	Reff	
Progesterone receptor									
Positive	195	73	122	**0.657 (0.433–0.999)**	**0.049**	49	146	0.853 (0.536–1.358)	0.503
Negative	170	81	89	Reff		48	122	Reff	
HER2 receptor									
Positive	77	34	43	1.107 (0.667–1.838)	0.694	24	53	1.334 (0.769–2.313)	0.304
Negative	288	120	168	Reff		73	215	Reff	
Adjuvant chemotherapy									
Yes	78	37	41	1.311 (0.793–2.168)	0.29	18	60	0.790 (0.439–1.421)	0.43
No	287	117	170	Reff		79	208	Reff	
Adjuvant hormone therapy									
Yes	63	33	30	1.645 (0.954–2.839)	0.072	20	43	1.359 (0.753–2.452)	0.307
No	302	121	181	Reff		77	225	Reff	
Clinopathology		Brain		Crude OR [95% confidence interval (CI)]	P	Lung		Crude OR [95% confidence interval (CI)]	P
		Yes	No			Yes	No		
Age									
< 40 years		2	36	Reff		7	31	Reff	
40–50 years		15	191	1.414 (0.310–6.449)	0.653	24	182	0.584 (0.232–1.471)	0.25
≥ 60 years		3	118	0.458 (0.074–2.846)	0.391	18	103	0.774 (0.296–2.023)	0.6
Estrogen receptor									
Positive		7	233	0**.259 (0.100–0.667)**	**0.003**	20	220	**0.301 (0.162–0.558)**	**0**
Negative		13	112	Reff		29	96	Reff	
Progesterone receptor									
Positive		7	188	0.450 (0.175–1.155)	0.089	13	182	**0.266 (0.136–0.521)**	**0**
Negative		13	157	Reff		36	134	Reff	
HER2 receptor									
Positive		2	75	0.400 (0.091–1.763)	0.211	10	67	0.953 (0.542–2.008)	0.899
Negative		18	270	Reff		39	249	Reff	
Adjuvant chemotherapy									
Yes		6	72	1.625 (0.603–4.378)	0.333	15	63	1.771 (0.909–3.453)	0.09
No		14	273	Reff		34	253	Reff	
Adjuvant hormone therapy									
Yes		6	57	1.414 (0.519–3.847)	0.496	**14**	**49**	**2.180 (1.093–4.348)**	**0.024**
No		14	288	Reff		35	267	Reff	

Bold indicates the significance of P value < 0.05

Reff represents the group with the lowest risk

Subgroup analyses for bone, brain and lung metastasis

We performed subgroup analyses by obtaining the three CBNs learned from datasets with patients who had (i) breast cancer without metastasis or with bone metastasis alone, (ii) breast cancer without metastasis or with brain metastasis alone, and (iii) breast cancer without metastasis and with lung metastasis alone.

Selection of patients and variables for CBN

Among the 365 patients, 228 (62.4%) had nonmetastatic breast cancer, 77 (21.0%) had bone metastasis alone, 8 (2.19%) had brain metastasis alone, and 25 (6.84%) had lung metastasis alone. The total numbers of enrolled patients in the subgroup analyses for bone, brain, and lung were 305, 236, and 253, respectively. Although the number of patients differed according to the site of metastasis, we used the same method to identify the candidate genes as described earlier in the overall analysis. The first three datasets (denoted as $Dbone^1$, $Dbrain^1$, and $Dlung^1$) included the top 10% of genes for bone, brain, and lung metastases (1323 genes), signature genes (144 genes), and clinical information (seven variables including age, ER, PR, HER2 receptor, adjuvant chemotherapy, adjuvant hormone therapy, and bone metastasis, brain metastasis, or lung metastasis) [15, 16].

Learning the CBN structures

After we learned CBNs using the first three datasets ($Dbone^1$, $Dbrain^1$, and $Dlung^1$) the same way we learned CBNs in the overall analysis, we obtained three CBNs with the highest log-likelihood score, which we denoted as $Sbone^1$, $Sbrain^1$ and $Slung^1$. In the second set of datasets for the follow-up CBN learning, which we denoted as $Dbone^2$, $Dbrain^2$ and $Dlung^2$, the datasets included the following variables: (i) genes within the third-degree MB of bone, brain, and lung metastasis variables in $Sbone^1$, $Sbrain^1$ and $Slung^1$; (ii) clinical information; (iii) signature genes having direct or indirect edges with bone, brain, or lung metastases nodes in the first three CBNs ($Sbone^1$, $Sbrain^1$, and $Slung^1$); and (iv) genes with top 10% correlations from the overall analysis using D^0. The follow-up CBN learning was performed following the same process as described in the overall analysis (see "Overall analysis: learning CBN structure" section).

Among the 12 best log-likelihood structures reported by each independent run, we chose the network with the highest log-likelihood score, which we denoted as $Sbone^2$, $Sbrain^2$, and $Slung^2$. We chose the variables within the third-degree MB of bone, brain, or lung metastasis variables in $Sbone^2$, $Sbrain^2$, and $Slung^2$. As we noted earlier, to include the important of genes that are closely connected to a disease, if second-degree MB of the disease included multiple number of genes, we used the variables in the second-degree MB instead of the third-degree MB of the disease. Three new datasets (denoted as $Dbone^3$, $Dbrain^3$, and $Dlung^3$) were created by selecting variables in the second- or third-degree MB of bone, brain, or lung metastasis variables in $Sbone^2$, $Sbrain^2$, and $Slung^2$ for 305, 236, and 253 patients, respectively. We then performed CBN learning using $Dbone^3$, $Dbrain^3$, and $Dlung^3$ following the same processes as used in the previous analyses. Finally, we obtained the structures with the highest log-likelihood score, which we denoted as $Sbone^3$, $Sbrain^3$, and $Slung^3$.

Learning CBN parameters

From the final highest log-likelihood CBNs ($Sbone^3$, $Sbrain^3$, and $Slung^3$), we represented the first-degree MB for the bone, brain, and lung metastasis variables and learned parameters (conditional probabilities) using $Dbone^3$, $Dbrain^3$, and $Dlung^3$, which we denoted as *CBNbone1MB*, *CBNbrain1MB*, and *CBNlung1MB* (note that the three structures include variables that each represent gene expression, pathological information, and metastasis location). In addition, we investigated the relationships between variables and the influence of the status of bone, brain, and lung metastasis of breast cancer on the expression of other genes in the Bayesian structure. GeNIe (BN Graphical Network Interface, version 2.2.1; BayesFusion, LLC, Pittsburgh, PA, USA) was used to learn these parameters.

Enrichment analysis for causal relationships and validation

We also ran an MCMC Order search (denoted as Order algorithm) with settings of the maximum number of parents as five without any prior knowledge given [24]. We report > 99% probable orders and, for each order, we report > 99% probable structures using $Dbone^3$, $Dbrain^3$, and $Dlung^3$ [25]. We then compared the results with the CBNs learned earlier.

We ran three independent runs of the Order algorithm of 4, 24, and 48 h (total of 76 h runs). After obtaining the three orders of groups with the best log-likelihood score (denoted as *PPbone*, *PPbrain*, and *PPlung*), the same the order had a group of structures that included the structure with best log-likelihood score. Finally, we chose the three structures with the best log-likelihood score using the structure code and denoted these as *PPSbone*, *PPSbrain*, and *PPSlung*. After comparing the log-likelihood scores between $Sbone^3$ and *PPSbone*, between $Sbrain^3$ and *PPSbrain*, and between $Slung^3$ and *PPSlung*, we chose the three structures with the higher log-likelihood score, which we denoted as *CBNbone*, *CBNbrain*, and *CBNlung*. The Order algorithm summarizes all plausible CBNs, which provided better information

for the mechanistic understanding underlying the roles of gene–gene and gene–environment interactions in development of cancer.

We further validated *CBNbone1MB*, *CBNbrain1MB*, and *CBNlung1MB* structures by (i) measuring how well the structures fit the data (maximum likelihood and conditional independencies); (ii) comparing the CBN structures with current knowledge in the published literature; (iii) using receiver-operating characteristic curve and relative risk to report the sensitivity and specificity of the CBN; and (iv) evaluating *CBNbone1MB*, *CBNbrain1MB*, and *CBNlung1MB* using leave-one-out cross-validation (LOOCV) and the area under receiver-operating characteristic curve (AUC). We determined the prediction rates of bone, brain, and lung metastasis of breast cancer for *CBNbone1MB*, *CBNbrain1MB*, and *CBNlung1MB* structures using the datasets (denoted as $Dbone^4$, $Dbrain^4$, and $Dlung^4$, respectively) by selecting variables that were parents, children, and coparents of bone, brain, or lung metastasis variables in *CBNbone1MB*, *CBNbrain1MB*, or *CBNlung1MB* of the patients included in $Dbone^3$, $Dbrain^3$, and $Dlung^3$. We calculated to what extent bone, brain, and lung metastases could be expected within the datasets for bone, brain, and lung metastasis based on the information in *Validation Variables* (*V**). *V** comprised (i) the direct cause (parent) and direct effect (children) genes and (ii) the disease node that showed the strongest influence in the conditional independency test [26]. Lastly, we investigated the degree of conditional independencies among variables in the *CBNbone1MB*, *CBNbrain1MB*, and *CBNlung1MB* structures and determined the associations with conditional independency between the variables [27].

Availability of data and material

The datasets generated and/or analyzed during the current study are available in the GEO public repository (https://www.ncbi.nlm.nih.gov/geo/, Gene Expression Omnibus, National Center for Biotechnology Information accessed in November 2017).

Results

We report here the following site-specific CBN structures for breast cancer metastasis: CBNs for bone, brain, and lung metastases of breast cancer (S^2), *CBNbone* for bone metastasis of breast cancer, *CBNbrain* for brain metastasis of breast cancer, and *CBNlung* for lung metastasis of breast cancer (Figs. 1, 2, 3, 4). We also report the results of the validation of the CBN structures.

Overall analysis: CBN for bone, brain, and lung metastases

Among the 12 CBNs, the CBN with the best log-likelihood score had a significantly better data fit than the second-best CBN (i.e., $\frac{P(D^1|S^1)}{P(D^1|S^1)+P(D^1|S')} > 99.999\%$ where S^1 and S' were the best and second-best CBNs, respectively, with 1476 variables, and D^1 was the dataset with the same number of variables for 365 patients. In the follow-up CBN learning comprising the 12 CBNs, one CBN with 139 variables (Fig. 1) was significantly better (> 99.999%) than the second-best CBN. The two genes CHPF and ARC were the direct plausible cause (parent) and plausible effect (child) of bone metastasis, respectively. In brain metastasis, the two genes NR2E1 and ANGPTL4 were plausible direct causes (parents). ADM, the plausible direct cause (parent) of ANGTL4, was also the plausible direct cause (parent) of the ER. PR and CTSW were the plausible direct cause (parent) and plausible effect (child) of lung metastasis, respectively. In addition, ER was the plausible direct cause (parent) of the PR node. Although the results of multivariate regression analysis identified correlations between variables and metastases of breast cancer, the best CBN with 139 variables (Fig. 1) suggested causal relationships between variables and bone, brain, and lung metastases as well as causal relationships between the intervening variables and metastases (Table 3).

Subgroup analyses—bone, brain and lung metastases

Learning three CBN structures using BANJO and Order code

Bone metastasis The CBN that best fit the datasets with 17 variables and 305 patients with no metastasis and bone metastasis is shown in Fig. 2a (*CBNbone*). In *CBNbone*, POLR2J4 (RNA polymerase II subunit JA) was the plausible direct cause of bone metastasis of breast cancer, and SPTLC1, ILK, and ALDH3B1 were the plausible direct effects of bone metastasis of breast cancer. Because the Order algorithm summarized the significant CBN structures identified here, the most likely summarized structure (*PPSbone*, shown in Fig. 2b) suggested that POLR2J4 was the more plausible direct effect of bone metastasis. Using the Order algorithm, we found that the following order was the most probable: bone metastasis, NRFKB, ALDH3B1, POLR2J4, ER, PLXNB1, TRPC1, ILK, CLUD2, PR, CXCL9, LMO4, DYNLL1, CD74, KLF5, SPTLC1, and BOLA2.

Brain metastasis The CBN that best fit the datasets with 21 variables and 236 patients with no metastasis and brain metastasis is shown in Fig. 3a (*CBNbrain*). The two genes PED6A and NR2E1 were plausible direct causes and

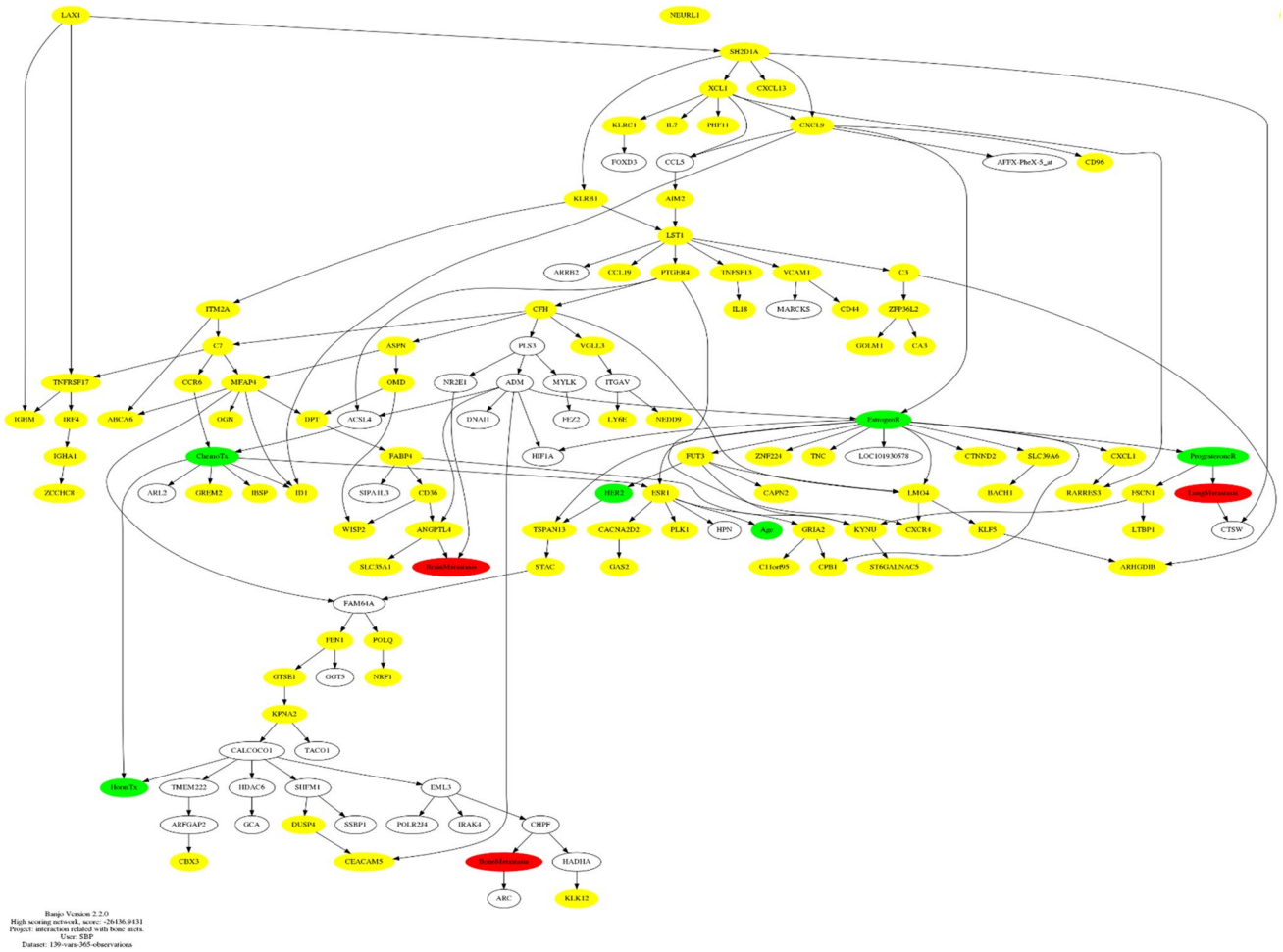

Fig. 1 Causal Bayesian network of genes relevant to breast cancer metastasis. The three nodes in red represent metastasis (bone, brain, or lung metastasis) nodes. The six nodes in green indicate clinical and pathological information connected to the metastasis nodes. The nodes in yellow denote signature genes. (Color figure online)

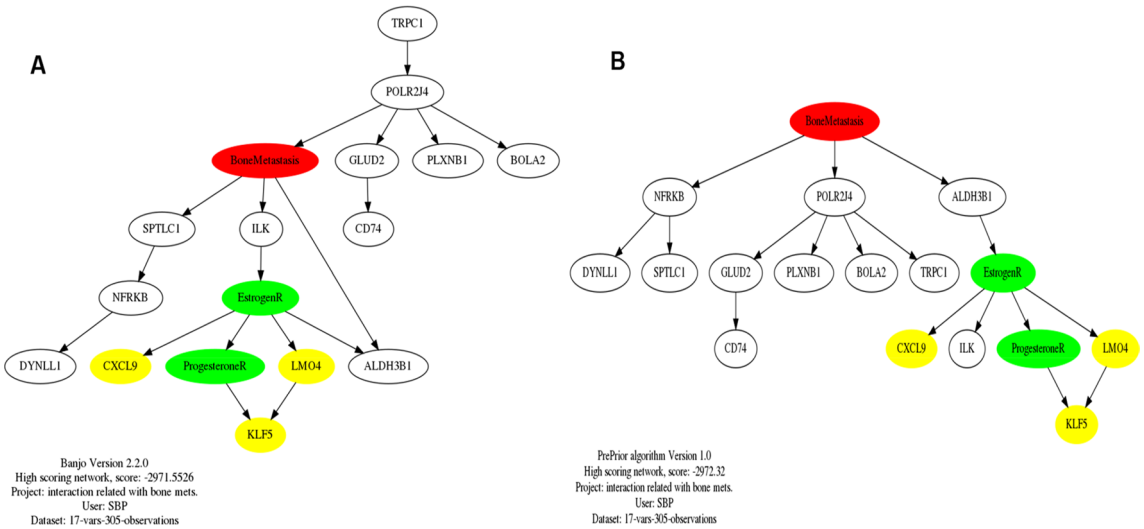

Fig. 2 Causal Bayesian network using BANJO analysis (**a**) and Order algorithm and structural code (**b**) with genes relevant to bone metastasis of breast cancer. The nodes in red represent bone metastasis. The two nodes in green indicate expression of the estrogen and progesterone receptors related to metastasis. The nodes in yellow indicate signature genes. (Color figure online)

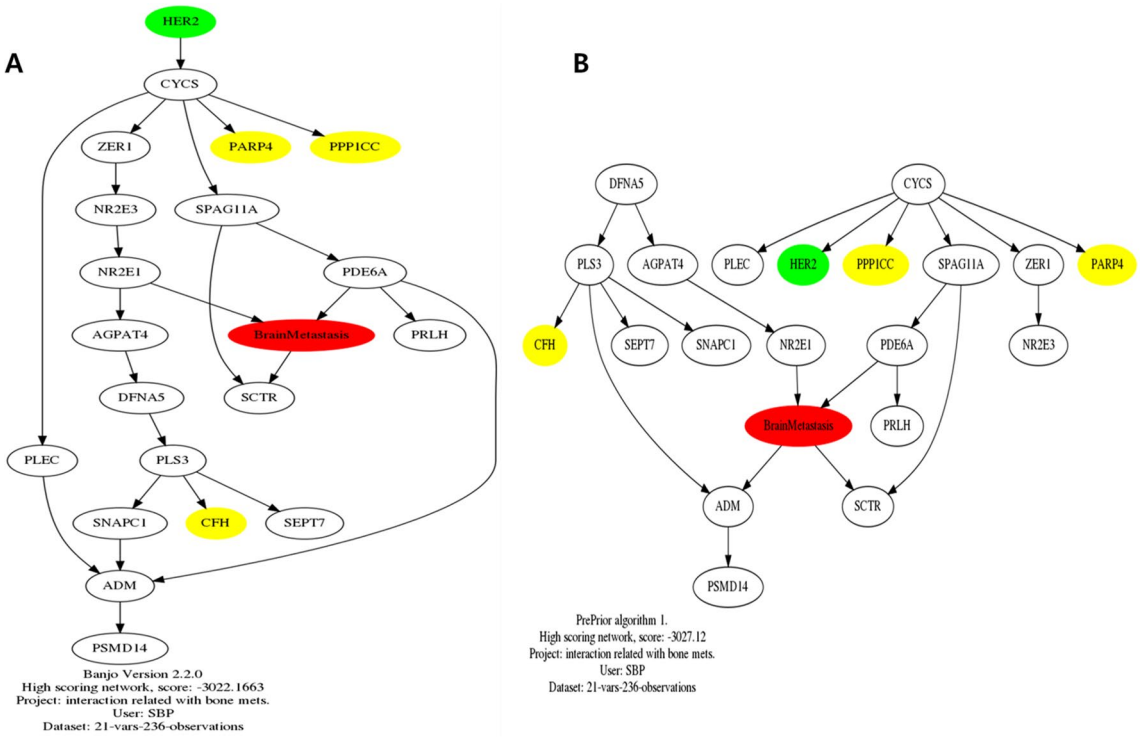

Fig. 3 Causal Bayesian network using BANJO analysis (**a**) and Order algorithm and structural code (**b**) with genes relevant to brain metastasis of breast cancer. The nodes in red represent the brain metastasis node. The nodes in green denote expression of the HER2 receptor related to metastases. The nodes in yellow represent the nodes of signature genes. (Color figure online)

SCTR was the plausible direct effect of brain metastasis of breast cancer in *CBNbrain*. HER2 was one of the ancestor nodes of brain metastasis and had connectivity with brain metastasis. The most likely summarized structure found using the Order algorithm is shown in Fig. 3b (*PPSbrain*). The order of parent and child nodes, and brain metastasis between *CBNbrain* and *PPSbrain* were similar, which provides further support for the plausible cause-and-effect relationships in *CBNbrain*. Using the Order algorithm, we found that the following order was the most probable: CYCS, DFNA5, SPAG11A, PDE6A, ZER1, AGPAT4, R2E3, PLS3, SEPT7, HER2, PPP1CC, PARP4, NR2E1, PRLH, brain metastasis, SNAPC1, SCTR, ADM, PLEC, CFH, and PSMD14.

Lung metastasis The CBN that best fit the datasets with 34 variables and 235 patients with no metastasis and lung metastasis is shown in Fig. 4a (*CBNlung*). The most likely summarized structure found using the Order algorithm is shown in Fig. 4b (*PPSlung*). HEY1 was a plausible direct cause, and KCNF1, SPAG7, ER, PR, and UVRAG were plausible direct effects of lung metastasis of breast cancer in *CBNlung*. Chemotherapy was connected with lung metastasis as a coparent node of the PR. Using the Order algorithm, we found that the following order was the most probable: INSIG2, STAC, chemotherapy, HEY1, ZBTB16, MAPKAP1, BICD1, ER, GALNT3, TFF1, CMC2, LMO4, ARFIP2, ACTR3, ESR1, ARHGEF9, TJP3, PR, FUT3, UVRAG, TFG, FSCN1, LYRM9, lung metastasis, IGHA1, KCNF1, LRIG1, PRRG1, FABP4, PTGER4, SLC35A1, SH3GLB2, ABLIM1, and SPAG7. HEY1 was identified as the plausible direct cause of lung metastasis of breast cancer in *CBNlung* and this effect was equally supported by *PPSlung*. However, the location of the plausible effects of lung metastasis of breast cancer differed between *CBNlung* and *PPSlung*.

Learning CBN parameters

Bone metastasis Parameters (probabilities) of the CBN with six variables that represented the first-degree MB of the group variable (*CBNbone1MB*) of *CBNbone* were learned from a new dataset that contained six variables and 305 patients (denoted as $Dbone^4$) extracted from the dataset $Dbone^3$ that contained 17 variables and 305 patients (Fig. 5a). Using the parameters learned in *CBNbone1MB*, we compared patients with bone metastasis of breast cancer (represented as *Bonemetastasis* with "State1") with those without metastasis including bone metastasis (represented as *Bonemetastasis* with "State0"). A high expres-

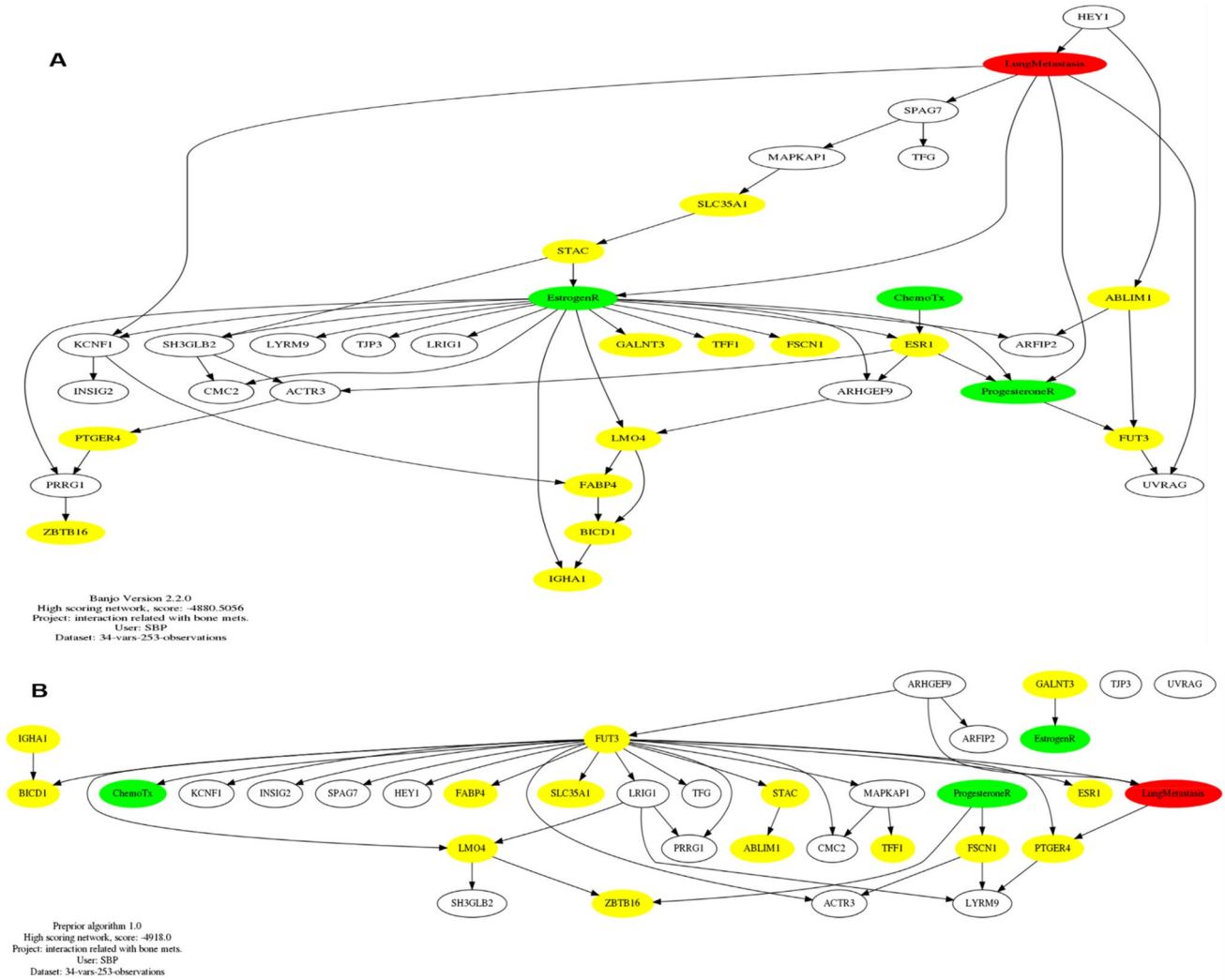

Fig. 4 Causal Bayesian network using BANJO analysis (**a**) and the Order algorithm and structural code (**b**) with genes associated with lung metastasis of breast cancer. The nodes in red represent bone metastasis. The three nodes in green denote the estrogen and progesterone receptors, and chemotherapy for metastasis. The nodes in yellow indicate the nodes of signature genes. (Color figure online)

sion level (denoted as "State2") of SPTLC1 showed that the largest change in probability for the four genes and ER that showed probability changes (probability of SPTLC1 having a high expression level ("State2") increased from 0.12 (Fig. 5b) to 0.32 (Fig. 5c). A high expression level of POLR2J4, a plausible cause (parent) of bone metastasis ("State2"), significantly decreased the risk of having bone metastasis to 0.05 (Fig. 5e) compared with a low expression level ("State0") of, e.g., 0.65 (Fig. 5d). Although the expression state ("State 0 and State 1") of the ER did not alter the change in probability of bone metastasis (21% and 27%, respectively), a high expression level ("State 2") of ILK and ALDH3B1 significantly decreased the risk of having bone metastasis to 0.06 and 0, and changed the probability of expression of the ER to 0.22 and 0.12, respectively.

Brain metastasis Parameters of the CBN with five variables that represented the first-degree MB of group variable (CBNbrain1MB) in CBNbrain were learned from a new dataset that contained five variables and 236 patients (denoted as Dbrain4) extracted from dataset Dbrain3 that contained 21 variables and 236 patients (Fig. 6a). We compared patients with brain metastasis of breast cancer (represented as Brainmetastasis with "State1") with those without bone metastasis (represented as Brainmetastasis with "State0"). We found that a high expression level (denoted as "State2") of NR2E1 produced the largest change in probability of the four genes that showed probability changes (probability of NR2E1 having high expression level ("State2"); the probability decreased from 0.95 (Fig. 6b) to 0.61 (Fig. 6c). When the expression level of NR2E1 was high ("State 2") and that of PDE6A low

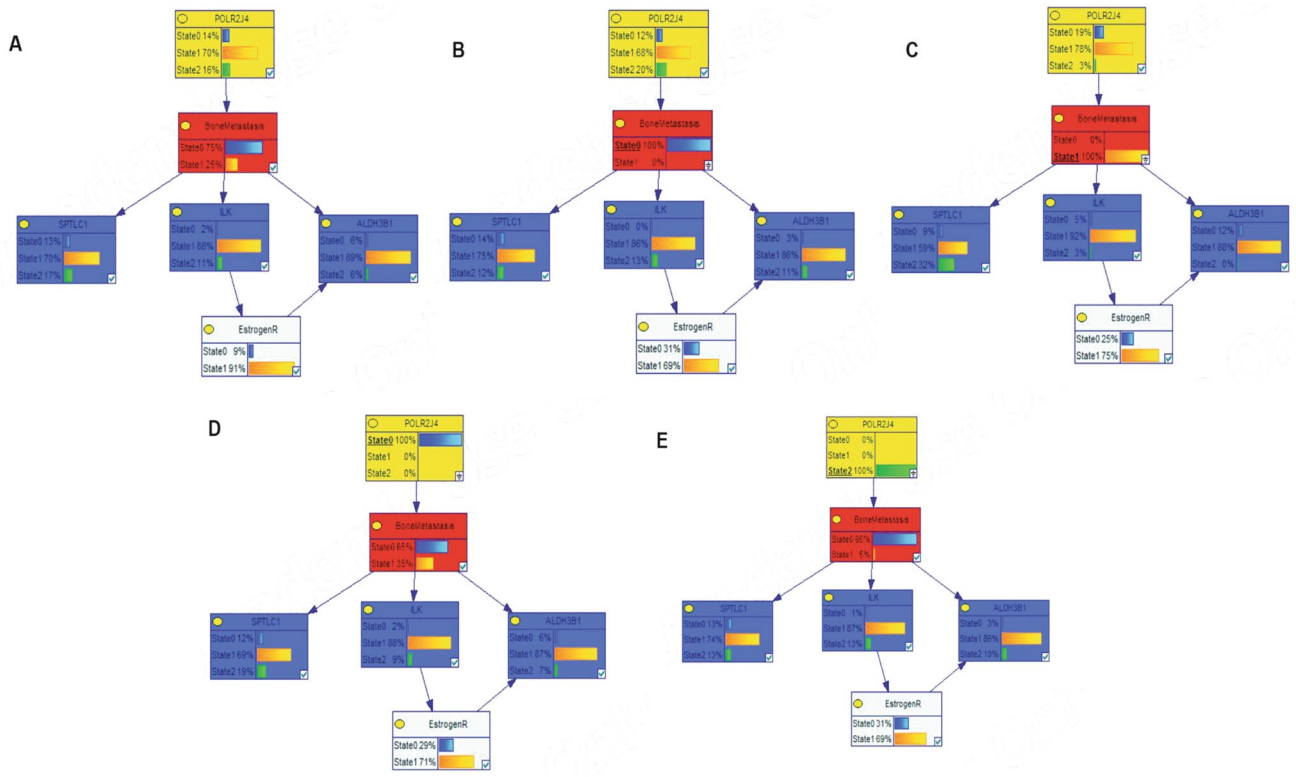

Fig. 5 Causal Bayesian network Structure Using GeNIe. The image shows the causal Bayesian structure learned with the existing data set using GeNIe (A). The parent, group, child, and coparent nodes are shown in yellow, red, light blue, and blue, respectively. States 0 and 1 in a group node represent the probability of bone metastasis and expression of the estrogen receptor, respectively. States 0, 1, and 2 in other nodes represent discretized values 0, 1, and 2, respectively. Images B, C, D, and E show nodal changes after artificial manipulation of the data in the parent and bone metastasis nodes. (Color figure online)

("State0"), the probability of brain metastasis was very small (< 0.01) (Fig. 6d). We also found that when NR2E1 expression was neither low nor high ("State 1") and PDE6A expression was low ("State0"), the probability of brain metastasis was high (> 0.76) (Fig. 6e). Although the results are limited by the small number of patients with brain metastasis (eight of 236 patients), our findings suggest that it may be worthwhile to follow up NR2E1 and PDE6A as candidates genes in clinical studies comparing those with and without brain metastasis of breast cancer.

Lung metastasis Parameters of the CBN with five variables that represented the first-degree MB of group variable (CBNlung1MB) in CBNlung were learned from a new dataset that contained 10 variables and 253 patients with 25 lung metastases (denoted as Dlung4) extracted from dataset Dlung3 that contained 34 variables and 253 patients (Fig. 7a). We compared patients with lung metastasis of breast cancer (represented as Lungmetastasis with "State1") with those without lung metastasis (represented as Lungmetastasis with "State0"). ER and PR status showed the largest change in probabilities among the seven genes and ER and PR expression that showed that the probability of ER expression ("State1") decreased from 0.69 (Fig. 7b) to 0.41 (Fig. 7c), and the probability of PR expression ("State1") decreased from 0.51 (Fig. 7b) to 0.23 (Fig. 7c). For neither high nor low expression of HEY1, a plausible cause of lung metastasis ("State1"), the risk of having bone metastasis increased to 0.43 (Fig. 7e) compared with a high expression level of 0.09 ("State2") (Fig. 7d).

Assessment and validations

Bone metastasis We used LOOCV to evaluate further the predictive performance of the *CBNbone* parameterized by *Dbone*[3]. This analysis produced a value of 75.69% (3925 correct predictions out of 17 × 305 cases). Using only direct causes and effects of bone metastasis of breast cancer in *CBNbone1MB* (six of 17 variables), the LOOCV value was 76.72% (1404 correct predictions out of 6 × 305 cases).The AUC for predicting bone metastasis of breast cancer was 67.68% in *CBNbone1MB*. *CBNbone1MB* parameterized by *Dbone*[4] predicted that two patients would have a very high probability (> 0.91) of having bone metastasis of breast

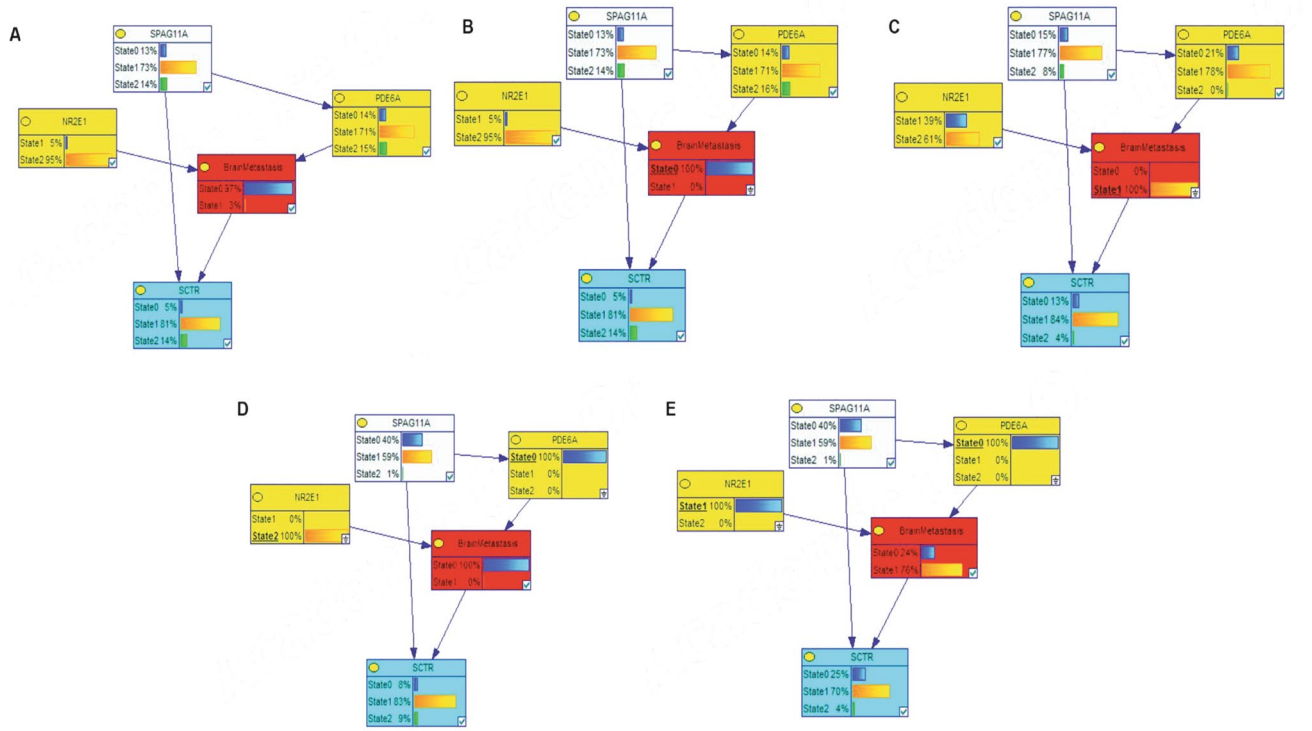

Fig. 6 Causal Bayesian network Structure using GeNIe. The image shows the causal Bayesian structure learned with the existing data set using GeNIe (A). The parent, group, child, and coparent nodes are shown in yellow, red, turquoise, and light blue, respectively. States 0 and 1 in the group node suggest the probability of brain metastasis, respectively. States 0, 1, and 2 in other nodes represent the gene expression levels of low, no change, and high, respectively. Images B–E show nodal changes after artificial manipulation of the data in the parent and brain metastasis nodes. (Color figure online)

cancer, and both of them had the metastasis. In addition, *CBNbone1MB* predicted that five patients would have a very low probability (<0.000372) of having bone metastasis of breast cancer, and none of them had the metastasis (Table 4).

The calculated probabilities of conditional independence between nodes showed that conditional independency relationships between five variables and the bone metastasis variable in *CBNbbone1MB* were consistent with the structure of *CBNbrain1MB* (Fig. 8a). Although the connectivity between all variables cannot be explained using the calculated probabilities of conditional independence, the independencies between the parent (POLR2JA) and children (SPTLC1 and ILK) nodes in the conditioned bone metastasis node were consistent with the *CBNbrain1MB* structure. Additionally, the data supporting the conditional independence of SPTLC1 and ALDH381 expression levels of the POLR2JA expression, and whether a patient had *Bonemetastasis* cancer ("2 4l0, 1" with the highest *p* value of 0.99538 in Fig. 8a) provide a plausible mechanistic understanding of the genes whose relationships are important to the development of bone metastasis of breast cancer.

Brain metastasis The prediction performance of CBNbrain parameterized by Dbrain3 analyzed using LOOCV was 78.97% (3914 correct predictions out of 21×236 cases). When we used only direct causes and effects of brain metastasis of breast cancer in CBNbrain1MB (five of 21 variables), the value was 83.14% (981 correct predictions out of 5×236 cases). The AUC value for predicting bone metastasis of breast cancer was 77.63% in CBNbrain1MB. CBNbrain1MB parameterized by Dbrain4 predicted that five patients would have a very low probability (<0.0012) of having brain metastasis of breast cancer, and none of the five had the metastasis (Table 5). No patients who actually had brain metastasis of breast cancer were predicted to have the metastasis with a significant probability (>0.5). This might have reflected the small number of patients with brain metastasis (eight) compared with the number of patients without the metastasis (228).

The calculated probabilities of conditional independence between nodes showed that the conditional independency relationships between the four variables and brain metastasis variable in CBNbrain1MB were consistent with the structure of CBNbrain1MB (Fig. 8b). There was little disagreement with the structure in the probabilities of conditional

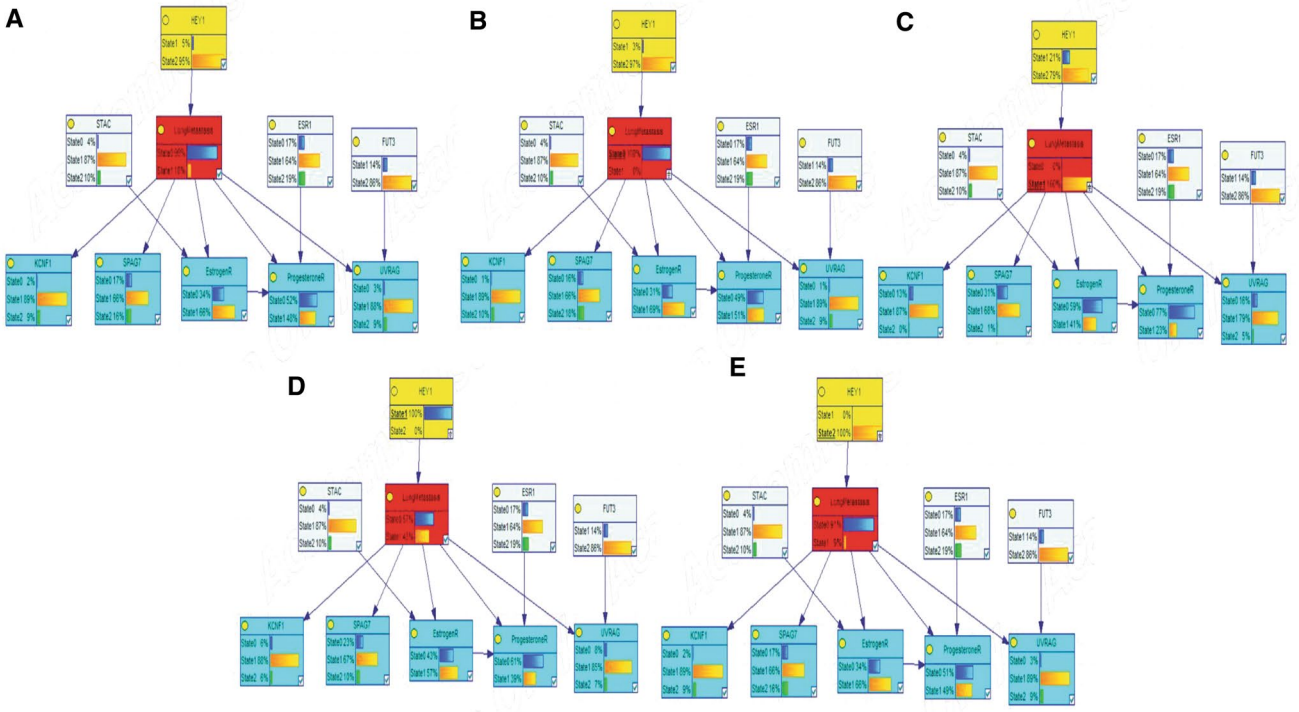

Fig. 7 Causal Bayesian network Structure Using GeNIe. The image shows the causal Bayesian structure learned with the existing data set using GeNIe (A). The parent, group, child, and coparent nodes are shown in yellow, red, turquoise, and light blue, respectively. States 0 and 1 in group node represent the probability of lung metastasis and expression of the estrogen and progesterone receptors, respectively. States 0, 1, and 2 in other nodes represent the gene expression levels of low, no change, and high, respectively. Images B–E show nodal changes after artificial manipulation of the data in the parent and lung metastasis nodes. (Color figure online)

Table 4 Top five and bottom five predicted probabilities of subjects having bone metastasis of breast cancer

Bone metastasis	POLR2J4	SPTLC1	ILK	ALDH3B1	Estrogen receptor	Prediction
Yes	No change	High	Low	No change	Yes	0.916315
Yes	No change	High	Low	No change	Yes	0.916315
Yes	Low	High	No change	Low	Yes	0.863853
Yes	Low	No change	Low	No change	Yes	0.818075
Yes	No change	High	No change	Low	Yes	0.817024
No	No change	No change	Low	No change	Yes	0.000247
No	Low	No change	No change	Low	Yes	0.000317
No	Low	No change	No change	Low	Yes	0.000317
No	No change	High	No change	No change	No	0.000371
No	No change	High	No change	No change	No	0.000371

independence between nodes. The weak stability of the structure associated with brain metastasis may be attributed to the small number of patients with brain metastasis of breast cancer.

Lung metastasis The prediction performance of CBNlung parameterized by Dlung3 analyzed by LOOCV was 78.55% (6757 correct predictions out of 34×206 cases). When we used only direct causes and effects of bone metastasis of breast cancer in CBNlungMB (10 out of 34 variables), the value was 78.69% (1991 correct predictions out of 10×206 cases). The AUC for predicting lung metastasis of breast cancer was 86.94% in CBNlung1MB. CBNlung1MB parameterized by Dlung4 predicted that seven patients would have a very high probability (> 0.99) of having lung metastasis of breast cancer, and all of them had the metastasis. CBNlung1MB also predicted that five patients would have a very low probability (< 0.00001) of

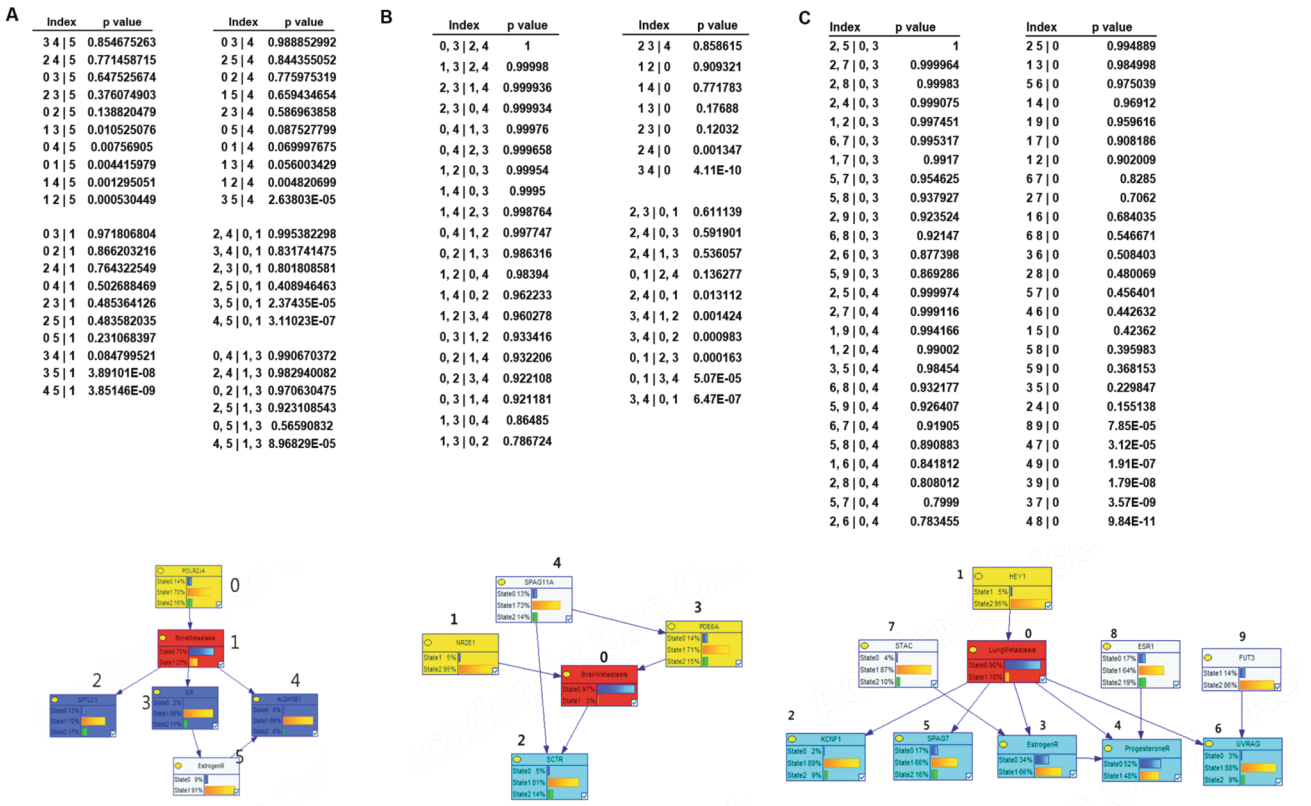

Fig. 8 Probability of conditional independence between nodes. Panels A–C illustrate the probabilities of conditional independence under different conditions and corresponding CBNbone1MB, CBNbrain1MB, and CBNlung1MB structures, respectively. All p values were calculated under the null hypothesis that the given statement is conditionally independent. For example, in Panel A, index 0 through 5 represents variables in the causal Bayesian network (e.g., 0 as POLRJ4 and 5 as EstrogenR), and "0 3|1" indicates that the expression level of POLRJ4 and expression level of ILK are conditionally independent of whether a given patient has BoneMetastasis cancer. A p value of 0.9718 for the statement "0 3|1" means that it is highly likely the statement is true given the data. (Color figure online)

having lung metastasis of breast cancer, and none of them had the metastasis (Table 6). The calculated probabilities of conditional independence between nodes involving nine variables and lung metastasis in CBNlung1MB were consistent with the structure of CBNlung1MB (Fig. 8c). A high degree of agreement with the structure was found in the probabilities of conditional independence between nodes and it was plausible. HEY1 was found to be a direct cause of lung metastasis.

Discussion

We conducted CBN analyses to build four CBNs from the GEO gene expression data obtained for patients with breast cancer with or without bone, brain, and lung metastases. Although we built the statistical models for causal inference, we have also incorporated prior medical and biological information, such as that obtainable in clinical and medical practice, and information involving known signature genes related to metastatic breast cancer. In the statistical analysis of the clinical and immunochemical parameters associated with bone, brain, and lung metastases in the 365 included patients with breast cancer, the general metastasis of breast cancer correlated significantly with the expression of the ER and PR in patients with brain metastasis, and with ER and PR expression, and adjuvant chemotherapy in patients with lung metastasis.

In the overall analysis of the CBN involving concurrent bone, brain, and lung metastases (Fig. 1), ER and PR expression was connected to brain and lung metastases. Although the connection between brain metastasis and ADM could be blocked in the conditioned ANGPTL4 node, brain metastasis (*Brainmetastasis* node) and ER expression had the same plausible causes (parents) as those of ADM. This suggest that ADM and ANGLT4 play roles in the interactions between brain metastasis and ER expression (Table 3).

The angiogenic peptide adrenomedullin is encoded by ADM and is secreted by breast cancer cells, and the peptide accelerates bone metastasis of breast cancer [28]. The ADM expression level is associated with neurodegenerative diseases such as Alzheimer's disease [29].

Table 5 Top five and bottom five predicted probabilities of subjects having brain metastasis of breast cancer

Brain metastasis	NR2E1	SCTR	PDE6A	SPAG11A	Prediction
Yes	No change	Low	No change	Low	0.443773
Yes	No change	No change	No change	No change	0.14702
Yes	No change	No change	No change	No change	0.14702
Yes	No change	No change	No change	No change	0.14702
Yes	High	No change	No change	No change	0.0291389
No	High	High	High	High	0.000278
No	High	No change	High	No change	0.000673
No	No change	High	High	No change	0.001180
No	No change	High	High	No change	0.001180
No	No change	High	High	No change	0.001180

ANGPLT4 encodes angiopoietin-like 4 protein, which is highly expressed in advanced breast cancer [30]. ADM appears to affect *Brainmetastasis* via angiopoietin-like 4 protein in the CBN used in the brain metastasis analysis (Fig. 3). In the brain metastasis CBN (Fig. 3), a direct parent node of *Brainmetastasis* was identified as NR2E1, which is known as nuclear receptor TLX or NR2E1 protein. TLX has been implicated in breast cancer and the initiation and progression of nervous system disorders in humans [31, 32]. These findings suggest that angiogenic peptides, which are encoded by ADM and ANGPLT4, and the regulator of neural stem cells, which is encoded by NR2E1, may be important causal factors underlying brain metastasis of breast cancer.

PR expression was identified as a plausible direct cause of lung metastasis of breast cancer (*Lungmetastasis* node), and ER expression as a plausible direct cause of PR expression in the overall analysis CBN (Fig. 1). These hormone receptors may be effective targets for lung metastasis of breast cancer. *Lungmetastasis* was also a plausible direct cause of CTSW, which encodes cathepsin W. Cathepsins are known contributors to invasive human cancers [33]. A plausible direct cause of CTSW was identified as SH2D1A, which encodes SH2 domain-containing protein 1A. SH2D1A is a prognostic stromal gene signature of breast cancer [34].

In the overall analysis CBN (Fig. 1), CHPF, which encodes chondroitin sulfate synthase 2, was identified as a plausible direct cause bone metastasis of breast cancer

Table 6 Top seven and bottom five predicted probabilities of subjects having lung metastasis of breast cancer

Lung metastasis	HEY1	KCNF1	Estrogen receptor	Progesterone receptor	SPAG7	UVRAG	STAC	ESR1	FUT3	Prediction
Yes	No change	No change	Low	Low	No change	Low	No change	No change	No change	0.999181
Yes	No change	No change	No change	Low	No change	Low	No change	High	No change	0.998622
Yes	No change	Low	Low	Low	Low	Low	No change	No change	High	0.996974
Yes	No change	Low	Low	Low	No change	High	No change	No change	No change	0.994273
Yes	No change	Low	Low	Low	No change	Low	No change	Low	High	0.994197
Yes	No change	No change	No change	No change	No change	No change	High	Low	No change	0.991409
Yes	No change	No change	No change	No change	No change	No change	High	Low	No change	0.991409
No	High	High	No change	No change	No change	No change	No change	No change	No change	3.28E−06
No	No change	High	No change	No change	No change	No change	No change	High	No change	2.25E−05
No	No change	High	No change	No change	No change	No change	No change	High	No change	2.25E−05
No	No change	High	No change	No change	No change	No change	No change	No change	No change	2.67E−05
No	No change	High	No change	No change	No change	No change	No change	No change	No change	2.67E−05

(*Bonemetastasis* node). It has been recently reported that when expressed abnormally in malignant tumors, CHPF promotes lung adenocarcinoma [35]. A plausible direct effect in the *Bonemetastasis* node was identified as ARC, whose expression correlates with liver metastasis of colorectal cancer [36]. Although the association between breast cancer and specific genes has yet to be specified, future studies examining associations between genes and breast cancer, and the mechanisms underlying these associations, will provide valuable information for the prevention and treatment of metastasis of breast cancer.

In the subgroup analysis of bone metastasis, POLR2J4, SPTLC1, ILK, and ALDH3B1 were the plausible direct causes or effects in the *Bonemetastasis* node (Fig. 4). POLR2J4, which is one predictor of recurrence-free survival in hepatocellular carcinoma patients, has been reported to have a direct relationship with bone metastasis [37]. In addition, the low expression of POLR2J4 affects bone metastasis. Although SPTLC1, ILK, and ALDH3B1 were identified as plausible direct effects in the *Bonemetastasis* node, the comparison between patients with *Bonemetastasis* ("State1") and those without *Bonemetastasis* ("State0") showed that SPTLC1 expression was more likely to be higher in those with *Bonemetastasis*. By contrast, expression of ILK and ALDHH3B1 was more likely to be lower in the same patients (Fig. 5b and c). In other words, high expression levels ("State 2") of ILK and ALDH3B1 significantly decreased the risk of having bone metastasis to 0.06 and 0 by changing the probability of ER expression to 0.22 and 0.12, respectively (*Bonemetastasis*, estrogenR = state0|ILK, ALDHH3B1 = state0).

Studies have reported that integrin-linked kinase, which is encoded by ILK, induces accelerated breast tumor development and regulates the migration of breast cancer cells by linkage with the ER [38, 39]. The changes in the gene expression ratios in the CBN structures of bone metastasis provide incomplete information when used with the current data set. However, we assume the biological mechanisms underlying bone metastasis involve the direct causal and effector genes, and hormone receptors in the CBN structure. Therefore, the genes POLR2J4, SPTLC1, ILK, and ALDH3B1 may be therapeutic candidates for targeted therapy of bone metastasis in breast cancer.

In the subgroup analysis of brain metastasis, NR2E1, PDE6A, and SCTR were identified as plausible genes that played important roles in the *Brainmetastasis* node (Fig. 3). SCTR is known to stimulate the proliferation and migration of breast cancer cells and, therefore, its inclusion explains the association between SCTR and brain metastasis of breast cancer [40]. NR2E1 was identified as the plausible direct cause (parent) of *Brainmetastasis* in the overall analysis CBN (Fig. 1). A very low probability (< 1.0E−10) was shown for *Brainmetastasis* ("State1") when NR2E1 expression was high ("State2") (Fig. 6). These findings suggest that the regulation of neural stem cells by NR2E1 may be an important therapeutic target. In the subgroup analysis of lung metastasis, the genes HEY1, KCNF1, and UVRAG, and ER and PR expression were identified as direct causes or effects of lung metastasis in breast cancer (*Lungmetastasis* node). Some studies have reported a relationship between breast cancer and UVRAG and ER and PR expression [4, 41, 42]. According to these studies, expression of HEY1, UVRAG, and the ER and PR appear to trigger lung metastasis in breast cancer. Our results support these earlier findings.

Conclusions

The CBNs of bone, brain, and lung metastases of breast cancer identified here appear to provide networks for reasonable causal inference. Many genes, including CHPF, ARC, ANGPTL4, NR2E1, SH2D1A, CTSW, POLR2J4, SPTLC1, ILK, ALDH3B1, PDE6A, SCTR, ADM, HEY1, KCNF1, and UVRAG, may be useful candidates for the early diagnosis and targeted therapy for metastasis of breast cancer. Although CBNs obtained by statistical inference and machine learning techniques might be limited by the small number of datasets currently available, the results of CBN analysis provide insight into the pathophysiology of metastasis of breast cancer. Future studies should collect more data about gene expression and clinical information, conduct validation studies in wet laboratory and clinical settings, and compare their findings to the current causal inference statistical model that we have developed, and the model should be updated accordingly.

Author contributions SBP has contribution to the design of the work and acquisition, analysis and interpretation of data. And he has drafted the work. KH has drafted the work. CKC has contribution to the design of the work. DR and CY have contribution to the conception of the work, interpretation of data and revised the draft.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Cosphiadi I, Atmakusumah TD, Siregar NC, Muthalib A, Harahap A, Mansyur M (2018) Bone metastasis in advanced breast cancer: analysis of gene expression microarray. Clin Breast Cancer 18:e1117–e1122. https://doi.org/10.1016/j.clbc.2018.03.001

2. Jin X, Mu P (2015) Targeting breast cancer. Metastasis Breast Cancer (Auckl) 9:23–34. https://doi.org/10.4137/BCBCR.S25460
3. Carioli G, Malvezzi M, Rodriguez T, Bertuccio P, Negri E, La Vecchia C (2018) Trends and predictions to 2020 in breast cancer mortality: Americas and Australasia. Breast 37:163–169. https://doi.org/10.1016/j.breast.2017.12.004
4. Waks AG, Winer EP (2019) Breast cancer treatment: a review. JAMA 321:288–300. https://doi.org/10.1001/jama.2018.19323
5. Patanaphan V, Salazar OM, Risco R (1988) Breast cancer: metastatic patterns and their prognosis. South Med J 81:1109–1112
6. Gupta GP, Massague J (2006) Cancer metastasis: building a framework. Cell 127:679–695. https://doi.org/10.1016/j.cell.2006.11.001
7. Randall RL (2014) A promise to our patients with metastatic bone disease. Ann Surg Oncol 21:4049–4050. https://doi.org/10.1245/s10434-014-4010-1
8. Chavez-MacGregor M, Mittendorf EA, Clarke CA, Lichtensztajn DY, Hunt KK, Giordano SH (2017) Incorporating tumor characteristics to the American Joint Committee on Cancer Breast Cancer Staging System. Oncologist 22:1292–1300. https://doi.org/10.1634/theoncologist.2017-0116
9. Bardia A, Mayer IA, Diamond JR, Moroose RL, Isakoff SJ, Starodub AN, Shah NC, O'Shaughnessy J, Kalinsky K, Guarino M, Abramson V, Juric D, Tolaney SM, Berlin J, Messersmith WA, Ocean AJ, Wegener WA, Maliakal P, Sharkey RM, Govindan SV, Goldenberg DM, Vahdat LT (2017) Efficacy and safety of Anti-Trop-2 antibody drug conjugate sacituzumab govitecan (IMMU-132) in heavily pretreated patients with metastatic triple-negative breast cancer. J Clin Oncol 35:2141–2148. https://doi.org/10.1200/JCO.2016.70.8297
10. Swain SM, Baselga J, Kim SB, Ro J, Semiglazov V, Campone M, Ciruelos E, Ferrero JM, Schneeweiss A, Heeson S, Clark E, Ross G, Benyunes MC, Cortes J, CLEOPATRA Study Group (2015) Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. N Engl J Med 372:724–734. https://doi.org/10.1056/NEJMoa1413513
11. Aktas B, Kasimir-Bauer S, Muller V, Janni W, Fehm T, Wallwiener D, Pantel K, Tewes M, CLEOPATRA Study Group (2016) Comparison of the HER2, estrogen and progesterone receptor expression profile of primary tumor, metastases and circulating tumor cells in metastatic breast cancer patients. BMC Cancer 16:522. https://doi.org/10.1186/s12885-016-2587-4
12. Cardoso F, Bedard PL, Winer EP, Pagani O, Senkus-Konefka E, Fallowfield LJ, Kyriakides S, Costa A, Cufer T, Albain KS, Force E-MT (2009) International guidelines for management of metastatic breast cancer: combination vs sequential single-agent chemotherapy. J Natl Cancer Inst 101:1174–1181. https://doi.org/10.1093/jnci/djp235
13. Robson M, Im SA, Senkus E, Xu B, Domchek SM, Masuda N, Delaloge S, Li W, Tung N, Armstrong A, Wu W, Goessl C, Runswick S, Conte P (2017) Olaparib for metastatic breast cancer in patients with a germline BRCA mutation. N Engl J Med 377:523–533. https://doi.org/10.1056/NEJMoa1706450
14. Fribbens C, O'Leary B, Kilburn L, Hrebien S, Garcia-Murillas I, Beaney M, Cristofanilli M, Andre F, Loi S, Loibl S, Jiang J, Bartlett CH, Koehler M, Dowsett M, Bliss JM, Johnston SR, Turner NC (2016) Plasma ESR1 mutations and the treatment of estrogen receptor-positive advanced breast cancer. J Clin Oncol 34:2961–2968. https://doi.org/10.1200/JCO.2016.67.3061
15. Mittempergher L, Saghatchian M, Wolf DM, Michiels S, Canisius S, Dessen P, Delaloge S, Lazar V, Benz SC, Tursz T, Bernards R, van't Veer LJ (2013) A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. Mol Oncol 7:987–999. https://doi.org/10.1016/j.molonc.2013.07.006
16. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365:671–679. https://doi.org/10.1016/S0140-6736(05)17947-1
17. Fazilaty H, Mehdipour P (2014) Genetics of breast cancer bone metastasis: a sequential multistep pattern. Clin Exp Metastasis 31:595–612. https://doi.org/10.1007/s10585-014-9642-9
18. Su G, Morris JH, Demchak B, Bader GD (2014) Biological network exploration with Cytoscape 3. Curr Protoc Bioinformatics 47:8.13.1-8.13.24. https://doi.org/10.1002/0471250953.bi0813s47
19. Deo RC, Nallamothu BK (2016) Learning about machine learning: the promise and pitfalls of big data and the electronic health record. Circ Cardiovasc Qual 9:618–620. https://doi.org/10.1161/Circoutcomes.116.003308
20. Nemzek JA, Hodges AP, He Y (2015) Bayesian network analysis of multi-compartmentalized immune responses in a murine model of sepsis and direct lung injury. BMC Res Notes 8:516. https://doi.org/10.1186/s13104-015-1488-y
21. Park SB, Chung CK, Gonzalez E, Yoo C (2018) Causal inference network of genes related with bone metastasis of breast cancer and osteoblasts using causal Bayesian networks. J Bone Metab 25:251–266. https://doi.org/10.11005/jbm.2018.25.4.251
22. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991-995. https://doi.org/10.1093/nar/gks1193
23. Lipton A, Theriault RL, Hortobagyi GN, Simeone J, Knight RD, Mellars K, Reitsma DJ, Heffernan M, Seaman JJ (2000) Pamidronate prevents skeletal complications and is effective palliative treatment in women with breast carcinoma and osteolytic bone metastases: long term follow-up of two randomized, placebo-controlled trials. Cancer 88:1082–1090
24. Friedman N, Koller D (2003) Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Mach Learn 50:95–125. https://doi.org/10.1023/A:1020249912095
25. Agostinho NB, Machado KS, Werhli AV (2015) Inference of regulatory networks with a convergence improved MCMC sampler. BMC Bioinformatics 16:306. https://doi.org/10.1186/s12859-015-0734-6
26. Scutari MDJ (2014) Bayesian networks with examples in R. Chapman and Hall, Boca Raton, p 16
27. Charniak E (1991) Bayesian networks without tears. Ai Mag 12:50–63
28. Siclari VA, Mohammad KS, Tompkins DR, Davis H, McKenna CR, Peng X, Wessner LL, Niewolna M, Guise TA, Suvannasankha A, Chirgwin JM (2014) Tumor-expressed adrenomedullin accelerates breast cancer bone metastasis. Breast Cancer Res 16:458. https://doi.org/10.1186/s13058-014-0458-y
29. Ferrero H, Larrayoz IM, Gil-Bea FJ, Martinez A, Ramirez MJ (2018) Adrenomedullin, a novel target for neurodegenerative diseases. Mol Neurobiol 55:8799–8814. https://doi.org/10.1007/s12035-018-1031-y
30. Shafik NM, Mohamed DA, Bedder AE, El-Gendy AM (2015) Significance of tissue expression and serum levels of angiopoietin-like protein 4 in breast cancer progression: link to NF-kappaB /P65 activity and pro-inflammatory cytokines. Asian Pac J Cancer Prev 16:8579–8587
31. Sobhan PK, Funa K (2017) TLX—its emerging role for neurogenesis in health and disease. Mol Neurobiol 54:272–280. https://doi.org/10.1007/s12035-015-9608-1

32. Lin ML, Patel H, Remenyi J, Banerji CR, Lai CF, Periyasamy M, Lombardo Y, Busonero C, Ottaviani S, Passey A, Quinlan PR, Purdie CA, Jordan LB, Thompson AM, Finn RS, Rueda OM, Caldas C, Gil J, Coombes RC, Fuller-Pace FV, Teschendorff AE, Buluwela L, Ali S (2015) Expression profiling of nuclear receptors in breast cancer identifies TLX as a mediator of growth and invasion in triple-negative breast cancer. Oncotarget 6:21685–21703. https://doi.org/10.18632/oncotarget.3942
33. Tan GJ, Peng ZK, Lu JP, Tang FQ (2013) Cathepsins mediate tumor metastasis. World J Biol Chem 4:91–101. https://doi.org/10.4331/wjbc.v4.i4.91
34. Winslow S, Leandersson K, Edsjo A, Larsson C (2015) Prognostic stromal gene signatures in breast cancer. Breast Cancer Res 17:23. https://doi.org/10.1186/s13058-015-0530-2
35. Hou XM, Zhang T, Da Z, Wu XA (2019) CHPF promotes lung adenocarcinoma proliferation and anti-apoptosis via the MAPK pathway. Pathol Res Pract. https://doi.org/10.1016/j.prp.2019.02.005
36. Toth C, Meinrath J, Herpel E, Derix J, Fries J, Buettner R, Schirmacher P, Heikaus S (2016) Expression of the apoptosis repressor with caspase recruitment domain (ARC) in liver metastasis of colorectal cancer and its correlation with DNA mismatch repair proteins and p53. J Cancer Res Clin Oncol 142:927–935. https://doi.org/10.1007/s00432-015-2102-3
37. Gu JX, Zhang X, Miao RC, Xiang XH, Fu YN, Zhang JY, Liu C, Qu K (2019) Six-long non-coding RNA signature predicts recurrence-free survival in hepatocellular carcinoma. World J Gastroenterol 25:220–232. https://doi.org/10.3748/wjg.v25.i2.220
38. Oloumi A, Maidan M, Lock FE, Tearle H, McKinney S, Muller WJ, Aparicio SA, Dedhar S (2010) Cooperative signaling between Wnt1 and integrin-linked kinase induces accelerated breast tumor development. Breast Cancer Res 12:R38. https://doi.org/10.1186/bcr2592
39. Acconcia F, Manavathi B, Mascarenhas J, Talukder AH, Mills G, Kumar R (2006) An inherent role of integrin-linked kinase-estrogen receptor alpha interaction in cell migration. Cancer Res 66:11030–11038. https://doi.org/10.1158/0008-5472.CAN-06-2676
40. Kang S, Kim B, Kang HS, Jeong G, Bae H, Lee H, Lee S, Kim SJ (2015) SCTR regulates cell cycle-related genes toward anti-proliferation in normal breast cells while having pro-proliferation activity in breast cancer cells. Int J Oncol 47:1923–1931. https://doi.org/10.3892/ijo.2015.3164
41. Wu T, Li Y, Gong L, Lu JG, Du XL, Zhang WD, He XL, Wang JQ (2012) Multi-step process of human breast carcinogenesis: a role for BRCA1, BECN1, CCND1, PTEN and UVRAG. Mol Med Rep 5:305–312. https://doi.org/10.3892/mmr.2011.634
42. McFall T, McKnight B, Rosati R, Kim S, Huang Y, Viola-Villegas N, Ratnam M (2018) Progesterone receptor A promotes invasiveness and metastasis of luminal breast cancer by suppressing regulation of critical microRNAs by estrogen. J Biol Chem 293:1163–1177. https://doi.org/10.1074/jbc.M117.812438