Systematic analysis of breast atypical hyperplasia-associated hub genes and pathways based on text mining

Wei Ma[a], Bei Shi[c], Fangkun Zhao[b], Yunfei Wu[a] and Feng Jin[a]

The purpose of this study was to describe breast atypical hyperplasia (BAH)-related gene expression and to systematically analyze the functions, pathways, and networks of BAH-related hub genes. On the basis of natural language processing, gene data for BAH were extracted from the PubMed database using text mining. The enriched Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways were obtained using DAVID (*http://david.abcc.ncifcrf.gov/*). A protein–protein interaction network was constructed using the STRING database. Hub genes were identified as genes that interact with at least 10 other genes within the BAH-related gene network. In total, 138 BAH-associated genes were identified as significant ($P<0.05$), and 133 pathways were identified as significant ($P<0.05$, false discovery rate<0.05). A BAH-related protein network that included 81 interactions was constructed. Twenty genes were determined to interact with at least 10 others ($P<0.05$, false discovery rate<0.05) and were identified as the BAH-related hub genes of this protein–protein interaction network. These 20 genes are *TP53, PIK3CA, JUN, MYC, EGFR, CCND1, AKT1, ERBB2, CTNN1B, ESR1, IGF-1, VEGFA, HRAS, CDKN1B, CDKN1A, PCNA, HGF, HIF1A, RB1,* and *STAT5A*. This study may help to disclose the molecular mechanisms of BAH development and provide implications for BAH-targeted therapy or even breast cancer prevention. Nevertheless, connections between certain genes and BAH require further exploration. *European Journal of Cancer Prevention* 28: 507–514 Copyright © 2018 The Author(s). Published by Wolters Kluwer Health, Inc.

European Journal of Cancer Prevention 2019, **28:**507–514

Keywords: breast atypical hyperplasia, hub genes, pathways, text mining

Department of [a]Breast Surgery, the First Affiliated Hospital of China Medical University, [b]Ophthalmology, the Fourth Affiliated Hospital of China Medical University and [c]Physiology, China Medical University, Shenyang, China

Correspondence to Feng Jin, PhD, Department of Breast Surgery, the First Affiliated Hospital of China Medical University, No. 155 Nanjingbei Street, Heping District, 110001 Shenyang, China
Tel: +86 24 8328 2618; e-mail: jinfeng@cmu.edu.cn

Received 4 September 2018 **Accepted** 10 October 2018

Introduction

Breast cancer has become one of the most common malignant tumors that threaten women's health and lives. However, the etiology of breast cancer is still unclear. A well-known hypothesis is the 'multistage development model theory' (Lakhani, 1999), in which breast cancer develops from normal tissue to general hyperplasia, atypical hyperplasia, carcinoma in situ, and then invasive carcinoma. Previous studies have shown that the cumulative risk of breast cancer among women with atypical hyperplasia approaches 30% at 25 years of follow-up (Hartmann et al., 2014, 2015). Therefore, the process may be driven by quantitative changes and qualitative transformation of some factors over an extended time period.

Atypical hyperplasia, as a premalignant disease, holds a transitional region between benign and malignant disease because it possesses some of the requisite features of a malignant tumor and may share a common ancestor with carcinoma on the basis of somatic mutations (Allred *et al.*, 2001; Santen and Mansel, 2005; Bombonati *et al.*, 2011; Newburger *et al.*, 2013; Degnim, 2015). In atypical hyperplasia, there is a proliferation of dysplastic, monotonous epithelial cell populations that include clonal subpopulations (Ellis, 2010). According to the microscopic appearance, two types of breast atypical hyperplasia (BAH) are found: atypical ductal hyperplasia and atypical lobular hyperplasia. These two types of BAH occur with similar frequency and confer equal risks of future breast cancer (Dupont and Page, 1985; Hartmann *et al.*, 2005; Degnim *et al.*, 2007; Page *et al.*, 2015). Although the risk of atypical hyperplasia becoming malignant is increasing, BAH will retrogress under certain conditions (Visscher *et al.*, 2017). Because of the high-risk features and high incidence of BAH, studies on the knowledge of atypical hyperplasia structure and BAH-related gene function may be valuable for diagnosing and determining targeted breast cancer prevention therapies.

Currently, there is a large body of biomedical literature in databases, and rapid growth of the research makes it impossible for researchers to address all of the information manually. Text mining tools are widely used in biomedical research to extract information about disease-related genes, proteins, molecular interactions, and pathways, and these tools allow for the generation of an enormous amount of information and the identification of relationships and

0959-8278 Copyright © 2019 The Author(s). Published by Wolters Kluwer Health, Inc..

DOI: 10.1097/CEJ.0000000000000494

structures that would otherwise not be possible. Previous studies have documented the use of these tools in the study of regulation mechanisms for different types of cancers, including breast cancer (Krallinger *et al.*, 2010). In the present study, we obtained BAH-related texts from PubMed by searching for ‘breast atypical hyperplasia’ or ‘atypical hyperplasia of mammary gland’ and retrieved 1777 publications. We identified sets of genes that were intensively investigated in relation to BAH; furthermore, we established a protein–protein interaction (PPI) network. We also identified enriched pathways and hub genes. These data may help to promote the understanding of BAH and substantially affect the treatment of this disease and may even have an effect on breast cancer prevention.

Materials and methods

The genes and proteins were automatically extracted from abstracts by natural language processing. We used ‘breast atypical hyperplasia’ or ‘atypical hyperplasia of mammary gland’, etc. as search terms (Fig. 1) and extracted literature published before July 2017 from the PubMed database. The genes and proteins mentioned in the abstracts were recognized and tagged by A Biomedical Named Entity Recognizer, which is used to tag genes, proteins, and biological entities (Settles, 2005). The Entrez Global Query Cross-Database Search System is a federated search engine that allows users to search health science databases on the NCBI website. This system was used to obtain genes and proteins with unified results to form a database (Maglott *et al.*, 2006). The number of hits for the search term in the database was counted. Hypergeometric distribution was used to calculate the co-occurrence probability of each gene name and BAH. If the co-occurrence probability of a gene exceeded the theoretical expectation ($P < 0.05$), this gene was considered relevant to BAH.

DAVID (*http://david.abcc.ncifcrf.gov/*) is a free, online bioinformatics resource that provides functional interpretation of large lists of genes derived from genomic studies. Gene Ontology enrichment analysis was performed using DAVID. Selected BAH-related genes from the aforementioned screening process were annotated and classified by biological processes, molecular functions, and cellular components.

Kyoto Encyclopedia of Genes and Genomes Orthology-Based Annotation System is an annotation system based on Kyoto Encyclopedia of Genes and Genomes that was applied for BAH-related signaling pathway enrichment annotation analysis.

The STRING database (*http://www.string-db.org/*) was used to construct the PPI network of BAH-related genes and select the hub genes. We selected the interactions with integrated scores of 0.9 to construct the PPI network. To select the hub genes from the PPI network, we calculated the number of genes directly interacting with each gene. We defined hub genes in the network as those genes with a degree of at least 10. A threshold of 0.05 was established for *P* values and the false discovery rate (FDR).

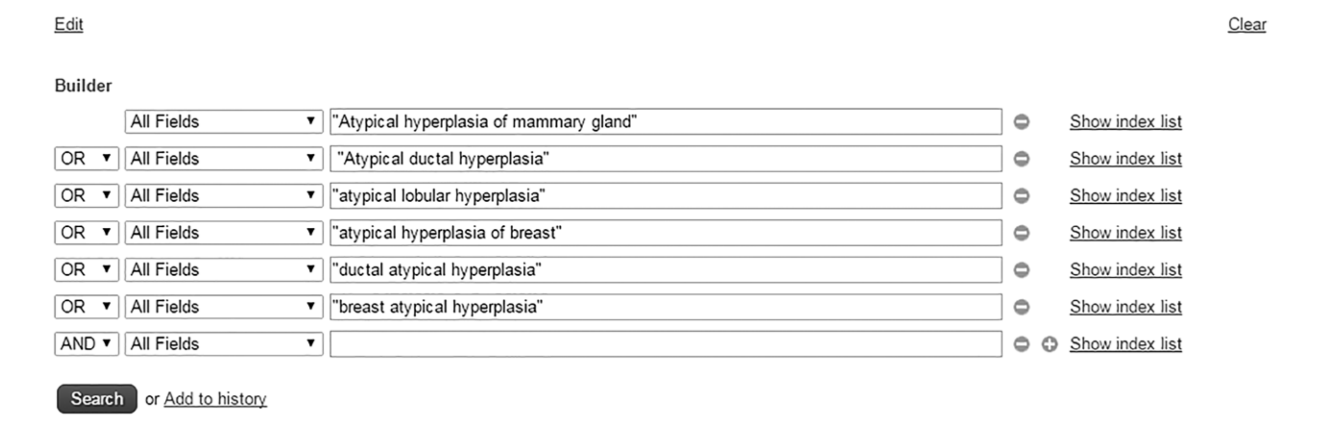
Results

Breast atypical hyperplasia-associated genes and Gene Ontology analysis

We examined 1777 abstracts and obtained 325 genes after the retrieval of contents from PubMed. Through hypergeometric distribution, a total of 138 genes were identified as BAH-related genes ($P < 0.05$). Among these BAH-related genes, the top 20 most frequently investigated genes are listed in Table 1.

ESR1 (ER-α), *TP53*, *ERBB2*, *CCND1*, and *TP63* were the most frequently mentioned genes (Table 1). The Gene Ontology analysis results of classification not only by biological processes and cellular components but also by molecular functions are presented in Table 2. Regulation of cell proliferation, apoptosis, programmed cell death, and cell death were the main biological processes

Fig. 1



	All Fields	"Atypical hyperplasia of mammary gland"
OR	All Fields	"Atypical ductal hyperplasia"
OR	All Fields	"atypical lobular hyperplasia"
OR	All Fields	"atypical hyperplasia of breast"
OR	All Fields	"ductal atypical hyperplasia"
OR	All Fields	"breast atypical hyperplasia"
AND	All Fields	

Search strategy.

Table 1 Top 20 most significant atypical hyperplasia-related genes based on text mining

Genes	Description	Count	*P* value
ESR1	Estrogen receptor 1	174	3.95E−171
TP53	Tumor protein p53	160	1.80E−221
ERBB2	Erb-b2 receptor tyrosine kinase 2 (HER2)	149	5.53E−43
CCND1	Cyclin D1	134	1.98E−42
TP63	Tumor protein p63	84	8.77E−79
BRCA1	BRCA1, DNA repair associated	81	8.36E−132
CDH1	Cadherin 1	75	3.17E−124
KRT5	Keratin 5	70	8.09E−60
MKI67	Marker of proliferation Ki-67	56	3.79E−169
BRCA2	BRCA2, DNA repair associated	56	9.28E−112
FEA	F9 embryonic antigen	49	6.13E−78
CDH13	Cadherin 13	49	0.000198302
PGR	Progesterone receptor	47	0
MUC1	Mucin 1, cell surface associated	47	2.48E−29
CDKN2A	Cyclin-dependent kinase inhibitor 2A	44	2.95E−08
KRT18	Keratin 18	40	0.000460611
VEGFA	Vascular endothelial growth factor A	39	1.76E−08
CCNB1	Cyclin B1	38	0.000410344
RB1	RB transcriptional corepressor 1	30	0.023265581
STAT5A	Signal transducer and activator of transcription 5A	29	0.004616211

associated with BAH-related genes. With respect to molecular function, the major activities of these genes included enzyme binding, structure-specific DNA binding, double-stranded DNA binding, and transmembrane receptor protein tyrosine kinase activity. These genes were related to various cellular components, including the plasma membrane, organelle lumens, and membrane rafts.

Pathway and protein–protein interaction analyses

Following the pathway analysis, 133 pathways were identified as significant ($P<0.05$, FDR <0.05). Among these pathways, pathways related to cancer, proteoglycans in cancer, and microRNAs in cancer involved the largest number of genes. The 20 most significant BAH-related pathways are presented in Table 3.

Meanwhile, we constructed a BAH-related PPI network (Fig. 2). The 20 genes that interact with at least 10 other genes ($P<0.05$, FDR <0.05) were identified as the hub genes of the BAH-related PPI network. These genes are *TP53*, *PIK3CA*, *JUN*, *MYC*, *EGFR*, *CCND1*, *AKT1*, *ERBB2*, *CTNN1B*, *ESR1*, *IGF-1*, *VEGFA*, *HRAS*, *CDKN1B*, *CDKN1A*, *PCNA*, *HGF*, *HIF1A*, *RB1*, and *STAT5A*. *TP53*, which interacts with 28 other genes, exhibited the greatest number of interactions (Fig. 3). The similarities and differences between BAH-related hub genes and the top 20 highest frequency genes were classified using a Venn diagram (Fig. 4).

Discussion

The remarkable increase in the morbidity and mortality of breast cancer is a major concern worldwide. BAH, as a precancerous disease, has attracted increasing attention. However, its biology is poorly understood. The multistage development model theory does not account for all breast cancer subtypes that stem from BAH on the basis of both genomic and histological observations (Gao *et al.*, 2009). Thus, a better understanding of BAH will advance not only our understanding of breast carcinogenesis but also our clinical management of these high-risk patients. Taking effective measures for treatment and intervention to reduce the incidence of breast cancer can greatly improve women's physical and mental health.

Text mining can help us derive implicit knowledge that may be hidden in unstructured literature and present the data in an organized form. Our knowledge of the pathophysiology of BAH allows us to propose possible candidate genes that could play a role in the development and progression of breast cancer. We generated an integrated approach to enrich the molecular context of BAH by applying text mining of events involving genes (presented as nodes) and pathways (presented as edges that correspond to interactions between nodes). By extracting information from PubMed, we present a comprehensive molecular interaction network for BAH (85 nodes and 291 edges) and discuss its properties using standard network metrics. All of the aforementioned 20 hub genes are known to be closely related to the typical pathological progression of BAH.

Atypical hyperplasia is a noncancerous cellular hyperplasia in which cells show some atypia. Therefore, some genes that affect cell proliferation, apoptosis, and signal transduction, such as *RB1*, *VEGF*, *STAT5A*, *CCND1*, *TP53*, *ESR1 (ER-α)*, and *ERBB*, could play an important role in the relationships between BAH and certain hub genes. These genes have been extensively studied, and all of the aforementioned genes are known to be closely related to the occurrence and development of BAH.

However, relative to these genes, *PCNA*, *CDKN1B*, *CTNNB1*, *EGFR*, *AKT1*, *MYC*, *JUN*, *CDKN1A*, *IGF-1*, *HIF1A*, *PIK3CA*, *HRAS*, and *HGF* have been reported less frequently in the context of BAH, which requires further research.

PIK3CA and *AKT1*

PIK3CA, *PIK3CB*, and *PIK3CD* encode a catalytic subunit (p110) of PI3K (Vogt *et al.*, 2010; Georgescu, 2011; Ersahin *et al.*, 2015). Activated PI3K can catalyze the formation of the second messenger phosphatidylinositol triphosphate, and then, phosphatidylinositol triphosphate plays a key role by recruiting Pleckstrin homology domain-containing proteins to the membrane, including AKT1 and PDPK1, and activating signaling cascades involved in cell growth, survival, proliferation, motility, and morphology (Karakas *et al.*, 2006; Engelman, 2009; Castaneda *et al.*, 2010). *PIK3CA* hotspot point mutations were identified in associated hyperplasia, even in usual ductal hyperplasia and columnar cell change, suggesting that *PIK3CA* mutations may play a role in breast epithelial proliferation and atypical changes (Kehr *et al.*,

Table 2 Classification results for biological process, cellular components, and molecular functions by Gene Ontology analysis

Terms	Count	*P* value
Biological process		
GO:0042127 – regulation of cell proliferation	48	3.61E−28
GO:0042981 – regulation of apoptosis	45	1.17E−24
GO:0043067 – regulation of programmed cell death	45	1.74E−24
GO:0010941 – regulation of cell death	45	2.03E−24
GO:0007242 – intracellular signaling cascade	40	3.07E−13
GO:0010033 – response to organic substance	36	1.00E−17
GO:0051252 – regulation of RNA metabolic process	36	1.87E−06
GO:0006355 – regulation of transcription, DNA-dependent	35	3.24E−06
GO:0010604 – positive regulation of macromolecule metabolic process	32	3.47E−12
GO:0043066 – negative regulation of apoptosis	29	1.19E−19
GO:0043069 – negative regulation of programmed cell death	29	1.73E−19
GO:0060548 – negative regulation of cell death	29	1.86E−19
GO:0008284 – positive regulation of cell proliferation	29	7.48E−18
GO:0031328 – positive regulation of cellular biosynthetic process	29	2.71E−12
GO:0009891 – positive regulation of biosynthetic process	29	3.83E−12
GO:0010557 – positive regulation of macromolecule biosynthetic process	28	5.71E−12
GO:0009719 – response to endogenous stimulus	27	4.71E−16
GO:0051173 – positive regulation of nitrogen compound metabolic process	26	1.49E−10
GO:0007049 – cell cycle	26	7.00E−09
GO:0009725 – response to hormone stimulus	25	5.44E−15
GO:0010628 – positive regulation of gene expression	25	1.05E−10
GO:0010647 – positive regulation of cell communication	24	5.11E−15
GO:0042325 – regulation of phosphorylation	24	7.71E−12
GO:0019220 – regulation of phosphate metabolic process	24	1.74E−11
GO:0051174 – regulation of phosphorus metabolic process	24	1.74E−11
GO:0045935 – positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	24	2.49E−09
GO:0008219 – cell death	24	3.57E−08
GO:0016265 – death	24	4.05E−08
GO:0051726 – regulation of cell cycle	23	6.11E−14
GO:0008285 – negative regulation of cell proliferation	23	3.57E−13
GO:0045893 – positive regulation of transcription, DNA-dependent	23	8.72E−11
GO:0051254 – positive regulation of RNA metabolic process	23	1.02E−10
GO:0045941 – positive regulation of transcription	23	2.06E−09
GO:0022402 – cell cycle process	23	2.13E−09
GO:0044093 – positive regulation of molecular function	23	4.18E−09
GO:0012501 – programmed cell death	23	8.97E−09
GO:0042592 – homeostatic process	23	3.42E−07
GO:0006915 – apoptosis	22	3.55E−08
GO:0006357 – regulation of transcription from RNA polymerase II promoter	22	8.29E−07
GO:0010605 – negative regulation of macromolecule metabolic process	22	9.68E−07
GO:0009967 – positive regulation of signal transduction	21	6.65E−13
GO:0043065 – positive regulation of apoptosis	21	6.04E−10
GO:0043068 – positive regulation of programmed cell death	21	6.82E−10
GO:0010942 – positive regulation of cell death	21	7.39E−10
GO:0006928 – cell motion	21	3.38E−09
GO:0001568 – blood vessel development	20	2.41E−13
GO:0001944 – vasculature development	20	3.72E−13
Other biological process	1349	<2.81E−05
Molecular function		
GO:0019899 – enzyme binding	17	1.43E−05
GO:0043566 – structure-specific DNA binding	10	5.43E−06
GO:0003690 – double-stranded DNA binding	9	2.22E−06
GO:0004714 – transmembrane receptor protein tyrosine kinase activity	8	2.00E−06
Cellular component		
GO:0044459 – plasma membrane part	39	5.15E−06
GO:0043233 – organelle lumen	33	2.71E−05
GO:0045121 – membrane raft	9	2.77E−05

2012). It is interesting that the rate of *PIK3CA* mutations in BAH is higher than it is in invasive carcinomas (Ang *et al*., 2014). This study provides some insight into the role of activating *PIK3CA* mutations in breast carcinogenesis and the precursor status of these early breast lesions. Subsequently, activated Akt such as AKT1 stimulates the regulation of cellular metabolism, growth, and survival by CCND1, MYC, NF-kB, and a variety of downstream factors (Koboldt *et al*., 2012; Khan *et al*., 2013; Deng *et al*., 2018). However, the rate of *AKT1* mutations is much higher in BAH than in breast cancer (Troxell *et al*., 2010). This phenomenon may suggest that *AKT1* mutations may play a role in precancerous disease. We could draw inspiration from these findings that the mutations of *PIK3CA* and *AKT1* play an important role in the early stage of malignant tumor formation, which may provide potential therapeutic targets for preventing the formation of malignant tumors and even precancerous lesions.

EGFR and *ERBB2 (HER2)*

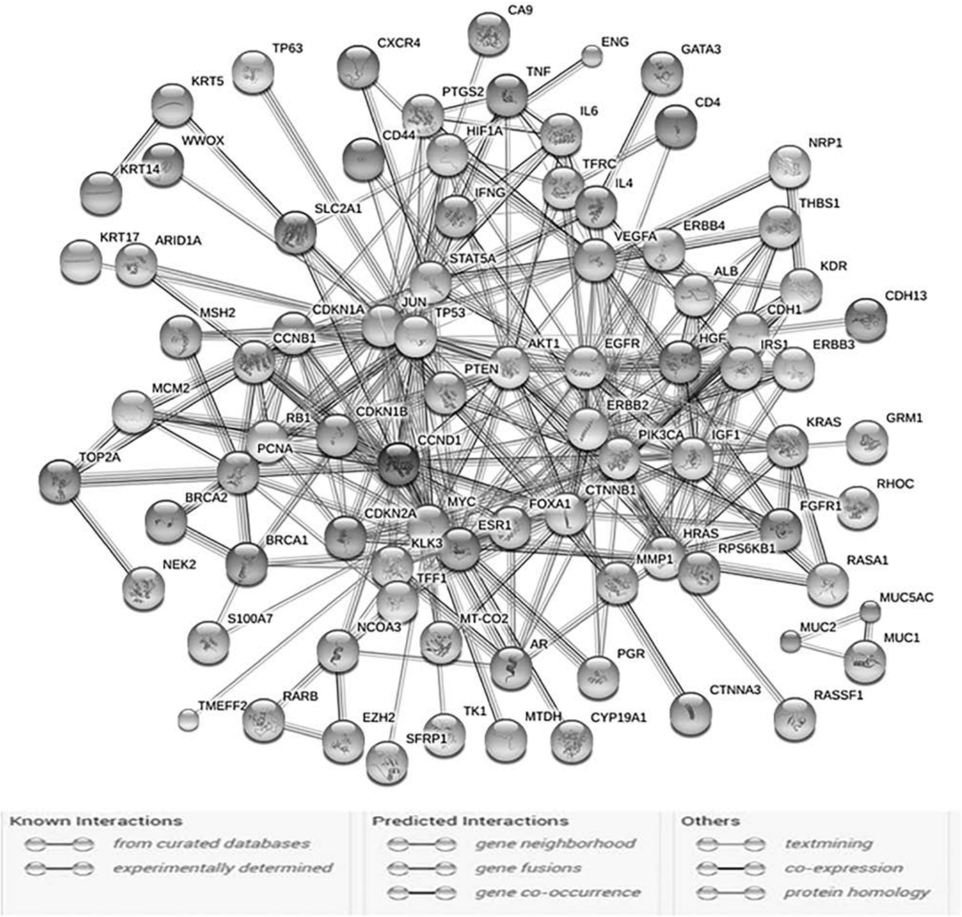
The oncogene *ERBB2 (c-erbB2/HER2)* is a well-established prognostic and predictive factor for invasive breast cancer and is a major driver of tumor development and progression in a subset of breast cancer following amplification (Popescu *et al*., 1989; Krishnamurti and Silverman, 2014). Following transphosphorylation, the dimerized receptor activates several intracellular signaling pathways, such as the Ras/MAPK pathway and the PI3K/Akt pathway, both of which subsequently affect cell proliferation, survival, motility, and adhesion (Moasser, 2007). It was reported that *ERBB2* amplification may predict substantially increased risk for subsequent breast cancer in women with benign breast diseases, including BAH (Stark *et al*., 2000). The overexpressed ERBB2 receptor may be a valuable therapeutic target not only for breast cancer but also for atypical hyperplasia.

Table 3 The 20 most significant pathways associated with atypical hyperplasia-related genes

Terms	Count	*P* value
Pathways in cancer	36	2.04E−40
Proteoglycans in cancer	25	5.09E−31
MicroRNAs in cancer	24	9.19E−26
PI3K–Akt signaling pathway	22	1.05E−21
Prostate cancer	18	3.97E−26
HTLV-I infection	18	1.79E−18
Endocrine resistance	17	8.86E−24
Hepatitis B	16	1.84E−19
Bladder cancer	15	2.77E−25
Melanoma	15	3.43E−22
EGFR tyrosine kinase inhibitor resistance	14	1.08E−19
ErbB signaling pathway	14	3.10E−19
HIF-1 signaling pathway	14	2.30E−18
FoxO signaling pathway	14	6.77E−17
Focal adhesion	14	1.46E−14
Endometrial cancer	13	3.24E−20
Non-small-cell lung cancer	13	7.54E−20
Rap1 signaling pathway	13	5.16E−13
Glioma	12	2.38E−17
Central carbon metabolism in cancer	12	3.30E−17

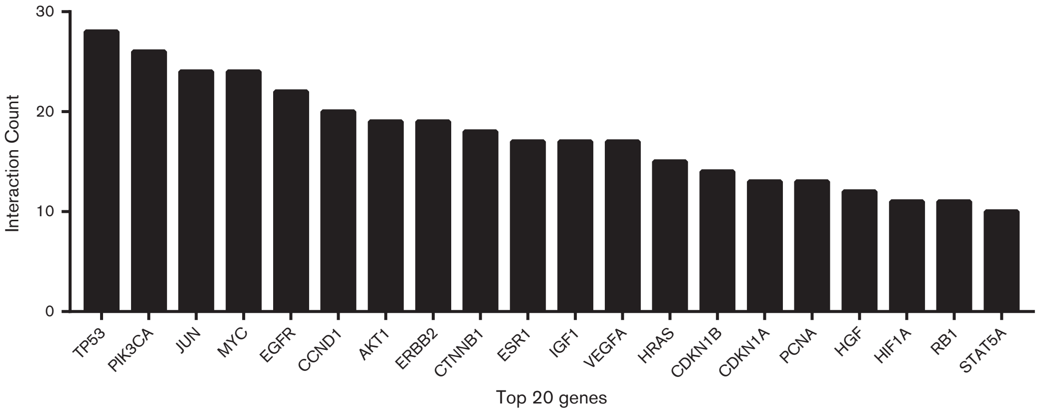
EGFR, also known as *ERBB1/HER1*, is another member of the epidermal growth factor receptor family. After ligand activation, phosphorylated EGFR provides a binding domain for PKC, PI3K/Akt/mTOR, SRC, STAT, and RAS/RAF/MEK1/ERK1/2 activation (Yarden and Sliwkowski, 2001). EGFR overexpression causes hyperplastic, dysplastic, and neoplastic changes in the mammary epithelium of transgenic mice (Brandt *et al*., 2000). Approximately 48% of primary human breast cancers exhibit EGFR overexpression (Klijn *et al*., 1992), and in women with atypical hyperplasia, fine needle aspiration

Fig. 2



Network analysis of breast atypical hyperplasia-related genes.

Fig. 3



Hub genes of breast atypical hyperplasia.

results showed that EGFR overexpression was 59% (Fabian *et al.*, 2015). However, the detection of EGFR in human biopsy tissue samples has not yet been reported, which may become a future direction of BAH research.

CTNNB1

CTNNB1 encodes β-catenin as a pivotal biomolecule that can not only combine with E-cadherin, T-cell factor, and lymphatic enhancement factor but also contact the complex composed of glycogen synthase kinase 3β, adenomatous polyposis coli, and axin. β-catenin is among a complex of proteins that constitute adherens junctions; it also plays a central role in transcriptional regulation in the Wnt signaling pathway (Hatsell *et al.*, 2003). Current evidence supports the disputation that the β-catenin/Wnt pathway is activated in a subgroup of breast cancers; however, the mechanisms leading to β-catenin nuclear accumulation in breast cancer remain elusive. There is a hypothesis that *CTNNB1*-activating gene mutations drive β-catenin nuclear expression (Hayes *et al.*, 2008; Geyer *et al.*, 2011). However, β-catenin/Wnt pathway activation in breast cancer is not commonly thought to be driven by *CTNNB1* mutations in the triple-negative phenotype. *CTNNB1* has been intensively studied in breast cancer, but its role in precancerous lesions requires further investigation in the future.

IGF-1

After insulin-like growth factor 1 (IGF-1) binding to insulin-like growth factor receptor 1 (IGF-1R), the complex activates numerous downstream pathways, such as the PI3K–AKT1–mTOR (Stewart *et al.*, 1990; Lee *et al.*, 1999; Rowinsky *et al.*, 2007; Naing *et al.*, 2011; Macaulay *et al.*, 2013; Iams and Lovly, 2015) and MAPK (Yamauchi and Pessin, 1994) pathways. IGF-1 plays a key role in the multistep process that leads from normal breast tissue to hyperplasia and then to malignancy (Kleinberg et al., 2009, 2011). However, in different mouse models, published data have shown that blockade of IGF-I action in the mammary gland prevents premalignant mammary lesion development (Hadsell and Bonnette, 2000; Carboni *et al.*, 2005; Singh *et al.*, 2014). The increased risk for breast cancer among women with benign breast diseases (including atypia) may be related to an apparent tendency to have lower levels of IGF-1 than those in healthy controls, notably among perimenopausal/postmenopausal women. The expression of IGF-1R is slightly increased in lesions (such as atypical ductal hyperplasia and columnar cell changes) that are hormonally driven, whereas it was significantly reduced in estrogen receptor-negative lesions (such as apocrine metaplasia). This observation may suggest that IGF-1 plays an important role in hyperplasia, even in atypical hyperplasia and breast cancer.

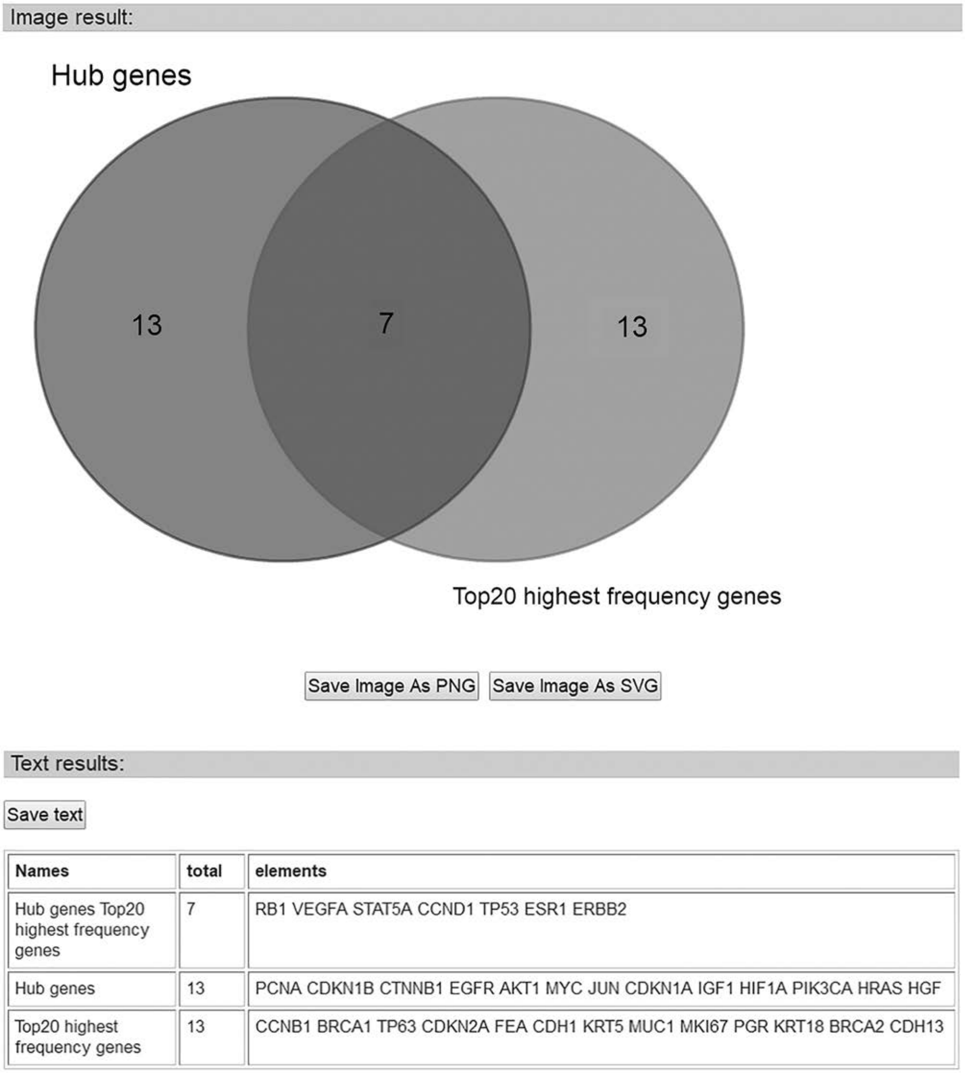
Typical hyperplasia of the breast is the key in the evolution from benign disease to malignancy. The levels of BAH-related gene expression and mutation are not the same as those in malignant tumors. Our study may provide some insight for future work in choosing the research topics; however, the results of our analyses are affected by some methodological limitations that should be considered. Much work remains for understanding the mechanism of progression from BAH to malignancy.

Although many epidemiology studies have clarified the risk associated with atypical hyperplasia and carcinoma in situ, there are no specific morphological or clinical features that help identify the high risk of developing invasive breast cancer. Further research into the molecular events occurring at the hyperplastic and in-situ stages is essential to understanding and identifying BAH as a high-risk disease for progression to invasive carcinoma.

Acknowledgements

This study was supported by the National Science Foundation of China (Grant no. 81773163).

Fig. 4



Names	total	elements
Hub genes Top20 highest frequency genes	7	RB1 VEGFA STAT5A CCND1 TP53 ESR1 ERBB2
Hub genes	13	PCNA CDKN1B CTNNB1 EGFR AKT1 MYC JUN CDKN1A IGF1 HIF1A PIK3CA HRAS HGF
Top20 highest frequency genes	13	CCNB1 BRCA1 TP63 CDKN2A FEA CDH1 KRT5 MUC1 MKI67 PGR KRT18 BRCA2 CDH13

Similarities and differences between breast atypical hyperplasia-related hub genes and the top 20 highest frequency genes.

Conflicts of interest

There are no conflicts of interest.

References

Allred DC, Mohsin SK, Fuqua SA (2001). Histological and biological evolution of human premalignant breast disease. *Endocr Relat Cancer* **8**:47–61.

Ang DC, Warrick AL, Shilling A, Beadling C, Corless CL, Troxell ML (2014). Frequent phosphatidylinositol-3-kinase mutations in proliferative breast lesions. *Mod Pathol* **27**:740–750.

Bombonati A, Sgroi DC (2011). The molecular pathology of breast cancer progression. *J Pathol* **223**:307–317.

Brandt R, Eisenbrandt R, Leenders F, Zschiesche W, Binas B, Juergensen C, *et al.* (2000). Mammary gland specific hEGF receptor transgene expression induces neoplasia and inhibits differentiation. *Oncogene* **19**:2129–2137.

Carboni JM, Lee AV, Hadsell DL, Rowley BR, Lee FY, Bol DK, *et al.* (2005). Tumor development by transgenic expression of a constitutively active insulin-like growth factor I receptor. *Cancer Res* **65**:3781–3787.

Castaneda CA, Cortes-Funes H, Gomez HL, Ciruelos EM (2010). The phosphatidyl inositol 3-kinase/AKT signaling pathway in breast cancer. *Cancer Metastasis Rev* **29**:751–759.

Hartmann LC, Degnim AC, Santen RJ, Dupont WD, Ghosh K (2015). Atypical hyperplasia of the breast: risk assessment and management options. *N Engl J Med* **372**:78–89.

Degnim AC, Visscher DW, Berman HK, Frost MH, Sellers TA, Vierkant RA, *et al.* (2007). Stratification of breast cancer risk in women with atypia: a Mayo cohort study. *J Clin Oncol* **25**:2671–2677.

Deng L, Zhu X, Sun Y, Wang J, Zhong X, Li J, *et al.* (2018). Prevalence and prognostic role of PIK3CA/AKT1 mutations in chinese breast cancer patients. *Cancer Res Treat* [Epub ahead of print]

Dupont WD, Page DL (1985). Risk factors for breast cancer in women with proliferative breast disease. *N Engl J Med* **312**:146–151.

Ellis IO (2010). Intraductal proliferative lesions of the breast: morphology, associated risk and molecular biology. *Mod Pathol* **23** Suppl 2: 1–7.

Engelman JA (2009). Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nat Rev Cancer* **9**:550–562.

Ersahin T, Tuncbag N, Cetin-Atalay R (2015). The PI3K/AKT/mTOR interactive pathway. *Mol Biosyst* **11**:1946–1954.

Fabian CJ, Kamel S, Zalles C, Kimler BF (2015). Identification of a chemoprevention cohort from a population of women at high risk for breast cancer. *J Cell Biochem* **63**:112–122.

Gao Y, Niu Y, Wang X, Wei L, Lu S (2009). Genetic changes at specific stages of breast cancer progression detected by comparative genomic hybridization. *J Mol Med (Berl)* **87**:145–152.

Georgescu MM (2011). PTEN tumor suppressor network in PI3K–Akt pathway control. *Genes Cancer* **1**:1170.

Geyer FC, Lacroixtriki M, Savage K, Arnedos M, Lambros MB, MacKay A, *et al.* (2011). β-Catenin pathway activation in breast cancer is associated with triple-negative phenotype but not with CTNNB1 mutation. *Mod Pathol* **24**: 209–231.

Hadsell DL, Bonnette SG (2000). IGF and insulin action in the mammary gland: lessons from transgenic and knockout models. *J Mammary Gland Biol Neoplasia* **5**: 19–30.

Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, *et al.* (2005). Benign breast disease and the risk of breast cancer. *N Engl J Med* **353**:229–237.

Hartmann LC, Radisky DC, Frost MH, Santen RJ, Vierkant RA, Benetti LL, *et al.* (2014). Understanding the premalignant potential of atypical hyperplasia through its natural history: a longitudinal cohort study. *Cancer Prev Res (Phila)* **7**:211–217.

Hartmann LC, Degnim AC, Santen RJ, Dupont WD, Ghosh K (2015). Atypical hyperplasia of the breast: risk assessment and management options. *N Engl J Med* **372**:78–89.

Hatsell S, Rowlands T, Hiremath M, Cowin P (2003). Beta-catenin and Tcfs in mammary development and cancer. *J Mammary Gland Biol Neoplasia* **8**:145–158.

Hayes MJ, Thomas D, Emmons A, Giordano TJ, Kleer CG (2008). Genetic changes of Wnt pathway genes are common events in metaplastic carcinomas of the breast. *Clin Cancer Res* **14**:4038–4044.

Iams WT, Lovly CM (2015). Molecular pathways: clinical applications and future direction of insulin-like growth factor-1 receptor pathway blockade. *Clin Cancer Res* **21**:4270.

Karakas B, Bachman KE, Park BH (2006). Mutation of the PIK3CA oncogene in human cancers. *Br J Cancer* **94**:455–459.

Kehr EL, Jorns JM, Ang D, Warrick A, Neff T, Degnin M, *et al.* (2012). Mucinous breast carcinomas lack PIK3CA and AKT1 mutations. *Hum Pathol* **43**:2207–2212.

Khan KH, Yap TA, Yan L, Cunningham D (2013). Targeting the PI3K–AKT–mTOR signaling network in cancer. *Chin J Cancer* **32**:253–265.

Kleinberg DL, Wood TL, Furth PA, Lee AV (2009). Growth hormone and insulin-like growth factor-i in the transition from normal mammary development to preneoplastic mammary lesions. *Endocr Rev* **30**:51–74.

Kleinberg DL, Ameri P, Singh B (2011). Pasireotide, an IGF-I action inhibitor, prevents growth hormone and estradiol-induced mammary hyperplasia. *Pituitary* **14**: 44–52.

Klijn JG, Berns PM, Schmitz PI, Foekens JA (1992). The clinical significance of epidermal growth factor receptor (EGF-R) in human breast cancer: a review on 5232 patients. *Endocr Rev* **13**: 3–17.

Cancer Genome Atlas N (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490**:61–70.

Krallinger M, Leitner F, Valencia A (2010). Analysis of biological processes and diseases using text mining approaches. *Methods Mol Biol* **593**:341–382.

Krishnamurti U, Silverman JF (2014). HER2 in breast cancer: a review and update. *Adv Anat Pathol* **21**:100–107.

Lakhani SR (1999). The transition from hyperplasia to invasive carcinoma of the breast. *J Pathol* **187**:272–278.

Lee AV, Jackson JG, Gooch JL, Hilsenbeck SG, Coronado-Heinsohn E, Osborne CK, *et al.* (1999). Enhancement of insulin-like growth factor signaling in human breast cancer: estrogen regulation of insulin receptor substrate-1 expression in vitro and in vivo. *Mol Endocrinol* **13**:787–796.

Macaulay VM, Middleton MR, Protheroe AS, Tolcher A, Dieras V, Sessa C, *et al.* (2013). Phase I study of humanized monoclonal antibody AVE1642 directed against the type 1 insulin-like growth factor receptor (IGF-1R), administered in combination with anticancer therapies to patients with advanced solid tumors. *Ann Oncol* **24**:784–791.

Maglott D, Ostell J, Pruitt KD, Tatusova T (2006). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33**(Database issue):D54-D58.

Moasser MM (2007). The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene* **26**:6469–6487.

Naing A, Kurzrock R, Burger A, Gupta S,Lei X, Busaidy N, *et al.* (2011). Phase I trial of cixutumumab combined with temsirolimus in patients with advanced cancer. *Clin Cancer Res* **17**:6052–6060.

Newburger DE, Kashefhaghighi D, Weng Z, Salari R, Sweeney RT, Brunner AL, *et al.* (2013). Genome evolution during progression to breast cancer. *Genome Res* **23**: 1097–1108.

Page DL, Dupont WD, Rogers LW, Rados MS (2015). Atypical hyperplastic lesions of the female breast. A long-term follow-up study. *Cancer* **55**: 2698–2708.

Popescu NC, King CR, Kraus MH (1989). Localization of the human erbB-2 gene on normal and rearranged chromosomes 17 to bands q12-21.32. *Genomics* **4**: 362–366.

Rowinsky EK, Youssoufian H, Tonra JR, Solomon P, Burtrum D, Ludwig DL (2007). IMC-A12, a human IgG1 monoclonal antibody to the insulin-like growth factor I receptor. *Clin Cancer Res* **13**:5549.

Santen RJ, Mansel R (2005). Benign breast disorders. *N Engl J Med* **353**:275–285.

Settles B (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**:3191–3192.

Singh B, Smith JA, Axelrod DM, Ameri P, Levitt H, Danoff A, *et al.* (2014). Insulin-like growth factor-I inhibition with pasireotide decreases cell proliferation and increases apoptosis in pre-malignant lesions of the breast: a phase 1 proof of principle trial. *Breast Cancer Res* **16**:463.

Stark A, Hulka BS, Joens S, Novotny D, Thor AD, Wold LE, *et al.* (2000). HER-2/neu amplification in benign breast disease and the risk of subsequent breast cancer. *J Clin Oncol* **18**:267–274.

Stewart AJ, Johnson MD, May FE, Westley BR (1990). Role of insulin-like growth factors and the type I insulin-like growth factor receptor in the estrogen-stimulated proliferation of human breast cancer cells. *J Biol Chem* **265**:21172–21178.

Troxell ML, Levine J, Beadling C, Warrick A, Dunlap J, Presnell A, *et al.* (2010). High prevalence of PIK3CA/AKT pathway mutations in papillary neoplasms of the breast. *Mod Pathol* **23**:27–37.

Visscher DW, Frank RD, Carter JM, *et al.* (2017). Breast cancer risk and progressive histology in serial benign biopsies. *J Natl Cancer Inst* **109**:10.

Vogt PK, Hart JR, Gymnopoulos M, Jiang H, Kang S, Bader AG, *et al.* (2010). Phosphatidylinositol 3-kinase: the oncoprotein. *Curr Top Microbiol Immunol* **347**:79–104.

Yamauchi K, Pessin JE (1994). Insulin receptor substrate-1 (IRS1) and Shc compete for a limited pool of Grb2 in mediating insulin downstream signaling. *J Biol Chem* **269**:31107–31114.

Yarden Y, Sliwkowski MX (2001). Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol* **2**:127–137.