Aus der Klinik für Gynäkologie und Geburtshilfe
der Heinrich-Heine-Universität Düsseldorf
Direktor: Univ.-Prof. Dr. med. Hans Neubauer

Expression, Clinical Significance, and Biological Functions of the Breast-specific Gene ANKRD30A in Breast Cancer

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Chen Chen

2022

Parts of this work have been published:

Chen Chen, Liwen Yang, Mahdi Rivandi, André Franken, Tanja Fehm, Hans Neubauer. Bioinformatic Identification of a Breast-Specific Transcript Profile. Proteomics Clinical Applications. 2020, Nov;14(6):e2000007. DOI: 10.1002/prca.202000007. https://pubmed.ncbi.nlm.nih.gov/32558282

Zusammenfassung

Hintergrund: Gewebespezifische Biomarker sind wertvoll für die Früherkennung eines Karzinoms, die Bewertung des Behandlungsansprechens, die Überwachung von Rezidiven, die Identifizierung von Tumorherden, die gezielte Verabreichung von Medikamenten und für die Tumor-Immuntherapie. Derzeit ist nur sehr wenig über Gene mit brustspezifischen Expression bekannt, insbesondere über die genaue Anzahl „brustspezifischer Gene“ im menschlichen Genom und ihre biologischen Funktionen. Die rasante Entwicklung von Hochdurchsatz-Genanalysetechniken hat enorme Daten und leistungsfähige bioinformatische Werkzeuge hervorgebracht, die die Identifizierung und Untersuchung brustspezifischer Gene erheblich erleichtern.

Ziele: 1) Identifizierung aller möglichen brustspezifischen Gene im menschlichen Genom mit bioinformatischen Methoden; 2) Selektion eines neuen brustspezifischen Kandidatengens, das zur Entstehung eines Mammakarzinoms (MaCa) beitragen kann, Untersuchung seiner biologischen Funktionen beim MaCa und Bewertung seiner potenziellen klinischen Bedeutung; 3) Etablierung eines Arbeitsablaufs zur Verwendung bioinformatischer Werkzeuge für die Unterstützung von Laborexperimenten und Validierung der Zuverlässigkeit der bioinformatischen Ergebnisse, wenn möglich.

Methoden: Gene mit brustspezifischen Expressionsmustern wurden mit dem Expressionsatlas identifiziert und mit dem Online-Server ShinyGO und Ensembl annotiert. Die Expression unseres Zielgens (ANKRD30A) in normalen menschlichen Geweben und Blutzellen wurde mit dem Human Protein Atlas überprüft und unter Verwendung der GEPIA-Datenbank bei 33 Tumorentitäten und entsprechenden normalen Vergleichsproben analysiert. Der Zusammenhang zwischen der ANKRD30A-Expression und den klinischen Merkmalen von MaCas wurde anhand des TCGA-BRCA-Datensatzes untersucht. Variationen in der Zahl der Genkopien und Mutationen von ANKRD30A beim MaCa wurden mit dem cBioPortal analysiert. Unter Verwendung des CCLE-BRCA-Datensatzes wurde die ANKRD30A-Expression in 51 MaCa-Zelllinien untersucht und in 6 MaCa-Zelllinien mittels RT-PCR und verschiedener Primer-Konfigurationen validiert. Die subzellulären Lokalisationen von ANKRD30A-Proteinen wurden unter Verwendung von Immunfluoreszenz-Assays analysiert. Die mögliche Funktion von ANKRD30A in MaCa-Zelllinien wurde mittels siRNA-vermitteltem „knock down“ der Genexpression analysiert und Effekte auf die Zellproliferation und Koloniebildung mit dem PrestoBlue-Zellviabilitäts- bzw. Plattenkoloniebildungsassay untersucht. Potenzielle biologische Funktionen von ANKRD30A beim MaCa und seiner assoziierten Gene wurden mithilfe der Gene Set Enrichment Analysis vorhergesagt. Die prognostische Signifikanz von ANKRD30A wurde mit dem Kaplan-Meier-Plotter bewertet.

Ergebnisse: Insgesamt wurden 96 potenziell brustspezifische Gene identifiziert. Unter ihnen wurde ANKRD30A für die weitere Analyse ausgewählt, da seine Sequenz die volle Länge fünf weiterer brustspezifischer Gene abdeckt, was ein einzigartiges Merkmal unter den identifizierten Kandidaten ist. Beim MaCa ist die Expression von ANKRD30A in ER-negativen Geweben und Zelllinien, insbesondere bei TNBC, signifikant herunterreguliert, mit anderen klinischen MaCa-Merkmalen in die Expression nicht signifikant assoziiert. Variationen in der Zahl der Genkopien und Mutationen von ANKRD30A sind sowohl bei primärem als auch beim metastasierten MaCa sehr selten. Die Sensitivität des ANKRD30A-mRNA-Nachweises durch PCR variiert stark mit den verwendeten Primerpaaren - insbesondere, wenn diese in verschiedenen Exon-Regionen lokalisiert sind. Die Lokalisation von ANKRD30A-Proteinen ist hauptsächlich zytoplasmatisch, während perinukleäres ANKRD30A in der MDA-MB-453-Zelllinie nachgewiesen werden konnte. Die Suppression von ANKRD30A steigerte die Fähigkeit von MaCa-Zellen zur Proliferation und Koloniebildung. Darüber hinaus ist die Herunterregulierung von ANKRD30A beim MaCa mit der Aktivierung von zellzyklus-bezogenen Signalen verbunden, insbesondere von CDC25A, MCM2, MCM4 und PLK1. Die Expression von LINC00993 und CDC25A war in ANKRD30A-supprimierten Zelllinien signifikant verringert. Eine hohe ANKRD30A-Expression beim MaCa weist auf bessere Überlebenswahrscheinlichkeit hin.

Schlussfolgerung: ANKRD30A ist ein brustspezifisch-exprimiertes Gen im menschlichen Genom, das insbesondere beim TNBC für eine verbesserte Behandlung sehr wertvoll sein könnte. In diesem Subtyp besteht ein Zusammenhang zwischen niedriger ANKRD30A-Expression und erhöhter Zellproliferation sowie der Aktivierung tumor-assoziierter Zellzyklus-Signalwege.

Summary

Background: Tissue-specific biomarkers are valuable for early cancer screening, treatment response assessment, recurrence monitoring, cancer source identification, targeted drug delivery, and cancer immunotherapy. Currently, very little is known about genes with breast specificity, especially the exact number of breast-specific genes in the human genome and their biological functions. The rapid development of high-throughput gene analysis techniques has yielded tremendous data and potent bioinformatics tools, greatly facilitating the identification and investigation of breast-specific genes.

Objectives: 1) To identify all possible breast-specific genes in the human genome using bioinformatics methods, and focus on one novel breast-specific gene that may contribute to the development of breast cancer. 2) To investigate biological functions of the candidate breast-specific gene in breast cancer and evaluate its potential clinical significance. 3) To establish a workflow of using bioinformatics tools to assist laboratory experiments, and validate the reliability of bioinformatics findings when possible.

Methods: Genes with breast-specific expression patterns were screened using the Expression Atlas and annotated using the ShinyGO and Ensembl online server. The expression of our targeted gene (ANKRD30A) in normal human tissues and blood cells was reviewed using the Human Protein Atlas. ANKRD30A expression in 33 types of cancers and corresponding normal counterparts were analyzed using the GEPIA database. The association between ANKRD30A expression and breast cancer clinical characteristics were assessed using the TCGA-BRCA dataset. Copy number variations and mutations of ANKRD30A in breast cancer were analyzed using the cBioPortal. ANKRD30A expression in 51 types of breast cancer cell lines was examined using the CCLE-BRCA dataset. ANKRD30A mRNA expression was validated in 6 breast cancer cell lines using RT-PCR with screened primers. Subcellular locations of ANKRD30A proteins were analyzed using the immunofluorescence assays. ANKRD30A mRNA expression was silenced using siRNAs. Roles of ANKRD30A in cell proliferation and colony formation were investigated using the PrestoBlue cell viability and plate colony formation assays, respectively. Potential ANKRD30A biological functions in breast cancer and its correlated genes were predicted using the Gene Set Enrichment Analysis. The prognostic significance of ANKRD30A was evaluated using the Kaplan-Meier Plotter.

Results: A total of 96 potential breast-specific genes were identified. Among them, ANKRD30A was selected for further analysis because its sequence covers the full length of other five breast-specific genes, which is an unique feature among the identified candidates. In different types of normal female tissues, cancer tissues, and blood cells, the expression of ANKRD30A is breast-specific. In breast cancer, the expression of ANKRD30A is significantly down-regulated in ER-negative tissues and cell lines, especially in TNBC, while its correlation with other clinical characteristics of breast cancer is not significant. Copy number variations and mutations of ANKRD30A are very rare in both primary and metastatic breast cancer. The sensitivity of ANKRD30A mRNA detection by PCR varies greatly if different primer pairs were used, especially primer pairs targeting its different exon regions. The location of ANKRD30A proteins is mainly cytoplasmic, whereas perinuclear ANKRD30A can be detected in the MDA-MB-453 cell line. Knocking down ANKRD30A with siRNA could promote the ability of breast cancer cells in proliferation and colony formation. Moreover, the down-regulation of ANKRD30A in breast cancer is associated with the activation of cell-cycle-related signaling, especially CDC25A, MCM2, MCM4, and PLK1. The expression of LINC00993 was significantly decreased while CDC25A was significantly increased in ANKRD30A-silenced cell lines. High ANKRD30A expression in breast cancer usually indicates better survival outcomes.

Conclusion: ANKRD30A is one of the rare breast-specific genes in the human genome, and its well-established breast specificity is valuable for improving the management of breast cancer. ANKRD30A is significantly and frequently down-regulated in TNBC, which is possibly oncogenic because of the association between low-ANKRD30A and increased cell proliferation as well as the activation of cancer-related cell cycle signaling.

List of Abbreviations

Abbreviation	Full Name
95%CI	95% confidence interval
aa	Amino acids
ANKRD30A	Ankyrin repeat domain-containing protein 30A
ASCO	American society of clinical oncology
bp	Base pair
BRCA	Breast cancer
bZIP	Basic region leucine zipper
CA15-3	Cancer antigen 15-3
CCLE	Cancer cell line encyclopedia
CCLE-BRCA	The cancer cell line encyclopedia breast cancer dataset
CDC25A	Cell division cycle 25A
cDNA	Complementary DNA
CEA	Carcinoembryonic antigen
cfDNA	Cell-free DNA
CNV	Copy number variation
CO2	Carbon dioxide
Cq	Quantification cycle
CTC	Circulating tumor cell
ctDNA	Circulating cell-free tumor DNA
DAPI	4′,6-diamidino-2-phenylindole
DCIS	Ductal carcinoma in situ
DFS	Disease-free survival
DMEM	Dulbecco’s modified eagle medium
DMFS	Distant metastasis free survival
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPBS	Dulbecco’s phosphate-buffered saline
DTC	Disseminated tumor cell

EDTA	Ethylenediaminetetraacetic acid
EGA	European genome-phenome archive
EMT	Epithelial-mesenchymal transition
EP tube	Eppendorf tube
ER	Estrogen receptor
ESR1	Estrogen receptor 1 (gene)
FDA	U.S. food and drug administration
FDR	False discovery rate
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase GATA3
GDC	Genomic data commons GEO
GEPIA	The gene expression profiling interactive analysis
GO	Gene ontology
GSEA	Gene set enrichment analysis
GTEx	Genotype-tissue expression
HER2	Human epidermal growth factor receptor 2
HPA	Human Protein Atlas
HR	Hazard ratio
IHC	Immunohistochemistry
KEGG	Kyoto encyclopedia of genes and genomes
KI67	Marker of proliferation Ki-67
LINC00993	Long intergenic non-protein coding RNA 993
MCM2	Minichromosome maintenance complex component 2
MCM4	Minichromosome maintenance complex component 4
METABRIC	Molecular taxonomy of breast cancer international consortium
MIQE	Minimum information for publication of quantitative real-time PCR experiments
mL	Milliliter
mRNA	Messenger RNA
MSigDB	Molecular signatures database
MUC1	Mucin 1
MUCL1	Mucin Like 1

mut	Mutation
NA	Not available
nc-siRNA	Negative control siRNA
NES	Normalized enrichment score
ng	Nanogram
nm	Nanometer
nM	Nanomolar
ns	Not significant
NX	Normalized expression
OS	Overall survival
PAM50	Prediction analysis of microarray 50
PCA3	Prostate-specific prostate cancer gene 3
PCR	Polymerase chain reaction
PFS	Progression-free survival
PLK1	Polo like kinase 1
PBMC	Peripheral blood mononuclear cells
PPS	Palliative performance scale
PRAD	Prostate adenocarcinoma
qRT-PCR	Real-time quantitative reverse transcription PCR
RFS	Relapse-free survival
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPPA	Reverse-phase protein array
RT-PCR	Reverse transcription polymerase chain reaction
SBEM	Small breast epithelial mucin
SD	Standard deviation
SEREX	Serological analysis of recombinant tumor cDNA expression libraries
siRNA	Small interfering RNA
SLN	Sentinel lymph node
SNP	Single nucleotide polymorphism
TAE	Tris-acetate-EDTA

TCGA	The cancer genome atlas
TF	Transcription factor
Tg	Thyroid-specific thyroglobulin
TNBC	Triple-negative breast cancer
TPM	Transcripts per million
UV	Ultraviolet
wt	Wild type
μg	Microgram
μL	Microliter
μM	Micromolar

Table of Contents

1 Introduction

1.1 Breast Cancer

1.1.1 Incidence

Breast cancer is the most prevalent female malignant tumor in 159/185 countries, representing 1 in 4 women cancer cases globally. [1] The lifetime rate of breast cancer incidence is 12.8%, which means 1 in 8 women will be diagnosed with invasive breast cancer in their lifetime. [2] Moreover, the incidence of breast cancer is usually higher in developed countries (55.9 per 100,000) than in developing countries (29.7 per 100,000), reflecting both the increased detection rates and changes of risk factors. [1] Despite the incidence of breast cancer in developed counties becomes stable, it rapidly increases in developing countries, especially in sub-Saharan Africa. [1] Several preventable non-genetic risk factors, well-established or newly discovered, may contribute to the increased incidence, including delayed childbearing, fewer births, lower rates of breastfeeding, overweight or obesity, physical inactivity, and alcohol intake. [3]

1.1.2 Mortality

Breast cancer is the fifth leading cause of cancer-related death worldwide, representing 1 in 6 female cancer deaths in 110 countries. [1] In contrast to its incidence, the rates of mortality are higher for patients with breast cancer in developing countries (15.0 per 100,000) than those in developed countries (12.8 per 100,000). In recent decades, breast cancer mortality rates decreased significantly due to breakthroughs in effective treatments, with a total of 40% decline from 1975 to 2017. [2] Currently, the 5-year breast cancer overall survival (OS) rates are relatively high for non-metastatic breast cancers (98%, 92%, and 75% for stage I, stage II, and stage III patients, respectively). [2] However, for patients with stage IV (metastatic) breast cancers, the 5-year breast cancer OS rate is only 27%, highlighting the significance of early detection and treatment. Moreover, about 1 in 4 patients with early-stage breast cancer will have inoperable locally advanced or metastatic diseases that remain virtually incurable at the current stage, with a 5-year OS rate of 25% and a median OS of 3 years. [4]

1.1.3 Subtype

Based on the multigene signatures determined by gene expression profiling, breast cancer can be categorized into five intrinsic molecular subtypes, namely Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal Breast-like. [5,6] However, immunohistochemical methods are the most commonly used alternatives in daily clinical practice. [7–9] Specifically, breast cancer can be immunohistochemically classified into four main subtypes (IHC4) based on the expression of estrogen receptor (ER), progesterone receptor (PG), human epidermal growth factor receptor 2 (HER2; also known as ERBB2), and marker of proliferation Ki-67 (KI67), namely the Luminal A, Luminal B, HER2-enriched, and triple-negative breast cancer (TNBC) subtype. Meanwhile, breast cancer can also be divided into three major subtypes based on expression status of ER, PR, and HER2, namely the hormone-receptor-positive (HR-positive) / HER2-negative subtype (ER-positive or PR-positive, HER2-negative), the HER2-positive subtype (HR-negative and HER2-positive), and the triple-negative tumors (HR-negative and HER2-negative). [10]

1.1.4 Treatment

Typically, there are two major types of treatment for patients with breast cancer, including local treatments and systematic therapies. [11] Specifically, local treatments include surgery and radiotherapy, while systematic therapies mainly involve chemotherapy, endocrine therapy, anti-HER2 targeted therapy, immunotherapy, and other targeted therapies on novel actionable targets. Actionable key molecular targets in breast cancer are critical for improving its treatment effectiveness, especially ER and HER2. ER-blocking endocrine therapies with tamoxifen reduce the 15-year breast cancer overall mortality by 30-40%. [12–14] Meanwhile, anti-HER2 targeted therapies with trastuzumab-containing regimens contribute to a 40% decrease in 10-year disease-free survival (DFS) and a 37% reduction in 10-year OS. [15] However, non-specific chemotherapy remains the most widely used strategy for treating the high-aggressive TNBC tumors due to the lack of effective therapeutic targets. [16,17]

1.2 Biomarkers for Breast Cancer

1.2.1 Limitations of Existing Breast Cancer Markers

Apart from treatments, biomarkers are also important for the screening, diagnosis, prognosis, and surveillance of breast cancer. The American Society of Clinical Oncology (ASCO) has issued clinical practice guidelines regarding the use of biomarkers to guide the management of women with early-stage and metastatic breast cancer. [18,19] However, limitations still exist regarding the clinical applications of existing markers. For example, methods for determining the status of most critical breast cancer biomarkers (such as ER, PR, and HER2) are generally invasive, such as core-needle biopsy or surgical resection. [20,21] Moreover, despite the efficacy of multigene signatures for aiding inconclusive clinical decisions have been well validated, the high costs indeed prohibit their more widespread use in clinical routine. [18] Besides, there is still no reliable biomarker for monitoring recurrence and treatment response, especially in metastatic breast cancer. Only some empirically chosen markers are available for these purposes, such as CEA and CA15-3, and their routinely clinical utility is even not recommended. [19] Therefore, considerable efforts have been devoted to the development and validation of novel biomarkers. [22–24]

1.2.2 CTCs and ctDNA

In recent decades, the non-invasive “real-time” liquid biopsy techniques aiming at the detection of circulating tumor cells (CTCs) and circulating cell-free tumor DNA (ctDNA) have become a promising and attractive research field. [25–27] In breast cancer, the prognostic significance of CTCs in the blood of patients with metastatic diseases has been well demonstrated. Generally, five or more detectable CTCs in 7.5 mL blood samples indicate a shorter progression-free survival (PFS) and OS. [28–30] However, the predictive values of CTCs have not been proven yet, as the survival outcomes of patients who received personalized treatment based on CTCs levels are not significantly different from the controls. [31] Moreover, CTCs are extremely rare in the blood, ranging from 1 to 10 CTCs in 10^6–10^8 white blood cells, especially in early-stage patients. [32–34] Meanwhile, the CTC heterogeneity caused by the epithelial-mesenchymal transition (EMT) plasticity further increases the difficulty of their detection. [35–37] Therefore, it is technically challenging to

enrich CTCs from an enormous background of leukocytes with both high sensitivity and satisfactory specificity. [25,26] Similar to CTCs, ctDNA is considered a valuable biomarker for predicting response and detecting early relapse. [34,38,39] However, several theoretical and technical issues also limit its clinical applications. [40] For example, the extremely low amount of ctDNA in blood samples poses the greatest challenge for ctDNA detection. [26,41,42] Besides, ctDNA-associated analyses are mainly focused on genetic mutations and epigenetic changes at the DNA level, hence other hallmarks of cancer will be technically omitted. [41,43] Moreover, some "cancer-associated" mutations can also be detected in the cell-free DNA (cfDNA) of healthy individuals, especially in the older population, which could reduce the specificity of ctDNA analysis. [26,44]

1.2.3 Tissue-specific Biomarkers

Notably, markers with tissue-specific expression patterns are promising candidates, as tissue-specific markers are generally considered valuable for screening early cancer, assessing treatment response, monitoring recurrence, identifying the source of carcinoma of unknown primary, and serving as targets for cancer immunotherapy. [45–48] For example, the thyroid-specific thyroglobulin (Tg) has been widely used in post-thyroidectomy surveillance for predicting treatment response and monitoring recurrence of patients with differentiated thyroid cancer. [49–51] Moreover, the first FDA-approved urine-based cancer detection assay PROGENSA, which was specially designed for the testing of the prostate-specific prostate cancer gene 3 (PCA3), has been confirmed helpful to reduce unnecessary repeat prostate biopsies in suspicious prostate cancer cases. [52,53]

1.2.4 Breast-specific Biomarkers

Previously, several genes with breast-specific expression patterns were identified. Among them, mammaglobin may be the most famous one and has been considered a promising biomarker for breast cancer diagnosis. [54–56] The positive mammaglobin expression in metastatic lesions was a helpful marker to identify carcinoma of breast origin, with a specificity of 85-100%. [57,58] Moreover, mammaglobin has also been proposed as a promising therapeutic molecular target for breast cancer. [59] However, the sensitivity of mammaglobin detection varies greatly, ranging from 26% to 84%, especially in TNBC tumors. [60–62] Besides, the breast specificity of mammaglobin has become controversial as the positive mammaglobin expression can also be detected in some non-breast tissues like endometrial, ovarian, and cervical tissues. [63,64] Despite some other genes are also considered as breast-specific, such as SBEM, MUC1, and GATA3, their breast specificity and potential clinical utility are rarely demonstrated in a large sample size. [54,65,66] To date, the exact number of genes with breast specificity in the human genome remains unknown. Therefore, a comprehensive view is needed to get a better understanding of breast-specific markers, especially for novel ones with potential clinical values.

1.3 Bioinformatics Tools

The rapid development of high-throughput gene analysis techniques has yielded tremendous data and potent bioinformatics tools, tremendously improving our understanding of the human genome, both physically and pathologically. Among them, several high-quality, large-scale,

genome-wide genetic studies on breast cancer, normal human tissues, and cancer cell lines greatly promote the identification of breast-specific genes and the prediction of their potential biological functions. The main bioinformatics tools used in this study were listed as follows, including Expression Atlas, cBioPortal, Human Protein Atlas, Ensembl, TCGA, METABRIC, CCLE, GTEx, Gene Set Enrichment Analysis, and Kaplan-Meier Plotter.

1.3.1 Expressing Atlas

Expression Atlas is a powerful online database that provides manually curated gene expression datasets from over 3,000 experiments across 40 different organisms (www.ebi.ac.uk/gxa/home). [67] Raw and normalized gene expression data from included RNA-seq or microarray studies are freely accessible, enabling customized analyses for various purposes. The three landmark datasets that described gene expression signatures in various types of normal human tissues were included in Expression Atlas, namely the The Genotype-Tissue Expression (GTEx), [68] RIKEN FANTOM5, [69] and Illumina Body Map. [70,71] Specifically, the curated GTEx dataset provides the expression data of 46,711 genes across 53 tissue types, the RIKEN FANTOM5 dataset provides the expression data of 21,105 genes across 76 tissue types, and the Illumina Body Map dataset provides the expression data of 46,754 genes across 16 tissue types.

1.3.2 Breast Cancer Datasets

The Cancer Genome Atlas (TCGA) is a landmark cancer genomics project (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) that characterized genomic, epigenomic, transcriptomic, and proteomic data from over 20,000 primary cancer and matched normal samples spanning 33 cancer types, including breast cancer (TCGA-BRCA). [72] Raw TCGA data can be accessed through the official Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov). However, since it is complicated to retrieve and process the raw data, many databases provided the ready-to-use datasets or service, such as the GEPIA, [73,74] TCGA Portal, [75] and cBioPortal [76] database.

In detail, the TCGA-BRCA dataset documented the expression of 60,478 genes in 1,217 breast cancer samples, including 502 Luminal A tumors, 199 Luminal B tumors, 78 HER2-enriched tumors, 172 Basal tumors, 36 Normal-like tumors, and 121 tumors with unknown subtype. Moreover, the corresponding clinical information of these samples is also available, which makes it possible to evaluate the association between gene expression and the clinical characteristics of breast cancer. Similarly, the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [77] dataset also comprehensively analyzed the genome of breast cancer tissues at the DNA and RNA levels, providing the microarray expression data of 24,368 genes in 1,904 breast cancer specimens. Among these samples, 700 were Luminal A tumors, 475 were Luminal B tumors, 224 were HER2-enriched tumors, 209 were Basal tumors, and 148 were Normal-like tumors.

1.3.3 Cancer Cell Line Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (CCLE) project performed a detailed genetic characterization to include genetic, RNA splicing, DNA methylation, histone H3 modification, microRNA expression, and reverse-phase protein array (RPAA) data of a large panel of cancer cell lines

(portals.broadinstitute.org/ccle). [78,79] The RNA-seq mRNA expression data of 57,820 genes in 1,019 cell lines were publicly available, including a total of 71 cell lines of breast source (CCLE-BRCA). Moreover, detailed phenotypes of these cell lines were also provided, such as histology, gender, race, age, site of finding, corresponding diseases, and so on.

1.3.4 GEPIA

The Gene Expression Profiling Interactive Analysis (GEPIA) database is an interactive online server for analyzing normalized RNA-seq gene expression data from the TCGA and GTEx datasets (gepia.cancer-pku.cn). [73,74] By simply entering the name of a certain gene for searching in the GEPIA website, it is quite easy to obtain its expression plot in 33 types of cancer and corresponding normal tissues, involving 9,736 tumor and 8,587 normal tissue samples. Besides, customizable co-expression analyses based on cancer types were also provided, and the correlation coefficients as well as corresponding figures will be auto-generated.

1.3.5 cBioPortal

The cBioPortal is an open-access interactive web server of multidimensional cancer genomics datasets, storing actively curated data regarding standardized mRNA expression, somatic mutation, DNA copy-number variation, microRNA expression, DNA methylation, RPPA or mass spectrometry based protein/phosphoprotein expression, and corresponding clinical characteristics (www.cbioportal.org). [76] Notably, it is not difficult to get the expression heatmap of one or several genes in a certain dataset using its “OncoPrint” function. Furthermore, customizable analyses for other purposes are also available, such as genetic alteration, co-expression, and survival analysis.

1.3.6 Human Protein Atlas

The Human Protein Atlas (HPA) is a program that aims to map all human proteins in cells, tissues, and organs using multi-omic techniques, including transcriptomics data, mass spectrometry-based proteomics data, corresponding high-quality images, and other biological annotations (www.proteinatlas.org). [80–84] The RNA expression overview in HPA integrated normalized RNA-seq data from the HPA, GTEx, and RIKEN FANTOM5 datasets, providing the expression data of 19,651 genes in 37 tissue types. Meanwhile, the HPA also summarized the expression data of 19,651 genes in 18 types of normal blood cells and peripheral blood mononuclear cells (PBMCs). Moreover, the expression data of 19,651 genes in 69 types of cell lines across 12 groups were also presented.

1.3.7 Ensembl

Ensembl is a genome database that describes multiple fields of vertebrate genomics, in particular gene annotation, comparative genomics, genetics and epigenomics (www.ensembl.org). [85,86] Annotations regarding gene synonyms, chromosome locations, gene types, details of transcripts and corresponding proteins are well integrated in the Ensembl database. Meanwhile, the location of a certain gene and its adjacent genes in a specific chromosome region can be viewed intuitively using its interactive genome browser.

1.3.8 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational algorithm to determine the statistical correlation between a priori defined gene set and different biological phenotypes (www.gsea-msigdb.org). [87] The Molecular Signatures Database (MSigDB) of GSEA stored a collection of well-annotated gene sets of nine major categories, where many cancer-related gene sets were available, such as the Hallmark gene set and the KEGG pathway gene set. Using the freely accessible GSEA software with corresponding datasets prepared as required, it is feasible to predict the potential biological functions of a previously unstudied gene in a certain disease or phenotype, effectively guiding the direction of further laboratory experiments.

1.3.9 Kaplan-Meier Plotter

The Kaplan Meier plotter is an online server to predict the impacts of over 54,000 genes on the survival outcomes of 21 cancer types (kmplot.com). [88] By integrating expression and matched survival data from the Gene Expression Omnibus (GEO), European Genome-Phenome Archive (EGA), and TCGA databases, the Kaplan Meier plotter enables the meta-analysis based discovery and validation of potential prognostic biomarkers. Specifically, the prognostic significance of a certain gene can be evaluated in 7,830 breast cancer patients for assessing its impacts on relapse-free survival (RFS; n=4,934), OS (n=1,880), distant metastases-free survival (DMFS; n=2,767), and palliative performance scale (PPS; n=458). Additionally, customizable subgroup analyses based on different ER status, PR status, HER2 status, intrinsic subtypes, lymph node status, grades, and TP53 status can also be performed.

1.4 Aims of this Study

The hypotheses and main objectives of this study were summarized as follows:

1. We hypothesized that the expression patterns of certain genes in the human genome are breast-specific, and some of these breast-specific genes are critical for the development of breast cancer. Therefore, we will identify all possible breast-specific genes in the human genome using bioinformatics methods, and focus on one novel breast-specific gene that may contribute to the development of breast cancer.
2. We hypothesized that some important breast-specific genes could be utilized to improve the management of breast cancer. Hence, we will investigate the biological functions of the candidate breast-specific gene in breast cancer and evaluate its potential clinical significance.
3. We hypothesized that *in silico* datasets and tools could facilitate laboratory experiments by guiding the direction of involved studies. Therefore, we will demonstrate the feasibility of using bioinformatics tools to assist laboratory experiments and validate the reliability of corresponding *in silico* findings when possible.

2 Material and Methods

2.1 Materials

2.1.1 Chemicals

Chemical	Source	Catalog Number
Accutase™ Cell Dissociation Reagent	Thermo Fisher Scientific	A1110501
Agarose	Sigma	A9539-500G
ANKRD30A Antibody	Affinity Biosciences	DF3421
Antibody Diluent	Agilent	S080983-2
Aurum™ Total RNA Mini Kit	Bio-Rad	732-6820
Bovine Serum Albumin (BSA)	Sigma	A7030-100G
Crystal Violet Solution	Sigma	V5625
CTS™ Opti-MEM™ I Medium	Thermo Fisher Scientific	A4124802
DAPI	Thermo Fisher Scientific	62248
DEPC-Treated Water	Thermo Fisher Scientific	750023
DMSO	Thermo Fisher Scientific	D12345
Donkey anti-Rabbit IgG AF647	Thermo Fisher Scientific	A31573
DPBS	Thermo Fisher Scientific	14190094
DreamTaq PCR Master Mix	Thermo Fisher Scientific	K1071
Dynabeads™ Goat anti-Mouse IgG	Thermo Fisher Scientific	11033
EDTA	Sigma	E5134-250G
Fetal Bovine Serum	Thermo Fisher Scientific	10270106
Fluorescence Mounting Medium	Agilent	S3023
GelRed	Sigma	SCT123
Ham’s F-12K (Kaighn’s) Medium	Thermo Fisher Scientific	21127022
IgG1 Isotype Control	Cell Signaling	8527S
iScript™ Advanced cDNA Synthesis Kit	Bio-Rad	1725037
LightCycler® 480 SYBR Green I Master	Roche	04707516001
Lipofectamine™ RNAiMAX Transfection Reagent	Thermo Fisher Scientific	13778100
MassRuler DNA Loading Dye	Thermo Fisher Scientific	R0621
NucBlue Live ReadyProbes™	Thermo Fisher Scientific	R37605
Penicillin-Streptomycin	Thermo Fisher Scientific	15140122
PrestoBlue™ Cell Viability Reagent	Thermo Fisher Scientific	A13261
Protein Block, Serum-Free	Agilent	X090930-2

RPMI 1640 Medium	Thermo Fisher Scientific	21875034
Saponin	Sigma	47036-50G-F
siRNA for ANKRD30A (s40567)	Thermo Fisher Scientific	4392420
siRNA Negative Control (NO.1, Silencer™ Select)	Thermo Fisher Scientific	4390843
siRNA Positive Control (GAPDH, Silencer™ Select)	Thermo Fisher Scientific	4390849
TrackIt™ 100 bp DNA Ladder	Thermo Fisher Scientific	10488058
Wash Buffer 10x	Agilent	S3006

Table 1. List of chemicals.

2.1.2 Consumables

Consumable	Source	Catalog Number
10/20 µl XL Graduated TipOne Filter Tip	TipOne®, StarLab	S1120-3810
100 µl UltraPoint® Graduated Tip	TipOne®, StarLab	S1120-1840
1000 µl TipOne® Filter Tip	TipOne®, StarLab	S1122-1830
12 Well Cell Culture Plate	CELLSTAR®, Greiner Bio-One	665180
200 µl Graduated TipOne® Filter Tip	TipOne®, StarLab	S1120-8810
24 Well Cell Culture Plate	CELLSTAR®, Greiner Bio-One	662160
48 Well Cell Culture Plate	CELLSTAR®, Greiner Bio-One	677180
6 Well Cell Culture Plate	CELLSTAR®, Greiner Bio-One	657160
Cell Culture Flask (T175)	CELLSTAR®, Greiner Bio-One	660160
Cell Culture Flask (T25)	CELLSTAR®, Greiner Bio-One	690160
Cell Culture Flask (T75)	CELLSTAR®, Greiner Bio-One	658170
Centrifuge Tube 15 mL	CELLSTAR®, Greiner Bio-One	188271
Centrifuge Tube 50 mL	CELLSTAR®, Greiner Bio-One	227261
Eppendorf PCR Tubes 0.1 mL	Eppendorf	0030124804
Eppendorf PCR Tubes 0.2 mL	Eppendorf	0030124707
Eppendorf Reference® - Mechanical Pipette	Eppendorf	2231302001
Eppendorf Safe-Lock Tubes 1.5 mL	Eppendorf	022363212
FrameStar® 480/96 qPCR Adhesive Seal	4titude®	4ti-0952-SBC
LightCycler® 480 Multiwell Plate 96	Roche	04729692001
Stripette® Serological Pipettes 10 mL	Corning® Costar®, Sigma	CLS4100
Stripette® Serological Pipettes 25 mL	Corning® Costar®, Sigma	CLS4250

Stripette® Serological Pipettes 5 mL	Corning® Costar®, Sigma	CLS4050

Table 2. List of consumables.

2.1.3 Devices

Device	Source	Catalog Number
-86 °C Ultra-low Temperature Freezers	Thermo Fisher Scientific	Forma™ 900 Series
ChemiDoc XRS+ Imaging System	Bio-Rad	1708265
CO2 Incubator	Thermo Fisher Scientific	HERACELL 150i
Digital Lab Scale Balance	DeltaRange® METTLER	PM480
Digital Lab Scale Balance	Toledo, METTLER	AG204
Electrophoresis Power Supply	Bio-Rad	PowerPac 300
Eppendorf	Eppendorf	5810 R
Fluorescence Microscope	Carl Zeiss	Axiolan 2 Imaging
Gel Electrophoresis System	Gibco, Life Technologies	1068BD
LightCycler® 480 Real-time PCR System	Roche	LightCycler® 480 II
Megafuge™ Centrifuge	Thermo Fisher Scientific	75004230
MiniStar Microcentrifuge	VWR	521-2844
Microplate Reader	TECAN	Spark®
Microscope	Olympus	CKX41
Microwave	BOSCH	HMT84M451B
NanoDrop™ 2000 Spectrophotometer	Thermo Fisher Scientific	ND-2000
Orbital Shaker	Köttermann	4010
Safety Cabinets	ScanLaf, LaboGene A/S	Mars
Shaking Water Baths	Kisker Biotech	GFL-1083
Thermal Cyclers	peqSTAR, VMR	2X Universal Gradient
Thermal Shake	VWR	89232-910
Vacuum Pump	Red Evac	PV 100
Vortex Mixer	VWR	VV3
Vortexer Stirrer Shaker	Heidolph Buchler	Reax-2000

Table 3. List of devices.

2.1.4 Cell Lines

Cell Line	ATCC ID	Disease
A549	CCL-185	Lung Cancer
BT-474	HTB-20	Breast Cancer
HCC-1500	CRL-2329	Breast Cancer
MCF-7	HTB-22	Breast Cancer
MDA-MB-231	HTB-26	Breast Cancer
MDA-MB-453	HTB-131	Breast Cancer
SK-BR-3	HTB-30	Breast Cancer

Table 4. List of cell lines.

2.1.5 Primers

Primer	Forward Sequence	Reverse Sequence
A174	GCGTGGCAAGAGTAACATCTAA	AGAGACTCCGAGAATCACAAGA
A315	CAAGAGCTCTGCAGTGTGAGATTG	CTGGTATTGGTGTTTCAGTGTGGC
A350	TCGAAGAGCAGCATAGGAAA	CAGAACTTTAAGCTGCCCACT
A488	GAGAGCAGATCAGATGTTCCCTTCA	TCACTTCTAACTCTTTCC-TATGCTGCTC
A697	GGCTAGCCTCACACCACTTTTACT	TGTTTCTTTTGCGGGACTCATA
A743	TTAGGGAAGAATTAGGAAGAATC	CATTTGACACTGTGTTTCACGTTG
A91	AACATGCACAAAGAGACCAACGT	TGTTTGTTCACATTATCTT-GTTCGTTT
A931	CAAAGCAGAGCCTCCCGAGAAG	CCTATGCTGCTCTTCGATTCTTCC
CDC25A	TTTGGACAGCAGCATTTCTGTG	AGCTACAGTGGGATGAACCAGC
ESR1	GGGAAGTATGGCTATGGAATCTG	TGGCTGGACACATATAGTCGTT
GAPDH	CCGGGAAACTGTGGCGTGATGG	AGGTGGAGGAG-TGGGTGTCGCTGTT
HER2	CCAGCCTTCGACAACCTCTATT	TGCCGTAGGTGTCCCTTTG
LINC00993	AGTGCGGGGCTCATCTAT	GCCCCATGTATTTTATGGCC
MCM2	AGAGGATCGTGGTACTGCTATGGC	TTATGGATGGCATAGGGCCTCAGA
MCM4	CCGAATCAACATGGAAACCT	AGTCCACCTGGCGAGTAGC
PLK1	ACGGGGCCCATGAGTGCTGCAGTGA	ACGACGCGTTTAGGAGGCCTT-GAGA

Table 5. List of primers. The primers for ANKRD30A were abbreviated as "A + the length of amplified PCR products," such as A91 (ANKRD30A-91bp). Moreover, "-F" and "-R" are the abbreviations for forward primers and reverse primers, respectively.

2.1.6 Bioinformatics Datasets and Tools

Dataset	Url
Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle
cBioPortal	https://www.cbioportal.org
CCLE (cBioPortal)	https://www.cbioportal.org/study/summary?id=ccle_broad_2019
Ensembl	https://www.ensembl.org
Expression Atlas	https://www.ebi.ac.uk/gxa/home
Gene Set Enrichment Analysis (GSEA)	https://www.gsea-msigdb.org/gsea/index.jsp
GEPIA	http://gepia.cancer-pku.cn/index.html
GEPIA2	http://gepia2.cancer-pku.cn
GTEx (E-MTAB-5214)	https://www.ebi.ac.uk/gxa/experiments/E-MTAB-5214
GTEx (GTEx Portal)	https://www.gtexportal.org/home/datasets
Illumina Body Map (E-MTAB-513)	https://www.ebi.ac.uk/gxa/experiments/E-MTAB-513
Kaplan-Meier Plotter	https://kmplot.com/analysis
METABRIC	https://www.cbioportal.org/study/summary?id=brca_metabric
RIKEN FANTOM5 (E-MTAB-3358)	https://www.ebi.ac.uk/gxa/experiments/E-MTAB-3358
ShinyGO	http://bioinformatics.sdstate.edu/go
TCGA	https://www.cancer.gov/tcga
TCGA-BRCA (Firehose Legacy)	https://www.cbioportal.org/study/summary?id=brca_tcga
TCGA-BRCA (GDC Xena Hub)	https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Breast%20Cancer%20(BRCA)
TCGA-BRCA (PanCancer Atlas)	https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018
The Human Protein Atlas	https://www.proteinatlas.org

Table 6. List of bioinformatics datasets and tools.

2.2 Methods

2.2.1 Identification of Breast-specific Genes

All 1,352 homo sapiens experiments in the expression atlas database [67] (https://www.ebi.ac.uk/gxa/experiments?species=homo%20sapiens) were screened for the search of potential breast-specific transcripts in human tissues. Only experiments that examined gene expression in healthy adult human tissues and including data in breast tissues were eligible. The GTEx (E-MTAB-5214), the illumina body map (E-MTAB-513), and the RIKEN FANTOM5 (E-MTAB-3358) were the only three projects that met the eligibility criteria, involving the expression of almost 49,311 transcripts across 88 human tissues. The normalized RNA-seq data of the three projects in TPM were integrated for further analysis. A potential breast tissue transcript was defined as follows: 1) its expression in non-breast tissues is less than 1 TPM; 2) its expression in breast tissues is 1 TPM or more; 3) its expression in breast tissues is at least 5-times greater than the highest in non-breast tissues. Data in fetal tissues and male-specific tissues were excluded. Details were listed as follows: 1) **GTEx**: GTEx: Brodmann (1909) area 24, Brodmann (1909) area 9, C1 segment of cervical spinal cord, EBV-transformed lymphocyte, transformed skin fibroblast, and testis; 2) **Illumina Body Map**: prostate and gland testis; 3) **RIKEN FANTOM5**: adult penis, adult prostate gland, adult testis, fetal colon, fetal diaphragm, fetal duodenum, fetal eye, fetal lung, fetal occipital lobe, fetal parietal lobe, fetal rectum, fetal skeletal muscle tissue, fetal small intestine, fetal spinal cord, fetal stomach, fetal temporal lobe, fetal throat, fetal thyroid gland, fetal tongue, fetal trachea, fetal umbilical cord, fetal uterus, and fetal zone of skin.

2.2.2 Gene Annotations

The basic annotations of candidate genes were obtained using the online ShinyGO [89] database (http://bioinformatics.sdstate.edu/go), mainly including gene types and chromosome locations. The chromosome distribution of identified breast-specific genes was plotted using the customized R scripts based on the RIdeogram [90] package. The chromosome locations of the 6 breast-specific genes on chromosome 10 were explored using the Ensembl [85] genome browser (https://www.ensembl.org/Homo_sapiens/Location/View?r=10:37101027-37410733).

2.2.3 Expression Correlation

The expression correlations of candidate genes were analyzed using the GEPIA [74] database (http://gepia.cancer-pku.cn/detail.php?clicktag=correlation). RNA-seq gene expression data from the TCGA-BRCA tumor, TCGA-BRCA normal, and GTEx-breast were selected to assess gene coexpression. Spearman's Correlation Coefficient was selected for evaluating the strength of correlation. [91] A correlation coefficient of 0.8 or more with a corresponding P-value of 0.05 or less indicates a strong correlation.

2.2.4 Expression in Normal Human Tissues and Blood Cells

The expression of ANKRD30A in normal solid human tissues and various types of blood cells were assessed using the Human Protein Atlas (HPA) [82] database (https://www.proteinatlas.org). The consensus RNA-seq data based on the normalized gene expression values (NX) from the HPA dataset, the GTEx [68] dataset, and the FANTOM5 [92] project were used to evaluate gene

expression in normal solid tissues (https://www.proteinatlas.org/ENSG00000148513-ANKRD30A/tissue). Similarly, the HPA RNA-seq data of 18 types of blood cells as well as total peripheral blood mononuclear cells (PBMC) was used for estimating gene expression in the blood (https://www.proteinatlas.org/ENSG00000148513-ANKRD30A/blood). 1 NX was considered as the cutoff for positive gene expression.

2.2.5 Expression in Cancer Tissues

The expression of ANKRD30A in cancer tissues was evaluated using the GEPIA2 [73] database based on the RNA-seq gene expression data from the TCGA (https://www.cancer.gov/tcga) and GTEx [68] dataset. Gene expression values were presented as transcripts per million (TPM), and 1 TPM was considered as the cutoff for predicting positive gene expression. The statistical difference between tumor tissues and corresponding non-paired normal tissues was auto-calculated by the GEPIA2 database. The abbreviations of involved cancer types are available at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations.

2.2.6 Expression in Breast Cancer Tissues

The expression of ANKRD30A in breast cancer tissues and its associations with breast cancer clinical characteristics were analyzed using TCGA-BRCA [72] dataset downloaded from the GDC Xena Hub [93] (https://xenabrowser.net/datapages/?cohort=GDC TCGA Breast Cancer (BRCA)). ANKRD30A expression was determined based on the TCGA-BRCA RNA-seq gene expression data (log2 TPM transformed based on "HTSeq - Counts"), while the corresponding clinical features were extracted from the phenotype files ("pam50 subtype," "phenotype," and "survival data"). Samples were split into different groups by subtypes, T stages, N stages, M stages, TNM stages, menstrual status, ethnicities, tumor histologic types, sex, and survival status. Expression differences between different cohorts was evaluated by Student's t-test, and a P-value less than 0.05 indicates a significant difference.

2.2.7 Genetic Alteration

Genetic alterations of ANKRD30A in breast cancer was analyzed using the cBioPortal [76] (https://www.cbioportal.org), including mutations and copy number variations. A total of 12 datasets containing 6,771 breast cancer samples were combined for analysis, including Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016), Breast Cancer (MSK, Cancer Cell 2018), Breast Cancer (MSK, Nature Cancer 2020), Breast Cancer (MSKCC, NPJ Breast Cancer 2019), Breast Cancer (SMC 2018), Breast Cancer Xenografts (British Columbia, Nature 2015), Breast Invasive Carcinoma (British Columbia, Nature 2012), Breast Invasive Carcinoma (Broad, Nature 2012), Breast Invasive Carcinoma (Sanger, Nature 2012), Metastatic Breast Cancer (INSERM, PLoS Med 2016), The Metastatic Breast Cancer Project (Provisional, February 2020), and Breast Invasive Carcinoma (TCGA, Firehose Legacy). Samples were split into primary tumors and metastatic tumors based on the annotations from the "Sample Type," involving 4,742 primary breast cancer samples and 1,264 metastatic samples. Mutation data were obtained by querying the "ANKRD30A" and then clicking the "Mutations" tab. Similarly, data regarding copy number variation were retrieved by clicking the "Plots" tab and then selecting the "Copy Number" item. According to the definitions from cBioPortal (https://docs.cbioportal.org/1.-general/faq), only deep deletion and amplification events were considered biologically relevant, whereas copy

number gain and shallow deletion were considered equivalent to diploid in our analysis. ANKRD30A mutations were documented in all the included samples, while its CNVs were available in 2,173 and 216 primary and metastatic breast tumors, respectively.

2.2.8 Expression in Breast Cancer Cell Lines (*In Silico* Data)

The expression of ANKRD30A in breast cancer cell lines was analyzed using the RNA-seq data of the Cancer Cell Line Encyclopedia (CCLE) dataset. [78,79] Specifically, the RNA-seq gene expression data named "CCLE_RNAseq_rsem_genes_tpm_20180929.txt.gz," the corresponding gene annotation file named "gencode.v19.genes.v7_model.patched_contigs.gtf.gz," and the cell line annotation file named "Cell_lines_annotations_20181226.txt" were downloaded from the CCLE official webpage (https://portals.broadinstitute.org/ccle/data). Breast-associated cell lines were extracted based on the suffixes of the corresponding sample IDs. 1 TPM was considered as the cutoff for positive gene expression. The subtype of involved breast cancer cell lines was determined according to the multiomically-confirmed annotations reported by Shari et. al. [94] The gene expression heatmap based on Z-scores was plotted using the cBioPortal (https://www.cbioportal.org/study/summary?id=ccle_broad_2019).

2.2.9 Cell Culture

The following cell lines were used in this study, including HCC-1500 (ATCC, CRL-2329), BT-474 (ATCC, HTB-20), A549 (ATCC, CCL-185), MDA-MB-231 (ATCC, HTB-26), MDA-MB-453 (ATCC, HTB-131), MCF-7 (ATCC, HTB-22), and SK-BR-3 (ATCC, HTB-30). Involved cell lines were cultured with recommended growth medium and incubated at 37°C in a humidified atmosphere with 5% CO2. Specifically, the complete growth medium containing 500 mL of RPMI-1640 (Thermo Fisher Scientific, 21875034), 50 mL fetal bovine serum (Thermo Fisher Scientific, 10270106), and 5 mL of penicillin-streptomycin (Thermo Fisher Scientific, 15140122) were used for culturing the MDA-MB-231, MCF-7, HCC-1500, and SK-BR-3 cell lines. Similarly, 500 mL of DMEM (Thermo Fisher Scientific, 41965062) supplemented with the same amount of fetal bovine serum and penicillin-streptomycin were used for culturing the MDA-MB-453 and BT-474 cells, while the A549 cells were cultured with F12K (Thermo Fisher Scientific, 21127022) added with the same supplements. All cells were digested with the Accutase Cell Dissociation Reagent (Thermo Fisher Scientific, A1110501). In addition, cells were cryopreserved with recommended freezing medium containing DMSO (Thermo Fisher Scientific, D12345) in liquid nitrogen.

2.2.10 RNA Isolation

Total RNA of adherent cells was isolated with the Aurum™ Total RNA Mini Kit (Bio-Rad, 7326820) according to the manufacturer's recommendations. Isolated RNA samples were diluted with 50-100 mL of elution solution provided in the kit. Next, the concentration and quality of the RNA samples were assessed using the NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific, ND-2000) by measuring the light absorbance at 230 nm (A230), 260 nm (A260), and 280 nm (A280). Typically, an RNA sample with an A260/A280 ratio of 1.8-2.2 and an A260/A230 ratio of 1.8-2.4 was considered to be qualified and suitable for downstream applications. The isolated total RNA was stored at -80°C for further use.

2.2.11 cDNA Synthesis

cDNA was synthesized using the iScript™ Advanced cDNA Synthesis Kit (Bio-Rad, 1725037) following the user manual. Briefly, for each reaction, a total of 20 µL mixed solution was prepared in a PCR-grade EP tube, including 4 µL of 5x iScript Advanced Reaction Mix, 1 µL of iScript Advanced Reverse Transcriptase, 13 µL of nuclease-free water, and 2 µL of RNA template (250 ng/µL). Next, the complete reaction mix was incubated using a thermal cycler according to the default protocol: 1) 46℃ for 20 minutes for reverse transcription; 2) 95℃ for 1 minute for terminating the reaction. Synthesized cDNA samples were stored at -20℃ for further use.

2.2.12 qRT-PCR

qRT-PCR was performed with LightCycler® 480 SYBR Green I Master (Roche, 04707516001) on a LightCycler® 480 Real-Time PCR System (Roche) in a 20 µL reaction mixture. All real-time PCR reactions were performed following the MIQE guidelines.[95] Specifically, the complete reaction mix was prepared with 10 µL of 2x Master Mix, 4 µL of nuclease-free water, 2 µL of the forward primers (3000 nM), 2 µL of the reverse primers (3000 nM), and 2 µL of cDNA (37.5 ng/µL). All prepared reaction mixes were added to the corresponding wells in a 96-well plate, with each reaction was repeated in triplicate. All reactions were performed in a LightCycler® 480 system under the following condition: 1) pre-incubation at 95℃ for 5 minutes; 2) amplification for 45 cycles of 95℃ for 10 seconds, 60℃ for 20 seconds, and 72℃ for 30 seconds; 3) 1 cycle for melting curves under 95℃ for 5 seconds, 65℃ for 60 seconds, and 97℃ for 10 seconds; 4) cooling at 40℃ for 10 seconds. The 2-△△Ct method was used to evaluate the relative expression levels of targeted genes compared with calibrator genes (GAPDH).

2.2.13 Normal RT-PCR

Normal RT-PCR was performed with the DreamTaq PCR Master Mix (Thermo Scientific, K1071) based on the recommended protocol. The following components were added to a PCR tube for each 50 µL reaction: 25 µL of 2x DreamTaq PCR Master Mix, 13 µL of nuclease-free water, 2 µL of the forward primers (3000 nM), 2 µL of the reverse primers (3000 nM), and 2 µL of cDNA (250 ng/µL). The reactions were run in a thermal cycler under the following conditions: 1) 95℃ for 3 minutes; 2) 40 cycles of 95℃ for 30 seconds, 60℃ for 30 seconds, and 72℃ for 60 seconds; 3) 72℃ for 15 minutes.

2.2.14 Nucleic acid electrophoresis

The amplified normal RT-PCR products were assessed using nucleic acid electrophoresis. Specifically, the 2.0% agarose gel was prepared with agarose (Sigma, A9539-500G) and 1x TAE buffer. GelRed® (Sigma, SCT123) was used for staining nucleic acids. DNA samples were mixed with the MassRuler DNA Loading Dye (Thermo Scientific, R0621) before loading. The TrackIt™ 100 bp DNA Ladder (Invitrogen, 10488058) was used for distinguishing the length of amplified products. The products were separated under 110 v for 90 minutes in 1x TAE buffer. Lastly, the separated DNA fragments were visualized under UV light using the ChemiDoc XRS+ System (Bio-Rad, 1708265).

2.2.15 Immunofluorescence

Immunofluorescence assays were performed following the established protocol for adherent cells in our laboratory. Briefly, the main processes were described as follows: 1) seed cells in 6-well plates at a density of 5×10^4 cells per well and incubate the cells for 48 hours under normal conditions; 2) wash the cells with ice-cold DPBS; 3) fix the cells with 4% paraformaldehyde for 10 minutes; 4) permeabilize the cells with 0.5% Saponin (Sigma, 47036-50G-F) for 10 minutes; 5) stain the nucleus with the diluted (1:1000) DAPI (Thermo Scientific, 62248) solution for 10 minutes; 6) block unspecific protein binding sites using the ready-to-use Protein Block solution (Agilent, X0909) for 40 minutes; 7) incubate the cells with diluted primary antibodies overnight at 4°C; 8) incubate the cells with diluted secondary antibodies for 60 minutes if necessary; 9) mount the washed cells with the DaKo Fluorescence Mounting Medium (Agilent, S3023); 10) examine and capture fluorescent images using a fluorescence microscope (Carl Zeiss, Axiolan 2 Imaging).

2.2.16 siRNA Transfection

Cells were transfected with siRNAs using the Lipofectamine® RNAiMAX Transfection Reagent (Thermo Fisher Scientific, 13778100) according to recommended procedures. The following siRNAs were used, including ANKRD30A-siRNAs (Thermo Fisher Scientific, 4392420, s40567), GAPDH-siRNAs (positive control; Thermo Fisher Scientific, 4390849), and scramble siRNAs (negative control; Thermo Fisher Scientific, 4390843). Briefly, seeded cells in 96-well plates will be processed when a 60-80% cell confluency was observed. The transfection mixture was prepared in two 1.5 mL EP tubes, with one containing 25 µL of Opti-MEM® Medium and 1.5 µL of Lipofectamine® RNAiMAX Reagent, and the other containing 25 µL of Opti-MEM® Medium and corresponding 10 µM siRNAs in various volumes (0.5 µL - 1.5 µL). The components in the two EP tubes were then mixed in a 1:1 ratio and incubated for 5 minutes at room temperature. Next, 10 µL of the siRNA-lipid complex was added to each well in a 96-well plate and then mixed with 100 µL of complete growth medium. Lastly, the transfected cells were incubated under normal conditions before further analysis.

2.2.17 Cell Viability and Proliferation

Cell viability and proliferation were evaluated using the PrestoBlue™ Cell Viability Reagent (Thermo Fisher Scientific, A13261) based on the user manual. The main steps were described as follows: 1) mix the complete growth medium and the PrestoBlue™ reagent in a 9:1 ratio; 2) add 100 µL of the mixture to a well of a 96-well plate and incubated the cells under standard conditions for 3 hours; 3) measure the reagent absorbance at 570 nm and the reference absorbance at 600 nm; 4) using the normalized absorbance to assess cell viability. Since the PrestoBlue™ reagent is a non-toxic compound for live cells, the tested PrestoBlue™ solution in 96-well plates will be removed and replaced with the growth medium (and siRNAs when necessary) after each measurement. The measured cells were then incubated under normal conditions until the next measurement.

2.2.18 Colony Formation Assay

The plate colony formation assays were performed with adherent BT-474, HCC-1500, and MDA-MB-453 cells in 6-well plates. Briefly, harvested cells were firstly resuspended in 5 mL of DPBS. The suspension was then diluted to 10,000 cells/mL based on the results of cell counting, and 750 cells were then seeded onto per well of a 6-well plate. 3 mL of growth medium supplemented with necessary ingredients was added to each well, and ANKRD30A-siRNA was concurrently added to the corresponding wells in a concentration of 3x. Thereafter, cells were incubated under normal conditions until colonies of 50 cells or more were observed (21 days). During this period, the culture medium was changed every 7 days, and ANKRD30A-siRNA will also be refilled at the same time point. Next, the cells were washed with ice-cold DPBS and then fixed with 4% paraformaldehyde for 60 minutes. After this, the cells were stained with 5% crystal violet solution (Sigma Aldrich, V5265) for 20 minutes and then washed with DPBS for 3 times. Lastly, the colony number was counted under a microscope in triplicates.

2.2.19 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) [87] was performed with the GSEA software (http://www.gsea-msigdb.org/gsea/downloads.jsp) using the TCGA-BRCA and METABRIC datasets according to the official guidelines. Briefly, using the median ANKRD30A expression as the cutoff, samples were divided into the ANKRD30A-high and ANKRD30A-low groups. The expression files and phenotype labels were prepared as recommended (https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html?Preparing_Data_Files). The hallmark gene sets and the KEGG pathway gene sets were used for enrichment. The auto-calculated normalized enrichment score (NES), normalized P-values, and false discovery rate (FDR) q-values were used to evaluate the statistical significance of enrichment. The top 5 positively-enriched and top 5 negatively-enriched gene sets were further analyzed and plotted. Core enriched genes were identified based on the annotations in the details of enriched gene sets. The expression of ANKRD30A and its correlated GSEA-enriched genes was reviewed with the cBioPortal [76,96] database (https://www.cbioportal.org) using the TCGA-BRCA, METABRIC, and CCLE-BRCA datasets.

2.2.20 Prognosis Analysis

The prognostic significance of ANKRD30A for breast cancer was evaluated based on microarray gene expression and corresponding survival data using the Kaplan-Meier Plotter [88] database (https://kmplot.com/analysis/index.php?p=service&cancer=breast). Three commonly used survival outcomes in clinical practice were analyzed, including relapse-free survival (RFS), overall survival (OS), and distant-metastases-free survival (DMFS). Patients were split into the ANKRD30A-high and ANKRD30A-low cohort by the median ANKRD30A expression values. The hazard ratio (HR) was computed by comparing the hazard in the ANKRD30A-high cohort to that in the ANKRD30A-low cohort. An HR greater than 1 was considered to be associated with unfavorable prognostic outcomes, whereas an HR smaller than 1 suggested lower survival risks. The auto-calculated 95% confidence intervals and log-rank test P-values were used to determine statistical significance. Data from the overall analysis and all subgroup analyses were integrated and plotted as forest plots for better viewing.

2.2.21 Statistics

The statistical difference between two groups was determined using the t-test, where a P-value of 0.05 or less indicates a significant difference. Moreover, t-test P-values were symbolized under certain circumstances for a better reading experience, where "*" stands for $P<0.05$, "**" for $P<0.01$, "***" for $P<0.001$, "****" for $P<0.0001$, and "ns" for not significant. Gene co-expression correlation was evaluated using Spearman's correlation coefficient, where a correlation coefficient of 0.8 or more with a corresponding P-value of 0.05 or less indicated a strong correlation. The normalized enrichment score (NES), normalized P-values, and false discovery rate (FDR) q-values auto-calculated by the GSEA software were used to assess the statistical significance of enriched gene sets. The log-rank test P-values were used to compare the survival difference between two groups. Statistical data were processed using Prism GraphPad or Microsoft Excel or auto-calculated by the involved online server or software.

3 Results

3.1 Identification of breast-specific genes

To search for eligible projects for the screening of breast-specific genes, we screened a total of 1,293 homo sapiens experiments in the Expression Atlas. Only three RNA-seq projects met our criteria, including the GTEx (43,539 genes across 52 human tissues), the Illumina body map (49,311 genes across 16 human tissues), and the RIKEN FANTOM5 project (21,105 genes across 76 human tissues). The normalized gene expression data in adult women tissues of the three experiments were integrated for further analysis. Based on our definition, a total of 96 breast-specific genes were identified **(Figure 1)**, with 7/43,539 genes in the GTEx, 84/49,311 genes in the Illumina body map, and 7/21,105 genes in the RIKEN FANTOM5 project. Only two breast-specific genes, ANKRD30A and RN7SL314P, lied in the overlap between the GTEx and the Illumina Body Map project, whereas none of the genes in the RIKEN FANTOM5 overlapped the list of the other two projects. These data suggested that only quite a small proportion (about 0.19%) of genes in human genome are breast-specific.

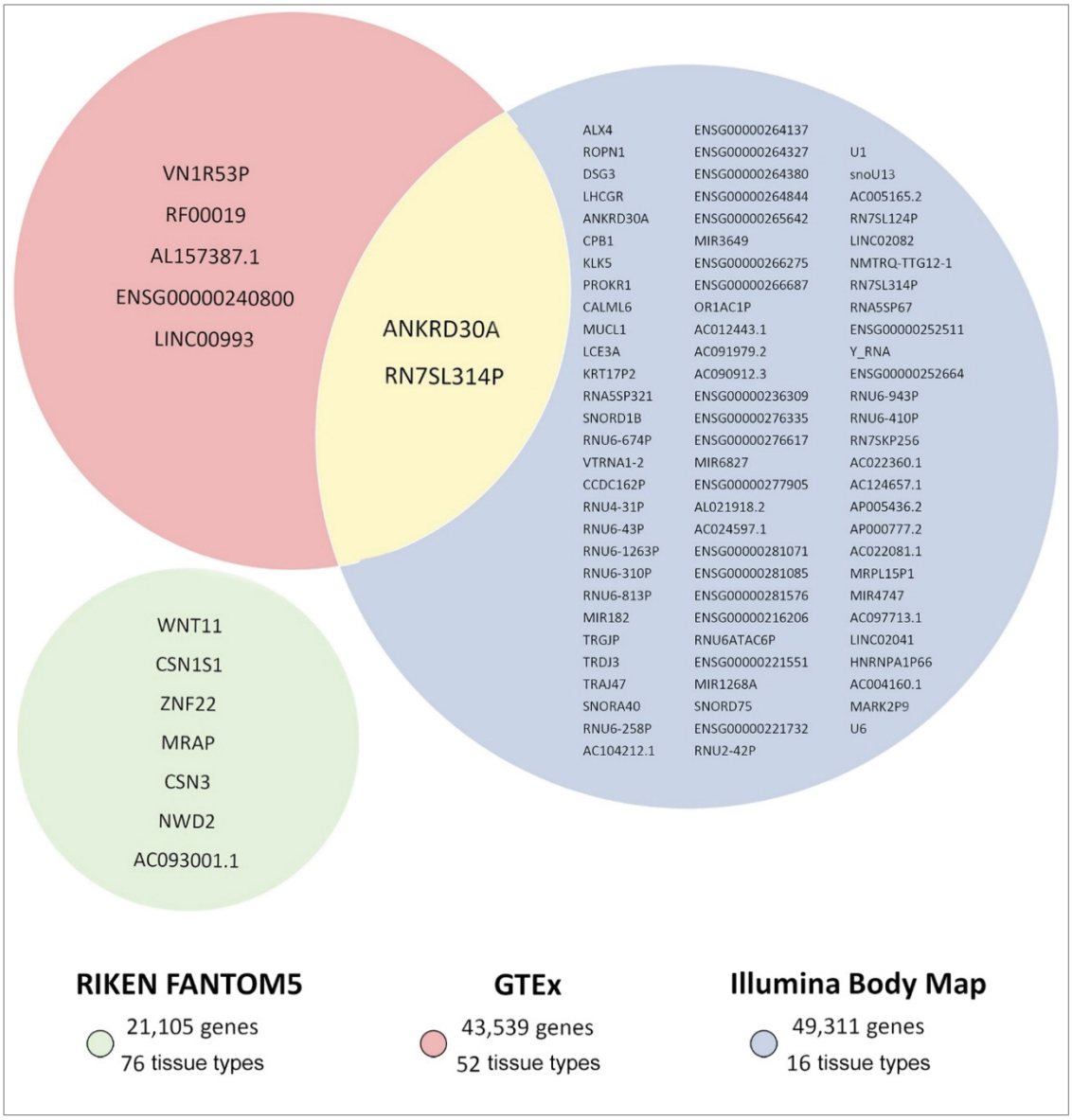
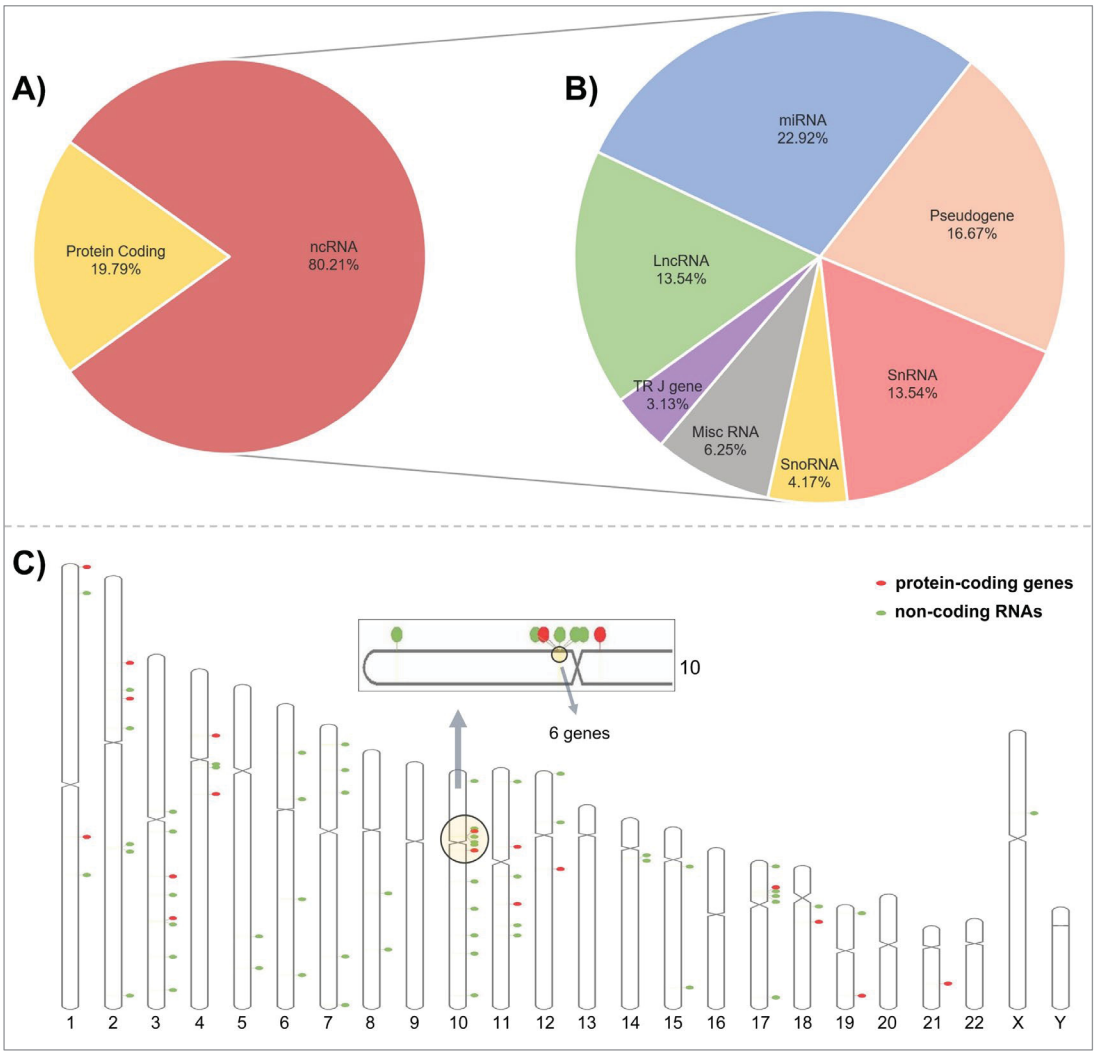


Figure 1. Breast-specific genes in the human genome. Genes with breast-specific expression patterns in human genome based on RNA-seq data from the RIKEN FANTOM5 (green), GTEx (red), and Illumina Body Map dataset (blue). The two shared breast-specific genes in both the GTEx and Illumina Body Map datasets were highlighted with a yellow circle.

3.2 Annotations of identified breast-specific genes

To get a better understanding of these breast-specific genes, we retrieved the ShinyGO database to get the annotations for their types and chromosome locations. Firstly, only 19/96 (19.79%) of these breast-specific genes are protein-coding genes, while 77/96 (80.21%) of them are non-coding RNAs (ncRNAs) (Figure 2A), including microRNAs (22/96, 22.92%), pseudogenes (16/96, 16.67%), lncRNAs (13/96, 13.54%), small nuclear RNAs (13/96, 13.54%), miscellaneous RNAs (6/96, 6.25%), small nucleolar RNAs (4/96, 4.17%), and joining chain T cell receptor gene (3/96, 3.13%) (Figure 2B). Secondly, the distribution of these genes on chromosomes is not uniform. Specifically, chromosome 10 had the highest number of breast-specific genes (14/96), while none of the identified genes were located on chromosome 9, 13, 16, 20, 22, and Y (Figure 2C). Notably, 6 out of the 14 breast-specific genes on chromosome 10 located in the same region in p11.21 (Figure 2D), including ANKRD30A, LINC00993, AL157387.1, RN7SL314P, Y-RNA, and VN1R53P. Among them, ANKRD30A is the only protein-coding gene and its sequence covers the full length of the other five genes.



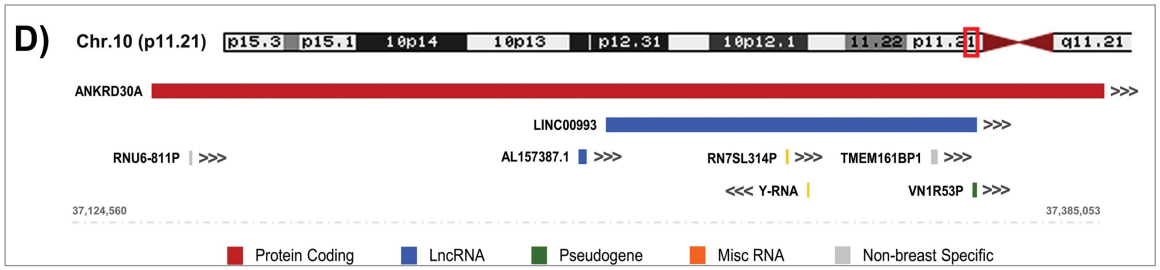


Figure 2. Types and chromosome locations of the 96 screened breast-specific genes. A) The primary types of the 96 genes. B) The subclassification of the 77 breast-specific ncRNAs. C) Chromosome locations of the screened breast-specific genes. The region marked with a semitransparent yellow circle on chromosome 10 was enlarged, where the focal distribution of the 6 breast-specific genes was also marked with a similar yellow circle. D) Chromosome locations and types of the 6 breast-specific genes (ANKRD30A, LINC00993, AL157387.1, RN7SL314P, Y-RNA, and VN1R53) on chromosome 10, in which the arrows represent their transcriptional directions.

3.3 Expression correlations of breast-specific genes on 10p11.21

Since genes located in the same cytoband are usually strongly correlated, especially at expression levels. [97–99] Therefore, we evaluated the expression correlation between ANKRD30A and its adjacent breast-specific genes using the gene expression profiling interactive analysis (GEPIA) database **(Figure 3)**. The gene Y-RNA was excluded in our further analysis as the term "Y-RNA" actually represents a class of non-coding RNAs other than a single gene. [100] According to the integrated RNA-seq data of the TCGA-BRCA and GTEx dataset, the expression of ANKRD30 was strongly correlated with its four adjacent breast-specific genes (LINC00993, RN7SL314P, VN1R53P, and AL157387) in both breast cancer tissues as well as normal breast tissues. Specifically, the Spearman's correlation coefficient between the ANKRD30A expression and LINC00993, RN7SL314P, VN1R53P, and AL157387 were 0.90 (P=0.0000), 0.92 (P=0.0000), 0.94 (P=0.0000), and 0.94 (P=0.0000), respectively.

Previously, we found LINC00993 acts like a tumor suppressor in breast cancer by inhibiting tumor proliferation and inducing cell apoptosis. [101,102] Combined with these previous findings, we predicted that ANKRD30A may have similar functions to LINC00993 in breast cancer and may regulate the expression of LINC00993 because 1) the two genes are always co-expressed in normal breast tissues and breast cancer; 2) the entire sequence of LINC00993 overlaps with the intron 36 of ANKRD30A. Meanwhile, the importance of RN7SL314P, VN1R53P, and AL157387 in breast cancer or other conditions remain unclear, as publications regarding the three genes were not available. Therefore, we narrowed down our focus to ANKRD30A in the following researches.

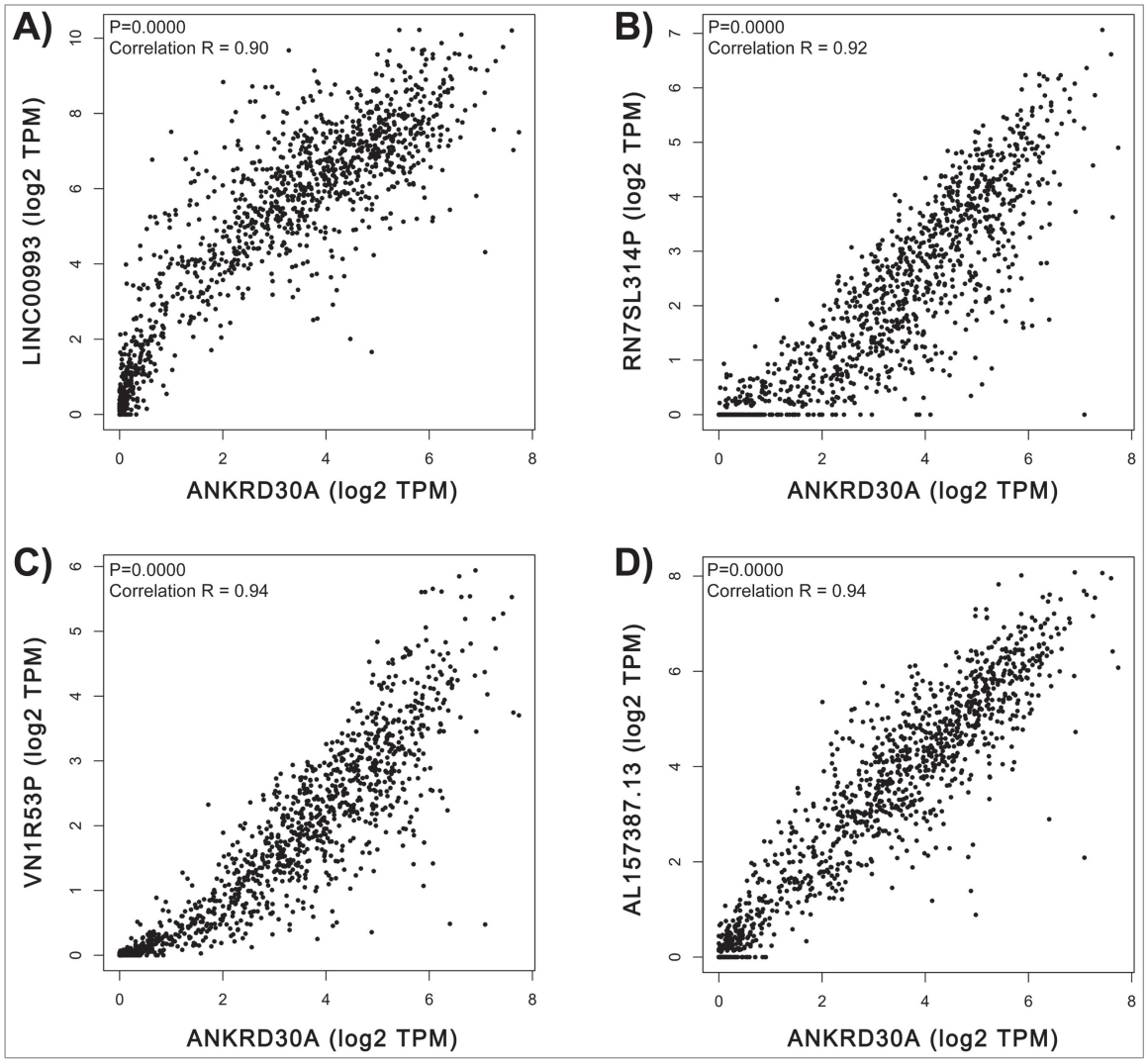


Figure 3. Expression correlations of breast-specific genes on 10p11.21. The mRNA expression correlation between ANKRD30A and A) LINC00993, B) RN7SL314P, C) VN1R53P, and D) AL157387 in normal breast tissues and breast cancers. Abbreviations: Correlation R, Spearman's Correlation Coefficient; TPM, Transcripts Per Million.

3.4 ANKRD30A expression in normal tissues and blood cells

Using the integrated RNA-seq gene expression data from the Human Protein Atlas database, we reviewed the expression of ANKRD30A in 16 integrated types of normal human tissues **(Figure 4A)** and 18 types of blood cells **(Figure 4B)**. In normal solid human tissues, the expression of ANKRD30A is mainly observed in male-specific testis tissues and breast tissues. Compared with testis tissues, the expression level of ANKRD30A in breast tissues is relatively higher, with the corresponding normalized expression values (NX) being 31.7 and 17.8, respectively. Meanwhile, despite ANKRD30A can be detected in some adipose tissues, its expression levels in these tissues are always below the positive cutoff (1 NX). Similarly, in the 18 types of normal blood cells, the expression of ANKRD30A is also always below the positive cutoff. Therefore, these data demonstrated that ANKRD30A is a protein-coding gene with breast-specific expression patterns in normal adult women.

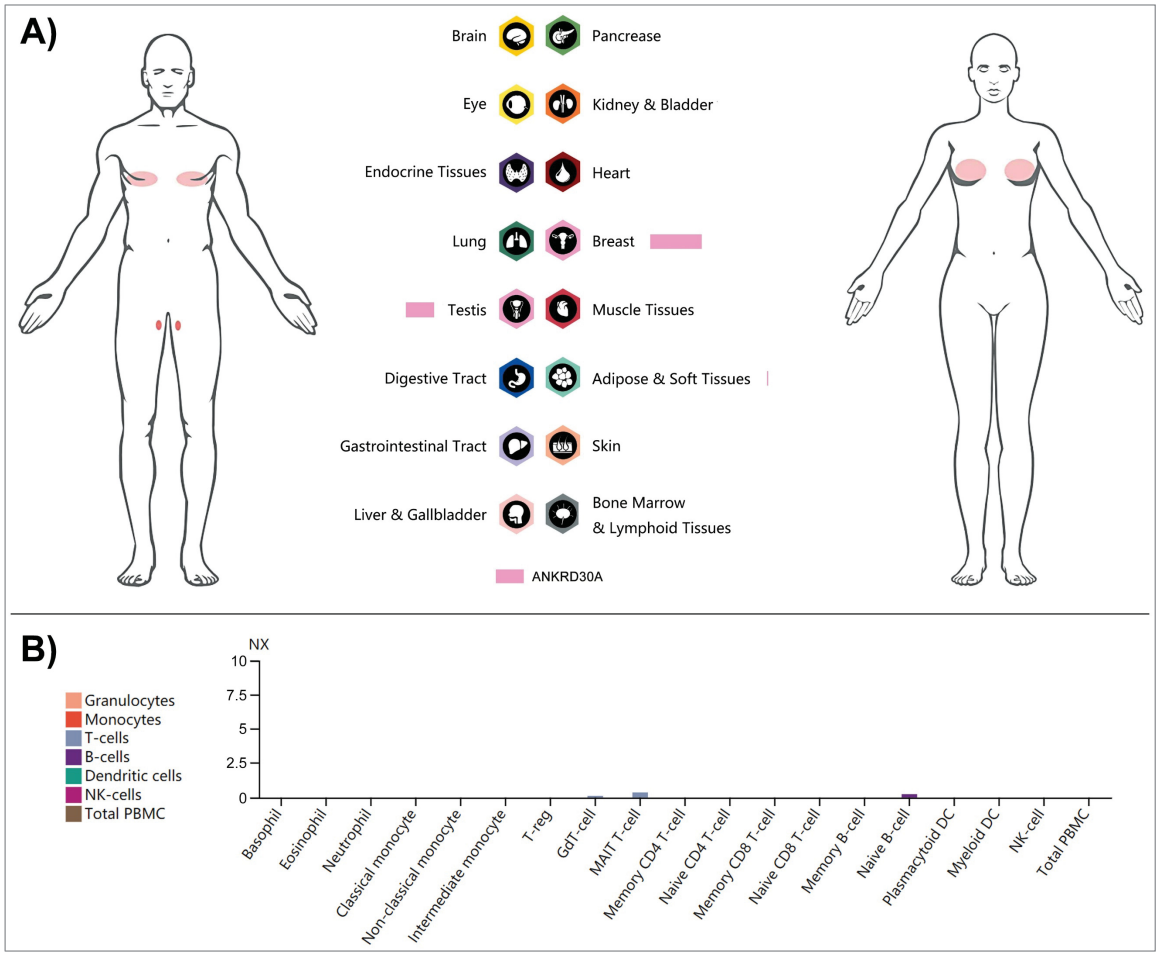


Figure 4. The expression of ANKRD30A in normal human tissues and blood cells. A) Normalized ANKRD30A mRNA expression in 16 integrated types of normal human solid tissues. B) Normalized ANKRD30A mRNA expression in 18 types of normal blood cells. See https://www.proteinatlas.org/ENSG00000148513-ANKRD30A/tissue for details. **Abbreviation**: NX, normalized expression.

3.5 ANKRD30A expression in cancers

To assess if ANKRD30A was also specific for breast cancer, we examined its expression in 33 types of cancer and corresponding normal tissues with the normalized RNA-seq data from the TCGA datasets and the GTEx dataset using the GEPIA database **(Figure 5)**. Interestingly, apart from some prostate adenocarcinoma tissues (PRAD) and normal testis tissues in males, the expression of ANKRD30A can only be detected in breast cancer (BRCA) and corresponding normal tissues. In non-breast tissues, cancer or normal, the expression levels ANKRD30A were always below the positive cutoff (1 transcript per million). In breast tissues, the median expression of ANKRD30A was 7.16 TPM in breast cancers and 5.37 TPM in normal breasts. These findings indicated that ANKRD30A is also specific for breast cancer in adult women.

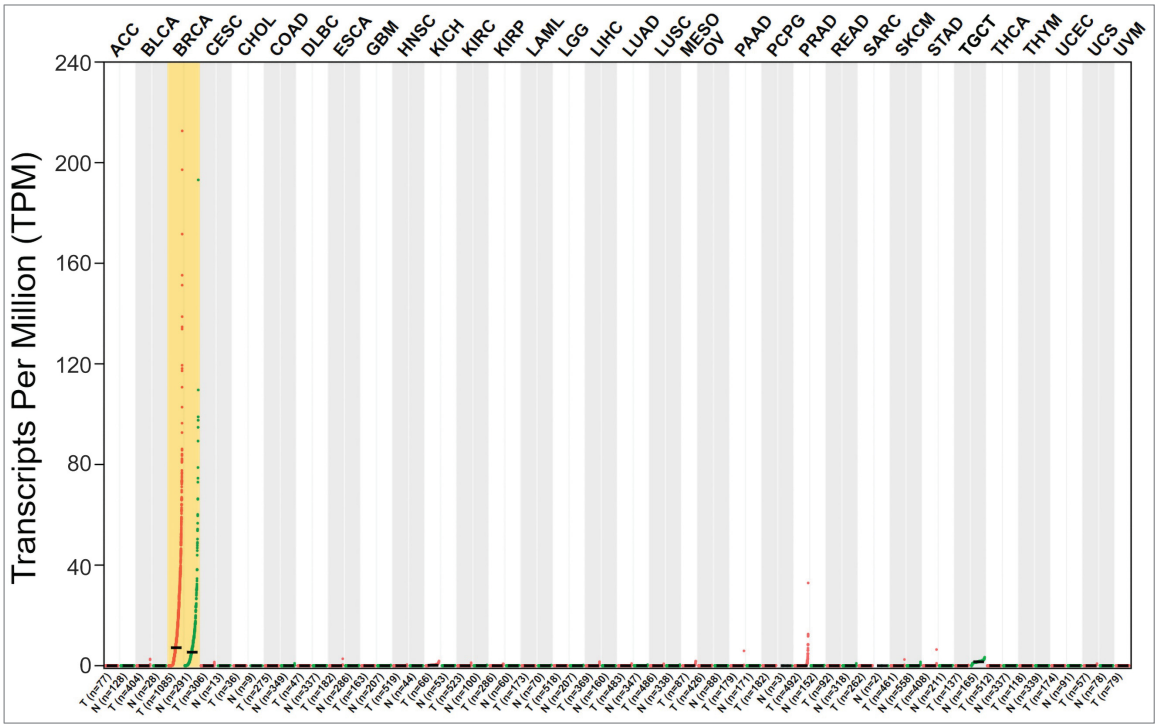


Figure 5. Expression of ANKRD30A in 33 types of cancer tissues and corresponding normal tissues. The expression of ANKRD30A in breast cancer (BRCA) was highlighted with a yellow background. The black dashes represent the median ANKRD30A expression values. It should be noted that samples from the same tissue type were not paired. **Abbreviations**: T, tumor; N, normal; see https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations for tissue type abbreviations.

3.6 ANKRD30A Expression in Breast Cancer Tissues

To further understand the association between ANKRD30A expression and breast cancer clinical characteristics, we analyzed the RNA-seq ANKRD30A expression data and the corresponding clinical features of 1,108 breast cancer tissue samples using the TCGA-BRCA dataset. Firstly, we noticed that the expression levels of ANKRD30A were significantly different in breast cancer subtypes **(Figure 6A)**. The median expression value of ANKRD30A was the highest in the Luminal A tumors, followed by the Luminal B tumors, and was similar in the HER2 subtype and the Normal-like subtype. Notably, its expression was the lowest and statistically down-regulated in the Basal subtype compared with any other subtype. Secondly, no significant difference was observed regarding ANKRD30A expression in breast cancer tumors of different T, N, M, and staging groups **(Figure 6, B-E)**. Thirdly, ANKRD30A expression was relatively lower in perimenopausal tumors compared with premenopausal and postmenopausal breast cancer tissues **(Figure 6F)**. Interestingly, compared with Asian and white patients, the expression of ANKRD30A was usually lower in black patients with breast cancer **(Figure 6G)**. For ANKRD30A expression in breast cancer tumors of different histologically types, no reliable conclusion can be drawn, as 71.4% (769/1,077) of the samples were infiltrating ductal carcinoma, while the ratio of the other 7 documented types was only 28.6% **(Figure 6H)**. Similarly, as 98.9% (1,075/1,087) of the TCGA-BRCA samples were from female patients, it still remains unknown whether ANKRD30A expression patterns are different between male and female breast cancer patients **(Figure 6I)**. Lastly, no obvious difference was observed regarding ANKRD30A

expression in breast cancer patients with different survival statuses **(Figure 6J)**. Briefly, the above findings suggested that the aberrant ANKRD30A expression in breast cancer was mainly associated with breast cancer subtypes, and its significant down-regulation in Basal tumors should be further investigated.

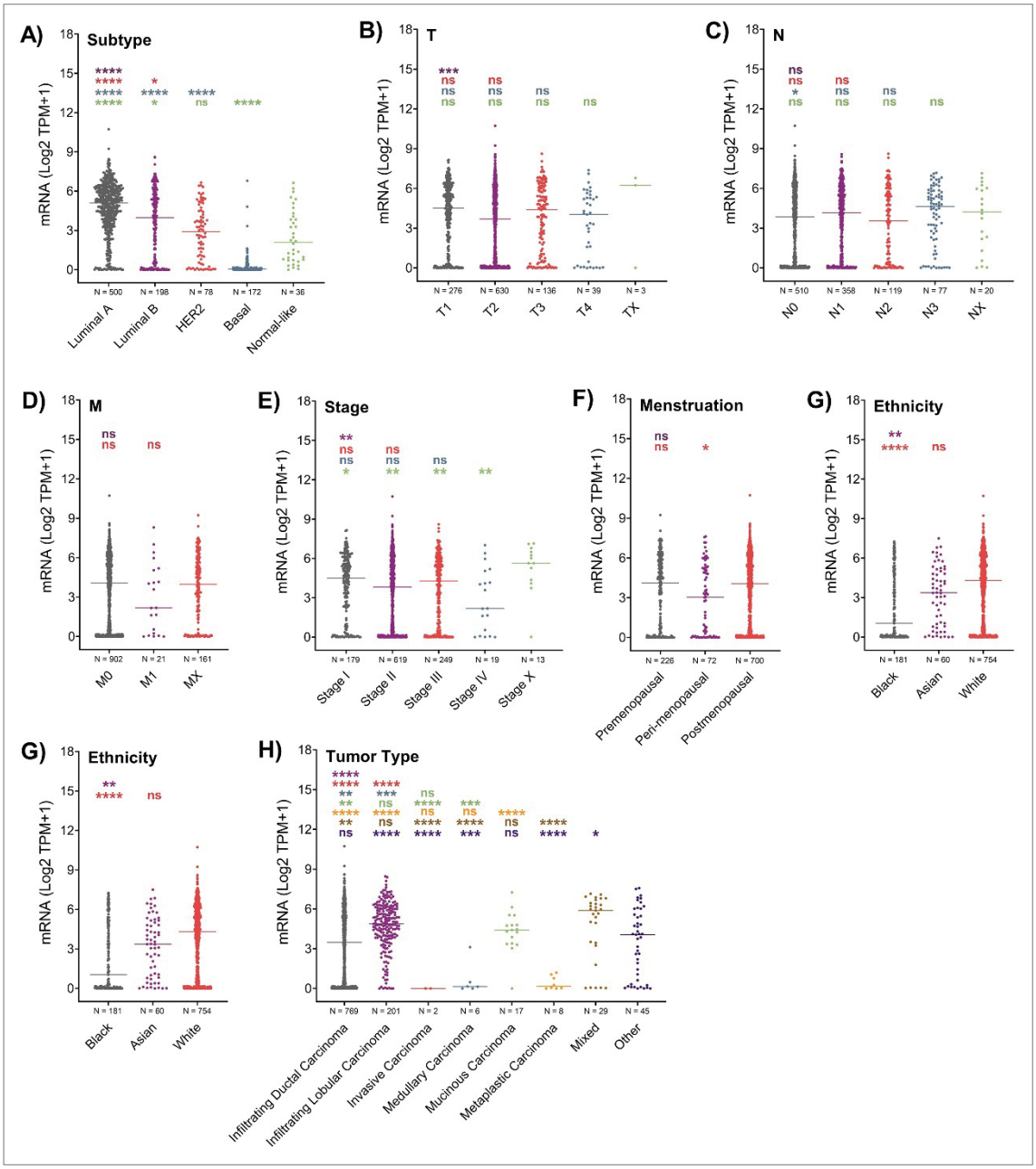


Figure 6. ANKRD30A expression and breast cancer clinical characteristics. The log2 ANKRD30A mRNA expression in breast cancer patients with different A) subtypes, B) T stages, C) N stages, D) M stages, E) TNM stages, F) menstrual status, G) ethnicities, H) tumor histologic types, I) sex, and J) survival status. The dashes represent the median expression of ANKRD30A in the corresponding cohort. The symbolized t-test P-values of ANKRD30A expression difference between different cohorts were presented as colorful asterisks. **Abbreviations**: *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001; ns, not significant.

3.7 Genetic Alterations of ANKRD30A in Breast Cancer

Genetic alteration events such as mutations and copy number variations (CNVs) often contribute to gene expression changes and cancer development. Therefore, we analyzed putative ANKRD30A mutations and CNVs in 4,745 primary and 1,264 metastatic breast cancer samples using the cBioPortal. In detail, only 9 (0.19%) and 3 (0.24%) ANKRD30A mutations were detected in these primary and metastatic tumors, respectively **(Figure 7, A-B)**. Moreover, all the detected mutations were non-recurrent as each type of mutation was observed only once. Similarly, ANKRD30A CNVs also can not be regarded as one of the major contributing factors in breast cancer. According to the definition of cBioPortal, only amplification and deep deletion events were considered biologically relevant. The data for ANKRD30A amplifications in primary and metastatic breast cancer are 38/2,173 (1.75%) and 4/216 (1.85%), respectively, while the figure for deep deletion is 0 in primary tumors and 1 in 216 metastatic breast cancer samples **(Figure 7, C-D)**. Taken together, mutations and CNVs of ANKRD30A are rarely observed in both primary and metastatic breast cancer. Hence, the two common types of genetic alterations should not be the focus of our further studies on ANKRD30A.

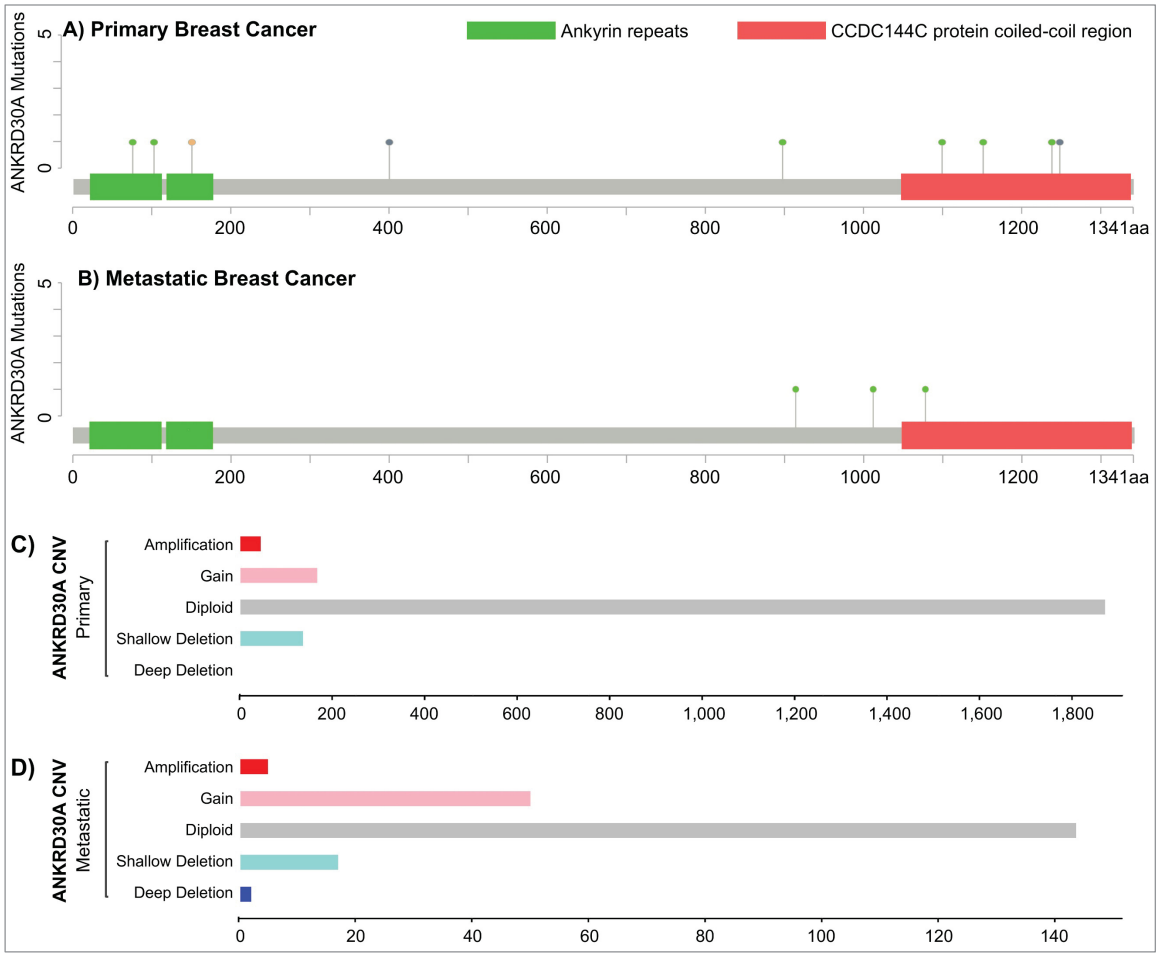


Figure 7. Putative ANKRD30A mutations and copy number variations in breast cancer. Mutations of ANKRD30A in A) primary and B) metastatic breast cancer tissues. The position of each mutation type on the sequence of ANKRD30A amino acids was labeled using vertical lines, and the height of these lines indicates the frequency of the corresponding mutation. Meanwhile, the special amino acid motifs of ANKRD30A were plotted in colored rectangles, where green represents the ankyrin repeats and red stands for the CCDC144C protein coiled-coil region. The count

of putative ANKRD30A copy number variations in C) primary and D) metastatic breast cancer. Only amplification and deep deletion events were considered biologically relevant. **Abbreviations**: CNV, copy number variation; aa, amino acid.

3.8 ANKRD30A Expression in Breast Cancer Cell Lines

Despite the mRNA and protein expression of ANKRD30A has been widely studied in previous publications, its expression in breast cancer cell lines remains unclear. [103–106] To find out the suitable cell models for investigating biological functions and molecular mechanism of ANKRD30A in breast cancer, we checked ANKRD30A expression using the RNA-seq data from the Cancer Cell Line Encyclopedia (CCLE). Of the 61 CCLE breast cancer cell lines, ANKRD30A expression data was available in 51 cell lines **(Table 1)**. Using 1 TPM as the cutoff for positive expression, we found that positive ANKRD30A expression was only observed in 13 breast cancer cell lines, including MDA-MB-134-VI, HCC1500, CAMA-1, ZR-75-30, MDA-MB-453, HCC1419, BT-474, BT-483, HCC202, MDA-MB-415, MDA-MB-361, and UACC-812. The highest expression of ANKRD30A was observed in the MDA-MB-134VI cell line (283.3 TPM), while the lowest positive expression was found in the UACC-812 cell lines (1.5 TPM). Moreover, similar to its expression patterns in breast cancer tissues, the cell lines that were annotated with positive ANKRD30A expression were either ER-positive or HER2-positive. Combined these findings with the availability of these cell lines, we selected three ANKRD30A-positive cell lines (HCC1500, MDA-MB-453, and BT-474) and three ANKRD30A-negative cell lines (MCF-7, MDA-MB-231, and SK-BR-3) for further researches **(Figure 8)**.

Cell Line	Subtype	ANKRD30A	ESR1	PGR	ERBB2
MDA-MB-134VI	ER+	283.3	94.21	2.51	22.3
HCC-1500	ER+	84.65	106.83	10.9	41.54
CAMA1	ER+	47.79	23.56	7.33	71.73
ZR-7530	HER2+	33.0	13.05	0.04	2809.5
MDA-MB-453	HER2+	9.69	0.01	0.02	265.84
HCC-1419	HER2+	9.59	13.57	0.01	3146.8
BT-483	ER+	4.24	32.79	0.6	284.52
BT-474	HER2+	4.04	18.18	51.5	1489.5
HCC-202	HER2+	3.98	0.22	0.02	2119.9
KPL1	NA	3.02	23.08	0.15	49.03
MDA-MB-415	ER+	2.62	8.94	0.56	57.83
MDA-MB-361	HER2+	1.84	18.04	1.43	810.09
UACC-812	HER2+	1.5	25.41	10.91	1990.8
EFM-192A	NA	0.4	5.52	0.08	2749.7
HCC-2218	HER2+	0.34	0.18	0.01	1841.6
MDA-MB-175VII	ER+	0.26	10.09	0.01	182.98
MDA-MB-231	TNBC	0.11	0.09	0.0	24.28
MCF7	ER+	0.08	41.35	4.47	44.39

UACC-893	HER2+	0.08	2.75	0.01	3238.6
HCC-1806	TNBC	0.07	1.3	0.01	31.0
HCC-1428	ER+	0.06	70.78	15.19	41.07
EFM-19	NA	0.05	44.35	15.61	152.58
CAL-851	NA	0.02	0.84	0.0	36.23
HCC-38	TNBC	0.01	0.03	0.69	139.16
T-47D	ER+	0.01	42.56	77.17	79.01
MDA-MB-436	TNBC	0.01	0.09	0.02	14.5
HCC-1395	TNBC	0.01	0.03	0.01	17.43
HCC-1187	TNBC	0.01	1.19	0.09	108.13
MDA-MB-157	TNBC	0.01	0.11	0.12	22.05
CAL-51	NA	0.01	0.33	0.0	80.61
HS578T	TNBC	0.0	0.05	0.0	25.52
HCC-1569	HER2+	0.0	0.05	0.0	2197.9
MDA-MB-468	TNBC	0.0	1.32	0.0	22.46
BT-549	TNBC	0.0	0.01	0.01	16.77
BT-20	TNBC	0.0	3.03	0.0	68.28
HMEL	NA	0.0	0.25	0.0	45.03
SKBR3	HER2+	0.0	0.25	0.0	1841.9
HCC-1143	NA	0.0	4.66	0.01	33.21
HCC-1937	TNBC	0.0	3.71	0.0	48.68
ZR-751	ER+	0.0	28.89	23.5	110.13
HCC-70	TNBC	0.0	1.77	0.02	41.9
HCC-1599	TNBC	0.0	2.06	0.03	41.39
HCC-2157	TNBC	0.0	0.01	0.0	91.57
HDQP1	NA	0.0	1.69	0.01	43.76
CAL-148	NA	0.0	0.03	0.01	201.89
DU-4475	TNBC	0.0	0.08	0.0	9.04
HCC-1954	HER2+	0.0	3.3	0.0	3836.4
HMC-18	NA	0.0	0.02	0.0	12.33
AU-565	HER2+	0.0	0.3	0.0	3138.0
CAL-120	NA	0.0	0.61	0.02	23.54
JIMT1	NA	0.0	2.09	0.0	355.49

Table 7. Expression of ANKRD30A in 51 CCLE breast cancer cell lines. Gene expression values were presented as TPM. The expression data of ESR1, PGR, and ERBB2 were also presented for distinguishing the subtype of corresponding cell lines. The six cell lines that were selected for further researches were highlighted with a yellow background. **Abbreviations**: ER+, ER-positive; HER2+, HER2-positive; TNBC, triple-negative breast cancer; NA, not available.

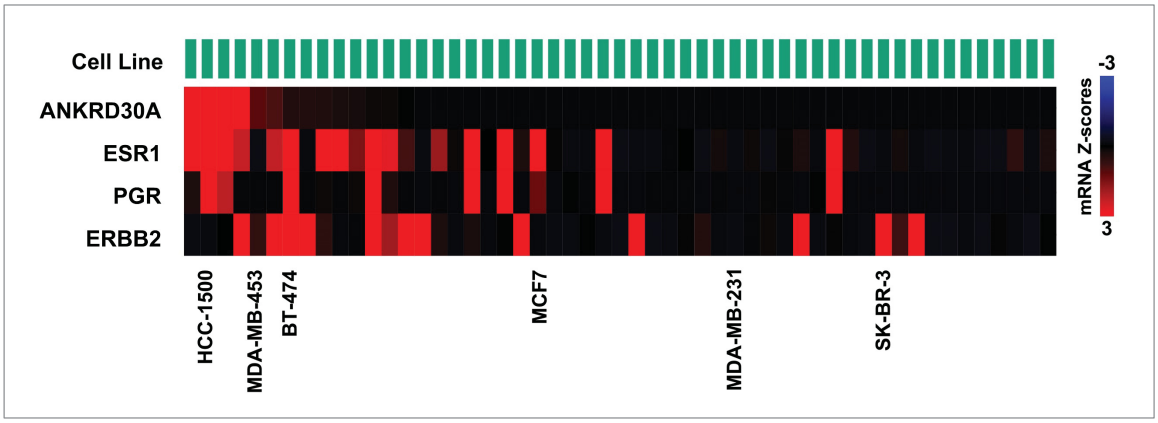


Figure 8. ANKRD30A expression in CCLE breast cancer cell lines. Gene expression was presented as Z-scores, where 3 represents the highest expression in included samples and -3 for the lowest. The expression of ESR1, PGR, and ERBB2 was plotted for distinguishing the subtype of corresponding cell lines. Gene expression in the six selected cell lines (HCC-1500, MDA-MB-453, BT-474, MCF7, MDA-MB-231, and SK-BR-3) was labeled on the X-axis.

3.9 Optimal Primers for ANKRD30A mRNA Detection

To validate our bioinformatics findings, we analyzed ANKRD30A mRNA expression in breast cancer cell lines using qRT-PCR assays. However, our preliminary experiments suggested that the sensitivity of the two empirical chosen primers was too low for ANKRD30A detection. Therefore, we reviewed all previous publications involving ANKRD30A mRNA detection by normal RT-PCR or qRT-PCR and collected all experimentally validated primers. A total of 8 pairs of primers were included, and these primers were named based on the length of the corresponding amplified products, such as A-91 (ANKRD30A-91bp). In addition, the A549 cell line was selected as the positive control as the positive expression of ANKRD30A has been validated in this cell line by Human Protein Atlas (https://www.proteinatlas.org/ENSG00000148513-ANKRD30A/cell).

Next, we performed the qRT-PCR assays to evaluate the sensitivity and specificity of the 8 primers under the same experimental conditions **(Figure 9, A-E)**. Firstly, we checked the melting curves to evaluate the specificity of the amplified qRT-PCR products. ANKRD30A mRNA can be successfully amplified using the primer pairs A-91, A-174, A-315, A-350, and A-488, while no specific melting curves were observed when ANKRD30A was amplified using the primer A-697, A-743, and A-931. Secondly, we found the Cq values are the lowest when ANKRD30A was amplified with the primer A-91, and acceptable Cq values (less than 35) were only observed in the A-91, A-174, A-315, and A-350 groups. Thirdly, we noticed that the Cq values seem to be positively correlated with the length of amplified PCR products. The longer the products are, the lower the Cq values. According to the qRT-PCR results, the primer A-91, A-174, A-315, and A-350 were selected for further evaluation by normal RT-PCR in the A549, BT-474, HCC-1500, and MDA-MB-453 cell lines **(Figure 9F)**. Despite all amplified products are specific, the sensitivity of these primers is quite different. A-315 was identified as the most sensitive primer pair, and the sensitivity of these primers is not correlated with the length of products. In addition, it has to be mentioned that both A-91 (exon 34) and A-315 (exon 32-34) are all targeting exon 34 of ANKRD30A. In summary, under our laboratory conditions, the primer pairs A-91 and A-315

were considered as the optimal for ANKRD30A mRNA expression using qRT-PCR and normal RT-PCR, respectively.

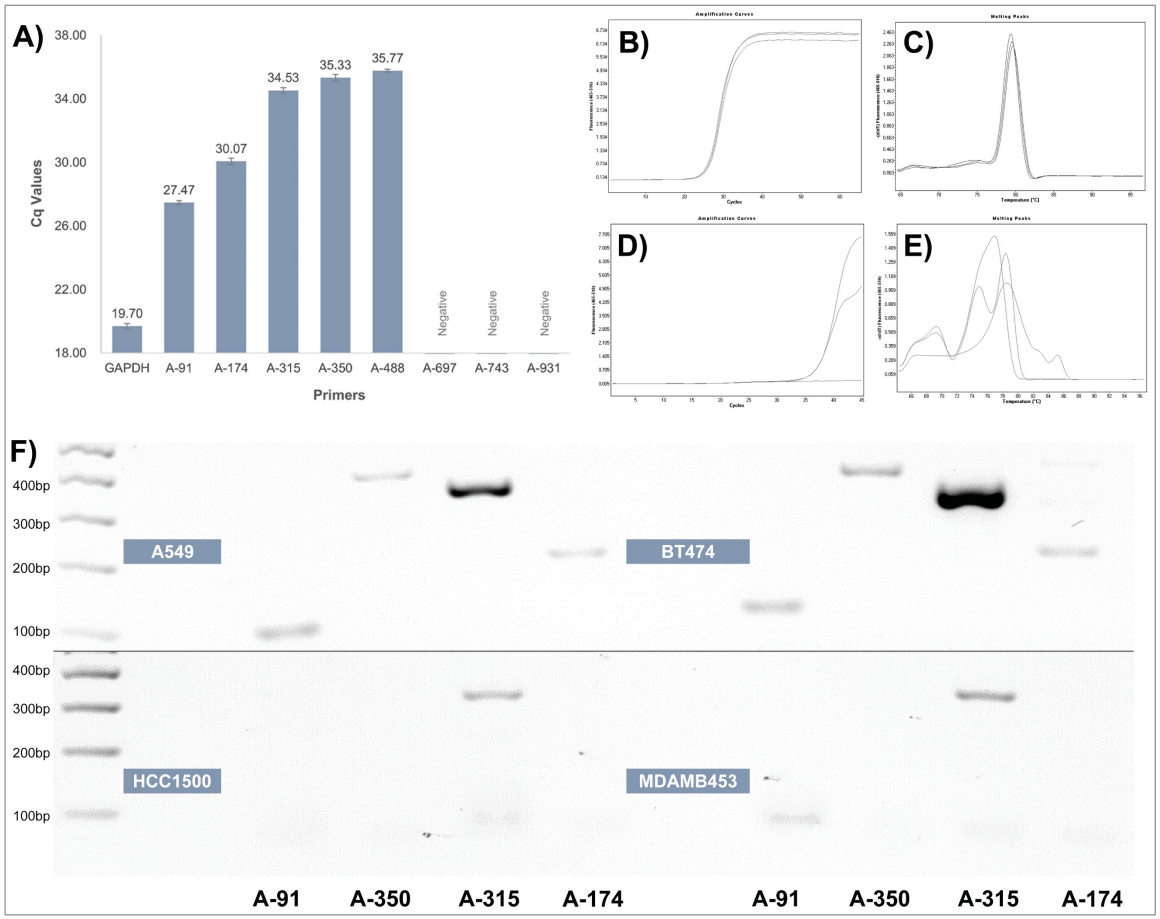


Figure 9. Optimal Primers for ANKRD30A mRNA Detection. A) qRT-PCR Cq values of the 8 ANKRD30A primers in A549 cells. B) Representative amplification curves and C) melting curves in successful qRT-PCR reactions performed with eligible ANKRD30A primers (A-91, A-174, A-315, and A-350). D) Representative amplification curves and E) melting curves in unsuccessful qRT-PCR reactions performed with ineligible ANKRD30A primers (A-488, A-697, A-743, and A-931). F) Agarose gel nucleic acid electrophoresis of normal RT-PCR products amplified with A-91, A-174, A-315, and A-350 in A549, BT-474, HCC-1500, and MDA-MB-453 cells.

3.10 ANKRD30A Expression Validation in Breast Cancer Cell Lines

After screening the optimal primers for detecting ANKRD30A mRNA, we then validated ANKRD30A mRNA expression in the six selected breast cancer cell lines **(Figure 10)**. In the three cell lines (HCC-1500, MDA-MB-453, and BT-474) in which the RNA-seq ANKRD30A expression values are more than 1 TPM, ANKRD30A can be effectively and specifically amplified in both the normal RT-PCR and qRT-PCR assays using the screened primers. Meanwhile, in the three cell lines (MCF-7, MDA-MB-231, and SK-BR-3) in which the RNA-seq ANKRD30A expression value is 0 TPM, no specific amplified PCR products were observed in both the normal RT-PCR and qRT-PCR assays. Moreover, the expression level of ANKRD30A analyzed by RT-PCR is not completely consistent with the CCLE-BRCA RNA-seq data. In the HCC-1500 cell line, the RNA-seq ANKRD30A mRNA expression value is the highest (84.65 TPM), followed

by the MDA-MB-453 cells (9.69 TPM), with the BT-474 cells (4.04 TPM) being the lowest. However, in either normal RT-PCR or qRT-PCR, the highest ANKRD30A expression was all observed in the BT-474 cells, while its expression was the lowest in the HCC-1500 cells. Briefly, data from this section suggested that it is feasible to use bioinformatics RNA-seq data to semi-quantitatively predict gene expression (positive or negative prediction), and the breast cancer cell lines HCC-1500, MDA-MB-453, and BT-474 are suitable cell models for further experimental studies on ANKRD30A.

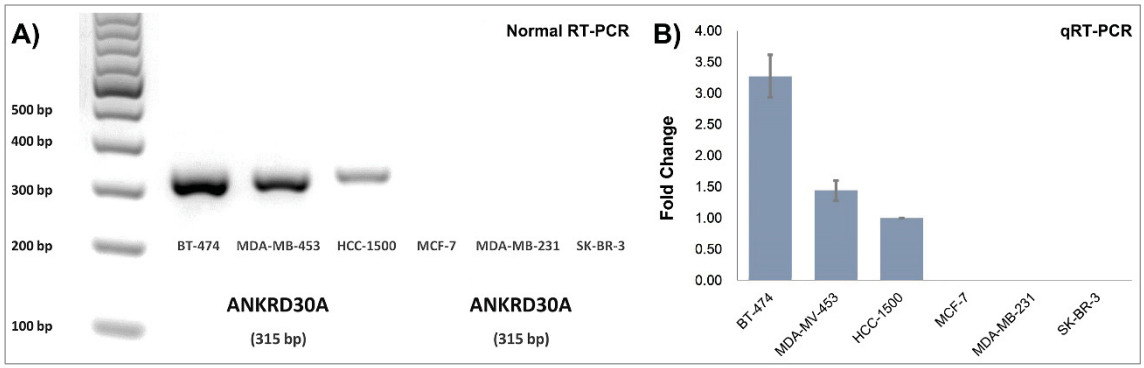


Figure 10. ANKRD30A expression validation in breast cancer cell lines. The expression of ANKRD30A mRNA was validated in the six selected breast cancer cell lines using A) normal RT-PCR and B) qRT-PCR with the screened primers. In qRT-PCR assays, the expression levels of ANKRD30A were presented as fold changes ± standard deviations by setting its relative expression in the HCC-1500 cells as 1.

3.11 ANKRD30A Subcellular Locations

Next, we analyzed the subcellular locations of ANKRD30A protein in breast cancer cell lines using immunofluorescence staining **(Figure 11)**. In the three ANKRD30A-mRNA-positive cell lines, the expression of ANKRD30A protein is also positive. Moreover, we noticed that the subcellular locations of ANKRD30A are different in these cell lines. In the HCC-1500 (ER+/HER2±) and BT-474 (ER+/HER2+) cell lines, the expression of ANKRD30A protein was mainly found in the cytoplasm, while no significant signals were observed in the nucleus. However, in the MDA-MB-453 (ER-/HER2+) cells, ANKRD30A protein can be detected in both the cytoplasm and perinuclear regions. Moreover, unlike its uniform distribution in the cytoplasm, the distribution of ANKRD30A protein in the perinucleus is in a focal pattern. These data suggested positive ANKRD30A mRNA expression usually translates into positive ANKRD30A protein expression, and the subcellular locations of ANKRD30A protein may be correlated with ER status in breast cancer.

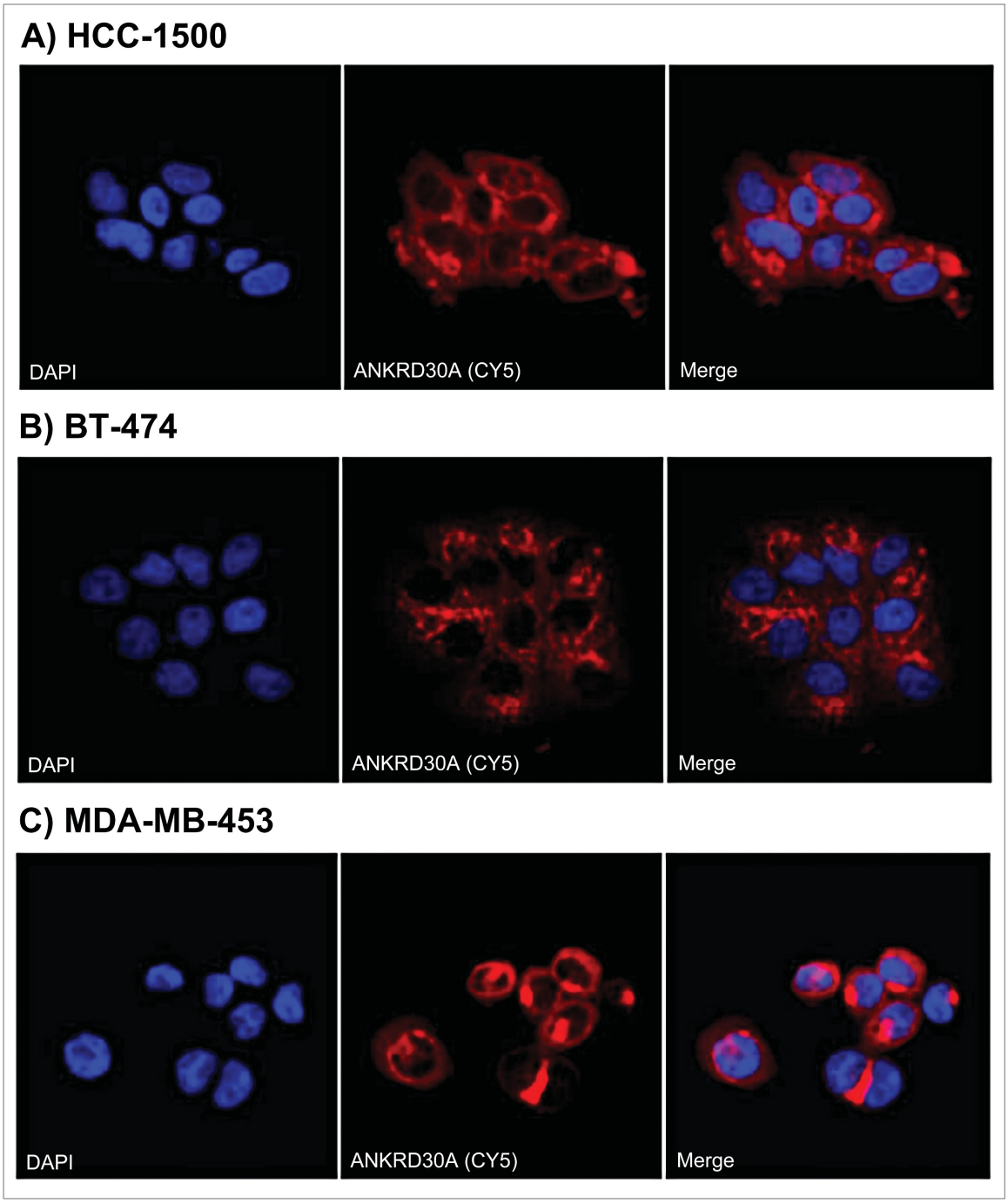


Figure 11. Subcellular locations of ANKRD30A protein in breast cancer cell lines. The expression of ANKRD30A protein was validated using immunofluorescence assays in the three ANKRD30A-mRNA-positive cell lines, namely the A) HCC-1500 (ER+/HER2±), B) BT-474 (ER+/HER2+), and C) MDA-MB-453 (ER-/HER2+) cell lines. The nucleus was stained with DAPI (blue), whereas ANKRD30A proteins were detected with Cy5-conjugated antibodies (red). The presented images were captured at 40x magnification.

3.12 Knockdown ANKRD30A using siRNA

To further investigate the potential biological functions of ANKRD30A in breast cancer, we constructed ANKRD30A-knockdown breast cancer cell lines using the siRNA method. Based on the annotations from the Ensembl database

(https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000148513), ANKRD30A is a protein-coding gene containing 4 protein-coding transcripts (ANKRD30A-201, ANKRD30A-202, ANKRD30A-204, and ANKRD30A-205) and 1 non-coding transcript (ANKRD30A-203) **(Figure 12)**. Notably, exon 34 is a shared region of all the four protein-coding transcripts. As previously mentioned, the screened primers are all targeting ANKRD30A exon 34. Therefore, to effectively knock down ANKRD30A, we selected a commercially available siRNA targeting its exon 34.

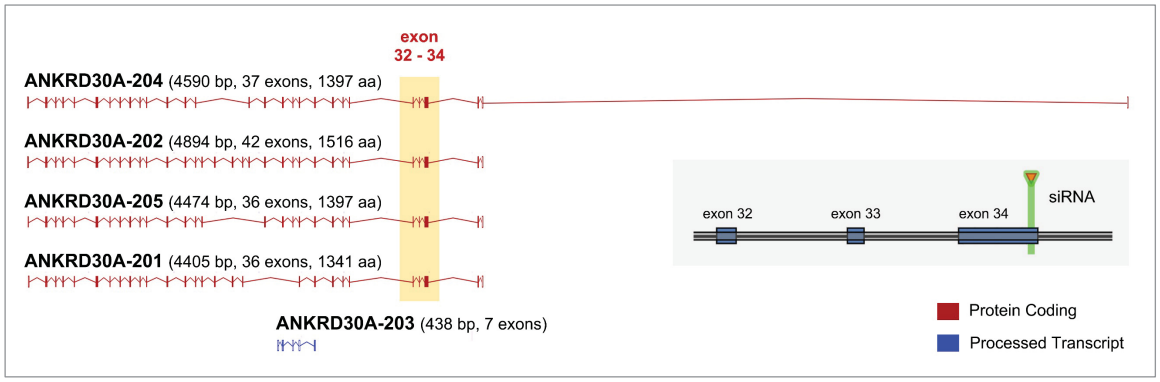


Figure 12. Transcripts and exons of ANKRD30A. The four protein-coding transcripts of ANKRD30A were plotted in red, including ANKRD30A-201, ANKRD30A-202, ANKRD30A-204, and ANKRD30A-205. The only non-coding ANKRD30A transcript ANKRD30A-203 was plotted in blue. The chromosome region of ANKRD30A exon 32-34 was highlighted with a yellow background, and the region was also enlarged for a better view. The target site of the selected siRNA was labeled with a green vertical line.

Next, we explored the feasibility and the optimal conditions for knocking down ANKRD30A using the siRNA method. Following the recommended workflow, we firstly incubated the BT-474 cells with ANKRD30A-siRNA, GAPDH-siRNA (positive control), and the scramble-siRNA (negative control) for 48 hours at various concentrations **(Figure 13, A-D)**. After this, the expression of ANKRD30A mRNA was evaluated using qRT-PCR assays. We found that the knockdown effect of ANKRD30A and GAPDH is stronger when the siRNA concentration is higher, and the most significant knockdown effect is observed in the 3x-default-concentration (3x) groups. Meanwhile, the expression of ANKRD30A and GAPDH is stable when cells were treated with the scramble-siRNA in a concentration of 0 to 3x. Moreover, the PrestoBlue assays suggested that up to 3x of siRNA has no significant toxicity to cell viability **(Figure 13E)**. Based on these data, we demonstrated that ANKRD30A can be effectively knocked down using the siRNA method, and the 3x siRNA concentration was considered optimal for further experiments.

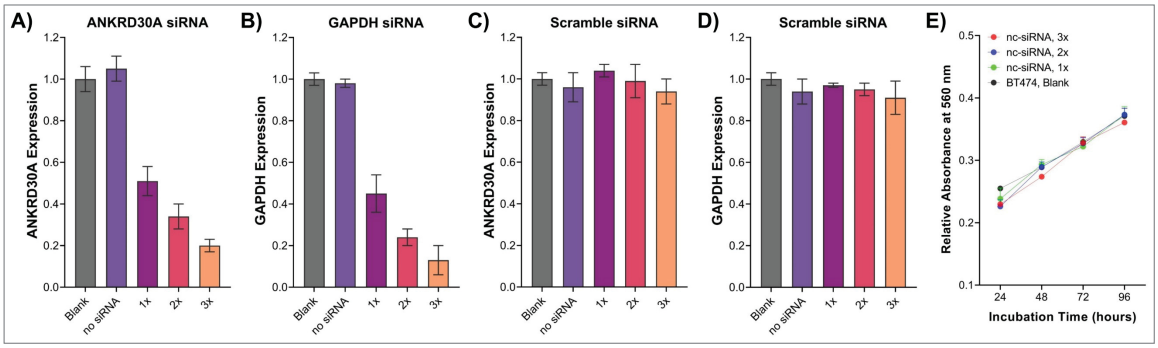


Figure 13. The optimal siRNA concentration for knocking down ANKRD30A. A) ANKRD30A expression in BT-474 cells incubated with ANKRD30A-siRNA for 48 hours at various concentrations. B) GAPDH expression in BT-474 cells incubated with GAPDH-siRNA for 48 hours at various concentrations. C) ANKRD30A expression in BT-474 cells incubated with scramble-siRNAs for 48 hours at various concentrations. D) GAPDH expression in BT-474 cells incubated with scramble siRNAs for 48 hours at various concentrations. E) The viability of BT-474 cells treated with different concentrations of scramble-siRNAs at various time points. **Abbreviations**: nc-siRNA, scramble-siRNA (negative control siRNAs).

After determining the optimal siRNA transfection conditions for ANKRD30A knockdown in breast cancer cells, we further explored the subcellular changes of ANKRD30A proteins using immunofluorescence **(Figure 14)**. Compared with the normal BT-474, HCC-1500, and MDA-MB-453 cells, the fluorescent signals of ANKRD30A proteins were significantly decreased in all of the three cell lines when cells were treated with ANKRD30A-siRNAs, which further confirmed the efficacy of ANKRD30A knockdown. Moreover, in the MDA-MB-453 cells where the focal perinuclear ANKRD30A expression can be observed under normal conditions, the fluorescent signals in the nucleus reduced more evidently than those in the cytoplasm. Therefore, these data suggested that nucleus-located ANKRD30A proteins may be more susceptible to knockdown effects.

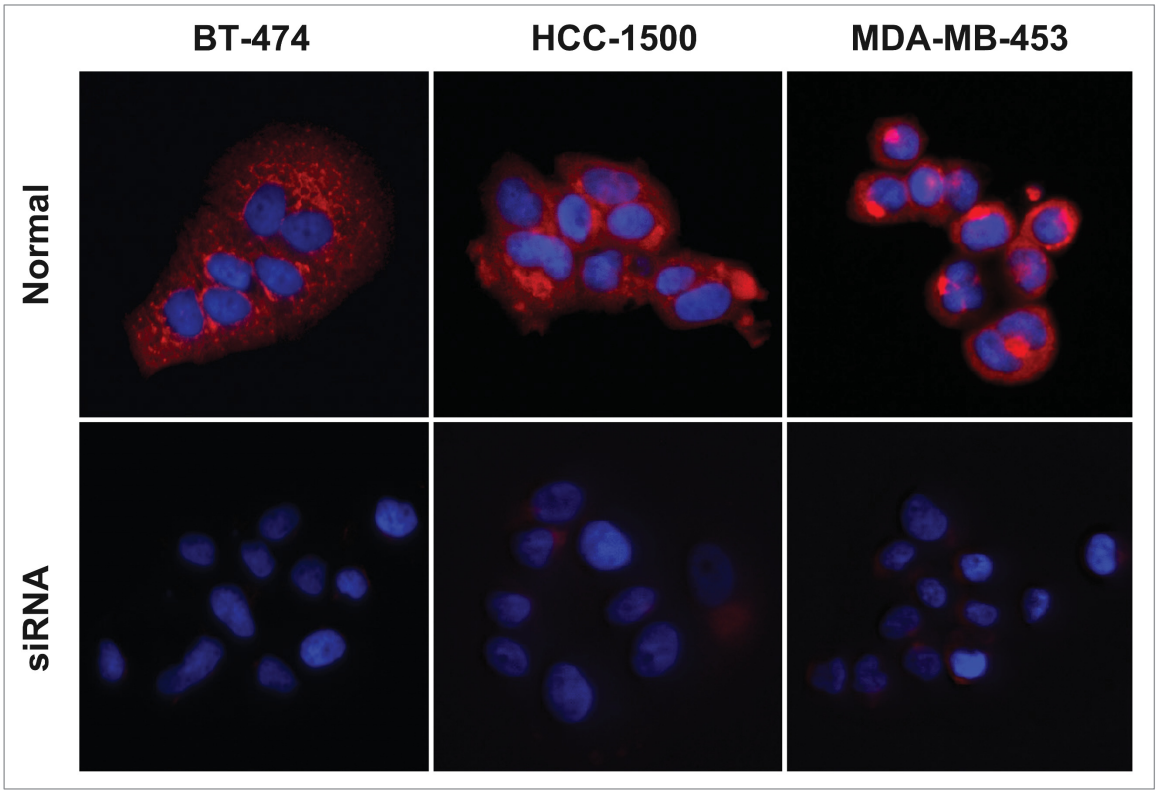


Figure 14. Immunofluorescence assays before and after silencing ANKRD30A in breast cancer cell lines. The merged fluorescent images in the HCC-1500 (ER+/HER2±), BT-474 (ER+/HER2+), and MDA-MB-453 (ER-/HER2+) cell lines were presented, where the nucleus was stained with DAPI (blue), and ANKRD30A proteins were detected with Cy5-conjugated antibodies (red). Cells in the siRNA group were incubated with 3x ANKRD30A-siRNA for 48 hours. The presented images were captured at 40x magnification.

3.13 ANKRD30A and Cell Proliferation

To explore the roles of ANKRD30A in cell proliferation, we performed cell viability assays with the PrestoBlue reagent using BT-474, HCC-1500, and MDA-MB-453 cells **(Figure 15)**. Compared with the controls, cells treated with ANKRD30A-siRNA grew faster, as the relative absorbance at 560 nm was usually higher in the siRNA groups. Moreover, the most significant difference was observed at the 48-72 hour time point, as the viability difference was usually the biggest and t-test P-values were the lowest. After 72 hours, the absorbance difference between the groups gradually decreased, especially in BT-474 and MDA-MB-453 cells. However, it should be pointed out that the overall viability difference is not so evident, because the biggest viability difference observed in the HCC-1500 cells at the 48-hour time point was only 20.62%. Overall, these data suggested that ANKRD30A ablation could promote the proliferation of breast cancer cells to some extent.

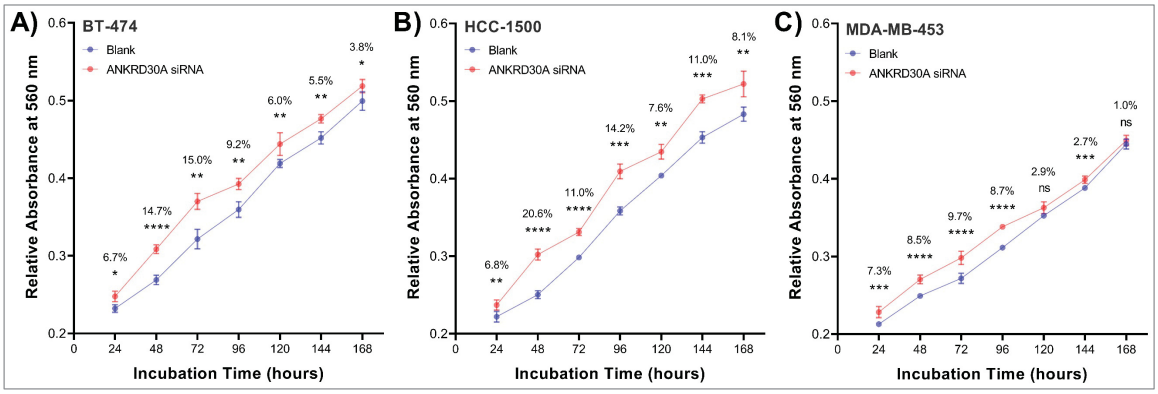


Figure 15. ANKRD30A for the proliferation of breast cancer cells. Cell viability was evaluated using the nontoxic PrestoBlue method by dynamically measuring the relative absorbance of live cells at 560 nm. The numbers above the trend lines represent the absorbance difference between the siRNA-treated cells and the control cells at each time point. Data were presented as mean ± standard deviation. **Abbreviations**: *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001.

3.14 ANKRD30A and Colony Formation

Next, the roles of ANKRD30A for colony growth of breast cancer cells were evaluated using the plate colony formation assays **(Figure 16)**. After an incubation period of 21 days, significantly more colonies were observed in cells treated with ANKRD30A-siRNA. In ANKRD30A-knockdown BT-474, HCC-1500, and MDA-MB-453 cells, the average number of cell colonies is 281±32, 619±29, and 512±49, respectively. In contrast, the colony number in corresponding normal cells is 143±34, 442±27, and 440±34, respectively. Moreover, the colony number difference between siRNA-treated cells and normal controls was statistically different in all of the three cell lines, with the t-test P-value of 0.0045 in BT-474 cells, 0.0205 in HCC-1500 cells, and 0.0447 in MDA-MB-453 cells, respectively. In brief, data from this section suggested ANKRD30A could inhibit the formation of cell colonies in breast cancer cells.

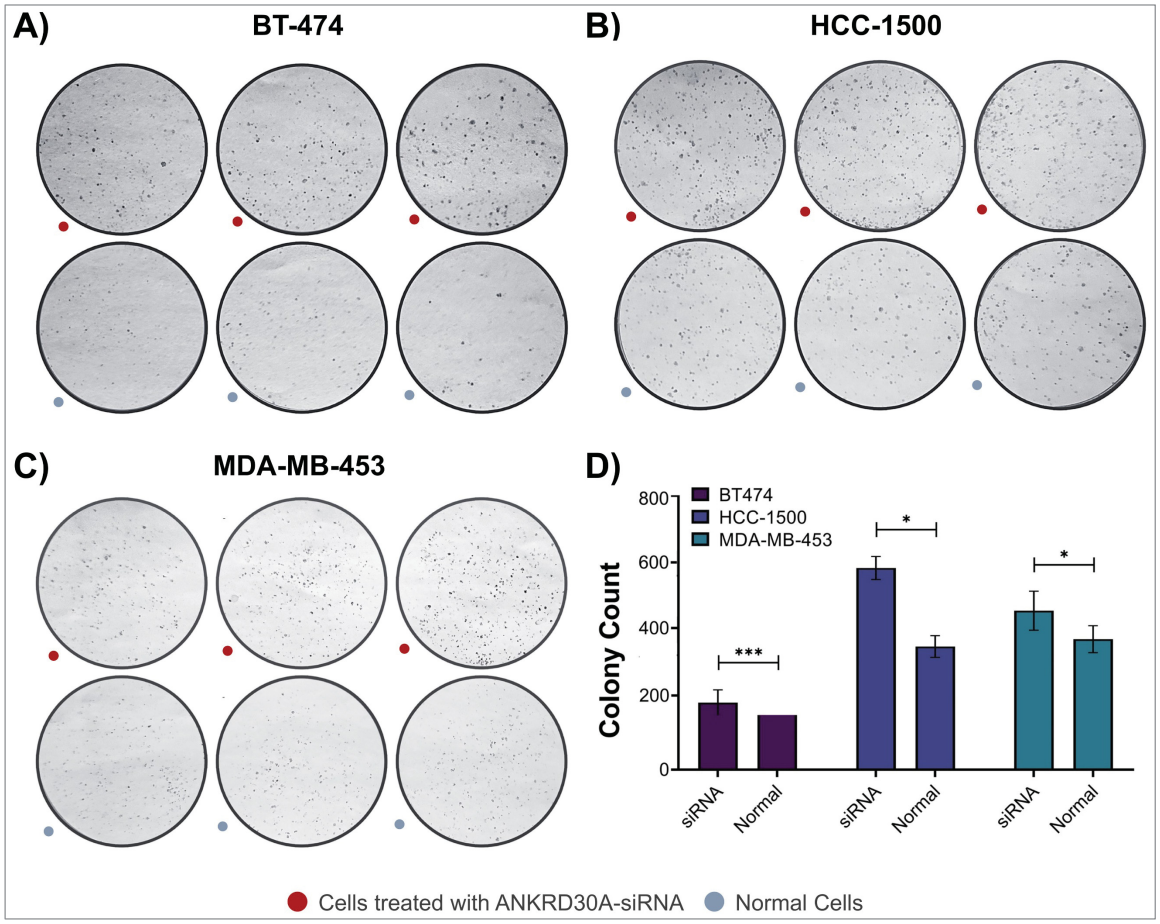


Figure 16. ANKRD30A for colony formation of breast cancer cells. The roles of ANKRD30A in colony formation of the A) BT-474 cells, B) HCC-1500 cells, and C) MDA-MB-453 cells were evaluated using the plate colony formation assay. The photos were taken on the 21-days after seeding cells. The red dots represent cells treated with ANKRD30A-siRNA, while the blue dots represent normal controls. D) The colony number in each cell line was presented as mean ± standard deviation, and the difference was assessed using the Student's t-test. **Abbreviations**: *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001.

3.15 Gene Set Enrichment Analysis

To predict potential biological functions of ANKRD30A in breast cancer, we performed the Gene Set Enrichment Analysis (GSEA) using the TCGA-BRCA and METABRIC dataset. We found that high ANKRD30A expression is significantly correlated with the activation of the Estrogen Response Early, UV Response, DN, and the Estrogen Response Late pathways, while low ANKRD30A expression is correlated with the activation of the E2F Targets, mROTC1 Signaling, and Cell Cycle pathways **(Figure 17, A-B)**. Since deregulation of cell cycle signaling is one of the hallmarks of breast cancer, [43,107] we then focused on the three cell-cycle-related gene sets that were commonly enriched in both TCGA-BRCA and METABRIC, namely the Cell Cycle, mROTC1 Signaling, and E2F Targets pathway **(Figure 17, C-E)**. By screening for genes that were significantly enriched in the three genesets, we found that MCM2, MCM4, CDC25A, and PLK1 are the four shared core enriched genes that are frequently activated in ANKRD30A-low breast cancer samples **(Figure 18)**. Next, we reviewed the expression of ANKRD30A and the four genes in breast cancer tissues and cell lines **(Figure 19)**. Briefly, the up-regulation of MCM2,

MCM4, CDC25A, and PLK1 is frequently observed in ANKRD30A-low samples, especially in Basal tumors. Meanwhile, their down-regulation is often detected in ANKRD30A-high samples, such as Luminal A tumors. Moreover, we also noticed that the activation of the 4 genes in breast cancer is usually oncogenic by interacting with cell-cycle-related signals. [108–113] Taken together, these data indicated that the aberrant inactivation of ANKRD30A may interact with one or more of the 4 cell cycle correlated genes and lead to tumorigenesis in breast cancer.

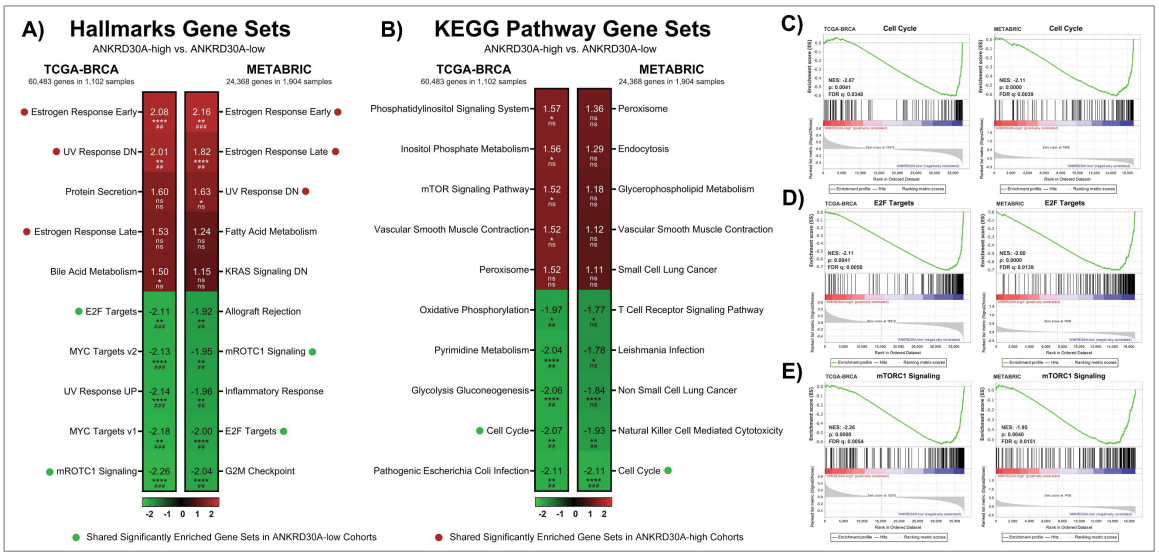


Figure 17. Gene Set Enrichment Analysis (GSEA) for predicting potential biological roles of ANKRD30A in breast cancer. The top 5 positively-enriched and top 5 negatively-enriched A) Hallmarks gene sets and B) KEGG pathway gene sets by comparing gene expression changes between ANKRD30A-high and ANKRD30A-low breast cancer samples. The median ANKRD30A values were used as the expression cutoff. The gene sets enriched in both the TCGA-BRCA and METABRIC datasets were labeled using the red (positively-enriched) or green (negatively-enriched) dots. The relationship between ANKRD30A expression and enrichment scores of C) the Cell Cycle, D) E2F Targets, and E) mTORC1 Signaling pathways. The enrichment scores were plotted in green curves, and high ANKRD30A expression was plotted in red rectangles, while low ANKRD30A expression in blue.

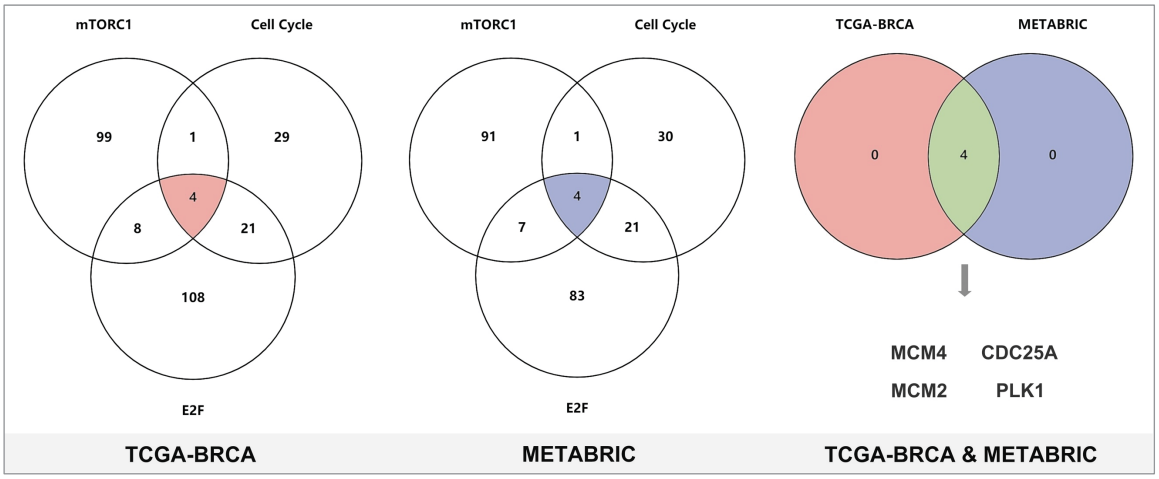


Figure 18. Associations between ANKRD30A and cell-cycle-related genes in breast cancer. The three cell-cycle-related gene sets were enriched in both the TCGA-BRCA and METABRIC datasets by comparing gene expression profiles in ANKRD30A-high and ANKRD30A-low breast cancer samples. Core enriched genes annotated in the three gene sets were screened for the shared ones. The four shared core enrichment genes were considered as ANKRD30A-correlated genes, namely MCM2, MCM4, PLK1, and CDC25A.

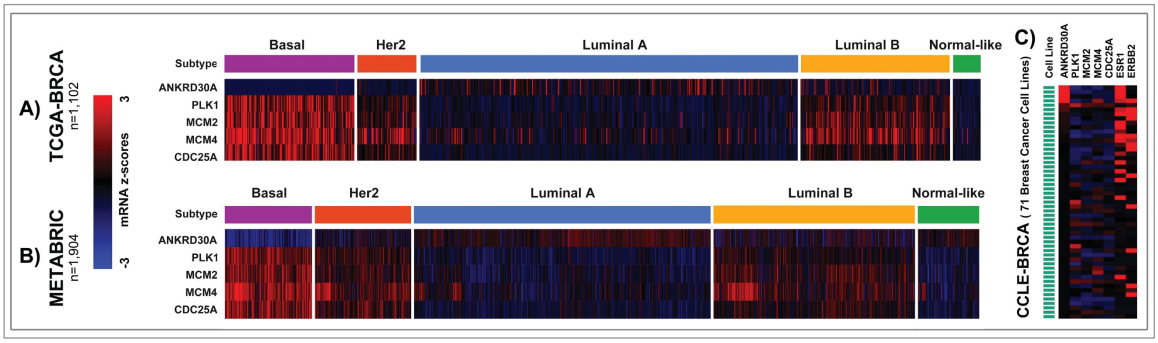


Figure 19. Expression of ANKRD30A and its correlated GSEA-enriched genes in breast cancer. In breast cancer tissues or cell lines from the A) TCGA-BRCA, B) METABRIC, and C) CCLE-BRCA datasets, the activation of the four cell-cycle-related genes was usually observed in ANKRD30A-low samples, while their expression was frequently down-regulated in ANKRD30A-high samples. The expression of ESR1 and ERBB2 was plotted for distinguishing the subtype of breast cancer cell lines.

3.16 Gene Expression Changes After ANKRD30A Silencing

Based on the above-mentioned findings, we noticed that ANKRD30A was co-expressed with LINC00993 in breast cancer, the expression of ANKRD30A was significantly associated with ER and HER2 status, and ANKRD30A may interact with the 4 cell-cycle-related genes (MCM2, MCM4, CDC25A, and PLK1) according to the GSEA results. Therefore, to confirm the association between ANKRD30A and the 8 genes in breast cancer, we compared their expression changes before and after ANKRD30A knockdown using qRT-PCR **(Figure 20)**. In both the BT-474 and HCC-1500 cells, the ablation of ANKRD30A will cause a significant down-regulation of LINC00993. Specifically, compared with the expression of ANKRD30A in untreated cells, the expression level of LINC00993 was only 41±8% ($P<0.001$) and 55±4% ($P<0.001$) in ANKRD30A-knockdown BT-474 and HCC-1500 cells, respectively. These data further confirmed the coexpression of the two genes in breast cancer and suggested LINC00993 is a downstream signal of ANKRD30A. Despite ANKRD30A is usually significantly down-regulated in ER-negative or HER2-negative breast cancer tissues, no significant changes were observed regarding the expression of ER (ESR1) and HER2 after ANKRD30A knockdown. This indicated that the aberrant ANKRD30A expression in breast cancer is more likely to be one of the consequences other than the causes of ER or HER2 expression changes. Moreover, among the 4 GSEA-predicted cell-cycle-related genes, only the expression level of CDC25A was significantly changed in ANKRD30A-ablated cells, and the corresponding relative expression levels were 151±7% ($P<0.01$) and 137±4% ($P<0.01$) in BT-474 and HCC-1500 cells, respectively. Meanwhile, the expression levels of MCM2, MCM4, and PLK1 remain stable between untreated and

ANKRD30A-siRNA-treated cells. Therefore, these results indicated LINC00993 and CDC25A may be regulated by ANKRD30A in breast cancer.

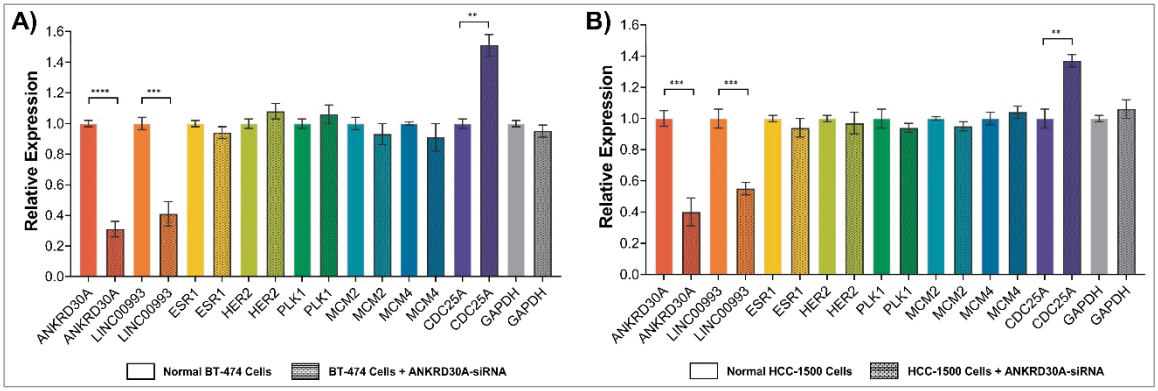


Figure 20. Gene expression changes after ANKRD30A silencing. qRT-PCR assessment of gene expression difference between normal and ANKRD30A-ablated A) BT-474 and B) HCC-1500 breast cancer cells. The bars filled with patterns represent gene expression in ANKRD30A-siRNA-treated cells, while the bars without patterns represent gene expression in corresponding untreated cells. The relative expression of each gene was presented as mean ± standard deviation, and the difference was assessed using the Student's t-test. **Abbreviations**: *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001.

3.17 Prognostic Significance of ANKRD30A in Breast Cancer

To explore the prognostic significance of ANKRD30A in breast cancer, we performed bioinformatics survival analysis using the Kaplan-Meier Plotter **(Figure 21)**. Using the median expression value of ANKRD30A as the cutoff, we found the high ANKRD30A expression in breast cancer is significantly associated with improved relapse-free survival (RFS; HR=0.75, 95%CI: 0.64-0.87, P=0.00, n=1,764) and overall survival (OS; HR=0.66, 95%CI: 0.48-0.91, P=0.01, n=626) in the overall analysis. Moreover, despite patients with a high ANKRD30A expression usually have fewer distant-metastases-free survival (DMFS) events, its difference compared with the ANKRD30A-low cohort is not statistically different (HR=0.74, 95%CI: 0.53-1.02, P=0.07, n=664). In subgroup analysis, we noticed that high ANKRD30A expression could be used as an effective prognostic marker for predicting better RFS in patients with ER-positive, HER2-negative, node-positive, and node-negative tumors. Similarly, ANKRD30A high expression is also associated with lower risks of OS and DMFS events in patients with p53-mutated and node-negative breast cancers, respectively. Meanwhile, ANKRD30A overexpression is also correlated with a shorter DMFS in ER-negative patients (HR=3.40, 95%CI: 1.32-8.72, P=0.01), but the number of patients in this cohort is so limited =68). Taken together, the high expression of ANKRD30A in breast cancer is generally a favorable prognosis predictor.

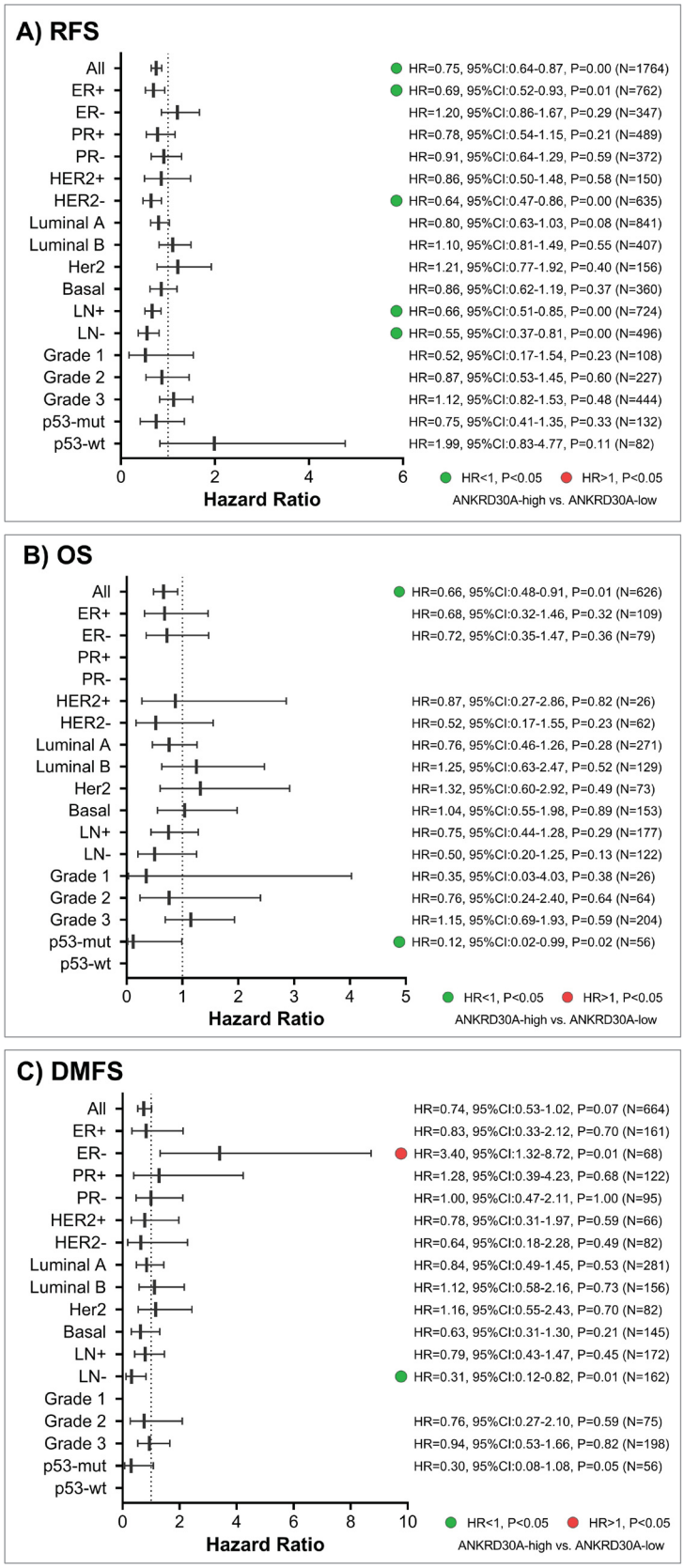


Figure 21. Prognostic significance of ANKRD30A in breast cancer. The forest plots were presented to illustrate the association between ANKRD30A expression and the A) RFS, B) OS, and C) DMFS outcomes of patients with breast cancer. Data from the overall analysis and subgroup analysis were integrated according to the results from the Kaplan-Meier Plotter. Patients were split into the ANKRD30A-high and ANKRD30A-low cohort by median ANKRD30A expression values. HR was computed by comparing the hazard in the ANKRD30A-high cohort to that in the ANKRD30A-low cohort. **Abbreviations**: RFS, relapse-free survival; OS, overall survival; DMFS, distant-metastases-free survival; HR, hazard ratio; 95%CI, 95% confidence interval; P, log-rank test P-value; LN, lymph node; mut, mutated; wt, wild type.

4 Discussion

4.1 Basic annotations of ANKRD30A

Ankyrin repeat domain-containing protein 30A (ANKRD30A), also known as NY-BR-1 or B726P, is a protein-coding gene located on the forward strand of chromosome 10 (37,125,725 - 37,384,111) **(Figure 22)**. [114] According to the annotations from the Ensembl genome browser, ANKRD30A has four protein-coding transcripts (ANKRD30A-201, ANKRD30A-202, ANKRD30A-204, and ANKRD30A-205) and one processed transcript (ANKRD30A-203). The length of its four protein-coding transcripts ranges from 4405 bp to 4894 bp, encoding proteins of 1341 aa to 1516 aa, while the length of the non-coding transcript ANKRD30A-203 is only 438 bp. Until now, very little is known about ANKRD30A, as only 47 papers can be retrieved from PubMed using the searching keywords “ANKRD30A” or “NY-BR-1” or “B726P” **(Figure 23)**.

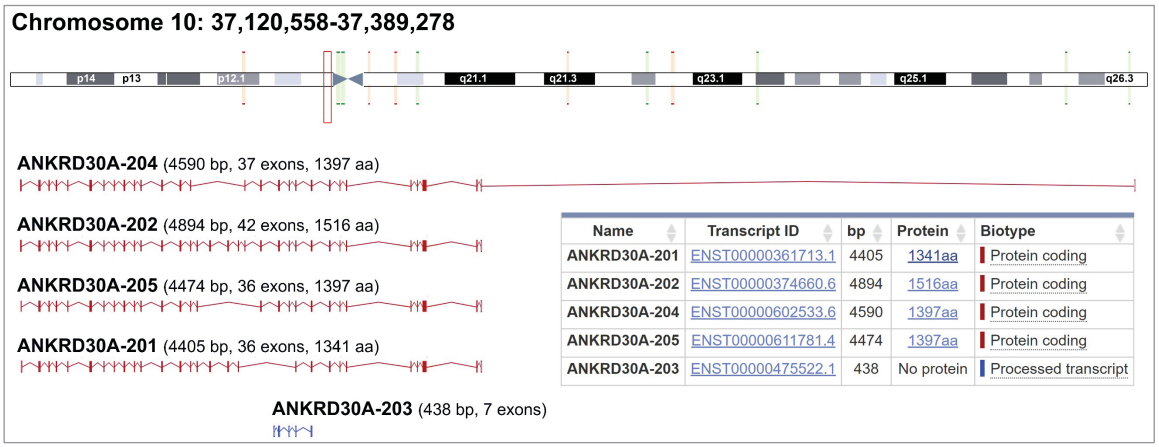


Figure 22. Basic annotations of ANKRD30A. The chromosome location, transcripts, and corresponding proteins of ANKRD30A were plotted based on the annotations from the Ensembl browser. The exons were plotted as red or blue vertical squares, while the introns were presented as horizontal lines. Data source: https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000148513

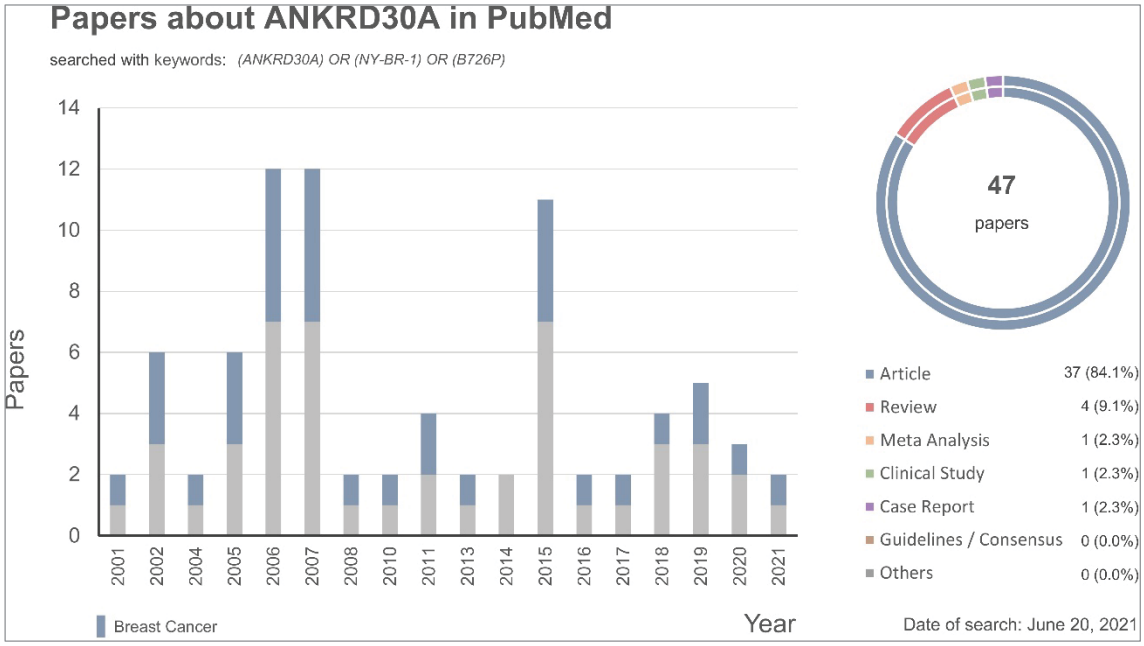


Figure 23. Publications about ANKRD30A in PubMed. PubMed was searched with the keywords "ANKRD30A" or "NY-BR-1" or "B726P." A total of 47 papers were retrieved, with the earliest literature published in 2001. Papers associated with breast cancer were highlighted in blue on the bar plot.

According to its full name (ankyrin repeat domain-containing protein 30A), ANKRD30A is one of the human proteins containing the ankyrin repeat, which is one of the most extensively existing amino acid structural motifs in nature, well known for mediating protein-protein interactions. [115,116] Ankyrin repeat is defined according to its structure rather than function. Typically, this repeat consists of 30-34 amino acids and occurs in at least four consecutive copies as a helix-loop-helix structure **(Figure 24)**. [117] Zeynep Kosaloglu et. al once reported that eight SNPs located on the ankyrin repeat domain of ANKRD30A were evaluated as damaging. [118] Therefore, it is reasonable to speculate that ANKRD30A is likely to be involved in some protein-protein reactions in breast cancer. However, most previous studies on ANKRD30A were focused on its expression, whereas its biological functions were rarely investigated.

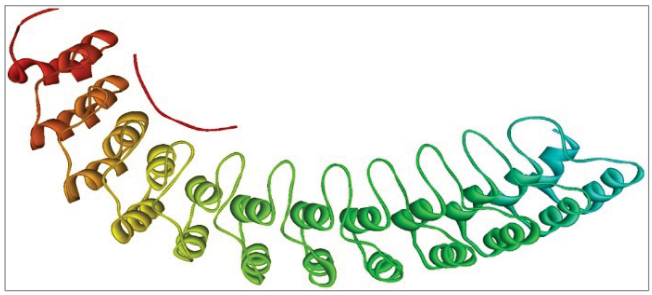


Figure 24. The structural diagram of the ankyrin repeat domain. Typically, an ankyrin repeat consists of 30-34 amino acids and occurs in at least four consecutive copies as a helix-loop-helix structure. **Figure source:** https://en.wikipedia.org/wiki/Ankyrin_repeat

4.2 Breast Specificity of ANKRD30A

4.2.1 Bioinformatics approaches versus SEREX for screening tissue-specific genes

ANKRD30A was first identified in 2001 by the screening of a breast cancer serological library using the SEREX (serological analysis of recombinant tumor cDNA expression libraries) technique, as ANKRD30A was found to have a tissue-restricted mRNA expression pattern in normal breast and testis. [119,120] SEREX is an immunological approach to identify tumor antigens with spontaneous humoral immune response. [121] Despite tissue-specific genes, such as breast-specific genes, can be identified using the SEREX method, it has several limitations compared with bioinformatics approaches. Firstly, the technical procedures in a SEREX assay are relatively complicated, requiring many laboratory reagents and devices, which is time and money consuming. [119] In contrast, screening potential breast-specific genes using bioinformatics methods with publicly available datasets is more cost-effective. Secondly, in a SEREX analysis, potential tissue-specific genes were screened from a cDNA library, which is usually generated from one or few tissue or cell line samples. [121,122] Therefore, the sample selection bias is inevitable and some important tissue-specific genes may be omitted because of sample heterogeneity. For example, ANKRD30A was firstly identified as a breast-specific antigen with the SEREX method using a cDNA library generated from one 60-year-old female with metastatic breast cancer. [119] Conversely, our bioinformatics screening is performed with many independent high-quality datasets, involving the expression data of at least 20,000 genes across tens of thousands of samples, greatly improving the screening accuracy and minimizing the sample selection bias. Thirdly, the SEREX-defined potential tissue-specific candidates must be validated using RT-PCR in normal and malignant samples. [119] However, RT-PCR is a relatively low-throughput technique compared with RNA-seq or microarray, and it takes time to collect enough specimens for a reliable validation. Hence, it is almost impossible to identify all potential breast-specific genes using the SEREX approach. By contrast, it is rather simple and fast to check the tissue specificity of candidate genes using bioinformatics tools.

4.2.2 Previously identified breast-specific genes

Previously, the expression patterns of several genes were also reported as breast-specific, including mammaglobin, GATA3, and MUCL1. [65,66,114,123,124] Among them, mammaglobin may be the most well-known gene with breast-specificity. [54] Moreover, these biomarkers were often used in clinical practice to identify the breast source of tumors with unknown origin. [125–131] However, their breast specificity has become controversial with the increase of related studies. For example, mammaglobin can also be detected in many non-breast tissues such as salivary, endometrial, ovarian, and cervical tissues. [63] Similarly, high GATA3 expression is often observed in clear cell papillary renal cell carcinoma, bladder cancer, squamous cell carcinoma, and prostatic adenocarcinoma. [132–135] Using the consensus RNA expression data from the Human Protein Atlas database, we reviewed the expression of mammaglobin, GATA3, and MUCL1 in normal human tissues **(Figure 25)**. Apart from breast tissues, the expression of mammaglobin can also be detected in the endometrium, cervix, and uterine. Similarly, MUCL1 can be detected in the salivary gland, which is not a sex-specific tissue type. Meanwhile, GATA3 cannot be regarded as breast-specific or even breast-selected, as its expression is widely distributed in many non-breast tissues. Notably, their expression patterns obtained from *in silico* RNA-seq data are consistent with laboratory-

confirmed cases. Based on our bioinformatics definition of genes with breast specificity, the there "breast-specific" genes were filtered out, while the more reliable candidate ANKRD30A was accurately identified. Briefly, these data demonstrated the reliability of our bioinformatics strategies and proved the superiority of these tools in certain research areas.

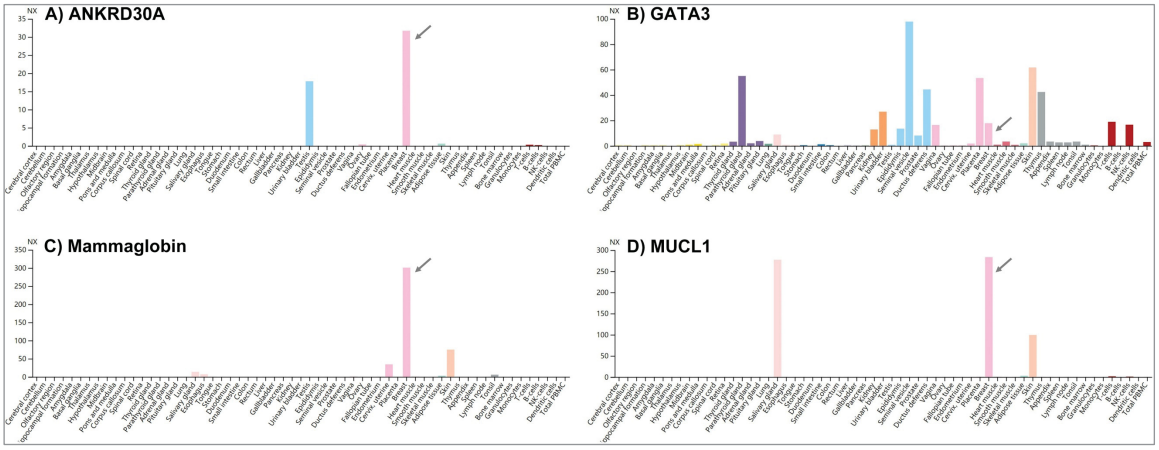


Figure 25. Expression of ANKRD30A and the three previously identified "breast-specific" genes in various types of normal human tissues. The normalized consensus RNA expression data from the Human Protein Atlas of A) ANKRD30A, B) GATA3, C) mammaglobin, and D) MUCL1. The expression of the four genes in breast tissues was indicated by arrows, and their expression in male-specific tissues was colored in blue. **Abbreviation:** NX, normalized expression.

4.2.3 ANKRD30A specificity in cancer scenario

Apart from normal breast tissues, data from several independent studies suggested that the positive ANKRD30A expression can also be frequently observed in breast cancer samples. For example, Neil O'Brien et. al detected the positive ANKRD30A mRNA expression in 44/108 (41%) breast cancer samples using RT-PCR assays, while its expression was undetectable in 0/20 non-breast tissues [65] Similarly, Anna H Woodard et. al detected the positive ANKRD30A protein expression in 111/190 (58.4%) breast cancer samples using immunohistochemical methods. [136] Besides, they also noticed that the staining intensity of ANKRD30A in cancer tissues is sometimes stronger than normal tissues, especially in ER-positive cases. In addition, Yuqiu Jiang et. al also found that the expression of ANKRD30A is breast-specific in normal breast tissues and malignant breast tumors using the customized PCR-based cDNA subtraction techniques and cDNA microarray. [137] Meanwhile, the positive ANKRD30A expression can sometimes be observed in non-breast cancers, such as sweat gland carcinomas and müllerian, but its positive rates in these tumors are low. [106,136] Despite previous studies did notice the potential breast specificity of ANKRD30A in cancer scenario, the conclusions were drawn through testing its expression in at most hundreds of samples across few cancer types using low-through techniques. In other words, the breast-cancer-specific expression patterns of ANKRD30A have never been systematically assessed before.

In this study, by presenting a comprehensive view of ANKRD30A expression in over 10,000 malignant samples spanning 33 common types of cancer using the TCGA pan-cancer dataset at individual levels, we demonstrated its breast specificity and highlighted the convenience of

bioinformatics tools. However, we also have to point out one limitation of our bioinformatics approaches. Almost all the presented bioinformatics gene expression data are mRNA expression data, while protein expression data are rarely shown. Despite TCGA also provided protein expression data measured by reverse-phase protein array (RPPA) or mass spectrometry (https://www.cbioportal.org/study/summary?id=brca_tcga), information regarding protein expression is very limited compared with mRNA expression. Specifically, the TCGA-BRCA RPPA dataset only documented the expression of 226 proteins, and the corresponding mass spectrometry assays were only performed in 74/1080 (6.7%) samples. Similarly, although the Human Protein Atlas provided standardized immunohistochemistrical protein expression data in normal tissues measured by tissue microarrays (https://www.proteinatlas.org/about/download), protein expression data in cancer tissues were not available. Some online databases such as the UALCAN (http://ualcan.path.uab.edu/index.html) provided integrated protein expression data for interactive analyses. Nevertheless, raw expression files were not accessible and the number of included samples was too limited. For example, the expression of ANKRD30A protein was only documented in 143 breast cancer samples (18 normal breast tissues and 125 primary breast tumors) in the USLCAN database, and the sample number is even smaller in subgroup analyses. Therefore, we believe it is difficult to generate reliable hypotheses using these small-size protein expression data.

4.3 Details regarding ANKRD30A expression

4.3.1 ANKRD30A exons and primers

The impacts of used primers on the detection rates of ANKRD30A are nonnegligible. In this study, after evaluating the sensitivity and specificity of the eight ANKRD30A RT-PCR primer pairs, we noticed that the positive rates of ANKRD30A varied greatly if different primers were used, and we also found that the shared sequences (exon 32-34 of ANKRD30A-204) on its four protein-coding transcripts are more likely to be detected than other exons **(Figure 26)**. Data reported by Jean-Philippe et. al supported our findings. [105] They demonstrated that the mRNA expression of ANKRD30A exon 30-33 (A30-33) is usually much stronger than its exon 4-7 (A4-7) in breast, testis, and their corresponding malignant specimens. Specifically, the positive rate of A30-33 is 69.5% in 442 breast cancer samples, while the figure for A4-7 is only 35.3% (n=439). Moreover, the positive expression of A4-7 is usually detected in A30-33-positive samples. Therefore, these data highlighted the distinct roles of A30-34 for the protein expression and biological functions of ANKRD30A.

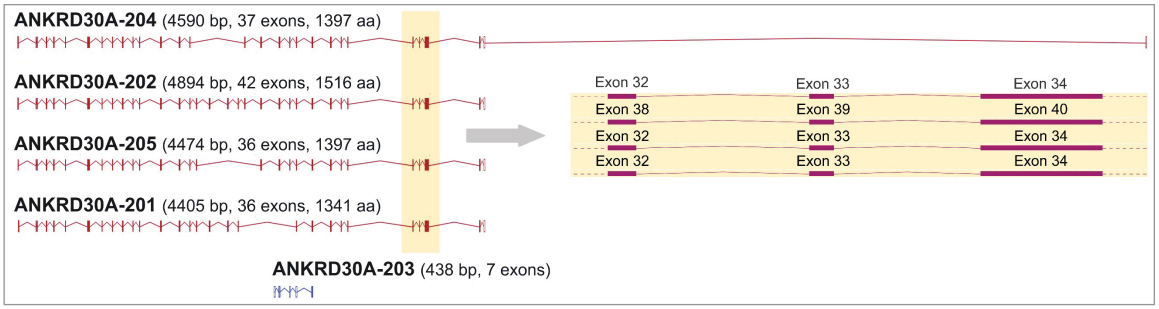


Figure 26. The shared sequences of the four ANKRD30A protein-coding transcripts. The four protein-coding ANKRD30A transcripts were plotted in red, while the non-coding one was plotted in blue. The shared sequences (exon 32-34 of ANKRD30A-204) of the four transcripts were highlighted in a yellow background and enlarged for a better view.

4.3.2 ANKRD30A expression in cell lines

Interestingly, despite the distinct expression of ANKRD30A has been extensively investigated in many independent studies, very little is known regarding its biological functions according to previous publications. This is probably because many commonly used breast cancer cell lines were ANKRD30A-negative. [138,139] We confirmed these findings by showing ANKRD30A can not be detected in the MCF-7, MDA-MB-231, and SK-BR-3 cell lines. Therefore, the selection of appropriate ANKRD30A-positive cell lines is of great importance to investigate its biological functions. Using the CCLE RNA-seq data, we checked ANKRD30A expression in 51 kinds of breast cancer cell lines and found several ANKRD30A-positive cell models. It is much easier to get these ANKRD30A-positive candidates using bioinformatics approaches than traditionally experimental screening. We demonstrated the reliability of these *in silico* predictions using traditional laboratory experiments, such as qRT-PCR and immunofluorescence staining. We have to admit these big-data-guided predictions greatly assist our choice, as many ANKRD30A-positive cell lines were rarely used, such as the CAMA1, ZR-7530, HCC-1419, HCC-202, and so forth.

However, according to our results, attention should be paid to the discrepancy between *in silico* gene expression data and gene expression levels measured by laboratory experiments. In detail, the CCLE-BRCA RNA-seq data suggested that ANKRD30A expression is higher in HCC-1500 cells (84.65 TPM) than BT-474 cells (4.04 TPM). But data from our RT-PCR assays indicated the contrary that ANKRD30A expression is about 3 times higher in BT-474 cells than HCC-1500 cells. The different sources, passage numbers, and culturing conditions between our cell lines and the CCLE cell lines may contribute to this quantitative discrepancy. Moreover, unlike gene expression in tissue samples, biological repeats or technical repeats were usually not available regarding gene expression in cell lines. For instance, we can generate a reliable hypothesis that ANKRD30A expression is significantly down-regulated in TNBC tissues compared to ER-positive breast cancer tissues, because this is based on data from at least hundreds of TNBC and ER-positive tissue samples. On the contrary, we can only compare its expression between one BT-474 cell line and one HCC-1500 cell line using the CCLE-BRCA RNA-seq data. Nevertheless, the CCLE-BRCA RNA-seq data did accurately predict the positivity or negativity of ANKRD30A in our tested cell lines. Hence, we believe it is reasonable to use these *in silico* data for qualitative or semi-quantitative analyses. In other words, these bioinformatics data are very helpful when predicting “yes or no,” but may not be that accurate when predicting “high or low” in cell lines.

4.3.3 Subcellular locations

Using immunohistochemical methods, Dirk Jäger et. al showed that the positive expression of ANKRD30A in normal breast tissues is mainly observed in the epithelia of ducts and acini of the mammary gland in a focal fashion. [106] This is further confirmed by the immunohistochemical results from the study of Zsuzsanna et. al in normal breast tissues. [140] Meanwhile, in breast tumor tissues (ductal carcinoma in situ, infiltrative breast ductal carcinoma, and metastatic breast

cancers), the positive ANKRD30A immunoreactivity can be observed in both the cytoplasm and nucleus. [140] Besides, Inka Seil et. al revealed that ectopically expressed ANKRD30A was mainly found in the cytoplasm and cell membrane, and its membrane localization is mediated by two cis-active membrane targeting domains. [138] To my best knowledge, our data described the subcellular locations of ANKRD30A in breast cancer cell lines for the first time. In detail, ANKRD30A proteins can be detected in the cytoplasm in all the three tested breast cancer cell lines (HCC-1500, BT-474, and MDA-MB-453), while positive perinuclear ANKRD30A expression can only be observed in the MDA-MB-453 cell line. Moreover, after silencing ANKRD30A, the perinuclear ANKRD30A decreased significantly while some fluorescent signals of cytoplasmic ANKRD30A can still be observed. However, factors that may influence its subcellular locations remain unclear.

4.3.4 ANKRD30A and histological grading

Data from several independent studies suggested that positive ANKRD30A expression was more likely to be detected in low-grade breast cancer tumors using immunohistochemical techniques. [136,140,141] Specifically, positive ANKRD30A staining can be observed in 77-82% of grade 1 and 63-63% of grade 2 carcinomas, while its positive rate in grade 3 specimens was only 29-50%. A study in 60 Egyptian females from Salma et. al suggested that ANKRD30A expression is significantly correlated with histological grading, progesterone receptor status, disease stages, menopausal status, and lymph node metastases. [142] However, in the results section, the correlation between ANKRD30A expression and histological grade of breast cancer was not presented. This is because data regarding tumor grading were not documented in the TCGA-BRCA and METABRIC datasets. Instead, we explored the Breast Cancer Gene-Expression Miner (http://bcgenex.ico.unicancer.fr/BC-GEM/GEM-Accueil.php?js=1), an online interactive tools for mining published breast cancer transcriptomic data, to check the association between ANKRD30A expression and tumor grading **(Figure 27)**. As expected, *in silico* data from 6,441 microarray and 3,617 RNA-seq samples are consistent with these experimental findings. In detail, the median ANKRD30A expression is the highest in grade 1 breast cancer samples, followed by grade 2 tumors, and is the lowest in grade 3 tumors. Besides, the difference between the three groups is statistically different. Taken together, these data again demonstrated the reliability of bioinformatics tools and confirmed the association between ANKRD30A and histological grading of breast cancer.

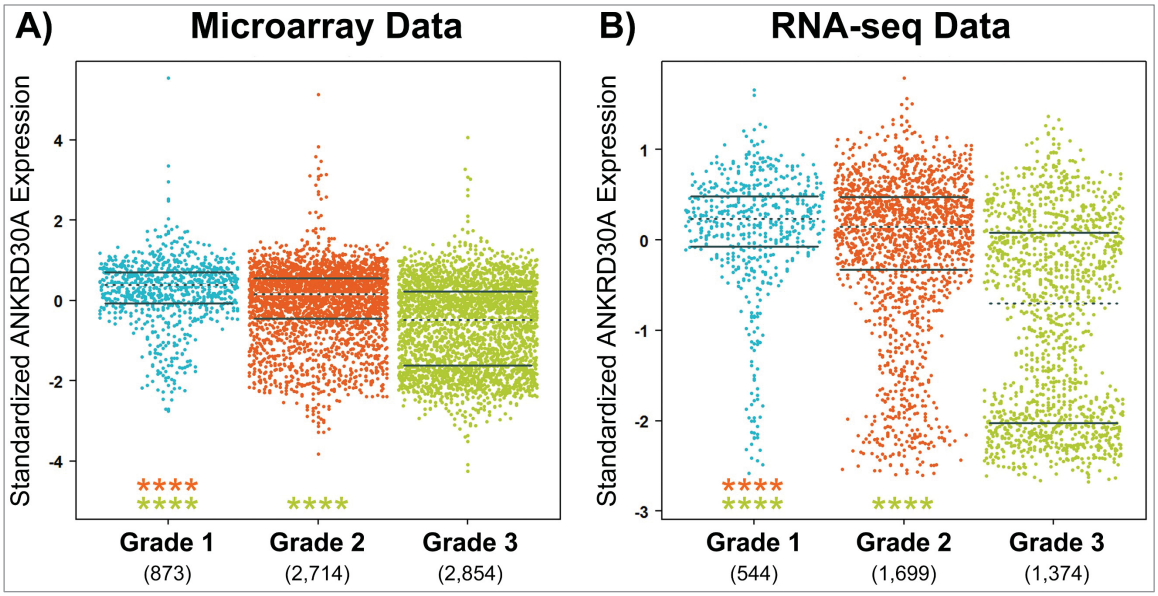


Figure 27. The association between ANKRD30A expression and histological grading of breast cancer. The integrated A) Microarray and B) RNA-seq expression data from the Breast Cancer Gene Expression Miner regarding ANKRD30A expression in breast cancer tissues of different histological grades. The statistical difference between every two groups was evaluated by Dunnett-Tukey-Kramer's test. The corresponding symbolized P-values were presented as colorful asterisks. **Abbreviation**: ****, P<0.0001.

4.4 ANKRD30A expression and prognosis

Consistent with our findings, other studies also proved that high ANKRD30A expression usually indicates better breast cancer prognostic outcomes. For example, Jean-Philippe et. al reported that the 5-year overall survival for breast cancer patients with strong ANKRD30A staining was 83% (n=357), while the figure for ANKRD30A-negative patients was 74% (n=337). [141] Similarly, Ya Wang et. al reported that high ANKRD30A expression in TNBC was associated with longer relapse-free survival (HR=0.57, 95%CI: 0.33-1.00, log-rank test P=0.047; n=161). [143] However, it must be pointed out that these survival data about ANKRD30A in breast cancer were all prognostic other than predictive. [144] In other words, the potential values of using ANKRD30A expression in guiding personalized management of patients with breast cancer remain unknown. Moreover, despite that at least 600 patients were included for analyzing the prognostic value of ANKRD30A in the overall analysis, the sample size in some subgroup analyses is relatively small. For example, in the OS subgroup analyses, the survival data were only available in 26 HER2-positive patients, which yielded less reliable conclusions. Meanwhile, since the significance of multigene signatures for assisting inconclusive clinical decisions has been demonstrated in many prospective clinical trials, such as the TAILORx trial for 21-gene [145] and the MINDACT trial for 70-gene, [146] it is reasonable to further evaluate the potential clinical values of ANKRD30A-containing multigene signatures in breast cancer. [147,148]

4.5 Clinical Utility of ANKRD30A

4.5.1 Identify the breast source of carcinoma of unclear primary

The potential clinical utilities of ANKRD30A have been evaluated in many studies. For example, ANKRD30A was demonstrated as a useful diagnostic biomarker to identify the breast source of primary or metastatic tumors, especially for carcinomas of unclear origin. Neil et. al evaluated the value of ANKRD30A for differentiating breast cancer from non-breast tissues. [65] They found that ANKRD30A can be detected in 44/108 (41%) of breast cancer tissues, while 0/20 of non-breast tissues were ANKRD30A-positive, and the specificity of ANKRD30A was the highest compared with other breast-specific biomarkers such as mammaglobin, SBEM, and MUC1. Olivier et. al reported a case of a 49-year-old woman that the positive ANKRD30A expression was used to confirm the breast source of a phyllodes tumor in the labium majus. [149] Besides, they further detected the positive ANKRD30A expression in 21/26 (81%) extramammary Paget's disease samples and 18/24 (75%) of their mammary counterparts. [150] Giovanni et. al analyzed the expression of ANKRD30A in 38 consecutive male breast cancer samples. [66] They found that ANKRD30A was positive in all the 30 primary tumors and all the 8 metastatic tumors, while the positive rates of the two breast-specific genes GATA3 and Mammaglobin were 75% in metastatic samples. Moreover, the presence of ANKRD30A in metastases is possibly deduced from the corresponding primary tumors. [141] According to data from Jean-Philippe et. al, the expression of ANKRD30A is positive in 15/17 (88%) of the recurrent breast tumors and corresponding primary breast cancer samples. Similarly, positive ANKRD30A can be detected in 10/14 (71%) of the metastatic breast tumors and matched primary counterparts. Notably, the low positive rate of ANKRD30A in ER-negative tumors may impede its application in the diagnosis of unclear primary carcinomas. Hence, the combination of ANKRD30A and other breast-specific markers that have high positive rates in ER-negative tumors, especially in TNBC tumors, is worth investigating.

4.5.2 Assist the diagnosis of sentinel lymph nodes metastases

Meanwhile, several studies tried to assess the value of ANKRD30A for diagnosing sentinel lymph node (SLN) metastases from breast cancer. Christian et. al measured ANKRD30A mRNA expression in 74 breast cancer SLNs using qRT-PCR assays, and they found that the sensitivity and specificity of ANKRD30A alone for diagnosing SLN metastases is 82.4% and 93.8%, respectively. [151] Furthermore, they also showed that the gene panel consists of CK19, MGB1, EPCAM, and ANKRD30A is optimal for SLN diagnosis, with a sensitivity of 95.9% and a specificity of 95.0%. Besides, the positive predictive value using this ANKRD30A-containing gene panel is 85.5%, while the negative predictive value is 98.7%, and the overall concordance rate with histology is 95.2%. Zsuzsanna Varga et. al reported that the positive ANKRD30A expression can be detected in 49% of metastatic lymph nodes from breast cancer, and its expression pattern in SLNs can be either focal or diffuse. [140] Similarly, another study performed with 30 breast cancer SLN samples suggested that positive ANKRD30A expression can be used to detect minimal residual diseases, and its sensitivity is higher than routine histological examinations or immunohistochemistrical staining. [152] In addition, a multigene panel of mammaglobin, GABApi, B305D, and ANKRD30A for RT-PCR is very sensitive and specific for diagnosing breast cancer. [153] The positive signals of this multigene panel can be detected in 27/27 primary breast cancer tumors and

50/50 metastatic breast cancer lymph nodes, while no positive signals were observed in the included 27 non-breast tumors, 22 non-breast normal tissues, and 14 colon tumors. Nevertheless, it must be pointed out that the expression of ANKRD30A was measured using RT-PCR assays in almost all of the above-mentioned studies. Since immunohistochemistrical approaches are more commonly used in clinical practice, the practical values of ANKRD30A for assisting the diagnosis of breast cancer SLN metastases still remain inconclusive.

4.5.3 ANKRD30A as a potential circulating biomarker for breast cancer

The feasibility of using ANKRD30A as a circulating biomarker has also been explored. In this study, we showed that the expression of ANKRD30A is always below the positive cutoff in normal blood cells according to data from the Human Protein Atlas. In contrast, its expression can be frequently detected in blood samples from patients with breast cancer. [154] Besides, its positive rate as well as gene copies were higher in patients with advanced diseases, especially those with metastatic lesions. However, the exact data regarding its detection rate is not available. Meanwhile, since ANKRD30A expression in females is breast-specific, it was considered as one potential biomarker for detecting disseminated or circulating breast cancer cells. [114] For example, Monica et. al assessed the feasibility of using ANKRD30A as one of the potential biomarkers for detecting circulating breast cancer cells. [155] By performing RT-PCR assays with epithelial cells enriched from 20 mL of blood from patients with breast abnormality, they showed that the sensitivity and specificity of ANKRD30A alone for detecting circulating breast cancer cells were 59.1% and 58.0%, respectively. Notably, ANKRD30A in their study was detected using Tagman probes other than primers targeting its exon 30-34. Since we already demonstrated the impacts of primers on the detection rates of ANRKD30A, the sensitivity and specificity of ANKRD30A for detecting circulating breast cancer cells might be higher if “optimal” primers were used. Briefly, the number of studies on circulating ANKRD30A was very small as no other relevant publications were available. Therefore, further studies are needed to evaluate the significance of ANKRD30A as a circulating biomarker in breast cancer, including its potential utilities for early cancer detection, response evaluation, and recurrence monitoring.

4.6 ANKRD30A SNPs

Apart from breast-specific expression, single nucleotide polymorphism (SNP) of ANKRD30A was also reported. For example, Zeynep Kosaloglu et. al systematically analyzed ANKRD30A SNPs using *in silico* methods. They showed that the majority (77.5%) of the 2,800 recorded ANKRD30A SNPs were non-coding intronic SNPs, and at least 16/191 non-synonymous SNPs were predicted as damaging, with the rs200639888, rs367841401 and rs377750885 evaluated as the most damaging SNPs. [118] Interestingly, ANKRD30A-related mutations were usually studied in non-breast diseases. For example, ANKRD30A mutations can be observed in Ewing sarcomas at a low frequency (9.6%), and five of its mutations in the 50 analyzed cases were recurrent mutations. [156] Besides, the rs11010435 SNP from the intergenic region of ANKRD30A was found to contribute to smoking behavior in a male-specific manner. [157] In addition, the SNP rs1192691 at 10p11.21 near ANKRD30A was identified as one of the susceptibility loci for epithelial ovarian cancer in Han Chinese women. [158] However, in breast cancer, genetic alterations of ANKRD30A have never been reported before. According to our results from over 6,000 breast cancer samples,

mutations and CNVs of ANKRD30A were rare events in both primary and metastatic breast tumors, with a mutation detection rate of 12/6,009 (0.20%) and a CNV detection rate o 43/2,389 (1.76%) in all analyzed samples. Moreover, these negative results effectively guided the direction of our further laboratory experiments, as it is very unlikely to detect ANKRD30A mutations or CNVs in clinical samples, let alone evaluating their impacts on the development of breast cancer. Therefore, we believe the correct use of these *in silico* data could avoid unnecessary waste of time and resources.

4.7 ANKRD30A is a Putative Transcription Factor

The expression of ANKRD30A is breast-specific in adult women, indicating corresponding tissue-specific functions. These tissue/cell-type-specific expression patterns are often observed in many transcription factors. [159] Transcription factors generally refer to proteins capable of binding DNA in a sequence-specific manner and regulating transcription. Currently, most transcription factors, confirmed or putative, have been identified and classified based on their sequence similarity to previously characterized DNA-binding domains. [159,160] ANKRD30A is annotated as a DNA-binding protein in the Gene Ontology (GO) database according to the findings from Jäger et. al in 2001, as they demonstrated that the leucine zipper sequence, a motif that belongs to the basic region leucine zipper (bZIP) transcription factors, is observed on the amino acid sequence of ANKRD30A **(Figure 28)**. [119,161,162] Therefore, the breast specificity of ANKRD30A and its leucine zipper sequence suggested that ANKRD30A is a putative tissue-specific transcription factor of the bZIP family. However, there is no laboratory evidence to support that ANKRD30A is a transcription factor.

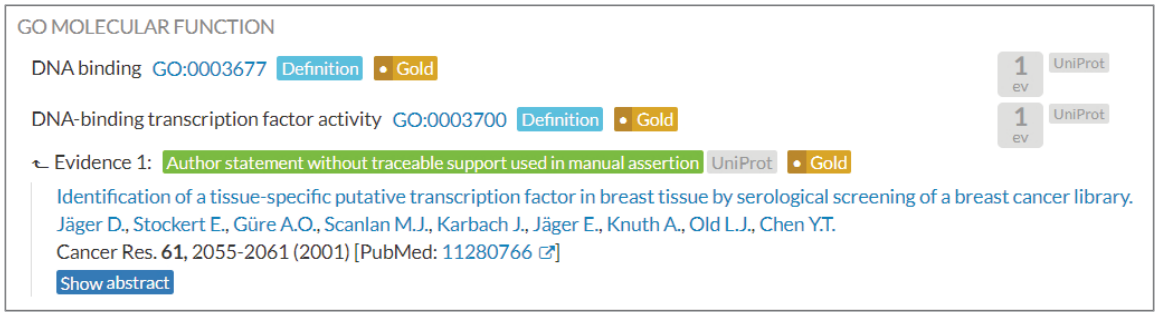


Figure 28. Function annotations for ANKRD30A in the Gene Ontology database. ANKRD30A was annotated as a putative transcription factor based on the evidence from Jäger et. al in 2001. **Figure source**: https://www.nextprot.org/entry/NX_Q9BXX3

4.8 The correlation between ANKRD30A and ER

In this study, our data suggested that the positive ANKRD30A expression was usually observed in ER-positive breast cancers tissues and cell lines, while its expression was obviously down-regulated in TNBC tumors. Moreover, we also found that the expression of ER mRNA remained unchanged after silencing ANKRD30A mRNA at exon 34 in two breast cancer cell lines. These

findings demonstrated the correlation between the two genes and also indicated that ER may be one of the upstream signals for ANKRD30A. Evidence from many previous publications could support our findings.

For example, Jean-Philippe et. al measured ANKRD30A expression in 1,350 breast cancer samples using tissue microarrays, and they found that 70% of the ER-positive tumors were also ANKRD30A-positive, whereas 67% of the ER-negative specimens were also ANKRD30A-negative. [141] Moreover, in one of their subsequent researches, they further proved that the expression of both ANKRD30A mRNA and protein are all significantly correlated with ER status in breast cancer, and this correlation in much stronger in premenopausal breast cancer patients than postmenopausal patients. [105] Nika C Gloyeske et. al also found that positive ANKRD30A expression can be detected more frequently in ER-positive breast cancers (92/152), while only 8/38 ER-negative tumors were ANKRD30A-positive. [103] Additionally, in ER-positive breast cancers, primary or metastatic, the positive rates of ANKRD30A are usually higher than previously identified breast-specific markers such as GATA3 and Mammaglobin. [66] In contrast, Tamás et. al reported that positive ANRKD30A staining can only be detected in 7/115 TNBC tumors, and its staining strength is usually very weak. [163] Similarly, Anita Sejben et. al reported that only 37/119 TNBC tumors can be immunohistochemically classified into ANKRD30A-positive tumors. [164]

Apart from expression correlation, Jean-Philippe et. al also identified four estrogen-response-like elements within the chromosome region 10,000 bp upstream to the transcriptional start site of ANKRD30A. [105] Moreover, they also noticed the impacts of Tamoxifen, one of the most extensively used selective estrogen receptor blockers, on ANKRD30A expression. Compared with primary tumors, the expression of ANKRD30A was significantly decreased in matched recurrent tumors of patients who received Tamoxifen treatment. Combined with our findings that ER may be one of the upstream signals for ANKRD30A, it is reasonable to hypothesize that the expression of ANKRD30A may be directly regulated by ER.

4.9 The mechanism of ANKRD30A Silencing

The expression of ANKRD30A is breast-specific in adult women, which also means ANKRD30A is always silenced in other non-breast normal tissues. Besides, the silence status of ANKRD30A can also be observed in many TNBC tumors. However, the exact molecular mechanisms of ANKRD30A silencing remain unclear. Data from Lanlan Shen et. al suggested that CpG islands methylation may contribute to the silence of ANKRD30A in normal tissues and cancer cell lines. [139] They demonstrated that the hypermethylation status of ANKRD30A can always be detected in samples where the expression of ANKRD30A is silenced, such as the breast cancer cell line MDA-MB-231, MCF-7, and SK-BR-3. Furthermore, they proved that DNA demethylating agents such as 5-aza-2′-deoxycytidine and deacetylase inhibitor trichostatin A can trigger the reactivation of silenced ANKRD30A in cancer cell lines, including the breast cancer cell line MCF-7 and HTB-126.

Moreover, since it is well recognized that methylation of CpG islands can cause stable gene silencing, [165–167] we explored the UCSC genome browser to search CpG islands adjacent to

ANKRD30A sequences. As expected, at about 1,000 bp upstream from the transcriptional start site of ANKRD30A, a 624 bp sequence consisting of 51 CpG islands was identified **(Figure 29)**. Therefore, these findings indicated that the methylation of the ANKRD30A-adjacent CpG islands may greatly influence its expression in both normal tissues and cancer samples. However, several critical questions must be appropriately addressed before drawing reliable conclusions, including but not limited to proving its methylation status is different in clinical breast cancer samples with different ANKRD30A expression, demonstrating the correlation between the predicted CpG islands and ANKRD30A methylation, and identifying or excluding other factors that may also contribute to its silence.

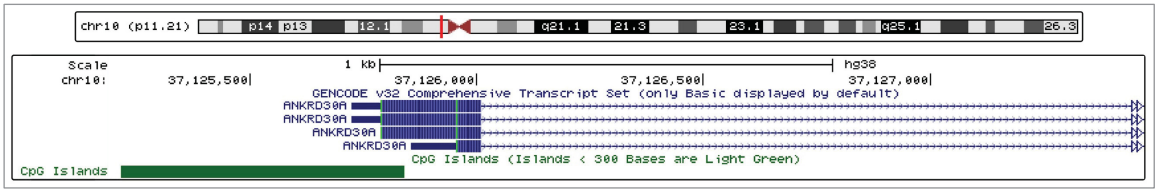


Figure 29. The closest CpG islands to ANKRD30A. The four protein-coding transcripts of ANKRD30A were plotted in blue. The 624 bp sequence consisting of 51 CpG islands located at the upstream of ANKRD30A transcriptional start site was plotted in green. Figure source: https://www.genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr10%3A37125100-37127472&hgsid=1010489939_jJkwd8mfyFrujzh6q6JprBBeazqB

4.10 The significance of ANKRD30A for breast cancer immunotherapy

ANKRD30A is generally regarded as an promising target for breast cancer immunotherapy. Firstly, ANKRD30A was demonstrated as a transmembrane protein as its expression is mainly detected in the cytoplasm and cell membrane, which indicates it can be specifically recognized by the anti-ANKRD30A antibodies on the surface of vital cells. [138] Our immunofluorescence staining assays in the HCC-1500 and BT-474 cell lines yielded the same results, and we also detected focal perinuclear ANKRD30A expression in the ER-negative MDA-MB-453 cell line. These findings makes it theoretically possible to apply ANKRD30A-specific antibody-based immunotherapy in breast cancer patients, especially for ER-positive breast cancers. Secondly, ANKRD30A expression in females is breast-specific, enabling targeted therapy and greatly minimizing off-target toxicities. [105,122,168,169] Besides, the tissue specificity of its expression is critical for the effective delivery of antibody-drug conjugates. [170,171] Additionally, we proved that only a very small proportion of genes in the human genome are breast-specific, and ANKRD30A was one of most reliable candidates.

Some studies have explored the feasibility of ANKRD30A-targeted immunotherapy. For example, Krishna et. al introduced the first ANKRD30A-specific H2-Db-resctricted CD8-positive T cell epitope and preliminarily demonstrated the efficacy of ANKRD30A-specific vaccine in murine tumor models. [172] The growth of ANKRD30A-positive tumors was partially inhibited when the mice received ANKRD30A-specific immunization. Similarly, Wei Wang et. al demonstrated that the p904 epitope derived from ANKRD30A is efficiently processed and endogenously

presented, and the stimulation from ANKRD30A peptides could trigger specific CD8-positive cytotoxic T cells responses. [173] Likewise, Dirk et. al identified two HLA-A2 restricted ANKRD30A epitopes (p158-167 and p960-968) that are recognized by CD8-positive T cell clones through ELISPOT analysis and tetramer staining, suggesting it is a potential strategy to develop active immunotherapy for HLA-A2 positive breast cancer patients with ANKRD30A-positive tumors. [169] However, it has to be pointed out that vaccine-induced immune response directly against ANKRD30A may also affect normal breast tissues, since positive ANKRD30A expression can also be detected in healthy breast epithelial cells. [174] Meanwhile, ANKRD30A was also found to be involved in some immune-related diseases. For instance, Nicholas et. al reported that ANKRD30A is one of the host factors promoting the replication of human immunodeficiency virus 1 (HIV-1) and that interferon could stimulate the rapid evolution of ANKRD30A. [175] Data from a meta-analysis indicated that ANKRD30A was one of the shared loci with immunoregulatory functions across ten pediatric autoimmune diseases. [176]

4.11 Summary and Conclusions

The main findings of this study can be summarized as follows: 1) A total of 96 genes were identified to have breast-specific expression patterns in the genome of normal adult females. 2) 6 breast-specific genes located on 10p11.21 in a focal pattern were focused on. Among them, ANKRD30A was selected for further analysis because its its sequence covers the full length of the other five genes and its potential roles as a tumor suppressor. 3) The expression of ANKRD30A is not only breast-specific in various types of normal female tissues, but also in more than 30 types of female carcinomas. Moreover, ANKRD30A expression is always below the positive cutoff in 18 types of normal blood cells. 4) In breast cancer, the expression of ANKRD30A is significantly down-regulated in ER-negative tissues and cell lines, especially in TNBC, while its correlation with other clinical characteristics of breast cancer is not significant. 5) Copy number variations and mutations of ANKRD30A are very rare in both primary and metastatic breast cancer. 6) The sensitivity of ANKRD30A mRNA detection by PCR varies greatly if different primer pairs were used, especially primer pairs targeting its different exon regions. 7) The location of ANKRD30A proteins is mainly cytoplasmic in the BT-474, HCC-1500, and MDA-MB-453 cell lines, whereas perinuclear ANKRD30A can only be detected in the MDA-MB-453 cell line in this study. 8) Knocking down ANKRD30A with siRNA could promote the ability of breast cancer cells in proliferation and colony formation to some extent. 9) The down-regulation of ANKRD30A in breast cancer is associated with the activation of cell-cycle-related signaling, especially CDC25A, MCM2, MCM4, and PLK1. 10) The expression of LINC00993 was significantly decreased while CDC25A was significantly increased in ANKRD30A-silenced cell lines. 11) High ANKRD30A expression in breast cancer usually indicates better survival outcomes.

In conclusion, ANKRD30A is one of the rare breast-specific genes in the human genome, and its well-established breast specificity is valuable for improving the management of breast cancer. ANKRD30A is significantly and frequently down-regulated in TNBC, which is possibly oncogenic because of the association between low-ANKRD30A and increased cell proliferation as well as the activation of cancer-related cell cycle signaling.

We believe our results presented a comprehensive view of breast-specific genes in the human genome. Moreover, we highlighted the clinical significance of ANKRD30A and explored its biological functions in breast cancer. Lastly, we demonstrated the feasibility and reliability of using bioinformatics tools to facilitate laboratory experiments.

5 References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a Cancer Journal for Clinicians*. 2021;71. doi:10.3322/caac.21660

2. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. *CA: a Cancer Journal for Clinicians*. 2019;69. doi:10.3322/caac.21583

3. Britt KL, Cuzick J, Phillips K-A. Key steps for effective breast cancer prevention. *Nature Reviews Cancer*. 2020;20. doi:10.1038/s41568-020-0266-x

4. Cardoso F, Spence D, Mertz S, et al. Global analysis of advanced/metastatic breast cancer: Decade report (2005-2015). *Breast (Edinburgh, Scotland)*. 2018;39. doi:10.1016/j.breast.2018.03.002

5. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406. doi:10.1038/35021093

6. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: Classification, prognostication, and prediction. *Lancet (London, England)*. 2011;378. doi:10.1016/S0140-6736(11)61539-0

7. Tang P, Tse GM. Immunohistochemical surrogates for molecular classification of breast carcinoma: A 2015 update. *Archives of Pathology & Laboratory Medicine*. 2016;140. doi:10.5858/arpa.2015-0133-RA

8. Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2020;38. doi:10.1200/JCO.19.02309

9. Jorns JM. Breast cancer biomarkers: Challenges in routine estrogen receptor, progesterone receptor, and HER2/neu evaluation. *Archives of Pathology & Laboratory Medicine*. 2019;143. doi:10.5858/arpa.2019-0205-RA

10. Waks AG, Winer EP. Breast cancer treatment: A review. *JAMA*. 2019;321. doi:10.1001/jama.2018.19323

11. Ruddy KJ, Ganz PA. Treatment of nonmetastatic breast cancer. *JAMA*. 2019;321. doi:10.1001/jama.2019.3927

12. Krauss K, Stickeler E. Endocrine therapy in early breast cancer. *Breast Care (Basel, Switzerland)*. 2020;15. doi:10.1159/000509362

13. Pan H, Gray R, Braybrooke J, et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *The New England Journal of Medicine*. 2017;377. doi:10.1056/NEJMoa1701830

14. Aromatase inhibitors versus tamoxifen in early breast cancer: Patient-level meta-analysis of the randomised trials. *Lancet (London, England)*. 2015;386. doi:10.1016/S0140-6736(15)61074-1

15. Wang J, Xu B. Targeted therapeutic options and future perspectives for HER2-positive breast cancer. *Signal Transduction and Targeted Therapy*. 2019;4. doi:10.1038/s41392-019-0069-2

16. Korde LA, Somerfield MR, Carey LA, et al. Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: ASCO guideline. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2021;39. doi:10.1200/JCO.20.03399

17. Yin L, Duan J-J, Bian X-W, Yu S-C. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research : BCR*. 2020;22. doi:10.1186/s13058-020-01296-5

18. Harris LN, Ismaila N, McShane LM, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American society of clinical oncology clinical practice guideline. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2016;34. doi:10.1200/JCO.2015.65.2289

19. Van Poznak C, Somerfield MR, Bast RC, et al. Use of biomarkers to guide decisions on systemic therapy for women with metastatic breast cancer: American society of clinical oncology clinical practice guideline.

Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2015;33. doi:10.1200/JCO.2015.61.1459

20. Wolff AC, Hammond MEH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2013;31. doi:10.1200/JCO.2013.50.9984

21. Hammond MEH, Hayes DF, Dowsett M, et al. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Archives of Pathology & Laboratory Medicine*. 2010;134. doi:10.1043/1543-2165-134.6.907

22. Nicolini A, Ferrari P, Duffy MJ. Prognostic and predictive biomarkers in breast cancer: Past, present and future. *Seminars in Cancer Biology*. 2018;52. doi:10.1016/j.semcancer.2017.08.010

23. Duffy MJ, Harbeck N, Nap M, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the european group on tumor markers (EGTM). *European Journal of Cancer (Oxford, England : 1990)*. 2017;75. doi:10.1016/j.ejca.2017.01.017

24. Colomer R, Aranda-López I, Albanell J, et al. Biomarkers in breast cancer: A consensus statement by the spanish society of medical oncology and the spanish society of pathology. *Clinical & Translational Oncology : Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*. 2018;20. doi:10.1007/s12094-017-1800-5

25. Shen Z, Wu A, Chen X. Current detection technologies for circulating tumor cells. *Chemical Society Reviews*. 2017;46. doi:10.1039/c6cs00803h

26. Alix-Panabières C, Pantel K. Clinical applications of circulating tumor cells and circulating tumor DNA as liquid biopsy. *Cancer Discovery*. 2016;6. doi:10.1158/2159-8290.CD-15-1483

27. Moon DH, Lindsay DP, Hong S, Wang AZ. Clinical indications for, and the future of, circulating tumor cells. *Advanced Drug Delivery Reviews*. 2018;125. doi:10.1016/j.addr.2018.04.002

28. Bidard F-C, Pierga J-Y. Clinical utility of circulating tumor cells in metastatic breast cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2015;33. doi:10.1200/JCO.2014.57.9714

29. Cabel L, Berger F, Cottu P, et al. Clinical utility of circulating tumour cell-based monitoring of late-line chemotherapy for metastatic breast cancer: The randomised CirCe01 trial. *British Journal of Cancer*. 2021;124. doi:10.1038/s41416-020-01227-3

30. Bidard F-C, Fehm T, Ignatiadis M, et al. Clinical application of circulating tumor cells in breast cancer: Overview of the current interventional trials. *Cancer Metastasis Reviews*. 2013;32. doi:10.1007/s10555-012-9398-0

31. Smerage JB, Barlow WE, Hortobagyi GN, et al. Circulating tumor cells and response to chemotherapy in metastatic breast cancer: SWOG S0500. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2014;32. doi:10.1200/JCO.2014.56.2561

32. Banys-Paluchowski M, Krawczyk N, Meier-Stiegen F, Fehm T. Circulating tumor cells in breast cancer–current status and perspectives. *Critical Reviews in Oncology/Hematology*. 2016;97. doi:10.1016/j.critrevonc.2015.10.010

33. Guo T, Stankiewicz E, Mao X, Lu Y-J. The isolation and analysis of circulating tumor cells. *Methods in Molecular Biology (Clifton, NJ)*. 2019;2054. doi:10.1007/978-1-4939-9769-5_7

34. Neumann MHD, Bender S, Krahn T, Schlange T. ctDNA and CTCs in liquid biopsy - current status and where we need to progress. *Computational and Structural Biotechnology Journal*. 2018;16. doi:10.1016/j.csbj.2018.05.002

35. Tashireva LA, Savelieva OE, Grigoryeva ES, et al. Heterogeneous manifestations of epithelial-mesenchymal plasticity of circulating tumor cells in breast cancer patients. *International Journal of Molecular Sciences*. 2021;22. doi:10.3390/ijms22052504

36. Alvarez Cubero MJ, Lorente JA, Robles-Fernandez I, Rodriguez-Martinez A, Puche JL, Serrano MJ. Circulating tumor cells: Markers and methodologies for enrichment and detection. *Methods in Molecular Biology (Clifton, NJ)*. 2017;1634. doi:10.1007/978-1-4939-7144-2_24

37. Yu M, Bardia A, Wittner BS, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science (New York, NY)*. 2013;339. doi:10.1126/science.1228522

38. Clatot F. Review ctDNA and breast cancer. *Recent Results in Cancer Research Fortschritte Der Krebsforschung Progres Dans Les Recherches Sur Le Cancer*. 2020;215. doi:10.1007/978-3-030-26439-0_12

39. Schiavon G, Hrebien S, Garcia-Murillas I, et al. Analysis of ESR1 mutation in circulating tumor DNA demonstrates evolution during therapy for metastatic breast cancer. *Science Translational Medicine*. 2015;7. doi:10.1126/scitranslmed.aac7551

40. Alimirzaie S, Bagherzadeh M, Akbari MR. Liquid biopsy in breast cancer: A comprehensive review. *Clinical Genetics*. 2019;95. doi:10.1111/cge.13514

41. Alix-Panabières C, Pantel K. Liquid biopsy: From discovery to clinical application. *Cancer Discovery*. 2021;11. doi:10.1158/2159-8290.CD-20-1311

42. Ignatiadis M, Lee M, Jeffrey SS. Circulating tumor cells and circulating tumor DNA: Challenges and opportunities on the path to clinical utility. *Clinical Cancer Research : an Official Journal of the American Association for Cancer Research*. 2015;21. doi:10.1158/1078-0432.CCR-14-1190

43. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144. doi:10.1016/j.cell.2011.02.013

44. Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *The New England Journal of Medicine*. 2014;371. doi:10.1056/NEJMoa1409405

45. Lu S-H, Tsai W-S, Chang Y-H, et al. Identifying cancer origin using circulating tumor cells. *Cancer Biology & Therapy*. 2016;17. doi:10.1080/15384047.2016.1141839

46. Wu L, Qu X. Cancer biomarker detection: Recent achievements and challenges. *Chemical Society Reviews*. 2015;44. doi:10.1039/c4cs00370e

47. Duffy MJ. Tumor markers in clinical practice: A review focusing on common solid cancers. *Medical Principles and Practice : International Journal of the Kuwait University, Health Science Centre*. 2013;22. doi:10.1159/000338393

48. Nair M, Sandhu SS, Sharma AK. Cancer molecular markers: A guide to cancer detection and management. *Seminars in Cancer Biology*. 2018;52. doi:10.1016/j.semcancer.2018.02.002

49. Schlumberger M, Leboulleux S. Current practice in patients with differentiated thyroid cancer. *Nature Reviews Endocrinology*. 2021;17. doi:10.1038/s41574-020-00448-z

50. Lamartina L, Grani G, Durante C, Borget I, Filetti S, Schlumberger M. Follow-up of differentiated thyroid cancer - what should (and what should not) be done. *Nature Reviews Endocrinology*. 2018;14. doi:10.1038/s41574-018-0068-3

51. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The diagnosis and management of thyroid nodules: A review. *JAMA*. 2018;319. doi:10.1001/jama.2018.0898

52. Duffy MJ. Biomarkers for prostate cancer: Prostate-specific antigen and beyond. *Clinical Chemistry and Laboratory Medicine*. 2020;58. doi:10.1515/cclm-2019-0693

53. Auprich M, Bjartell A, Chun FK-H, et al. Contemporary role of prostate cancer antigen 3 in the management of prostate cancer. *European Urology*. 2011;60. doi:10.1016/j.eururo.2011.08.003

54. Al Joudi FS. Human mammaglobin in breast cancer: A brief review of its clinical utility. *The Indian Journal of Medical Research*. 2014;139.

55. Span PN, Waanders E, Manders P, et al. Mammaglobin is associated with low-grade, steroid receptor-positive breast tumors from postmenopausal patients, and has independent prognostic value for relapse-free

survival time. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2004;22. doi:10.1200/JCO.2004.01.072

56. Zach O, Lutz D. Mammaglobin remains a useful marker for the detection of breast cancer cells in peripheral blood. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2005;23. doi:10.1200/JCO.2005.05.176

57. Radwan WM, Moussa HS, Essa ES, Kandil SH, Kamel AM. Peripheral blood mammaglobin gene expression for diagnosis and prediction of metastasis in breast cancer patients. *Asia-pacific Journal of Clinical Oncology*. 2013;9. doi:10.1111/j.1743-7563.2012.01556.x

58. Watson MA, Dintzis S, Darrow CM, et al. Mammaglobin expression in primary, metastatic, and occult breast cancer. *Cancer Research*. 1999;59.

59. Liu Z, Yang X, Duan C, et al. Identification and characterization of mammaglobin-a epitope in heterogenous breast cancers for enhancing tumor-targeting therapy. *Signal Transduction and Targeted Therapy*. 2020;5. doi:10.1038/s41392-020-0183-1

60. Gown AM, Fulton RS, Kandalaft PL. Markers of metastatic carcinoma of breast origin. *Histopathology*. 2016;68. doi:10.1111/his.12877

61. Fleming TP, Watson MA. Mammaglobin, a breast-specific gene, and its utility as a marker for breast cancer. *Annals of the New York Academy of Sciences*. 2000;923. doi:10.1111/j.1749-6632.2000.tb05521.x

62. Monsalve-Lancheros A, Ibáñez-Pinilla M, Ramírez-Clavijo S. Detection of mammagloblin by RT-PCR as a biomarker for lymph node metastasis in breast cancer patients: A systematic review and meta-analysis. *Plos One*. 2019;14. doi:10.1371/journal.pone.0216989

63. Hagemann IS, Pfeifer JD, Cao D. Mammaglobin expression in gynecologic adenocarcinomas. *Human Pathology*. 2013;44. doi:10.1016/j.humpath.2012.07.013

64. Bhargava R, Beriwal S, Dabbs DJ. Mammaglobin vs GCDFP-15: An immunohistologic validation survey for sensitivity and specificity. *American Journal of Clinical Pathology*. 2007;127. doi:10.1309/TDP92PQLDE2HLEET

65. O'Brien N, O'Donovan N, Foley D, et al. Use of a panel of novel genes for differentiating breast cancer from non-breast tissues. *Tumour Biology : the Journal of the International Society for Oncodevelopmental Biology and Medicine*. 2007;28. doi:10.1159/000115527

66. Biserni GB, Di Oto E, Moskovszky LE, Foschini MP, Varga Z. Preferential expression of NY-BR-1 and GATA-3 in male breast cancer. *Journal of Cancer Research and Clinical Oncology*. 2018;144. doi:10.1007/s00432-017-2542-z

67. Papatheodorou I, Moreno P, Manning J, et al. Expression atlas update: From tissues to single cells. *Nucleic Acids Research*. 2020;48. doi:10.1093/nar/gkz947

68. The genotype-tissue expression (GTEx) project. *Nature Genetics*. 2013;45. doi:10.1038/ng.2653

69. Alam T, Agrawal S, Severin J, et al. Comparative transcriptomics of primary cells in vertebrates. *Genome Research*. 2020;30. doi:10.1101/gr.255679.119

70. Asmann YW, Necela BM, Kalari KR, et al. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Research*. 2012;72. doi:10.1158/0008-5472.CAN-11-3142

71. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*. 2012;22. doi:10.1101/gr.132159.111

72. Berger AC, Korkut A, Kanchi RS, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*. 2018;33. doi:10.1016/j.ccell.2018.03.014

73. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Research*. 2019;47. doi:10.1093/nar/gkz430

74. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research*. 2017;45. doi:10.1093/nar/gkx247

75. Xu S, Feng Y, Zhao S. Proteins with evolutionarily hypervariable domains are associated with immune response and better survival of basal-like breast cancer patients. *Computational and Structural Biotechnology Journal*. 2019;17. doi:10.1016/j.csbj.2019.03.008

76. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*. 2012;2. doi:10.1158/2159-8290.CD-12-0095

77. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486. doi:10.1038/nature10983

78. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*. 2019;569. doi:10.1038/s41586-019-1186-3

79. Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483. doi:10.1038/nature11003

80. Berglund L, Björling E, Oksvold P, et al. A genecentric human protein atlas for expression profiles based on antibodies. *Molecular & Cellular Proteomics : MCP*. 2008;7. doi:10.1074/mcp.R800013-MCP200

81. Uhlen M, Oksvold P, Fagerberg L, et al. Towards a knowledge-based human protein atlas. *Nature Biotechnology*. 2010;28. doi:10.1038/nbt1210-1248

82. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science (New York, NY)*. 2015;347. doi:10.1126/science.1260419

83. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. *Science (New York, NY)*. 2017;356. doi:10.1126/science.aal3321

84. Uhlen M, Zhang C, Lee S, et al. A pathology atlas of the human cancer transcriptome. *Science (New York, NY)*. 2017;357. doi:10.1126/science.aan2507

85. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Research*. 2018;46. doi:10.1093/nar/gkx1098

86. Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. *Nucleic Acids Research*. 2020;48. doi:10.1093/nar/gkz966

87. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102. doi:10.1073/pnas.0506580102

88. Győrffy B, Lanczky A, Eklund AC, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Research and Treatment*. 2010;123. doi:10.1007/s10549-009-0674-9

89. Ge SX, Jung D, Yao R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics (Oxford, England)*. 2020;36. doi:10.1093/bioinformatics/btz931

90. Hao Z, Lv D, Ge Y, et al. RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *Peerj Computer Science*. 2020;6. doi:10.7717/peerj-cs.251

91. Schober P, Boer C, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*. 2018;126. doi:10.1213/ANE.0000000000002864

92. Kawaji H, Kasukawa T, Forrest A, Carninci P, Hayashizaki Y. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Scientific Data*. 2017;4. doi:10.1038/sdata.2017.113

93. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature Biotechnology*. 2020;38. doi:10.1038/s41587-020-0546-8

94. Smith SE, Mellor P, Ward AK, et al. Molecular characterization of breast cancer cell lines through multiple omic approaches. *Breast Cancer Research : BCR*. 2017;19. doi:10.1186/s13058-017-0855-0

95. Bustin SA, Benes V, Garson JA, et al. The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*. 2009;55. doi:10.1373/clinchem.2008.112797

96. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*. 2013;6. doi:10.1126/scisignal.2004088

97. Matsui A, Ihara T, Suda H, Mikami H, Semba K. Gene amplification: Mechanisms and involvement in cancer. *Biomolecular Concepts*. 2013;4. doi:10.1515/bmc-2013-0026

98. Dam S van, Võsa U, Graaf A van der, Franke L, Magalhães JP de. Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics*. 2018;19. doi:10.1093/bib/bbw139

99. Triska M, Ivliev A, Nikolsky Y, Tatarinova TV. Analysis of cis-regulatory elements in gene co-expression networks in cancer. *Methods in Molecular Biology (Clifton, NJ)*. 2017;1613. doi:10.1007/978-1-4939-7027-8_11

100. Gulìa C, Signore F, Gaffi M, et al. Y RNA: An overview of their role as potential biomarkers and molecular targets in human cancers. *Cancers*. 2020;12. doi:10.3390/cancers12051238

101. Chen C, Li Z, Yang Y, Xiang T, Song W, Liu S. Microarray expression profiling of dysregulated long non-coding RNAs in triple-negative breast cancer. *Cancer Biology & Therapy*. 2015;16. doi:10.1080/15384047.2015.1040957

102. Guo S, Jian L, Tao K, Chen C, Yu H, Liu S. Novel breast-specific long non-coding RNA LINC00993 acts as a tumor suppressor in triple-negative breast cancer. *Frontiers in Oncology*. 2019;9. doi:10.3389/fonc.2019.01325

103. Gloyeske NC, Woodard AH, Elishaev E, et al. Immunohistochemical profile of breast cancer with respect to estrogen receptor and HER2 status. *Applied Immunohistochemistry & Molecular Morphology : AIMM*. 2015;23. doi:10.1097/PAI.0000000000000076

104. Balafoutas D, Hausen A zur, Mayer S, et al. Cancer testis antigens and NY-BR-1 expression in primary breast cancer: Prognostic and therapeutic implications. *BMC Cancer*. 2013;13. doi:10.1186/1471-2407-13-271

105. Theurillat J-P, Zürrer-Härdi U, Varga Z, et al. Distinct expression patterns of the immunogenic differentiation antigen NY-BR-1 in normal breast, testis and their malignant counterparts. *International Journal of Cancer*. 2008;122. doi:10.1002/ijc.23241

106. Jäger D, Filonenko V, Gout I, et al. NY-BR-1 is a differentiation antigen of the mammary gland. *Applied Immunohistochemistry & Molecular Morphology : AIMM*. 2007;15. doi:10.1097/01.pai.0000213111.05108.a0

107. Thu KL, Soria-Bretones I, Mak TW, Cescon DW. Targeting the cell cycle in breast cancer: Towards the next phase. *Cell Cycle (Georgetown, Tex)*. 2018;17. doi:10.1080/15384101.2018.1502567

108. Sadeghi H, Golalipour M, Yamchi A, Farazmandfar T, Shahbazi M. CDC25A pathway toward tumorigenesis: Molecular targets of CDC25A in cell-cycle regulation. *Journal of Cellular Biochemistry*. 2019;120. doi:10.1002/jcb.26838

109. Tu X, Kahila MM, Zhou Q, et al. ATR inhibition is a promising radiosensitizing strategy for triple-negative breast cancer. *Molecular Cancer Therapeutics*. 2018;17. doi:10.1158/1535-7163.MCT-18-0470

110. Ma X, Wang L, Huang D, et al. Polo-like kinase 1 coordinates biosynthesis during cell cycle progression by directly activating pentose phosphate pathway. *Nature Communications*. 2017;8. doi:10.1038/s41467-017-01647-5

111. Yousef EM, Furrer D, Laperriere DL, et al. MCM2: An alternative to ki-67 for measuring breast cancer cell proliferation. *Modern Pathology : an Official Journal of the United States and Canadian Academy of Pathology, Inc*. 2017;30. doi:10.1038/modpathol.2016.231

112. Issac MSM, Yousef E, Tahir MR, Gaboury LA. MCM2, MCM4, and MCM6 in breast cancer: Clinical utility in diagnosis and prognosis. *Neoplasia (New York, NY)*. 2019;21. doi:10.1016/j.neo.2019.07.011

113. Abe S, Yamamoto K, Kurata M, et al. Targeting MCM2 function as a novel strategy for the treatment of highly malignant breast tumors. *Oncotarget*. 2015;6. doi:10.18632/oncotarget.5408

114. Lacroix M. Significance, detection and markers of disseminated breast cancer cells. *Endocrine-related Cancer*. 2006;13. doi:10.1677/ERC-06-0001

115. Li J, Mahajan A, Tsai M-D. Ankyrin repeat: A unique motif mediating protein-protein interactions. *Biochemistry*. 2006;45. doi:10.1021/bi062188q

116. Sedgwick SG, Smerdon SJ. The ankyrin repeat: A diversity of interactions on a common structural framework. *Trends in Biochemical Sciences*. 1999;24. doi:10.1016/s0968-0004(99)01426-7

117. Mosavi LK, Cammett TJ, Desrosiers DC, Peng Z-Y. The ankyrin repeat as molecular architecture for protein recognition. *Protein Science : a Publication of the Protein Society*. 2004;13. doi:10.1110/ps.03554604

118. Kosaloglu Z, Bitzer J, Halama N, et al. In silico SNP analysis of the breast cancer antigen NY-BR-1. *BMC Cancer*. 2016;16. doi:10.1186/s12885-016-2924-7

119. Jäger D, Stockert E, Güre AO, et al. Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library. *Cancer Research*. 2001;61.

120. Jäger D, Unkelbach M, Frei C, et al. Identification of tumor-restricted antigens NY-BR-1, SCP-1, and a new cancer/testis-like antigen NW-BR-3 by serological screening of a testicular library with breast cancer serum. *Cancer Immunity*. 2002;2.

121. Jäger D. Potential target antigens for immunotherapy identified by serological expression cloning (SEREX). *Methods in Molecular Biology (Clifton, NJ)*. 2007;360. doi:10.1385/1-59745-165-7:319

122. Jäger D, Taverna C, Zippelius A, Knuth A. Identification of tumor antigens as potential target antigens for immunotherapy by serological expression cloning. *Cancer Immunology, Immunotherapy : CII*. 2004;53. doi:10.1007/s00262-003-0470-z

123. Zach O, Kasparu H, Krieger O, Hehenwarter W, Girschikofsky M, Lutz D. Detection of circulating mammary carcinoma cells in the peripheral blood of breast cancer patients via a nested reverse transcriptase polymerase chain reaction assay for mammaglobin mRNA. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 1999;17. doi:10.1200/JCO.1999.17.7.2015

124. Sangoi AR, Shrestha B, Yang G, Mego O, Beck AH. The novel marker GATA3 is significantly more sensitive than traditional markers mammaglobin and GCDFP15 for identifying breast cancer in surgical and cytology specimens of metastatic and matched primary tumors. *Applied Immunohistochemistry & Molecular Morphology : AIMM*. 2016;24. doi:10.1097/PAI.0000000000000186

125. Tang T, Zhang L, Li C, Zhou T. Gastric and adrenal metastasis from breast cancer: Case report and review of literature. *Medicine*. 2020;99. doi:10.1097/MD.0000000000018812

126. Dulskas A, Al Bandar M, Choi YY, et al. A case of gastric cancer metastasis to the breast in a female with BRCA2 germline mutation and literature review. *Acta Chirurgica Belgica*. 2019;119. doi:10.1080/00015458.2017.1411554

127. Maeshima Y, Osako T, Morizono H, et al. Metastatic ovarian cancer spreading into mammary ducts mimicking an in situ component of primary breast cancer: A case report. *Journal of Medical Case Reports*. 2021;15. doi:10.1186/s13256-020-02653-w

128. Wong YP, Tan GC, Muhammad R, Rajadurai P. Occult primary breast carcinoma presented as an axillary mass: A diagnostic challenge. *The Malaysian Journal of Pathology*. 2020;42.

129. Liu Y-F, Liu L-Y, Xia S-L, Li T, Li J. An unusual case of scalp metastasis from breast cancer. *World Neurosurgery*. 2020;137. doi:10.1016/j.wneu.2020.01.230

130. Bishop JA, Yonescu R, Batista D, Begum S, Eisele DW, Westra WH. Utility of mammaglobin immunohistochemistry as a proxy marker for the ETV6-NTRK3 translocation in the diagnosis of salivary mammary analogue secretory carcinoma. *Human Pathology*. 2013;44. doi:10.1016/j.humpath.2013.03.017

131. Skalova A, Michal M, Simpson RH. Newly described salivary gland tumors. *Modern Pathology : an Official Journal of the United States and Canadian Academy of Pathology, Inc*. 2017;30. doi:10.1038/modpathol.2016.167

132. Williamson SR. Clear cell papillary renal cell carcinoma: An update after 15 years. *Pathology*. 2021;53. doi:10.1016/j.pathol.2020.10.002

133. Leivo MZ, Tacha DE, Hansel DE. Expression of uroplakin II and GATA-3 in bladder cancer mimickers: Caveats in the use of a limited panel to determine cell of origin in bladder lesions. *Human Pathology*. Published online 2021. doi:10.1016/j.humpath.2021.04.005

134. Goto K, Ishikawa M, Hamada K, et al. Comparison of immunohistochemical expression of cytokeratin 19, c-KIT, BerEP4, GATA3, and NUTM1 between porocarcinoma and squamous cell carcinoma. *The American Journal of Dermatopathology*. Published online 2021. doi:10.1097/DAD.0000000000001901

135. McDonald TM, Epstein JI. Aberrant GATA3 staining in prostatic adenocarcinoma: A potential diagnostic pitfall. *The American Journal of Surgical Pathology*. 2021;45. doi:10.1097/PAS.0000000000001557

136. Woodard AH, Yu J, Dabbs DJ, et al. NY-BR-1 and PAX8 immunoreactivity in breast, gynecologic tract, and other CK7+ carcinomas: Potential use for determining site of origin. *American Journal of Clinical Pathology*. 2011;136. doi:10.1309/AJCPUFNMEZ3MK1BK

137. Jiang Y, Harlocker SL, Molesh DA, et al. Discovery of differentially expressed genes in human breast cancer using subtracted cDNA libraries and cDNA microarrays. *Oncogene*. 2002;21. doi:10.1038/sj.onc.1205278

138. Seil I, Frei C, Sültmann H, et al. The differentiation antigen NY-BR-1 is a potential target for antibody-based therapies in breast cancer. *International Journal of Cancer*. 2007;120. doi:10.1002/ijc.22620

139. Shen L, Kondo Y, Guo Y, et al. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *Plos Genetics*. 2007;3. doi:10.1371/journal.pgen.0030181

140. Varga Z, Theurillat J-P, Filonenko V, et al. Preferential nuclear and cytoplasmic NY-BR-1 protein expression in primary breast cancer and lymph node metastases. *Clinical Cancer Research : an Official Journal of the American Association for Cancer Research*. 2006;12. doi:10.1158/1078-0432.CCR-05-2192

141. Theurillat J-P, Zürrer-Härdi U, Varga Z, et al. NY-BR-1 protein expression in breast carcinoma: A mammary gland differentiation antigen as target for cancer immunotherapy. *Cancer Immunology, Immunotherapy : CII*. 2007;56. doi:10.1007/s00262-007-0316-1

142. Abu El-Nazar SY, Ghazy AA, Ghoneim HE, et al. NY-BR-1 antigen expression and anti-NY-BR-1 IgG in egyptian breast cancer patients: Clinicopathological and prognostic significance. *The Egyptian Journal of Immunology*. 2015;22.

143. Wang Y, Li H, Ma J, et al. Integrated bioinformatics data analysis reveals prognostic significance of SIDT1 in triple-negative breast cancer. *Oncotargets and Therapy*. 2019;12. doi:10.2147/OTT.S215898

144. Ballman KV. Biomarker: Predictive or prognostic? *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2015;33. doi:10.1200/JCO.2015.63.3651

145. Sparano JA, Gray RJ, Makower DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *The New England Journal of Medicine*. 2018;379. doi:10.1056/NEJMoa1804710

146. Cardoso F, Veer LJ van't, Bogaerts J, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *The New England Journal of Medicine*. 2016;375. doi:10.1056/NEJMoa1602253

147. Qian Y, Daza J, Itzel T, et al. Prognostic cancer gene expression signatures: Current status and challenges. *Cells*. 2021;10. doi:10.3390/cells10030648

148. Mamounas EP, Mitchell MP, Woodward WA. Molecular predictive and prognostic markers in locoregional management. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2020;38. doi:10.1200/JCO.19.02905

149. Giger OT, Lacoste E, Honegger C, Padberg B, Moch H, Varga Z. Expression of the breast differentiation antigen NY-BR-1 in a phyllodes tumor of the vulva. *Virchows Archiv : an International Journal of Pathology*. 2007;450. doi:10.1007/s00428-007-0377-8

150. Giger O, Caduff R, O'Meara A, et al. Frequent expression of the breast differentiation antigen NY-BR-1 in mammary and extramammary paget's disease. *Pathology International*. 2010;60. doi:10.1111/j.1440-1827.2010.02591.x

151. Wallwiener CW, Wallwiener M, Kurth RR, et al. Molecular detection of breast cancer metastasis in sentinel lymph nodes by reverse transcriptase polymerase chain reaction (RT-PCR): Identifying, evaluating and establishing multi-marker panels. *Breast Cancer Research and Treatment*. 2011;130. doi:10.1007/s10549-011-1710-0

152. Nissan A, Jager D, Roystacher M, et al. Multimarker RT-PCR assay for the detection of minimal residual disease in sentinel lymph nodes of breast cancer patients. *British Journal of Cancer*. 2006;94. doi:10.1038/sj.bjc.6602992

153. Zehentner BK, Dillon DC, Jiang Y, et al. Application of a multigene reverse transcription-PCR assay for detection of mammaglobin and complementary transcribed genes in breast cancer lymph nodes. *Clinical Chemistry*. 2002;48.

154. Zehentner BK, Secrist H, Hayes DC, et al. Detection of circulating tumor cells in peripheral blood of breast cancer patients during or after therapy using a multigene real-time RT-PCR assay. *Molecular Diagnosis & Therapy*. 2006;10. doi:10.1007/BF03256441

155. Reinholz MM, Nibbe A, Jonart LM, et al. Evaluation of a panel of tumor markers for molecular detection of circulating cancer cells in women with suspected breast cancer. *Clinical Cancer Research : an Official Journal of the American Association for Cancer Research*. 2005;11. doi:10.1158/1078-0432.CCR-04-1483

156. Agelopoulos K, Richter GHS, Schmidt E, et al. Deep sequencing in conjunction with expression and functional analyses reveals activation of FGFR1 in ewing sarcoma. *Clinical Cancer Research : an Official Journal of the American Association for Cancer Research*. 2015;21. doi:10.1158/1078-0432.CCR-14-2744

157. Li M, Chen Y, Yao J, et al. Genome-wide association study of smoking behavior traits in a chinese han population. *Frontiers in Psychiatry*. 2020;11. doi:10.3389/fpsyt.2020.564239

158. Chen K, Ma H, Li L, et al. Genome-wide association study identifies new susceptibility loci for epithelial ovarian cancer in han chinese women. *Nature Communications*. 2014;5. doi:10.1038/ncomms5682

159. Lambert SA, Jolma A, Campitelli LF, et al. The human transcription factors. *Cell*. 2018;172. doi:10.1016/j.cell.2018.01.029

160. Todeschini A-L, Georges A, Veitia RA. Transcription factors: Specific DNA binding and specific gene regulation. *Trends in Genetics : TIG*. 2014;30. doi:10.1016/j.tig.2014.04.002

161. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics*. 2000;25. doi:10.1038/75556

162. The gene ontology resource: Enriching a GOld mine. *Nucleic Acids Research*. 2021;49. doi:10.1093/nar/gkaa1113

163. Zombori T, Cserni G. Immunohistochemical analysis of the expression of breast markers in basal-like breast carcinomas defined as triple negative cancers expressing keratin 5. *Pathology Oncology Research : POR*. 2018;24. doi:10.1007/s12253-017-0246-y

164. Sejben A, Vörös A, Golan A, Zombori T, Cserni G. The added value of SOX10 immunohistochemistry to other breast markers in identifying cytokeratin 5-positive triple negative breast cancers as of mammary origin. *Pathobiology : Journal of Immunopathology, Molecular and Cellular Biology*. Published online 2021. doi:10.1159/000512006

165. Bird A. DNA methylation patterns and epigenetic memory. *Genes & Development*. 2002;16. doi:10.1101/gad.947102

166. Lentini A, Nestor CE. Mapping DNA methylation in mammals: The state of the art. *Methods in Molecular Biology (Clifton, NJ)*. 2021;2198. doi:10.1007/978-1-0716-0876-0_4

167. Bates SE. Epigenetic therapies for cancer. *The New England Journal of Medicine*. 2020;383. doi:10.1056/NEJMra1805035

168. Vanneman M, Dranoff G. Combining immunotherapy and targeted therapies in cancer treatment. *Nature Reviews Cancer*. 2012;12. doi:10.1038/nrc3237

169. Jäger D, Karbach J, Pauligk C, et al. Humoral and cellular immune responses against the breast cancer antigen NY-BR-1: Definition of two HLA-A2 restricted peptide epitopes. *Cancer Immunity*. 2005;5.

170. Zhao Z, Ukidve A, Kim J, Mitragotri S. Targeting strategies for tissue-specific drug delivery. *Cell*. 2020;181. doi:10.1016/j.cell.2020.02.001

171. Ryaboshapkina M, Hammar M. Tissue-specific genes as an underutilized resource in drug discovery. *Scientific Reports*. 2019;9. doi:10.1038/s41598-019-43829-9

172. Das K, Eisel D, Vormehr M, et al. A transplantable tumor model allowing investigation of NY-BR-1-specific t cell responses in HLA-DRB1*0401 transgenic mice. *BMC Cancer*. 2019;19. doi:10.1186/s12885-019-6102-6

173. Wang W, Epler J, Salazar LG, Riddell SR. Recognition of breast cancer cells by CD8+ cytotoxic t-cell clones specific for NY-BR-1. *Cancer Research*. 2006;66. doi:10.1158/0008-5472.CAN-05-3529

174. Jäger D, Knuth A. Antibodies and vaccines–hope or illusion? *Breast (Edinburgh, Scotland)*. 2005;14. doi:10.1016/j.breast.2005.08.029

175. Meyerson NR, Rowley PA, Swan CH, Le DT, Wilkerson GK, Sawyer SL. Positive selection of primate genes that promote HIV-1 replication. *Virology*. 2014;454-455. doi:10.1016/j.virol.2014.02.029

176. Li YR, Li J, Zhao SD, et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine*. 2015;21. doi:10.1038/nm.3933

6 Appendix

6.1 List of Figures

6.2 List of Tables

Acknowledgements

First and foremost I am extremely grateful to my supervisor, Prof. Dr. Hans Neubauer, for his invaluable advice, continuous support, and patience during my study. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. Additionally, I would like to express my sincere gratitude to my co-supervisor Prof. Dr. Georg Pongratz and Prof. Dr. Fehm for guiding and supporting me well throughout the research work. I would also like to extend my deepest gratitude to all the members of our lab for their technical support and encouragement. Their friendship and the warmth they extended to me always making me feel so welcome. I'm deeply grateful to receive the scholarship from the China Scholarship Council (CSC). It wouldn't have been possible to conduct my research in Germany without their financial support. Finally, I gratefully acknowledge my parents and my wife. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.