UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato di ricerca in Oncologia e Chirurgia Sperimentali
Dipartimento di Medicina di Precisione in Area Medica, Chirurgica e Critica (Me.Pre.C.C.)

Dissecting the transcriptome of CD44 in triple negative breast cancer

Doctoral Dissertation of:
Davide Vacca

Tutor:
Prof. Claudio Tripodo

The Chair of the Doctoral Program:
Prof. Antonio Russo

2023-2024 – Cycle XXXVI

Abstract

Background: The CD44 receptor is a transmembrane glycoprotein involved in the recognition of specific extracellular matrix (ECM) ligands. It has a complex transcriptional regulation characterized by different splicing variants, many of which play a role in promoting tumorigenesis. Therefore, CD44 variants may serve as prognostic biomarkers and therapeutic targets.

Triple-negative breast cancer (TNBC) is a poor prognosis subtype of breast adenocarcinomas characterized by the lack of druggable molecular hallmarks. TNBC represents a prototypical model for studying CD44's pro-tumoral activity. However, the regulation of the CD44 transcriptome and its influence on TNBC aggressiveness remains elusive.

In this study, we dissected the CD44 transcriptome in the TNBC cell lines MDA-MB-231 and MDA-MB-468, which exhibit different malignant properties. MDA-MB-231 cells display higher extracellular matrix invasive potential, more prominent metastatic potential, and poorer response to chemotherapy compared to MDA-MB-468 cells.

Finally, we induced the overexpression of a differentially expressed CD44 isoform between the two cell lines using transfection assays. We then performed whole transcriptome RNA-sequencing analysis of the transfected cells and their control.

Methods: On RNA extracted from two TNBC cell lines and the non-tumoral epithelial cell line MCF10A used as a control, we conducted a reverse transcription targeting polyA-containing transcripts, followed by long-PCR amplification with primers specific for highly conserved regions between CD44 mRNA sequences. The resulting cDNA was then sequenced using Oxford Nanopore Technology (ONT), and the obtained reads were aligned to the CD44 reference sequence.

Furthermore, we transfected the TNBC cell line MDA-MB-468 with a CD44 isoform (XM_017018585.2) that was found to be differentially expressed between the two TNBC lines. We then performed whole transcriptome sequencing analysis on the transfected cells.

Results: We identified distinct patterns in the CD44 transcriptome between the two TNBC cell lines. Specifically, among the top 5 differentially expressed (DE) CD44 isoforms compared to the control cell line, we detected the predicted transcript variant XM_017018585.2. This variant will be further confirmed through PCR and Sanger sequencing assays. To functionally characterize XM_017018585.2, we performed overexpression of the exogenous transcript in the MDA-MB-468 cell line using viral transfection. The overexpression of XM_017018585.2 induced upregulation of transcripts involved in the MAPK8 signaling pathway and downregulation of interferon response programs. Interestingly, the same downregulated genes were also observed in the estrogen-responsive breast cancer cell line MCF-7 upon silencing of specific pathways.

Conclusions: The CD44 transcriptome analysis of TNBC cell models revealed novel transcriptional variants with potential significance in the modulation of the malignant phenotype. This study could shed light on the role of previously unidentified CD44 isoforms in promoting tumorigenesis. Furthermore, it could uncover new candidate diagnostic and prognostic biomarkers based on CD44 variant expression.

Summary

Introduction

The human CD44 gene, also known by various aliases including IN, LHR, MC56, MDU2, MDU3, MIC4, Pgp1, CDW44, CSPG8, H-CAM, HCELL, ECM-III, HUTCH-1, HUTCH-I, ECMR-III, and HERMES-1, is a single, highly conserved gene spanning 93,232 base pairs (bp). It is located on chromosome 11p13 (Chromosome 11: 35,138,882-35,232,402, forward strand) and encodes a single-chain transmembrane glycoprotein/part-time proteoglycan classified as a cell adhesion molecule (CAM) [1,2].

In the mouse genome, the orthologous protein is encoded by the Cd44 gene, also known as Ly-24, Pgp-1, and Hermes. This gene spans 90,529 bp and is located on chromosome 2q (Chromosome 2: 102,641,486-102,732,010, reverse strand) [2,3].

Both the human and mouse CD44 genes consist of 19 coding exons. In humans, these exons span 90,030 bp, while in mice, they span 87,214 bp, as illustrated in Figure 1 A-B. The coordinates of each exon in the human and mouse reference genomes are provided in Supplementary Tables 1A and 1B, respectively. Notably, in the human genome, the ENSE00001598274 exon is exclusively present in the ENST00000442151.6 transcript, where it is both the ninth and final coding exon, as shown in Figure 1 A.

Within this exon, only the initial four nucleotides "5'UUGA3'" participate in protein translation. Specifically, the first "U" pairs with the final two nucleotides of the eighth exon, forming the AGU codon, which encodes Ser-294 of the CD44 antigen isoform H0Y5E4. The terminal three nucleotides "UGA" constitute a stop codon in the messenger RNA [2,4], as illustrated in Figure 1 C.

CD44 expression is initiated during embryonic development at day 9.5 post-coitum, with detectable expression already present in the egg cylinder by day 6.5. During these early stages, a highly specialized expression pattern encompassing numerous CD44 isoforms is observed [5]. In adult tissues, CD44 expression is more confined, appearing predominantly in specific cell types during particular stages of activation, maturation, or development. These include macrophages, activated lymphocytes, keratinocytes, and certain epithelial cells found in the stomach, bladder, and uterine cervix. CD44 plays a crucial role in the specialized biological functions of these cell types, participating in various processes such as lymphocyte homing, cell adhesion and aggregation, cell migration, leukocyte activation, lymphopoiesis, myelopoiesis, angiogenesis, and cytokine release [3].

Significant levels of CD44 expression are also detected in the respiratory system, gastrointestinal tract, and kidney, as depicted in Figure 2A. When comparing male and female tissues, females exhibit higher CD44 protein expression levels than males, particularly in the endometrium, vagina, breast, and cervix.

In males, the prostate is the only tissue with expression levels comparable to those mentioned above, as shown in Figure 2B.

Analysis of the transcriptome using single-cell RNA sequencing reveals that CD44 mRNA expression predominantly clusters with cells originating from the blood and immune systems, as illustrated in Figure 3.

The regulation of CD44 expression is orchestrated by a variety of transcription factors, protein kinases, cytokines, epigenetic mechanisms, and microRNAs (miRNAs). These factors can act as either repressors, exemplified by ZMYND8, or activators, such as NUP98-HOXA9, SNAI2, F2R, F2RL1, NOTCH1, TWIST1, and TWIST2 [6]. Additional regulators are depicted in Figure 4.

Furthermore, cytokines and growth factors can stimulate signaling cascades pivotal for activating transcription factors that enhance CD44 promoter activity, as shown in Figure 5. For instance, treatment with IL6 and IGF1 leads to increased CD44 promoter activity [7].

CD44 gene expression has been extensively studied in the context of cancer. While the role of CD44 in tumors remains incompletely defined, accumulating evidence links CD44 expression with aggressive biological behavior and poor prognosis in various malignant tumors [3]. Elevated levels of CD44 mRNA have been documented in a range of cancers, including lymphoma, breast, colon, and endometrial cancer [8].

For instance, lung adenocarcinoma cells exhibit heightened expression of various CD44 isoforms, which correlates with enhanced characteristics of cancer stem cells (CSCs), increased proliferation, and resistance to chemotherapeutic agents [9]. Specific variant transcripts, notably variant 6 (CD44V6), have been closely associated with metastatic behavior in pancreatic carcinoma cells [10]. In both mouse and human pancreatic ductal adenocarcinoma, CD44 expression is upregulated by activation of the β-catenin signaling pathway, leading to increased expression of Zeb1 and Snail1 and subsequent induction of an epithelial-to-mesenchymal transition (EMT) phenotype [11].

In human colorectal adenocarcinoma cells, CD44 expression is upregulated by Wnt, β-catenin, and Tcf-4 factors. In lung carcinoma cells, Godar et al. demonstrated that the inhibition of CD44 expression results from the binding of a noncanonical p53 sequence to its promoter. Indeed, knocking out p53 in these cells led to elevated CD44 levels and increased resistance to apoptosis [12].

In human breast cancer cells, elevated protein synthesis levels correlate with the activity of the transcription factor NF-κB, particularly in cells exhibiting a triple-negative phenotype. This regulation is strengthened by the interaction of NF-κB with the conserved cis-regulatory element, known as the conserved region 1 (CR1), located upstream of the CD44 promoter. Inhibition of NF-κB leads to a

decrease in CD44 expression [3; 13]. Conversely, the Forkhead box protein 3 (Foxp3) transcription factor has been shown to negatively regulate CD44 expression, thereby suppressing invasion and metastatic capabilities [14]. Upregulation of CD44 expression has also been associated with activation of the β-catenin and AKT signaling pathways in both breast and cervical cancer cells. In contrast, mRNA levels were significantly reduced in cells where these pathways were knocked out or inhibited [15]. Additionally, transforming growth factor-beta 1 (TGF-β1) enhances CD44 expression in prostate cancer cells [16].

Other studies have highlighted the role of microRNAs (miRNAs) in regulating CD44 expression in cancer. For instance, overexpression of miR-328 reduced CD44 mRNA expression in gastrointestinal cancer cells, leading to inhibited cancer cell growth and increased sensitivity to chemotherapy and reactive oxygen species (ROS). Conversely, treatment with miR-328 inhibitors resulted in enhanced growth of gastrointestinal cancer cells [17].

In prostate CD44+ Cancer Stem Cells (CSCs), both miR-34a and miR-141 have been shown to suppress tumor growth and metastasis by targeting the CD44 gene [18, 19]. CD44 was also identified as a direct molecular target of miR-520b, which interferes with the tumorigenesis process of head and neck cancer by regulating cancer stemness conversion [20].

In ovarian cancer-initiating cells, miR-199a-mediated targeting of CD44 suppresses tumorigenesis and multidrug resistance, while miR-143 inhibits progression and stemness features in breast cancer cells [21, 22].

Conversely, a positive correlation has been observed between CD44 and miR-221 expression in hepatocellular carcinoma cells and tumors. Inhibition of miR-221 leads to decreased CD44 protein expression, whereas miR-221 mimic enhances protein levels [23].

FUNCTIONAL STRUCTURES OF THE CD44 RECEPTOR

The CD44 antigen is a cell-surface protein identified as the primary cell surface receptor for hyaluronate, a major component of the extracellular matrix (ECM) [3]. As a member of the CAMs family, CD44 plays crucial roles in cellular communication and adhesion between cells and the ECM [24]. Functionally, the CD44 receptor can be divided into three characteristic domains:

Extracellular region: This region, also known as the ectodomain, comprises seven extracellular domains, including an N-terminal signal sequence. These domains are involved in the formation of the link-homology hyaluronan-binding module, or ligand-binding region. This region is encoded by exons 1

to 5 and 16 to 17. The ectodomain facilitates the transduction of external signals from the microenvironment by binding to various ECM components such as hyaluronic acid (HA), collagen, osteopontin, growth factors, cytokines, and proteases. Within this domain, four overlapping regions exist: the HA binding domain, the Link motif which contains the HA and glycosaminoglycan binding sites, a basic motif and a stem structure. The stability and folding of the Link motif depend on interchain disulfide links in four highly conserved cysteine residues and two additional flanking cysteine residues.

The stem region is of variable length among isoforms. It is crucial for the conformational changes at the protein binding sites, thereby influencing the biological activity of CD44 [25, 26]. This structure is absent in the CD44std protein only.

Transmembrane region (TM): This hydrophobic single-pass domain is encoded by exon 18 and anchors the receptor to the cell membrane.

Intracellular cytoplasmic domain (ICD): It refers to the short C-terminal tail of CD44. It is derived from the translation of exons 19, which contain codifying regions of varying dimensions, as observed in different CD44 isoforms. The ICD interacts with various intracellular signaling molecules to transmit signals from the extracellular matrix (ECM) [27].

The exons encoding for the three HA binding domain, Link and basic motifs ectodomain regions and those translating for TM region and ICD tail, are essential for the biological function of the receptor and are therefore present in all CD44 protein-coding transcripts, thus termed as constant exons.

Conversely, the stem region encodes a juxtamembrane domain that regulates conformational changes at the protein binding sites, influencing CD44's biological activity [26, 28, 1]. This domain is frequently exposed to alternative splicing events, resulting in different exon patterns within the constant exons 5 and 16. For immature CD44 mRNA, 10 exons from 6 to 15 can undergo splicing, termed variable exons v1 to v10 [26, 28].

Each CD44 isoform is encoded by a specific mRNA pattern comprising the nine constant exons and one or more variable exons (see Figure 6A). The most common isoform, known as CD44 standard (CD44std), has a molecular weight of 85–95 kDa. This isoform lacks all 10 variable exons (6 to 15) due to splicing events. Consequently, exons 1-5 and 16-19 are referred to as stable (or standard) and enumerated as s1 to s5 and s6 to s9, respectively [26, 28].

Orthotopic transplantation experiments in mice have shown that Epithelial Splicing Regulatory Protein 1 (ESRP1) is crucial for maintaining variable exons. In these experiments, transplantation of 4T1 breast cancer cells expressing various CD44v isoforms resulted in efficient lung metastasis dependent on the

activity of the cystine-glutammate transporter xCT. CD44v isoforms control the stability of xCT. Knocking down ESRP1 in CD44v-expressing cells led to an isoform switch to CD44s. This has reduced the xCT cell surface expression, suppressing lung colonization [7; 29].

Ectodomain Modifications.

The highly regulated post-transcriptional modifications, driven by alternative splicing mechanisms, underscore the intricate regulatory landscape governing the CD44 receptor.

In addition to these post-transcriptional modifications, CD44 undergoes various post-translational modifications (PTMs). These PTMs include variable glycosylation with O-glycans, N-glycans, and glycosaminoglycans (GAGs) such as chondroitin sulfate and heparan sulfate [30]. Due to these side-chain attachments, the typical size of the CD44 core protein (37 kDa) expands to 80–100 kDa for some isoforms, with heavily glycosylated forms exceeding 200 kDa [31]. Each PTM contributes to the maturation of the CD44 antigen receptor, imparting it with specific biological functions [7].

The majority of potential N-linked glycosylation sites (Asn residues) are located at the extreme N-terminal of the extracellular domain. In contrast, potential O-linked glycosylation sites (Ser/Thr residues) and GAG attachment sites (Ser-Gly motifs) are found in more internal regions, including the stem and membrane-proximal domains [32].

This extensive post-transcriptional and post-translational maturation is further complicated by the observation that tumor cells may express multiple CD44v isoforms, each at varying levels of expression. This diversity adds another layer of complexity to the regulatory mechanisms governing CD44, underscoring its multifaceted role in cellular physiology and pathology [7].

Transmembrane Domin and ICD interactions.

The CD44 transmembrane domain consists of a single-pass TM domain comprising 23 amino acids. This domain facilitates CD44 oligomerization and its recruitment into glycolipid-enriched membrane domains (GEMs). GEMs play a pivotal role in mediating the interaction of the membrane receptor with extracellular ligands and in association with other transmembrane and cytoplasmic molecules. Notably, the intramembrane domain contains highly conserved amino acids, such as Cys286 and Cys295, which can undergo reversible palmitoylation. This modification enhances the localization of CD44 to cellular membranes and promotes its interactions with cell surface and cytoplasmic proteins. Among the proteins that interact with CD44 in lipid rafts are various receptor tyrosine kinases (RTKs), innate receptors (TLRs), ABC transporters, EMMPRIN, PI3K, Src kinases, ezrin, and hyaluronidase 2.

Moreover, the transmembrane domain is susceptible to specific proteolytic cleavage by enzymes like membrane type 1 matrix metalloprotease (MT1-MMP) and presenilin-1/γ-secretase. This results in the release of a CD44β-like peptide into the extracellular matrix (ECM) and it has also effect on the ICD conformation. ICD is a 72-amino-acid cytoplasmic tail of CD44 receptor, which following to presenilin activity generates a free cytoplasmic peptide (ICD_f), which can have either a short or long-tail configuration, as will be described in the next paragraph.

Although the CD44 ICD lacks enzymatic activity, it plays a crucial role in signaling pathways by interacting with various kinases and other molecules involved in multiple signaling cascades [1].

ICD can be divided into four structural motifs. These motifs facilitate interactions with the cytoskeleton, suggesting a role for CD44 in regulating actin cytoskeleton reorganization (see Figure 6B). Consistent with this, CD44 exhibits preferential localization in actin polymerization regions like lamellipodia, filopodia, and apical microvilli in migrating cells. Among these structural motifs, two positively charged amino acid clusters (292RRRCGQKKK300) constitute the FERM-binding domain. This domain facilitates interactions between CD44 ICD and ERM (ezrin/radixin/moesin) cytoskeletal proteins as well as merlin/NF2 [33, 34]. Following this, the ankyrin-binding domain (304NSGNGAVEDRKPSGL318) allows CD44 ICD to interact with ankyrin, promoting CD44 clustering on the membrane and facilitating its interaction with E-selectin on activated endothelial cells. This interaction is crucial for neutrophil tethering, subsequent rolling arrest, and emigration into perivascular tissues [1]. Towards the C-terminal end, a dihydrophobic basolateral targeting motif (331LV332) and a PDZ (PSD-95/Dlg/ZO-1)-domain-binding peptide represented by the four amino acids 358KIGV361 are present [1], as shown in Figure 7.

The ERM proteins and merlin/NF2 contain specific domains for binding cortical actin filaments within their conserved three-lobe N-terminal FERM domain (head). This domain represents the region of highest homology between these two proteins.

Upon phosphorylation, ERM proteins transition from an inactive state characterized by head-to-tail self-association to an active state where the three-lobed structure dissociates from the C-terminal domain. Activated ERM binds to the FERM-binding domain of CD44 ICD phosphorylated at Ser325, a site constitutively phosphorylated by CaMKII. This interaction leads to a dynamic association between CD44 and the cytoskeleton [35]. However, complete dephosphorylation of Ser325, followed by phosphorylation of other serine residues such as Ser291 and Ser316 in the FERM-binding domain upon PKC signaling activation, results in the dissociation of CD44 from the ezrin complex and its detachment from the actin cytoskeleton [1].

In contrast, merlin/NF2 binds to CD44 in its dephosphorylated state. Phosphorylation of merlin by various Ser/Thr kinases, including PKA, Akt, and p21-activated kinase (PAK), leads to the inactivation of its tumor-suppressing activity [36].

Consequently, ERM proteins and merlin/NF2 may employ distinct regulatory mechanisms in actin cytoskeleton organization, acting as cross-linkers between the CD44 membrane protein and F-actin [1]. Experiments conducted under high-density conditions in cell cultures reveal hypophosphorylation of both ERM proteins and merlin. Under these circumstances, ERM proteins remain inactive, while merlin adopts an active, unphosphorylated closed conformation. Activated merlin inhibits the proteolytic processing of the CD44 extracellular domain, thereby preserving cell density signaling, suppressing the ras-mediated pathway, and inhibiting cell proliferation and migration [37]. Conversely, reduced cell density leads to merlin inactivation through PAK-induced phosphorylation. Inactivated merlin dissociates from CD44 ICD, thereby negating the growth-suppressive effects of merlin [1].

CD44 ICD also enhances inhibition of contact processes through the Hippo signaling pathway in two ways. Firstly, CD44 ICD indirectly promotes Hippo signaling through its interaction with merlin. Merlin's association with the cytoplasmic domain of CD44 facilitates the recruitment of LATS, a core kinase in the Hippo pathway, thereby stimulating Hippo signaling [38]. Secondly, CD44 ICD directly interacts with the PAR1b/microtubule affinity-regulating kinase 2 (MARK2) complex due to receptor clustering induced by HMW-Hyaluronic acid. The MARK2 complex typically inactivates MST1 and MST2 kinases, leading to Hippo signaling activation and contributing to contact inhibition of cell growth [39].

The CD44 ICD contains seven serine residues (at positions 291, 305, 316, 323, 325, 336, and 337) that could potentially serve as phosphorylation targets. However, only Ser291, Ser316, and Ser325 undergo phosphorylation. Mutagenesis studies have shown that Ser325 is the primary site of constitutive CD44 phosphorylation, occurring on approximately one-third of CD44 molecules [40]. This phosphorylation is mediated by Ca2+/calmodulin-dependent protein kinase II (CaMKII), which is activated by increased cytosolic Ca2+ levels. Elevated cytoplasmic Ca2+ concentrations, resulting from ECM influx and/or intracellular store release, activate CaMKII, leading to CD44 ICD phosphorylation at Ser325 [1].

ROLE OF CD44 ICD

Functionally, the CD44 ICD exhibits diverse mechanisms to exert its biological roles.

Firstly, when it is anchored at the plasmatic membrane can act as an intracellular scaffold for proteins already bound to the receptor or, when localized to lipid rafts due to post-translational modifications like palmitoylation, it can interact with additional cytoplasmic proteins. (When complexed within lipid rafts, CD44 inhibits CD3-mediated signaling in immune cells [41]).

Moreover, when it is released in the cytoplasm, CD44 ICD_f can translocate to the nucleus, where it serves as a transcriptional regulator for genes involved in cellular metabolic pathways, cell cycle progression, and stemness properties [42], see Figure 8. Depending on the isoform of the CD44 antigen, cleavage by presenilin-1/γ-secretase can release an ICD_f with either a short-tail or long-tail configuration.

CD44 short-tail isoforms encoded by exon 18 and short exon 19 coding region, are generally less abundant than long-tail isoforms. They contain the three amino acids 292RRR295 encoded by exon 18 and arise from an in-frame termination signal in exon 19 (CD44HD68) [43].

Conversely, long-tail CD44 ICD deriving by translation of the exon 18 and longer exon 19 coding sequences, contains a nuclear localization signal (NLS) that facilitates its import into the nucleus via the nuclear pore complex. The NLS comprises two clusters of basic amino acids (292RRRCGQKKK300), identical to the FERM-binding domain described earlier [1], as illustrated in Figure 7.

For instance, CD44 ICD binds to the promoter regions of MMP-9 and HIF-1 alpha, recognizing consensus sequences like the 12-O-tetradecanoylphorbol 13-acetate (TPA)-response element or CCTGCG [44]. Notably, CD44 exerts positive feedback on its own transcription by inducing the expression of full-length CD44 through its binding to the CBP/p300 transcriptional activator [45].

In thyroid cancer cells, the interaction between CD44 ICD_f and the CREB transcription factor leads to the phosphorylation of CREB, thereby activating its promoter and increasing cyclin D1 expression, which supports cell cycle progression [46].

In glioma cells, CD44 ICD_f is essential for hypoxia-induced stemness, particularly in hypoxic perinecrotic and perivascular regions, where it interacts with HIF-2 alpha, thereby enhancing the expression of HIF target genes [47].

Additionally, CD44 ICD_f interacts with individual stem cell transcription factors such as Oct4, Sox2, and Nanog to promote self-renewal, clonal formation, and therapeutic resistance. When it binds to the entire complex (ICD–Oct4–Sox2–Nanog), it facilitates their coordinated recruitment to the promoter response elements of target genes, thereby enhancing stemness properties in the cell [48; 49].

CD44 Receptor And Its Ligands

The CD44 receptor can interact with a variety of ligands, including Hyaluronic Acid (HA), fibronectin, osteopontin, serglycin, and chondroitin.

Hyaluronic Acid (HA):

HA, a linear glycosaminoglycan composed of alternating disaccharide units of D-glucuronic acid and N-acetyl-D-glucosamine, is one of the primary ligands for CD44. Binding of HA to CD44 induces conformational changes, facilitating the recruitment of adaptor molecules to its cytoplasmic domain. This triggers several signaling pathways like Ras, MAPK, and PI3K, which promote cell adhesion, migration, and proliferation [7].

In breast cancer, HA binding to CD44 enhances cell survival and proliferation by modulating β-catenin and NFκB signaling, leading to increased expression of Bcl-xl and P-glycoprotein in MCF-7 cells [28].

Fibronectin:

Although fibronectin doesn't directly bind to CD44, it interacts with the receptor after forming complexes with HA in the ECM. Evidence suggests that inhibition of hyaluronan synthesis, elicits higher fibronectin and collagen deposition in lung myofibroblasts [50].

Osteopontin (OPN):

OPN, a plasma protein, can bind to CD44, initiating signaling pathways associated with tumor progression and metastasis. The interaction between OPN and CD44 activates the Akt pathway, promoting cancer cell motility, survival, stem-like properties, and resistance to radiation. Inhibiting CD44 or reducing its expression decreases OPN secretion, leading to reduced clonogenicity in colorectal cancer cells. Additionally, CD44-positive colorectal cancer cells co-cultured with macrophages produce higher OPN levels, enhancing tumorigenicity [7].

Serglycin:

Serglycin binds specifically to CD44 in hematopoietic cells when combined with glycosaminoglycans like chondroitin sulfate (CS). This interaction with modified CD44 molecules leads to degranulation of CD44-positive cytotoxic T lymphocytes and influences lymphoid cell adhesion and activation [28].

Versican (VCAN):

VCAN, a chondroitin sulfate proteoglycan, binds to HA, resulting in the formation of structural aggregates. Secreted by tumor-associated fibroblasts and cancer cells, increased VCAN expression is linked to higher tumor grade and invasiveness in breast carcinomas with malignant appearing microcalcifications (MAMCs). When combined with elevated CD44 expression, VCAN promotes motility and invasion in ovarian cancer. Thus, disrupting the HA-CD44 interaction through VCAN might offer a strategy to curb tumor growth and metastasis in certain cancers [7].

CD44 And Its Role In Cancer

Various studies highlighted that CD44 plays a pivotal role in cancer cell metabolism [1].

A notable interaction exists between CD44 and the M2 splice isoform of pyruvate kinase isoenzyme type M2 (PKM2), a key enzyme in glycolysis. PKM2 catalyzes the conversion of phosphoenolpyruvate and ADP into pyruvate and ATP, facilitating the Warburg effect. This metabolic pathway allows tumor cells to produce energy (ATP) through glucose fermentation, even in oxygen-low environments [51, 52, 53].

Under conditions of hypoxia or in p53-mutated glycolytic cancer cells, CD44 interacts with PKM2, reducing its enzymatic activity through tyrosine phosphorylation by RTKs. This leads to increased glycolysis and the pentose phosphate pathway, a major source of NADPH [51].

CD44 also influences the lactate dehydrogenase isoenzyme ratio (LDHA/LDHB) via the AMPKα/mTOR and HIF-1α signaling pathways. Elevated CD44 expression boosts lactate glycolysis and the Warburg effect, while reduced CD44 expression favors mitochondrial respiration [51, 54, 55].

The CD44 ICD has also a role in the regulation of glycolytic enzymes. Indeed, it modulates the transcription of 6-phosphofructo-2-kinase/fructose-2,6-biphosphate 4 (PFKFB4), a critical enzyme in glycolysis. By interacting with CREB at the PFKFB4 promoter region, CD44 ICD_f regulates cellular levels of fructose-2,6-biphosphate (F-2,6-BP), an allosteric activator of phosphofructokinase 1 (PFK1), a key enzyme in the glycolytic pathway [1].

Moreover, CD44 ICD promotes the expression of ALDOC and PDK1, genes encoding enzymes involved in aerobic glycolysis. ALDOC facilitates ATP and glucose biosynthesis, while PDK1 inhibits pyruvate dehydrogenase, increasing lactate production and contributing to an acidified microenvironment that facilitates invasion and migration [56]. The role of CD44 ICD in the regulation of transcriptome and cell metabolism is shown in Figure 9.

CD44 has been implicated in solid neoplasia progression, influencing tumor growth, invasion, and therapeutic resistance.

In glioblastoma, CD44std enhances EGFR levels, promoting AKT signaling by inhibiting Rab7A. Combined inhibition of CD44 and EGFR results in significant glioblastoma cell death [57].

In head and neck cancers, CD44 binding with HA regulates stemness and CSCs development via the PI3K/4EBP1/SOX2 pathway. Additionally, the CD44/VCAM-1 interaction promotes cell invasion through the ezrin/PI3K pathway [58].

CD44 activates the oncogenic KRAS signaling pathway in lung adenocarcinoma, promoting tumor cell proliferation and survival [3].

In breast cancer, CD44std drives a positive feedback loop that supports tumor cell survival by enhancing PI3K/AKT signaling through hyaluronan synthase 2 (HAS2) activation. It also interacts with STAT3, influencing PDGFRβ/STAT3 pathways and modulating c-SRC levels to regulate breast CSCs [59].

In gastric adenocarcinoma, CD44 pairs up with FGFR2 to maintain cancer stemness by regulating c-Myc transcription [60].

In colorectal adenocarcinoma the HA/CD44v6 complex activates RTK signaling, boosting cancer cell survival, invasion, and resistance to apoptosis [61, 62, 63].

In cholangiocarcinoma, CD44std and CD44v8-10 interact with xCT, increasing reduced glutathione (GSH) levels and suppressing the p38 MAPK pathway, leading to a poorer prognosis [64].

CD44 knockdown in pancreatic neoplasia and ductal adenocarcinoma reduces Zeb1 and Snail1 expression, CD44 knockout in pancreatic cancer cells shows a reduction of either Zeb1 and Snail1 expression, impairing cell proliferation and invasion [11].

In hepatocellular carcinoma, CD44std boosts Twist1 and AKT signaling, increasing anoikis resistance [65].

Prostate cancer patients with a larger CD44+ cell population exhibit increased invasion and migration and active Hippo-YAP signaling, correlating with poor survival [66].

Finally, CD44 plays a role in chronic myeloid leukemia by modulating the Wnt/β-catenin signaling pathway, which is crucial for cell proliferation [67].

Overall, elevated CD44 expression is linked to higher tumor grade, advanced stage, and shorter survival in many cancers, due to its role in modulating several signaling pathways.

Overall, studies consistently highlighted that elevated CD44 expression is linked to cancers presenting an aggressive clinic-pathological behavior. Indeed, increased CD44 levels are closely tied to higher tumor

grades, advanced stages of the disease, and reduced survival rates [3]. This strong association underscores CD44 as a negative prognostic indicator across multiple cancer types.

The reason behind CD44's impact on cancer aggressiveness lies in its ability to influence various signaling pathways.

CD44 is enabled to drive cancer aggressiveness stems through its capacity to modulate multiple signaling pathways by activating different CD44 isoforms, these pathways are modulated, leading to enhanced tumor growth, invasion, and resistance to therapies. A visual summary of some of these pivotal signaling pathways influenced by CD44 can be found in Figure 10.

CD44 And Cancer Stem Cell

The prevailing theories on the origin of cancer stem cells (CSCs) can be broadly categorized into two definitions. The first posits that CSCs arise from normal stem cells undergoing genetic mutations or exposure to environmental factors. The second theory suggests that mutations in genes crucial for cell growth and differentiation can occur in any cell within a tumor population, leading to the emergence of sub-clones with proliferative and resistance advantages. Some of these sub-clones regain stem-cell-like properties, including self-renewal, multilineage differentiation potential, and resistance to standard cancer therapies [68, 69].

CSCs inhabit specialized microenvironments within tumors known as niches. These niches maintain the unique characteristics of CSCs, shielding them from immune surveillance [68].

The HA signalling in these niches might hinder the effectiveness of chemotherapeutic agents [70]. HA, abundant in both normal tissues and stem-cell niches, likely creates a conducive environment for CSC self-renewal and maintenance.

The glycoprotein CD44 is a commonly recognized marker for CSCs. The correlation between CD44 expression and tumor aggressiveness, including chemoresistance, likely mirrors the role of the hyaluronan/CD44 axis in maintaining CSC populations [69]. Indeed, CD44+ stem-like cells in epithelial tumors have demonstrated resistance to chemotherapy agents like paclitaxel and platinum [71].

CD44's regulatory functions span a broad spectrum, from integrating signals from the tumor microenvironment to transmitting them to the cell nucleus, thereby facilitating self-renewal.

CD44 plays a pivotal role in the EMT process. Binding of HA to CD44 promotes EMT, while inhibiting HA synthesis or using CD44-specific antibodies can mitigate EMT and reduce metastasis [69].

Stromal cells within niches also secrete various cytokines and growth factors (e.g., EGF, FGF, HGF, VEGF, TGF-b, and MMPs) that support the formation of CD44/RTK complexes and their downstream signaling. For instance, CD44v6 facilitates the interaction between HGF and c-Met, resulting in the activation of c-Met, MEK, and ERK signaling, promoting CSC migration and invasion [25].

CD44-positive CSCs display a heightened tumorigenic potential, forming tumors more effectively in xenograft models across various cancer types compared to CD44-negative counterparts. This suggests that CD44 plays a role in maintaining a subset of cells with stem-like properties [72].

In breast CSCs, CD44std, rather than CD44v, is predominantly expressed, promoting tumor initiation and CSC gene traits through the PDGFRβ/STAT3 signaling pathway. Eliminating CD44std diminishes CSC signatures, while shifting splicing from CD44v to CD44std enhances CSC traits [73].

Hypoxia, a hallmark of neoplasia, plays a crucial role in various aspects of carcinogenesis, including angiogenesis, tumor invasion, and metastasis, mediated by hypoxia-inducible factors (HIFs).

CD44 is central to regulating signaling pathways that maintain CSC characteristics under hypoxic conditions. There appears to be a synergistic relationship between CD44 and HIFs. For instance, Osteopontin-CD44 signaling regulates HIF-2a expression in gliomas, promoting aggressive growth and stem cell-like characteristics. Conversely, HIF-1a, induced during hypoxia, upregulates the expression of specific CD44 variants [74].

CD44 protects CSCs from stress-induced damage or apoptosis triggered by reactive oxygen species (ROS). CD44v isoforms appear to modulate ROS metabolism in CSCs through interactions with PKM2 and xCT [75, 76].

In conclusion, CD44, often associated with CSC properties such as tumor-initiating capabilities, immunosuppression, and chemoresistance, serves as a marker for CSCs in many cancers, notably breast and head and neck cancers. However, it's crucial to recognize the heterogeneity within CSC populations, and CD44 expression isn't universal. Targeted therapies against CD44 aim not only to eliminate tumor-initiating cells but also to disrupt the supportive microenvironment and signaling networks sustaining CSCs and therapy resistance.

CD44 Isoforms

CD44 exists in multiple isoforms due to post-transcriptional modifications. The presence of splice variants in normal tissues suggests that CD44 isoforms also play a role in normal cellular functions. Conversely, specific isoforms have been associated with various cancers, with distinct biological

functions and clinical implications, see Supplementary Table 2. Many CD44 variants contribute to tumorigenesis and may serve as prognostic biomarkers [1, 3].

For example, while normal colonic mucosa lacks CD44 isoforms, tumors in this location often display a wide range of CD44 isoforms. Numerous studies have linked some CD44 variants to specific stages of cancer progression [77].

In breast cancer cells, Brown et al. demonstrated a shift in CD44 profiling from variant to standard isoforms, leading to increased protein levels. This shift accelerated the EMT process by activating AKT signaling, resulting in metastatic behavior and resistance to apoptosis [78].

Other studies indicate that CD44std interacts with phosphorylated cortactin to activate invadopodia, facilitating ECM degradation and metastasis in breast and lung tumors. Reduced CD44std levels inhibit invadopodia activity, leading to reduced tumor invasion and metastasis [79].

In hepatocellular carcinoma, CD44std expression correlates with poor patient survival. Similarly, in pancreatic cancer, it is implicated in radio-resistance by maintaining ERK phosphorylation and promoting radiation-induced EMT [3].

Up-regulation of CD44v6 is associated with invasion in endocrine-resistant breast cancer cells via EGFR signaling and MAP kinase activation, promoting cell survival and growth. Chemotherapy efficacy is reduced in colon cancer cells overexpressing CD44v6 [80].

In tongue squamous cell carcinoma, high CD44v6 expression increases relapse risk [81].

CD44v9 overexpression is found in gastric adenocarcinoma, while in a prostate cancer cell model (PC3), CD44v9's effect synergizes with CD44v6.

Overexpression of CD44v6 and CD44v10 is associated with poor prognosis, increased cell migration, and metastasis in colorectal cancer. The CD44v8-10 variant enhances resistance to reactive oxygen species (ROS) in gastrointestinal cancer cells and, in combination with CD44v4-10, promotes tumorigenesis [3].

CD44v2 overexpression predicts poor prognosis in primary and xenograft tumors [82]. CD44v5-6 and CD44v8-9 amplify the KRAS/MAPK signaling pathway, leading to increased tumor cell proliferation in non-small cell lung adenocarcinoma. These patients typically have a worse prognosis due to CD44 alternative splicing induction [83].

In neck squamous cell carcinoma (HNSCC), CD44v3 overexpression significantly increases cell migration. Inhibiting this variant with specific antibodies reverses cisplatin resistance and reduces cell proliferation [84].

Hu and colleagues identified a positive correlation between ESRP1 levels and the likelihood of distant metastasis in breast cancer. The expression of ESRP1 was detected to align with particular mRNA levels of CD44v3-10 and CD44s. Notably, the ratio of CD44v3-10 to CD44s expression emerged as a predictive marker for distant metastasis [85].

In conclusion, profiling CD44 expression in cancer patients has become crucial for determining prognosis, especially when EMT plays a pivotal role in metastasis. Clinical evidence suggests that both CD44std and CD44v isoforms contribute to tumor aggressiveness and metastasis across various cancers. Thus, CD44 isoforms expressed by patients could serve as potential prognostic markers and therapeutic targets to prevent metastasis.

CD44 IN TRIPLE-NEGATIVE BREAST CANCER (TNBC)

According to the World Health Organization's 2020 data, approximately 2.3 million women worldwide were diagnosed with breast cancer, resulting in 685,000 deaths in that year alone.

Of all breast cancer cases, 10-15% are identified as triple-negative breast cancer (TNBC). This subtype earns its name because it lacks expression of estrogen receptor (ER), progesterone receptor (PR), and does not overexpress the HER2 gene—thus testing "negative" on all three counts. This characteristic makes TNBC particularly aggressive, with a propensity for metastasis to the bones and brain.

TNBC predominantly affects younger, premenopausal women under 40, especially those of African ethnicity or those with a BRCA1 mutation.

Women with TNBC often face higher recurrence rates, shorter disease-free intervals, and reduced overall survival (OS) when not treated appropriately [86, 87].

Research on TNBC extends beyond its aggressive nature to its intricate molecular profile. Recent guidelines have categorized TNBC into two main subtypes: basal-like, representing 70% of TNBC cases, and non-basal subtypes.

Further molecular classifications have been proposed by experts Lehmann and Burstein, identifying six TNBC subgroups: immunomodulatory (IM), mesenchymal (NI), mesenchymal stem-like (MSL), luminal androgen receptor (LAIR), and two basal-like subtypes (BL1 and BL2). Burstein et al. also contributed to molecular classification by distinguishing four TNBC clusters based on DNA profiling: luminal AR (LAR), mesenchymal (MES), basal-like immunosuppressed (BLIS), and basal-like immune-activated (BLIA). This molecular categorization aids in identifying TNBC subtypes associated with breast cancer

stem cells (BCSCs) a diverse population capable of self-renewal and differentiation, and known for features like chemotherapy resistance, tumor recurrence, and metastatic potential [86].

The characterization of BCSCs remains a complex area of study. Stemness markers commonly linked to BCSCs include CD44 and CD24. CD24, an adhesion molecule, binds to P-selectin a protein found on activated platelets and endothelial cells. CD24 has been observed to promote metastasis by facilitating interactions between proliferating tumor cells and the extracellular matrix via P-selectin and fibronectin in mammary carcinoma cells.

In TNBC, CD44+/CD24+ BCSCs are strongly associated with distant metastatic subtypes, whereas CD44+/CD24- cells display an EMT phenotype, known for its invasive and metastatic capabilities.

Regarding OS, patients with CD44+/CD24- BCSCs generally have a poorer prognosis compared to CD44-/CD24- patients, albeit less severe than some other subtypes [86].

Aim

The primary objective of this study was to explore the transcriptome of CD44 in MDA-MB-231 and MDA-MB-468 triple-negative breast cancer (TNBC) cell lines using long-read RNA sequencing. The aim was to identify distinct expression patterns that could not only explain the variability in CD44 immunophenotypes, but also shed light on the differential behavior of these TNBC cell lines in xenograft models. This approach could potentially uncover novel transcript variants relevant to the biology of human breast cancer, thereby expanding our understanding of the diagnostic, prognostic, and therapeutic implications of the CD44 repertoire.

Results

Heterogeneous localization of CD44 in Triple-Negative Breast Cancer Samples and Xenograft Models

We conducted an immunohistochemical (IHC) analysis of 90 formalin-fixed, paraffin-embedded human triple-negative breast cancer (TNBC) samples to evaluate the expression of the CD44 receptor. For the analysis, antibodies specific for the receptor's ectodomain were used, as described in the materials and methods.

Surprisingly, our comparative analysis revealed variable staining patterns compared to the canonical membrane localization. Indeed, in some cases, CD44 displayed a cytoplasmic or even nuclear localization. Interestingly, we also observed concurrent expression in both membrane and other compartments within the same samples, as illustrated in Figure 11.

We subsequently analyzed breast tumor tissue samples derived from xenograft models, including MDA-MB-231 and MDA-MB-468 xenograft TNBC tumors [88, 89]. The MDA-MB-231 cell line is known for its high proteolytic activity on the extracellular matrix (ECM), making it a highly aggressive and metastatic tumor capable of spreading to the bones, brain, and lungs. In contrast, the MDA-MB-468 cell line exhibits morphological features consistent with a basal-like phenotype and strong expression of EGFR and cytokeratin 5/6, accompanied by a high Ki-67 index.

Our analysis revealed distinct differences in the localization of CD44 between the two cell line-derived xenograft models. Specifically, IHC analysis on the MDA-MB-468 xenograft tumors displayed the typical membranous staining pattern for CD44. In contrast, MDA-MB-231 tissue samples exhibited the atypical nuclear CD44 localization, as illustrated in Figure 12.

These observations highlight the heterogeneity in terms of CD44 protein localization, which could reflect different molecular mechanisms that contribute to the distinct biological behaviors of TNBCs.

Characterization of a Novel CD44 Isoform in Triple-Negative Breast Cancer Cell Lines

To obtain sufficient RNA for Nanopore library preparation, cultures of the TNBC cell lines MDA-MB-231 and MDA-MB-468, as well as the non-tumorigenic control cell line MCF10A, were propagated to a density of 10^5 cells. The MCF10A line is an immortalized epithelial cell line derived from the mammary gland of a 36-year-old Caucasian female with fibrocystic breast disease [94].

Total RNA from each sample was enriched for mRNA by reverse transcription of polyA-tailed transcripts. Subsequently, long-range PCR targeting the CD44 transcriptome was conducted. Given the extensive transcriptional diversity of the CD44 gene, PCR primers were designed to amplify specific exons that are highly conserved across annotated CD44 isoforms, refer to Supplementary Table 3.

Electrophoretic analysis of the PCR amplicons revealed product sizes ranging from 300 bp to 2 kb for the MCF10A control, and 700 bp to 7.5 kb for the TNBC lines, indicating a broader CD44 expression profile in the cancer cells compared to the non-tumorigenic control. Furthermore, the TNBC lines exhibited CD44 mRNAs of higher molecular weight relative to MCF10A (Figure 13).

The PCR amplicons from each sample were then barcoded and pooled for sequencing on the Nanopore platform. Bioinformatic processing, including transcript alignment, reconstruction, and quantification [91, 92] yielded approximately 4 million, 3.2 million, and 3 million reads for MCF10A, MDA-MB-231, and MDA-MB-468 respectively, showing a coverage suitable for targeted RNA expression analysis [93]. as reported in Supplementary Table 4.

To validate the specificity of the primers designed to amplify the CD44 transcriptome, the reads were filtered for the region of chromosome 11 encompassing the targeted exons, as indicated in the materials and methods section. This analysis confirmed an accuracy close to 100% for the reads, validating the enrichment approach used during library preparation, as reported in the columns 3 and 4 of the Supplementary Table 4.

After filtering, 36 transcripts with greater than 1 read per sample were identified.

For differential expression analysis (DEA), Reads Per Million (RPM) indices were calculated for each transcript, and the normalized data were compared between the experimental cell line libraries. The heatmap shown in Figure 14, illustrates the DEA results. Lastly, the fold change (FC) parameter was calculated by comparing the filtered RPM values of the MDA-MB-231 and MDA-MB-468 cell lines to the control cell line.

Of the 36 transcripts, 7 clusters were excluded from further analysis. Three clusters did not align with any annotated CD44 isoforms (BambuxT2, BambuxT3, BambuxT6), and four had the lowest RPM values (XM_011520487.4, NM_001001392.2, XM_011520489.4, XM_047427904.1, BambuxT2, BambuxT3). Additionally, one transcript corresponding to CD44 noncoding RNA (NR_174965.1) was not considered in this study.

FC analysis revealed that 6 out of the remaining 28 identified transcripts were upregulated (5-114 fold) and 19 were downregulated (2-51 fold) in MDA-MB-231 compared to the control. Three transcripts with FC < 2 were excluded (see Figure 15 A and Supplementary Table 5). For MDA-MB-468, 5 out of 28 transcripts were upregulated (2-53 fold) and 10 were downregulated (2-14 fold) relative to the control line. Thirteen transcripts with FC < 2 were excluded (see Figure 15B).

Comparison of the fold change (FC) values showed that only 2 isoforms were upregulated in both TNBC cell lines: NM_001001391.2 and XM_005253238.4. While the former had similar FC in both lines, the latter showed expression levels approximately 2-fold higher in MDA-MB-231 than in MDA-MB-468.

Conversely, among the predicted overexpressed transcripts, only XM_017018585.2 showed differential expression between the TNBC lines when compared to the control MCF10A cell line. Notably,

XM_017018585.2 was approximately 4-5 fold more highly expressed in MDA-MB-231 than in MDA-MB-468, as shown in Figure 15A-B and Supplementary Table 5.

Regarding downregulated transcripts, MDA-MB-231 showed transcripts with lower FC compared to MDA-MB-468. Nine isoforms were downregulated in both TNBC lines, while 10 were exclusively repressed in MDA-MB-231, as reported in Supplementary Table 5.

Next, to analyze the assembly of the identified isoforms, BAM files resulting from read alignments for each cell line were loaded into the Integrative Genomic Viewer (IGV – version 2.8.2). By comparing the reads to the reference genome (GRCh38 – hg38), the exact exon genomic coordinates, coverage, and variant calling for each read were obtained.

A distinct read cluster exhibited a specific junction between exons ENSE00003678600 (Chr11:35,198,121-246) and ENSE00003526469 (Chr11:35,214,852-914), corresponding to EX8 and EX17 of the human CD44 gene, as depicted in Figure 16, (refer to Supplementary Table 1 for the exons genomic coordinates). The ENSE00003526469 exon is more conserved than ENSE00003678600. Indeed, it is retained as EX15 in the unspliced longest isoform NM_000610.4 (CD44v2-10) and as exon s6 in CD44std. In contrast, ENSE00003678600 is conserved in only 17 CD44 transcripts: 2 annotated and 15 predicted. For instance, it is spliced in CD44std, but it is EX7 in the CD44v2-10 isoform.

The consensus sequence (CS) of this cluster was obtained using samtools consensus [94] and blasted with the NCBI Blast tool [95] to identify transcripts with the highest similarity. The CS confirmed 100% identity with the above-mentioned isoform XM_017018585.2 (data not shown).

This transcript consists of 10 exons and has a coding sequence (CDS) of 1212 bp, encoding a 403-amino acid protein (deposited in GenBank with accession XP_016874074). The 27 bases spanning the exon 9 to exon 10 junction encode a nuclear localization signal (NLS), as shown in Supplementary Table 6.

Splicing modifications result in a unique exon-exon junction formed by exons 6 and 7 (exons 6/7) of this transcript, conserved only by another rare CD44 isoform known as CD44-220 (Ensembl Transcript ID ENST00000525688.5). However, unlike XM_017018585.2, the CD44-220 isoform lacks the first 4 exons at the 5' end after post-translational modifications.

To sequence the XM_017018585.2 isoform in TNBC cells using Sanger methodology, two pairs of primers were designed: one to amplify a 5' region and the other for a 3' region of the target, from the exons 6/7 junction, of 721 bp and 220 bp, respectively (as described in the Materials and Methods section).

RNA from both breast cancer cell lines was therefore reverse-transcribed and used in PCR to amplify the two target regions in two separate reactions, as described in the Materials and Methods section. After purifying the PCR products, their lengths were confirmed by electrophoresis using the Bioanalyzer 2100, as shown in Figure 17 A-B.

Subsequently the amplicons corresponding to the 721 bp 5' region and 220 bp 3' region, were sequenced using the Sanger methodology, as described in the Materials and Methods section. (The Sanger row data are available as supplementary file).

The resulting sequences from Sanger for each target region in both TNBC lines were blasted, confirming a 99% identity with the XM_017018585.2 transcript, as shown in Figure 18 A-B.

Finally, a Sybr RT-qPCR analysis was conducted to explore the gene expression (GE) of the CD44 isoform in the TNBC cell lines. In addition to assessing XM_017018585.2 expression, the expression of the CD44std isoform was examined to further clarify the differences in the gene transcriptome between these breast cancer cell types. Utilizing the 2^ΔΔCT method, mRNA expression levels were quantified, using the MCF10A non-tumoral cell line as a reference for the analysis. The findings revealed a tenfold increase in XM_017018585.2 expression in MDA-MB-231 compared to MDA-MB-468, with a more pronounced difference between the standard isoform and the variant observed in the former rather than the latter, as reported in Figure 19.

MOLECULAR ANALYSIS OF THE XM_017018585.2 OVER-EXPRESSION IN TRANSFECTED MDA-MB-468 CELL LINE

To investigate the potential effects of increased expression of the XM_017018585.2 isoform on the TNBC cell line MDA-MB-468, which is known for its lower invasiveness and heightened sensitivity to chemotherapy compared to MDA-MB-231, the consensus sequence obtained from Sanger sequencing (with 99% identity to the predicted isoform) was cloned into a lentiviral vector system, as described in the Materials and Methods section. Following this, three cultures of MDA-MB-468 cells were expanded. One culture was transfected to consistently express XM_017018585.2 mRNA (referred to as $CD44_{vector}$). As a control for the transfection assay, another culture was transfected with the same system engineered to express green fluorescent protein (GFP) continuously, labeled as GFP_{vector}. The fluorescence emitted from this control was quantified to assess transfection efficiency in the cells.

RNA extraction was then performed on both transfected cells and a third culture not subjected to transfection (referred to as parental 468), following the protocol outlined in the Materials and Methods

section. After the extraction, GE qPCR analysis by Sybr RT-qPCR was conducted for both CD44std and XM_017018585.2 mRNA to assess the impact of transfection on the CD44 transcriptome. The 2^ΔΔCT method was employed to quantify mRNA expression levels, with the parental 468 sample serving as the calibrator for the analysis. The results indicated that exclusive induction of the XM_017018585.2 isoform expression was observed only in cells transfected with $CD44_{vector}$. Specifically, compared to parental 468 cells, only the predicted isoform exhibited upregulation in cells treated with $CD44_{vector}$, while the expression of CD44std remained relatively stable, as reported in Figure 20.

Finally, RNA from all three experimental groups was processed for bulk sequencing by the INVIEW Transcriptome Discover service at Eurofins Genomics, as detailed in the Materials and Methods section. Alignment of the generated fastq files was performed using the STAR software [96], and the resulting BAM files were converted to a count matrix. Data normalization was conducted by calculating counts per million (CPM) using the limma package. Due to the limited number of samples, calculation of p-values for comparison analysis was not feasible. Therefore, differentially expressed (DE) genes were identified based solely on $\log_2$ fold change ($\log_2$ FC), with a threshold value set at > 1.

Subsequently, DE genes common to cells treated with $CD44_{vector}$ and either GFP_{vector}-transfected cells or parental 468 cells were excluded from the analysis to eliminate false positives. The remaining genes were categorized as up- or down-regulated. This analysis resulted in two gene sets comprising 164 up-regulated and 175 down-regulated genes, which were utilized for clustering analysis. Heat map that depicts the results of this analysis is reported in Figure 21. Files containing the differentially expressed (DE) gene sets (up- and down-regulated) are provided in .xlsx format as supplementary materials.

Each gene set (up- and down-regulated) underwent analysis using the Gene Set Enrichment Analysis (GSEA) tool, which compares the list of genes to a collection of gene sets in the database [97]. For the up-regulated gene set, two overlaps were identified. The first overlap was found in the C2 collection, linking 16 genes with the "YOSHIMURA_MAPK8_TARGETS_UP" set, which was identified in a study on the role of MAPK8 in abdominal aortic aneurysm disease [98]. The second overlap consisted of 10 genes belonging to the "DESCARTES_MAIN_FETAL_CILIATED_EPITHELIAL_CELLS" set in the C8 collection. This gene set is expressed in organ-specific specializations of broadly distributed cell types in fetal erythropoiesis sites, particularly in the adrenal gland [99]. The overlaps for the up-regulated gene set are shown in Figure 22.

In contrast, the down-regulated DE gene set exhibited more overlaps with various GSEA collections compared to the up-regulated DE gene set. Among these overlaps, several were related to alpha and gamma interferon-responsive programs, including those observed in cancer patients [100], as well as

fatty acid metabolism. Importantly, there were overlaps with three genes that shown downregulation when KRAS was silenced in epithelial human cancer cell lines with abnormal expression of the oncogene.

Additionally, overlaps were identified with 10 genes found down-regulated in another type of breast cancer cell line, MCF-7, which exhibits features commonly seen in luminal subtype breast cancers, as reported in Figure 23. MCF-7 is estrogen receptor (ER) positive, progesterone receptor positive, and HER2 negative, and its responsiveness to estrogen makes it a common model for studying the effects of estrogen and anti-estrogen therapies on breast cancer cell growth and survival [101].

Discussion

The expression profiles of CD44 isoforms exert significant influence on pivotal aspects of cancer cell behavior, including tumorigenicity, tumor-initiating capacity, metastasis, and response to therapeutic interventions. Recent investigations have underscored the pivotal role of the standard CD44 isoform (CD44std), in contrast to CD44 variant (CD44v) isoforms, in driving tumor initiation and the acquisition of cancer stem cell (CSC) traits through activation of the PDGFR/STAT3 signaling cascade. Depletion of CD44std leads to attenuation of CSC signatures, while modulation of the splicing regulator ESRP1 to favor CD44std over CD44v splicing induces CSC phenotypes [73]. Additionally, CD44std promotes epithelial-mesenchymal transition (EMT) by activating the AKT pathway, thereby fostering the emergence of EMT-associated recurrent tumors and conferring resistance to apoptosis [78]. Likewise, CD44std, but not CD44v, engages phosphorylated cortactin to facilitate invadopodia activation, promoting extracellular matrix (ECM) degradation and metastasis of breast tumor cells to distant sites such as the lungs. Depletion of CD44std curtails invadopodia activity, impedes ECM degradation, and diminishes tumor cell invasion and metastasis [79].

Conversely, some investigations suggest that tumors expressing certain CD44v isoforms exhibit heightened aggressiveness compared to those expressing CD44std. For instance, numerous studies have associated CD44v6 expression with the progression and prognosis of various cancer types [3]. In head and neck squamous cell carcinoma (HNSCC), overexpression of the CD44v3 isoform has been demonstrated to significantly augment cell migration [102].

Furthermore, CD44v10 has been identified as a major promoter of invasiveness in MDA-MB-231 cells, with anti-CD44v10 antibodies markedly inhibiting tumor migration towards type I collagen, implicating its involvement in TNBC migration and proliferation, albeit without affecting adhesion [103].

Breast cancer ranks as the most prevalent cancer among women, excluding skin malignancies, accounting for approximately 30% of all new female cancer diagnoses annually in the United States alone. Incidence rates have shown a modest annual increase, particularly among women under 50. Despite advancements in screening, awareness, and treatment modalities, breast cancer remains the second leading cause of cancer-related mortality in women after lung cancer, although mortality rates have declined by 42% since 1989 [87].

TNBCs comprising 10-15% of breast cancer cases, exhibit aggressive behavior and metastatic potential due to the absence of estrogen receptor (ER), progesterone receptor (PR), and overexpression of the HER2 gene.

The hypothesis driving this study stemmed from histological observations in human TNBC samples, wherein variable CD44 receptor staining patterns were detected using antibodies derived from the same clone (VFF-18) but synthesized via an inflammatory response to different CD44 peptides. Although predominant membrane staining was observed as anticipated, focal cytoplasmic and nuclear expression patterns were also noted. This aberrant phenotype was similarly observed in basal-like MDA-MB-468 and claudin-low MDA-MB-231 TNBC cell lines, which exhibit distinct in vivo behaviors. MDA-MB-231 cells demonstrate high invasiveness via extracellular matrix (ECM) proteolysis, facilitating premetastatic cell extravasation, while MDA-MB-468 displays less aggressive behavior and heightened responsiveness to chemo- and radiotherapy compared to MDA-MB-231.

To investigate the CD44 in TNBC, we employed long-read sequencing technology known as Oxford Nanopore Technology (ONT). This approach offers significant advantages over conventional sequencing methods like Illumina or Ion Torrent, which are limited in their ability to sequence regions longer than 300 base pairs. In contrast, ONT enables the generation of lengthy sequencing reads, up to 206.5 kilobases for unique reads. This capability facilitates comprehensive genome assembly without gaps and allows for accurate detection of repetitive regions and structural variants. These properties make ONT technology well-suited for distinguishing between different transcriptional variants of a gene [104, 105]. The sequencing analysis revealed distinct expression profiles in the TNBC cell line MDA-MB-231, characterized by both up- and downregulation of CD44 isoforms, markedly differing from the non-tumoral control line. In contrast, the TNBC cell line MDA-MB-468 exhibited a transcriptional profile more akin to the control, with approximately 48% of isoforms displaying expression levels comparable to the Mcf10A cell line, as reported in Figure 15A-B.

Among the differentially expressed transcripts, the predicted isoform XM_017018585.2 emerged as upregulated in both TNBC cell lines, albeit with higher expression levels observed in MDA-MB-231 compared to MDA-MB-468, as corroborated by both next-generation sequencing (NGS) and quantitative PCR (qPCR) approaches. This transcript spans 10 exons, featuring a predicted coding sequence (CDS) of 1212 base pairs, encoding a membrane receptor comprising 403 amino acids (XP_016874074). Notably, it exhibits a distinctive junction between exons 6 and 7 (EXs 6/7), which was consistently

identified in the sequence obtained via Sanger analysis. The experimental validation in this study confirmed the expression of the XM_017018585.2 isoform.

During the course of data collection for this study, the NCBI record pertaining to XM_017018585.2 was temporarily suppressed from NCBI text-based search and comparative analysis results. However, such suppressed data typically undergo reinstatement to public status at a future date [106]. The General Transfer Format (GTF) file containing annotation details for XM_017018585.2 is accessible via the following link:

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/

To elucidate the functional implications of the CD44 isoform XM_017018585.2, its RNA sequence was synthetically synthesized and cloned into a lentiviral plasmid for transfection into the MDA-MB-468 TNBC cell line. This approach aimed to observe the potential influence of increased protein expression on various transcriptional pathways.

Gene enrichment analysis, conducted using a web-based tool on genes exhibiting a $\log_2$ fold change >1 and filtered by differentially expressed genes identified in either GFP-transfected or parental cells, revealed overlaps with annotated gene sets. Among the upregulated gene set, enrichment was observed for mitogen-activated protein kinases (MAPKs) implicated in extracellular matrix (ECM) degradation signaling and genes associated with fetal stem cells. Conversely, genes involved in interferon response programs or downregulated in interference experiments conducted on estrogen-responsive breast cancer cells, such as the MCF-7 line, were identified in the downregulated gene set, as reported in Figures 22 and 23.

To continue the study, in silico analysis is being conducted to predict the folding and interactome of the protein encoded by the XM_017018585.2 isoform, thereby enhancing our understanding of its biological role. These investigations could explore whether the specific junction between exons 6 and 7 confers unique functional properties to the receptor, particularly in relation to the biology of stem or mature cancer cells.

Furthermore, in vitro migration assays conducted on various extracellular matrix (ECM) substrates, comprising primary ECM molecules such as collagen, fibronectin, and osteonectin, among others, could provide insights into the invasive potential of cells constitutively expressing the XM_017018585.2 isoform.

Moreover, in vivo experiments involving the transplantation of transformed cells into animal models could yield additional information, offering a deeper understanding of the underlying biology of TNBC and potentially revealing new prognostic and therapeutic insights for patients.
Finally, the CD44 transcriptome analysis using ONT in the TNBC models of this study revealed the presence of additional non-annotated transcriptional variants. These variants could potentially hold clinical significance as biomarkers. However, the specific roles of these CD44 isoforms remain unknown, and further investigations are required to elucidate their clinical implications.

Material and methods

SAMPLE COLLECTION AND IMMUNOLOCALIZATION ANALYSES

Formalix-fixed and paraffin-embedded samples of human TNBCs and of xenografts of MDA-MB-231 and MDA-MB-468 were collected as part of a study conducted at Fondazione IRCCS Istituto Nazionale dei Tumori [107].
Four-micrometer–thick human and mouse tissue sections were deparaffinized, rehydrated, and unmasked using Novocastra Epitope Retrieval Solutions at pH9 (Leica Biosystems) in a thermostatic bath at 98°C for 30 minutes. The sections were then brought to room temperature and washed in PBS. After neutralization of the endogenous peroxidase with 3% H_2O_2 and Fc blocking with 0.4% casein in PBS (Leica Biosystems), the sections were incubated with CD44v6 mouse monoclonal (clone VFF-18,1:500) primary antibody. IHC staining was developed using the Novolink Polymer Detection Systems (Leica Biosystems) or IgG (H&L)-specific secondary antibodies (Life Technologies, 1:500) and AEC (3,3′-Diaminobenzidine, Novocastra) as substrate chromogen. IHC-stained slides were analyzed and imaged under a Zeiss Axioscope-A1 and Axiocam 503 Color camera (Zeiss).

RNA EXTRACTION AND CDNA SYNTHESIS

RNA was extracted from three cellular lines: two models of Human Triple Negative Breast Cancer, MDA-MB-231 and MDA-MB-468, and an epithelial cell line MCF10A.
For MDA-MB-231 and MDA-MB-468, cells were cultured in RPMI-1640 medium supplemented with 10% serum, 2 mM L-glutamine, 1 mM sodium pyruvate, and 1× non-essential amino acids. MCF10A cells were cultured in DMEM-F12 medium supplemented with 10% FBS, 20 ng/ml EGF, 10 µg/ml insulin, 500 ng/ml hydrocortisone, L-glutamine, and Pen/Strep.

To extract RNA, 10^5 cells were obtained from each culture and processed using the miRNeasy Mini Kit (Qiagen, Cat. No./ID: 217004) following the manufacturer's instructions.

Next, 1 µg of RNA per sample was reverse transcribed using 100 pmoles of Oligo(dt)20 primers (Smobio, cat no RP1300). The reverse transcription reaction mixture consisted of 200 U of Maxima H Minus Reverse Transcriptase (200 U/µl) in 20% of 5x RT Buffer (ThermoFisher, cat # EP0751), 20 U of RNase Inhibitor (Smobio, cat no RP1300), 0.5 mM of dNTP mix (each at a starting concentration of 10mM), and nuclease-free water to achieve a volume of 20 µl. Before adding the Reverse Transcriptase, the RNA was incubated at 65 °C for 5 min to reduce secondary structures. Each solution was then heated at 50 °C for 30 min to prime the reverse transcription reaction, followed by heating at 85 °C for 5 min to terminate the reaction, as described in the manufacturer's Reverse Transcriptase protocol.

CD44 TRANSCRIPTOME LIBRARY PREPARATION AND SEQUENCING

CD44 target region cDNA Synthesis:

cDNAs were used in a tailing PCR reaction employing a long-PCR DNA Polymerase and tailed primers to barcode and enrich for CD44 transcripts. The tailed primers were designed to recognize two non-spliced exons of the CD44 gene: ENSE00002197227 for the forward primer and ENSE00002190979 for the reverse primer, refer to Supplementary Table 3 to exons details. The primers were previously modified as described in the "Introduction to the PCR Barcoding" section of the protocol "ligation-sequencing-gdna-pcr-barcoding-sqk-lsk114-with-exp PBC_9182_v114_revG_07Mar2023-minion". It protocol is available through ONT community web page at the link:

https://community.nanoporetech.com/docs/prepare/library_prep_protocols

All the primers used in this study are listed in Table 7 in the supplementary material.

These primers amplified a region within Chr11, spanning from 35,175,69 to 35,220,387 bases.

To perform the amplification, an amplification mix was prepared containing 10% cDNA per sample, 0.4 µM of each primer, 32% LongAmp Taq 2X Master Mix (NEB, cat # M0287), and nuclease-free water to achieve a volume of 50 µl. The PCR thermal profile consisted of 1 step at 95 °C for 30 sec, 14 steps at 95 °C for 15 sec and 65 °C for 104 sec, followed by a final step at 65 °C for 6 min.

After the amplification, the libraries were treated with exonuclease I (NEB, cat # M0293) to remove any excess primers that had not successfully annealed. Then, the libraries were purified using AMPure XP Beads provided in the Ligation Sequencing Kit V14 (Nanopore, cat. no SQK-LSK114). Finally each library was eluted in 12 µl of nuclease-free water, following the instructions described in the "cdna-pcr-

sequencing-sequence-specific-sqk-pcs111-CPSS_9185_v111_revB_19Apr2023-minion" protocol, from page 10 step 7 to page 11 step 22, available through ONT community web page at the link: https://community.nanoporetech.com/docs/prepare/library_prep_protocols
Subsequently, 2 µl per library were used for quantification using a qubit 3.0 fluorimeter (ThermoFisher Scientific, Catalog Number Q33216) and to perform a quality check through electrophoresis profiling using an Agilent 2100 Bioanalyzer (Agilent, G2938B), as shown in Figure 13.

Long target sequencing and CD44 reads filtering:

The library was prepared using the Ligation Sequencing Kit V14 with PCR Barcoding Expansion Pack 1-96 (Nanopore, cat. no EXP-PBC096), and R10.4.1 MinION Flow Cells (Nanopore, cat. no FLO-MIN114), following the instructions in the "ligation-sequencing-gdna-pcr-barcoding-sqk-lsk114-with-exp PBC_9182_v114_revG_07Mar2023-minion" protocol, starting from the Barcoding PCR step on page 17 available through ONT community web page at the link: https://community.nanoporetech.com/docs/prepare/library_prep_protocols
In brief, nuclease-free water was added to each library to achieve a volume of 48 µl. Then, 2 µl of Nanopore PCR barcode 08, 09, and 10 were mixed with the respective libraries for MDA-MB231, MDA-MB468, and MCF10A. Additionally, 50 µl of LongAmp Taq 2x master mix was added to each library. The PCR thermal profile included 1 step at 95 °C for 3 min, followed by 22 cycles of 95 °C for 15 sec, 62 °C for 15 sec, and 65 °C for 4 min, with a final step at 65 °C for 4 min.
Following the barcoding step, the libraries were purified using Ampure XP and eluted in 25 µl of nuclease-free water. Each library was diluted to a concentration of approximately 19.16 fmol/µl. Then, 16.34 µl per sample were mixed to create a library pool with a final volume of 49 µl. This library pool was further processed for End-prep and Adapter ligation steps, as described in the protocol. To calculate the molarity, 1 µl of the eluted pool was analyzed using an Agilent 2100 Bioanalyzer.
Finally, 20 fmols of the library were added to 75 µl of the library sequencing mix, which was loaded into a previously primed nanopore flow cell.
The MinKNOW nanopore software was set up to generate raw data POD5 files.Basecalling and alignment of nanopore sequences. The raw data basecalling and demultiplexing were performed using DORADO version 0.6.0 [91] .
The following commands were used:

- dorado basecaller --emit-fastq --no-trim dna_r10.4.1_e8.2_400bps_hac@v4.3.0 file.fastq - This command was used to start the basecalling process.

- dorado demux --emit-fastq --output-dir fastq_folder --kit-name EXP-PBC001 file.fastq - This command was used to start the demultiplexing process and store the output fastq files in the specified folder.

Next, fastq alignment, transcript reconstruction, and quantification were performed using the nfcore nanoseq pipeline version 3.1.0 [92]. The transcripts that aligned with the CD44 Chr11 enrichment region were identified and filtered using the PrintReads tool [108].

For the comparison analysis, each read was normalized as Read Per Million (RPM) by calculating the ratio between the number of a specific read and the total number of reads, and then multiplying it by 10^6.

PCR AND SANGER SEQUENCING

To synthesize the XM_017018585.2 isoform through PCR, two pairs of primers were designed to amplify an upstream region of 721 bps (from EX2 to EXs6/7) and a downstream region of 220 bps (from exon EXs6/7 to exon EX9) from the specific exon-exon junction. Please refer to in Supplementary Table 7, for the primer details.

For the PCR reaction, 5 µl of each reverse-transcribed sample was added to a mix containing 50% Hot Start Taq 2X master mix (NEB, M0496), 0.2 µM of each primer, and nuclease-free water to achieve a final volume of 50 µl. The thermal profile was set up according to the manufacturer's instructions.

After the PCR amplification, each reaction was purified using Ampure XP beads (Beckman Coulter, Product No: A63881), and the size of the PCR products was verified using the Agilent DNA 12000 kit on the Bioanalyzer 2100, as shown in figure 18.

The amplicons obtained from both TNBC cell lines were diluted with nuclease-free water to a final concentration of 2.5 ng/µl for the 721 bp amplicon and 0.5 ng/µl for the 220 bp amplicon. Each PCR product was mixed with 25 pmol of a single primer (either forward or reverse) in a final volume of 10 µl. Samples with the forward primer and the reverse primer, for each TNBC cell line, were sent to Eurofins Genomics service for sequencing using the Sanger methodology (Mix2Seq service [109]).

GENE EXPRESSION ANALYSIS

Gene expression (GE) analysis of the XM_017018585.2 isoform was performed using RT-qPCR. For each sample, 2 µl of cDNA was added to a 20 µl reaction mix consisting of 0.5 µM of primers, 50%

PowerUp™ SYBR™ Green Master Mix (2X) (Thermo Fisher, Catalog Numbers A25741), and nuclease-free water to achieve the desired volume. The primers used to amplify the studied isoform were the same as those used for Sanger analysis, targeting a 220 bp amplicon. Additionally, the expression of the 249 bp CD44std amplicon was also measured. For CD44std, specific primers targeting the exon-exon junctions s4/s5 for the forward primer and s5/s7 for the reverse primer, were used for amplification. The expression of a 157 bp β-actin amplicon was considered as the endogenous control (please, refer to Supplementary Table for primers details). The accuracy of each primer pair used in this study was validated by assessing their electrophoretic profile after endpoint PCR (data not shown).

The gene expression levels were analyzed using the 2^ΔΔCT method [110].

To compare the expression of CD44 XM_017018585.2 and CD44std in TNBC cell lines, the MCF10A non-tumoral cell line was used as the sample calibrator. In contrast, to evaluate the transfection effect on MDA-MB-468 cells, the parental line was chosen as the reference.

TRANSFECTION AND GENE ENRICHMENT ANALYSIS OF THE BULK RNA-SEQ READS

The synthetic consensus sequence of XM_017018585.2, obtained through the Sanger approach, was purchased from IDT and cloned into the lentiviral plasmid pCCL using the AgeI and SalI restriction sites. To produce lentivirus, HEK293T cells were transfected with Lipofectamine 2000 along with second-generation packaging plasmids (pMDLg/pRRE, pRSV rev, and pMD2.VSVG), and the pCCL CD44 XM_017018585.2 plasmid. The virus-containing supernatant was centrifuged and filtered through 0.45 μm filters. The lentiviral particles were then added to cultures of MDA-MB-468 cells for transformation. The overexpression of CD44 was confirmed through western blot analysis (data not shown).

Subsequently, RNA was extracted from $CD44_{vector}$ and GFP_{vector} transfected cells, as well as the parental line, from two biological replicates (plate A and plate B). The extracted RNA was processed for bulk sequencing using the Illumina platform by the INVIEW Transcriptome Discover service at Eurofins Genomics company. The requested sequencing parameters were 2x 150 bp Paired-end (PE) reads with a guaranteed coverage of 30 million read pairs (+/- 3%).

The resulting sequencing reads were subjected to gene enrichment analysis. First, the quality of the reads was assessed using fastqc. Then, they were aligned using STAR and converted into a count matrix using the GenomicAlignments package in R. The count matrix data was normalized using the counts per million (CPM) function in the R package "limma" [111]. Finally, the log_2FC (log_2 fold change) was

calculated between differentially expressed genes in $CD44_{vector}$ and those commonly DE in GFP_{vector} and the parental line (2 Vs 4) using limma.

Figures

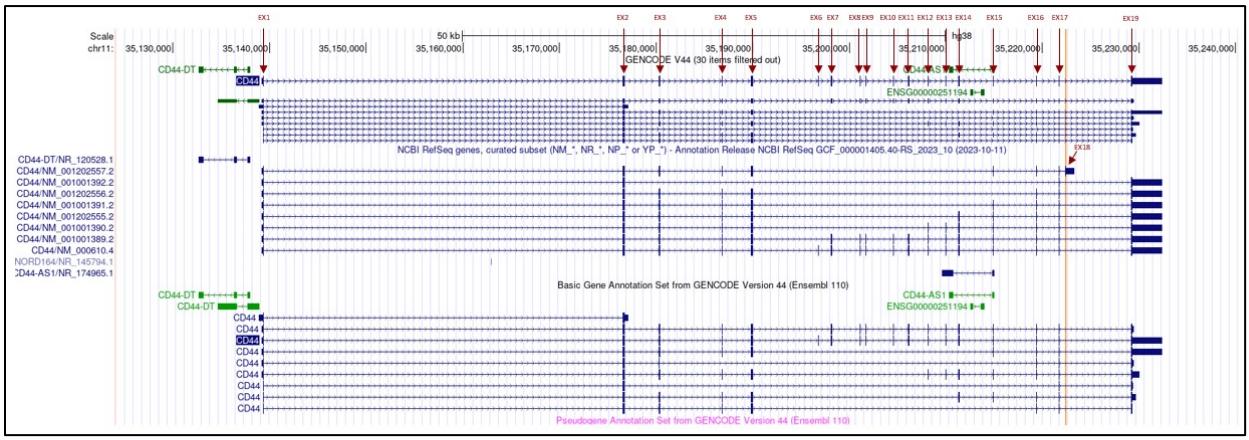

***Figure 1A**: The localization of the nineteen exons of Homo sapiens CD44 can be observed using the UCSC web tool. Each coding exon is indicated by an arrow. It is noteworthy that the ENSE00001598274 exon is present exclusively in transcript NM_001202557.2 (Ensembl_ID: ENST00000442151.6). (Back to text)*

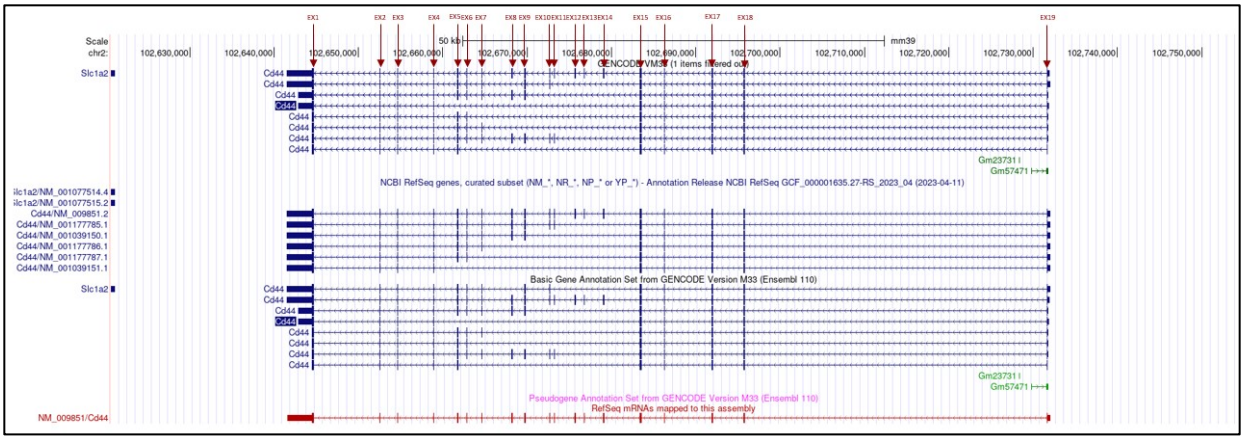

***Figure 1B**: Shows Mus musculus CD44's nineteen exons localization, from UCSC web tool.*

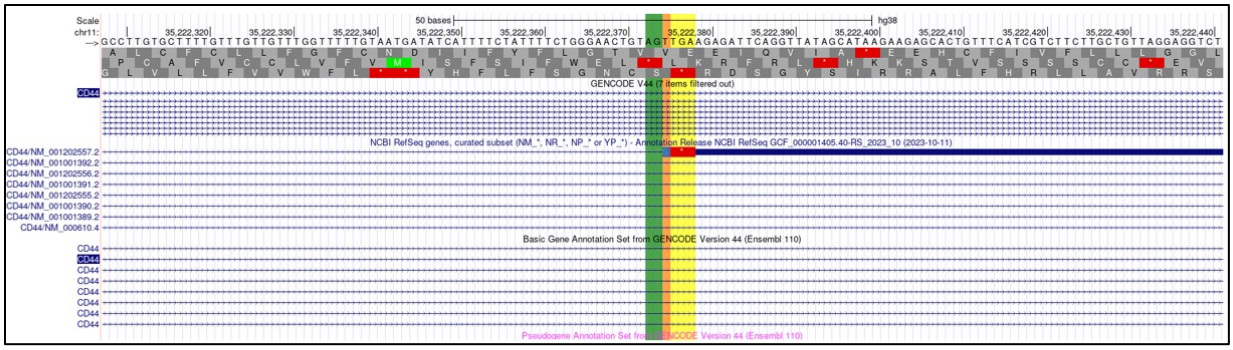

***Figure 1C:** Shows the junction between the last two nucleotides of exon 8 (green) and the first base of exon 9 (orange), which are transcribed as the 5'-AGU-3' codon, encoding for Ser-294 in CD44 antigen isoform H0Y5E4. This is followed by a three-nucleotide sequence, 5'-TGA-3' (yellow), which serves as the stop codon, transcribed as 5'-UGA-3'. (Back to text)*

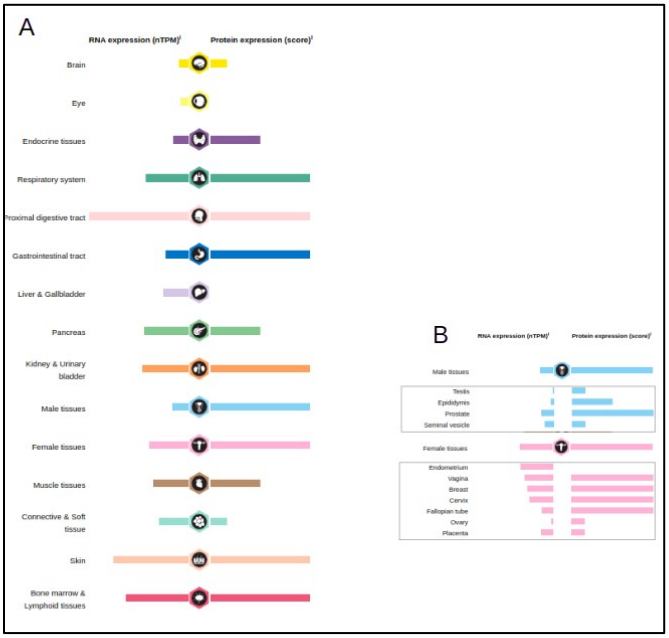

Figure 2*: The image displays the expression levels of CD44 mRNA and protein in various human tissues (A) and the comparison between male and female tissues (B). The RNA expression data was obtained from normalized expression values (nTPM) derived from both the Human Protein Atlas (HPA) RNA-seq data and the Genotype-Tissue Expression (GTEx) project's RNA-seq data. This image has been obtained from the Human Protein Atlas web tool. (Back to text)*

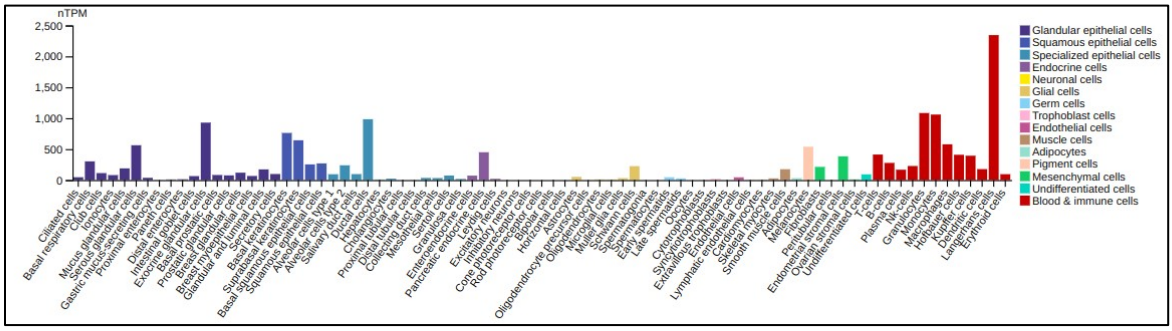

Figure 3: *This image illustrates the distribution of CD44 mRNA across different cytotypes as measured in single-cell sequencing. The expression of CD44 was found to be clustered in blood and immune cells. The image has been obtained from the Human Protein Atlas web tool. (Back to text)*

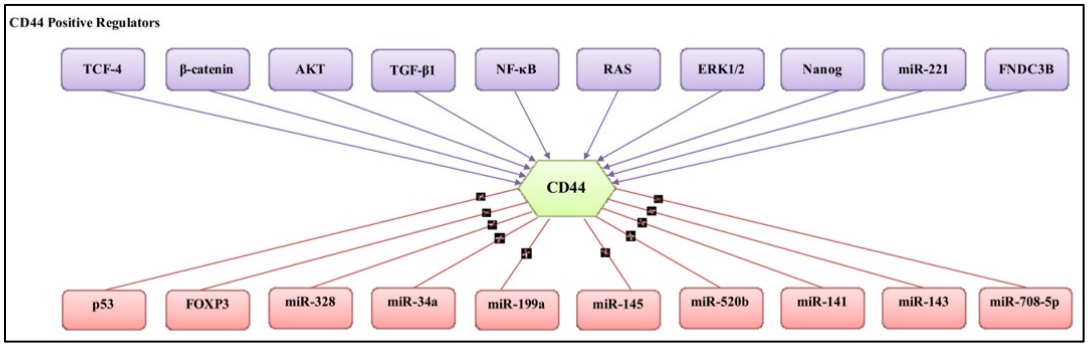

***Figure 4**: The scheme illustrates a selection of factors, protein kinases, cytokines, and miRNA involved in the regulation of CD44 activity. Mesrati et al. "Biomolecules" (2021 Dec 9;11(12):1850) provides further details. The corresponding article's DOI is 10.3390/biom11121850. (Back to text)*

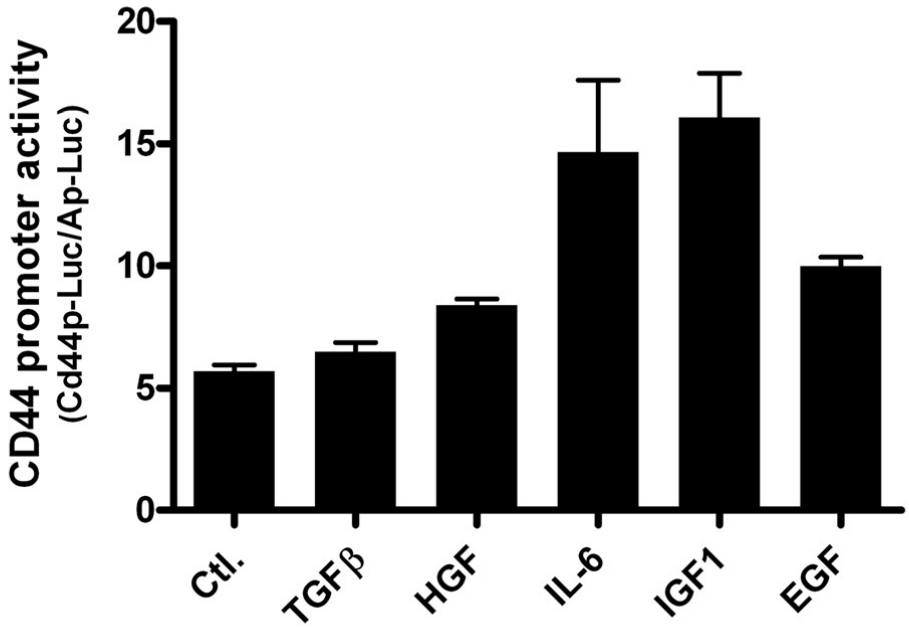

***Figure 5**: CD44 promoter activity comparison among different cytokines treatment. Abbreviation: TGFβ transforming growth factor β, HGF hepatocyte growth factor, IL-6 interleukin-6, IGF1 insulin-like growth factor 1, EGF epidermal growth factor. Chen et al. Oncol. 2018 May 10;11(1):64. provides further details. The corresponding article's DOI is 10.1186/s13045-018-0605-5. (Back to text)*

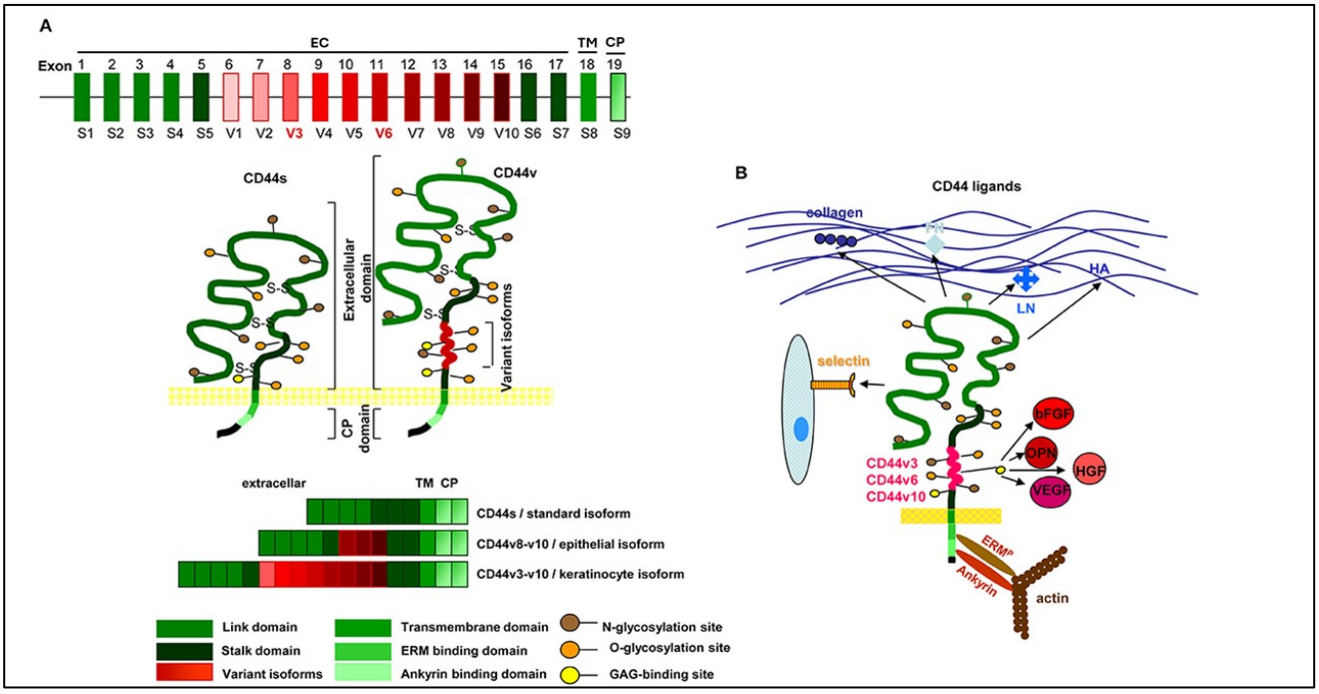

Figure 6: *showcases the genomic organization and protein structure of CD44, including its glycosylation sites, cellular membrane localization, and various combinations of frequently observed CD44v exon products (A). The ectodomain and stem region of the CD44v receptor are responsible for recognizing certain matrix proteins and cellular ligands. Additionally, the C-terminal tail of CD44 binds to cytoskeletal elements such as ERM and ankyrin proteins (B). It can also interact with specific signaling proteins in pathways (not depicted). For more in-depth information, please refer to the study by Wang Z et al (2018 Aug 28;6:97). The article's DOI is 10.3389/fcell.2018.00097.* *(Back to text)*

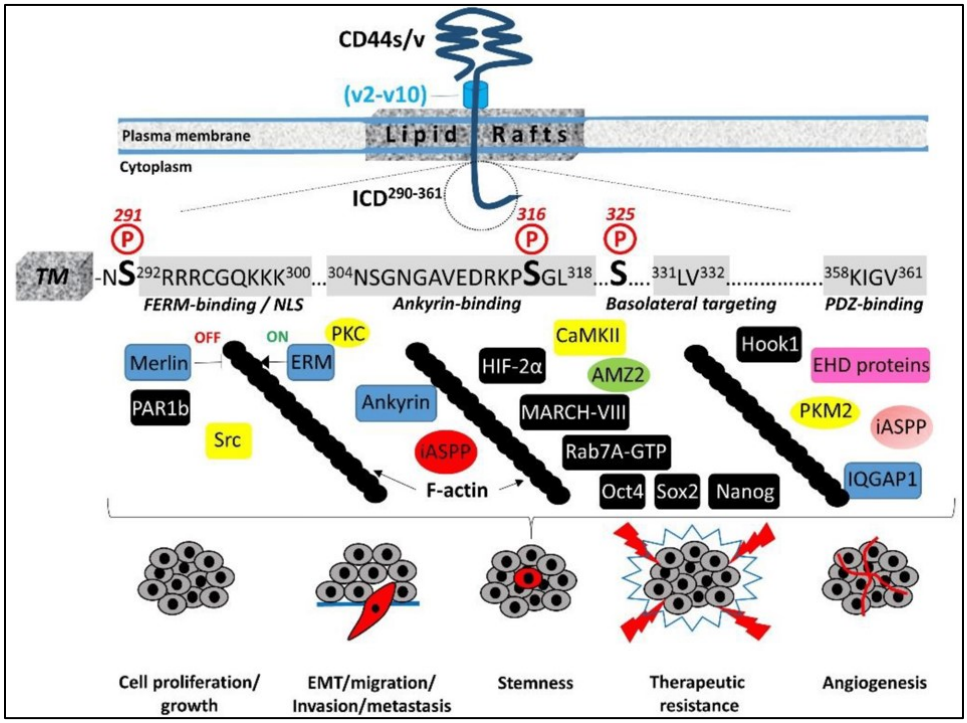

***Figure** 7: Presents the structural motifs of CD44 ICD (intracellular domain) and their interactions with cytoplasmic proteins. The highlighted Ser residues (Ser291, Ser316, and Ser325) are sites prone to phosphorylation. Cytoskeletal proteins are represented in blue, kinases in yellow, and proteins involved in cell trafficking, metabolism, and transcription are depicted (refer to the text for further details). Proteins with interactions yet to be characterized with specific domains or sites of CD44 ICD are indicated in black boxes. The Figure also illustrates the functional significance of CD44 ICD interactions in physiological and pathological processes. For more comprehensive information, please consult the study by Skandalis SS (2023 Oct 18;15(20):5041). The article's DOI is 10.3390/cancers15205041. (Back to text)*

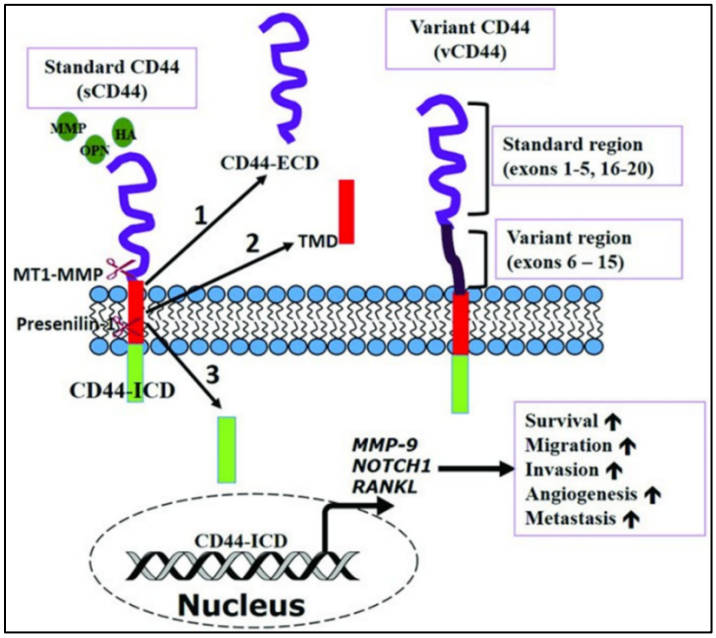

***Figure 6**: CD44 can be sequentially cleaved by membrane type 1 matrix metalloprotease (MT1-MMP) and then presenilin-1/γ secretase induced by ligands [osteopontin (OPN), hyaluronic acid (HA), etc] binding. Cleavage produces (1) extracellular domain (ECD) fragment. (2) CD44β like peptide or transmembrane domain (TMD), and (3) CD44 intracellular domain (ICD) fragment. CD44—ICD translocates into the nucleus to activate transcription of genes important in metastasis and cell survival. Senbanjo LT et al., Front Cell Dev Biol. 2017 Mar 7;5:18. doi: 10.3389/fcell.2017.00018. (Back to text)*

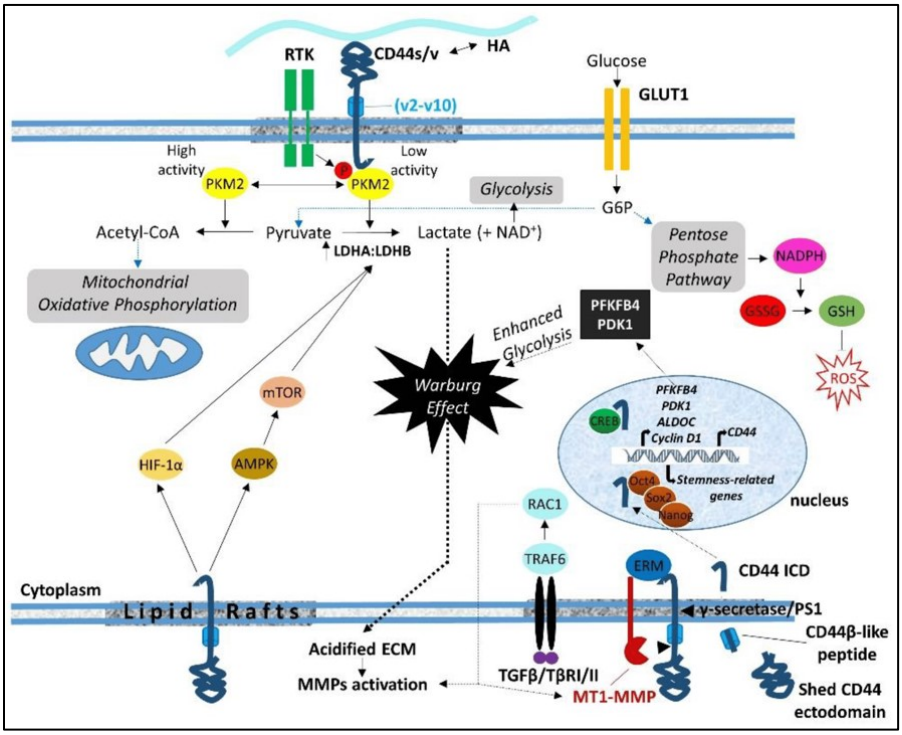

Figure 9: *provides a schematic representation of the roles of CD44 ICD (intracellular domain) in the regulation of the transcriptome and cell metabolism. CD44, through its ICD, influences the metabolic shift from mitochondrial oxidative phosphorylation to aerobic glycolysis, commonly known as the Warburg effect. This regulatory function involves the control of key metabolic enzymes such as PKM2, PFKFB4, and LDH. Additionally, CD44 ICD plays a role in enhancing the flux of the pentose phosphate pathway, leading to NADPH production, which promotes the synthesis of glutathione (GSH) and subsequently suppresses the accumulation of reactive oxygen species (ROS).The increased production of lactate resulting from enhanced aerobic glycolysis contributes to the creation of an acidified pericellular microenvironment, facilitating the activation of matrix metalloproteinases (MMPs). Activation of MMPs can also be triggered by the TGFβ/TβR/TRAF6/RAC1 axis. This activation, in turn, leads to the cleavage of CD44 and the translocation of CD44 ICDf to the nucleus. Once in the nucleus, CD44 ICD regulates the transcription of target genes, including CD44 itself. For further details, please refer to the study by Skandalis SS, with the article's DOI being 10.33/cancers15205041. (Back to text)*

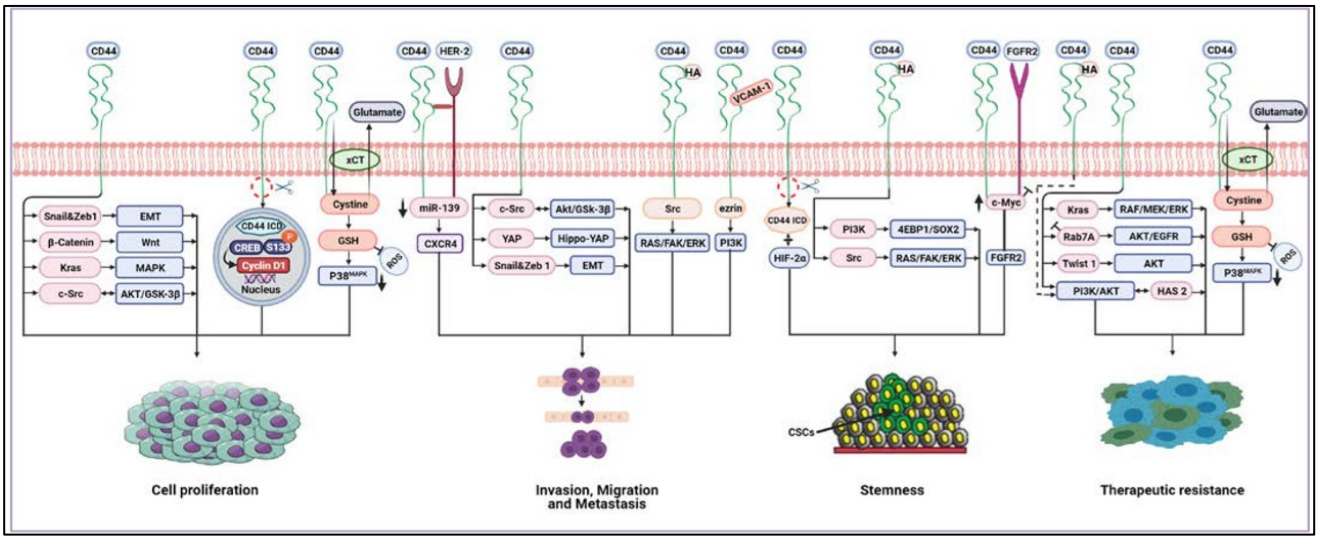

Figure 10*: depicts the cancer-associated signaling pathways that are influenced by CD44. CD44 plays a role in upregulating epithelial-mesenchymal transition (EMT) biomarkers such as Snail1 and Zeb1, thereby promoting proliferation and invasion. Through the MAPK pathway, CD44 activates KRAS, which further enhances tumor cell proliferation and survival. By binding to HER2, CD44 mediates invasion and metastasis, inhibiting miR-139 and upregulating CXCR4. Additionally, there is a positive interaction between CD44 and FGFR2 that sustains cancer stemness.*

CD44 modulates tumor cell proliferation, invasion, and migration by regulating c-Src via the AKT/GSK-3β signaling pathway. The interaction of CD44 with VCAM-1 promotes invasion signaling through the ezrin/PI3K pathway. Binding to hyaluronic acid (HA), CD44 contributes to stemness development via the PI3K/4EBP1/SOX2 pathway. Furthermore, the HA/CD44 interaction drives tumor invasion, metastasis, and stemness by activating the Src/RAS/FAK/ERK pathways. Similarly, HA/CD44 stimulates the PI3K/AKT pathway, leading to increased therapeutic resistance.

CD44 inhibits Rab7A to sustain EGFR and AKT signaling, resulting in therapeutic resistance. CD44 promotes the expression of HAS2 through activation of the PI3K/AKT pathway. HAS2, in turn, enhances CD44-mediated PI3K/AKT signaling, creating a positive feedback loop that drives tumor cell resistance and survival. CD44 stabilizes the cystine/glutamate antiporter (xCT), leading to increased glutathione (GSH) levels and decreased reactive oxygen species (ROS) levels, facilitating tumor cell proliferation and therapeutic resistance by suppressing the p38 pathway.

CD44 regulates tumor cell proliferation via the Wnt/β-catenin signaling pathway and promotes invasion and migration through the Hippo-YAP oncogene signaling pathway. By upregulating Twist1 and AKT signaling, CD44 mediates tumor cell resistance. The CD44 intracellular domain (CD44-ICD) binds to CREB, enhancing S133 phosphorylation and promoting cyclin D1 activity, thereby facilitating cell proliferation. In a hypoxic environment, CD44-ICD is released and binds to HIF-2α, inducing stemness.

For further details, please refer to the study by Hassn Mesrati et al., accessible through the following DOI: https://doi.org/10.3390/biom11121850. (Back to text)

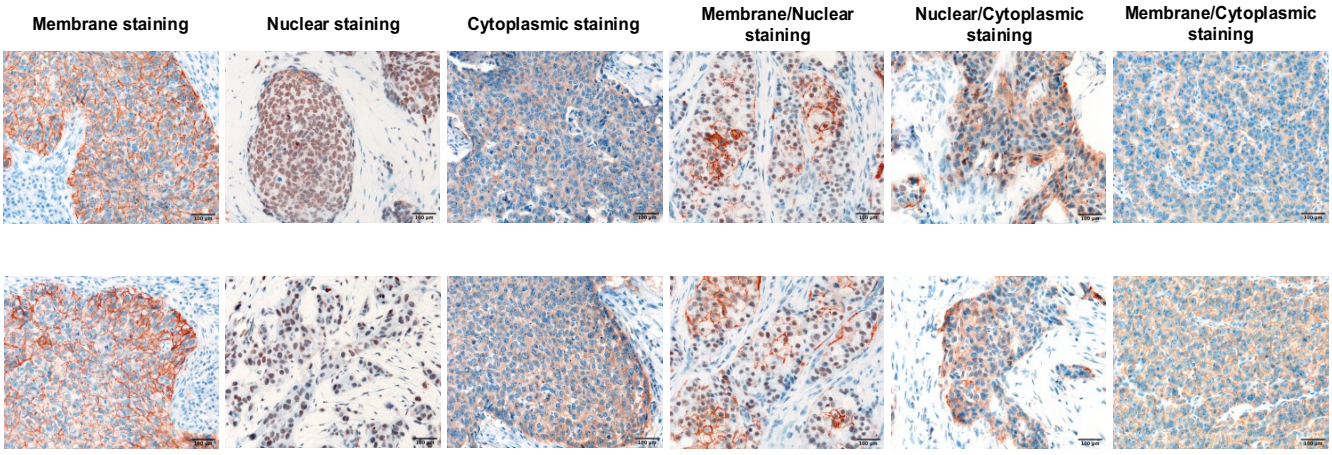

***Figure 7**: Representative microphotographs of IHC staining for CD44v6 showing heterogeneous membrane, nucleus cytoplasmic patterns in TNBC human samples. Original magnification, x200. Scale bar, 100 µm. (Back to text)*

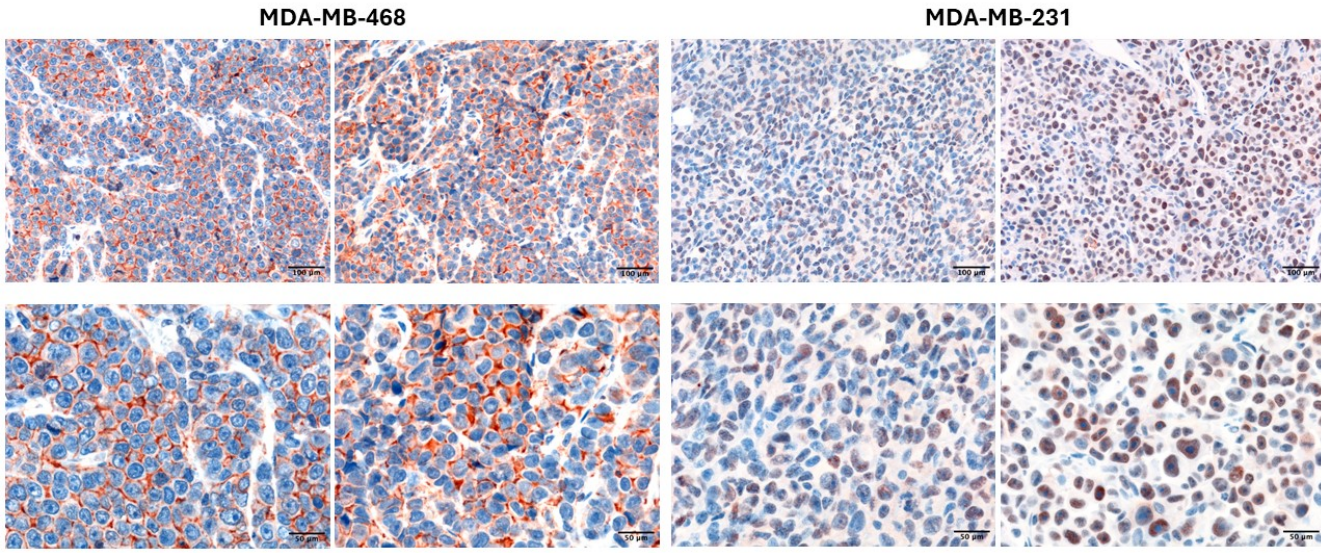

***Figure 8**: Representative microphotographs of IHC staining for CD44v6 showing heterogeneous patterns in MDA-MB-231 and MDA-MB-468 xenograft tumor samples. at 10X, 20X and 40X magnification. Original magnification, x200 and x4000. Scale bar, 100 and 50 µm. (Back to text)*

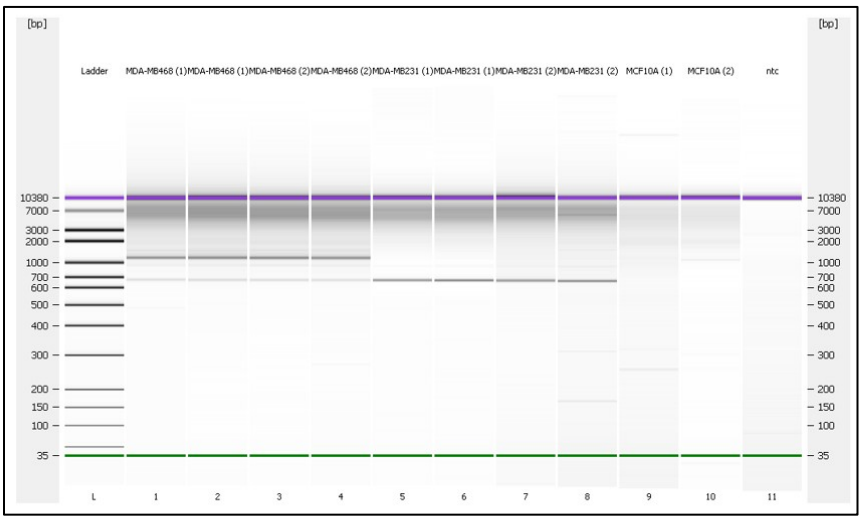

***Figure 9**: Reported the obtained electrophoresis fragments by Bionalyzer DNA 12000 assay. TNB lines have a CD44 PCR profile similar and wider than control line MCF10A. (Back to text)*

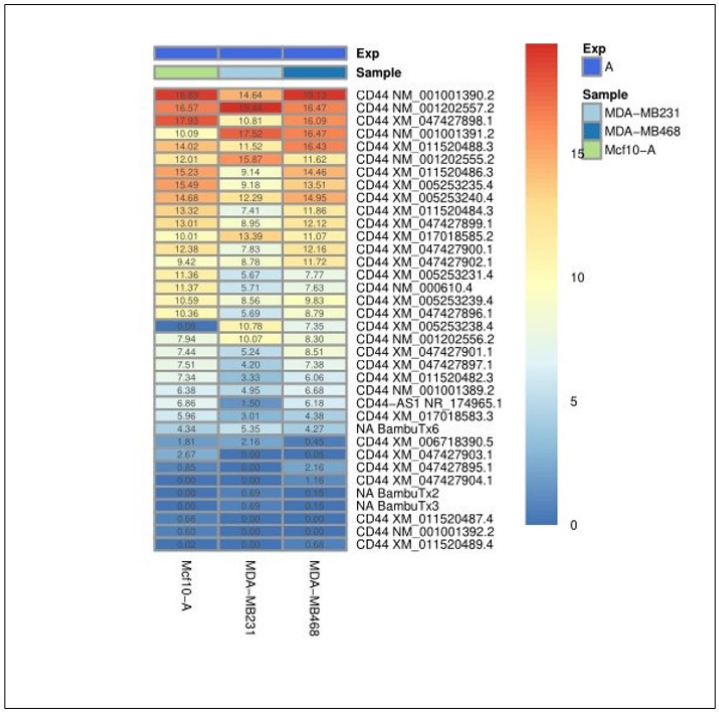

***Figure 10**: Heat map that compare each normalized value per transcript (RPM) alignment to CD44 locus, in the three libraries. Out of the 36 identified reads clusters, three did not align to any annotated transcript.(Back to text)*

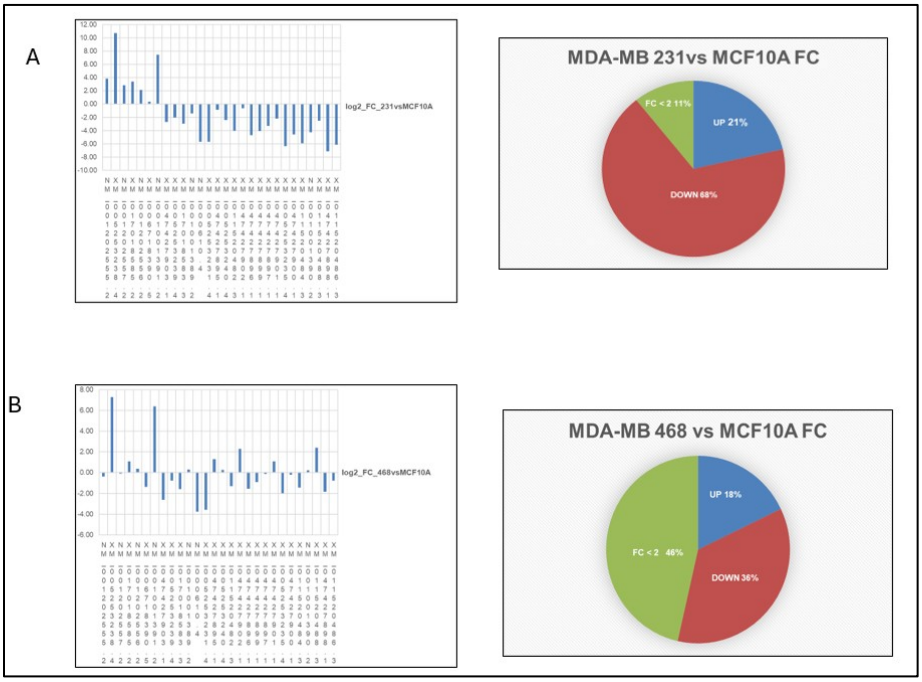

***Figure 11**: Graphs representing the log2 FC values (calculated on RPM index): A) log2 FD of MDA-MB-231 on MCF10A; B) log2 FD of MDA-MB-468 on MCF10.* *(Back to text)*

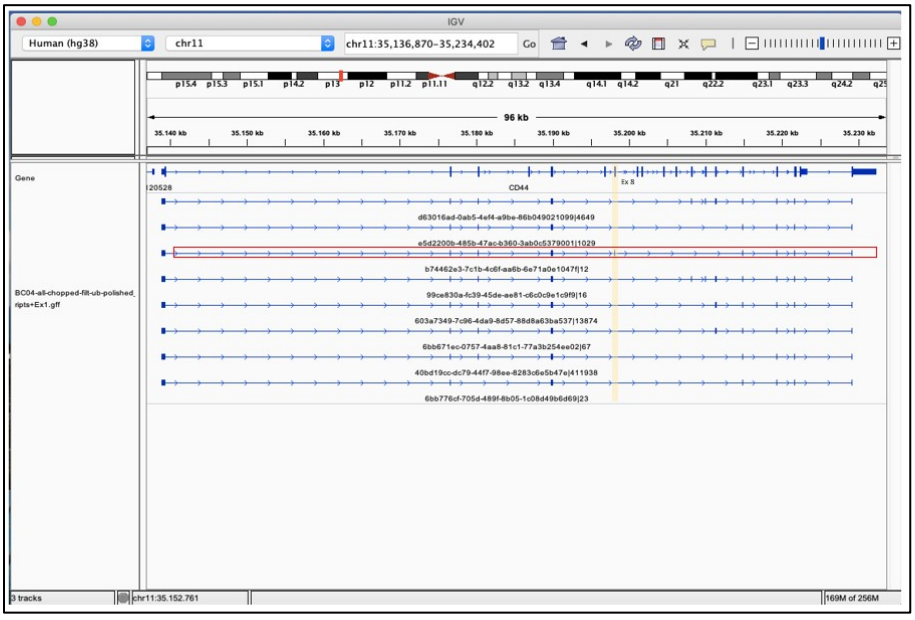

***Figure 16**: IGV window which shows MDA-MB-231 BAM file. A reads clusters maintain the exon junction between CD44 reference EX8 and EX17.* *(Back to text)*

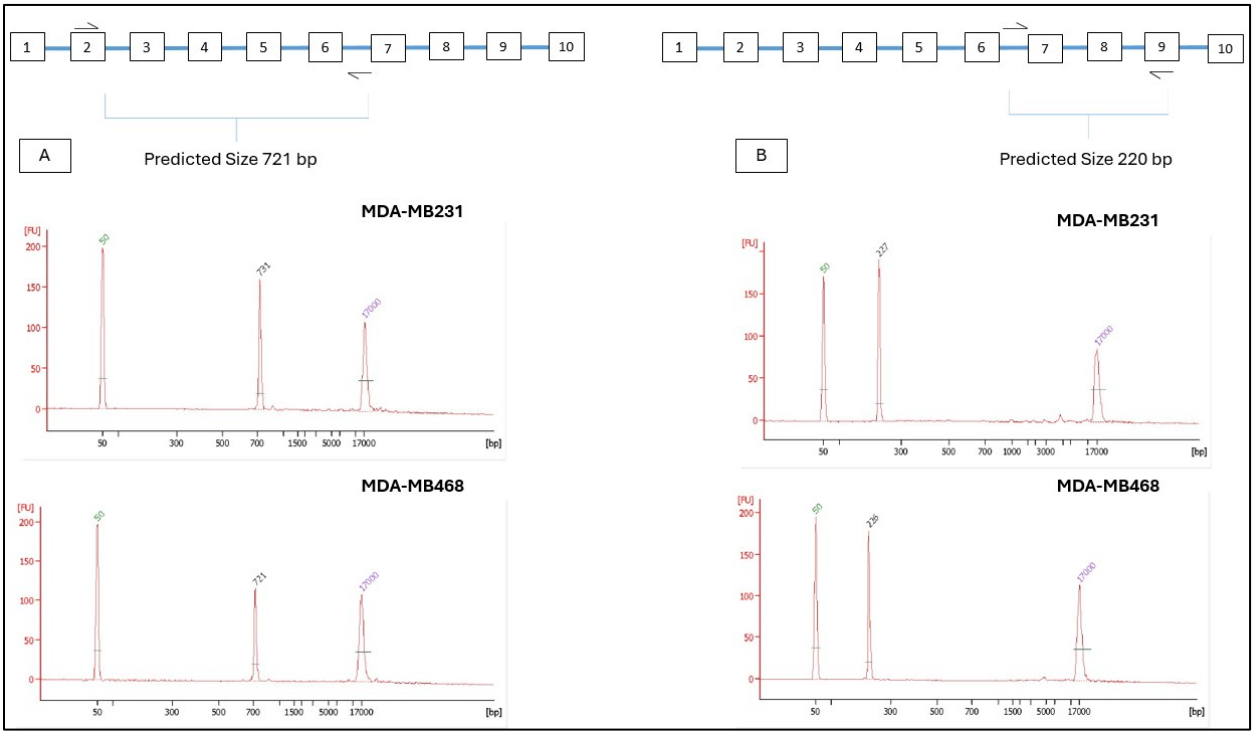

***Figure 17**: Electrophoresis profile to amplicons produced by PCR using primer specific to 5' and 3' region from the exon junction EXs6/7, in MDA-MB-231 (A) and MDA-MB-468 (B) TNBC cells. (Back to text)*

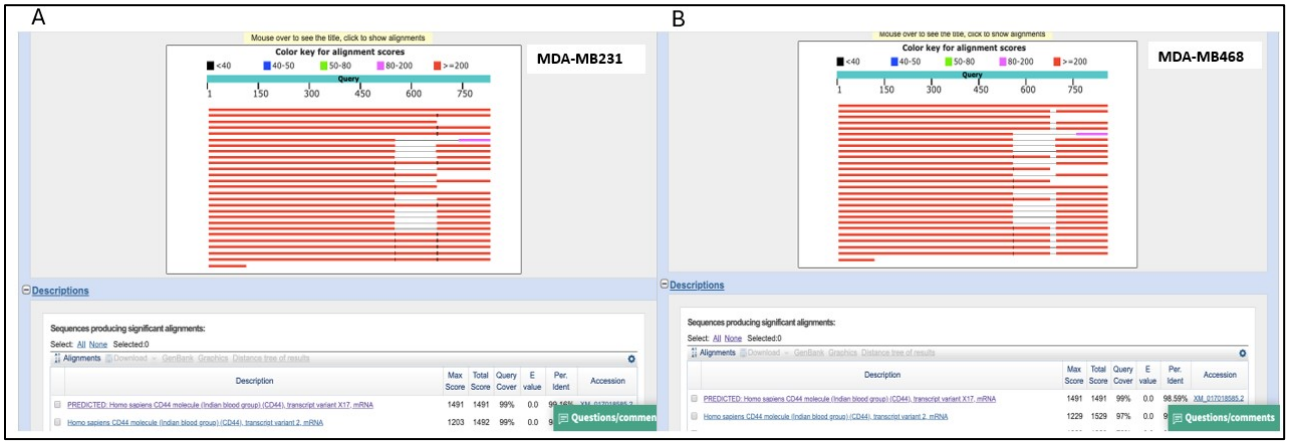

***Figure 18**: Blast output of Sanger sequences obtained with primer specific for the EXs6/7 junction, in MDA-MB-231 (A) and MDA-MB-468 (B). (Back to text)*

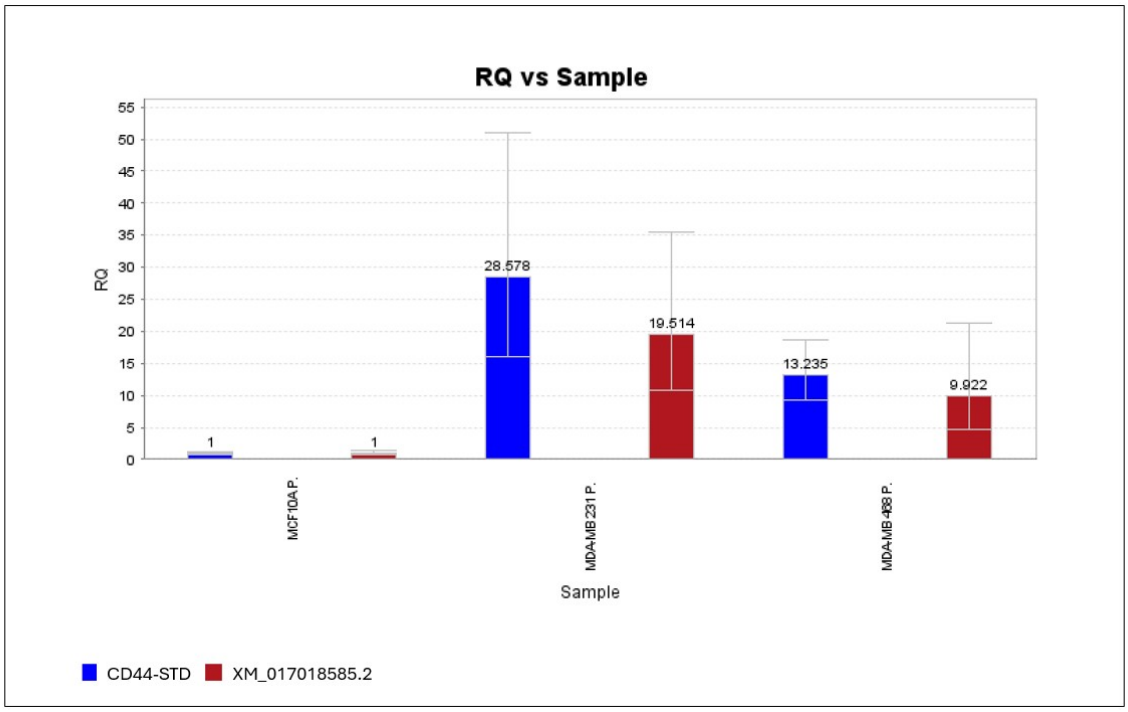

***Figure 19**: Gene expression analysis of the transcripts CD44-STD and XM_017018585.2 in TNBC lines compared to line MCF10A, using a RT-qPCR and 2^ΔΔCT method to the quantification. (Back to text)*

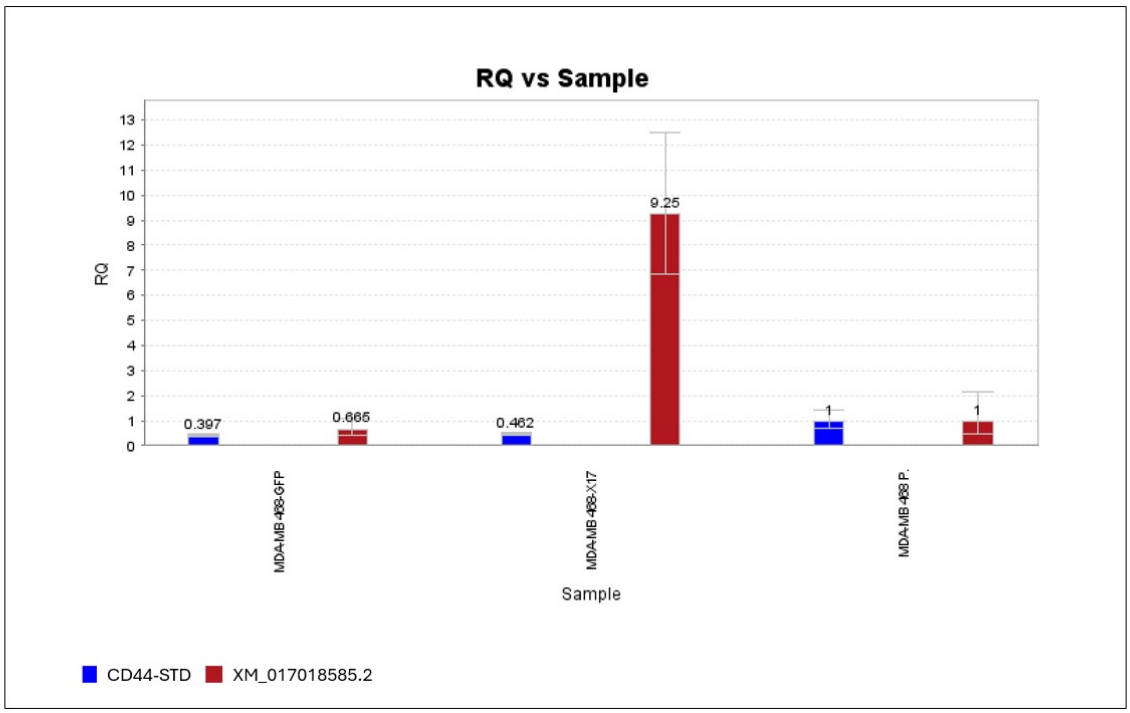

***Figure 20**: Gene expression analysis for the transcripts CD44-STD and XM_017018585.2 in MDA-MB468 transfected with expression vector specific for GFP or CD44 XM_017018585.2 mRNA, using a RT-qPCR approach. For analysis, 2^ΔΔCT method and parental MDA-MB-468 line (as calibrator) were used for the quantification. (Back to text)*

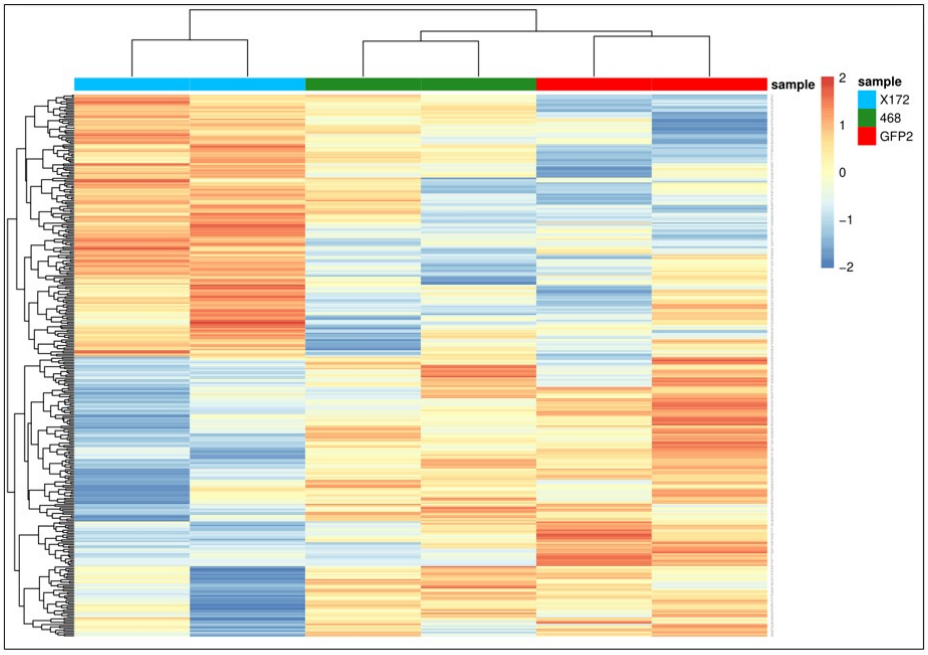

***Figure 21**: Heat map that clusters differentially expressed (DE) genes. The analysis excludes genes that are common between cells treated with Cd44vector (represented by the sky blue cluster) and either cells transfected with GFPvector (represented by the red cluster) or the parental 468 cells (represented by the green cluster).(Back to text)*

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDRq-value
YOSHIMURA_MAPK8_TARGETS_UP [1282]	Genes up-regulated in vascular smooth muscle cells (VSMC) by MAPK8 (JNK1) [GeneID=5599].	16		$7.34\ e^{-7}$	$5.31\ e^{-3}$

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDRq-value
DESCARTES_MAIN_FETAL_CILIATED_EPITHELIAL_CELLS [679]	descartes DE_gene_77_main_cell_type.csv, fold.change>=1.5, qval<0.05, pval<0.05	10		$2.5\ e^{-5}$	$2.08\ e^{-2}$

***Figure 22**: Schematizes the overlaps between the DE up gene set of the CD44vector cells and GSEA collections.(Back to text)*

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDRq-value
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	6		$5.57\ e^{-7}$	$2.78\ e^{-5}$
HALLMARK_FATTY_ACID_METABOLISM [158]	Genes encoding proteins involved in metabolism of fatty acids.	5		$1.28\ e^{-4}$	$3.2\ e^{-3}$
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	5		$3.81\ e^{-4}$	$6.35\ e^{-3}$

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDRq-value
MEK_UP.V1_DN [194]	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] MCF-7 cells (breast cancer) stably over-expressing constitutively active MAP2K1 [Gene ID=5604] gene.	8		$1.58\ e^{-7}$	$1.12\ e^{-5}$
LTE2_UP.V1_DN [195]	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] MCF-7 cells (breast cancer) and long-term adapted for estrogen-independent growth.	8		$1.65\ e^{-7}$	$1.12\ e^{-5}$
ERBB2_UP.V1_DN [197]	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] and engineered to express ligand-activatable ERBB2 [Gene ID=2064].	8		$1.78\ e^{-7}$	$1.12\ e^{-5}$
EGFR_UP.V1_DN [196]	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] and engineered to express ligand-activatable EGFR [Gene ID=1956].	6		$3.21\ e^{-5}$	$1.52\ e^{-3}$
KRAS.DF.V1_UP [191]	Genes up-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	5		$3.09\ e^{-4}$	$1.17\ e^{-2}$
ALK_DN.V1_DN [139]	Genes down-regulated in DAOY cells (medulloblastoma) upon knockdown of ALK [Gene ID=238] gene by RNAi.	4		$8.88\ e^{-4}$	$2.46\ e^{-2}$
KRAS.300_UP.V1_DN [140]	Genes down-regulated in four lineages of epithelial cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	4		$9.12\ e^{-4}$	$2.46\ e^{-2}$
STK33_UP [286]	Genes up-regulated in NOMO-1 and SKM-1 cells (AML) after knockdown of STK33 [Gene ID=65975] by RNAi.	5		$1.88\ e^{-3}$	$4.44\ e^{-2}$

***Figure 23**: Schematizes the overlap between the DE down gene set of the CD44vector cells and GSEA collections.**(Back to text)*

References

1. Skandalis, S. S. CD44 Intracellular Domain: A Long Tale of a Short Tail. *Cancers* **15**, 5041 (2023).

2. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).

3. Hassn Mesrati, M., Syafruddin, S. E., Mohtar, M. A. & Syahir, A. CD44: A Multifunctional Mediator of Cancer Progression. *Biomolecules* **11**, 1850 (2021).

4. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).

5. Ruiz, P., Schwärzler, C. & Günthert, U. CD44 isoforms during differentiation and development. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **17**, 17–24 (1995).

6. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1-1.30.33 (2016).

7. Chen, C., Zhao, S., Karnad, A. & Freeman, J. W. The biology and role of CD44 in cancer progression: therapeutic implications. *J. Hematol. Oncol.J Hematol Oncol* **11**, 64 (2018).

8. PubMed entry.

9. Nishino, M. *et al.* Variant CD44 expression is enriching for a cell population with cancer stem cell-like characteristics in human lung adenocarcinoma. *J. Cancer* **8**, 1774–1785 (2017).

10. Fujita, N. *et al.* Expression of CD44 in normal human versus tumor endometrial tissues: possible implication of reduced expression of CD44 in lymph-vascular space involvement of cancer cells. *Cancer Res.* **54**, 3922–3928 (1994).

11. Wang, L. *et al.* ATDC induces an invasive switch in KRAS-induced pancreatic tumorigenesis. *Genes Dev.* **29**, 171–183 (2015).

12. Godar, S. *et al.* Growth-inhibitory and tumor- suppressive functions of p53 depend on its repression of CD44 expression. *Cell* **134**, 62–73 (2008).

13. Smith, S. M., Lyu, Y. L. & Cai, L. NF-κB affects proliferation and invasiveness of breast cancer cells by regulating CD44 expression. *PloS One* **9**, e106966 (2014).

14. Zhang, C. *et al.* FOXP3 suppresses breast cancer metastasis through downregulation of CD44. *Int. J. Cancer* **137**, 1279–1290 (2015).

15. Li, J. & Zhou, B. P. Activation of β-catenin and Akt pathways by Twist are critical for the maintenance of EMT associated cancer stem cell-like characters. *BMC Cancer* **11**, 49 (2011).

16. Shang, Z. *et al.* A switch from $CD44^+$ cell to EMT cell drives the metastasis of prostate cancer. *Oncotarget* **6**, 1202–1216 (2015).

17. Ishimoto, T. *et al.* Macrophage-derived reactive oxygen species suppress miR-328 targeting CD44 in cancer cells and promote redox adaptation. *Carcinogenesis* **35**, 1003–1011 (2014).

18. Liu, C. *et al.* The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. *Nat. Med.* **17**, 211–215 (2011).

19. Liu, C. *et al.* MicroRNA-141 suppresses prostate cancer stem cells and metastasis by targeting a cohort of pro-metastasis genes. *Nat. Commun.* **8**, 14270 (2017).

20. Lu, Y.-C. *et al.* MiR-520b as a novel molecular target for suppressing stemness phenotype of head-neck cancer by inhibiting CD44. *Sci. Rep.* **7**, 2042 (2017).

21. Cheng, W., Liu, T., Wan, X., Gao, Y. & Wang, H. MicroRNA-199a targets CD44 to suppress the tumorigenicity and multidrug resistance of ovarian cancer-initiating cells. *FEBS J.* **279**, 2047–2059 (2012).

22. Yang, Z. *et al.* MicroRNA-143 targets CD44 to inhibit breast cancer progression and stem cell-like properties. *Mol. Med. Rep.* **13**, 5193–5199 (2016).

23. Kim, J., Jiang, J., Badawi, M. & Schmittgen, T. D. miR-221 regulates CD44 in hepatocellular carcinoma through the PI3K-AKT-mTOR pathway. *Biochem. Biophys. Res. Commun.* **487**, 709–715 (2017).

24. Ouhtit, A., Rizeq, B., Saleh, H. A., Rahman, M. M. & Zayed, H. Novel CD44-downstream signaling pathways mediating breast tumor invasion. *Int. J. Biol. Sci.* **14**, 1782–1790 (2018).

25. Williams, K., Motiani, K., Giridhar, P. V. & Kasper, S. CD44 integrates signaling in normal stem cell, cancer stem cell and (pre)metastatic niches. *Exp. Biol. Med. Maywood NJ* **238**, 324–338 (2013).

26. Wang, Z., Zhao, K., Hackert, T. & Zöller, M. CD44/CD44v6 a Reliable Companion in Cancer-Initiating Cell Maintenance and Tumor Progression. *Front. Cell Dev. Biol.* **6**, 97 (2018).

27. Bourguignon, L. Y. W., Earle, C. & Shiina, M. Activation of Matrix Hyaluronan-Mediated CD44 Signaling, Epigenetic Regulation and Chemoresistance in Head and Neck Cancer Stem Cells. *Int. J. Mol. Sci.* **18**, 1849 (2017).

28. Xu, H., Niu, M., Yuan, X., Wu, K. & Liu, A. CD44 as a tumor biomarker and therapeutic target. *Exp. Hematol. Oncol.* **9**, 36 (2020).

29. Yae, T. *et al.* Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. *Nat. Commun.* **3**, 883 (2012).

30. Chen, K.-L., Li, D., Lu, T.-X. & Chang, S.-W. Structural Characterization of the CD44 Stem Region for Standard and Cancer-Associated Isoforms. *Int. J. Mol. Sci.* **21**, 336 (2020).

31. Mishra, M. N., Chandavarkar, V., Sharma, R. & Bhargava, D. Structure, function and role of CD44 in neoplasia. *J. Oral Maxillofac. Pathol. JOMFP* **23**, 267–272 (2019).

32. Naor, D., Nedvetzki, S., Golan, I., Melnik, L. & Faitelson, Y. CD44 in cancer. *Crit. Rev. Clin. Lab. Sci.* **39**, 527–579 (2002).

33. Legg, J. W. & Isacke, C. M. Identification and functional analysis of the ezrin-binding site in the hyaluronan receptor, CD44. *Curr. Biol. CB* **8**, 705–708 (1998).

34. Yonemura, S. *et al.* Ezrin/radixin/moesin (ERM) proteins bind to a positively charged amino acid cluster in the juxta-membrane cytoplasmic domain of CD44, CD43, and ICAM-2. *J. Cell Biol.* **140**, 885–895 (1998).

35. Ren, M. *et al.* Molecular basis of PIP2-dependent conformational switching of phosphorylated CD44 in binding FERM. *Biophys. J.* **122**, 2675–2685 (2023).

36. Sherman, L. *et al.* Interdomain binding mediates tumor growth suppression by the NF2 gene product. *Oncogene* **15**, 2505–2509 (1997).

37. Hartmann, M. *et al.* Tumor Suppressor NF2 Blocks Cellular Migration by Inhibiting Ectodomain Cleavage of CD44. *Mol. Cancer Res. MCR* **13**, 879–890 (2015).

38. Yin, F. *et al.* Spatial organization of Hippo signaling at the plasma membrane mediated by the tumor suppressor Merlin/NF2. *Cell* **154**, 1342–1355 (2013).

39. Ooki, T., Murata-Kamiya, N., Takahashi-Kanemitsu, A., Wu, W. & Hatakeyama, M. High-Molecular-Weight Hyaluronan Is a Hippo Pathway Ligand Directing Cell Density-Dependent Growth Inhibition via PAR1b. *Dev. Cell* **49**, 590-604.e9 (2019).

40. Lewis, C. A., Townsend, P. A. & Isacke, C. M. Ca(2+)/calmodulin-dependent protein kinase mediates the phosphorylation of CD44 required for cell migration on hyaluronan. *Biochem. J.* **357**, 843–850 (2001).

41. Guo, Y. J., Lin, S. C., Wang, J. H., Bigby, M. & Sy, M. S. Palmitoylation of CD44 interferes with CD3-mediated signaling in human T lymphocytes. *Int. Immunol.* **6**, 213–221 (1994).

42. Senbanjo, L. T. & Chellaiah, M. A. CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells. *Front. Cell Dev. Biol.* **5**, 18 (2017).

43. Lee, J.-L., Wang, M.-J. & Chen, J.-Y. Acetylation and activation of STAT3 mediated by nuclear translocation of CD44. *J. Cell Biol.* **185**, 949–957 (2009).

44. Miletti-González, K. E. *et al.* Identification of function for CD44 intracytoplasmic domain (CD44-ICD): modulation of matrix metalloproteinase 9 (MMP-9) transcription via novel promoter response element. *J. Biol. Chem.* **287**, 18995–19007 (2012).

45. Okamoto, I. *et al.* Proteolytic cleavage of the CD44 adhesion molecule in multiple human tumors. *Am. J. Pathol.* **160**, 441–447 (2002).

46. De Falco, V. *et al.* CD44 proteolysis increases CREB phosphorylation and sustains proliferation of thyroid cancer cells. *Cancer Res.* **72**, 1449–1458 (2012).

47. Johansson, E. *et al.* CD44 Interacts with HIF-2α to Modulate the Hypoxic Phenotype of Perinecrotic and Perivascular Glioma Cells. *Cell Rep.* **20**, 1641–1653 (2017).

48. Bourguignon, L. Y. W., Wong, G., Earle, C. & Chen, L. Hyaluronan-CD44v3 interaction with Oct4-Sox2-Nanog promotes miR-302 expression leading to self-renewal, clonal formation, and cisplatin resistance in cancer stem cells from head and neck squamous cell carcinoma. *J. Biol. Chem.* **287**, 32800–32824 (2012).

49. Cho, Y. *et al.* Cleaved CD44 intracellular domain supports activation of stemness factors and promotes tumorigenesis of breast cancer. *Oncotarget* **6**, 8709–8721 (2015).

50. Evanko, S. P., Potter-Perigo, S., Petty, L. J., Workman, G. A. & Wight, T. N. Hyaluronan Controls the Deposition of Fibronectin and Collagen and Modulates TGF-β1 Induction of Lung Myofibroblasts. *Matrix Biol. J. Int. Soc. Matrix Biol.* **42**, 74–92 (2015).

51. Tamada, M. *et al.* Modulation of glucose metabolism by CD44 contributes to antioxidant status and drug resistance in cancer cells. *Cancer Res.* **72**, 1438–1448 (2012).

52. Skandalis, S. S., Kozlova, I., Engström, U., Hellman, U. & Heldin, P. Proteomic identification of CD44 interacting proteins. *IUBMB Life* **62**, 833–840 (2010).

53. Christofk, H. R. *et al.* The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230–233 (2008).

54. Nam, K., Oh, S. & Shin, I. Ablation of CD44 induces glycolysis-to-oxidative phosphorylation transition via modulation of the c-Src-Akt-LKB1-AMPKα pathway. *Biochem. J.* **473**, 3013–3030 (2016).

55. Semenza, G. L. *et al.* Hypoxia response elements in the aldolase A, enolase 1, and lactate dehydrogenase A gene promoters contain essential binding sites for hypoxia-inducible factor 1. *J. Biol. Chem.* **271**, 32529–32537 (1996).

56. Gatenby, R. A. & Gillies, R. J. Why do cancers have high aerobic glycolysis? *Nat. Rev. Cancer* **4**, 891–899 (2004).

57. Wang, W. *et al.* Internalized CD44s splice isoform attenuates EGFR degradation by targeting Rab7A. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8366–8371 (2017).

58. Gomez, K. E. *et al.* Cancer Cell CD44 Mediates Macrophage/Monocyte-Driven Regulation of Head and Neck Cancer Stem Cells. *Cancer Res.* **80**, 4185–4198 (2020).

59. Nam, K., Oh, S., Lee, K., Yoo, S. & Shin, I. CD44 regulates cell proliferation, migration, and invasion via modulation of c-Src transcription in human breast cancer cells. *Cell. Signal.* **27**, 1882–1894 (2015).

60. Park, J. *et al.* A reciprocal regulatory circuit between CD44 and FGFR2 via c-myc controls gastric cancer cell growth. *Oncotarget* **7**, 28670–28683 (2016).

61. Orian-Rousseau, V. & Sleeman, J. CD44 is a multidomain signaling platform that integrates extracellular matrix cues with growth factor and cytokine signals. *Adv. Cancer Res.* **123**, 231–254 (2014).

62. Orian-Rousseau, V. CD44, a therapeutic target for metastasising tumours. *Eur. J. Cancer Oxf. Engl. 1990* **46**, 1271–1277 (2010).

63. Ma, L., Dong, L. & Chang, P. CD44v6 engages in colorectal cancer progression. *Cell Death Dis.* **10**, 30 (2019).

64. Thanee, M. *et al.* CD44 variant-dependent redox status regulation in liver fluke-associated cholangiocarcinoma: A target for cholangiocarcinoma treatment. *Cancer Sci.* **107**, 991–1000 (2016).

65. Okabe, H. *et al.* CD44s signals the acquisition of the mesenchymal phenotype required for anchorage-independent cell survival in hepatocellular carcinoma. *Br. J. Cancer* **110**, 958–966 (2014).

66. Lai, C.-J. *et al.* CD44 Promotes Migration and Invasion of Docetaxel-Resistant Prostate Cancer Cells Likely via Induction of Hippo-Yap Signaling. *Cells* **8**, 295 (2019).

67. Chang, G. *et al.* CD44 targets Wnt/β-catenin pathway to mediate the proliferation of K562 cells. *Cancer Cell Int.* **13**, 117 (2013).

68. Gomari, M. M. *et al.* CD44 polymorphisms and its variants, as an inconsistent marker in cancer investigations. *Mutat. Res. Rev. Mutat. Res.* **787**, 108374 (2021).

69. Cirillo, N. The Hyaluronan/CD44 Axis: A Double-Edged Sword in Cancer. *Int. J. Mol. Sci.* **24**, 15812 (2023).

70. Gerardo-Ramírez, M. *et al.* CD44 Contributes to the Regulation of MDR1 Protein and Doxorubicin Chemoresistance in Osteosarcoma. *Int. J. Mol. Sci.* **23**, 8616 (2022).

71. Qu, C. *et al.* Extensive CD44-dependent hyaluronan coats on human bone marrow-derived mesenchymal stem cells produced by hyaluronan synthases HAS1, HAS2 and HAS3. *Int. J. Biochem. Cell Biol.* **48**, 45–54 (2014).

72. Endo, K. & Terada, T. Protein expression of CD44 (standard and variant isoforms) in hepatocellular carcinoma: relationships with tumor grade, clinicopathologic parameters, p53 expression, and patient survival. *J. Hepatol.* **32**, 78–84 (2000).

73. Zhang, H. *et al.* CD44 splice isoform switching determines breast cancer stem cell state. *Genes Dev.* **33**, 166–179 (2019).

74. Yan, Y., Zuo, X. & Wei, D. Concise Review: Emerging Role of CD44 in Cancer Stem Cells: A Promising Biomarker and Therapeutic Target. *Stem Cells Transl. Med.* **4**, 1033–1043 (2015).

75. Tamada, M., Suematsu, M. & Saya, H. Pyruvate kinase M2: multiple faces for conferring benefits on cancer cells. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **18**, 5554–5561 (2012).

76. Ishimoto, T. *et al.* CD44 variant regulates redox status in cancer cells by stabilizing the xCT subunit of system xc(-) and thereby promotes tumor growth. *Cancer Cell* **19**, 387–400 (2011).

77. Chaudhry, G.-E.-S. *et al.* Understanding Hyaluronan Receptor (CD44) Interaction, HA-CD44 Activated Potential Targets in Cancer Therapeutics. *Adv. Pharm. Bull.* **11**, 426–438 (2021).

78. Brown, R. L. *et al.* CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J. Clin. Invest.* **121**, 1064–1074 (2011).

79. Zhao, P. *et al.* The CD44s splice isoform is a central mediator for invadopodia activity. *J. Cell Sci.* **129**, 1355–1365 (2016).

80. Lv, L. *et al.* Upregulation of CD44v6 contributes to acquired chemoresistance via the modulation of autophagy in colon cancer SW480 cells. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **37**, 8811–8824 (2016).

81. Yanamoto, S. *et al.* Expression of the cancer stem cell markers CD44v6 and ABCG2 in tongue cancer: effect of neoadjuvant chemotherapy on local recurrence. *Int. J. Oncol.* **44**, 1153–1162 (2014).

82. Ozawa, M. *et al.* Prognostic significance of CD44 variant 2 upregulation in colorectal cancer. *Br. J. Cancer* **111**, 365–374 (2014).

83. Zhao, P. *et al.* CD44 promotes Kras-dependent lung adenocarcinoma. *Oncogene* **32**, 5186–5190 (2013).

84. Wang, S. J., Wreesmann, V. B. & Bourguignon, L. Y. W. Association of CD44 V3-containing isoforms with tumor cell growth, migration, matrix metalloproteinase expression, and lymph node metastasis in head and neck cancer. *Head Neck* **29**, 550–558 (2007).

85. Hu, J., Li, G., Zhang, P., Zhuang, X. & Hu, G. A CD44v+ subpopulation of breast cancer stem-like cells with enhanced lung metastasis capacity. *Cell Death Dis.* **8**, e2679 (2017).

86. Escudero Mendez, L. *et al.* Evaluation of CD44+/CD24- and Aldehyde Dehydrogenase Enzyme Markers in Cancer Stem Cells as Prognostic Indicators for Triple-Negative Breast Cancer. *Cureus* **14**, e28056 (2022).

87. American Cancer Society. *Book American Cancer Society. Cancer Treatment & Survivorship Facts & Figures 2022-2024*. (Atlanta, 2022).

88. Cailleau, R., Young, R., Olivé, M. & Reeves, W. J. Breast tumor cell lines from pleural effusions. *J. Natl. Cancer Inst.* **53**, 661–674 (1974).

89. Cailleau, R., Olivé, M. & Cruciger, Q. V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* **14**, 911–915 (1978).

90. Soule, H. D. & McGrath, C. M. A simplified method for passage and long-term growth of human mammary epithelial cells. *Vitro Cell. Dev. Biol. J. Tissue Cult. Assoc.* **22**, 6–12 (1986).

91. Oxford Nanopore Technologies. Dorado.

92. Oxford Nanopore Technologies. nfcore nanoseq.

93. Illumina. Considerations for RNA Seq read length and coverage. (2024).

94. htslib. Samtools.

95. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

96. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

97. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

98. Yoshimura, K. *et al.* Regression of abdominal aortic aneurysm by inhibition of c-Jun N-terminal kinase. *Nat. Med.* **11**, 1330–1338 (2005).

99. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).

100. Einav, U. *et al.* Gene expression analysis reveals a strong signature of an interferon-induced pathway in childhood lymphoblastic leukemia as well as in breast and ovarian cancer. *Oncogene* **24**, 6367–6375 (2005).

101. Lee, A. V., Oesterreich, S. & Davidson, N. E. MCF-7 cells--changing the course of breast cancer research and care for 45 years. *J. Natl. Cancer Inst.* **107**, djv073 (2015).

102. Bellerby, R. *et al.* Overexpression of Specific CD44 Isoforms Is Associated with Aggressive Cell Features in Acquired Endocrine Resistance. *Front. Oncol.* **6**, 145 (2016).

103. Iida, J. *et al.* DNA aptamers against exon v10 of CD44 inhibit breast cancer cell migration. *PloS One* **9**, e88712 (2014).

104. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinforma. Oxf. Engl.* **35**, 2193–2198 (2019).

105. Depledge, D. P. & Wilson, A. C. Using Direct RNA Nanopore Sequencing to Deconvolute Viral Transcriptomes. *Curr. Protoc. Microbiol.* **57**, e99 (2020).

106. ncbi. NLM GenBank and SRA Data Processing. (2023).

107. L, C. *et al.* WNT signaling modulates PD-L1 expression in the stem cell compartment of triple-negative breast cancer. *Oncogene* **38**, (2019).

108. GATK broadistitute. PrintReads.

109. eurofins Genomics. Mix2Seq service.

110. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* **3**, 1101–1108 (2008).

111. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).