UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Curso de Biotecnologia

Trabalho de Conclusão de Curso

Análise, *in silico*, dos padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs e seus possíveis papéis no câncer de mama.

Monize Provisor

Pelotas, 2016

Monize Provisor

Análise, *in silico*, dos padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs e seus possíveis papéis no câncer de mama.

Trabalho de Conclusão de Curso apresentado ao Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Biotecnologia

Orientador Acadêmico: Prof. Dr. Luciano da Silva Pinto
Orientador de Estágio: Prof. Dr. Wilson Araújo da Silva Junior

Pelotas, 2016

Dados de catalogação na fonte:
Ubirajara Buddin Cruz – CRB-10/901
Biblioteca de Ciência & Tecnologia - UFPel

S237a Santos, Monize Nakamoto Provisor
Análise, in silico, dos padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs e seus possíveis papéis no câncer de mama / Monize Nakamoto Provisor Santos. – 47f. : il. – Trabalho de conclusão de curso (Graduação em Biotecnologia). Universidade Federal de Pelotas. Centro de Desenvolvimento Tecnológico. Pelotas, 2015. – Orientador Luciano da Silva Pinto.

1.Biotecnologia. 2.Metástase. 3.RNA-Seq. 4.Bioinformática. 5.Expressão. I.Pinto, Luciano da Silva. II.Título.

CDD: 616.994

Dedico este trabalho à minha família, aos meus queridos amigos e a todos que contribuiram para minha formação profissional.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Luciano da Silva Pinto, pela orientação, paciência, amizade e confiança depositada. Por ter sido o responsável por despertar meu interesse na área de bioinformática e sempre me incentivar a procurar por novos conhecimentos e caminhos;

Ao meu orientador, Prof. Dr. Wilson Araújo da Silva Junior, pela oportunidade, generosidade e disponibilidade de orientação.

À minha co-orientadora, Jessica Rodrigues Plaça, pela paciência, amizade, generosidade e disponibilidade para uma rica orientação. Pela disposição em compartilhar seu conhecimento e pelo tempo em que passamos juntas. Muito obrigada.

Aos meus pais, Clara e Alberto, pelo apoio, carinho, amor, educação e confiança. Por serem minha base e, apesar da distância, sempre me apoiaram nos momentos mais difíceis. Tudo o que eu sou hoje devo a vocês.

Ao Laboratório de Bioinformática e Proteômica e ao apoio de todos os colegas de trabalho, em especial ao Marcus Eslabão, Frederico Kremer, Rafael Woloski, Julia Labonde e Paulo Ricardo Porto. Agradeço pela ajuda, pelo conhecimento compartilhado, pela paciência e amizade, pelas risadas e, principalmente, por terem me recebido com muito carinho;

Ao Laboratório de Genética Molecular e Bioinformática e ao apoio e carinho das colegas de trabalho Juliana, Maithê e Marcelo. Muito obrigada!

Aos meus queridos amigos que construí durante a graduação, em especial Giovanni, Luiza, Bruno, Carlos e Roberta. Obrigada pela amizade, companheirismo, paciência e por serem a minha segunda família;

Aos meus amigos que construí ao longo da vida, em especial Gisele, Renê, Manuella e Felipe. Obrigada pelo amor, carinho, companheirismo, conselhos e, principalmente, por, apesar da distância, sempre me apoiarem nos momentos difíceis e torcerem pelo meu sucesso a cada etapa.

Ao Ivan, pelo apoio, paciência, amizade, carinho e amor. Obrigada por estar sempre ao meu lado nos momentos difíceis, me incentivar a cada desafio e por ter se tornado parte da minha família;

A todos aqueles que foram fundamentais para a minha formação profissional, pelos conselhos e puxões de orelha, companheirismo, amizade, carinho e disposição em compartilhar seus conhecimentos;

À Universidade Federal de Pelotas e à Universidade de São Paulo.

RESUMO

PROVISOR, Monize. **Análise, *in silico*, dos padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs e seus possíveis papéis no câncer de mama.**. 2016. 48 f. Trabalho de Conclusão de Curso (Biotecnologia) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2016.

A metástase no câncer de mama é considerada o fator responsável por causar a maior taxa de letalidade entre mulheres. A atuação conjunta das proteínas Metaloproteinases de Matriz (MMPs), que apresentam forte interação com componentes da Matriz Extracelular, das Proteínas Desintegrinas e Metaloproteínases (ADAMTSs) e das proteínas TIMPS, capazes de inibir a ação das MMPs, é responsável por exercer um importante papel proteolítico no ambiente tumoral. Com a introdução da bioinformática e das novas tecnologias de sequenciamento tornou-se possível a busca por assinaturas gênicas para auxiliar nos métodos prognósticos. O presente estudo objetiva verificar, *in silico*, os padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs, em tumores metastáticos, a partir de amostras de tecidos de carcinoma invasivo de mama e de tecidos saudáveis adjacentes. As análises diferenciais de expressão globais e específicas, de clusterização e funcionais foram realizadas a partir de dados de RNA mensageiro (mRNA) contendo amostras tumorais e normais pareadas, através da linguagem R de programação, de testes (exactTest e Mann-Whitney) e parâmetros estatísticos (Correlação de Pearson, Fold Change e p-value), além de pacotes presentes no software Bioconductor (TCGAbiolinks, edgeR e EDAseq). A análise funcional considerando todos os genes demonstrou, para os principais genes diferencialmente expressos, funções e vias característicos das família gênicas alvo do estudo, ressaltando a importância destas na progressão tumoral, sendo as famílias gênicas de interesse capazes de diferenciar amostras normais de amostras tumorais.

Palavras-chave: Metástase, RNA-Seq, Bioinformática, Metaloproteinases.

ABSTRACT

PROVISOR, Monize. **Analysis of differential expression patterns of TIMPs, ADAMTs e MMPs genes and their possible roles in Breast Cancer**. 2016. 48 f. Trabalho de Conclusão de Curso (Biotecnologia) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2016.

The metastasis in breast cancer is considered the responsible factor for causing the highest rate of mortality among women. The combined action of Metalloproteinases Matrix proteins (MMPs), which have strong interactions with components of the Extracellular Matrix, the Disintegrins and Metalloproteinases Proteins (ADAMTSs) and TIMP proteins, capable of inhibiting the action of MMPs, is responsible for an important proteolytic role in the tumor environment. Through the introduction of bioinformatics and new sequencing technologies, it's now possible to search gene signatures to support the prognostic methods. This study aims to verify, in silico, the patterns of expression of the gene families TIMPs, ADAMTSs and MMPs in metastatic tumors from invasive breast carcinoma tissue and adjacent healthy tissues samples. Both analyzis of differential expression, global and specific, besides clustering and functional were made with messenger RNA (mRNA) data from tumor and normal samples paired, through the R programming language, tests (exactTest and Mann-Whitney) and statistical parameters (Correlation Pearson, fold change and p-value), and software packages present in Bioconductor (TCGAbiolinks, edgeR and EDAseq). The visualization of the results was obtained by Graphics like Volcano plot, Heatmap and Barplot. The functional analysis considering all genes showed, for the main differentially expressed genes, characteristic functions and pathways of gene families that are target of this study, highlighting the importance of these in tumor progression, where the gene families of interest are able to differentiate normal samples of tumor samples.

Keywords: Metastasis, RNA-Seq, Bioinformatics, Metalloproteases.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DE	Diferencialmente Expressos
DNA-Seq	Sequenciamento de DNA
ECM	Matriz Extracelular
ELISA	Ensaio Imunoenzimático ligado a Enzimas
ER	Receptor de estrogênico
HER2	Fator de Crescimento Epidermal 2
MLG	Modelos Lineares Generalizados
MMPs	Metaloproteinases de Matriz
NGS	Sequenciamento de nova geração
PCR	Reação em cadeia de polimerase
PGR	Receptor de progesterona
RNA-Seq	Sequenciamento de RNA
SAGE	Análises Seriais de Expressão Gênica
TIMPs	Inibidores Teciduais de Metaloproteinases
TMA	Técnica de Microarranjo Tecidual

SUMÁRIO

1 INTRODUÇÃO

O câncer de mama é considerado o segundo câncer mais frequente a nível mundial e a principal causa de morte por câncer entre mulheres (JEMAL et al., 2010). São estimados 57.120 novos casos para o ano de 2014, o que indica uma taxa de incidência de 56,1 casos por 100.000 mulheres (INCA, 2014). É uma doença heterogênea e complexa, sendo a heterogeneidade muitas vezes atribuída a existência de diferentes subtipos moleculares e assinaturas gênicas, gerando uma alta variedade de prognósticos e de respostas terapêuticas (CIANFROCCA; GRADISHAR, 2009).

Segundo (LU et al., 2009), a metástase no câncer de mama é considerada o fator responsável por causar a maior taxa de letalidade entre mulheres. Os elementos principais durante o processo de invasão e metástase são a degradação do tecido conectivo estromal e de componentes da matriz extracelular (HERSZèNYI et al., 1999) Diferentes tipos de enzimas proteolíticas (metalo-, aspartic-, cisteína-, serina- e treoninaproteinases) exercem papel fundamental para que esta degradação ocorra. A atuação conjunta das chamadas Metaloproteinases de Matriz (MMPs), responsáveis por clivar grande parte das proteínas presentes na Matriz Extracelular (ECM), das Proteínas Desintegrinas e Metaloproteínases (ADAMs) e da classe de proteínas capazes de inibir a ação das MMPs, as chamadas proteínas TIMPs, é responsável por exercer um importante papel proteolítico no ambiente tumoral (CRUZ-MUNOZ; KHOKHA, 2008).

Desta forma, é considerada de extrema importância a identificação de mecanismos moleculares e celulares regulatórios envolvidos no processo metastático para, posteriormente, auxiliar na geração de métodos prognósticos mais eficazes (LEREBOURS et al., 2003).

Com o passar dos anos, as tecnologias de Sequenciamento de Última Geração (Next Generation Sequencing - NGS) têm revolucionado a pesquisa biomédica, devido a sua rapidez e capacidade de analisar grandes quantidades de dados genômicos e transcriptômicos (KOBOLDT et al., 2013) (RIVERA; REN, 2013). A introdução destas novas tecnologias permitiu a análise de RNA através de sequenciamento em larga escala (RNA-seq), superando as limitações presentes nas tecnologias mais antigas

(BERRETTA; MORILLON, 2009) (KAPRANOV; WILLINGHAM; GINGERAS, 2009). Além disso, a introdução desta novidade no meio cientifico, contribuiu para o surgimento de projetos colaborativos em todo o mundo, como, por exemplo, o TCGA (The Cancer Genome Atlas), um banco de dados que fornece informações clínicas e dados de caracterização genômica de diversos tipos tumorais, gerados através de sequenciamento de alta cobertura (PUENTE et al., 2009).

Portanto, considera-se de grande interesse para o meio científico que sejam realizados mais estudos, utilizando esta ampla gama de informações disponíveis, a fim de melhor compreender a patogênese da doença e, assim, obter assinaturas gênicas suficientes para diagnosticar lesões pré-malignas de alto risco de progressão, gerando um impacto significativo nos processos de gestão e tratamento da saúde humana.

2 REVISÃO

2.1 Câncer de Mama – Aspectos fundamentais

Dentre os diversos tipos de câncer existentes, o carcinoma de mama se destaca por demonstrar taxas elevadas de incidência e mortalidade em todo o mundo (JEMAL et al., 2010). Este carcinoma é definido como uma doença heterogênea complexa, principalmente devido a diversidade de subtipos moleculares e assinaturas gênicas existentes, resultando em comportamentos clínicos diversificados. (CIANFROCCA; GRADISHAR, 2009). Esta heterogeneidade é considerada um obstáculo para a classificação dos tumores de mama, o que dificulta o desenvolvimento de métodos prognósticos eficazes (SIMPSON et al., 2005).

Ensaios moleculares que buscam traçar dados de perfis de expressão gênica têm exercido papel fundamental no processo de classificação e discriminação destes tumores (SOTIRIOU; PUSZTAI, 2009). Estes ensaios também têm auxiliado no processo de discriminação de subtipos existentes, uma vez que estes subtipos podem estar associados com diferenças em resultados clínicos, respostas à tratamentos, prognósticos e propagação metastática (PRAT; PEROU, 2011).

Aproximadamente 90% das mortes por câncer, incluindo câncer de mama, se devem a ocorrência de metástase (LU et al., 2009) De acordo com a literatura, as células cancerosas associadas com o processo metastático normalmente possuem alterações em sua forma e em sua capacidade de fixação às demais células e à matriz extracelular (BERX; ROY, 2009). A diminuição da expressão de E-caderina, molécula considerada chave no processo de adesão célula-célula, possui fundamental importância no desencadeamento destas alterações e, consequentemente, nos processos de invasão e metástase.

O processo metastático é descrito como uma sequência de passos descontínuos onde, através de mudanças biológicas na célula, inicia-se com uma invasão local, seguida pelo intravasamento das células cancerígenas em vasos linfáticos próximos, possibilitando que estas células transitem pelos sistemas hematógeno e linfático (TALMADGE; FIDLER, 2010). O evento seguinte, denominado extravasamento, consiste

na fuga destas células do lúmen dos vasos para o parênquima de tecidos distantes. A partir deste momento, inicia-se a formação de micrometástases, que são pequenos nódulos de células tumorais que podem crescer e se transformar em tumores macroscópicos. Esta última etapa denomina-se colonização.

Segundo (SøRLIE et al., 2001), há correlação entre o risco de ocorrência de metástase à distância e o local de preferência destas lesões, com o subtipo tumoral no câncer de mama. Sabe-se que os receptores de estrogênio (ER), (OSBORNE et al., 1980), de progesterona (PGR) e de fator de crescimento epidermal (HER2) são de grande importância no estudo do câncer de mama.

Através da técnica de Imunohistoquímica (IHC) tornou-se possível diferenciar e classificar amostras de pacientes em três subgrupos de acordo com estes receptores: grupo de positivos para ER e/ou PGR, grupo de positivos para HER2 e, por fim, grupo de receptores triplo negativos, ou seja, aqueles que são negativos para ER, PGR e HER2 (SøRLIE et al., 2001). Estudos comprovaram que tumores de mama classificados como positivos para ER possuem menor probabilidade de ocorrência de metástase à distância, formando geralmente metástases ósseas. Enquanto que os tumores classificados como positivos para HER2 e os considerados triplo negativos possuem maiores chances de progressão metastática, formando geralmente metástases cerebrais ou viscerais.

O surgimento da técnica de microarranjo tecidual (Tissue Microarray – TMA) permitiu avanços em estudos moleculares baseados em perfis de expressão gênica, através de microarranjo de DNA complementar (cDNA) microarray. Esta técnica teve importância fundamental para a evolução da pesquisa científica na área médica, sendo descrita pela primeira vez por (KONONEN et al., 1998).

Com o uso da técnica de TMA, (PEROU et al., 2000) propuseram uma nova subdivisão dos subtipos tumorais, baseando-se na correlação entre a classificação molecular e parâmetros clínicos relevantes, como o tempo de sobrevida e o tempo livre da doença. De acordo com estes parâmetros, foram definidos cinco subgrupos distintos: luminal A (RE positivo e HER2 negativo), luminal B (RE e HER2 positivos), superexpressão do HER2 (RE negativo e HER2 positivo), basal (RE e HER2 negativos) e normal breast-like. O subtipo luminal A associa-se à assinatura de melhor prognóstico, respondendo a terapias com medicamentos antiestrogênicos (SøRLIE et al., 2001); o subtipo luminal B associa-se ao pior prognóstico, principalmente por se associar a recidivas tumorais; o subtipo superexpressão de HER2 é visto como um excelente biomarcador de prognóstico e, como o próprio nome já diz, caracteriza-se por superexpressar o receptor de fator de crescimento epidérmico ou HER2 (BERTUCCI; BIRNBAUM; GONCALVES, 2006); o subtipo basal se encontra associado ao pior prognóstico e é associado a mutações genéticas no gene BRCA (PAREDES et al., 2006). Este, por não possuir alvo terapêutico definido como os demais tipos, faz com

que se torne dificultoso o desenvolvimento de drogas para o tratamento de pacientes classificados com este subtipo.

A existência de um grande número de fenótipos moleculares no câncer de mama é um dos fatores que dificulta sua classificação (PEROU et al., 2000). Sendo assim, vê-se a necessidade de uma classificação mais definida destes tumores para que se possa traçar uma relação estatística significativa com parâmetros clínicos relevantes. E, apesar de existir algumas variáveis prognósticas já disponíveis atualmente, tais como: o estado nodal, o tamanho tumoral, a idade e o estado dos receptores hormonais, é necessário que sejam identificados novos fatores prognósticos a fim de diminuir os fatores de risco e possibilitar a seleção da terapia mais adequada para cada paciente (EIRó et al., 2012).

2.2 Matriz Extracelular e Câncer

Os processos de invasão e metástase no câncer de mama são considerados como fatores responsáveis por ocasionar a maior taxa de letalidade entre pacientes, sendo, portanto, determinantes no quadro de evolução da doença (LU et al., 2009). Com isso, moléculas envolvidas nestes processos são vistas como potenciais candidatas para a identificação de novos marcadores prognósticos neste tipo de câncer (EIRó et al., 2012).

A ocorrência de angiogênese, invasão e metástase é dependente dos mecanismos de interação entre as células tumorais e o estroma, envolvendo moléculas de adesão, fatores de crescimento e de coagulação, proteinases, dentre outros (RUNDHAUG, 2005). O elemento considerado chave durante estes processos é a degradação dos componentes presentes na membrana basal e na matriz extracelular (ECM) sendo esta degradação dependente da ação de diferentes tipos de enzimas proteolíticas (as metalo-, aspartic-, cisteína-, serina- e treoninaproteinases) (COMOGLIO; TRUSOLINO, 2005).

De acordo com (BISSELL; LABARGE, 2005), é estabelecido que microambientes locais ou nichos exercem papel fundamental na regulação celular. E o que era um dos temas centrais apenas da embriologia básica, passa ser aceito também na biologia do câncer. O microambiente tumoral é composto por uma matriz extracelular insolúvel (ECM), um estroma composto por fibroblastos, adipócitos, células endoteliais e células do sistema imunológico, além de fatores de crescimento e citocinas (LU; WEAVER; WERB, 2012). Estudos, como o de (BHOWMICK; NEILSON; MOSES, 2004), têm sido dedicados para determinar como os componentes celulares do microambiente tumoral promovem o desenvolvimento do câncer. Grande parte destes evidencia a importância de componentes não celulares do nicho, especialmente a matriz extracelular, durante a progressão tumoral (LEVENTAL et al., 2009).

A matriz extracelular em si é composta por uma mistura complexa de componentes, incluindo proteínas, glicoproteínas, proteoglicanas e polissacarídeos, que exercem atividades físicas e bioquímicas distintas na célula (WHITTAKER et al., 2006). Estruturalmente, estes componentes formam também a membrana basal, que é produzida em conjunto por células epiteliais, endoteliais e estromais, para separar o epitélio ou endotélio do estroma, e pela matriz intersticial, que é produzida primeiramente pelas células estromais. A membrana basal é uma ECM especializada composta por colágeno do tipo IV, lamininas, fibronectina e ligantes capazes de interligar o colágeno aos demais componentes protéicos, além de ser mais compacta e menos porosa do que a matriz intersticial (EGEBLAD et al., 2010). Em contrapartida, a matriz intersticial é rica em colágenos fibrilares, proteoblicanas e diversas glicoproteínas, tais como tenascina C e fibronectina, que contribuem para dar resistência à força de tração dos tecidos.

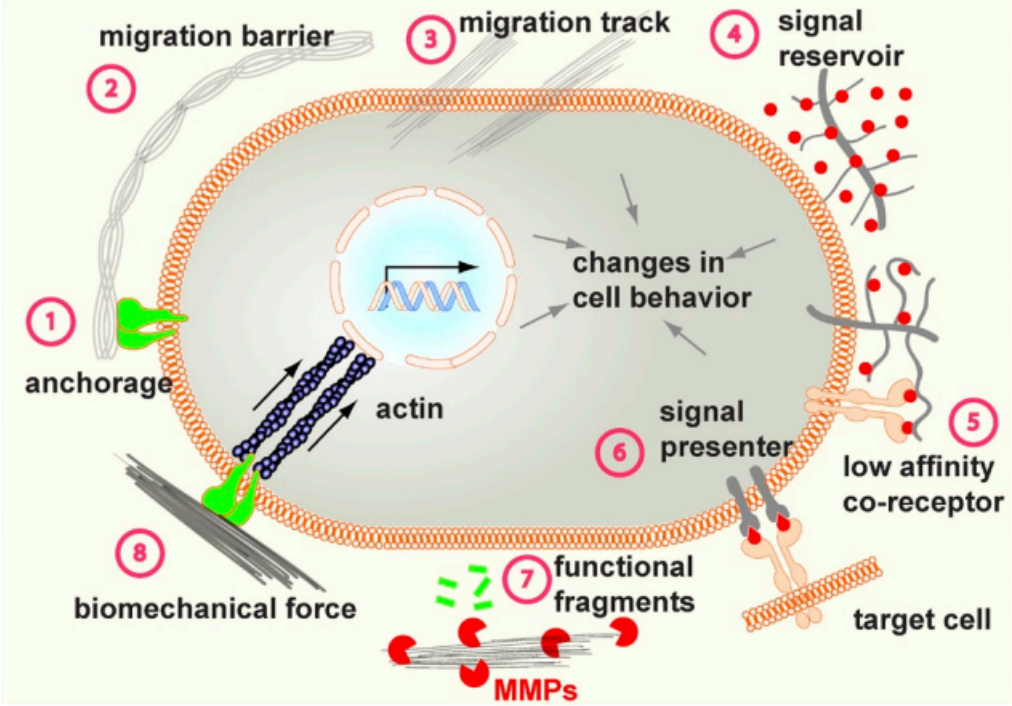

Figura 1: Mecanismos funcionais da ECM (LU; WEAVER; WERB, 2012)

Quando os componentes da ECM são colocados de forma ordenada, por possuírem uma considerável diversidade estrutural e bioquímica, estes conferem propriedades físicas, bioquímicas e biomecânicas únicas à matriz. Propriedades estas que atuam no comportamento celular e são fundamentais no processo de regulação. Por exemplo, as propriedades físicas da ECM são referentes à rigidez, porosidade, insolubilidade, disposição e orientação, que auxiliam na manutenção e suporte do tecido, bem como em sua integridade. Além disso, ao possuir função de barreira e sítio de ancoragem, as propriedades físicas da ECM desempenham tanto papéis positivos quanto negativos no processo de migração celular Figura 1, nos itens.

Dentre as propriedades bioquímicas da ECM estão suas capacidades diretas e

indiretas de sinalização, que permitem que as células detectem e interajam com o ambiente, resultando na expressão de genes ou em alterações no comportamento da célula (LU; WEAVER; WERB, 2012). É, também, desta forma que a ECM se torna capaz de aumentar a acessibilidade e a capacidade de sinalização direta de ligantes para seus receptores cognatos, como pode ser observado na Figura 1, nos itens 4 à 6. Com relação aos eventos de sinalização direta, segundo HYNES (2009), a ECM utiliza fatores de crescimento endógenos ou fragmentos funcionais derivados após serem processados por proteases, tais como as Metaloproteinases (MMPs), Figura 1 item 7.

O estudo das propriedades biomecânicas da ECM é ainda uma área em desenvolvimento e se baseia em questionar como estas propriedades como, por exemplo a elasticidade, podem contribuir para o desenvolvimento da doença Como resultado destes estudos constatou-se que as propriedades biomecânicas da ECM regulam diversos comportamentos celulares essenciais, incluindo a determinação das células de destino, além da diferenciação e função do tecido Figura 1 item 8 (ENGLER et al., 2006).

Devido a grande quantidade de papeis importantes exercidos pela ECM, existem diversos mecanismos reguladores para garantir que sua dinâmica, em termos de produção, degradação e remodelagem (PAGE-MCCAW; EWALD; WERB, 2007). Com isso, a ruptura desses mecanismos de controle tende a desregular e desorganizar a ECM, levando a comportamentos anormais das células que residem no nicho. A dinâmica anormal da ECM é considerada um dos desfechos clínicos mais evidentes no câncer (COX; ERLER, 2011).

Nos estudos de (PUPA et al., 2002) relatou-se alterações anormais nas propriedades bioquímicas e físicas destas moléculas durante o desenvolvimento do carcinoma de mama, contribuindo para a progressão tumoral e resistência à terapias. De acordo com (SHAPIRO; EYRE, 1982), a alteração mais bem reconhecida da ECM que ocorre no tecido tumoral é o aumento da deposição de colágeno. O colágeno é o componente mais abundante presente na ECM, constituindo 90% da ECM e 30% do total das proteínas presentes em humanos, sendo responsável por promover integridade estrutural e resistência à tração nos órgãos e tecidos humanos (REST; GARRONE, 1991). No contexto biológico do câncer, o colágeno regula propriedades físicas e bioquímicas do microambiente tumoral, atuando na modulação da polaridade celular, migração e sinalização (LEVENTAL et al., 2009).

Além disso, estudos envolvendo abordagens moleculares para correlacionar características clínicas a padrões de expressão gênica, têm destacado genes que codificam componentes da ECM associados ao tumor (RAMASWAMY et al., 2003). A partir destes estudos associou-se o aumento da expressão de genes que codificam proteínas com papel no remodelamento da ECM com a alta taxa de mortalidade em pacientes

com câncer de mama.

2.3 Genes MMPs, ADAMTs e TIMPs e metástase tumoral

Os elementos principais durante os processos de invasão tumoral e metástase no câncer são a degradação do tecido conectivo estromal e de componentes da matriz extracelular (HERSZèNYI et al., 1999). No entanto, alguns componentes, sobretudo os colágenos intersticiais, demonstram forte resistência à ação de proteolíticos, podendo ser degradados por tipos específicos de enzimas proteolíticas, como as metalo-, aspartic-, cisteína-, serenina- e treoninaproteinases.

As metaloproteinases de matriz (MMPs), sobretudo as MMP-2 e MMP-9 (RUNDHAUG, 2005), sintetizadas sob a forma de zimogênios, são coletivamente capazes de degradar as proteínas presentes na matriz extracelular (ECM) De acordo com (MURRAY, 2001), as MMPs foram classificadas em quatro subgrupos: colagenases, gelatinases, estromelisinas e de membrana. Esta classificação foi feita de acordo com o tipo de molécula que as MMPs são capazes de degradar. As MMP-2 e MMP-9, classificadas como colagenases ou gelatinases do tipo IV, degradam colágeno desnaturado (gelatina) e os tipos IV, V, VII, IX e X de colágeno (CHAMBERS; MATRISIANN, 1997), enquanto que as MMP-1, MMP-8 e MMP-13, classificadas como colagenases, degradam colágenos nativos do tipo I, II, III e IV (WESTERMARCK; KAHARI, 1999). As MMP-3, MMP-7, MMP-10 e MMP-11, classificadas como estromalisinas, catalisam o processo degradativo de diferentes substratos presentes na matriz extracelular, incluindo os proteoglicanos, lamininas, fibrinectinas, bem como as regiões não helicoidais do colágeno IV (CHAMBERS; MATRISIANN, 1997).Por fim, as MMP-T1 e MMP-T6, metaloproteinases classificadas como de membrana, também são capazes de degradar grande parte das moléculas presentes na matriz extracelular, tais como: colágenos I e III, lamininas, fibronectinas, vitronectinas e proteoglicanos (YANA; SEIKI, 2002).

Segundo STERNLICHT e WERB (2001), a expressão das MMPs é induzida por diferentes estímulos externos, tais como citocinas e fatores de crescimento, incluindo os interferons, interleucinas, fatores de crescimento endotelial vascular (VEGF), fatores de necrose tumoral alfa (TNF-) e beta (TNF-), fatores de crescimento epidermal (EGF) e indutores extracelulares de metaloproteinases de matriz (EMMPRIN). Em contrapartida, as atividades destas proteínas podem ser inibidas especialmente por outra classe de proteínas denominadas inibidoras teciduais das metaloproteinases (TIMPs).

Atualmente são conhecidos quatro diferentes tipos de TIMPs (TIMP-1, 2, 3 e 4), sendo estas proteínas responsáveis por executar a principal etapa de regulação sobre a atividade proteolítica das MMPs, após serem ativadas pela enzima latente, além de serem inibidores endógenos de desintegrinas da família das metaloproteinases

(ADAMTs) (CRUZ-MUNOZ; KHOKHA, 2008). Uma vez que a superexpressão de MMPs se encontra relacionada com a conversão de células normais para células tumorais malignas, a atividade inibidora dos genes TIMPs é considerada importante na inibição dos processos de progressão tumoral, invasão e metástase (BAKER et al., 1999; BERNHARD et al, 2004; COUSSENS et al., 2000; WANG et al., 1997).

As proteínas ADAMTS são compostas por 19 proteases secretadas, que possuem papeis de importância nos processos de montagem das matrizes de colágeno, na homeostase vascular e na clivagem das proteoglicanas da matriz extracelular (APTE, 2009; PORTER et al., 2005). Encontra-se bem estabelecido na literatura que estas proteínas estão relacionadas com a patogênese do câncer, exercendo influência nos processos de migração das células tumorais, sinalização celular e controle da angiogênese (PORTER et al., 2004). Porém, estudos de WAGSTAFF et al (2011), ao comparar a expressão dos 19 tipos de genes ADAMTs em tecidos normais e em tecidos tumorais de mama, demonstraram que alguns destes genes possuem papéis controversos por atuarem na supressão do tumor.

Um dos principais subgrupos dentro da família das ADAMTs são as proteoglicanases capazes de clivar moléculas estruturais da ECM, sobretudo as ADAMTS-1, que possuem atividades inibidoras de tumor e de promoção da angiogênese, além de participar de diversos processos biológicos de importância, tais como inflamação e desenvolvimento do sistema urogenital (OVERALL; LóPEZ-OTíN, 2008). Apesar de alguns papeis das ADAMTS-1 já terem sido elucidados, alguns estudos como o de ROCKS et al (2008) demonstram que os efeitos dessas moléculas durante a progressão tumoral permanecem controversos (HOJILLA; WOOD; KHOKHA, 2008). Este estudo demonstrou que as ADAMTS-1 contribuem para o desenvolvimento do tumor por estas atraírem fibroblastos e remodelarem a ECM.

Coletivamente, as MMPS são responsáveis pela clivagem da maior parte dos componentes proteicos da ECM e sua interação equilibrada com as proteínas TIMPs garante sua homeostase (OVERALL; LóPEZ-OTíN, 2008). Estas proteínas já foram relatadas na literatura como fatores de risco para o desenvolvimento do câncer de mama, progressão e sobrevivência tumoral (WIECZOREK et al., 2012). Em contrapartida, alguns estudos discordantes descrevem estas proteínas como fatores prognósticos ou até mesmo demonstram que não existe associação alguma com parâmetros clínicos no câncer de mama (HIRVONEN et al., 2003). A existência de dados tão controversos e de mecanismos ainda não elucidados faz com que o estudo dos perfis de expressão destas proteínas seja de interesse para a melhor compressão do comportamento tumoral e dos parâmetros clínicos associados à doença.

2.4 Tecnologias de Sequenciamento de RNA de Nova Geração (RNA-seq)

Um dos grandes desafios da biologia molecular é explicar como células com uma mesma composição genética são capazes de originar tipos celulares tão distintos, que desempenham papeis únicos e fundamentais para o funcionamento de um organismo multicelular (BRENTANI et al., 2005). Constatou-se, então, que essa diversidade fenotípica se deve a capacidade das células ou tecidos de expressarem diferentes conjuntos de genes (transcriptoma), refletindo em sua função, fenótipo e em suas respostas a estímulos ambientais (MOROZOVA; HIRST; MARRA, 2009). A partir deste momento, o estudo da correlação entre padrões de expressão gênica com os diferentes destinos e funções das células tornou-se de extremo interesse para o meio científico.

O desenvolvimento da tecnologia de microarray ou chip de DNA, através do método de hibridização, tornou possível a caracterização simultânea da expressão de milhares de transcritos de uma amostra (RUSSO et al., 2003). Este avanço científico incentivou o surgimento de diversos projetos a fim de caracterizar assinaturas gênicas de tipos celulares em diferentes estágios de determinadas patologias (MOCKLER et al., 2005). Apesar da eficácia desta técnica, ela também apresenta limitações consideráveis que podem diminuir a qualidade de uma análise. A necessidade de um conhecimento prévio do genoma ou de suas características pode ser considerada uma limitação desta técnica, pois afeta, por exemplo, a qualidade da análise em casos de genomas não completamente anotados (HURD; NELSON, 2009). Além disso, a variedade de formatos, metodologias e abordagens analíticas existentes para microarrays pode limitar a reprodutibilidade dos dados nesta técnica.

Com o passar dos anos e com o aumento massivo dos dados biológicos, passou-se a existir, também, a necessidade de novas tecnologias que permitissem sequenciar um maior número de bases em um curto espaço de tempo (MARDIS, 2013). O surgimento de sequenciadores de DNA automatizados demonstrou a importância da interação entre diversas áreas do conhecimento, tais como química, engenharia, bioinformática e biologia molecular, e de sua aplicação sobre métodos mais antigos de sequenciamento, como o Método de Sanger et al (1977).

Os sequenciadores que utilizam a tecnologia de nova geração (Next-Generation Sequencing – NGS), introduzida por (PASZEK et al., 2005) têm sido fundamentais para o avanço de estudos científicos, uma vez que permitem a geração de informações sobre milhões de pares de bases de forma paralela e em uma única corrida, promovendo um excelente custo-benefício (ASHLEY; SCHLUETER, 2012).

Esta capacidade de sequenciar paralelamente em massa é capaz de gerar reads a partir de bibliotecas fragmentadas: a partir de um genoma específico, no caso de

sequenciamento de genomas; a partir de um conjunto de fragmentos de cDNA, originados através da transcrição reversa de moléculas de RNA (RNAseq); ou a partir de um conjunto de moléculas amplificadas pela técnica de PCR. Apesar das tecnologias de NGS disponíveis atualmente utilizarem compostos químicos e ferramentas para detecção de bases diferentes, estas compartilham de dois passos fundamentais: o preparo da biblioteca e a detecção dos nucleotídeos incorporados (GLENN, 2011; ZHANG et a., 2011). Além disso, as novas tecnologias de sequenciamento podem ser classificadas em dois grupos principais: o grupo de tecnologias baseadas em PCR, que inclui as plataformas Roche 454 (Roche Diagnostics Corp., Branford, CT, USA), HiSeq 2000 (Illumina Inc., San Diego, CA, USA), AB SOLiD System (Life Technologies Corp., Carlsbad, CA, USA) e Ion Personal Genome Machine (Life Technologies, South San Francisco, CA, USA); e o grupo que não inclui o passo de amplificação antes do sequenciamento, sendo baseado na tecnologia de Sequenciamento de Molécula Única (Single Molecule Sequencing - SMS).

Os sequenciadores com a tecnologia Illumina foram introduzidos em 2007 e, devido a sua alta capacidade, passaram a ser muito utilizados em projetos de ressequencimeanto completo de genomas, incluindo o do genoma humano (SHOKRALLA et al., 2012). A plataforma Illumina, anteriormente conhecida como Solexa, possui uma abordagem de sequenciamento por síntese utilizando DNA polimerase e nucleotídeos terminadores marcados com fluoróforos (FEDURCO et al., 2006; TURCATTI et al., 2008). Para o sequenciamento com Illumina é utilizada a técnica de PCR de fase sólida, onde se realiza a clonagem in vitro dos fragmentos em uma plataforma sólida de vidro.

A introdução de tecnologias NGS, como a descrita anteriormente, revolucionou a transcriptômica, uma vez que estas se demonstram capazes de medir o nível de expressão de genes, eliminando dificuldades encontradas em tecnologias com abordagens mais antigas (METZKER, 2010). Um sequenciamento de RNA (RNA-Seq) típico consiste basicamente de quatro etapas: o isolamento do RNA; a conversão de RNA para DNA complementar (cDNA); o preparo da biblioteca de sequenciamento; e o sequenciamento propriamente dito, através de plataformas NGS (FINOTELLO; DI CAMILLO, 2014) (NAGALAKSHMI et al., 2008). O conjunto de RNAs transcritos sob uma determinada condição e tempo, podem revelar informações de mecanismos patológicos relevantes para determinadas doenças.

O estudo de perfis de expressão diferenciais de genes torna possível a comparação a partir de tecidos em diferentes condições como, por exemplo, a de tecidos saudáveis versus tecidos tumorais, visando identificar genes que desempenham papéis de importância na determinação de um fenótipo e que de alguma forma se encontram envolvidos na patologia. Por esse motivo, bancos de dados biológicos exercem papel fundamental em estudos bioinformáticos, uma vez que oferecem a oportunidade de

acesso a uma grande variedade de dados biológicos relevantes, incluindo sequências de genomas de diversas espécies, incluindo a humana. O projeto Atlas do The Cancer Genome Atlas (TCGA) faz parte de um esforço colaborativo em todo o mundo, sendo responsável por catalogar o panorama genômico de diversas linhagens tumorais geradas por sequenciamento de alta cobertura (PUENTE et al., 2009) (VERHAAK et al., 2010).

A bioinformática é uma ferramenta fundamental para o estudo do câncer, pois possibilita a exploração de biomarcadores relacionados ao câncer, tais como marcadores de predisposição genética, prognósticos, diagnósticos e terapêuticos (VERHAAK et al., 2010). A análise bioinformática supervisionada de dados transcriptômicos é aplicada para selecionar os genes mais diferencialmente expressos em pacientes com resultados discrepantes, sendo seu principal objetivo a identificação de assinaturas gênicas relevantes (MICHIELS; KOSCIELNY; HILL, 2005). Apesar das descobertas nesta área serem relativamente recentes, a bioinformática já tem sido vista como uma tecnologia promissora e como parte indispensável em estudos voltados para compreender as mudanças genéticas do câncer nesta era pós-genômica.

2.5 Análise de Expressão Gênica

A análise de perfis de expressão gênica, através de RNA-Seq, tem fornecido uma visão global do cenário da transcrição. Além do levantamento de níveis de expressão gênica, este tipo de análise pode ser aplicado na descoberta de novas estruturas genicas ou isoformas de splicing alternativo, dentre outras aplicações (KUKURBA; MONTGOMERY, 2015; PICKRELL et al., 2010).

O armazenamento de dados e análises de RNA-seq resultam em arquivos com formatos padronizados, que são cruciais tanto no processo de compartilhamento de informações entre laboratórios quanto no momento de reprodução de dados experimentais (TRAFNELL et al., 2012). O formato de arquivos mais comumente encontrados é o FASTQ, uma versão do arquivo FASTA estendida com a escala Phred de qualidade, indicando a qualidade das bases.

A medida da qualidade de um sequenciamento é fundamental, pois fornece informações importantes sobre a acurácia de cada passo do processo, incluindo preparação da biblioteca, identificação de bases e variantes e alinhamento de reads (NAKAGAWA et al., 2010). A identificação de bases, medida através da escala de qualidade Phred (Q score), é a medida mais comum para verificar a acurácia da plataforma de sequenciamento, indicando a probabilidade que uma determinada base foi identificada incorretamente pelo sequenciador. Por exemplo, se uma base possui um Phred de 30 (Q30), isso indica que a probabilidade desta base estar incorreta é 1 em 1000 vezes, indicando uma acurácia de 99,9% e assim por diante. São exem-

plos de ferramentas e softwares desenvolvidos para gerar dados de qualidade, a ferramenta FASTX (http://hannonlab.cshl.edu/fastx_toolkit), o software FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) e o pacote RobiNA (LOHSE et al., 2012).

Após a análise de qualidade das reads, a próxima etapa consiste no mapeamento destas reads contra um genoma de referência ou através do método *de novo* (ZHENG; MORTAZANA, 2012). Este passo no RNA-seq é considerado mais desafiador quando comparado ao DNA-seq, uma vez que muitas reads podem estar contidas em junções de splicing. Por esse motivo, não se recomenda o uso de algoritmos tradicionais de mapeamento, como o Bowtie (LANGMEAD et al., 2009) e BWA (LI; DURBIN, 2009), para mapeamento de reads contra um genoma de referência. Uma estratégia para minimizar este problema consiste em aderir, ao genoma de referência, sequências derivadas de junções de união exon-exon adquiridos de genes conhecidos (MORTAZAVI et al., 2008).

Com o aumento do uso de dados de RNA-Seq, aumentou também os esforços em desenvolver ferramentas de alinhamento especificas para mapear dados transcriptômicos, sendo capazes de reconhecer eventos de splicing. As ferramentas mais comumente utilizadas para este fim são MapSplice (WANG et al., 2010), GSNA (WU; NACU, 2010), STAR (DOBIN et al., 2013) e TopHat (TRAPNELL et al., 2009). A seleção destas ferramentas vai de acordo com os objetivos de estudo de RNA-Seq, uma vez que cada uma delas possui diferentes vantagens em desempenho, velocidade e memória. (ENGSTROM et al., 2013). O projeto denominado Avaliação da Anotação Genômica por RNA-seq (RGASP3), iniciado pelo GENCODE, teve como objetivo identificar as principais diferenças entre as ferramentas de alinhamento, utilizando como base diferentes parâmetros, tais como rendimento de alinhamento, precisão na identificação de bases, números de mismatches e capacidade de identificar sítios de splicing alternativo.

Após as reads serem alinhadas, realiza-se a montagem destas em transcritos. Grande parte dos programas computacionais, como Cufflinks (http://cufflinks.cbcb.umd.edu/) e RSEM (http://deweylab.biostat.wisc.edu/rsem), realiza este processo através do acúmulo de reads alinhadas contra um genoma de referência (LI et al., 2011; MEZLINI et al.2012; ROBERTS et al., 2011;TRAPNELL et al., 2010). Uma abordagem alternativa para a montagem de transcritos é a chamada montagem de novo, onde as contigs alinhadas podem ser comparadas a transcritos anotados para identificar novos transcritos (BIRZELE et al., 2010). Esta abordagem possui as vantagens de não necessitar de um conhecimento prévio das junções exon-exon e de poder ser usada em casos onde um genoma de referência não está disponível ou quando se encontra mal anotado. A plataforma Trinity (http://trinityrnaseq.sf.net) é um exemplo de software que utiliza a abordagem de novo.

Anteriormente ao processo de detecção de genes diferencialmente expressos, é importante considerar que se deve normalizar os dados, uma vez que a normalização é um passo fundamental para a analise de dados de RNA-seq (DILLIES et al., 2013; RISSO et al., 2011). Para se obter medidas de expressão gênica e comparar estas medidas entre grupos de *lanes* é necessário primeiramente normalizar as contagens de reads para ajustá-las quanto a variação da profundidade de sequenciamento destas *lanes* e quanto a outros potenciais efeitos técnicos característicos de cada plataforma de sequenciamento (BULLARD et al., 2010). Nos últimos anos, diversas estratégias de normalização vêm sendo propostas para corrigir as diferenças entre a distribuição de amostras presentes nas contagens de reads, como a contagem total ou profundidade de sequenciamento, e para corrigir os efeitos específicos de genes presentes nas amostras, como o tamanho do gene ou conteúdo GC (HANSEN et al., 2012).

A causa mais comum de variação entre *lanes* é a diferença no tamanho de bibliotecas, sendo assim a forma mais simples de normalização entre amostras consiste na medida da escala de contagem de reads em cada lane, através de um único fator especifico da lane, refletindo o tamanho da biblioteca. De acordo com a literatura, existem cinco métodos diferentes que calculam estes fatores escalares, sendo estes divididos em dois subgrupos. Um subgrupo consiste em métodos que normalizam a partir do conceito de tamanho de biblioteca, como TMM (ROBINSON; OSHLACK, 2010) e DESeq (ANDERS; HUBER, 2010) e o outro são métodos que normalizam considerando semelhanças na distribuição de contagem de reads, esteja esta distribuição em um único quantil (TC, UQ, Med e RPKM) ou em todos os quantis (Q).

Tanto TMM quanto DESeq utilizam como hipótese que a maioria dos genes não são Diferencialmente Expressos (DE) e propões um fator de análise baseado na proporção entre parâmetros como media ou mediana.

Além dos métodos propostos anteriormente, existem estratégias de normalização com foco em eliminar o conteúdo GC (FLISEK et a., 2012; LISTGARTEN et al., 2010), uma vez que é descrito na literatura a ocorrência de vieses relacionados com a eficácia no sequenciamento de regiões genômicas. Este fato demonstra que a contagem de reads não depende apenas do tamanho do gene, mas depende também de características da sequencia, sendo o conteúdo GC um exemplo (BENJAMINI; SPEED, 2011; BULLARD et al., 2010; ZHENG et al., 2011). Estes estudos relatam que tanto fragmentos muito ricos em conteúdo GC como aqueles muito pobres em conteúdo GC tendem a ser pouco representados em RNA-Seq, influenciando no processo de contagem de reads. (ANDERS;HUBER, 2010). Com isso, constatou-se que vieses relacionados ao comprimento e conteúdo GC possuem alto impacto nos resultados de expressão diferencial (DE) e, sendo assim, é importante que seja realizada uma normalização adequada para permitir a inferência exata das diferenças nos níveis de expressão presentes na amostra. O pacote EDASeq (RISSO et al., 2011), presente

no software Bioconductor, é um exemplo de ferramenta que realiza este tipo de normalização.

Com relação a RNA-seq, devido a questões biológicas, a abundância relativa de cada gene irá variar entre as amostras de RNA (PAN et al., 2008). Além disso, um problema muito difundido é que a contagem dos dados tipicamente mostra uma forte relação de média-variância, que não é respeitado por análises normais existentes, levando a inferências estatísticas ineficientes. Transformações como raiz-quadrada (HOEN et al., 2008) podem reduzir, mas não remover inteiramente os problemas envolvendo a relação de média-variância.

O modelo de Poisson não considera a variabilidade biológica ou qualquer fonte que faça com que a abundancia relativa de diferentes genes varie entre diferentes amostras de RNA. Quando a abundância não é constante entre amostras, as contagens de reads se apresentar totalmente dispersas em relação à de Poisson, ou seja, a variância deve ser maior do que a média. O modelos de super-dispersão binomial (BAGGERLY et al., 2004) ou Poisson (AUER; DOERGE, 2011) vem sendo proposto para Análises Seriais de Expressão Gênica(SAGE) ou para dados de RNA-Seq.

De acordo com ZHOU et al (2011), um método simples para compartilhar informações entre genes é partir do princípio de que todos os genes possuem a mesma relação de média-variância, em outras palavras, todos os genes possuem a mesma dispersão. Além disso, ROBINSON e SMYTH (2008) desenvolveram o modelo Bayesiano empírico, utilizando a probabilidade ponderada para estimar a variação biológica. Este modelo é implementado no pacote edgeR, presente no software Bioconductor. Outros métodos que utilizam a abordagem Bayesiana, como o pacote baySeq, são propostos para SAGE ou dados de RNA-Seq. Comparações entre edgeR e baySeq e outros métodos alternativos demonstraram que os primeiros são mais recomendados, uma vez que os demais não permitem a variabilidade específica do gene ou não compartilham informações entre genes (HARDCASTLE; KELLY, 2010).

Além disso, Modelos Lineares Generalizados (MLG) foram sugeridos para dados de contagem de SAGE ou RNA-Seq. Estes são conhecidos por necessitarem de montagem iterativa e, portanto, um problema comum de se utilizar esta abordagem é o tempo computacional requerido e a insuficiência algorítmica para alguns conjuntos de dados (MCCARTHY et al., 2012). Com base nisso, novas abordagens vêm sendo desenvolvidas com o objetivo de suprir estas necessidades e fazer análises mais acuradas. (KATOH et al., 2006; MICHIELS et al., 2005). A expectativa do aprimoramento de métodos que realizam análises diferenciais de expressão gênica em larga escala, é promover uma compreensão molecular mais significativa dos tumores e, consequentemente, fornecer meios para melhor classificar os pacientes e para o desenvolvimento de terapias personalizadas.

3 OBJETIVOS

3.1 Objetivo geral

Verificar, *in silico*, os padrões de expressão das famílias gênicas TIMPs, ADAMTs e MMPs em tumores metastáticos, a partir de amostras de tecidos de carcinoma invasivo de mama e de tecidos saudáveis adjacentes.

3.2 Objetivos Específicos

- Analisar as diferenças nos padrões de expressão entre os tecidos dos grupos saudáveis adjacentes e tumor primário correspondente;
- Investigar se os genes de interesse são capazes de diferenciar amostras normais das tumorais;
- Verificar se há correlação de expressão entre os genes das famílias MMPs, TIMPs e ADAMTs.

4 MATERIAIS E MÉTODOS

4.1 Amostras e Dados Clínicos

A partir dos dados de expressão gênica (mRNA) de 1094 pacientes com Carcinoma Invasivo de Mama, adquiridos através do Banco de Dados TCGA (The Cancer Genome Atlas), as amostras foram selecionadas quanto a plataforma de sequenciamento Ilumina Hiseq2000; e quanto ao algoritmo de análise RNASeqV2, utilizado tanto para o mapeamento quanto para a detecção dos transcritos.Após, uma nova seleção com amostras tumorais e normais pareadas correspondentes utilizadas nas análises supervisionadas foi realizada através de seus barcodes "TN" (Tumor, matched Normal) "NT" (Normal, matched Tumor), respectivamente. Além disso, também foram considerados para a seleção os dados clínicos: gênero (sexo feminino) e presença de metástase linfonodal e/ou à distância, resultando em um total de 64 amostras normais e 64 amostras tumorais (Tabela 1). O download das amostras foi realizado em julho/2015.

4.2 Análise RNA-SeqV2

RNASeq Versão 2 é uma metodologia de análise definida pelo TCGA (The Cancer Genome Atlas), um projeto de colaboração mundial, onde diferentes tipos tumorais são coletados e caracterizados, através de diferentes plataformas gênicas (CHANG, 2011). Esta metodologia tem por objetivo utilizar dados de sequenciamento para determinar níveis de expressão gênica. Para isso, primeiramente, são gerados arquivos no formato FASTQ, contendo as reads sequenciadas a partir de Illumina HiSeq 2000. A etapa posterior consiste na análise da qualidade e trimagem destas reads através do estabelecimento de um cut-off de Q20, na escala Phred.

A próxima etapa consiste no mapeamento ou alinhamento das reads a um genoma de referência, neste caso o genoma humano, através do programa MapSPlice (WANG et al., 2010). Esta etapa gera arquivos nos formatos SAM/BAM, contendo a posição genômica de cada gene.

O passo seguinte consiste na anotação gênica e quantificação dos transcritos, atra-

Tabela 1: Dados clínicos das amostras selecionadas para o estudo de expressão diferencial.

Barcode	Gênero	Marg.	Tam.	Metast. (Linf.)	Metast. (Dist.)	Estágio	Tipo Histológ.
TCGA-A7-A13E	Feminio	Neg	> 2cm	1 a 3	Indefinido	IIB	Ductal
TCGA-A7-A13F	Feminio	Neg	> 5cm	1 a 3	Ausência	IIA	Ductal
TCGA-AC-A2FF	Feminio	Neg	> 2cm	1 a 3	Indefinido	IIB	Lobular
TCGA-AC-A2FM	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Lobular
TCGA-BH-A0AZ	Feminio	Neg	> 5cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A0B3	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A0B5	Feminio	Pos	> 2cm	1 a 3	Ausência	IIA	Lobular
TCGA-BH-A0B7	Feminio	Neg	> 2cm	4 a 9	Ausência	IIB	Ductal
TCGA-BH-A0BA	Feminio	Neg	> 5cm	1 a 3	Ausência	IIC	Indefinido
TCGA-BH-A0BC	Feminio	Neg	> 2cm	> 10	Ausência	IIC	Ductal
TCGA-BH-A0BJ	Feminio	Neg	> 2cm	> 10	Ausência	IIB	Ductal
TCGA-BH-A0BM	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A0BS	Feminio	Neg	> 5cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A0BT	Feminio	Neg	> 1cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A0BV	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A0BZ	Feminio	Neg	> 5cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A0C0	Feminio	Neg	> 1cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A0DH	Feminio	Pos	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A0DQ	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A0DT	Feminio	Neg	> 1cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A0DV	Feminio	Neg	> 2cm	4 a 9	Ausência	IIA	Ductal
TCGA-BH-A0DZ	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A0E0	Feminio	Neg	> 5cm	> 10	Ausência	IIC	Ductal
TCGA-BH-A0E1	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A18U	Feminio	Neg	> 2cm	4 a 9	Ausência	IIIA	Ductal
TCGA-BH-A18V	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A1EO	Feminio	N.A.	> 1cm	1 a 3	Ausência	IIA	Indefinido
TCGA-BH-A1EV	Feminio	N.A.	> 5cm	1 a 3	Ausência	IIIA	Ductal
TCGA-BH-A1EW	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A1F0	Feminio	N.A.	>0.1cm	1 a 3	Ausência	IIA	Ductal
TCGA-BH-A1F2	Feminio	N.A.	Indef.	4 a 9	Ausência	IIIB	Ductal
TCGA-BH-A1F6	Feminio	N.A.	Indef.	4 a 9	Indefinido	Indefinido	Ductal
TCGA-BH-A1F8	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIIB	Ductal
TCGA-BH-A1FB	Feminio	N.A.	> 1cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A1FC	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIA	Medular
TCGA-BH-A1FE	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A1FH	Feminio	N.A.	> 2cm	4 a 9	Presença	IV	Ductal
TCGA-BH-A1FJ	Feminio	N.A.	> 5cm	1 a 3	Ausência	IIIA	Ductal
TCGA-BH-A1FM	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIIA	Ductal
TCGA-BH-A1FR	Feminio	N.A.	Indef.	1 a 3	Ausência	IIIB	Indefinido
TCGA-BH-A203	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A204	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-BH-A208	Feminio	N.A.	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-E2-A153	Feminio	Neg	> 1cm	1 a 3	Ausência	IIB	Ductal
TCGA-E2-A158	Feminio	Neg	> 1cm	1 a 3	Ausência	IIA	Ductal
TCGA-E2-A15K	Feminio	Neg	> 2cm	1 a 3	Ausência	IIA	Ductal
TCGA-E2-A1IG	Feminio	Neg	> 2cm	4 a 9	Ausência	IIB	Ductal
TCGA-E2-A1L7	Feminio	Neg	> 2cm	1 a 3	Ausência	IIIA	Ductal
TCGA-E2-A1L8	Feminio	Neg	> 2cm	4 a 9	Ausência	IIB	Lobular
TCGA-E9-A1N4	Feminio	Neg	> 2cm	1 a 3	Ausência	IIIA	Indefinido
TCGA-E9-A1N5	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Indefinido
TCGA-E9-A1N6	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-E9-A1ND	Feminio	Neg	> 2cm	1 a 3	Ausência	IIB	Ductal
TCGA-E9-A1RC	Feminio	Neg	Indef.	>10	Ausência	IIIC	Indefinido
TCGA-E9-A1RF	Feminio	Neg	> 2cm	4 a 9	Ausência	IIIA	Ductal
TCGA-E9-A1RI	Feminio	Neg	> 1cm	4 a 9	Ausência	IIIA	Indefinido

vés do pacote RSEM (versão 0.4.6), gerando arquivos no formato GTF. Este é o modo como o TCGA diponibiliza os dados para que análises posteriores possam ser realizadas.

4.3 Normalização dos Dados e Análise de Expressão Diferencial

O processo de normalização de dados foi realizado pelo pacote EDAseq (versão 2.5.0), que normaliza as reads levando em consideração o conteúdo GC, a partir de dados de contagens brutas (raw counts). A partir dos dados de quantificação resultantes realizou-se análises supervisionadas de expressão diferencial utilizando a linguagem estatística de programação R (versão 3.2.2), além do pacote TCGAbiolinks (versão 1.0.1), presente no software Bioconductor, que integra outros pacotes, como edgeR (versão 3.2) e EDAseq (versão 2.5.0).

Posteriormente, foi realizada uma análise global de expressão diferencial entre os genes presentes no tecido tumoral de mama e no tecido normal adjacente. Para o teste de expressão diferencial entre os dois grupos foi utilizado o teste de exactTest, ajustado pelo método de Benjamini-Hochberg. Através da comparação entre amostras de tecido normais e tumorais, foram considerados genes diferencialmente expressos aqueles que apresentaram um valor inferior ao cut-off estabelecido de 0.01 e fold-change menor/igual ou maior/igual a -1 e 1, respectivamente. Para a visualização desta análise foi gerado o gráfico de Volcano plot.

Posteriormente, realizou-se a clusterização hierárquica entre as amostras em conjunto com a análise de correlação entre os genes, considerando a Correlação de Pearson entre as amostras, no eixo x, e a distância euclidiana, no eixo y. A correlação de Pearson foi gerada a partir do pacote R. Foram considerados genes correlacionados aqueles que possuem valor de R >= 0.6 ou <=-0.6 e p-value <= 0.01.

4.4 Análise Funcional

O processo de enriquecimento de vias (Kegg) e ontologias (GO) foi realizado com base nos resultados diferenciais de expressão dos genes de interesse, através da função TCGAanalyze_EAcomplete presente no pacote TCGAbiolinks (versão 1.0.1).

A análise funcional foi realizada com base no Sistema de Classificação PANTHER(Protein Analysis Through Evolutionary Relationships), que possui conexão com o banco de dados Gene Ontology (GO). Esta análise resultou em Barplot's, baseados nos valores de p-valeu ajustado por log10FDR e de log2 fold-change (FC), ambos gerados com base no teste estatístico de Mann-Whitney, contido no pacote edgeR (versão 3.2).

5 RESULTADOS E DISCUSSÃO

Atualmente, muitos estudos relacionados ao estágio do tumor, grau histológico, Receptores de Estrogênio (ER), Receptores de Progesterona (PR) e ao Fator de Crescimento Epitelial Humano 2 (HER2), mutações (BRCA1 e BRCA2), dentre outros, têm sido considerados fatores importantes no processo de prognóstico do câncer de mama (JEZIERSKA; MOTYL, 2009; PEPPERCORN et al., 2008). No entanto, recentes descobertas envolvendo técnicas moleculares têm sido vistas como um novo caminho promissor para o desenvolvimento de novas classificações para este tipo de tumor, auxiliando na caracterização do prognóstico, através da descoberta de assinaturas gênicas.

Com isso, o foco principal do presente estudo foi verificar, através de perfis transcriptômicos, se as famílias gênicas MMPs, TIMPs e/ou ADAMTSs desempenham papel biológico relevante no Carcinoma Invasivo de Mama. Para responder a estas questões, o passo fundamental foi identificar se as famílias apresentam diferenças significativas em seus níveis de expressão, quando comparados aos tecidos normais adjacentes.

5.1 Análise global de genes diferencialmente expressos

5.1.1 Análise de distribuição

Para um estudo supervisionado de expressão gênica diferencial, é interessante que seja realizada, previamente, uma análise incluindo todos os genes considerados diferencialmente expressos envolvidos no câncer de mama, possibilitando visualizar os padrões de distribuição destes genes nos dois tecidos (normais e tumorais). Esta análise, realizada pelo pacote TCGAbiolinks, pôde ser visualizada através do Volcano plot (Figura 2).

Este gráfico demonstra que há uma divisão clara entre os genes presentes nas amostras normais e tumorais, indicando que estes possuem fortes características que os tornaram capazes de se agrupar entre si. Todos os genes que estão à direita (com Fold Change maior que 2), os genes que se encontram mais expressos no tecido

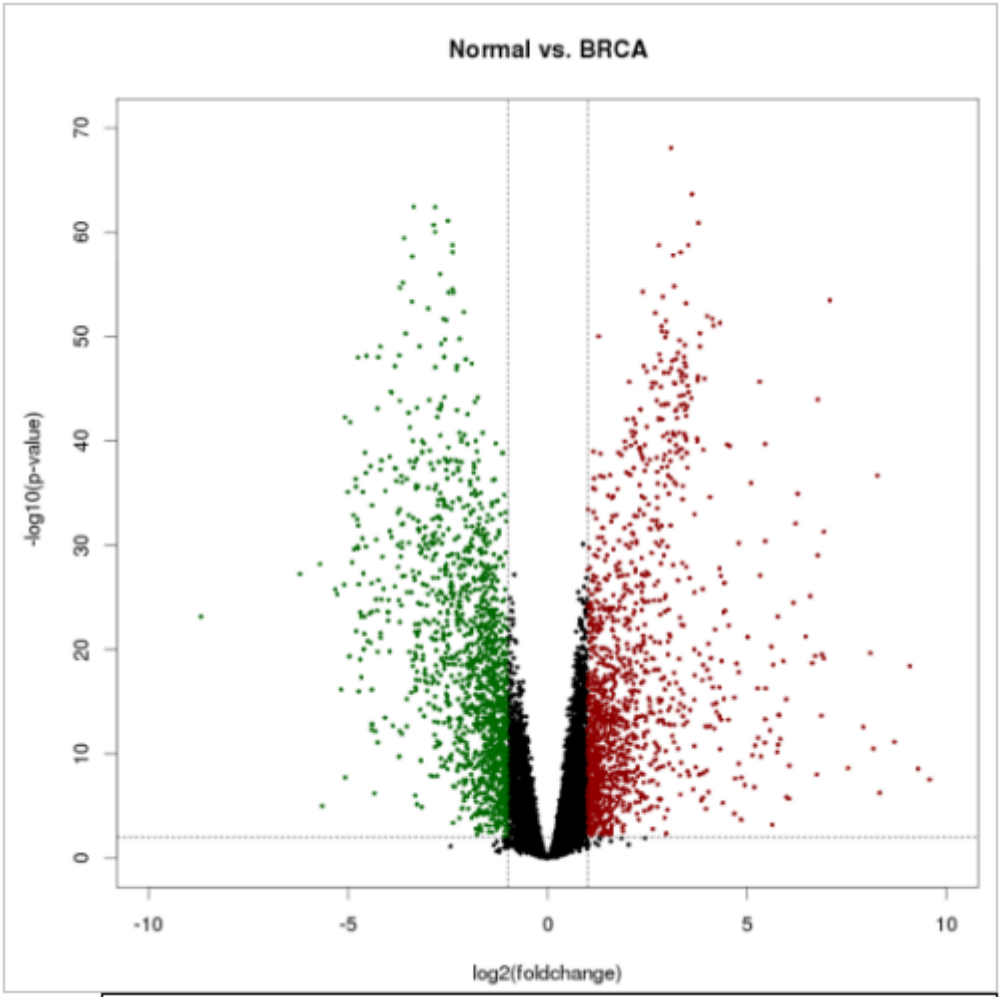

Figura 2: Análise global da diferença de expressão gênica entre tecidos tumorais de mama e normais adjacentes com base nos valores de p (-log10 (0.01)) e de Fold Change (log2(2)).

tumoral e, aqueles que estão à esquerda (com Fold Change menor que 2), os genes que se encontram mais expressos no tecido normal. O p-value extremamente alto, representado por -log10(0.01), demonstra que os resultados diferenciais de expressão são significativos.

5.1.2 Análise Funcional

Posteriormente, ainda de forma global, foi realizada a análise funcional, com base no sistema de classificação PANTHER (Protein Analysis Through Evolutionary Relationships), que possui conexão com o banco de dados de vias (Kegg) e de funções e processos biológicos (Gene Ontology – GO). Esta análise resultou em um gráfico Barplot (Figura 3), baseados nos valores de p igual a 0.01 ajustados em log10 por FDR, e nos valores de Log2 de Fold Change, para Fold Change igual a 2.

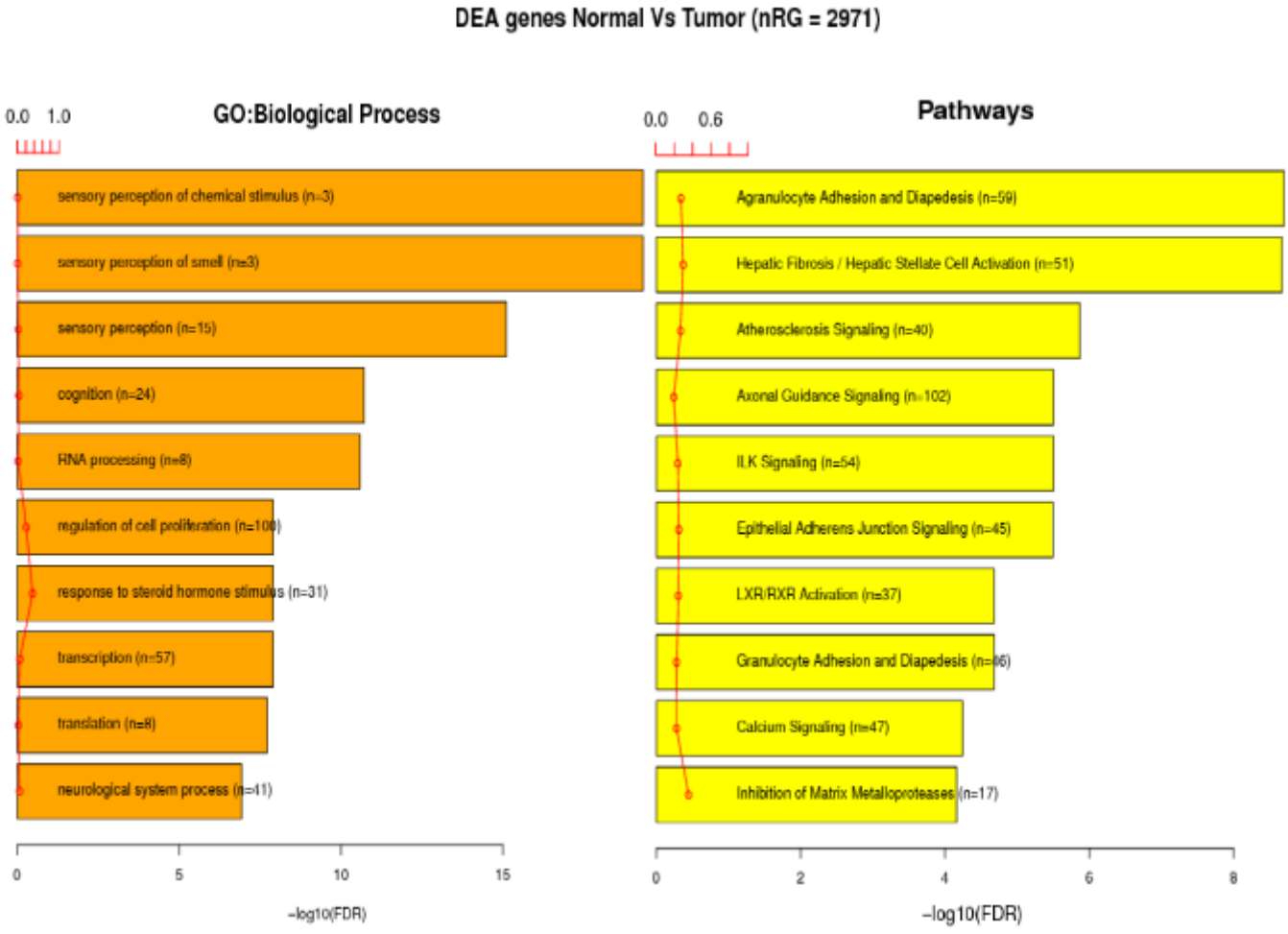

Figura 3: Análise funcional de processos biológicos (Gene Ontology) e vias (Kegg), baseada na diferença de expressão gênica entre tecidos tumorais de mama e normais adjacentes. Ambos gerados com base no teste estatístico de Mann-Whitney, contido

Através deste gráfico, é possível observar que, apesar de não ter havido seleção das famílias gênicas de interesse (MMPs, TIMPs e ADAMTSs) para a análise

funcional, grande parte dos genes considerados diferencialmente expressos quando comparados tecidos tumorais de mama e tecidos normais apresentaram funções biológicas e vias, descritos na literatura, como sendo característicos destas famílias. A regulação da proliferação celular (n=100), transcrição (n=57), resposta a estimulos de hormonios esteroidais (n=31), a aderência epitelial à junções sinalizadoras (n=45), sinalização de cálcio (n=47) e inibição de metaloproteinases de matriz (n=17) são processos biológicos e vias relacionados às famílias gênicas de interesse, sugerindo que estas exercem papel fundamental no processo de invasão e progressão tumoral no carcinoma invasivo de mama.

5.2 Análise diferencial de expressão das famílias gênicas (MMPs, TIMPs e ADAMTSs) no câncer de mama

A análise de expressão diferencial das famílias gênicas de interesse (MMPs, TIMPs e ADAMTs) comparando tecidos tumorais de mama e normais adjacentes, através do pacote edgeR, identificou 22 genes diferencialmente expressos[1] (Tabela 2).

5.2.1 Análise de Clusterização Hierárquica

Utilizando os 24 genes diferencialmente expressos realizou-se a análise de clusterização hierárquica para verificar se os genes de interesse são capazes de agrupar as amostras e diferenciá-las. A visualização desta análise foi possível através do gráfico HeatMap gerado (Figura 4), utilizando a função aheatmap, presente no pacote NMF, do software Bioconductor.

Através desta análise verificou-se que todas as metaloproteinases diferencialmente expressas, com exceção da MMP-28, se encontram mais expressas em tecidos tumorais quando comparados a tecidos normais adjacentes. As MMPs estão envolvidas, em sua maioria, nos processos de degradação da matriz extracelular, migração celular, proliferação, apoptose e remodelação de tecidos, em numerosos estados patológicos e biológicos, sobretudo no câncer (MORRISON et al., 2009; RAY et al., 2009). As MMP-1, classificadas como colagenases, são relatadas na literatura como uma das metaloproteinases mais expressas em tecidos de carcinoma invasivo de mama, além de estar presente em outros tipos de câncer (BALDUYC et al., 2000; MIGITA et al., 1999) e o aumento de sua expressão, juntamente com as MMP-9, -13 e -14 está associado a diminuição da taxa de sobrevida de pacientes com este tipo de câncer. Além disso, a degradação de ECM por MMP-1 também demonstrou influência nas interações célula-célula e célula-ECM através de desassociação, levando ao aumento da divisão celular e diminuição da apoptose e à tumorigênese (HOJILLA et al., 2003).

[1] Os Valores gerados com base no teste estatístico de Mann-Whitney, para False Discovery Rate (FDR) igual a - log10(p-value), para p-value < 0.01 e Fold Change(FC) igual a log2(FC), para FC=2.

Tabela 2: Parâmetros de expressão diferencial dos genes MMPs, TIMPs e ADAMTSs quando comparadas amostras de tecidos tumoral e normal adjacente.

mRNA	Log2FC	FDR	Tumor	Normal
MMP11	6.362986	1.843337e-200	26446.6579	312.30702
MMP9	3.323533	5.826557e-31	13182.9386	949.81579
MMP14	1.250872	1.830563e-29	10291.51754	32191.0972
MMP13	6.611346	8.732888e-67	2633.3772	23.45614
MMP1	6.721741	2.692464e-67	2161.2281	17.50000
MMP3	2.229512	1.172367e-21	873.8158	167.77193
MMP10	3.332908	2.091292e-22	372.2193	32.67544
MMP12	2.032505	8.332584e-11	298.5877	60.68421
MMP28	-2.114608	8.594177e-44	222.2632	909.24561
ADAMTS1	-2.191659	3.287921e-51	3194.28070	13472.90351
ADAMTS5	-3.018220	3.470234e-122	889.02632	6613.00000
ADAMTSL4	-1.884194	5.574871e-36	669.13158	2300.42105
ADAMTS9	-1.307638	4.595578e-28	947.29825	2190.96842
ADAMTS14	3.118975	8.894581e-52	381.04386	39.26316
ADAMTSL2	1.782829	1.481585e-37	492.76316	139.09649
ADAMTS7	1.059599	1.374851e-16	515.67544	237.18421
ADAMTS6	1.355943	1.550592e-26	217.16667	79.06140
ADAMTS19	2.398059	5.981589e-16	118.24561	19.89474
ADAMTS18	-1.729423	2.014138e-14	123.05263	409.36842
ADAMTS3	-1.556254	5.165225e-30	122.33333	330.03509
ADAMTS8	-1.692810	4.85402e-13	91.73684	155.2931
TIMP4	-3.381693	3.684634e-19	315.746	3822.667

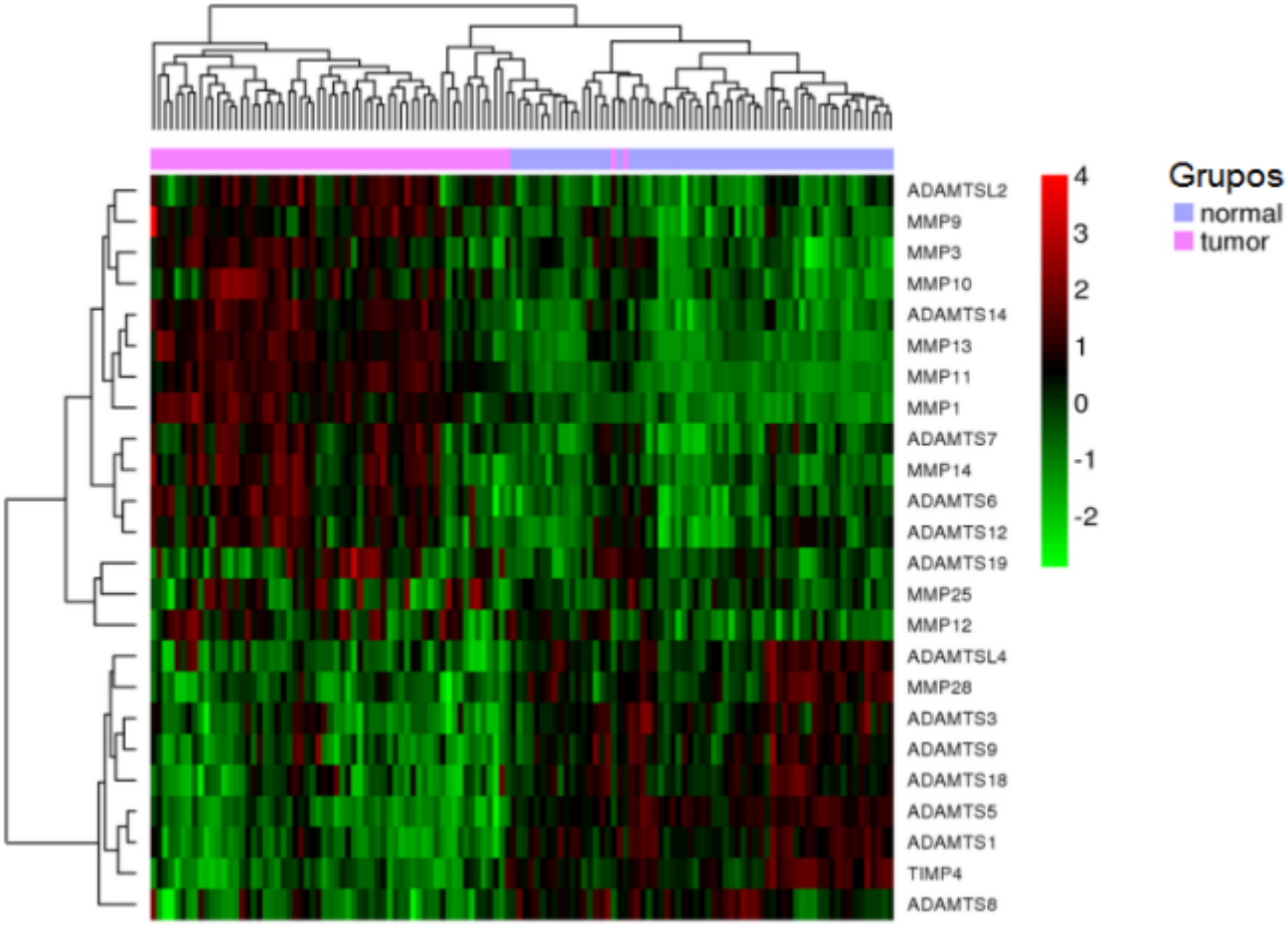

Figura 4: Heatmap com a representação da clusterização entre amostras e correlação entre genes diferencialmente expressos no câncer de mama (n=128).

Em estudos in vivo, através da técnica de ELISA (Enzyme-Linked Immunosorbent Assay), foi observada uma alta expressão das proteínas MMP-1 e MMP-9 em tecidos tumorais quando comparados a tecidos normais (Przybylowska et al., 2006).

As MMP-2 e MMP-9 têm sido amplamente estudadas como biomarcadoras e alvos terapêuticos no câncer de mama (EGEBLAD;WERB, 2002; STERNLICHT; WERB, 2001). Isso principalmente por estas gelatinases se relacionarem com a invasão tumoral e metástase, devido a sua capacidade de remodelamento tecidual via matriz extracelular e, também, por induzir a degradação da membrana basal e a angiogênese (SUN et al., 2014). Apesar de relatos de alta expressão por parte destas metaloproteinases, em tecidos tumorais de mama, seus papeis como fatores prognósticos ainda permanecem pouco elucidados.

As MMP-11, classificadas como estromalisinas, são conhecidas por desempenhar um papel de importância na tumorigênese da mama (ENGEL et al., 1994). Ao contrário de todas as outras MMPs que são secretadas sob a forma de proenzimas, as MMP-11 são ativadas antes mesmo de serem secretadas, através de proteases do tipo furina associadas ao Complexo de Golgi. De acordo com estudos de Köhrmann et al (2009) e Kossakowska et al (1996), observou-se um aumento significativo de expressão de MMP-11 em tecidos tumorais de mama em comparação a tecidos normais. Além disso, a superexpressão de estromalisinas se encontra significantemente relacionada a parâmetros clinicopatológicos, o que pode ser fundamental para prever a evolução da doença.

As MMP-13, colagenases, demonstram estar bastante expressas em tecidos tumorais de mama e ainda se encontram associadas ao aumento da taxa de metástase à distância (VIZOSO et al., 2007). Além disso, de acordo com estudos de ZHANG et al (2008), as MMP-13 possuem correlação com fenótipos tumorais agressivos, estando inversamente correlacionadas com a taxa de sobrevivência de pacientes com este tipo de câncer. Estas podem ser vistas como potenciais fatores prognósticos para o câncer de mama, mas seus mecanismos precisam ser mais estudados.

No presente estudo, o fato desta família gênica ter demonstrado uma expressão significativa em tecidos tumorais em relação aos normais indica concordância com o que há na literatura. As MMPs-1, -9, -10, -11 e -13 foram as que se demonstraram mais expressas no tecido tumoral, apresentando um valor alto de log2 Fold Change, sendo seus respectivos valores de q-value (p-value ajustado em log10 por FDR), são inferiores ao cut-off estabelecido de 0.01, ressaltando que estas diferenças de expressão são estatisticamente significativas. Apesar das MMP-28 apresentarem, na literatura, uma alta expressão em tecidos tumorais de mama, neste estudo ela demonstrou resultados controversos, encontrando-se mais expressas em tecidos normais quando comparadas a tecidos normais. Este resultado indica que o grande número de amostras é necessário para a compreensão do papel dos genes. Além disso, mais estudos

devem ser realizados para investigar outros possíveis papeis que esta metaloproteinase pode estar exercendo como, por exemplo, de supressão tumoral.

Os genes TIMPs são genes conhecidos por possuirem atividade proteolítica inibitória e, sendo assim, diversos estudos relatam a importância desta família gênica na inibição da progressão tumoral e, consequentemente, dos processos de invasão e metástase (ALBINI et al., 1991; BAKER et al., 1999; MATSUSAWA et al., 1996). A superexpressão do gene TIMP-4 já foi relatada em linhagens celulares de câncer de mama e associada com a inibição dos processos de invasão, metástase e crescimento tumoral (WANG et al., 1997). Além disso, foram relatados os mesmos efeitos de inibição quando estes foram implantados em camundongos. Porém, também existem alguns estudos com resultados conflitantes, como o relato de que a terapia com este gene, em camundongos, promoveu a formação tumoral e a associação deste gene com a transição de carcinoma dutal in situ para carcinoma invasivo, em humanos (JIANG et al., 2001; ZHAO et al., 2004).

O resultado deste estudo indicou que, dentre os quatro tipos de genes TIMPs existentes, o TIMP-4 apresentou uma diferença significativa de expressão para estas amostras. Nesta análise, o Log2 de fold-change se apresentou negativo, indicando que há maior expressão deste gene nos tecido normal quando comparado ao tecido tumoral. Além disso, o p-value ajustado por FDR indica a confiabilidade desta análise, uma vez que o cut-off estabelecido foi de 0.01. Apesar do resultado encontrado contradizer a literatura com relação a sua atividade tumoral inibitória, é possível perceber pelos genes diferencialmente expressos que a degradação da matrix celular está muito mais presente nas amostras tumorais. Assim, os resultados compõem o esperado.

A arquitetura das proteínas ADAMTS, com domínios que conferem atividades proteolíticas, além da habilidade de ligação a diversas moléculas associadas a matriz celular e extracelular (ECM), sugerem sua participação relevante na progressão tumoral e metástase (PORTER et al., 2005). No entanto, na presente análise, verificou-se que maior parte dos ADAMTS (ADAMTS-1, ADAMTS-5, ADAMTS-L4, ADAMTS-9, ADAMTS-18, ADAMTS-3 e ADAMTS-8) demonstrou maior expressão no tecido normal que no tecido tumoral. Uma possível explicação para estes resultados consiste na presença do domínio metaloproteinase ativo nestas proteínas. Este domínio, por exercer papel fundamental na degradação de componentes da matriz extracelular, pode estar degradando moléculas de fatores de crescimento e citoquinas (MARETZKI et al., 2005; OVERALL;KLEIFELD, 2006), contribuindo assim para o controle da proliferação, migração celular e angiogênese. Além disso, a presença dos domínios C-terminais descritos em algumas destas proteínas, como ADAMTS-1 (ENGLE et al., 2001), ADAMTS-4 (GAO et al., 2002), ADAMTS-8 (VAZQUES et al., 1999), ADAMTS-9 (SOMERVILLE et al., 2003) e ADAMTS-12 (CAL et al., 2002), pode ser responsável por ações anti-tumorais e anti-metastáticas (ZHENG et al., 2003).

As proteínas ADAMTS começaram a ganhar atenção muito recentemente, sendo assim, existem informações muito limitadas na literatura a seu respeito, principalmente com relação a seus possíveis papeis no câncer de mama. Por este motivo, considera-se fundamental que sejam realizados mais estudos com o objetivo de melhor definir suas funções biológicas e investigar seus possíveis papeis na progressão tumoral.

A análise de clusterização hierárquica demonstrou que os 22 genes considerados diferencialmente expressos, pertencentes às famílias gênicas (MMPs, TIMPs e ADAMTSs), foram capazes de clusterizar as amostras normais e tumorais, separando-as. Além disso, a análise através da Correlação de Pearson demonstrou coerência no agrupamento das amostras de acordo com as características dos genes. Este resultado indica que o grande número de amostras utilizadas neste estudo foi relevante para descobrirmos novos genes possivelmente envolvidos com a progressão tumoral e, a partir dai, investigarmos se estes são capazes de atuarem como marcadores moleculares, auxiliando em métodos prognósticos ou colaborando para terapias específicas e personalizadas.

6 CONCLUSÕES

Através deste estudo, ao comparar amostras tumorais de mama e normais adjacentes do mesmo paciente, foi possível demonstrar que existem diferenças significativas de expressão entre os genes. Isso foi visto tanto na análise global quanto na análise restrita das famílias de interesse no estudo. Além disso, as caracteristicas biológicas presentes nestas famílias foram ressaltadas durante a análise diferencial de expressão global, ressaltando a importância destes genes em processos relacionados ao carcinoma invasivo de mama.

Foi possível identificar que os genes MMPs, TIMPs e ADAMTSs foram capazes de diferenciar as amostras normais das tumorais. Porém, a análise de correlação entre eles indica que mais estudos serão necessários tanto para elucidar os papeis exercidos em cada família gênica, uma vez que diversos estudos demonstram resultados controversos com relação a atuação destes genes no câncer de mama, quanto para investigar novos papeis.

REFERÊNCIAS

ASHLEY, N. E.; SCHLUETER, J. **Applications of next-generation sequencing in plant biology**. Howell Science Complex N303a, Mailstop 551, Greenville, North Carolina 27858 USA: East Carolina University, Department of Biology, 2012.

BERRETTA, J.; MORILLON, A. **Pervasive transcription constitutes a new level of eukaryotic genome regulation**. 91198 Gif-sur-Yvette, Paris 6, France: Centre de Génétique Moléculaire-Centre National de la Recherche Scientifique, Université of Pierre et Marie Curie, 2009.

BERTUCCI, F.; BIRNBAUM, D.; GONCALVES, A. **Proteomics of breast cancer: principles and potential clinical applications.** Marseille, France: Centre de Recherche en Cancérologie de Marseille, Département d'Oncologie Moléculaire, Institut Paoli-Calmettes, 2006.

BERX, G.; ROY, F. van. **Involvement of Members of the Cadherin Superfamily in Cancer**. [S.l.]: Cold Spring Harb Perspect Biol, 2009.

BHOWMICK, N. A.; NEILSON, E. G.; MOSES, H. L. **Review article Stromal fibroblasts in cancer initiation and progression**. [S.l.]: Nature Publishing Group, 2004.

BISSELL, M.; LABARGE, M. **Context, tissue plasticity, and cancer: are tumor stem cells also regulated by the microenvironment?** USA: Department Cancer Biology, Life Sciences Division, Lawrence Berkeley National Laboratory, University of California, 2005.

BRENTANI, R.; CARRARO, D.; VERJOVSKI-ALMEIDA, S.; REIS, E.; NEVES, E.; SOUZA, S. de. **Gene expression arrays in cancer research: methods and applications.** São Paulo, Brasil: Ludwig Institute for Cancer Research, 2005.

CHAMBERS, A.; MATRISIANN, L. **Changing views of the role of matrix metalloproteinases in metastasis.** Canada: Department of Oncology, University of Western Ontario, 1997.

CIANFROCCA, M.; GRADISHAR, W. New Molecular Classifications of Breast Cancer. **CA: A Cancer Journal for Clinicians**, [S.l.], v.59, n.5, p.303–313, 2009.

COMOGLIO, P. M.; TRUSOLINO, L. **Cancer: the matrix is now in control**. 10060 Candiolo (Torino), Italy: Division of Molecular Oncology, IRCC, Institute for Cancer Research and Treatment, University of Torino School of Medicine, 2005.

COX, T. R.; ERLER, J. T. Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. **Disease Models and Mechanisms**, [S.l.], v.4, n.2, p.165–178, 2011.

CRUZ-MUNOZ, W.; KHOKHA, R. The role of tissue inhibitors of metalloproteinases in tumorigenesis and metastasis. **Crit Rev Clin Lab Sci**, [S.l.], v.45, p.291–338, 2008.

EIRó, N.; GONZáLEZ, L.; GONZáLEZ, L. O.; FERNANDEZ-GARCIA, B.; LAMELAS, M. L.; MARíN, L.; GONZáLEZ-REYES, S.; CASAR, J. M. d.; VIZOSO, F. J. Relationship between the Inflammatory Molecular Profile of Breast Carcinomas and Distant Metastasis Development. **PLoS ONE**, [S.l.], v.7, n.11, p.e49047, 11 2012.

ENGLER, A. J.; SEN, S.; SWEENEY, H. L.; DISCHER, D. E. Matrix Elasticity Directs Stem Cell Lineage Specification. **Cell**, [S.l.], v.126, n.4, p.677 – 689, 2006.

FINOTELLO, F.; DI CAMILLO, B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. **Briefings in Functional Genomics**, [S.l.], 2014.

HERSZèNYI, L.; PLEBANI, M.; CARRARO, P.; DE PAOLI, M.; ROVERONI, G.; CARDIN, R.; TULASSAY, Z.; NACCARATO, R.; FARINATI, F. **The role of cysteine and serine proteases in colorectal carcinoma.** Budapest, Hungary.: Second Department of Medicine, Semmelweis University Medical School, 1999.

HIRVONEN, R.; TALVENSAARI-MATTILA, A.; PääKKö, P.; TURPEENNIEMI-HUJANEN, T. **Matrix metalloproteinase-2 (MMP-2) in T(1-2)N0 breast carcinoma.** Finland: Department of Oncology and Radiotherapy, University Hospital of Oulu, 2003.

HOJILLA, C.; WOOD, G.; KHOKHA, R. **Inflammation and breast cancer: metalloproteinases as common effectors of inflammation and extracellular matrix breakdown in breast cancer.** Toronto, M5G 2M9 Canada: Department of Medical Biophysics, Ontario Cancer Institute, 2008.

HURD, P. J.; NELSON, C. J. **Advantages of next-generation sequencing versus the microarray in epigenetic research**. Canada: BC Cancer Agency, Genome Sciences Center, 2009.

INCA. **Estimativa 2014. Incidência do Câncer no Brasil.** Rio de Janeiro: INSTITUTO NACIONAL DE CÂNCER, 2014.

JEMAL, A.; SIEGEL, R.; XU, J.; WARD, E. Cancer Statistics, 2010. **CA: A Cancer Journal for Clinicians**, [S.l.], v.60, n.5, p.277–300, 2010.

KAPRANOV, P.; WILLINGHAM, A.; GINGERAS, T. **Genome-wide transcription and the implications for genomic organization.** 3420 Central Expressway, Santa Clara, California 95051, USA.: Affymetrix, Inc., 2009.

KOBOLDT, D. C.; MELTZ STEINBERG, K.; LARSON, D. E.; WILSON, R. K.; MARDIS, E. R. **The Next-Generation Sequencing Revolution and Its Impact on Genomics**. St. Louis, MO 63108, USA: INSTITUTO NACIONAL DE CÂNCER, 2013.

KONONEN, J.; BUBENDORF, L.; KALLIONIEMI, A.; BäRLUND, M.; SCHRAML, P.; LEIGHTON, S.; KALLIONIEMI, O. **Tissue microarrays for high-throughput molecular profiling of tumor specimens.** USA: Laboratory of Cancer Genetics, National Human Genome Research Institute, National Institutes of Health, 1998.

LEREBOURS, F.; BERTHEAU, P.; BIECHE, I.; PLASSA, L.; CHAMPEME, M.; HACENE, K.; TOULAS, C.; ESPIE, M.; MARTY, M. **Two prognostic groups of inflammatory breast cancer have distinct genotypes.** France: Institut National de la Santé et de la Recherche Médicale (INSERM) E0017/Oncogénétique, 2003.

LEVENTAL, K. R.; YU, H.; KASS, L.; LAKINS, J. N.; EGEBLAD, M.; ERLER, J. T.; FONG, S. F.; CSISZAR, K.; GIACCIA, A.; WENINGER, W.; YAMAUCHI, M.; GASSER, D. L.; WEAVER, V. M. **Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling**.

LU, J.; STEEG, P.; PRICE, J.; KRISHNAMURTHY, S.; MANI, S.; REUBEN, J.; CRISTOFANILLI, M.; DONTU, G.; BIDAUT, L.; VALERO, V.; HORTOBAGYI, G.; YU, D. **Breast cancer metastasis: challenges and opportunities.** M. D. Anderson Cancer Center, Houston, Texas 77030, USA: Department of Molecular and Cellular Oncology, University of Texas, 2009.

LU, P.; WEAVER, V.; WERB, Z. **The extracellular matrix: a dynamic niche in cancer progression.** Manchester M20 4BX, England, UK.: Breakthrough Breast Cancer Research Unit, University of Manchester, 2012.

METZKER, M. **Sequencing technologies - the next generation.** Houston, Texas 77030, USA.: Human Genome Sequencing Center and Department of Molecular Human Genetics, Baylor College of Medicine, 2010.

MICHIELS, S.; KOSCIELNY, S.; HILL, C. **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** Villejuif, France.: Biostatistics and Epidemiology Unit, Institut Gustave Roussy, 2005.

MOCKLER, T.; CHAN, S.; SUNDARESAN, A.; CHEN, H.; JACOBSEN, S.; ECKER, J. **Applications of DNA tiling arrays for whole-genome analysis.** USA: Plant Biology Laboratory, The Salk Institute for Biological Studies, 2005.

MOROZOVA, O.; HIRST, M.; MARRA, M. A. **Applications of New Sequencing Technologies for Transcriptome Analysis**. Canada: BC Cancer Agency, Genome Sciences Center, 2009.

MURRAY, G. I. **Matrix metalloproteinases: a multifunctional group of molecules**. [S.l.]: The Journal of Pathology, 2001.

NAGALAKSHMI, U.; WANG, Z.; WAERN, K.; SHOU, C.; RAHA, D.; GERSTEIN, M.; SNYDER, M. **The transcriptional landscape of the yeast genome defined by RNA sequencing.** New Haven, CT 06520, USA.: Department of Molecular, Cellular, and Developmental Biology, Yale University, 2008.

OSBORNE, C.; YOCHMOWITZ, M.; KNIGHT WA, r.; MCGUIRE, W. **The value of estrogen and progesterone receptors in the treatment of breast cancer.**

OVERALL, C.; LóPEZ-OTíN, C. **Strategies for MMP inhibition in cancer: innovations for the post-trial era.** Canada: Department of Oral Biological and Medical Sciences, C.I.H.R. Group in Matrix Dynamics, University of British Columbia, 2008.

PAGE-MCCAW, A.; EWALD, A. J.; WERB, Z. **Matrix metalloproteinases and the regulation of tissue remodelling**. Finland: Nature Publishing Group, 2007.

PAREDES, J.; ALBERGARIA, A.; CARVALHO, S.; SCHMITT, F. C. **Basal-like Breast Carcinomas: Identification by P-cadherin, P63and EGFR Basal Cytokeratins Expression**.

PASZEK, M.; ZAHIR, N.; JOHNSON, K.; LAKINS, J.; ROZENBERG, G.; GEFEN, A.; REINHART-KING, C.; MARGULIES, S. **Tensional homeostasis and the malignant phenotype.** Philadelphia, 19104, USA.: Department of Bioengineering, University of Pennsylvania, 2005.

PEROU, C.; SøRLIE, T.; EISEN, M.; RIJN M van de; JEFFREY, S.; REES, C.; POLLACK, J. **Molecular portraits of human breast tumours.** California 94305, USA.: Department of Genetics, Stanford University School of Medicine, 2000.

PRAT, A.; PEROU, C. **Deconstructing the molecular portraits of breast cancer.** Chapel Hill, NC 27599, USA: Lineberger Comprehensive Cancer Center, University of North Carolina, 2011.

PUENTE, X.; PINYOL, M.; QUESADA, V.; CONDE, L.; ORDóñEZ, G.; VILLAMOR, N. **Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia.** 3420 Central Expressway, Santa Clara, California 95051, USA.: Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, 2009.

PUPA, S.; MéNARD, S.; FORTI, S.; TAGLIABUE, E. **New insights into the role of extracellular matrix during tumor onset and progression.** Milan, Italy: Molecular Targeting Unit, Department of Experimental Oncology, Istituto Nazionale Tumori, 2002.

RAMASWAMY, S.; ROSS, K.; LANDER, E.; GOLUB, T. **A molecular signature of metastasis in primary solid tumors.** USA: Whitehead Institute/MIT Center for Genome Research, 2003.

REST, M. van der; GARRONE, R. **Collagen family of proteins.** Villeurbanne, France: Institute of Biology and Chemistry of Proteins, 1991.

RIVERA, C.; REN, B. **Mapping Human Epigenomes**. 9500 Gilman Drive, La Jolla, CA 92093-0653, USA: Ludwig Institute for Cancer Research, Institute of Genomic Medicine, UCSD Moores Cancer Center, University of California School of Medicine, 2013.

RUNDHAUG, J. **Matrix metalloproteinases and angiogenesis.** Department of Carcinogenesis, Science Park–Research Division, The University of Texas M. D. Anderson Cancer Center: Division of Molecular Oncology, IRCC, Institute for Cancer Research and Treatment, University of Torino School of Medicine, 2005.

SHAPIRO, F.; EYRE, D. Collagen polymorphism in extracellular matrix of human osteosarcoma. , [S.l.], 1982.

SIMPSON, P. T.; REIS-FILHO, J. S.; GALE, T.; LAKHANI, S. R. Molecular evolution of breast cancer. **The Journal of Pathology**, [S.l.], v.205, n.2, p.248–254, 2005.

SOTIRIOU, C.; PUSZTAI, L. **Gene-expression signatures in breast cancer.** Brussels, Belgium: Medical Oncology Department, Translational Research Unit, Jules Bordet Institute, Université Libre de Bruxelles, 2009.

SøRLIE, T.; PEROU, C.; TIBSHIRANI, R.; AAS, T.; GEISLER, S.; JOHNSEN, H.; HASTIE, T.; EISEN, M.; RIJN, M. van de. **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** Oslo, Norway: Department of Genetics, The Norwegian Radium Hospital, 2001.

TALMADGE, J.; FIDLER, I. **AACR centennial series: the biology of cancer metastasis: historical perspective.** Omaha, Nebraska, USA: University of Nebraska Medical Center, Transplantation Immunology Laboratory, 2010.

VERHAAK, R.; HOADLEY, K.; PURDOM, E.; WANG, V.; WILKERSON, M. **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** Cambridge, MA 02142, USA.: The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, 2010.

WHITTAKER, C. A.; BERGERON, K.-F.; WHITTLE, J.; BRANDHORST, B. P.; BURKE, R. D.; HYNES, R. O. **The echinoderm adhesome**. USA: Center for Cancer Research, Massachusetts Institute of Technology, 2006.

WIECZOREK, E.; RESZKA, E.; GROMADZINSKA, J.; WASOWICZ, W. **Genetic polymorphism of matrix metalloproteinases in breast cancer.** Lodz, Poland: Department of Toxicology and Carcinogenesis, Nofer Institute of Occupational Medicine, 2012.

7 GLOSSÁRIO

lanes

São regiões da "*flow cell*"aonde são realizadas as reações de sequenciamento, o que inclui a formação dos *clusters* e o "sequenciamento por síntese"(*sequencing by synthesis* - SBS) propriamente dito. Cada "*flow cell*"pode contêm várias lanes, sendo oito o número mais comum.

Flow cell

Placa aonde são carregadas as bibliotecas de sequenciamento. *Flow cells* são usadas por diferentes plataformas se sequenciamento de nova-geração, como ABI SOLiD, Illumina Solexa e Roche 454. Diferentes do Chips usados na plataforma Ion Torrents. Normalmente é utilizada como superfície de armazenamento para alocar os fragmentos a serem sequenciados.

read

Sequencia de DNA identificada a partir de um fragmento após o processo de "*base-calling*"("chamada de bases"). Pode ser "*raw*"(bruta), quando não passou por nenhum processamento, ou "*filtered*"(filtrada), quando foi modificada (ex: trimada, mascarada).

Tecnologia de microarray

Também chamada de chip de DNA é uma ferramenta de análise de expressão gênica que permite investigar a expressão de centenas ou milhares de genes em uma amostra com uma reação de hibridização.A tecnologia é baseada na hibridação de alvos marcados derivados de amostras biológicas e uma série de sondas de DNA imobilizadas em uma matriz sólida, que representam os genes de interesse.

Cobertura

Também denominada "*coverage*"ou "*depth*"é uma medida de quantas vezes uma base foi sequenciada, sendo equivalente ao números de *reads* que cobrem aquela região. Essa informação é obtida durante a montagem, seja esta "de novo"ou "por referência"(re-sequenciamento).

Análise, *in silico*, dos padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs e seus possíveis papéis no câncer de mama. – Monize Provisor

UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Curso de Biotecnologia

Trabalho de Conclusão de Curso

Análise, *in silico*, dos padrões de expressão das famílias gênicas TIMPs, ADAMTSs e MMPs e seus possíveis papéis no câncer de mama.

MONIZE PROVISOR

Pelotas, 2016