UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

**Genetic profile analysis of tumor stem cells in locally advanced
breast cancer**

Willian Abraham da Silveira

Ribeirão Preto
2015

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

**Genetic profile analysis of tumor stem cells in locally advanced
breast cancer**

Tese de Doutorado apresentado ao
Programa de Pós-Graduação em
Ginecologia e Obstetrícia para obtenção
do Título de Doutor em Ciências Médicas

Área de Concentração: Ginecologia e
Obstetrícia.

Orientado: Willian Abraham da Silveira

Orientador: Daniel Guimarães Tiezzi

Ribeirão Preto
2015

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Silveira, Willian Abraham

Análise do perfil genético de células tronco tumorais no câncer de mama localmente avançado, 2015.

99 p. : il. ; 30cm.

Tese de Doutorado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP – Área de concentração: Ginecologia e Obstetrícia.

Orientador: Tiezzi, Daniel Guimarães

1.Cancêr de Mama.2. Célula-Tronco.3. Transcriptoma . Ribeirão Preto.

FOLHA DE APROVAÇÃO

Nome do Aluno: Willian Abraham da Silveira

Título do trabalho: Análise do perfil genético de células tronco tumorais no câncer de mama localmente avançado

Tese de Doutorado apresentado ao
Programa de Pós-Graduação em
Ginecologia e Obstetrícia para obtenção
do Título de Doutor em Ciências Médicas

Área de Concentração: Ginecologia e
Obstetrícia.

Orientado: Willian Abraham da Silveira

Orientador: Daniel Guimarães Tiezzi

Aprovado em:

Banca Examinadora

Prof.Dr. _____

Instituição: _____ Assinatura: _____

Prof.Dr. _____

Instituição: _____ Assinatura: _____

Prof.Dr. _____

Instituição: _____ Assinatura: _____

Prof.Dr. _____

Instituição: _____ Assinatura: _____

DEDICATION / DEDICATÓRIA

I dedicate this work to the patients who took part in this study, especially to one, the look of hope from her eyes I will not forget. And to my grandfather, José Francisco da Silveira, whose disease angered the child I was and whose life example inspire the man I am.

Dedico este trabalho as pacientes que participaram deste estudo, principalmente a uma, cujo olhar de esperança não esquecerei. E a meu avó paterno, José Francisco da Silveira, cuja doença causou revolta na criança que fui e cujo exemplo de vida inspira o homem que sou.

**ACKNOWLEDGEMENTS /
AGRADECIMENTOS**

This work is the result of the support of many, without which it would never have been written. Firstly I want to thank my mother, Gilda Ap. de Moraes Silveira, and my brother, Wayne Ambrósio da Silveira Jr, for having always supported me in my dreams and for believing in me when times became difficult. I also want to thank my father, Wayne Ambrósio da Silveira, who taught me too early about the brevity and fragility of life, its value and the necessity to make our best with the time we have.

I thank Professor Daniel Guimarães Tiezzi for the chance he gave me and for the confidence he deposited in me at the time we barely knew each other. His sense of practicability, his sharp intelligence and the way he takes care of the group, personally and professionally, aroused my admiration. I thank the members of the group for the support and for the good moments, Renata Danielle Sicchieri, Larissa Raquel Mouro Mandarano, Tatiane Mendes Gonçalves de Oliveira, Heriton Marcelo Ribeiro Antonio, Fernanda Marques Rey, Fernanda Carvalho and Angélica Pires da Costa. I also thank the people who shared the laboratory with us, Fermino Netto, Paulo Novais, Luana Lourenço, Renata Collares, Andressa Romualdo, Patricia Fadel and Vagner Schiavoni. I want to thank the resident doctors of the “Hospital das Clínicas” at the time, principally to Bruno André, Raphael Bettero, Paola Rodrigues Menani and Isabela Panzeri Carlotti. I want to say thanks to Professor Jurandyr Moreira de Andrade and Hélio Humberto Angotti Carrara, both always gave myself and the group help and support and the latter adopted me as his student in the end when Prof. Daniel needed to go abroad.

I thank to Patricia Vianna Bonini Palma from the Laboratory of Flow Cytometry of the Hemocenter of Ribeirão Preto, for our long talks while the experiments ran, for the advice and support, without whose help this project would have died at the beginning.

My Thanks to Professor Sílvia Regina Rogatto from UNESP of Botucatu and from the “Hospital A.C Camargo” of São Paulo, and to her group, principally to Rolando André Rios Villacis. Their kindness in agreeing to help us was essential for the completion of this work.

I thank the people from the Institute of Cancer Research of Montpellier, in France. To Dr. Charles Theillet for the opportunity, Dr. Stanislas Du Manoir for

all the teachings, for the friendship and for the laughs. To Alejandra Damian, Amanda Abi Khalil, Augusto Faria, Béatrice Orsetti, Benoit Beganton, Berfin Seyran, Coralie Lefreve, Florence Cammas, Hanine Oubari, Hélène Delpech, Joelle Azzi, Laurence Lasorsa, Marianne Le Gall, Meryem Brital, Mohammad Hamyeh, Mona Houhou, Pauline Mayonove, Patrick Augereau, Rahila Rahimova, Rana Melhem, Rui Bras-Gonçalves, Sara Cherradi, Shefqat Hajdari, Thibauld Houles and Toufic (“The Wise”) Kassouf, for all the friendship and good moments. I arrived in France alone, I did not leave that way.

I thank FAPESP for the financial support and the Department of Gynecology and Obstetrics of the Ribeirão Preto Medical School and its staff, principally Suellen Soares, Gabriela Sica and Reinaldo.

I want to thank my family and all the people who passed through my life in the period. No one works alone, no one lives alone, no one learns alone, and I am happy to have too many people to name.

My final thanks go to the masters I have found on the way. Among them, Dr. Marcelo Dias Baruffi, who taught me a lot about the scientific world and what the meaning of the word “scientist” is; Dr. Antonio Caliri, who made me realize the great difference between what we can understand of reality and what reality really is; and to Dr. Richard John Ward, the first to open a door for me, and continues to help me whenever the necessity arises. The final thanks go to “Mrs. Sônia” my first grade teacher. She gave the best answer an adult can give to a curious child, the answer which made me look for other answers, the one which marked the start of the journey that led me here: I don’t know.

Este trabalho é o resultado do apoio de muitos, sem os quais ele nunca teria sido escrito. Primeiramente gostaria de agradecer a minha mãe e a meu irmão, Gilda Ap. de Moraes Silveira e Wayne Ambrósio da Silveira Jr, por sempre me apoiarem em meus sonhos e sempre acreditarem em mim quando a situação se tornava difícil. Gostaria também de agradecer ao meu pai, Wayne Ambrósio da Silveira, que me ensinou muito cedo sobre a brevidade e a fragilidade da vida, seu valor e a necessidade de fazer o melhor que pudermos com o tempo que temos.

Agradeço ao Prof. Dr. Daniel Guimarães Tiezzi, pela chance e pela confiança que depositou em mim em um momento em que mal nos conhecíamos. Seu senso prático, sua inteligência afiada e seu cuidado com o grupo, tanto pessoal quanto profissionalmente, despertaram minha admiração. Agradeço pelo apoio e pelos momentos de descontração aos membros do grupo, Renata Danielle Sicchieri, Larissa Raquel Mouro Mandarano, Tatiane Mendes Gonçalves de Oliveira, Heriton Marcelo Ribeiro Antonio, Fernanda Marques Rey, Fernanda Carvalho e Angélica Pires da Costa. e aos amigos com quem dividimos o laboratório, Fermino Netto, Paulo Novais, Luana Lourenço, Renata Collares, Andressa Romualdo, Patricia Fadel e Vagner Schiavoni e também aos residentes do Dept. de Ginecologia e Obstetrícia que trabalham enquanto o projeto era feito, to Bruno André, Raphael Bettero, Paola Rodrigues Menani e Isabela Panzeri Carlotti. Quero agradecer ao Professor Jurandyr Moreira de Andrade e Hélio Humberto Angotti Carrara, ambos sempre apoiaram nosso grupo e o último me adotou no final, quando o Prof. Daniel teve que viajar para fora do país.

Agradeço a Patricia Vianna Bonini Palma do Laboratório de Citometria de Fluxo do Hemocentro de Ribeirão Preto por seu apoio, por nossas longas conversas enquanto o experimento corria e por seus conselhos. Sem sua ajuda este trabalho teria sido encerrado logo em seu início.

Agradeço a Prof.Dra. Sílvia Regina Rogatto da UNESP de Botucatu e do Hospital A.C Camargo de São Paulo e a seu grupo, principalmente a Rolando André Rios Villacis. Sua gentileza em aceitar ser nossa colaboradora foi essencial para que efetuássemos este trabalho.

Agradeço as pessoas do Instituto de Pesquisa do Câncer de Montpellier, na França. Ao Dr. Charles Theillet pela oportunidade, ao Dr. Stanislas Dumanoir por todo ensinamento, pela amizade e pelas risadas. A Alejandra Damian, Amanda Abi Khalil, Augusto Faria, Béatrice Orsetti, Berfin Seyran, Coralie Lefreve, Florence Cammas, Hanine Oubari, Hélène Delpech, Joelle Azzi, Laurence Lasorsa, Marianne Le Gall, Meryem Brital, Mohammad Hamyeh, Mona Houhou, Patrick Augereau, Pauline Mayonove, Rahila Rahimova, Rana Melhem, Rui Bras-Gonçalves, Sara Cherradi, Shefqat Hajdari, Thibauld Houles and Toufic (“O Sábio”) Kassouf, pela amizade e pelos bons momentos. Cheguei a França solitário, mas não a deixei assim.

Agradeço a FAPESP pelo apoio financeiro e ao Departamento de Ginecologia e Obstetrícia da Faculdade de Medicina de Ribeirão Preto e ao seu Serviço de Pós-Graduação, principalmente Suelen Soares, Gabriela Sica and Reinaldo.

Agradeço a toda minha família e a todas as pessoas que passaram por minha vida neste período, ninguém trabalha só, ninguém vive só, ninguém aprende só e posso me considerar feliz por ter pessoas demais para nomear.

Meus agradecimentos finais vão aos mestres que encontrei pelo caminho. Entre eles, o Prof. Dr. Marcelo Dias Baruffi, que me ensinou muito do mundo científico e o significado de ser um cientista, o Prof. Antonio Caliri, que me fez realizar a grande diferença entre o que podemos entender da natureza e o que ela realmente é, e que a natureza não tem nenhuma obrigação de concordar com nossa opinião e o Prof. Dr. Richard John Ward que me abriu a primeira porta e que ainda me ajuda, sempre que necessário. E o agradecimento final a “Dona Sônia”, minha professora da primeira série, ela deu a melhor resposta que um adulto poderia dar a uma criança curiosa, a resposta que me fez procurar, a resposta que marcar para mim o começo da caminhada que me levou até aqui: Não sei.

EPIGRAPH / EPÍGRAFE

“He is nothing, but adaptable. In his profession he has to be. Those who are not, die early.”

Stephen King, The Gunslinger – The Dark Tower Vol. 1

“Ele não é nada, a não ser adaptável. Em sua profissão, tem que ser. Aqueles que não o são, morrem cedo.”

Stephen King . O Pistoleiro – A Torre Negra Vol.1

RESUMO

da Silveira WA. **Análise do perfil genético de células tronco tumorais no câncer de mama localmente avançado.** 2015. 99f. Thesis (Doctoral) - Faculdade de Medicina, Universidade de São Paulo, Ribeirão Preto, 2015.

INTRODUÇÃO: O cancer de mama é no mundo o câncer mais comum em mulheres e a disseminação metastática é o principal fator relacionado com a morte pela doença. Acreditasse que as células tronco do câncer de mama - bCSC, na sigla em inglês e definida neste trabalho com a população $ALDH1^{high}/LIN^{-}/ESA^{+}$ - é responsável pela metástase e pela quimioresistência. O objetivo deste trabalho é encontrar genes que são essenciais para o controle do fenótipo das bCSC, em particular fatores de transcrição. **MATERIAIS E MÉTODOS:** Nesse trabalho nós utilizamos dois grupos de datasets com dados do transcriptoma, o grupo de datasets de descoberta contém um dataset gerado por nós com 3 amostras pareadas comparando as bCSC com o tumor total (My Data - bCSC/Bulk dataset), um dataset com 8 amostras pareadas comparando as bCSC com as células cancerígenas (Wicha - bCSC/CC dataset) e um dataset com 115 amostras de tecido de câncer de mama (Clinical Response dataset). O segundo grupo, grupo de validação, contém o dataset BRCA-TCGA com 621 amostras, as 4142 amostras de câncer de mama da ferramenta Kmpilot, as 17 amostras humanas primárias do subtipo BasL e sua informação sobre a geração, ou não, de tumores em camundongos imunossuprimidos e a análise de linhagens celulares (MF10A e HMLE). Para a análise dos dataset utilizamos o test-t pareado no pacote Limma da linguagem R, o algoritmo ARACNE para a inferência de regulons no dataset “Clinical Response”, a análise MRA-FET para definir os Reguladores Mestres para o fenótipo das bCSC e a análise GSEA para identificar o significado biológico de nossos achados nos diferentes datasets. **RESULTADOS E DISCUSSÃO:** Nós identificamos 12 TFs como reguladores mestres, com 9 deles formando duas redes altamente conectadas, uma positivamente relacionada ao fenótipo bCSC formada por SNAI2, TWIST, PRRX1, BNC2 e TBX5 com seus regulons, e definida aqui como a “rede de transcrição mesenquimal”, e uma rede correlacionada negativamente, formada por SCML4, ZNF831, SP140 e IKZF3, definida aqui como a “rede de transcrição da resposta imune” e totalmente desconhecida da literatura no contexto do câncer de mama. Embora ainda com fraca evidencia, ZEB1 para controlar as duas redes e ser responsável pela expressão de ALDH1 e dos 3 TFs restantes: ID4, HOXA5 e TEAD1. Como mostram seus nomes, e independente do dataset, do subtipo molecular ou da plataforma utilizada, a “rede de transcrição mesenquimal”, parece ser responsável pela manutenção do fenótipo de células tronco cancerígenas e a “rede de transcrição da resposta imune” pela resposta imune adaptativa ao tumor e a um bom

prognóstico para as pacientes. **CONCLUSÃO:** Nós encontramos e descrevemos duas redes de fatores de transcrição que parecem controlar o fenótipo das bCSC, uma delas totalmente desconhecida até agora e relacionada a um bom prognóstico. Nosso achados possuem um claro potencial para uso clínico.

Palavras Chave: Câncer de Mama, Célula-tronco, Transcriptoma, Biologia Sistêmica.

ABSTRACT

da Silveira WA. **Genetic profile analysis of tumor stem cells in locally advanced breast cancer.** 2015. 99f. Thesis (Doctoral) - Faculdade de Medicina, Universidade de São Paulo, Ribeirão Preto, 2015.

INTRODUCTION: Breast cancer is the most common cancer in women worldwide and metastatic dissemination is the principal factor related to death by this disease. Breast cancer stem cells (bCSC), defined in this work as the $ALDH1^{high}/LIN^{-}/ESA^{+}$ population, are thought to be responsible for metastasis and chemoresistance. The objective of this work is to find gene master regulators, in particular transcription factors (TFs), which are controlling the bCSC phenotype. **METHODS:** We used in this work two groups of datasets with transcriptome data, the discovery dataset group contains one dataset obtained by ourselves containing three paired samples comparing the bCSC and the bulk of the tumor (My Data - bCSC/Bulk dataset), a dataset with eight paired samples comparing the bCSC and cancer cells (Wicha - bCSC/CC dataset) and a dataset with 115 samples of breast cancer tissue (clinical response dataset). The second group, validation datasets, contains the BRCA-TCGA dataset with information of 621 samples, 4142 breast cancer samples of the Kmpot tool, 17 primary samples of BasL subtype and their information of grafting in patient derived xenografts and analyzes of cell lines (MF10A and HMLE). For the analyzes we used the paired t-test in the Limma R package, the ARACNE algorithm for the inference of regulons in the “clinical response” dataset, MRA-FET to define the master regulators of the bCSC phenotype, and GSEA to identify the biological meaning of the findings in the different datasets. **RESULTS:** We identified 12 TFs as master regulators of the bCSC phenotype, with nine of them forming two highly interconnected networks, one positively related with the bCSC phenotype formed by SNAI2, TWIST, PRRX1, BNC2 and TBX5 with its regulons, defined here as the “mesenchymal transcription network” and one negative correlated to the phenotype formed by SCML4, ZNF831, SP140 and IKZF3, defined as the “immune response transcription network”, totally unknown in the context of breast cancer in the literature. Although still with weak evidence, ZEB1 seems to control the two networks and can be responsible for the expression of ALDH1 and of the three remaining TFs: ID4, HOXA5 and TEAD1. As their names portray, our data showed in the different datasets, and independently of the molecular subtype and of the platform used, that the “mesenchymal transcription network” seems to be responsible for the bCSC phenotype and the “immune response transcription network” to the adaptive immune response in the tumor and a better prognosis for the patients. We also defined 10 membrane proteins as new

markers and/or therapeutic targets of the bCSC. CONCLUSION: We found and described two TF networks that seem to control the bCSC phenotype, one of them totally unknown until now and correlated to a good prognosis. Our findings have a clear potential for clinical use.

Keywords: Breast cancer, stem cell, transcriptome, System Biology

LIST OF FIGURES:

Figure 1: bCSC.....	4
Figure 2: Cancer stem cell markers in breast neoplasias.	5
Figure 3: Transcription Factors..	7
Figure 4: Network Motifs..	8
Figure 5: Transcription Networks and Master gene Regulators.....	9
Figure 6: Standardization FACS essay of 4T1 cell line using ALDEFLUOR® kit..	23
Figure 7: Cell integrity after FACS.	23
Figure 8: Searching for bCSC specific Transcription Factors.	24
Figure 9: bCSC specific Transcription Factors.....	26
Figure 10: Graphical representation of the regulons of the 17 Transcription Factors..	27
Figure 11: Mesenchymal Transcription Network.	30
Figure 12: Immune Response Transcription Network.	31
Figure 13: Hierarchical Clustering	32
Figure 14: : Hierarchical Clustering – “Wicha - bCSC/CC” dataset	32
Figure 15: Coordinated expression of the TFs in the two Validated networks are correlated to the Complete disappearance of the tumor under Treatment	34
Figure 16: GSEA Datasets positively correlated with High values for the Metagene in in the “Pathological response” dataset..	35
Figure 17: GSEA Datasets Negatively correlated with High values for the Metagene in the “Pathological response” dataset.	36
Figure 18: TCGA-BRCA dataset - Coordinated expression of the TFs in the two Validated networks.	38
Figure 19: GSEA Analysis, TCGA-BRCA. Positive correlation – Regulons of the Mesenchymal Network Transcription Factors.	39
Figure 20: GSEA Analysis, TCGA-BRCA. Negative correlation – Regulons of the Immune Response Network Transcription Factors.	40
Figure 21: GSEA Analysis, TCGA-BRCA. Positive correlation.	41
Figure 22: GSEA Analysis, TCGA-BRCA. Negative correlation.	42
Figure 23: PAM50. TCGA-BRCA dataset - Coordinated expression of the TFs in the two Validated networks.	43
Figure 24: Mammary Stem Cell signature. GSEA Analysis, PAM50. TCGA-BRCA. Positive correlation.	45
Figure 25: EMT signature. GSEA Analysis, PAM50. TCGA-BRCA. Positive correlation.	46
Figure 26: Natural Killer cell Mediated Cytotoxicity signature. GSEA Analysis, PAM50. TCGA-BRCA. Negative correlation..	47
Figure 27: EMT signature. GSEA Analysis, PAM50. TCGA-BRCA. Positive correlation. ,.....	48
Figure 28: Positive correlation between the expression of the 60 gene differentiation signature and Survival in Breast Cancer..	49
Figure 29: BasL Xenografts. Graphics of the enrichment results the Positive correlation of the basL samples in the Xenograft dataset of the expression of Mesenchymal Transcription Factors networks regulons in GSEA Analysis..	50

Figure 30: BasL Xenografts. Graphics of the enrichment results of the Negative correlation of the basL samples in the Xenograft dataset of the expression of Immune Respose	
Transcription Factor network regulons in GSEA Analysis	51
Figure 31: ZEB1 as possible Master Regulator of both networks of the bCSC phenotype in bCSC/CC dataset.	55

LIST OF TABLES:

Table 1: Sorting by FACs - Characteristics of the samples.....	14
Table 2: My Data - bCSC/Bulk dataset Clinical Data..	16
Table 3: Immunohistochemical status from the samples in “Wicha - bCSC/CC ” dataset (GSE52327).....	16
Table 4: Clinical Data from the “Clinical Response” dataset (GSE32646) tissue samples.	17
Table 5: PAM50 subsetting of the “TCGA-BRCA” dataset tissue samples.....	18
Table 6: Standardization	22
Table 7: Standardization	22
Table 8: bCSC Master Regulators.....	28
Table 9: Pathological Response Group, Chi2 Analysis.	34
Table 10: GSEA Analysis, TCGA-BRCA. Positive correlation – Regulons of the Mesenchymal Network Transcription Factors.	44
Table 11: GSEA Analysis, TCGA-BRCA. Negative correlation – Regulons of the Immune Response Network Transcription Factors.....	44
Table 12: BasL Xenografts, Take vs Not Take – Fold Change and p value fom T-test for the TFs of both networks.	51
Table 13: bCSC Membrane Protein Markers Candidates. A – Filters used in the selection. GSEA_pos_bCSC/CC: gene cointained in the Leading edge of Mesenchymal network positively correlated with the bCSC phenotype.....	53
Table 14: GSEA Analysis. MRA-FET TFs of the Mesenchymal Transcription Network in CD44+/CD24- in MCF10A cell line.....	54
Table 15: GSEA Analysis. MRA-FET TFs of the Mesenchymal Transcription Network in GSE24202 perturbation dataset of HMLE cell line	54
Table 16: GSEA Analysis. MRA-FET TFs of the Immune Response Transcription Network in GSE24202 perturbation dataset of HMLE cell line	54

LIST OF ABBREVIATIONS:

ALDH1	aldehyde dehydrogenase 1
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
bCSC	Breast Cancer Stem Cells
Bulk	tumor tissue sample as collected
CC	Cancer Cell
CCLE	Cell Line Encyclopedia
CSC	Cancer Stem Cell
EMT	Epithelial to Mesenchymal Transition
ER	Estrogen Receptor
ESA	Epithelial Specific Antigen, also known as EPCAM, Epithelial cell adhesion molecule
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
HER2	Human Epidermal Growth Factor Receptor 2
LIN	anti-human Lineage Cocktail (CD3, CD14, CD19, CD20, CD56)
MRA-FET	Fisher's Exact Test method of Master Regulator Analysis
nCR	non-pathological response
pCR	pathological Complete Response
PDXs	Patient derived xenografts
PgR	Progesterone Receptor
REP	Replicate, Sample
TFs	Transcription factors
TNBC	triple-negative breast cancer

SUMMARY

RESUMO	v
ABSTRACT	iii
LIST OF FIGURES:	v
LIST OF TABLES:	vii
LIST OF ABBREVIATIONS:	viii
1. INTRODUCTION	1
1.1 Breast cancer and intrinsic subtypes	2
1.3 Transcription Networks – simplicity becoming complexity	6
2. OBJECTIVES	10
3. MATERIALS AND METHODS	12
3.1 Cell Culture	13
3.2 Patients and tissue samples	13
3.2.1 Patients	13
3.2.3 Obtention of the samples	14
3.2.4 Enzymatic digestion and cell sorting	14
3.2.5 RNA extraction and quality control	15
3.3 Datasets	15
3.3.1 – Discovery datasets	15
3.3.2 – Validation datasets	18
4. RESULTS	21
4.1 Standardization	22
4.2 – Discovery of bCSC specific transcription networks	24
4.2.1 – Transcription Factor networks in the breast cancer stem cell Phenotype	24
4.2.2– Transcription Factor regulons and master regulators of the breast cancer stem cell phenotype.	25
4.2.3 – The expression of the TFs from both networks in the “My Data - bCSC/Bulk dataset”, “Wicha - bCSC/CC dataset”	29
4.2.4 Coordinated expression of TFs into two networks is correlated to the complete disappearance of the tumor under treatment, to epithelial to mesenchymal transition and to the immune response.	33
4.3 Test of the hypothesis: The expression of the transcription factors in different datasets.	33

4.3.1 The TCGA-BRCA dataset : TF network behavior and its biological meaning is independent of the molecular subtype and of the platform used.	33
4.3.2 The expression of the genes of the “immune response transcription network” is related to better survival in all PAM50 molecular subtypes.....	37
4.3.3 The grafting of basal-Like (basL) tumor samples in Xenografts	37
4.4 - Potential membrane protein markers for breast cancer stem cells	52
4.5 ZEB1 as a possible Master Regulator of the two transcriptions network.	52
4.6 Transcription Factor networks in the breast cancer stem cell Phenotype : Summary of the Results.....	56
5. DISCUSSION	59
5 – DISCUSSION	60
5.1 The mesenchymal transcription network	60
5.2 The immune response transcription network.....	61
5.3 The TCGA-BRCA dataset	64
5.4 Patient-derived tumour xenografts (PDXs)	65
5.5 The networks and the cell lines.....	66
5.6 New membrane proteins candidates for bCSC markers.....	67
5.7 ZEB1: A possible Master Regulator of the two networks	70
6. CONCLUSION	72
7. REFERENCES	74
8. SUPPLEMENTARY TABLES	94

1. INTRODUCTION

1. INTRODUCTION

1.1 Breast Cancer and Intrinsic Subtypes

Breast cancer is the most common cancer in women worldwide (Lee *et al.*, 2012) and metastatic dissemination is the principal factor related to death by this disease (The World Cancer Report - the major findings, 2003; Jemal *et al.*, 2010). The histological characteristics, protein expression patterns and the genetic profile of the cancer cells allow the characterization of different subtypes of the disease (Sørbye *et al.*, 2001; The World Cancer Report - the major findings, 2003; Sorlie *et al.*, 2003; Sørbye *et al.*, 2006; Guedj *et al.*, 2012). Although these biological markers have been described as prognostic factors, the mechanisms that underlie why similar tumors show a distinct biological behavior are still not elucidated. Tumor intrinsic heterogeneity is an acceptable hypothesis to explain the treatment failure and metastatic dissemination (Razzak *et al.*, 2008).

Locally advanced breast cancer generally refers to large primary tumors (>5cm) associated with skin or chest-wall involvement or with fixed (matted) axillary nodes (Society *et al.*, 2000). Breast tumors are highly heterogeneous and are classified based on: (1) histologically into in situ or invasive carcinomas and their subdivisions, (2) the expression of estrogen (ER) and progesterone (PR) receptors, respectively - and human epidermal growth factor receptor 2 (HER2) into ER+, HER2+, and ER-PR-HER2- (triple-negative breast cancer (TNBC)) subtypes, and (3) differentiation state/gene expression profiles into subtypes (Malhotra *et al.*, 2010; Polyak and Metzger Filho, 2012).

The first model to use expression profiles to subtype breast cancer was made by Sorlie and colleagues in 2003 (Sorlie *et al.*, 2003), this method evolved to the PAM50 classification (Nielsen *et al.*, 2014; Györfy *et al.*, 2015) and it is most commonly used nowadays. PAM50 classification is made of 5 subtypes: Normal-Like, Luminal A, Luminal B, HER2-enriched and Basal. The Normal-Like subtype is not totally accepted, when a sample falls in that group it is principally for a great proportion of normal mammary tissue in its composition (Prat *et al.*, 2010). The Luminal A tumor represents 50 %-60 % of invasive breast cancers, it frequently has low histological grade, low degree of nuclear pleomorphism, low mitotic activity and good prognosis; it is characterized by high expression of hormone receptor and associated genes (Schnitt, 2010; Prat and Perou, 2011; Yersal and Barutca, 2014). The Luminal B tumor comprises 15 %-20 % of invasive breast cancers and has a more aggressive phenotype, higher

histological grade, proliferative index and a worse prognosis than Luminal A, its expression of hormone receptor and associated genes is also lower than the Luminal A subtype and ~30 % of them express HER2 (Schnitt, 2010; Prat and Perou, 2011; Yersal and Barutca, 2014). HER2-enriched tumors represent ~15 % of invasive breast cancers and, as described by Yersal and Barutca, 2014, are characterized by high expression of the HER2 gene and other genes associated with the HER2 pathway and/or HER2 amplicon located in the 17q12 chromosome. Morphologically, these tumors are highly proliferative, 75 % have a high histological and nuclear grade and more than 40 % have p53 mutations (Carey, 2010; Schnitt, 2010; Prat and Perou, 2011; Yersal and Barutca, 2014). Basal subtypes represent ~15 % of invasive breast cancers, they are with high histological and nuclear grade, lymphocytic infiltrate and medullary features with exceptionally high mitotic and proliferative indices. Most of these tumors are infiltrating ductal tumors with a solid growth pattern, aggressive clinical behavior and high rate of metastasis to the brain and lung, they generally have poor prognosis (Heitz *et al.*, 2009; Schnitt, 2010; Prat and Perou, 2011; Yersal and Barutca, 2014).

There are other models of molecular classification of breast cancers (Kristensen *et al.*, 2014). The model defined by the group of Dr. Charles Theillet in 2012 (Guedj *et al.*, 2012), with a core dataset of 537 samples and test dataset of more than 3000 samples defined, from ascending aggressiveness: Normal-like, Luminal A, Luminal B, Luminal C, mApo, and Basal-Like.

1.2 The Breast Cancer Stem Cell

Mammalian cells have the ability to form tissues which require the sequential and overlapping activation and deactivation of numerous cellular programs in conjuncts of cells (Han, 2008; Benfey, 2011; El-Samad and Madhani, 2011; Skibinski and Kuperwasser, 2015). The malignant cancer cells differ very little from their normal counterparts, their difference lies principally in their incapacity to respond normally to environmental inputs and the partial loss of information stored in the DNA sequence which impairs, alters or prevents the execution of differentiation, tissue organization and multiplication programs (Wang, 2010; Hanahan and Weinberg, 2011; Ferguson *et al.*, 2015).

In the basis of the cancer cells hypothesis lies the idea of a treatment-resistant subpopulation of tumor cells that possess the capacity to self-renew and to cause the heterogeneous lineages of cancer cells that comprise the tumor (Clarke *et al.*, 2006; Al-Ejeh *et al.*, 2011; Skibinski and Kuperwasser, 2015).

Figure 1: bCSC. Hypothetical model of breast cancer stem cells. Modified from Shipitsin & Polyak, 2008

Nonetheless, although the idea of disruption or corruption in the process of breast tissue formation as a cause to cancer and emergence of the breast cancer stem cells (bCSC) is well accepted (**Figure 1**), the exact process that originated it and maintain it are not totally elucidated (Shipitsin and Polyak, 2008). The process of epithelial-to-mesenchymal transition, a highly conserved process of cellular reprogramming that transforms epithelial cells into mesenchymal cells, has a close relationship with bCSC formation and maintenance (Mallini *et al.*, 2014; Li and Li, 2015; Liu and Fan, 2015; Tan *et al.*, 2015).

Figure 2: Cancer stem cell markers in breast neoplasias. Modified from Schmitt *et al.*, 2012.

In breast cancer, a subpopulation of cells enriched for bCSC can be identified by the expression of proteins in the cellular membrane. The $ESA^+/CD44^+/CD24^-$ phenotype cells have a high initiation and self-renew capabilities in breast carcinoma (Al-Hajj *et al.*, 2003). Recent studies have shown other ways to identify the bCSC population (**Figure 2**) (Alison *et al.*, 2010; Tsukabe *et al.*, 2013). Aldehyde dehydrogenase 1 (ALDH1) activity seems to be a better marker than immunophenotyping for bCSC identification (Ginestier *et al.*, 2007; Charafe-Jauffret *et al.*, 2010). The high activity of the ALDH1 selects cells with a bipotential capacity in normal breast tissue, this intermediate state of differentiation between the basal and luminal axis represents the point with highest phenotypic plasticity and therefore of stem-like functionality (Yu *et al.*, 2013; Condiotti *et al.*, 2014; Granit *et al.*, 2014). The percentage of the bCSC can vary widely in invasive carcinomas and a correlation between its proportion on the tumor and patient prognostic was expected (Abraham *et al.*, 2005; Mylona *et al.*, 2008; De Beça *et al.*, 2013), nonetheless, as some works of our own group show, it is still an open issue (Tiezzi *et al.*, 2013; Mandarano, 2013).

Neoadjuvant chemotherapy has been used as the standard treatment for locally advanced breast cancer (Fisher *et al.*, 1997; Beriwal *et al.*, 2006). The group of patients who

objectively respond to systemic treatment, especially with complete pathologic response, have an improvement in disease free and overall survival rates when compared to unresponsive patients (Fisher *et al.*, 1998; Amat *et al.*, 2005). Nevertheless, the determinant factor of tumor chemotherapy sensitivity is unknown. Recent evidence shows that the presence of bCSC in solid tumors can be responsible for the lack of treatment response (Li *et al.*, 2008; Gottschling *et al.*, 2012; Gangopadhyay *et al.*, 2013).

Thus, the bCSC identification and the study of its expression profile can bring useful information to predict the outcome of the cytotoxic treatment in breast cancer, open new perspectives for drug development, diminish the phamaco-economic impact of the treatment and, last but not least, can lead to attenuate side effects from neoadjuvant chemotherapy.

1.3 Transcription Networks – Simplicity becoming complexity

Transcription Factors are the principal regulators of expression in mammalian cells (Alon, 2006; Vaquerizas *et al.*, 2009; Carro *et al.*, 2010; Theunissen and Jaenisch, 2014). A TF is defined as any protein required to initiate or regulate transcription in eukaryotes (Rédei, 2008). In normal cells the expression of TFs are regulated by environmental states, and the activities of the TFs can be considered as an internal representation of the environment (**Figure 3**) (Alon, 2006). A conjoint of TFs act together to build the adequate states, regulating their target genes to mobilize the appropriate protein response according to specific signaling (Alon, 2006). The existence of 1700 to 1900 TF coding genes in the human genome was estimated, close to only 6 % of the protein-coding genes (Vaquerizas *et al.*, 2009).

One useful concept to bear in mind when studying TFs and expression networks is the concept of regulons. In eukaryotes, a regulon is a genetic unit consisting of a noncontiguous group of genes under the control of a single regulator gene (Medical Subject Headings, 2015). The regulons of each TF are not a static list of genes, each gene in our genome can have multiple TF binding sites that can be accessible or inaccessible depending of the cellular context. The regulon of a specific TF will vary in different cells and in the same cell type, in different conditions (Alon, 2006; Bruce Alberts, 2007; Nickel and Stadler, 2015).

The TFs in transcription networks are organized in logical circuits with repetitive patterns, a few examples of that are depicted in **Figure 5**. A breakdown in this regulatory

system can cause a great number of diseases (Habener and Stoffers, 1998; Martin *et al.*, 2005; Hannenhalli *et al.*, 2006) and transcription factors are overrepresented among oncogenes (Vaquerizas *et al.*, 2009).

Figure 3: Transcription Factors. Mapping between the environmental signals, transcription factors inside the cell and the genes they regulate. From Alon, 2006.

Within one transcription network there is a hierarchy of TF regulation, with the one at the top, the master regulator, regulating great part of the network and being regulated by a few other TFs. More specifically one “master regulator is a gene that is expressed at the inception of a developmental lineage or cell type, participates in the specification of that lineage by regulating multiple downstream genes either directly or through a cascade of gene expression changes, and critically, when misexpressed, has the ability to specify the fate of cells destined to form other lineages” (Chan and Kyba, 2013).

Using the concept of regulons, gene circuits and master regulators, is possible to interpret a large transcription network, such as the 149 genes and several hundreds of interactions of the “mesenchymal transformation network” of high-grade glioma brain tumors (**Figure 5.A**) in a simple network with five master regulators with 10 interactions (**Figure 5.B**) (Carro *et al.*, 2010)

Considering the regulatory importance of TFs and the intrinsic noise data coming from the analyzes of multiple transcriptome datasets, we decided to focus on the expression of TFs as a proxy of the internal state of bCSC and as a signal of the mechanisms these cells are using to maintain their phenotype.

The way you construct your evaluation method defines which type of information you can acquire from it. With that in mind, we compared bCSC with the bulk of the tumor and bCSC with cancer cell strategies to infer two TF networks that seem to switch off the bCSC phenotype, with clear potential for clinical use.

Figure 4: Network Motifs. Examples of network motifs commonly found in transcription networks. Modified from Alon, 2006.

Figure 5: Transcriptions Networks and Master gene Regulators. A: relationships between the genes in the transcriptional network for mesenchymal transformation of high-grade glioma brain tumors. Pink: TF network activators, Purple: Repressive TFs. B: Master Regulators of the networks responsible for the regulation of 74 % of the genes of the signature. Modified from Carro et al., 2010.

2. OBJECTIVES

2 - OBJECTIVES:

2.1 General Objective:

The objective of this work was to find gene master regulators, in particular TFs, which are controlling the bCSC phenotype.

2.2: Specific Objectives:

- Define the genes regulated by each master regulator;
- Identify the biological impact of the expression of the bCSC master regulators in the entire breast cancer tissue in different datasets;
- Propose new possible markers for the bCSC phenotype and signaling pathways involved with the expression of the master regulators

3. MATERIALS AND METHODS

3 - MATERIALS AND METHODS:

In this work we defined the bCSC as the ALDH1^{high}/LIN-/ESA+ cell population in the tumor tissue. Using equipment coupled to the flow cytometer known as a fluorescence-activated cell sorter, we were able to sort this population and study it individually.

3.1 Cell Culture

In order to standardize the RNA extraction and fluorescence activated cell sorting (FACS) methods we used the cell lines ZR75-1 and 4T1. Culture conditions were performed as described by the American Type Culture Collection (ATCC). For both, we used RPMI-1640 Medium with 10 % fetal bovine serum and 1 % of antibiotics, as recommended by the ATCC.

3.2 Patients and Tissue samples

3.2.1 Patients

We prospectively sampled breast tumors from 40 patients at the Hospital das Clínicas, in the city of Ribeirão Preto, Brazil. The local ethics committee in has approved the study with the protocol number: 2467/09. All the patients were informed of the objectives of the study and signed a free and informed consent document before their inclusion. One patient refused to be part of the study for religious reasons.

From the 40 patients, 21 tissue samples were sorted by FACS (**Table 1**). From these 21 samples, two were excluded because of a negative diagnostic for invasive ductal carcinoma (samples 11 and 15), two were excluded because they presented a low quality of extracted RNA (RIN below 6, samples 2 and 19) and eight were excluded because they have not reached the minimum amount of 50 ng of RNA necessary for the microarray procedure (samples 1, 6, 7, 13, 16, 17, 18 and 21). From the 21 samples, only nine had RNA in sufficient quality and quantity for microarray procedure. Three samples were excluded due to technical problems during or after the procedure. Thus, only six samples were suitable for whole gene expression analysis. We used samples 12, 14 and 20 to generate the data.

Patient	Initials	%bCSC	Conc. bCSC (ng/ul)	Conc. Bulk (ng/ul)	RIN bCSC	RIN Bulk
1	AASF	0.1	6.3	578.5	7.7	<6
2	SC	0.2	15.7	-	4	-
3	EFMN	1.9	19.2	-	8.5	-
4	ASR	2.9	30	-	7.7	-
5	KAD	0.7	60	-	8.8	-
6	CAR	1.1	20	95.5	7.1	8.4
7	SML	13.6	2.14	-	6	-
8	IPM	0.8	12.67	-	8	-
9	RXA	2.9	44.3	-	8	-
10	ICSM	2.4	54.8	-	7.6	-
11	RAPL	5.2	5	51.2	-	-
12	MFN	3.9	46.2	36.06	9.1	<6
13	MJS	4.3	5.2	22.3	8.6	<6
14	LST	2.4	69.3	258.4	8.4	9.8
15	EAM	2	51.3	29.5	-	-
16	ALV	0.7	7.6	161.3	8.3	<6
17	MMMS	3.4	7.35	238.3	8.6	<6
18	ARSS	1.8	3.8	235	8.6	<6
19	DACRM	2.6	10.5	30.6	3.2	-
20	MHF	1.8	58	141.1	8.6	9.8
21	APRC	3.25	7.3	789.2	7.7	<6

Table 1: Sorting by FACs - Characteristics of the samples. Patient: Identification of the sample in the study. Initials: Initials of the name of the patients who donate the samples. %bCSC: % of the ALDH1high/Lin-/ESA+ population in the sample. Conc. bCSC: Concentration of RNA extracted from the bCSC cells. Conc. Bulk: Concentration of RNA extracted from the entire tissue sample. RIN bCSC: RNA integrity number of the RNA from the bCSC. RIN Bulk: RNA integrity number of the RNA from the entire tissue sample.

3.2.3 Samples collection

All samples were obtained by percutaneous ultrasound-guided biopsy as a routine procedure in the hospital. One core fragment per patient was used in this study. The fragments were separated into halves. One half was used for cell sorting and the other for RNA purification.

3.2.4 Enzymatic digestion and Cell Sorting

Fresh tissue samples were minced with a scalpel blade and then mixed with a final volume of 1 ml of a solution of Collagenase IV (1mg/ml) at 37 °C for 1 hour in agitation. After digestion, the cellular suspension was filtered (mesh BD 70µm), washed twice with RPMI-1640 Medium, sedimented by centrifugation, and the pellet was re-suspended in 500 µL of Aldefluor Assay Buffer[®]. The total amount of live cells was estimated by trypan blue exclusion. The final concentration of live cells was adjusted to 1 x 10⁶ live cells/ml using the Aldefluor Assay Buffer[®].

To identify the bCSC and sort them by FACS we used the ALDEFLUOR[®] kit (Aldagen) was used, as specified by the manufacturer, to evaluate the activity of the enzyme ALDH1. The antibody anti-ESA (anti-EpCam) was also used, to identify epithelial cells and a pool of antibodies called Lin (eBioscience, containing anti-CD2, anti-CD3, anti-CD14, anti-CD16, anti-CD19, anti-CD56, anti-CD235a) plus anti-CD31 and anti-CD45, the pool of antibody identify cells from the hematopoietic lineage. In this way we were able to separate a cell with high activity of ALDH1 from the luminal lineage ALDH1^{high}/LIN⁻/ESA⁺ cell population. The flow cytometry assay was performed in the FACS Aria II (BD Biosciences, San Jose, CA) and the analysis of the data was performed using the FlowJo software (TreeStar, USA).

3.2.5 RNA extraction and quality control

The total RNA was extracted using the Mirvana Kit (Ambion, USA) as specified by the manufacturer. We estimated the RNA quantity and purity by UV spectrometry by A220/A260 and A260/A230 ratios. RNA integrity was evaluated using the RNA 6000 Nano Kit, RNA 6000 Pico Kit and 2100 Bioanalyzer (Agilent Technologies, USA).

3.3 Datasets

3.3.1 – Discovery Datasets

3.3.3.1 – My Data - bCSC/Bulk dataset

RNA from samples of patients described in item 3.2.1 were extracted, both from bCSC and from the bulk of the tumor, and analyzed utilizing the GeneChip[®] Human Gene 2.0 ST Array (Affymetrix, USA) in the International Research Center (CIPE), A.C. of the Camargo Cancer Center, São Paulo, SP, Brazil. This analysis was performed with the transcriptome information from 3 samples (ER+/HER2+, HER2+ and TN) paired as bCSC against the bulk, and the two groups were analyzed one against the other, as described below. The clinical data from the patients are depicted in **Table 2**.

Patient Number	Age	Histologic Grade	ER	PgR	HER2	T	N	M	Clinical Stage
12	50	3	POS	POS	POS	4B	3C	1	IV
14	61	3	NEG	NEG	POS	4D	2	1	IV
20	40	3	NEG	NEG	NEG	2	0	0	IIA

Table 2: My Data - bCSC/Bulk dataset Clinical Data. Clinical data from the patients whose samples fulfilled all the requirements of the study. These patients were included in the “My data - bCSC/CC” dataset. ER: Estrogen receptors, PgR: Progesterone receptors, HER2: HER2 receptors. TNM staging as used by the American Joint Committee on Cancer. T.N.M – AJCC classification, T = Tumor stage (The tumour being 2cm across or less = 1, The tumor being larger than 5cm and Inflammatory = 4D), N = Lymphonod stage (No cancer cells found in any nearby nodes = 0, cancer cells in lymph nodes above the collarbone = 3c), M = Metastasis stage (No signals of mestastasis = 0, Metastasis = 1). Clinical stage: Anatomic stage/prognostic groups based on the T.N.M evaluation. Histological Grade: Nottingham–Bloom–Richardson system(Tavassoli Fa, 2003), Grade 1 = well differentiated, Grade 3 = poorly differentiated.

3.3.3.2 – Wicha - bCSC/CC dataset

The “Wicha - bCSC/CC” dataset was taken from the GEO repository, its accession number is GSE52327. It is composed of paired samples of 8 patients. The samples were divided into bCSC (ALDH1+/LIN-/ESA+) and cancer cells (CC, ALDH1-/LIN-/ESA+) and the two groups analyzed one against the other, as described below. The RNA extract was analyzed with the Human Genome U133 Plus 2.0 Array chip (Affymetrix, USA)(Liu *et al.*, 2014). The Immunohistochemical status of the samples are described in **Table 3**.

Replicates	ER	PR	HER2
REP 1	+	+	-
REP 2	-	-	-
REP 3	-	-	+
REP 4	+	+	-
REP 5	ND	ND	ND
REP 6	+	+	-
REP 7	+	+	-
REP 8	+	+	-

Table 3: Immunohistochemical status from the samples in “Wicha - bCSC/CC ” dataset (GSE52327). Breast Cancer Stem Cell X Cancer Cell, ER = Estrogen Receptor, PR = Progesterone Receptor, HER2 = human epidermal growth factor receptor 2, POS = Positive, NEG = Negative. + = positive, - = negative, ND = information Not Available.

3.3.3.3 – Clinical Response dataset

The clinical response dataset was taken from the GEO repository, its accession number is GSE32646. The dataset consists of transcriptome data of 115 tissue samples from patients with breast cancer acquired by core biopsy prior to chemotherapy (Miyake *et al.*, 2012). The RNA extract was analyzed with the Human Genome U133 Plus 2.0 Array chip (Affymetrix, USA) (Liu *et al.*, 2014). As described by Miyake et al:

“Primary breast cancer patients (n = 123, T1-4b N0-1 M0) who were consecutively recruited for the present study had been treated with NAC consisting of paclitaxel (80 mg/m²) weekly for 12 cycles followed by 5-FU (500 mg/m²), epirubicin (75 mg/m²) and cyclophosphamide (500 mg/m²) every 3 weeks for four cycles (paclitaxel followed by 5-fluorouracil/epirubicin/cyclophosphamide [P-FEC]) at Osaka University Hospital between 2004 and 2010.”

The dataset was divided into two groups by response to the treatment, pathological complete response (pCR) and non-pathological response (nCR) groups. The immunohistochemical status of the samples in the dataset is described in **Table 4**.

Status	pCR	nCR	Total
ER+	5	50	55
HER2+	12	22	34
TN	10	16	26
Total	27	88	115

Table 4: Clinical Data from the “Clinical Response” dataset (GSE32646) tissue samples. Status = Expression of Immunohistochemical Markers, ER+ = Estrogen Receptor positive, HER2+ = human epidermal growth factor receptor 2 positive, TN = Triple Negative, negative for the expression of ER, HER2 and Progesterone Receptor; pCR= Pathological Complete Response to paclitaxel, nCR = Non-Complete Response to paclitaxel.

PAM50 Subtype	TCGA-BRCA Samples
Normal-Like	15
Luminal A	189
Luminal B	162
HER2-enriched	107
Basal	148
Total	621

Table 5: PAM50 subsetting of the “TCGA-BRCA” dataset tissue samples. The samples used were from females diagnosed with Invasive ductal carcinoma of the breast.

3.3.2 – Validation Datasets

3.3.2.1 TCGA – BRCA Dataset

From “The Cancer Genome Atlas” (TCGA), we used the public available transcriptomic data acquired by RNAseq from the Illumina platform. The RNAseq dataset from 1000 patients was downloaded using the R environment using the TCGA-Assembler package using the function `DownloadRNASeqData`, with the following arguments: `cancerType = "BRCA"`, `assayPlatform = "RNASeqV2"`, `dataType = "gene.quantification"`. RNAseq data was processed using the `ProcessRNASeqData` function and the argument `verType = RNASeqV2`. RNASeqV2 pipeline in TCGA extracts the normalized count values. Clinical and pathological data were downloaded with the TCGA-Assembler package and we selected 621 samples with histological confirmed invasive ductal carcinoma from females (Ma and Ellis, 2013; Zhu *et al.*, 2014). We classified the samples in the PAM50 subset using the “`intrinsic.cluster`” function in the “`genefu`” R package (Table 5) (Perou *et al.*, 2000; Sørlie *et al.*, 2001; Haibe-Kains B, 2014) with the arguments: `do.mapping=F`, `std="scale"`, `number.cluster=5`, `mins=5`, `method.cor="spearman"`, `method.centroids="mean"` (Haibe-Kains B, 2014).

3.3.2.2 Survival Dataset

The survival analysis was made using the web tool “Kaplan Meyer Plotter” (Györfy and Schäfer, 2009; Györfy *et al.*, 2010), which uses 4142 breast samples from the GEO repository, the technology used was the HGU133 Plus 2.0 array from Affymetrix (Li *et al.*, 2011; Mihály *et al.*, 2013). To make the survival curve we used the mean expression of 60 genes derived from the “immune response transcription network” regulons and up-regulated in the Wicha-bCSC/CC and the clinical response dataset (**Supplementary Table 2**).

3.3.2.3 Xenografts

We used the data from breast cancer samples generated in the Theillet group at the Institut de Recherche en Cancérologie de Montpellier (IRCM) for the building of a collection of patient derived xenografts (PDXs) and kindly shared with us by Dr. Stanislas Du Manoir (Du Manoir, 2013). We compared the initial transcriptome profile of samples that have generated PDXs (Take) against those that did not (No Take).

3.3.2.4 Cell Lines

We also evaluated two datasets of cell lines, the GSE15192 (Bhat-Nakshatri *et al.*, 2010), which contains affymetrix microarray data of MCF10A cell line from CD44+/CD24- and CD44-/CD24+ cells, and the GSE24202 (Taube *et al.*, 2010), which contains immortalized HMLE breast epithelial cells which were retrovirally transduced in culture with vectors encoding epithelial to mesenchymal transition (EMT) inducing genes or control vectors.

3.4 - Molecular Profile Analysis of the Datasets.

We normalized the samples values by robust multi-array average (RMA) using the "affy" package in R (Gautier *et al.*, 2004; Team, 2012) in the case of the "Wicha - bCSC/CC" and the "clinical response" dataset, and with the "oligo" package (Carvalho and Irizarry, 2010) in the case of the "My Data - bCSC/Bulk dataset". We annotated the samples using the "hgu133plus2.db" and "hugene20stprobeset.db" packages and summarized the data by max strategy (M). Using the "limma" function on the POMELOS2 website (Morrissey and Diaz-Uriarte, 2009; Ritchie *et al.*, 2015) we applied a paired t-test in the "My Data - bCSC/Bulk" and "Wicha - bCSC/CC" datasets and t-test in the "clinical response" dataset.

In the "My Data - bCSC/Bulk" and "Wicha - bCSC/CC" datasets we selected the transcription factor genes (Vaquerizas *et al.*, 2009) with $p \leq 0,05$ and a fold change greater than 2, that were going in the same direction in the two datasets.

From that we inferred the regulons of these genes using the ARACNE algorithm as a method to perform reverse engineering of cellular networks (Margolin, Wang, *et al.*, 2006; Carro *et al.*, 2010). Each of the regulons was analyzed by Fisher's exact test method of master

regulator analysis (MRA-FET) (Carro *et al.*, 2010). Based on the relationships of the TFs inferred by ARACNE and confirmed by MRA-FET we constructed the networks depicted in Figure 11 and Figure 12. We then calculated in the “clinical response dataset”, the value of the metagene of the expression of the nine TFs evaluated as master regulators using the as coefficient for each gene its fold change in the bCSC/CC dataset, using the genefu package in R (Haibe-Kains B, 2014). We ordered the samples by that score and depicted them as a heatmap using the gplots package in R (Gregory R. Warnes *et al.*, 2015), the hierarchical clustering was performed using the complete linkage method (Murtagh and Contreras, 2012), we divided the ordered dataset into two halves and calculated the difference in pathological complete response between them using Chi-square analyzes. Finally, we used the score of the metagene as a continuous phenotype (using Pearson metrics) in GSEA analyzes to evaluate what is positively and negatively correlated with its expression. This same method was used in the analysis of the “TCGA-BRCA” dataset.

In the xenograft and the cell lines dataset we evaluated the expression of the TF regulons by GSEA analysis we normalized the data in the same way described above, only changing the annotation for the affymetrix chips, when necessary.

4. RESULTS

4 – RESULTS

4.1 Standardization

Before sorting cells from tumor tissues and extracting their RNA we performed a first experiment with cell lines ZR7531 and 4T1. These procedures were used as a standard protocol for sorting tumor cells. We extracted RNA from different numbers of cells (25000 and 400000). With RNA purification, sufficient RNA quantity and purity was achieved with 4×10^5 cells (**Table 6**).

Test	Number of Cells	RNA Concentration	Total RNA	260nm/280nm
1	25000	1.97ng/ μ L	197ng	NA
2	400000	2744.00ng/ μ L	274400ng	2

Table 6: Standardization . RNA extraction from ZR7531 cell line using the Mirvana Kit.

We utilized the 4T1 cell line of murine breast cancer to establish the protocol for sorting the ALDH^{high} cells by FACS. We used two different sets in the FACS Aria II flow cytometer (**Table 10**), in both tests we used the same protocol for cell culturing and staining protocols with distinct flow speed. Although we haven't found a great difference in ALDH^{high} cell phenotyping, we observed an expressive decrease in sorting efficiency with increased flow velocity. For this work we have used the sets established in test 3. Using this same protocol in the cells from the tumor tissue samples we achieved an efficiency of 90 %.

Test	Cell Initial Number	%ALDH1high cells phenotyped	%ALDH1high cells Sorted	Efficiency (%)
3	1000000	11.2	5.7	50.5
4	1000000	9.65	0.2	1.8

Table 7: Standardization . Test of two different specifications for sorting ALDH1^{high} cells.

The stress caused by the trypsinization, labeling and the flux of the cell in the flow cytometer before the sensors caused close to 10 % of cell death (**Figure 6B** and **Figure 6E**). Nonetheless, after that step we still have the ionization of the drop and the impact of the drop with the internal part of the collection tube. We feared an expressive break in the membrane of the cells, causing a premature degradation of RNA before the start of the RNA extraction process, but a large number of the cells maintain their integrity after cell sorting (**Figure 7**).

Figure 6: Standardization FACS essay of 4T1 cell line using an ALDEFLUOR® kit. A, B and C: Use of DEAB to inhibit ALDH1 activity. D, E and F: no inhibition of ALDH1. P1: total population, P2: PI- cells, Living cells defined by the absence of Propidium Iodide inside the cell, P3: PI-/ALDH^{low} population. P4: PI-/ALDH^{high} population.

Figure 7: Cell integrity after FACS. HE of Cytopsin from 4T1 PI-/ALDH^{high} after separation for FACS. 100x.

4.2 – Discovery of bCSC specific transcription networks

4.2.1 – Transcription factor networks in the breast cancer stem cell phenotype

The transcriptome analysis of the “My Data-bCSC/Bulk” dataset gives us the genes differentially expressed when we compare the bCSC population against the tumor microenvironment, with the heterogenous cellular population consisting of fibroblasts, endothelial cells, immune cells and the more differentiated cancer cells. Making the same analysis with the “Wicha - bCSC/CC” dataset give us the genes differentially expressed between the bCSC and the more differentiated cancer cells, but ignores the microenvironment.

We based the choice of which transcriptor factor genes to study on a simple mathematical idea from set theory:

“The intersection of sets x and y is the set consisting of those objects that are members of both x and y .” (Devlin, 2012)

As depicted in **Figure 8**, the intersection set of genes differentially expressed in both the “My Data-bCSC/Bulk” and “Wicha - bCSC/CC” datasets, leaves us with the TFs differentially expressed only in bCSC.

Figure 8: Searching for bCSC specific transcription factors. The Intersection of the TFs differentially expressed in the “My Data-bCSC/Bulk” and “Wicha - bCSC/CC” datasets, leaves us with the TFs differentially expressed only in bCSC.

Using this strategy we found 17 TF genes differentially expressed in both datasets with $p \leq 0.05$ and fold change ≥ 2.0 (**Figure 9**): PRRX1, SNAI2, TWIST1, ID4, BNC2, GATA6, ZNF503, FOXF2, TBX5, HOXA5, HOXB3, TSC22D1, CREB3L1, SCML4, ZNF831, IKZF3 and SP140. The Pearson correlation coefficient of the expression of these TFs in the datasets is 0.84 (**Figure 9.B**). There are 245 TF genes with $p \leq 0.05$ and fold change ≥ 2.0 in the “My Data-bCSC/Bulk” dataset (**Figure 9.C**) and 119 in the “Wicha - bCSC/CC” (**Figure 9.D**). This difference in numbers of TF genes differentially expressed was expected, given the nature of the comparison of the two datasets.

4.2.2 – Transcription factor regulons and master regulators of the breast cancer stem cell phenotype.

Considering the way the TF works in the cell (Alon, 2006), the next logical step was to infer which genes are regulated by each TF – the TF regulons – in breast cancer. We did this by applying the ARACNE algorithm to the clinical response dataset (Margolin, Wang, *et al.*, 2006; Miyake *et al.*, 2012). This dataset has 115 samples, and we used it because to make an ARACNE inference a dataset with at least 100 samples is necessary (**Figure 10**) (Margolin and Califano, 2007). It is interesting to note that the inferred network is divided in two separated highly connected blocks, which agree with the expression of the TFs in the “My Data-bCSC/Bulk” and “Wicha - bCSC/CC” datasets (**Figure 9**). The complete set of genes in each inferred regulon is listed in **Supplementary Table 1**.

We then evaluated by master regulator analysis – Fisher’s exact test (MRA-FET) (Carro *et al.*, 2010) which of these regulons can be considered a master regulator of the bCSC, comparing with cancer cells, in the “Wicha - bCSC/CC” dataset (**Table 8**). MRA-FET gave us 12 master regulators, eight positively correlated with the bCSC phenotype (BNC2, PRRX1, TBX5, SNAI2, TWIST1, ID4, HOXA5, TEAD1), and four negatively correlated (IKZF3, SCML4, SP140, ZNF831). It is interesting to note here that the TF regulons negatively correlated with the bCSC phenotype have a much lower p-value than the positively correlated, even when we compare regulons with almost the same size, such as IKZF3 ($p = 4.60 \times 10^{-152}$) and BNC2 ($p = 7.24 \times 10^{-22}$).

Figure 9: bCSC specific Transcription Factors. **A** – Table containing the list with the 17 transcription factors differentially expressed going in the same direction in both “My Data-bCSC/Bulk” and “Wicha - bCSC/CC” datasets, comparing the fold change presented in each dataset. Red = Up-regulated, Green = Down-regulated. $p \leq 0.05$. **B** – Graphic depicting the expression patterns of the TFs in both datasets, r = Pearson correlation coefficient. **C** and **D** – Volcano plots depicting the expression behavior, in fold change and p-value, of the TFs in “My Data-bCSC/Bulk” (**C**) and “Wicha - bCSC/CC” (**D**) datasets. Green dots = TFs with $p \leq 0.05$ and fold change ≥ 2.0 , Red dots: TFs with $p \leq 0.05$, Yellow dots = TFs with fold change ≥ 2.0 , Blue Dots = The 17 TFs in both datasets with $p \leq 0.05$ and fold change ≥ 2.0 , with special attention to SCML4, ZNF831, SP140, IKZF3, SNAI2, TWIST1, BNC2, TBX5 and PRRX1. Black dots = All the others. Paired t-test, limma.

Figure 10: Graphical representation of the regulons of the 17 Transcription Factors. Inferred in the “pathological response” dataset by ARACNE algorithm. Green triangles represent genes, the lines linking them represent a relationship between the expression of them. **Regulons:** **1** – SNAI2, TWIST1, PRRX1, BNC2, TBX5, **2** – SCML4, ZNF831, SP140, IKZF3, **3** – ID4, **4** – HOXA5, **5** – ZNF503, **6** – TSC22D1, **7** – HOXB3, **8** – CREBL1, **9** – GATA6. The Regulon of FOXF2 has 32 genes, these genes are shared by the regulons of SNAI2, TWIST1, PRRX1, BNC2, TBX5, ID4 and HOXA5 and are not identifiable in the figure.

Master Regulator	FET P-Value	Markers in regulon	Markers in intersection set	Mode
IKZF3	4.60E-152	375	255	-
SCML4	4.49E-147	419	266	-
SP140	9.48E-145	407	260	-
ZNF831	1.74E-138	308	220	-
BNC2	7.24E-22	369	108	+
HOXA5	4.37E-20	471	124	+
PRRX1	2.42E-18	341	96	+
SNAI2	1.23E-12	90	36	+
TBX5	8.01E-10	222	57	+
TWIST1	3.34E-09	162	45	+
ID4	9.06E-09	248	60	+
TEAD1	7.69E-08	86	30	+

Table 8: bCSC Master Regulators – Genes selected as master regulators in the “bCSC/CC” dataset by MRA-FET. Mode “-” = negative correlated with the bCSC phenotype, “+” = positive correlated with the bCSC phenotype.

Of the 12 TFs selected as master regulators, five of the positively correlated (BNC2, PRRX1, TBX5, SNAI2, TWIST1) formed a well-connected and logical circuit network (**Figure 11**). BNC2, PRRX1 and TBX5 form three pairs of double-positive feedback loop (**Figure 4, Figure 11.A**), which means that the activation of one of them is possibly sufficient to lock the high expression of the three without the necessity of any other stimuli. TWIST1, SNAI2, PRRX1 and BNC2 form a motif called Bifan (**Figure 4, Figure 11.A**), an example of a simple overlapping regulation pattern, when two genes (TWIST1 and SNAI2) regulate the expression of the other two (BNC2 and PRRX1). The increase in complexity of this network is that SNAI2 is at the same time an “input” and an “output”. With this first analysis it was not possible to infer the existence of logical gates (Alon, 2006) or of analogical computation (Sarpeshkar, 2014). The genes in the regulons of these five TFs, selected by MRA-FET, presents a high degree of overlapping (**Figure 11.B and 11.C**) BNC2, PRRX1 and TBX5 regulons have 40 overlapped genes, which represents 70 % (40/57) of the genes of the TBX5 regulon, 37 % (40/96) of the BNC2 regulon and 41 % (40/108) of the PRRX1 regulon. If we consider only BNC2 and PRRX1, they have 73 overlapped genes, which means 67.6 % (73/108) and 76.0 % (73/96) of their regulons, respectively. The regulon of SNAI2 overlaps completely with BNC2, PRRX1 and TBX5 regulons, having 50 % (18/36) of the genes overlapping with all three, 25 % (9/36) only with BNC2 and PRRX1 and 25 % only with BNC2. The TWIST1 regulon overlaps 44 % (20/45) with BNC2, PRRX1 and TBX5, and has 33 % with no overlapping. TWIST1 and SNAI2 have 14 overlapped genes, all of them also in common with BNC2, PRRX1 and TBX5 regulons.

The evaluation of the biological meaning of this network (**Figure 11.D**) shows a great correlation of these genes and signatures related to invasiveness, EMT and stem cell properties, using GSEA analysis in the list of the 40 genes in common with PRRX1, BNC2 and TBX5 regulons, selected by MRA-FET. From now on we will call these five TFs factors and their interactions as the “mesenchymal transcription network”.

The list of the four master regulators that negatively correlate to the bCSC phenotype (IKZF3, SCML4, SP140, ZNF831) also assume a well-connected and logical circuit format shape, **Figure 12.A**. ZNF831 and SCML4 assume a double-positive feedback loop, and the four genes themselves form a multi-input feed-forward loop (**Figure 4**). There is a large overlap (**Figure 12.B**) of 169 genes in common in the four regulons, this means 66.3 % of IKZF3, 65 % of SP140, 63.5 % of SCML4 and 76.8 % of ZNF831 regulons. Once evaluated with GSEA we encounter that the 169 overlapped genes are related to the immune response, **Figure 12. C**. From now on we will call these four TFs and their interactions the “immune response transcription network”.

4.2.3 – The expression of the TFs from both networks in the “My Data - bCSC/Bulk dataset”, “Wicha - bCSC/CC dataset”.

The expression of these nine TFs is sufficient to separate the bCSC phenotype from the bulk phenotype in the “My Data - bCSC/Bulk dataset” (**Figure 13**) and the bCSC phenotype from the cancer cell phenotype in the “Wicha - bCSC/CC dataset” (**Figure 14**). Although in the latter, four groups are formed, one group is composed from cancer cell samples with the TFs from the “mesenchymal transcription network” up-regulated and the TFs of the “immune response transcription network”, down-regulated. In another group with the opposite behavior composed mostly of bCSC samples, one group composed of two cancer cell samples down-regulated for the TFs of the “immune response transcription network” only and the last composed of bCSC samples down-regulated by the expression of TFs from both networks. The expression of the “immune response transcription network” seems to be sufficient to separate the bCSC from the cancer cells, with the cancer cells having, but not the bCSC.

Figure 11: Mesenchymal Transcription Network : **A** – TF network inferred by the ARACNE algorithm in the “pathological response” dataset and validated by MRA-FET in the bCS/CC dataset. **B** – Venn Diagram depicting the intersection between all the genes validated in regulons of BNC2, SNAI2, PRRX1 and TBX5. **C** – Venn Diagram depicting the intersection between all the genes validated in regulons of BNC2, PRRX1, TBX5 and TWIST1. **D** – GSEA analysis of curated gene sets of the 40 genes in common between PRRX1, BNC2, TBX5 and TWIST1.

Figure 12: Immune Response Transcription Network : **A** – TF network inferred by the ARACNE algorithm in the “pathological response” dataset and validated by MRA-FET in the bCS/CC dataset. **B** – Venn Diagram depicting the intersection between all the genes validated in regulons of ZNF831, SCML4, SP140 and IKZF3. **C** – GSEA analysis of curated gene sets of the 169 genes in common between ZNF831, SCML4, SP140 and IKZF3.

Figure 13: Hierarchical Clustering – “My Data - bCSC/Bulk” dataset: TFs of the networks. Based on the list of 9 transcription factors validated by MRA-FET. bCSC = breast cancer stem cell (ALDH+/ESA+/LIN- population). Bulk = bulk of the tumor

Figure 14: : Hierarchical Clustering – “Wicha - bCSC/CC” dataset: TFs of the networks. Based on the list of 9 transcription factors validated by MRA-FET. A – bCSC/Bulk dataset. B – bCSC/CC dataset. bCSC = breast cancer stem cell (ALDH+/ESA+/LIN- population). Bulk = bulk of the tumor. CC = cancer cell, REP = Biological Replicate.

4.2.4 Coordinated expression of TFs into two networks is correlated to the complete disappearance of the tumor under treatment, to epithelial to mesenchymal transition and to the immune response.

Until now we used the clinical response dataset only to infer the regulons of the transcription factors (**Figure 10**). But once MRA-FET was performed (**Table 4, Figures 11 and 12**) and the transcription networks of the bCSC inferred, we started to wonder what would be the behavior of the expression of TFs genes in cancer tissues.

We ordered the samples as described in section **3.4**, using a metagene taking the expression of all the nine TFs, positively for the “mesenchymal transcription network” and negatively for the “immune response network”. When we ordered the samples in this way (**Figure 15**) we can see that the expression of the two networks are, in the great majority of samples, mutually exclusive. When we divided the samples into two groups (58 vs 57 samples, **table 9**), there is a ratio of 3.57 of pCR in the group with a lower expression of TFs from the “mesenchymal transcription network” ($p=0.0008$).

We find that a high score in the metagene correlates with angiogenesis, Stem cell and mesenchymal characteristics. We used the metagene score as label of phenotype to define the Pearson correlation between the transcriptome of each sample and the signatures in the GSEA website databank (**Figure 16**). A low score correlates to adaptive immune response (**Figure 17**). We want to emphasize that these data are coming from the analyzes of tissue samples.

4.3 Test of the hypothesis: The expression of the transcription factors in different datasets.

4.3.1 The TCGA-BRCA dataset: TF network behavior and its biological meaning is independent of the molecular subtype and of the platform used.

Using the samples from females with invasive ductal carcinoma of the TCGA-BRCA dataset with expression data acquired by RNAseq (**Figure 18, 21 and 22**) we can see exactly the same pattern presented by the “clinical response” dataset (**Figure 15-17**). The TCGA-BRCA is a

dataset completely independent from the “pathological complete response” dataset, in samples and in technology. The first one uses RNAseq and the second Affymetrix microarray, to acquire the transcriptome data. We evaluated if the regulons inferred were being expressed in the same way in the two datasets (**Figures 19 and 20**) All the subtypes also present this pattern (**Tables 10 and 11**).

Figure 15: Coordinated expression of the TFs in the two validated networks are correlated to the complete disappearance of the tumor under treatment. Heatmap of 115 breast cancer tissue samples from the “pathological response” dataset ordered by the rank of the metagene of all TFs in the two validated networks depicted and their relation with pathological complete response to paclitaxel (pCR). When divided into two groups (58 vs 57 samples), there is a ratio of 3.57 of pCR in the group expressing less cancer stem cell TFs, $p=0.0008$. MscTFs = mesenchymal transcription factors.

	pCR	nCR	MarginalRowTotals
Metagene Score High	6 (13.62) [4.26]	52 (44.38) [1.31]	58
Metagene Score Low	21 (13.38) [4.34]	36 (43.62) [1.33]	57
Marginal Column Totals	27	88	115 (GrandTotal)

Table 9: Pathological Response Group, Chi2 Analysis. When ordered by the rank of the metagene of all TFs in the two validated networks and divided into two groups there is a ratio of 3.57 of pCR in the group expressing less cancer stem cell TFs. $p=0.0008$. The Chi-square statistic is 11.2346. pCR = pathological complete response to paclitaxel, nCR = No pathological complete response to paclitaxel.

Figure 16: GSEA datasets positively correlated with high values for the metagene in the “pathological response” dataset. High expression of the “mesenchymal transcription factors Network” and low expression of the “immune response transcription network”, evaluated by GSEA in a continuous way using Pearson metrics to rank the genes. $p \leq 0.05$, $FDR \leq 0.01$.

Figure 17: GSEA datasets negatively correlated with high values for the metagene in the “pathological response” dataset. Low expression of the “mesenchymal transcription factors network” and high expression of the “immune response transcription network”, evaluated by GSEA in a continuous way using Pearson metrics to rank the genes. $p \leq 0.05$, $FDR \leq 0.01$.

With a bigger dataset, the BRCA-TCGA dataset, we were able to make the analyzes in all the molecular subtypes of the PAM50 classification (**Figure 23**). As a rule, all of the subtypes presented the same pattern, but with some differences. We have only 15 samples from the Normal-Like subtype, **Figure 23.A**, which made the analyzes more difficult but (**Figures 24 and 25**) the biological meaning of the metagene, high levels linked to stem cells property, and low levels linked to immunological response, was maintained. In Luminal A and Luminal B samples (**Figures 23.B and 23.C**) when the TFs from the “immune response transcription network” begin to start to be expressed, the TFs from the “mesenchymal transcription network” ceased to be, and vice-versa. In the Her2-enriched subtype (**Figure 23.D**) this pattern is not so clear and in the Basal subtype the link between the expression of the two networks seems weaker (**Figure 23.E**). But all the same, the biological meaning of the levels of the expression of the metagene remains the same (**Figures 24 to 27**).

4.3.2 The expression of the genes of the “immune response transcription network” is related to better survival in all PAM50 molecular subtypes

We defined a signature of 60 genes derived from the “immune response transcription network” regulons and up-regulated in the Wicha-bCSC/CC and the “clinical response” dataset (**Supplementary Table 2**). The group that has high expression of these genes shows a better survival in all the molecular types of cancer, with a greater difference in HER2 and Basal cancers (**Figure 28**).

4.3.3 The grafting of basal-Like (basL) tumor samples in xenografts

The capacity of tissue samples coming from basL breast cancers (Du Manoir *et al.*, 2014) to grow in immunosuppressed mice is positively correlated with the expression of the regulons of the TF from the “mesenchymal transcription network” and negatively correlated with the “immune response transcription network” (**Figures 29 and 30**). For the TFs when compared with the groups Take vs Not Take, by a t-test, only SNAI2, ZNF831, IKZF3 and SP140 presented p values above 0.05 (**Table 12**).

Figure 18: TCGA-BRCA dataset - coordinated expression of the TFs in the two validated networks. Heatmap of 621 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset ordered by the rank of the metagene of all TFs in the two validated networks.

Figure 19: GSEA analysis, TCGA-BRCA. Positive correlation – regulons of the mesenchymal network transcription factors. Snapshot of enrichment results of the positive correlation of GSEA analysis using the mesenchymal network transcription factor regulons in the 621 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. FDR < 0.005, nominal pvalue < 0.01.

Figure 20: GSEA analysis, TCGA-BRCA. Negative correlation – regulons of the immune response network transcription factors. Snapshot of enrichment results of the negative correlation of GSEA analysis using the immune response network transcription factor regulons in the 621 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. FDR < 0.003, nominal pvalue < 0.01.

Figure 21: GSEA analysis, TCGA-BRCA. Positive correlation. Snapshot of enrichment results of the positive correlation of GSEA analysis using the Hallmarks (H5) signatures in the 621 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. FDR < 0.05, nominal pvalue < 0.01.

Figure 22: GSEA analysis, TCGA-BRCA. Negative correlation. Snapshot of enrichment results of the positive correlation of GSEA analysis using the KEGG signatures in the 707 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. FDR < 0.05, nominal pvalue < 0.01.

Figure 23: PAM50. TCGA-BRCA dataset - Coordinated expression of the TFs in the two validated networks . Heatmap of invasive ductal breast cancer tissue from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. **A:** 15 Normal-Like samples, **B:** 189 Luminal A Samples, **C:** 162 Luminal B Samples, **D:** 107 Her2-enriched Samples, **E:** 148 Basal Samples.

	Normal-Like		Luminal A		Luminal B		HER2-enriched		Basal	
NAME	p-value	FDR	p-value	FDR	p-value	FDR	p-value	FDR	p-value	FDR
BNC2_REGULON	0.015	0.088	<0.001	<0.001	<0.001	0.001	<0.001	0.004	0.010	0.025
PRRX1_REGULON	0.039	0.076	<0.001	<0.001	<0.001	0.001	<0.001	0.005	0.051	0.069
SNAI2_REGULON	0.011	0.062	<0.001	0.002	<0.001	0.001	<0.001	0.003	0.002	0.005
TBX5_REGULON	0.066	0.078	<0.001	<0.001	<0.001	0.001	<0.001	0.003	0.145	0.156
TWIST1_REGULON	0.110	0.129	0.002	0.013	<0.001	0.005	0.012	0.013	0.434	0.457

Table 10: GSEA analysis, TCGA-BRCA. Positive correlation – regulons of the mesenchymal network transcription factors. Enrichment results of the positive correlation of GSEA analysis using the mesenchymal network transcription factor regulons in the 621 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks and divided by the PAM50 classification.

	Normal-Like		Luminal A		Luminal B		HER2-enriched		Basal	
NAME	p-value	FDR	p-value	FDR	p-value	FDR	p-value	FDR	p-value	FDR
IKZF3_REGULON	0.017	0.044	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SCML4_REGULON	0.015	0.022	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SP140_REGULON	0.029	0.025	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
ZNF831_REGULON	0.023	0.019	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 11: GSEA analysis, TCGA-BRCA. Negative correlation – regulons of the immune response network transcription factors. Enrichment results of the negative correlation of GSEA analysis using the mesenchymal network transcription factors regulons in the 621 invasive ductal breast cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks and divided by the PAM50 classification.

Figure 24: Mammary stem cell signature. GSEA analysis, PAM50. TCGA-BRCA. Positive correlation. Snapshot of enrichment results of the positive correlation of GSEA analysis using the “**LIM mammary stem cell up signature**” in cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. **A: Normal-Like**, 32 Samples, $p \leq 0.1$, $FDR \leq 0.25$. **B: Luminal A**, 404 samples $FDR \leq 0.01$, nominal p value < 0.01 . **C: Luminal B**. $FDR \leq 0.005$, nominal p value < 0.001 . **D: HER2-enriched** $FDR < 0.05$, nominal p value < 0.01 , **E: Basal**: $FDR \leq 0.1$, nominal p value < 0.05 .

Figure 25: EMT signature. GSEA analysis, PAM50. TCGA-BRCA. Positive correlation. Snapshot of enrichment results of the positive correlation of GSEA analysis using EMT Hallmark signature in cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. A: Normal-Like, 32 Samples, $p < 0.05$. B: Luminal A, 404 samples $FDR \leq 0.001$, nominal pvalue < 0.001 . C: Luminal B. $FDR \leq 0.005$, nominal pvalue < 0.001 . D: HER2-enriched $FDR < 0.005$, nominal pvalue < 0.001 , E: Basal: $FDR < 0.01$, nominal pvalue < 0.01 .

Figure 26: Natural killer cell mediated cytotoxicity signature. GSEA analysis, PAM50. TCGA-BRCA. Negative correlation. Snapshot of enrichment results of the positive correlation of GSEA analysis using NK cell mediated cytotoxicity signature in cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. **A: Normal-Like**, 32 Samples, $p < 0.05$. **B: Luminal A**, 404 samples $FDR \leq 0.01$, nominal p value < 0.001 . **C: Luminal B**, $FDR \leq 0.005$, nominal p value < 0.001 . **D: HER2-enriched** $FDR < 0.001$, nominal p value < 0.001 , **E: Basal**: $FDR < 0.001$, nominal p value < 0.01 .

Figure 27: EMT signature. GSEA analysis, PAM50. TCGA-BRCA. Positive correlation. Snapshot of enrichment results of the positive correlation of GSEA analysis using **EMT** Hallmark signature in cancer tissue samples from females of the TCGA-BRCA dataset, ordered by the rank of the metagene of all TFs in the two validated networks. **A: Normal-Like**, 32 Samples, FDR ≤ 0.1 , $p < 0.01$. **B: Luminal A**, 404 samples FDR < 0.001 , nominal pvalue < 0.001 . **C: Luminal B**, FDR ≤ 0.005 , nominal pvalue < 0.001 . **D: HER2-enriched** FDR < 0.001 , nominal pvalue < 0.001 . **E: Basal**: FDR < 0.01 , nominal pvalue < 0.001 .

Figure 28: Positive correlation between the expression of the 60 gene differentiation signature and survival in breast cancer. Transcriptomic data from 3554 tissue samples of breast invasive carcinoma from the Cancer Genome Atlas was analyzed comparing the expression of 60 genes derived from the differentiation network regulons and up-regulated in the Wicha-bCSC/CC and the clinical response dataset. The PAM 50 classification is used.

Figure 29: BasL xenografts. Graphics of the enrichment results of the positive correlation of the basL samples in the xenograft dataset of the expression of mesenchymal transcription factor network regulons in GSEA analysis. GSEA method using the TF regulons applied in the Xenograft dataset, comparing the “Taken” samples with the “Not-Taken”. FDR <0.001 nominal pvalue < 0.001 %).

Figure 30: BasL xenografts. Graphics of the enrichment results of the negative correlation of the basL samples in the Xenograft dataset of the expression of immune response transcription factor network regulons in GSEA analysis. GSEA method using the TF regulons applied in the Xenograft dataset, comparing the “Taken” samples with the “Not-Taken”. FDR <0.001 nominal pvalue < 0.001 %).

BasL - Take vs Not Take		
SYMBOL	FoldChange	p_value
SNAI2	-2.54	≤ 0.001
ZNF831	1.68	≤ 0.001
IKZF3	2.94	0.005
SP140	4.10	0.001

Table 12: BasL Xenografts, Take vs Not Take – fold change and p value from T-test for the TFs of both networks. Only TFs with value below 0.05 are depicted.

All the lines in the experimental conditions analyzed by GSEA analyzes showed a positive correlation of the mesenchymal transcription factor regulons in all the “stem” phenotypes (**Table 14** and **15**).

This is true for CD44+/CD24- cells in MCF10A (**Table 14**) and for the up-regulation of TWIST1, TGF- β and Gsc (**Table 15**) transcription factors known to drive the epithelial to mesenchymal transition (Taube *et al.*, 2010). TWIST1 is part of the mesenchymal network and its up-regulation is up-regulating the regulons of PRRX1 and BNC2 as predicted.

4.4 - Potential membrane protein markers for breast cancer stem cells

We defined 10 new potential membrane protein markers for bCSC. First, we took all the genes identified as relevant in the regulons of PRRX1, BNC2 and TBX5 to characterize the difference between the bCSC cells and the cancer cells in the “Wicha - bCSC/CC” dataset by GSEA analyzes. Second, we filtered only the ones identified as “integral component of plasma membrane” by Gene Ontology classification (Ashburner *et al.*, 2000). We then filtered again only the ones with $p \leq 0.05$ in the paired t-test in “Wicha - bCSC/CC” dataset as described in item 3.4, we ranked the genes by their fold change. For the fourth filter, we selected in the Protein Atlas (Pontén *et al.*, 2008; Uhlén *et al.*, 2015), only the genes whose proteins have a small expression in health and in tumor mammary tissue. These four filters were the ones which decided the inclusion of the gene as a potential marker, the next two were, in order of importance for ranking, the fifth filter was the p value of the gene in the paired t-test in the “My Data - bCSC/Bulk dataset” and the sixth the p value of the gene in the t-test in the “clinical response dataset” comparing the pCR and nCR groups. We ended with 10 new possibilities: ROR1, CDH11, CS248, IL1R1, DDR2, AXL, CD109, PCDH7, CORIN and JAM3 (**Table 12**).

4.5 ZEB1 as a possible master regulator of the two transcription networks.

The expression of the two networks in all of the situations analyzed shows an inverse behavior - when one is activated the other is not - the question of the existence of a master regulator of the two networks appeared. Analyzing the promoter regions of SP140, SCML4 and

IKZF3 we encounter at least two binding sites for ZEB1 in each, ZNF831 has nine. The regulon of ZEB1 is considered a master regulator of the bCSC phenotype, in the “Wicha - bCSC/CC” dataset ($p=6,58 \times 10^{-8}$). This gave us the possibility that ZEB1 as the master regulator of the two networks. The Regulon of Zeb1 links three previously unconnected genes also selected as master regulators of the bCSC phenotype in Item 4.2 – ID4, HOXA5 and TEAD1 – and regulate the expression of ALDH1, whose activity was the main parameter of bCSC classification in this work. But unfortunately, the gene of ZEB1 itself does not have statistically significant shift in “My Data - bCSC/Bulk”, “Wicha - bCSC/CC” and in the “clinical response” datasets. a greater evaluation of ZEB1 is necessary in this process.

	Filters:	First	Second	Third	Fourth	Fifth	Sixth
Rank	Gene	GSEA_pos_bCSC/CC	Membrane Protein	$p \leq 0.05$ _bCSC/CC	ProteinAtlasLowNormalLowTumor	$p \leq 0.05$ _bCSC/Bulk	GSEA_neg_pCRdataset
1	ROR1	yes	yes	yes	yes	yes	yes
2	CDH11	yes	yes	yes	yes	yes	no
3	CD248	yes	yes	yes	yes	yes	no
4	IL1R1	yes	yes	yes	maybe	yes	no
5	DDR2	yes	yes	yes	yes	no	yes
6	AXL	yes	yes	yes	yes	no	yes
8	CD109	yes	yes	yes	yes	no	yes
7	PCDH7	yes	yes	yes	maybe	no	yes
9	CORIN	yes	yes	yes	maybe	no	no
10	JAM3	yes	yes	no	yes	no	yes

Table 13: bCSC Membrane protein marker candidates. A – Filters used in the selection.
GSEA_pos_bCSC/CC: gene contained in the leading edge of the mesenchymal network positively correlated with the bCSC phenotype. **Membrane Protein:** The gene encodes a membrane protein. **$p \leq 0.05$ _bCSC/CC:** gene expression with a $p \leq 0.05$ in a paired t-test in the bCSC/CC dataset. **ProteinAtlasLowNormalLowTumor:** The protein has a low expression in normal and cancer breast tissues. **$p \leq 0.05$ _bCSC/Bulk:** gene expression with a $p \leq 0.05$ in a paired t-test in the bCSC/Bulk dataset. **GSEA_neg_pCRdataset:** gene contained in the leading edge of mesenchymal network negatively correlated with the non complete pathological response phenotype in the pCR dataset.

MCF10A - CD44+/CD24- vs CD44-/CD24+		
NAME	p-value	FDR
PRRX1_REGULON	<0.001	0.023
BNC2_REGULON	<0.001	0.020
TBX5_REGULON	<0.001	0.017
SNAI2_REGULON	<0.001	0.015
TWIST1_REGULON	<0.001	0.032

Table 14: GSEA analysis. MRA-FET TFs of the mesenchymal transcription network in CD44+/CD24- in MCF10A cell line. MRA-FET results in the GSE15192 dataset comparing the CD44+/CD24- and CD44-/CD24+ phenotype.

GSE24202	TWIST1 vs Control		TGFb vs Control		GSC vs Control	
NAME	p-value	FDR	p-value	FDR	p-value	FDR
BNC2_REGULON	<0.001	0.182	<0.001	0.185	<0.001	0.271
PRRX1_REGULON	<0.001	0.173	<0.001	0.167	<0.001	0.242
SNAI2_REGULON	<0.001	0.171	<0.001	0.186	<0.001	0.241
TBX5_REGULON	<0.001	0.189	<0.001	0.158	<0.001	0.296
TWIST1_REGULON	<0.001	0.213	<0.001	0.129	<0.001	0.241

Table 15: GSEA analysis. MRA-FET TFs of the mesenchymal transcription network in GSE24202 perturbation dataset of HMLE cell line. Positive correlation of the expression of mesenchymal transcription factor network regulons in the perturbation dataset.

GSE24202	TWIST1 vs Control		TGFb vs Control		GSC vs Control	
NAME	p-value	FDR	p-value	FDR	p-value	FDR
IKZF3_REGULON	<0.001	0.136	0.097	0.206	<0.001	0.065
SCML4_REGULON	<0.001	0.261	0.098	0.216	<0.001	0.060
SP140_REGULON	<0.001	0.150	0.204	0.237	<0.001	0.063
ZNF831_REGULON	<0.001	0.151	0.097	0.185	<0.001	0.070

Table 16: GSEA analysis. MRA-FET TFs of the immune response transcription network in GSE24202 perturbation dataset of HMLE cell line. Negative correlation of the expression of immune response transcription factor network regulons in the perturbation dataset.

Figure 31: ZEB1 as possible master regulator of both networks of the bCSC phenotype in bCSC/CC dataset. ZEB1 regulon is classified as master regulator of the bCSC phenotype by MRA-FET in bCSC/CC dataset. In the promoter analysis, all the four TFs have at least two possible binding sites for ZEB1 in their promoters, ZNF831 have nine possible binding sites.

4.6 Transcription factor networks in the breast cancer stem cell phenotype : Summary of the Results.

Using the data of our three paired samples of bCSC and total tissues (**Table 2**) and of the GSE52327 (**Table 3**) respectively defined in this work as the “My Data – bCSC/Bulk” and the “Wicha – bCSC/CC” datasets; combined with the simple idea of intersection (**Figure 8**) we were able to identify 17 TFs that are specifically differentiated expressed in the bCSC (**Figure 9**), 13 of the up-regulated in bCSC in relation to cancer cells and the bulk of the tumor (PRRX1, SNAI2, TWIST1, ID4, BNC2, GATA6, ZNF503, FOXF2, HOXA5, HOXAB1, TSC22D1 and CREBL1) and four of them down-regulated in the same situation (SCML4, ZNF831, IKZF3 and SP140). The Pearson correlation coefficient of the expression of each one in the two datasets is 0.84, with a $p < 0.001$, **Figure 9.B**, which means that the pattern of the expression of these TFs in the two datasets is strongly correlated.

Considering the way TFs work in the cell (Alon, 2006), the next logical step was to infer which genes are regulated by each TFs – the TF regulons – in breast cancer. We did this by applying the ARACNE algorithm to the clinical response dataset (Margolin, Wang, *et al.*, 2006; Miyake *et al.*, 2012). This dataset has 115 samples, and was used because for ARACNE inference, at least 100 samples are necessary (Margolin and Califano, 2007). **Figure 10**. It is interesting to note that the inferred network is divided into two separated highly connected blocks, which agree with the expression of the TFs in the bCSC/Bulk and bCSC/CC datasets, **Figure 9**.

We then evaluated by master regulator analysis – Fisher’s exact test (MRA-FET) (Carro *et al.*, 2010) which one of these regulons could be considered a master regulator of the bCSC phenotype, comparing cancer cells, in the “Wicha - bCSC/CC” dataset (**Table 8**). MRA-FET gave us 12 master regulators, eight positively correlated with the bCSC phenotype (BNC2, PRRX1, TBX5, SNAI2, TWIST1, ID4, HOXA5, TEAD1), and maintained all the four negative correlated (IKZF3, SCML4, SP140, ZNF831).

In the 12 genes list, five (BNC2, PRRX1, TBX5, SNAI2, TWIST1) formed a well-connected and logical circuit network. BNC2, PRRX1 and TBX5 form three pairs of double-positive feedback loop (**Figure 4**, **Figure 11.A**), which means that the activation of one of them is possibly sufficient to lock the high expression of the three without the necessity of any other stimuli.

TWIST1, SNAI2, PRRX1 and BNC2 form a motif called Bifan (**Figure 4, Figure 11.A**), an example of a simple overlapping regulation pattern, when two genes (TWIST1 and SNAI2) regulate the expression of another two (BNC2 and PRRX1). The increase in complexity of this network is that SNAI2 is at the same time an “input” and an “output”. With this first analysis it was not possible to infer the existence of logical gates (Alon, 2006) or of analogical computation (Sarpeshkar, 2014). The regulons of these five genes, validated by MRA-FET, presents a high degree of overlapping, with BNC2, PRRX1, TBX5 being the greater, SNAI2 totally overlapped with these three (**Figure 11.B**), and TWIST regulon with 67 % of overlapping (30/45), (**Figure 11.C**). BNC2, PRRX1 and TBX5 validated regulons have 40 overlapped genes. When we analyze by gene enrichment analysis (GSEA) from which signatures these genes are correlated we encountered signatures related to invasive breast cancer, mammary stem cells and epithelial-to-mesenchymal transition (**Figure 11.D**). We named these five transcription factors and their interactions the “mesenchymal transcription network”.

The list of the four negative correlated (IKZF3, SCML4, SP140, ZNF831) genes also assumes a well-connected and logical circuit format shape, **Figure 12.A**. ZNF831 and SCML4 assume a double-positive feedback loop, and the four genes themselves form a multi-input feed-forward loop (**Figure 4**). There is a large gene overlap (**Figure 12.B**) and once evaluated with GSEA we encountered that the 169 overlapped genes are related to differentiation and immune response . **Figure 12.C**. We named these four transcription factors and their interactions the “immune response transcription network”.

When we used only the expression of these nine TFs to classify the samples in the “My Data - bCSC/Bulk” and “Wicha - bCSC/CC” datasets we separated quite well the bCSC from the cancer cells and from the Bulk of the tumor, **Figure 14** and **15**. Using the level of expression of these nine TFs as a metagene, as explained in item **3.4**, to rank the samples in the “clinical response dataset”, **Figure 16**, we see that the TFs from the “mesenchymal transcription network” and the “immune response transcription network” and have an inverse, and probably coordinated, pattern of expression, which was expected considering the way the regulons are organized, **Figure 10**. As the expression values are reversed, in the direction of lower expression of the mesenchymal TFs and higher expression of the others, there is an increase in the pathological complete response to chemotherapy (**Table 9**).

We evaluated the biological characteristics linked to each side of the spectrum of the metagene expression (**Figures 16 and 17**). In the “clinical response dataset”, a dataset formed with transcriptional data of breast tumor tissue, we see that a higher expression of the SNAI2, TWIST1, BNC2, PRRX1 and TBX5 means an increase in the expression of genes related to stem cells, EMT and aggressiveness, (**Figure 17**). A lower expression of the metagene, which means an increase in the expression of SCML4, ZNF831, SP140 and IKZF3, means an increase in genes related to the immune response (**Figure 18**).

5. DISCUSSION

5 – DISCUSSION

5.1 The mesenchymal transcription network

The better known network responsible for the epithelial-to-mesenchymal transition is built around the TF families of SNAIL, ZEB and TWIST (De Craene and Berx, 2013), so it is not surprising that we found SNAI2 and TWIST1 as the principal inputs of our “mesenchymal transcription network” (**Figure 11**) and ZEB1 as a possible coordinator of both “immune response” and mesenchymal networks (**Figure 31**) in bCSC.

SNAI2 and TWIST1 are involved in the mechanisms of EMT induced by TGF- β , a key component to EMT and stem cell maintenance (Itoh *et al.*, 2014; Ajani *et al.*, 2015), both induce a down-regulation of E-cadherin and Claudins, key proteins to cell-to-cell interaction and maintenance of an epithelial phenotype (Wang, 2010; Lamouille *et al.*, 2014). TWIST1 and PRRX1 are known to cooperate in EMT in embryos and cancer cells (Ocaña *et al.*, 2012). TBX5 protein forms a complex with YAP1 and β -Catenin stimulating cell survival and tumorigenesis (Rosenbluh *et al.*, 2012), TBX5 also forms a complex with TAZ (WWTR1 gene), critical to its activation (Murakami *et al.*, 2005). YAP1 and TAZ are the key effectors of the Hippo pathway (Kodaka and Hata, 2015) and the Hippo pathway is a key pathway to stem cells maintenance in general (Mo *et al.*, 2014) and epithelial stem cells maintenance in particular (Yin and Zhang, 2015). TAZ has an increase of 2.21 (p=0.04) in its fold change in the “My Data - bCSC/Bulk dataset” and an increase of 3.82 in the “Wicha - bCSC/CC dataset”, but with a p=0.07. This opens the possibility of a role for TBX5, YAP and TAZ cooperation in the bCSC phenotype.

Of the importance of BNC2 little is known. BNC2 is required for proper mitotic arrest, prevention of premature meiotic initiation, and meiotic progression in male mouse germ cells (Vanhoutteghem *et al.*, 2014) and its gene can produce, by alternative splicing, more than 2000 different proteins (Vanhoutteghem and Djian, 2007). In 2004, when the gene was discovered, it was declared that “The extreme conservation of the basenuclin 2 amino acid sequence across vertebrates suggests that basenuclin 2 serves an important function, presumably as a regulatory protein of DNA transcription” (Vanhoutteghem and Djian, 2004). Here, for the first time, we propose that BNC2 is one of the master regulators of the bCSC phenotype.

TEAD1, ID4 and HOXA5 were also selected as master regulator (**Table 8**) but are only linked with the “mesenchymal transcriptional network” if we assume the possibility of ZEB1 as the coordinator of the two networks (**Figure 31**) as is be discussed above.

By the time we drew the networks there was no information in the literature of the importance of ID4 for mammary stem cells and cancer, as is still the case with BNC2 and with all TFs from the “immune response transcriptional network”. Nevertheless, very recently ID4 was recognized to have a key role in these two events (Junankar *et al.*, 2015). TEAD1 is part of the Hippo pathway, as TBX5, and its expression induces metastasis (Lamar *et al.*, 2012; Mo *et al.*, 2014). Interestingly, the expression of both TEAD1 and HOXA5 are related to the developmental process and induction of apoptosis and their role in bCSC need to be better understood (Chen *et al.*, 2005; Stasinopoulos *et al.*, 2005; Landin Malt *et al.*, 2012; Xie *et al.*, 2013). In our context when we analyze in the GSEA, only the genes in the regulon of both TFs that were selected as important in the MRA-FET analysis were found, and as expected, both related to EMT (data not shown).

5.2 The immune response transcription network

Discussing this network is much more difficult than discussing the mesenchymal. First, we have little information in the literature of the way the four TFs that are part of it work; second, the effects we saw are probably a mix of events happening in the cancer cell itself and with its interaction with the stroma.

IKZF3 is by far the best-known TF of the four, the third member of the Ikaros family of transcription factors, also known as Aiolos. The Ikaros family controls cell fate decision, as in hematopoiesis, via chromatin remodeling (Rebollo and Schmitt, 2003; Kioussis, 2007; John and Ward, 2011). There are at least 16 splicing possibilities and at least one site for phosphorylation for IKZF3 (John and Ward, 2011), therefore it is difficult to make generalizations about the specific role of this gene in each cell type and/or situation. Another layer of complexity, considering this gene, is that its location is adjacent to the HER2 gene, the amplicon of HER2 in breast cancer often has its boundaries in two different regions of the IKZF3 gene (Matsenko and Kovalenko, 2013), considering the literature and what we have

found, the study of the role of IKZF3 and its different isoforms in breast cancer can constitute a line of research by itself.

SP140 is part of the “promyelocytic leukemia protein nuclear body” (PML-NB) (Bernardi *et al.*, 2008; Granito *et al.*, 2010). PML-NB is a tumor suppressor, linked with regulation of apoptosis and senescence (Bernardi *et al.*, 2008; Bourdeau *et al.*, 2009), SP140 itself is recognized as an auto antigen in primary biliary cirrhosis (Granito *et al.*, 2010). So we can link SP140 with tumor suppression and immune response in some models, but this is as far as we can go with the literature.

There is really scarce information about SCML4, Sex Comb On Midleg-Like Protein 4, in the literature. The gene is involved in neurogenesis in *N. furzeri*, a fish model (Baumgart *et al.*, 2014) and by similarity (Doron Lancet, 1996-2014), is related to the Polycomb group complexes. Polycomb proteins act in cancer development, stem cell plasticity and cell fate decision, acting in chromatin remodeling (Pasini *et al.*, 2004; Schuettengruber and Cavalli, 2009) as is the case with the Ikaros family. Polycomb proteins are also known to regulate cell proliferation, in a context dependent way (Piunti *et al.*, 2014).

Of ZNF831 the literature has almost no information, besides its sequence and the possible relation with HIV infection (Brass *et al.*, 2008), interestingly, SP140 is implicated with the innate immune response to HIV1 (Madani *et al.*, 2002), but is difficult with this sparse information to establish a relationship between the two events.

So, the literature information leaves us with one gene that controls cell fate by chromatin remodeling, IKZF3; another, SCML4, probably doing the same by the same mechanism and with the possibility to act in synergy with IKZF3; a third, SP140, that is a possible tumor suppressor and that can act, by now in a specific context, as an antigen; and the last, ZNF831, with practically unknown properties.

In the context of breast cancer, these four transcription factors are related to the immune response against the tumor (**Figure 17**). We see an increase in presentation and processing of antigens, an increase in the T cell receptor pathway and an increase in cytotoxicity mediated by natural killer cells in a way that resembles the rejection of transplanted organs. This indicates an active fighting of the immune system against a tumor in a Th1 immune response (Abbas, 2008). Aggressive tumors normally subvert the immune response to tolerance making naïve T cells differentiate to Treg cells, which induce tolerance, rather than T helper 1 cells,

which orchestrate the local adaptive immune response to the tumor (Wang, 2010; Gabrilovich *et al.*, 2012; Giraldo *et al.*, 2014; Shekarian *et al.*, 2015).

It was surprising to us to realize that the specific TFs of the differentiated cancer cell, compared with the bCSC and the stroma, were responsible to mount the immune response. There is of course still the possibility that these genes are also implicated in the differentiation process, although at this time it is a topic we cannot discuss much.

The analysis of the overall survival in patients expressing 60 of the most expressed genes of the “immune response transcriptional network” gives us some clues of what is happening.

Figure 28 shows the correlation of the expression of 60 genes from the “immune response transcriptional network” with the overall survival of patients with breast cancer (**Supplementary Table 2**). There is a good prognosis for the patients with a high expression these genes precisely in the more aggressive forms of breast cancer, Luminal B, HER2 and Basal. Which once again time emphasizes the possible therapeutical value of the regulation of IKZF3, SCML4, SP140 and ZNF831. We compared these 60 genes with two signatures based in the expression in the stroma of breast cancer, one to predict prognosis (Finak *et al.*, 2008) and the other for predicting resistance to chemotherapy (Farmer *et al.*, 2009). The stroma good prognosis signature has nine of its 33 genes in common with our 60 genes (CD2, CD247, CD3D, CD48, CD52, CD8A, GZMA, RUNX3, XCL1) and none with the poor prognosis signature (Finak *et al.*, 2008), there is also no intersection between the 60 genes and the stroma signature for resistance to chemotherapy (Farmer *et al.*, 2009). This is in accord with the data shown in **Figure 15**, which show a greater pathological complete response to chemotherapy in patients in which the tumor has a great expression of the TFs from the “immune response transcriptional network”.

The expression of membrane proteins in the cancer cells linked with activation and recruiting of naïve T cells can explain the phenomenon we are seeing. The expression of cited membrane proteins and the others also controlled by the network (CD2, CD226, CD247, CD38, CD3D, CD3E, CD3G, CD48, CD5, CD52, CD53, CD6, CD69, CD8A, CD96, CDC42SE2, CCL5, CCR2, CCR5, CXCR3, CXCR6) are probably responsible for the cancer cell capacity to orchestrate the immune response, unfortunately, until now we cannot evaluate the contribution of each one.

The “immune response transcription network” is probably the greater contributor of this work. The test of the potential in using IKZF3, SP140, ZNF831 and SCML4 in a gene therapy or

as the base of a more complex circuit made in the context of synthetic biology is the natural next step of this work.

5.3 The TCGA-BRCA dataset

The two networks and their functions fit well in our discovery datasets, but, principally for the “immune response network” we still had doubts: Will these patterns and properties be held in other breast cancer datasets? If we subset the samples into molecular subtypes, will we still see the same thing happening? To answer these we used the TCGA-BRCA dataset, with the transcriptome data of 621 invasive ductal breast tumors from female patients. Fortunately, the answer to the questions is: Yes.

We can see that the samples in the BRCA-TCGA dataset have the same inversed pattern of expression for the TFs of the two networks (**Figure 18** and **15**) the molecular subtypes are evenly distributed in the heatmap (**Figure 18**) with the exception of the Basal subtype. Basal breast cancer normally expresses *SNAI2* but does not express other TFs linked to EMT (Guo *et al.*, 2012; Condiotti *et al.*, 2014) as will be better explained below.

Once the regulons were defined in a different dataset that used a different technology - Microarray chips from Affymetrix for the “clinical response dataset” and RNAseq technology from Illumina for the “BRCA-TCGA dataset” – we needed to confirm if they showed the same pattern of expression in the BRCA-TCGA dataset. The expression of the regulons of the “mesenchymal transcription network” are related to high values of metagene and the expression of regulons of the “immune response transcription network” is related with low values of the metagene, as is the case with the “clinical response dataset” (**Figures 19** and **20**). Analyzing the samples by their molecular subtypes (**Tables 10** and **11**) gives us the same result.

We also see that the biological meaning of the networks remains the same in the dataset as a whole (**Figures 21** and **22**) and in the subtypes (**Figures 24** to **27**). The principal question that the analyzes of the “clinical response dataset” was not able to answer is that what we see (**Figures 15** to **17**) is independent from the molecular subtypes of breast cancer. The analyzes of the TCGA-BRCA dataset shows us this independence with again, the strong value of being made not only with a different dataset but also with different technology.

Of course, even if the expression of the regulons and the biological meaning holds, there are differences in the expression of the TFs in the subtypes (**Figure 23**). There are only 15

samples of the Normal-Like subtype (**Figure 23.A**) which give us some problems with the analyzes, but we can see that in the extremes there are samples expressing the TFs of one network and not from the other, samples from the Luminal A and Luminal B are the ones showing the better pattern of mutual exclusive expression of the genes of the two networks (**Figure 23.B and 23.C**), the fact of the luminal subtypes being the more differentiated than the subtypes (Sorlie *et al.*, 2003; Nielsen *et al.*, 2014; Györfy *et al.*, 2015) possibly has some influence in that. In the HER2-enriched subtype (**Figure 23.D**) there is a great presence of samples expressing the genes of the two networks at the same time, but in the extremes, the mutual exclusion continues. In the Basal subtype (**Figure 23.E**) we see a lower relationship between the two networks, there is a great correlation of the low values of the metagene with high expression of IKFZ3, SCML4, SP140 and ZNF831 but the genes from the “mesenchymal transcription network” does not seem to be well correlated with the metagene, with the exception of SNAI2, the second gene in the heatmap, its expression is inversely correlated to the four TFs of the “immune response transcription network”. SNAIL2 is not only a marker for the Basal subtype but also essential for maintenance of the basal identity (Guo *et al.*, 2012; Condiotti *et al.*, 2014) and we can infer, based on the literature and on our own data, the basal identity is inversely correlated with the expression of the “immune response network” (**Figures 23 to 25**) and that the expression of this network is correlated with a better prognosis in the Basal subtype (**Figure 28**).

The question still unanswered is whether the induced expression of the TFs of the “immune response network” in the different subtypes of breast cancer will be sufficient to give patients a better prognosis.

5.4 Patient-derived tumor xenografts (PDXs)

PDXs are being used as a model for cancer because they much better reflect the tumor heterogeneity and the environment than *in vitro* methodologies (Choi *et al.*, 2014; Du Manoir *et al.*, 2014; Cassidy *et al.*, 2015), as described by Tentler in 2012 “The approach is very straightforward, consisting of obtaining fresh surgical tissue, sectioning it into approximately 3 mm³ pieces, followed by subcutaneous or orthotopic implantation into the flank of an immunodeficient mouse or rat” (Tentler *et al.*, 2012). Not all the cancer samples which pass through this process produce PDXs, the grafting of the PDXs is correlated with the aggressiveness of the tumor (Choi *et al.*, 2014; Du Manoir *et al.*, 2014; Cassidy *et al.*, 2015).

Using the expression data from the primary tumors used by Du Manoir *et al.* to construct a repository of PDXs (Du Manoir *et al.*, 2014), data kindly given by the author, we can see that there is a difference of expression of the regulons of the TFs of the two networks in the grafting of samples from the Basal-Like subtype (**Figure 29** and **30**). The regulons from the “mesenchymal transcription network” are related to the samples that have produced PDXs and the regulons from the “immune response transcription network” are related to the samples that were unable to produce PDXs. With the exception of SCML4, the other three genes of the immune network are differentially expressed when the groups Take vs Not Take are compared (Table 12) and only SNAI2 from the mesenchymal network, which agrees with the data encountered in the TCGA-BRCA dataset, with the Kmpot dataset (**Figure 23.E** and **Figure 28**) and with the literature. The Basal-Like subtype from the Guedj classification resembles greatly the Basal subtype of the PAM50 classification (Guedj *et al.*, 2012).

As postulated by DeRose and collaborators in their seminal study, “the ability of a tumor to survive and grow in a foreign host might reflect a more aggressive phenotype that is independent of known clinical variables” (Derose *et al.*, 2011). The relationship of the tumor with its microenvironment, including the presence and active state of the hematopoietic cells is one of the factors involved in the engraftment frequency and growth rate of implanted tumors (Williams *et al.*, 2013). Dr Du Manoir and collaborators identified an IL8 signature in the grafted the BasL (Du Manoir *et al.*, 2014), showing the importance of the immune apparatus in the grafting. Here, under the supervision of Dr. Du Manoir himself, and analyzing the same samples in our context, we found the immune system playing a role also in the direct opposite effect. We found the influence of the networks only in the engraftment of the BasL samples, this could indicate a stronger influence of the immune network in this specific subtype. The Basal-like subtype, such as the Basal from PAM50, was the most aggressive subtype of breast cancer, finding a method to activate our immune network for this kind of tumor seems promising.

5.5 The networks and the cell lines

The first interesting thing to bring to attention about the cell lines and the TFs of networks came from the analyzes of the breast cancer cell lines in the “Cancer Cell Line Encyclopedia (CCLE)” (Barretina *et al.*, 2012). Analyzing the expression of the TFs in the 56 cell lines from human breast tumors we see that TFs from the immune response network have low

values for the average expression and for the standard deviation – 4.0, sd $\pm$ 0.2 for SP140; 3.6, sd $\pm$ 0.1 for ZNF831; 3.9, sd $\pm$ 0.5 for IKFZ3 and 3.8, sd $\pm$ 0.1 for SCML4 – and mixed values for the TFs of the mesenchymal network – 5.0, sd $\pm$ 2.1 for PRRX1; 4.9, sd $\pm$ 1.1 for BNC2; 6.9, sd $\pm$ 3.1 for SNAI2; 4.3, sd $\pm$ 0.5 for TBX5 and 6.4, sd $\pm$ 2.8 for TWIST1.

To continue this work it is essential to define what will be the best tools to use, and genetic alteration of breast cell lines is the easy choice, if they proved suited for the experiments. The low expression of the immune response network TFs is intriguing, and even more intriguing is the low deviation of expression in the 56 different human cell lines in the CCLE. It is not clear if the expression of these four genes is incompatible with the growth of cells *in vitro*, or if they are just unnecessary in the context. This will need to be tested if we are to use modification in the expression of the cell lines to biologically validate our findings.

The information we do have is that when we compare CD44+/CD24- cells and CD44-/CD24+, another way to compare bCSCs and cancer cells, in the MCF10A cell line (**Table 14**) we have positive correlation of the mesenchymal transcription factor regulons in the CD44+/CD24- cells.

We also have perturbation data of HMLE cell lines, when genes known to drive EMT - TWIST1, TGF β and GSC are expressed in the cells, we have positive correlation of the mesenchymal transcription factor regulons in the group of more mesenchymal characteristics (**Table 15**) and a negative correlation in this group of the expression of regulons of the immune response network.

5.6 New membrane protein candidates for bCSC markers

As explained in the introduction and depicted in **Figure 2**, there is not a definitive marker for the bCSC phenotype, the fact that it is possible to have a plethora of steps in the bCSC phenotype, gave us another layer of complexity in the problem (Shipitsin and Polyak, 2008; Schmitt *et al.*, 2012; Liu *et al.*, 2014).

As explained in section 4.4, analyzing the genes controlled by the “mesenchymal transcription network” we ended up with 10 possible new membrane protein markers for bCSC, in order of probability: ROR1, CDH11, CD248, IL1R1, DDR2, AXL, CD109, PCDH7, CORIN and JAM3 (**Table 13**).

Receptor Tyrosine Kinase-Like Orphan Receptor 1 (ROR1), is expressed in ovarian cancer stem cells and plays a crucial role in developmental morphogenesis, principally in the brain (Endo *et al.*, 2012; Zhang, S. *et al.*, 2014). ROR1 signaling is involved with survival of the cell by a ROR1/MEK/ERK signaling in cooperation with AKT, with the potential as an immunotherapeutic agent being tested in lung cancers (Karachaliou *et al.*, 2014; Shabani *et al.*, 2015). So, if the protein is also proved to be a marker for bCSC too we can test the therapeutical value of inhibiting it.

Cadherin 11 (CDH11), is expressed in invasive breast cancer cell lines, is a marker of the mesenchymal phenotype and is involved with EMT in melasma (Pishvaian *et al.*, 1999; Kaur *et al.*, 2012; Kim *et al.*, 2014). The expression of CDH11 is involved with cell motility and promotes metastasis in the bone in prostate cancer and in renal cell carcinoma (Chu *et al.*, 2008; Kaur *et al.*, 2012; Schulte *et al.*, 2013; Satcher *et al.*, 2014). The expression of this protein is related to promote angiogenesis (Park *et al.*, 2014) and is commonly viewed as a marker for bad prognosis in malignancies (Assefnia *et al.*, 2014). As is the case with ROR1, the inhibition of CDH11 by immunotherapy has a great potential.

CD248, also known as Endosialin, is a marker for mesenchymal stem cells, is expressed in cancer stem cells in human sarcoma cell lines and in human high grade sarcomas and brain tumors (Carson-Walter *et al.*, 2009; Rouleau *et al.*, 2011; Naylor *et al.*, 2012; Rouleau *et al.*, 2012). The protein is involved with angiogenesis through its interaction with PDGF, Platelet-Derived Growth Factor, and can be related to immunosuppression (Bagley *et al.*, 2008; Tomkowicz *et al.*, 2010; Ochs *et al.*, 2013).

There is not much information about the link of IL1R1, Interleukin 1 Receptor Type 1, and cancer, there is a relation between the expression of different isoforms of IL1R1 and comorbidities in patients with breast cancer (Mccann *et al.*, 2012; Merriman *et al.*, 2014), but there is no present correlation as a marker for stem cells. IL1 is a key cytokine in diverse signaling pathways, as in TLR, MAPK, NLR and NF- κ B (Kuno and Matsushima, 1994; Abbas, 2008; Acuner Ozbabacan *et al.*, 2014). The study of the role of IL1R1 in bCSC is a potential new line of research.

The expression of Discoidin Domain Receptor Tyrosine Kinase 2 (DDR2), is a marker for bad prognosis in breast cancer, principally in triple negative cancers (Ren *et al.*, 2013; Toy *et al.*, 2015). Its expression is related to EMT in breast cancer in conditions of hypoxia (Ren *et al.*, 2014) and induces metastasis in breast cancer, prostate cancer, melanoma and head and neck

squamous cell carcinoma (Zhang *et al.*, 2013; Ren *et al.*, 2014; Yan *et al.*, 2014; Poudel *et al.*, 2015).

The over expression and activation of AXL, AXL receptor tyrosine kinase, is correlated with invasiveness and poor prognosis in breast, prostate and lung cancers (Mishra *et al.*, 2012; Asiedu *et al.*, 2014; Leconet *et al.*, 2014; Wu *et al.*, 2014). There are data showing that the expression of this receptor induces EMT and regulates the function of bCSC (Asiedu *et al.*, 2014) and an antibody Anti-AXL already passed the preclinical test for its use in immunotherapy for pancreatic cancer (Leconet *et al.*, 2014). Considering the literature, this is the protein with a better and faster chance to be used in translational research of breast cancer as a marker for bCSC and as a therapeutic target.

The expression of CD109 was reported as a useful marker for the diagnosis of invasive breast and prostate carcinomas (Hasegawa *et al.*, 2007; Hasegawa *et al.*, 2008), is considered a possible target for triple-negative breast cancer (Tao *et al.*, 2014) and its high expression regulates the cancer stem cell phenotype in the epithelioid sarcoma cell line ESX (Emori *et al.*, 2013).

There is also not much information on the role of protocadherin-7 (PCDH7), in cancers. PCDH7 has a high expression in triple negative breast cancer (Tao *et al.*, 2014), is reported to induce bone metastasis in breast cancer (Li *et al.*, 2013), and its presence is necessary for the beginning of mitosis in HeLa cells (Özlü *et al.*, 2015).

There is no information in the literature regarding the role of CORIN, also known as atrial natriuretic peptide-converting enzyme, in cancer cells. In the adult, individual CORIN expression is principally related to the circulatory system and its diseases (Armaly *et al.*, 2013; Zhang, Y. *et al.*, 2014; Liu *et al.*, 2015). Nonetheless, CORIN expression promotes trophoblast invasion and uterine spiral artery remodeling in pregnancy (Cui *et al.*, 2012). Trophoblast invasion was once defined as “as a tightly regulated battle between the competing interests of the survival of the fetus and those of the mother” (Anin *et al.*, 2004) and is a finely controlled process containing the steps: adhesion and detachment from the extracellular matrix (ECM), invasion of the ECM and maternal vessels by proteolysis, proliferation and death by apoptosis differentiation, and interaction with the maternal immune system (Goldman-Wohl and Yagel, 2002; Anin *et al.*, 2004). This process very much resembles the process of carcinogenesis and the growth of a tumor in metastatic sites, and cancer is known to corrupt systems used in

embryonic development (Bruce Alberts, 2007; Wang, 2010), so the role of CORIN in bCSC also has also the potential to be a new line of research in breast cancer.

Junctional Adhesion Molecule 3 (JAM3), is a marker for neural stem cells (Stelzer *et al.*, 2012), its expression regulates metastasis in melanoma and non-small cell lung cancer (Arcangeli *et al.*, 2012; Hao *et al.*, 2014), and its soluble form induces angiogenesis (Rabquer *et al.*, 2010).

5.7 ZEB1: A possible master regulator of the two networks

Considering our results and the pattern of expression of the two networks, there is a strong possibility of the TFs of the mesenchymal and of the immune response networks being regulated by a TF hierarchically above of them. This TF needs to be both an activator and a repressor of transcription, and because of the importance of the process, EMT and cancer stem cell phenotype, would be probably well known,

Zinc Finger E-Box Binding Homeobox 1 (ZEB1), meet all of our requisites. As discussed above the ZEB, SNAIL and TWIST families have the best known genes which induce EMT and maintain a stem cell phenotype (Wang, 2010; De Craene and Berx, 2013; Lamouille *et al.*, 2014). ZEB1 can act both as an activator and/or a repressor of transcription, depending of the gene and of the context (Chaffer *et al.*, 2013). The effects of ZEB1, TWIST1 and SNAIL2 expression in EMT are being reported to occur in parallel, independently, and in cooperation (Scheel and Weinberg, 2012; De Craene and Berx, 2013; Lamouille *et al.*, 2014). TWIST1 and SNAI2 are the inputs of our “mesenchymal transcription network” (**Figure 11**) so we were not exactly surprised when we defined the regulon of ZEB1, as described in items **3.4** and **4.5**, and found ZEB1 as a master regulator of the bCSC phenotype in the “Wicha - bCSC/CC dataset” (**Figure 31**). What began to surprise us was the link with ZEB1 and TEAD1, HOXA5 and ID4, the three TFs found to be specific and a master regulator of bCSCs (**Figure 9** and **Table 8**) but were not connected to the mesenchymal network (**Figure 11**).

The use of the ARACNE algorithm has its limitations (Margolin, Nemenman, *et al.*, 2006; Margolin, Wang, *et al.*, 2006; Margolin and Califano, 2007), and we found very little information on which genes are repressed by ZEB1 and SNAI2 using ARACNE, and both genes are known to have dual roles in transcription (Chaffer *et al.*, 2013; De Craene and Berx, 2013; Lamouille *et al.*, 2014).

Analyzing the promoter region of ZNF831, SCML4, SP140 and IKZF3 we found multiple sites for ZEB1 binding in all the four genes (**Item 4.5** and **Figure 31**). This opens the possibility that they are also under the influence of ZEB1 expression.

The last surprise for us is the fact that ALDH1 is part of the regulon of ZEB1. The activity of ALDH1 is the principal parameter used in this work to identify bCSC, considering ZEB1 as the central TF for the two networks would explain all our results, including the sorting by FACS.

Nonetheless, is difficult for us to affirm it with our data because the ZEB1 gene has practically no alteration in the fold change in “Wicha - bCSC/CC” and “My Data - bCSC/Bulk” datasets, with a p value of 0.2 and 0.7, respectively. We need more information to be able to establish a direct link between ZEB1 and our two networks.

6. CONCLUSION

6 - CONCLUSION:

We have discovered two networks with inverse expression behavior in breast cancer tissues and bCSC. The First, the “mesenchymal transcription network” composed of SNAI2, TWIST, PRRX1, BNC2 and TBX5.

The second, the “immune response transcription network” composed of SCML4, ZNF831, SP140 and IKZF3, is totally unknown in the context of breast cancer in the literature and is responsible for immune response phenotype and better prognosis.

Both networks seem to be regulated by ZEB1.

This data was analyzed and confirmed using different datasets, technologies and experimental contexts.

We generated a hypothesis about master regulator TFs that can be experimentally validated to switch off the bCSC phenotype, with clear potential for clinical use, and established several possible new lines of research for breast cancer.

7. REFERENCES

7 - References:

ABBAS, A. K., ANDREW H. LICHTMAN, AND SHIV PILLAI. *Imunologia celular e molecular*. Elsevier Brasil, 2008.

ABRAHAM, B. K. et al. Prevalence of CD44+/CD24-/low cells in breast cancer may not be associated with clinical outcome but may favor distant metastasis. **Clin Cancer Res**, v. 11, n. 3, p. 1154-9, Feb 2005. ISSN 1078-0432. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/15709183> >.

ACUNER OZBABACAN, S. E. et al. The structural pathway of interleukin 1 (IL-1) initiated signaling reveals mechanisms of oncogenic mutations and SNPs in inflammation and cancer. **PLoS Comput Biol**, v. 10, n. 2, p. e1003470, Feb 2014. ISSN 1553-7358. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24550720> >.

AJANI, J. A. et al. Cancer stem cells: the promise and the potential. **Semin Oncol**, v. 42 Suppl 1, p. S3-17, Apr 2015. ISSN 1532-8708. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25839664> >.

AL-EJEH, F. et al. Breast cancer stem cells: treatment resistance and therapeutic opportunities. **Carcinogenesis**, v. 32, n. 5, p. 650-8, May 2011. ISSN 1460-2180. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21310941> >.

AL-HAJJ, M. et al. Prospective identification of tumorigenic breast cancer cells. **Proc Natl Acad Sci U S A**, v. 100, n. 7, p. 3983-8, Apr 2003. ISSN 0027-8424. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/12629218> >.

ALISON, M. R. et al. Finding cancer stem cells: are aldehyde dehydrogenases fit for purpose? **J Pathol**, v. 222, n. 4, p. 335-44, Dec 2010. ISSN 1096-9896. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20848663> >.

ALON, U. **An Introduction to Systems Biology: Design Principles of Biological Circuits**. 1st. London, UK.: CRC Press, Taylor & Francis Group, 2006.

AMAT, S. et al. High prognostic significance of residual disease after neoadjuvant chemotherapy: a retrospective study in 710 patients with operable breast cancer. **Breast Cancer Res Treat**, v. 94, n. 3, p. 255-63, Dec 2005. ISSN 0167-6806. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16267618> >.

ANIN, S. A.; VINCE, G.; QUENBY, S. Trophoblast invasion. **Hum Fertil (Camb)**, v. 7, n. 3, p. 169-74, Sep 2004. ISSN 1464-7273. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/15590570> >.

ARCANGELI, M. L. et al. The Junctional Adhesion Molecule-B regulates JAM-C-dependent melanoma cell metastasis. **FEBS Lett**, v. 586, n. 22, p. 4046-51, Nov 2012. ISSN 1873-3468. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23068611> >.

ARMALY, Z.; ASSADY, S.; ABASSI, Z. Corin: a new player in the regulation of salt-water balance and blood pressure. **Curr Opin Nephrol Hypertens**, v. 22, n. 6, p. 713-22, Nov 2013. ISSN 1473-6543. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24100222> >.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nat Genet**, v. 25, n. 1, p. 25-9, May 2000. ISSN 1061-4036. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/10802651> >.

ASIEDU, M. K. et al. AXL induces epithelial-to-mesenchymal transition and regulates the function of breast cancer stem cells. **Oncogene**, v. 33, n. 10, p. 1316-24, Mar 2014. ISSN 1476-5594. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23474758> >.

ASSEFNIA, S. et al. Cadherin-11 in poor prognosis malignancies and rheumatoid arthritis: common target, common therapies. **Oncotarget**, v. 5, n. 6, p. 1458-74, Mar 2014. ISSN 1949-2553. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24681547> >.

BAGLEY, R. G. et al. Endosialin/TEM 1/CD248 is a pericyte marker of embryonic and tumor neovascularization. **Microvasc Res**, v. 76, n. 3, p. 180-8, Nov 2008. ISSN 1095-9319. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18761022> >.

BARRETINA, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. **Nature**, v. 483, n. 7391, p. 603-7, Mar 2012. ISSN 1476-4687. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22460905> >.

BAUMGART, M. et al. RNA-seq of the aging brain in the short-lived fish *N. furzeri* - conserved pathways and novel genes associated with neurogenesis. **Aging Cell**, v. 13, n. 6, p. 965-74, Dec 2014. ISSN 1474-9726. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25059688> >.

BENFEY, P. N. Taking a developmental perspective on systems biology. **Dev Cell**, v. 21, n. 1, p. 27-8, Jul 2011. ISSN 1878-1551. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21763604> >.

BERIWAL, S. et al. Breast-conserving therapy after neoadjuvant chemotherapy: long-term results. **Breast J**, v. 12, n. 2, p. 159-64, 2006 Mar-Apr 2006. ISSN 1075-122X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16509842> >.

BERNARDI, R.; PAPA, A.; PANDOLFI, P. P. Regulation of apoptosis by PML and the PML-NBs. **Oncogene**, v. 27, n. 48, p. 6299-312, Oct 2008. ISSN 1476-5594. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18931695> >.

BHAT-NAKSHATRI, P. et al. SLUG/SNAI2 and tumor necrosis factor generate breast cells with CD44+/CD24- phenotype. **BMC Cancer**, v. 10, p. 411, 2010. ISSN 1471-2407. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20691079> >.

BOURDEAU, V.; BAUDRY, D.; FERBEYRE, G. PML links aberrant cytokine signaling and oncogenic stress to cellular senescence. **Front Biosci (Landmark Ed)**, v. 14, p. 475-85, 2009. ISSN 1093-4715. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19273079> >.

BRASS, A. L. et al. Identification of host proteins required for HIV infection through a functional genomic screen. **Science**, v. 319, n. 5865, p. 921-6, Feb 2008. ISSN 1095-9203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18187620> >.

BRUCE ALBERTS, A. J., JULIAN LEWIS, MARTIN RAFF, KEITH ROBERTS, PETER WALTER. **Molecular Biology of the Cell**. 5th. Garland Science, 2007.

CAREY, L. A. Through a glass darkly: advances in understanding breast cancer biology, 2000-2010. **Clin Breast Cancer**, v. 10, n. 3, p. 188-95, Jun 2010. ISSN 1938-0666. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20497917> >.

CARRO, M. S. et al. The transcriptional network for mesenchymal transformation of brain tumours. **Nature**, v. 463, n. 7279, p. 318-25, Jan 2010. ISSN 1476-4687. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20032975> >.

CARSON-WALTER, E. B. et al. Characterization of TEM1/endosialin in human and murine brain tumors. **BMC Cancer**, v. 9, p. 417, 2009. ISSN 1471-2407. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19948061> >.

CARVALHO, B. S.; IRIZARRY, R. A. A framework for oligonucleotide microarray preprocessing. **Bioinformatics**, v. 26, n. 19, p. 2363-7, Oct 2010. ISSN 1367-4811. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20688976> >.

CASSIDY, J. W.; CALDAS, C.; BRUNA, A. Maintaining Tumor Heterogeneity in Patient-Derived Tumor Xenografts. **Cancer Res**, Jul 2015. ISSN 1538-7445. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/26180079> >.

CHAFFER, C. L. et al. Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. **Cell**, v. 154, n. 1, p. 61-74, Jul 2013. ISSN 1097-4172. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23827675> >.

CHAN, S. S.; KYBA, M. What is a Master Regulator? **J Stem Cell Res Ther**, v. 3, May 2013. ISSN 2157-7633. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23885309> >.

CHARAFE-JAUFFRET, E. et al. Aldehyde dehydrogenase 1-positive cancer stem cells mediate metastasis and poor clinical outcome in inflammatory breast cancer. **Clin Cancer Res**, v. 16, n. 1, p. 45-55, Jan 2010. ISSN 1078-0432. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20028757> >.

CHEN, H. et al. Identification of transcriptional targets of HOXA5. **J Biol Chem**, v. 280, n. 19, p. 19373-80, May 2005. ISSN 0021-9258. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/15757903> >.

CHOI, S. Y. et al. Lessons from patient-derived xenografts for better in vitro modeling of human cancer. **Adv Drug Deliv Rev**, v. 79-80, p. 222-37, Dec 2014. ISSN 1872-8294. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25305336> >.

CHU, K. et al. Cadherin-11 promotes the metastasis of prostate cancer cells to bone. **Mol Cancer Res**, v. 6, n. 8, p. 1259-67, Aug 2008. ISSN 1541-7786. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18708358> >.

CLARKE MF, D. J., DIRKS PB, EAVES CJ, JAMIESON CH, JONES DL, VISVADER J, WEISSMAN IL, WAHL GM. Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells **Cancer Research**, 2006.

CONDIOTTI, R.; GUO, W.; BEN-PORATH, I. Evolving views of breast cancer stem cells and their differentiation States. **Crit Rev Oncog**, v. 19, n. 5, p. 337-48, 2014. ISSN 0893-9675. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25404149> >.

CUI, Y. et al. Role of corin in trophoblast invasion and uterine spiral artery remodelling in pregnancy. **Nature**, v. 484, n. 7393, p. 246-50, Apr 2012. ISSN 1476-4687. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22437503> >.

DE BEÇA, F. F. et al. Cancer stem cells markers CD44, CD24 and ALDH1 in breast cancer special histological types. **J Clin Pathol**, v. 66, n. 3, p. 187-91, Mar 2013. ISSN 1472-4146. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23112116> >.

DE CRAENE, B.; BERX, G. Regulatory networks defining EMT during cancer initiation and progression. **Nat Rev Cancer**, v. 13, n. 2, p. 97-110, Feb 2013. ISSN 1474-1768. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23344542> >.

DEROSE, Y. S. et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. **Nat Med**, v. 17, n. 11, p. 1514-20, 2011. ISSN 1546-170X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22019887> >.

DEVLIN, K. *The joy of sets: fundamentals of contemporary set theory*. . Springer Science & Business Media, 2012

DORON LANCET, M. S., NAOMI ROSEN, GIL STELZER, NOAM NATIV, TSIPPI INY-STEIN, FRIDA BELINKY, IRIS BAHIR, NOA RAPPAPORT, MICHAL TWIK, SHAHAR ZIMMERMAN, SIMON FISHILEVICH, ASHER KOHN, YAIR YEHUDA, MICHAEL REBHAN, AVITAL ADATO, JUSTIN ALEXANDER, DANIELLA BAR, HILA BENJAMIN, EMILY BREWSTER, YANA BROMBERG, URI BENDOR, ASAF CARMİ, VERED CHALIFA-CASPI, GERSHON CELNIKER, IRINA DALAH, TIRZA DONIGER, NIR ESTERMAN, ZIV FRANKENSTEIN, MARIA FRONCZAK, OFIR GOLDBERGER, OHAD GREENSHPAN, ILYA GRINBLATT, SHIRA GROSSMAN, ARIK HAREL, SHIRLEY HORN-SABAN, PAVEL KATS, MIKHAIL KORMAN, HAGIT KRUG, MICHAL LAPIDOT, SHLOMI LEVITAN, ASAF LEVY, GUY LEVY, RUSSELL LEVY, ASAF MADI, ELENA MATUSEVICH, KARIN NOY, TSVIYA OLENDER, RON OPHIR, YAKOV PERLMANN, INGA PETER, JAIME PRILUSKY, SHANY RON, MICHAL RONEN, REVITAL ROSENBERG, HERSHEL SAFER, YIGEAL SATANOWER, ANDREAS SCHNEIDER, TALİ SEFTİ, SHAI SHEN-ORR, MAXİM SHKLAR, MICHAEL SHMOISH, ORIT SHMUELI, ALEXANDRA SIROTA MADI, YARON SOLE, JULIE STAMPNITZKY, LIORA STRICHMAN-ALMASHANU, LIORA YAAR, ITAI YANAI, IDO ZAK, ZIV ZEIRA. GeneCards - The Human Gene Compendium. 1996-2014. Available at: < <http://www.genecards.org/> >. Accessed on: 25/07.

DU MANOIR, S. et al. Breast tumor PDXs are genetically plastic and correspond to a subset of aggressive cancers prone to relapse. **Mol Oncol**, v. 8, n. 2, p. 431-43, Mar 2014. ISSN 1878-0261. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24394560> >.

DU MANOIR, S. O., B. BRAS-GONÇALVES, R. NGUYEN, T. LASORSA, L. BOISSIÈRE, F. MASSEMIN, B. COLOMBO, P. BIBEAU, F. JACOT, W. THEILLET, C. Breast tumor PDXs are genetically plastic and correspond to a subset of aggressive cancers prone to relapse. **(article in press)**, 2013.

EL-SAMAD, H.; MADHANI, H. D. Can a systems perspective help us appreciate the biological meaning of small effects? **Dev Cell**, v. 21, n. 1, p. 11-3, Jul 2011. ISSN 1878-1551. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21763599> >.

EMORI, M. et al. High expression of CD109 antigen regulates the phenotype of cancer stem-like cells/cancer-initiating cells in the novel epithelioid sarcoma cell line ESX and is related to poor prognosis of soft tissue sarcoma. **PLoS One**, v. 8, n. 12, p. e84187, 2013. ISSN 1932-6203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24376795> >.

ENDO, M. et al. Ror family receptor tyrosine kinases regulate the maintenance of neural progenitor cells in the developing neocortex. **J Cell Sci**, v. 125, n. Pt 8, p. 2017-29, Apr 2012. ISSN 1477-9137. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22328498> >.

FARMER, P. et al. A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. **Nat Med**, v. 15, n. 1, p. 68-74, Jan 2009. ISSN 1546-170X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19122658> >.

FERGUSON, L. R. et al. Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. **Semin Cancer Biol**, Apr 2015. ISSN 1096-3650. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25869442> >.

FINAK, G. et al. Stromal gene expression predicts clinical outcome in breast cancer. **Nat Med**, v. 14, n. 5, p. 518-27, May 2008. ISSN 1546-170X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18438415> >.

FISHER, B. et al. Effect of preoperative chemotherapy on local-regional disease in women with operable breast cancer: findings from National Surgical Adjuvant Breast and Bowel Project B-18. **J Clin Oncol**, v. 15, n. 7, p. 2483-93, Jul 1997. ISSN 0732-183X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/9215816> >.

_____. Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. **J Clin Oncol**, v. 16, n. 8, p. 2672-85, Aug 1998. ISSN 0732-183X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/9704717> >.

GABRILOVICH, D. I.; OSTRAND-ROSENBERG, S.; BRONTE, V. Coordinated regulation of myeloid cells by tumours. **Nat Rev Immunol**, v. 12, n. 4, p. 253-68, Apr 2012. ISSN 1474-1741. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22437938> >.

GANGOPADHYAY, S. et al. Breast cancer stem cells: a novel therapeutic target. **Clin Breast Cancer**, v. 13, n. 1, p. 7-15, Feb 2013. ISSN 1938-0666. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23127340> >.

GAUTIER, L. et al. affy-analysis of Affymetrix GeneChip data at the probe level. **Bioinformatics**, v. 20, n. 3, p. 307-15, Feb 2004. ISSN 1367-4803. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/14960456> >.

GINESTIER, C. et al. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. **Cell Stem Cell**, v. 1, n. 5, p. 555-67, Nov 2007. ISSN 1934-5909. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18371393> >.

GIRALDO, N. A. et al. The immune contexture of primary and metastatic human tumours. **Curr Opin Immunol**, v. 27, p. 8-15, Apr 2014. ISSN 1879-0372. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24487185> >.

GOLDMAN-WOHL, D.; YAGEL, S. Regulation of trophoblast invasion: from normal implantation to pre-eclampsia. **Mol Cell Endocrinol**, v. 187, n. 1-2, p. 233-8, Feb 2002. ISSN 0303-7207. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/11988332> >.

GOTTSCHLING, S. et al. Are we missing the target? Cancer stem cells and drug resistance in non-small cell lung cancer. **Cancer Genomics Proteomics**, v. 9, n. 5, p. 275-86, 2012 Sep-Oct 2012. ISSN 1790-6245. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22990107> >.

GRANIT, R. Z.; SLYPER, M.; BEN-PORATH, I. Axes of differentiation in breast cancer: untangling stemness, lineage identity, and the epithelial to mesenchymal transition. **Wiley Interdiscip Rev**

Syst Biol Med, v. 6, n. 1, p. 93-106, 2014 Jan-Feb 2014. ISSN 1939-005X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24741710> >.

GRANITO, A. et al. PML nuclear body component Sp140 is a novel autoantigen in primary biliary cirrhosis. **Am J Gastroenterol**, v. 105, n. 1, p. 125-31, Jan 2010. ISSN 1572-0241. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19861957> >.

GREGORY R. WARNES, B. B., LODEWIJK BONEBAKKER, ROBERT GENTLEMAN, WOLFGANG HUBER; ANDY LIAW, T. L., MARTIN MAECHLER, ARNI MAGNUSSON, STEFFEN MOELLER, MARC; VENABLES, S. A. B. **gplots: Various R Programming Tools for Plotting Data. R package version 2.16.0** 2015.

GUEDJ, M. et al. A refined molecular taxonomy of breast cancer. **Oncogene**, v. 31, n. 9, p. 1196-206, Mar 2012. ISSN 1476-5594. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21785460> >.

GUO, W. et al. Slug and Sox9 cooperatively determine the mammary stem cell state. **Cell**, v. 148, n. 5, p. 1015-28, Mar 2012. ISSN 1097-4172. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22385965> >.

GYÖRFFY, B. et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. **Breast Cancer Res Treat**, v. 123, n. 3, p. 725-31, Oct 2010. ISSN 1573-7217. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20020197> >.

GYÖRFFY, B.; SCHÄFER, R. Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. **Breast Cancer Res Treat**, v. 118, n. 3, p. 433-41, Dec 2009. ISSN 1573-7217. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19052860> >.

GYÖRFFY, B. et al. Multigene prognostic tests in breast cancer: past, present, future. **Breast Cancer Res**, v. 17, p. 11, 2015. ISSN 1465-542X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25848861> >.

HABENER, J. F.; STOFFERS, D. A. A newly discovered role of transcription factors involved in pancreas development and the pathogenesis of diabetes mellitus. **Proc Assoc Am Physicians**, v. 110, n. 1, p. 12-21, 1998 Jan-Feb 1998. ISSN 1081-650X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/9460079> >.

HAIBE-KAINS B, S. M., BONTEMPI G, SOTIRIOU C, QUACKENBUSH J. **genefu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer. R package version 1.16.0** 2014.

HAN, J. D. Understanding biological functions through molecular networks. **Cell Res**, v. 18, n. 2, p. 224-37, Feb 2008. ISSN 1748-7838. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18227860> >.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. **Cell**, v. 144, n. 5, p. 646-74, Mar 2011. ISSN 1097-4172. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21376230> >.

HANNENHALLI, S. et al. Transcriptional genomics associates FOX transcription factors with human heart failure. **Circulation**, v. 114, n. 12, p. 1269-76, Sep 2006. ISSN 1524-4539. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16952980> >.

HAO, S. et al. JAM-C promotes lymphangiogenesis and nodal metastasis in non-small cell lung cancer. **Tumour Biol**, v. 35, n. 6, p. 5675-87, Jun 2014. ISSN 1423-0380. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24584816> >.

HASEGAWA, M. et al. CD109, a new marker for myoepithelial cells of mammary, salivary, and lacrimal glands and prostate basal cells. **Pathol Int**, v. 57, n. 5, p. 245-50, May 2007. ISSN 1320-5463. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/17493171> >.

_____. CD109 expression in basal-like breast carcinoma. **Pathol Int**, v. 58, n. 5, p. 288-94, May 2008. ISSN 1440-1827. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18429827> >.

HEITZ, F. et al. Triple-negative and HER2-overexpressing breast cancers exhibit an elevated risk and an earlier occurrence of cerebral metastases. **Eur J Cancer**, v. 45, n. 16, p. 2792-8, Nov 2009. ISSN 1879-0852. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19643597> >.

ITO, H.; WATABE, T.; MIYAZONO, K. Roles of TGF- β family signals in the fate determination of pluripotent stem cells. **Semin Cell Dev Biol**, v. 32, p. 98-106, Aug 2014. ISSN 1096-3634. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24910449> >.

JEMAL, A. et al. Cancer statistics, 2010. **CA Cancer J Clin**, v. 60, n. 5, p. 277-300, 2010 Sep-Oct 2010. ISSN 1542-4863. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20610543> >.

JOHN, L. B.; WARD, A. C. The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. **Mol Immunol**, v. 48, n. 9-10, p. 1272-8, May 2011. ISSN 1872-9142. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21477865> >.

JUNANKAR, S. et al. ID4 controls mammary stem cells and marks breast cancers with a stem cell-like phenotype. **Nat Commun**, v. 6, p. 6548, 2015. ISSN 2041-1723. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25813983> >.

KARACHALIOU, N. et al. ROR1 as a novel therapeutic target for EGFR-mutant non-small-cell lung cancer patients with the EGFR T790M mutation. **Transl Lung Cancer Res**, v. 3, n. 3, p. 122-30, Jun 2014. ISSN 2218-6751. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25806291> >.

KAUR, H. et al. Cadherin-11, a marker of the mesenchymal phenotype, regulates glioblastoma cell migration and survival in vivo. **Mol Cancer Res**, v. 10, n. 3, p. 293-304, Mar 2012. ISSN 1557-3125. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22267545> >.

KIM, N. H. et al. Cadherin 11, a miR-675 target, induces N-cadherin expression and epithelial-mesenchymal transition in melasma. **J Invest Dermatol**, v. 134, n. 12, p. 2967-76, Dec 2014. ISSN 1523-1747. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24940649> >.

KIOUSSIS, D. Aiolos: an ungrateful member of the Ikaros family. **Immunity**, v. 26, n. 3, p. 275-7, Mar 2007. ISSN 1074-7613. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/17376390> >.

KODAKA, M.; HATA, Y. The mammalian Hippo pathway: regulation and function of YAP1 and TAZ. **Cell Mol Life Sci**, v. 72, n. 2, p. 285-306, Jan 2015. ISSN 1420-9071. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25266986> >.

KRISTENSEN, V. N. et al. Principles and methods of integrative genomic analyzes in cancer. **Nat Rev Cancer**, v. 14, n. 5, p. 299-313, May 2014. ISSN 1474-1768. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24759209> >.

KUNO, K.; MATSUSHIMA, K. The IL-1 receptor signaling pathway. **J Leukoc Biol**, v. 56, n. 5, p. 542-7, Nov 1994. ISSN 0741-5400. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/7964161> >.

LAMAR, J. M. et al. The Hippo pathway target, YAP, promotes metastasis through its TEAD-interaction domain. **Proc Natl Acad Sci U S A**, v. 109, n. 37, p. E2441-50, Sep 2012. ISSN 1091-6490. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22891335> >.

LAMOUILLE, S.; XU, J.; DERYNCK, R. Molecular mechanisms of epithelial-mesenchymal transition. **Nat Rev Mol Cell Biol**, v. 15, n. 3, p. 178-96, Mar 2014. ISSN 1471-0080. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24556840> >.

LANDIN MALT, A. et al. Alteration of TEAD1 expression levels confers apoptotic resistance through the transcriptional up-regulation of Livin. **PLoS One**, v. 7, n. 9, p. e45498, 2012. ISSN 1932-6203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23029054> >.

LECONET, W. et al. Preclinical validation of AXL receptor as a target for antibody-based pancreatic cancer immunotherapy. **Oncogene**, v. 33, n. 47, p. 5405-14, Nov 2014. ISSN 1476-5594. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24240689> >.

LEE, B. L. et al. Breast cancer in Brazil: present status and future goals. **Lancet Oncol**, v. 13, n. 3, p. e95-e102, Mar 2012. ISSN 1474-5488. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22381937> >.

LI, A. M. et al. Protocadherin-7 induces bone metastasis of breast cancer. **Biochem Biophys Res Commun**, v. 436, n. 3, p. 486-90, Jul 2013. ISSN 1090-2104. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23751349> >.

LI, L.; LI, W. Epithelial-mesenchymal transition in human cancer: comprehensive reprogramming of metabolism, epigenetics, and differentiation. **Pharmacol Ther**, v. 150, p. 33-46, Jun 2015. ISSN 1879-016X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25595324> >.

LI, Q. et al. Jetset: selecting the optimal microarray probe set to represent a gene. **BMC Bioinformatics**, v. 12, p. 474, 2011. ISSN 1471-2105. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22172014> >.

LI, X. et al. Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. **J Natl Cancer Inst**, v. 100, n. 9, p. 672-9, May 2008. ISSN 1460-2105. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18445819> >.

LIU, S. et al. Breast Cancer Stem Cells Transition between Epithelial and Mesenchymal States Reflective of their Normal Counterparts. **Stem Cell Reports**, v. 2, n. 1, p. 78-91, Jan 2014. ISSN 2213-6711. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24511467> >.

LIU, X.; FAN, D. The epithelial-mesenchymal transition and cancer stem cells: functional and mechanistic links. **Curr Pharm Des**, v. 21, n. 10, p. 1279-91, 2015. ISSN 1873-4286. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25506898> >.

LIU, Y. et al. Increased Serum Soluble Corin in Mid Pregnancy Is Associated with Hypertensive Disorders of Pregnancy. **J Womens Health (Larchmt)**, v. 24, n. 7, p. 572-7, Jul 2015. ISSN 1931-843X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/26086065> >.

M, C. *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.0.0.*

MA, C. X.; ELLIS, M. J. The Cancer Genome Atlas: clinical applications for breast cancer. **Oncology (Williston Park)**, v. 27, n. 12, p. 1263-9, 1274-9, Dec 2013. ISSN 0890-9091. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24624545> >.

MADANI, N. et al. Implication of the lymphocyte-specific nuclear body protein Sp140 in an innate response to human immunodeficiency virus type 1. **J Virol**, v. 76, n. 21, p. 11133-8, Nov 2002. ISSN 0022-538X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/12368356> >.

MALHOTRA, G. K. et al. Histological, molecular and functional subtypes of breast cancers. **Cancer Biol Ther**, v. 10, n. 10, p. 955-60, Nov 2010. ISSN 1555-8576. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21057215> >.

MALLINI, P. et al. Epithelial-to-mesenchymal transition: what is the impact on breast cancer stem cells and drug resistance. **Cancer Treat Rev**, v. 40, n. 3, p. 341-8, Apr 2014. ISSN 1532-1967. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24090504> >.

MANDARANO, L. R. M. **Avaliação da porcentagem de células com alta atividade de aldeído-desidrogenase no tumor primário não prediz a resposta a quimioterapia neoadjuvante em câncer de mama.** . Dissertação de Mestrado apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo. Área de concentração: Tocoginecologia 2013.

MARGOLIN, A. A.; CALIFANO, A. Theory and limitations of genetic network inference from microarray data. **Ann N Y Acad Sci**, v. 1115, p. 51-72, Dec 2007. ISSN 0077-8923. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/17925348> >.

MARGOLIN, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. **BMC Bioinformatics**, v. 7 Suppl 1, p. S7, 2006. ISSN 1471-2105. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16723010> >.

_____. Reverse engineering cellular networks. **Nat Protoc**, v. 1, n. 2, p. 662-71, 2006. ISSN 1750-2799. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/17406294> >.

MARTIN, T. A. et al. Expression of the transcription factors snail, slug, and twist and their clinical significance in human breast cancer. **Ann Surg Oncol**, v. 12, n. 6, p. 488-96, Jun 2005. ISSN 1068-9265. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/15864483> >.

MATSENKO, N. I. U.; KOVALENKO, S. P. [DNA structural features on the borders of ERBB2 amplicons in breast cancer]. **Mol Biol (Mosk)**, v. 47, n. 5, p. 818-27, 2013 Sep-Oct 2013. ISSN 0026-8984. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25509354> >.

MCCANN, B. et al. Associations between pro- and anti-inflammatory cytokine genes and breast pain in women prior to breast cancer surgery. **J Pain**, v. 13, n. 5, p. 425-37, May 2012. ISSN 1528-8447. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22515947> >.

MEDICAL SUBJECT HEADINGS, M. Regulon (definition). 2015. Available at: < <http://www.ncbi.nlm.nih.gov/mesh/?term=regulon> >. Accessed on: 08/07/2015.

MERRIMAN, J. D. et al. Association between an interleukin 1 receptor, type I promoter polymorphism and self-reported attentional function in women with breast cancer. **Cytokine**, v. 65, n. 2, p. 192-201, Feb 2014. ISSN 1096-0023. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24315345> >.

MIHÁLY, Z. et al. A meta-analysis of gene expression-based biomarkers predicting outcome after tamoxifen treatment in breast cancer. **Breast Cancer Res Treat**, v. 140, n. 2, p. 219-32, Jul 2013. ISSN 1573-7217. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23836010> >.

MISHRA, A. et al. Hypoxia stabilizes GAS6/Axl signaling in metastatic prostate cancer. **Mol Cancer Res**, v. 10, n. 6, p. 703-12, Jun 2012. ISSN 1557-3125. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22516347> >.

MIYAKE, T. et al. GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. **Cancer Sci**, v. 103, n. 5, p. 913-20, May 2012. ISSN 1349-7006. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22320227> >.

MO, J. S.; PARK, H. W.; GUAN, K. L. The Hippo signaling pathway in stem cell biology and cancer. **EMBO Rep**, v. 15, n. 6, p. 642-56, Jun 2014. ISSN 1469-3178. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24825474> >.

MORRISSEY, E. R.; DIAZ-URIARTE, R. Pomelo II: finding differentially expressed genes. **Nucleic Acids Res**, v. 37, n. Web Server issue, p. W581-6, Jul 2009. ISSN 1362-4962. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19435879> >.

MURAKAMI, M. et al. A WW domain protein TAZ is a critical coactivator for TBX5, a transcription factor implicated in Holt-Oram syndrome. **Proc Natl Acad Sci U S A**, v. 102, n. 50, p. 18034-9, Dec 2005. ISSN 0027-8424. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16332960> >.

MURTAGH, F.; CONTRERAS, P. **Algorithms for hierarchical clustering**

an overview, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery Volume 2, Issue 1. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2: 86-97 p. 2012.

MYLONA, E. et al. The clinicopathologic and prognostic significance of CD44+/CD24(-/low) and CD44-/CD24+ tumor cells in invasive breast carcinomas. **Hum Pathol**, v. 39, n. 7, p. 1096-102, Jul 2008. ISSN 1532-8392. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18495204> >.

NAYLOR, A. J. et al. The mesenchymal stem cell marker CD248 (endosialin) is a negative regulator of bone formation in mice. **Arthritis Rheum**, v. 64, n. 10, p. 3334-43, Oct 2012. ISSN 1529-0131. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22674221> >.

NICKEL, A.; STADLER, S. C. Role of epigenetic mechanisms in epithelial-to-mesenchymal transition of breast cancer cells. **Transl Res**, v. 165, n. 1, p. 126-42, Jan 2015. ISSN 1878-1810. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24768944> >.

NIELSEN, T. et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. **BMC Cancer**, v. 14, p. 177, 2014. ISSN 1471-2407. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24625003> >.

OCAÑA, O. H. et al. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. **Cancer Cell**, v. 22, n. 6, p. 709-24, Dec 2012. ISSN 1878-3686. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23201163> >.

OCHS, K. et al. Immature mesenchymal stem cell-like pericytes as mediators of immunosuppression in human malignant glioma. **J Neuroimmunol**, v. 265, n. 1-2, p. 106-16, Dec 2013. ISSN 1872-8421. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24090655> >.

PARK, S. J. et al. Interaction of mesenchymal stem cells with fibroblast-like synoviocytes via cadherin-11 promotes angiogenesis by enhanced secretion of placental growth factor. **J Immunol**, v. 192, n. 7, p. 3003-10, Apr 2014. ISSN 1550-6606. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24574497> >.

PASINI, D.; BRACKEN, A. P.; HELIN, K. Polycomb group proteins in cell cycle progression and cancer. **Cell Cycle**, v. 3, n. 4, p. 396-400, Apr 2004. ISSN 1538-4101. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/14752272> >.

PEROU, C. M. et al. Molecular portraits of human breast tumours. **Nature**, v. 406, n. 6797, p. 747-52, Aug 2000. ISSN 0028-0836. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/10963602> >.

PISHVAIAN, M. J. et al. Cadherin-11 is expressed in invasive breast cancer cell lines. **Cancer Res**, v. 59, n. 4, p. 947-52, Feb 1999. ISSN 0008-5472. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/10029089> >.

PIUNTI, A. et al. Polycomb proteins control proliferation and transformation independently of cell cycle checkpoints by regulating DNA replication. **Nat Commun**, v. 5, p. 3649, 2014. ISSN 2041-1723. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24728135> >.

POLYAK, K.; METZGER FILHO, O. SnapShot: breast cancer. **Cancer Cell**, v. 22, n. 4, p. 562-562.e1, Oct 2012. ISSN 1878-3686. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23079664> >.

PONTÉN, F.; JIRSTRÖM, K.; UHLEN, M. The Human Protein Atlas--a tool for pathology. **J Pathol**, v. 216, n. 4, p. 387-93, Dec 2008. ISSN 1096-9896. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18853439> >.

POUDEL, B.; LEE, Y. M.; KIM, D. K. DDR2 inhibition reduces migration and invasion of murine metastatic melanoma cells by suppressing MMP2/9 expression through ERK/NF- κ B pathway. **Acta Biochim Biophys Sin (Shanghai)**, v. 47, n. 4, p. 292-8, Apr 2015. ISSN 1745-7270. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25733533> >.

PRAT, A. et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. **Breast Cancer Res**, v. 12, n. 5, p. R68, Sep 2010. ISSN 1465-542X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20813035> >.

PRAT, A.; PEROU, C. M. Deconstructing the molecular portraits of breast cancer. **Mol Oncol**, v. 5, n. 1, p. 5-23, Feb 2011. ISSN 1878-0261. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21147047> >.

RABQUER, B. J. et al. Junctional adhesion molecule-C is a soluble mediator of angiogenesis. **J Immunol**, v. 185, n. 3, p. 1777-85, Aug 2010. ISSN 1550-6606. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20592283> >.

RAZZAK, A. R.; LIN, N. U.; WINER, E. P. Heterogeneity of breast cancer and implications of adjuvant chemotherapy. **Breast Cancer**, v. 15, n. 1, p. 31-4, 2008. ISSN 1880-4233. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18224391> >.

REBOLLO, A.; SCHMITT, C. Ikaros, Aiolos and Helios: transcription regulators and lymphoid malignancies. **Immunol Cell Biol**, v. 81, n. 3, p. 171-5, Jun 2003. ISSN 0818-9641. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/12752680> >.

REN, T. et al. Increased expression of discoidin domain receptor 2 (DDR2): a novel independent prognostic marker of worse outcome in breast cancer patients. **Med Oncol**, v. 30, n. 1, p. 397, Mar 2013. ISSN 1559-131X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23307244> >.

_____. Discoidin domain receptor 2 (DDR2) promotes breast cancer cell metastasis and the mechanism implicates epithelial-mesenchymal transition programme under hypoxia. **J Pathol**, v. 234, n. 4, p. 526-37, Dec 2014. ISSN 1096-9896. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25130389> >.

RITCHIE, M. E. et al. limma powers differential expression analyzes for RNA-sequencing and microarray studies. **Nucleic Acids Res**, Jan 2015. ISSN 1362-4962. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25605792> >.

ROSENBLUH, J. et al. β -Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. **Cell**, v. 151, n. 7, p. 1457-73, Dec 2012. ISSN 1097-4172. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23245941> >.

ROULEAU, C. et al. Endosialin expression in side populations in human sarcoma cell lines. **Oncol Lett**, v. 3, n. 2, p. 325-329, Feb 2012. ISSN 1792-1074. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22740905> >.

_____. Endosialin is expressed in high grade and advanced sarcomas: evidence from clinical specimens and preclinical modeling. **Int J Oncol**, v. 39, n. 1, p. 73-89, Jul 2011. ISSN 1791-2423. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/21537839> >.

RÉDEI, G. P. **Encyclopedia of Genetics, Genomics, Proteomics, and Informatics**. 3rd. Columbia. Missouri. USA: Springer, 2008. 2201.

SARPESHKAR, R. Analog synthetic biology. **Philos Trans A Math Phys Eng Sci**, v. 372, n. 2012, p. 20130110, Mar 2014. ISSN 1364-503X. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24567476> >.

SATCHER, R. L. et al. Cadherin-11 in renal cell carcinoma bone metastasis. **PLoS One**, v. 9, n. 2, p. e89880, 2014. ISSN 1932-6203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24587095> >.

SCHEEL, C.; WEINBERG, R. A. Cancer stem cells and epithelial-mesenchymal transition: concepts and molecular links. **Semin Cancer Biol**, v. 22, n. 5-6, p. 396-403, Oct 2012. ISSN 1096-3650. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22554795> >.

SCHMITT, F. et al. Cancer stem cell markers in breast neoplasias: their relevance and distribution in distinct molecular subtypes. **Virchows Arch**, v. 460, n. 6, p. 545-53, Jun 2012. ISSN 1432-2307. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22562130> >.

SCHNITT, S. J. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. **Mod Pathol**, v. 23 Suppl 2, p. S60-4, May 2010. ISSN 1530-0285. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20436504> >.

SCHROEDER, A. et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. **BMC Mol Biol**, v. 7, p. 3, 2006. ISSN 1471-2199. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16448564> >.

SCHUETTENGROBER, B.; CAVALLI, G. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. **Development**, v. 136, n. 21, p. 3531-42, Nov 2009. ISSN 1477-9129. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19820181> >.

SCHULTE, J. D. et al. Cadherin-11 regulates motility in normal cortical neural precursors and glioblastoma. **PLoS One**, v. 8, n. 8, p. e70962, 2013. ISSN 1932-6203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23951053> >.

SHABANI, M.; NASERI, J.; SHOKRI, F. Receptor tyrosine kinase-like orphan receptor 1: a novel target for cancer immunotherapy. **Expert Opin Ther Targets**, v. 19, n. 7, p. 941-55, Jul 2015. ISSN 1744-7631. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25835638> >.

SHEKARIAN, T. et al. Paradigm shift in oncology: targeting the immune system rather than cancer cells. **Mutagenesis**, v. 30, n. 2, p. 205-11, Mar 2015. ISSN 1464-3804. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25688113> >.

SHIPITSIN, M.; POLYAK, K. The cancer stem cell hypothesis: in search of definitions, markers, and relevance. **Lab Invest**, v. 88, n. 5, p. 459-63, May 2008. ISSN 1530-0307. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/18379567> >.

SKIBINSKI, A.; KUPERWASSER, C. The origin of breast tumor heterogeneity. **Oncogene**, Feb 2015. ISSN 1476-5594. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25703331> >.

SOCIETY, A. C.; WINCHESTER, D. J.; WINCHESTER, D. P. **Breast cancer : atlas of clinical oncology**. first. Canada: Hamilton, Ont. : B C Decker, 2000. 294 ISBN 97815500911201550091123.

SORLIE, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. **Proc Natl Acad Sci U S A**, v. 100, n. 14, p. 8418-23, Jul 2003. ISSN 0027-8424. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/12829800> >.

STASINOPOULOS, I. A. et al. HOXA5-twist interaction alters p53 homeostasis in breast cancer cells. **J Biol Chem**, v. 280, n. 3, p. 2294-9, Jan 2005. ISSN 0021-9258. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/15545268> >.

STELZER, S. et al. JAM-C is an apical surface marker for neural stem cells. **Stem Cells Dev**, v. 21, n. 5, p. 757-66, Mar 2012. ISSN 1557-8534. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22114908> >.

SØRLIE, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. **Proc Natl Acad Sci U S A**, v. 98, n. 19, p. 10869-74, Sep 2001. ISSN 0027-8424. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/11553815> >.

_____. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyzes across three different platforms. **BMC Genomics**, v. 7, p. 127, 2006. ISSN 1471-2164. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16729877> >.

TAN, E. J.; OLSSON, A. K.; MOUSTAKAS, A. Reprogramming during epithelial to mesenchymal transition under the control of TGFβ. **Cell Adh Migr**, v. 9, n. 3, p. 233-46, 2015. ISSN 1933-6926. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25482613> >.

TAO, J. et al. CD109 is a potential target for triple-negative breast cancer. **Tumour Biol**, v. 35, n. 12, p. 12083-90, Dec 2014. ISSN 1423-0380. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25149155> >.

TAUBE, J. H. et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. **Proc Natl Acad Sci U S A**, v. 107, n. 35, p. 15449-54, Aug 2010. ISSN 1091-6490. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20713713> >.

TAVASSOLI FA, D. P. **World Health Organization Classification of Tumours: Tumors of the Breast and Female Genital Organs** Oxford: Oxford University Press, 2003. 314.

TEAM, R. C. **R: a language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing 2012.

TENTLER, J. J. et al. Patient-derived tumour xenografts as models for oncology drug development. **Nat Rev Clin Oncol**, v. 9, n. 6, p. 338-50, Jun 2012. ISSN 1759-4782. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/22508028> >.

The World Cancer Report--the major findings. **Cent Eur J Public Health**, v. 11, n. 3, p. 177-9, Sep 2003. ISSN 1210-7778. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/14514174> >.

THEUNISSEN, T. W.; JAENISCH, R. Molecular control of induced pluripotency. **Cell Stem Cell**, v. 14, n. 6, p. 720-34, Jun 2014. ISSN 1875-9777. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24905163> >.

TIEZZI, D. G. et al. Expression of aldehyde dehydrogenase after neoadjuvant chemotherapy is associated with expression of hypoxia-inducible factors 1 and 2 alpha and predicts prognosis in locally advanced breast cancer. **Clinics (Sao Paulo)**, v. 68, n. 5, p. 592-8, May 2013. ISSN 1980-5322. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23778413> >.

TOMKOWICZ, B. et al. Endosialin/TEM-1/CD248 regulates pericyte proliferation through PDGF receptor signaling. **Cancer Biol Ther**, v. 9, n. 11, p. 908-15, Jun 2010. ISSN 1555-8576. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/20484976> >.

TOY, K. A. et al. Tyrosine kinase discoidin domain receptors DDR1 and DDR2 are coordinately deregulated in triple-negative breast cancer. **Breast Cancer Res Treat**, v. 150, n. 1, p. 9-18, Feb 2015. ISSN 1573-7217. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25667101> >.

TSUKABE, M. et al. Clinicopathological Analysis of Breast Ductal Carcinoma in situ with ALDH1-Positive Cancer Stem Cells. **Oncology**, v. 85, n. 4, p. 248-256, Oct 2013. ISSN 1423-0232. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24192633> >.

UHLÉN, M. et al. Proteomics. Tissue-based map of the human proteome. **Science**, v. 347, n. 6220, p. 1260419, Jan 2015. ISSN 1095-9203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25613900> >.

VANHOUTTEGHEM, A.; DJIAN, P. Basonuclin 2: an extremely conserved homolog of the zinc finger protein basonuclin. **Proc Natl Acad Sci U S A**, v. 101, n. 10, p. 3468-73, Mar 2004. ISSN 0027-8424. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/14988505> >.

_____. The human basonuclin 2 gene has the potential to generate nearly 90,000 mRNA isoforms encoding over 2000 different proteins. **Genomics**, v. 89, n. 1, p. 44-58, Jan 2007. ISSN 0888-7543. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/16942855> >.

VANHOUTTEGHEM, A. et al. The zinc-finger protein basonuclin 2 is required for proper mitotic arrest, prevention of premature meiotic initiation and meiotic progression in mouse male germ cells. **Development**, v. 141, n. 22, p. 4298-310, Nov 2014. ISSN 1477-9129. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25344072> >.

VAQUERIZAS, J. M. et al. A census of human transcription factors: function, expression and evolution. **Nat Rev Genet**, v. 10, n. 4, p. 252-63, Apr 2009. ISSN 1471-0064. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/19274049> >.

WANG, E. **Cancer System Biology**. CRC Press, 2010. 419.

WILLIAMS, S. A. et al. Patient-derived xenografts, the cancer stem cell paradigm, and cancer pathobiology in the 21st century. **Lab Invest**, v. 93, n. 9, p. 970-82, Sep 2013. ISSN 1530-0307. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23917877> >.

WU, X. et al. AXL kinase as a novel target for cancer therapy. **Oncotarget**, v. 5, n. 20, p. 9546-63, Oct 2014. ISSN 1949-2553. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25337673> >.

XIE, Q. et al. YAP/TEAD-mediated transcription controls cellular senescence. **Cancer Res**, v. 73, n. 12, p. 3615-24, Jun 2013. ISSN 1538-7445. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23576552> >.

YAN, Z. et al. Discoidin domain receptor 2 facilitates prostate cancer bone metastasis via regulating parathyroid hormone-related protein. **Biochim Biophys Acta**, v. 1842, n. 9, p. 1350-63, Sep 2014. ISSN 0006-3002. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24787381> >.

YERSAL, O.; BARUTCA, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. **World J Clin Oncol**, v. 5, n. 3, p. 412-24, Aug 2014. ISSN 2218-4333. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25114856> >.

YIN, M. X.; ZHANG, L. Hippo signaling in epithelial stem cells. **Acta Biochim Biophys Sin (Shanghai)**, v. 47, n. 1, p. 39-45, Jan 2015. ISSN 1745-7270. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25476205> >.

YU, M. et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. **Science**, v. 339, n. 6119, p. 580-4, Feb 2013. ISSN 1095-9203. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23372014> >.

ZHANG, K. et al. The collagen receptor discoidin domain receptor 2 stabilizes SNAIL1 to facilitate breast cancer metastasis. **Nat Cell Biol**, v. 15, n. 6, p. 677-87, Jun 2013. ISSN 1476-4679. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/23644467> >.

ZHANG, S. et al. Ovarian cancer stem cells express ROR1, which can be targeted for anti-cancer-stem-cell therapy. **Proc Natl Acad Sci U S A**, v. 111, n. 48, p. 17266-71, Dec 2014. ISSN 1091-6490. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25411317> >.

ZHANG, Y. et al. A corin variant identified in hypertensive patients that alters cytoplasmic tail and reduces cell surface expression and activity. **Sci Rep**, v. 4, p. 7378, 2014. ISSN 2045-2322. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25488193> >.

ZHU, Y.; QIU, P.; JI, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. **Nat Methods**, v. 11, n. 6, p. 599-600, Jun 2014. ISSN 1548-7105. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/24874569> >.

ÖZLÜ, N. et al. Quantitative comparison of a human cancer cell surface proteome between interphase and mitosis. **EMBO J**, v. 34, n. 2, p. 251-65, Jan 2015. ISSN 1460-2075. Available at: < <http://www.ncbi.nlm.nih.gov/pubmed/25476450> >.

8 SUPPLEMENTARY TABLES

SNAI2_Regulon	TWIST1_Regulon	TBX5_Regulon	ZNF831_Regulon	PRRX1_Regulon	BNC2_Regulon	IKZF3_Regulon	SP140_Regulon	SCML4_Regulon
SEPT11	SEPT11	SEPT11	SEPT1	SEPT11	ABCA6	SEPT1	MARCL1	MARCL1
ADAM12	ABCA6	ABCC9	SEPT6	ABCC9	ABHD6	SEPT6	SEPT6	SEPT1
AGPAT4	ABCA8	ACACB	ACAP1	ACTA2	ABI2	ABI3	ACAP1	SEPT6
AKT3	ABCC9	ADAMTS7	ADCY7	ADAM12	ADAM12	ACAP1	ACSL5	ACAP1
ANAPC7	ADCY4	ADRA2A	AIM2	ADAMTS12	ADAMTS12	ADAMDEC1	ADAMDEC1	ADAM28
ANTXR1	ADRA2A	AKAP12	AKNA	ADAMTS2	ADAMTS2	AIM2	ADCY7	AIM2
ARHGAP28	AKAP12	AKT3	AMICA1	ADAMTS5	ADAMTS6	AKNA	AIM2	AKNA
BICC1	AKAP2	ANGPTL2	ANKRD44	ADAMTS6	ADAMTS9.AS2	ANKRD36BP2	AKNA	AMICA1
BNC2	AKT3	ANK2	AOAH	ADAMTSL1	AEBP1	ANKRD44	ANKRD44	ANKRD44
CALD1	ANGPTL2	ANTXR1	APOBEC3G	AEBP1	AKAP12	AOAH	ANKRD55	AOAH
CDH11	ANK2	ANTXR2	ARHGAP15	AGPAT4	AKT3	APOBEC3F	AOAH	APOBEC3G
COL3A1	ANTXR2	ANXA1	ARHGAP25	AHNAK2	ALDH1L2	APOBEC3G	APBB1IP	ARHGAP15
COL6A3	APOB	ARHGAP20	ARHGAP30	AKT3	AMPH	ARHGAP15	APOBEC3G	ARHGAP25
CRISPLD2	ARHGAP20	ARHGAP24	ARHGAP9	ALDH1L2	ANGPTL2	ARHGAP25	APOL3	ARHGAP27
CTSK	ATAD2	ASCL1	B2M	ANGPTL2	ANK2	ARHGAP30	APOL6	ARHGAP30
CYTH3	ATOH8	AXL	BANK1	ANK2	ANTXR1	ARHGAP9	ARHGAP15	ARHGAP9
DACT1	AVPR1A	BHLHE41	BCL11B	ANTXR1	ANTXR2	B2M	ARHGAP25	ATM
DCN	BENDE6	BIRC5	BIN2	ANTXR2	ARHGAP20	BANK1	ARHGAP30	ATP8A1
DDR2	BICC1	BMP1	BIRC3	ARHGAP24	ARHGAP24	BCL11B	ARHGAP9	B2M
DZIP1	BNC2	BNC2	BTLA	ARHGAP28	ARHGAP28	BCL2A1	ATM	BANK1
EFEMP2	C13orf33	C1orf151-NBL1	BTN3A1	ARHGAP31	ARHGEF40	BIN2	B2M	BCL11B
EMP1	CCDC50	C1QTNF5	BTN3A2	ARHGEF40	ARL10	BIRC3	BANK1	BCL2A1
FBLN1	CCDC80	C3orf62	BTN3A3	ASPN	ASPN	BMP6	BCL11B	BEND4
FBN1	CCL14	C5orf62	BZRAP1-AS1	AXL	AXL	BTX	BCL2A1	BIN2
FERMT2	CDC14B	CACNA2D1	C11orf21	BACE1	BACE1	BTLA	BEND4	BIRC3
FEZ1	CDH5	CCDC25	C16orf54	BGN	BEND6	BTN3A1	BIN2	BTLA
FIBIN	CEP112	CCDC36	C1orf186	BHLHE22	BGN	BTN3A3	BIRC3	BTN2A2
FKBP7	CFH	CCDC80	C5orf58	BICC1	BHLHE22	BZRAP1-AS1	BTX	BTN3A2
FKBP9	CLEC1A	CDH11	CAMK4	BMP1	BICC1	C11orf21	BTLA	BTN3A3
FSTL1	CNRIP1	CDK13	CCDC69	BMPR2	BMP1	C16orf54	BTN2A2	C11orf21
GAS1	COLEC12	CDR1	CCL19	BNC2	BNC2.1	C1orf186	BTN3A1	C16orf54
GAS7	CPED1	CFH	CCL5	C13orf33	C14orf37	C5orf20	BTN3A2	C16orf62
GJA1	CPXM1	CHL1	CCR2	C14orf37	C1orf135	C5orf58	BTN3A3	C5orf20
GLS3	CXCL12	CKAP2	CCR5	C1orf151-NBL1	C1orf151.NBL1	CAMK4	BZRAP1-AS1	C5orf58
GLT8D2	DAB2	CLMP	CD19	C1QTNF5	C1QTNF5	CARD11	C11orf21	C7orf44
GOLIM4	DCN	CLU	CD2	C15	C20orf194	CCDC88B	C16orf54	CAMK4
GXYLT2	DDR2	CNRIP1	CD200R1	C5orf62	C5orf62	CCL19	C1orf162	CARD11
HIST1H2BD	DLG1	COL5A3	CD226	C9orf3	CACNA2D1	CCL4	C1orf186	CARD8
HMCN1	DLG3	COLEC12	CD247	CALD1	CALD1	CCL5	C5orf20	CASP1
HOXA10-HOXA9	DST	COP22	CD27	CAV2	CCDC8	CCR2	C5orf58	CASP10
HTRA1	EBF1	CTSK	CD28	CCDC69	CCDC80	CCR4	CAMK4	CCDC69
KDELCL1	EBF3	CWF19L2	CD38	CCDC80	CD99	CCR5	CARD11	CCDC88B
LAMB1	ECSCR	CXCL12	CD3D	CD109	CDCA5	CCR7	CARD16	CCL19
LAMC1	EFEMP1	DAAM2	CD3E	CD248	CDH11	CD19	CASP1	CCL5
LOC100132116	EFHA2	DAB2	CD3G	CD99	CDH5	CD2	CASP10	CCR2
LOC100505813	EGFLAM	DAD1	CD40	CDH11	CDK6	CD200R1	CCL4	CCR4
LOX	EHD2	DCHS1	CD40LG	CDR1	CFH	CD226	CCL5	CCR5
LUM	ENPEP	DCN	CD48	CEL2	CHRNA1	CD247	CCR2	CCR7
MEG3	EPDR1	DDR2	CD5	CFH	CHSY3	CD27	CCR4	CD19
MEG8	ERG	DFFB	CD52	GFI	CLMP	CD274	CCR5	CD1D
MIR100HG	ESAM	DLC1	CD53	CHSY3	CMTM3	CD28	CCR7	CD2
MSRB3	ESRP1	DNM3OS	CD6	CLDN5	CNRIP1	CD37	CCR2	CD200R1
MXRA7	ESYT1	DPY	CD69	CLMP	CNTN1	CD38	CD19	CD226
NAP1L3	EVC2	DPYSL2	CD74	CMTM3	COL10A1	CD3D	CD2	CD244
NID1	F13A1	DPYSL3	CD84	CNRIP1	COL12A1	CD3E	CD226	CD247
NID2	FAM13A-AS1	DSEL	CD8A	CNTN1	COL14A1	CD3G	CD247	CD27
NTSE	FAM162B	DST	CD8B	COL10A1	COL16A1	CD48	CD274	CD274
OLFML2A	FBLN1	EFEMP2	CD96	COL12A1	COL18A1	CD5	CD28	CD28
PAM	FBLN2	EFHA2	CD42SE2	COL1A1	COL1A1	CD52	CD38	CD38
PARVA	FBN1	EHD2	CEACAM21	COL1A2	COL1A2	CD53	CD3D	CD3D
PCDH7	FERMT2	EMP1	CEL2	COL3A1	COL3A1	CD6	CD3E	CD3E
PDGFRL	FEZ1	EMX2	CIITA	COL5A1	COL5A1	CD69	CD3G	CD3G
PLS3	FGF7	ERG	CKAP2	COL5A2	COL5A2	CD72	CD40	CD40LG
POSTN	FLRT2	ESRP1	CLEC2D	COL6A1	COL6A1	CD74	CD48	CD48
PRKD1	FOXN3	ESYT2	CLECL1	COL6A2	COL6A2	CD79A	CD5	CD5
PRRX1	FOXO1	F13A1	CLUAP1	COL6A3	COL6A3	CD79B	CD52	CD52
PTPRG	FSTL1	FAM41AY1	CORO1A	COL8A1	COL8A1	CD80	CD53	CD53
QKI	GALNTL2	FAM82A1	COTL1	COL8A2	COL8A2	CD84	CD6	CD6
RECK	GAS1	FAP	CR1	COP22	COP22	CD8A	CD69	CD69
RFTN2	GNG11	FAT4	CR2	CPXM1	CORIN	CD88	CD72	CD72
SACS	GPR116	FBLN1	CRTAM	CPZ	CPXM1	CD96	CD79A	CD79A
SFRP2	GPR124	FBN1	CSF2RB	CRISPLD2	CPZ	CD42SE2	CD79B	CD79B
SGCD	GPR146	FBXL5	CST7	CRTAP	CRTAP	CDON	CD80	CD84
SH3PXD2A	HDAC9	FETUB	CTLA4	CSRNP3	CSTF3	CEACAM21	CD83	CD8A
SH3PXD2B	HS2ST1	FEZ1	CTSW	CSRNP2	CTDSP2	CEL2	CD84	CD8B
SHOX2	HSPB6	FFAR3	CXCL9	CTSK	CTHRC1	CFLAR	CD86	CD96
SNX18	IGF2	FGF10	CXCR3	CXCL12	CTSK	CIITA	CD8A	CD42SE2
SPARCL1	IPO11	FHL1	CXCR5	CXorf36	CTTNBP2	CKAP2	CD8B	CEACAM21
SPON1	ITGA1	FIBIN	CXCR6	CYP2U1	CXCL12	CLEC2D	CD96	CECR1
SRPX	KIAA1462	FMOD	CXorf65	CYS1	CYBRD1	CLECL1	CD42SE2	CEL2
SSPN	KLHL3	FOXN3	CYAT1	CYTH3	CYGB	CLUAP1	CEACAM21	CIITA
TCF4	LAMA4	FOXO1	CYLD	DAB2	CYS1	CORO1A	CEP128	CLEC2D
THBS2	LDB2	FSTL1	CYTH4	DACT1	DAB2	CPNE5	CIITA	CLEC4C
TIMP2	LEPR	GALNTL2	CYTIIP	DCN	DACT1	CR1	CKAP2	CLECL1
TRPC1	LEPROT	GAS1	DENND1C	DDR2	DCHS1	CR2	CLEC2D	CLUU1
TSHZ3	LHFP	GAS7	DERL3	DLC1	DCN	CRTAM	CLECL1	CLNK

SNAI2_Regulon	TWIST1_Regulon	TBX5_Regulon	ZNF831_Regulon	PRRX1_Regulon	BNC2_Regulon	IKZF3_Regulon	SP140_Regulon	SCML4_Regulon
WWTR1	LOC100505679	GJC1	DGKA	DNM3O5	DDR2	CSF2R8	CLLU1	CLUAP1
XYLT1	LRP1	GOLGA8IP	DOCK2	DPT	DKK3	CST7	CLUAP1	CNTRL
ZEB1	MAF	GPR124	DOCK8	DPYSL3	DLC1	CTLA4	CMKL1	COL4A4
ZNF521	MAGI2-AS3	HIC1	DTHD1	DST	DNM3O5	CTSC	CORO1A	CORO1A
-	MAP1B	HIST1H2BC	EOMES	EDNRA	DPT	CTSW	COTL1	COTL1
-	MAP4	HIST1H2BH	EV12B	EFEMP2	DPYSL3	CXCL10	CR1	CR1
-	MAP4K4	HMCN1	FAM26F	EHD2	DSEL	CXCL11	CR2	CRTAM
-	MCAM	HNMT	FAM65B	ELK3	DST	CXCL13	CRTAM	CSF2R8
-	MECOM	HOXA10-HOXA9	FASLG	EMCN	DZIP1	CXCL9	CSF2R8	CST7
-	MEF2C	HOXA7	FCRL1	EMILIN2	EBF1	CXCR3	CTLA4	CTLA4
-	MEG3	HSPA12B	FCRL2	EMP1	EDNRA	CXCR4	CTSC	CTSC
-	MIR100HG	IFFO1	FCRL3	ERG	EFEMP2	CXCR5	CTSS	CTSS
-	MME	IGFBP4	FCRL5	ESAM	EFHA2	CXCR6	CXCL10	CTSW
-	MMRN2	ISLR	FCRLA	EVC	EHD2	CXorf65	CXCL11	CUX2
-	MRV1	ITGBL1	FGD2	F13A1	EIF5A2	CYAT1	CXCL13	CXCL13
-	MSRB3	JAM2	FGL2	FAM26E	EMX2	CYLD	CXCL9	CXCL9
-	MXRA7	JAM3	FLJ32255	FAP	ENTPD1	CYTIP	CXCR3	CXCR3
-	MYCT1	KANK2	FNBP1	FBLN1	EVC	DAPP1	CXCR5	CXCR4
-	NDN	KCTD12	FYB	FBLN2	EVC2	DEF6	CXCR6	CXCR5
-	NLRP1	KIAA1462	GBP1	FBN1	FABP4	DENND1C	CXorf21	CXCR6
-	NR3C1	KIRREL	GBP2	FERMT2	FAM198B	DOCK2	CXorf65	CXorf65
-	NRP1	KRR1	GBP4	FEZ1	FAM26E	DOCK8	CYBB	CYTH4
-	NTSE	KRT84	GBP5	FGD5	FANCA	DTHD1	CYLD	CYTIP
-	OLFML1	LAMA4	GF11	FHL1	FAP	EAF2	CYTH4	DEF6
-	PARVA	LAMB1	GIMAP1-GIMAP5	FIBIN	FAS	EML4	CYTIP	DOCK10
-	PCDH18	LDB2	GIMAP6	FILIP1L	FAT3	EOMES	DAPP1	DOCK2
-	PCSK5	LHFP	GIMAP7	FLRT2	FBLN1	EV12B	DDX26B	DOCK8
-	PDE3A	LIMA1	GLYR1	FMOD	FBLN2	FAM129C	DENND1C	DPEP2
-	PDGFRA	LOC100505813	GNLY	FN1	FBN1	FAM26F	DOCK10	DTHD1
-	PEAK1	LOC100507630	GPR174	FOXN3	FER	FAM46C	DOCK2	EMLA
-	PECAM1	LPAR1	GPR183	FOXO1	FERMT2	FAM65B	DOCK8	EOMES
-	PEL1	LRP1	GPR65	FSTL1	FEZ1	FAM78A	EMLA	EPST11
-	PGM5	LUM	GVINP1	GAS1	FGD5	FASLG	EOMES	EV12B
-	PLAGL1	MAF	GZMA	GAS7	FGF1	FCRL1	EPHA1-AS1	FAIM3
-	PRDM5	MAGI2-AS3	GZMB	GFPT2	FGF14	FCRL2	EPST11	FAM129C
-	PRRX1	MAP1A	GZMH	GLT8D2	FGF18	FCRL3	ETS1	FAM26F
-	PTPRB	MAP4K4	GZMK	GNG2	FHL1	FCRL5	EV12B	FAM65B
-	PTPRG	MEG3	HCLS1	GOLGA8IP	FHL5	FCRLA	F5	FAM71B
-	PTRF	MEG8	HLA-DOA	GPR124	FIBIN	FGD2	FAM26F	FAM78A
-	RARRES2	MEIS2	HLA-DOB	GXYLT2	FILIP1L	FGL2	FAM65B	FASLG
-	RBMS2	MIR10A	HLA-DPA1	HEG1	FKBP7	FLJ32255	FAM78A	FCRL1
-	RECK	MIR143HG	HLA-DPB1	HMCN1	FLRT2	FMNL1	FASLG	FCRL2
-	RHOJ	MLL3	HLA-DQA1	HNMT	FMOD	FNBP1	FCHSD2	FCRL3
-	ROBO4	MMP19	HLA-DQB1	HOTAIRM1	FN1	FYB	FCRL1	FCRL5
-	RUNX1T1	MRV11	HLA-DRA	HOXA10-HOXA9	FRMD6	GABBR1	FCRL2	FCRLA
-	S1PR1	MSRB3	HLA-DRB1	HRH1	FSTL1	GBP1	FCRL3	FERMT3
-	SEMA6A	NAP1L3	HLA-E	HSD17B11	GAS1	GBP2	FCRL5	FGD2
-	SERPING1	NAV3	HLA-F	HSPG2	GAS7	GBP4	FCRLA	FGL2
-	SH3RF3	NDN	HMHA1	HTRA1	GFPT2	GBP5	FGD2	FMNL1
-	SLC16A7	NEGR1	ICOS	IFFO1	GJB2	GF11	FGL2	FNBP1
-	SLIT3	NID1	IDO2	IGF2	GLIS3	GIMAP1-GIMAP5	FLU1	FYB
-	SOD3	NID2	IFNG	IGFBP7	GLT8D2	GLYR1	FMNL1	FYN
-	SOX7	NNMT	IGFLR1	IL1R1	GPC6	GNLY	FNBP1	GABBR1
-	SRPX	NPR2	IGHA1	ISLR	GPX8	GP18A	FYB	GBP4
-	SSPN	NRP2	IGHG1	ITGA1	GXYLT2	GPR171	GAB3	GBP5
-	SULT1C4	NUAK1	IGHM	ITGBL1	H2AFZ	GPR174	GABBR1	GF11
-	SVEP1	OLFML1	IGK@	JAM2	HEG1	GPR52	GBP1	GIMAP1
-	SYNE1	OLFML3	IGKC	JAM3	HEPH	GRAP2	GBP2	GIMAP1-GIMAP5
-	TCEAL7	PALM	IGLC1	KANK2	HGF	GVINP1	GBP4	GIMAP4
-	TCF4	PARP1	IGLV1-44	KCTD12	HMCN1	GZMA	GBP5	GIMAP6
-	TEK	PARVA	IKZF1	KIAA1462	HRH1	GZMB	GF11	GIMAP7
-	TENC1	PCSK5	IKZF3	KIRREL	HSD17B12	GZMK	GIMAP1	GIMAP8
-	TFPI	PDE11A	IL10RA	LAMA4	HSPA12B	HCLS1	GIMAP1-GIMAP5	GLYR1
-	TGFBR2	PDE3A	IL12RB1	LAMB1	HTRA1	HLA-DMB	GIMAP2	GMFG
-	THSD7A	PDGFRB	IL16	LAMC1	HTRA3	HLA-DOA	GIMAP4	GNLY
-	TIMP2	PDGFRL	IL18R1	LATS2	IGDCC4	HLA-DOB	GIMAP6	GP18A
-	TRPC1	PEAK1	IL18RAP	LDB2	IGF2	HLA-DPA1	GIMAP7	GPR171
-	TSHZ3	PEL12	IL21R	LEPROT	IGFBP4	HLA-DPB1	GLS	GPR174
-	TSPAN7	PLSCR4	IL23A	LHFP	IL1R1	HLA-DQA1	GMFG	GPR18
-	TWIST2	PNMA6A	IL26	LINC00478	ISLR	HLA-DQB1	GNLY	GPSM3
-	VIM	PRDM2	IL2RB	LINC00514	ITGA1	HLA-DRA	GPR171	GRAP2
-	WHAMMP2	PRICKLE2	IL2RG	LIX1L	ITGB5	HLA-F	GPR174	GVINP1
-	ZEB1	PRKD1	IL7R	LOC100132116	ITGBL1	HMHA1	GPR183	GZMA
-	ZEB2	PRPF4B	INPP4A	LOC100132891	JAM2	ICOS	GPR65	GZMK
-	ZFHx4	PRRX1	INPP5D	LOC100505813	JAM3	IDO1	GRAP2	HCLS1
-	ZNF471	PTGIS	IPCEF1	LOC100506941	JAZF1	IDO2	GVINP1	HIGD1A
-	-	PTPRG	IRF1	LOC100653217	KANK2	IFNG	GZMA	HLA-DMB
-	-	PTPRM	IRF4	LOC728392	KCN2D	IGFLR1	GZMB	HLA-DOA
-	-	PTRF	IRF8	LOX	KCNJ12	IGH@	GZMH	HLA-DOB
-	-	PURG	ITGA4	LOXL1	KCNMB4	IGHA1	GZMK	HLA-DPB1
-	-	OKI	ITGAL	LPAR1	KIAA1199	IGHD	HCLS1	HLA-DRA
-	-	RAP1A	ITGB7	LRCH2	KIAA1462	IGHG1	HLA-A	HLA-E
-	-	RARRES2	ITK	LRP1	KIF26B	IGHM	HLA-C	HLA-F
-	-	RASGRF2	ITM2A	LRRC15	KLHL3	IGHV3-48	HLA-DMB	HMHA1
-	-	RBMS2	JAK3	LRRC32	LAMA2	IGK@	HLA-DOA	HS3T3B1
-	-	RCAN2	KCNA3	LUM	LAMA4	IGKC	HLA-DOB	HVCN1

SNAI2_Regulon	TWIST1_Regulon	TBX5_Regulon	ZNF831_Regulon	PRRX1_Regulon	BNC2_Regulon	IKZF3_Regulon	SP140_Regulon	SCML4_Regulon
-	-	RCBTB2	KIAA0125	MAF	LAMB1	IGLC1	HLA-DPB1	ICOS
-	-	RECK	KIF21B	MAGEL2	LATS2	IGLJ3	HLA-DQA1	IDO1
-	-	RFTN2	KLHL6	MAGI2-AS3	LDB2	IGLL3P	HLA-DQB1	IDO2
-	-	RNF180	KLRC1	MAN1A1	LEPREL1	IGLL5	HLA-DRA	IFI16
-	-	RSPO3	KLRC3	MAN2A1	LHFP	IGLV1-44	HLA-DRB1	IFNG
-	-	RUNX1T1	KLRC4	MAP1A	LMO1	IGLV6-57	HLA-E	IGFLR1
-	-	RUNX2	KLRC4-KLRK1	MAP4K4	LMOD1	IKZF1	HLA-F	IGH@
-	-	SERPINF1	KLRD1	MBNL1	LOC100130872	IKZF3	HLA-F-AS1	IGHA1
-	-	SFRP4	KLRG1	MCAM	LOC100132891	IL10RA	HMHA1	IGHD
-	-	SGCD	LAMP3	MEG3	LOC100287221	IL12B	HS3ST3B1	IGHG1
-	-	SHOX2	LAP3	MEG8	LOC100505813	IL12RB1	ICOS	IGHM
-	-	SUT2	LAT	MEIS2	LOC100506941	IL16	IDO1	IGK@
-	-	SUT3	LAX1	MFAP4	LOC100509635	IL18RAP	IDO2	IGKC
-	-	SMC4	LCK	MFAP5	LOC283867	IL21R	IFI16	IGLC1
-	-	SNCA	LCP1	MGC24103	LOX	IL23A	IFNAR2	IGLL5
-	-	SORBS1	LCP2	MIR100HG	LOXL1	IL26	IFNG	IKZF1
-	-	SPARCL1	LIMD2	MIR143HG	LOXL2	IL2RA	IGFLR1	IKZF3
-	-	SPATA9	LOC100287723	MME	LPAR1	IL2RB	IGH@	IL10RA
-	-	SPON1	LOC100506776	MMP2	LRCH1	IL2RG	IGHA1	IL12RB1
-	-	SRSF2	LOC100507463	MN1	LRCH2	IL7R	IGHD	IL16
-	-	SSPN	LOC100653340	MNF1	LRP1	INPP4A	IGHG1	IL18RAP
-	-	STON1	LOC285972	MRGPRF	LRRC15	INPP5D	IGHM	IL21R
-	-	SVEP1	LOC642838	MRV11	LRRC17	IPCEF1	IGHV3-48	IL23A
-	-	SYNE1	LRRC18	MSRB3	LRRC32	IRF1	IGK@	IL26
-	-	TBC1D2B	LTB	MTMR7	LSAMP	IRF4	IGKC	IL2RB
-	-	TBX15	LY9	MXRA5	LTBP2	IRF8	IGLL5	IL2RG
-	-	TBX5	MAP4K1	MXRA7	LUM	ITGA4	IKZF1	IL36B
-	-	TBX5-AS1	MBNL1	MXRA8	MAGEL2	ITGAL	IKZF3	IL7R
-	-	TCF4	MCOLN2	MYCT1	MAGI2-AS3	ITGB7	IL10RA	INPP5D
-	-	TEK	MGC40069	MYL9	MAN1A1	ITK	IL12B	IPCEF1
-	-	TGFBR2	MGC42157	NAP1L3	MAN2A1	JAK3	IL12RB1	IRF1
-	-	TIMP2	MIR155	NBLA00301	MAP4K4	JAKMIP1	IL16	IRF4
-	-	TLN2	MIR3929	NDN	MED17	KCNA3	IL18RAP	IRF8
-	-	TNS1	MPEG1	NEGR1	MEG3	KCNJ10	IL21R	ITGA2B
-	-	TRPC1	MS4A1	NHS	MEG8	KIAA0125	IL23A	ITGA4
-	-	TSHZ3	MYO1G	NID1	MFAP4	KIF21B	IL2RB	ITGAL
-	-	TWIST2	NCKAP1L	NID2	MFAP5	KIF2A	IL2RG	ITGAX
-	-	UBE3C	NEDD9	NNMT	MGC24103	KLHL6	IL7R	ITGB7
-	-	USP1	NFATC2	NOX4	MGC4294	KLRB1	INPP5D	ITK
-	-	VCAN	NKG7	NRN1	MIR100HG	KLRC3	IPCEF1	ITM2A
-	-	VGLL3	NLRCS	NRP1	MIR143HG	KLRC4	IRF1	JAK3
-	-	VSTM4	P2RX5	NRP2	MMP19	KLRC4-KLRK1	IRF4	KCNA3
-	-	ZEB1	P2RY10	NTSE	MMP2	KLRD1	IRF8	KIAA0125
-	-	ZFHx4	P2RY8	NUAK1	MN1	KLRG1	ITGA4	KIAA1191
-	-	ZFHx4-AS1	PAG1	ODZ4	MRC2	LAMP3	ITGAL	KIF21B
-	-	ZFPM2	PARP15	OLFML1	MRGPRF	LAP3	ITGB7	KIF2A
-	-	ZMAT3	PARP8	OLFML2B	MRV11	LAT	ITK	KIR3DL1
-	-	ZNF423	PARVG	OLFML3	MSRB3	LAX1	JAK3	KLHL6
-	-	ZNF697	PATL2	OSBP2	MTHFD2	LCK	KBTBD8	KLRB1
-	-	-	PAX5	PALLD	MTMR2	LCP1	KCNA3	KLRC4-KLRK1
-	-	-	PIK3CD	PAM	MXRA5	LCP2	KCNJ10	KLRD1
-	-	-	PIK3CG	PARVA	MXRA7	LGALS2	KIAA0125	KLRG1
-	-	-	PIK3R5	PCDH7	MXRA8	LILRB2	KIF21B	LAMP3
-	-	-	PIM2	PCOLCE	MYH2	LIMD2	KLHL6	LAT
-	-	-	PLA2G2D	PCSK5	MYL6	LOC100130100	KLRC3	LAX1
-	-	-	PLAC8	PDE3A	MYL9	LOC100287723	KLRC4	LCK
-	-	-	PLCKD2	PDE7B	MYLK	LOC100505658	KLRC4-KLRK1	LCP1
-	-	-	POUZAF1	PDGFC	NAP1L3	LOC100505746	KLRD1	LCP2
-	-	-	POUF2F2	PDGFRA	NAV1	LOC100506776	LAMP3	LGALS2
-	-	-	PPP1R16B	PDGFRB	NAV3	LOC100507463	LAP3	LINC00184
-	-	-	PPP2R5C	PDGFRL	NBL1	LOC100508797	LAPTM5	LOC100128843
-	-	-	PRF1	PDLIM2	NBLA00301	LOC100653340	LAT	LOC100287723
-	-	-	PRKCB	PDN	NDN	LOC285628	LAT2	LOC100505746
-	-	-	PRKCH	PEAK1	NDNF	LOC285972	LAX1	LOC100505812
-	-	-	PRKCQ	PECAM1	NEGR1	LOC439949	LCK	LOC100506776
-	-	-	PSTPIP1	PLAGL1	NID1	LOC642838	LCP1	LOC100507164
-	-	-	PTCSC1	PLXDC1	NID2	LOC96610	LCP2	LOC100507206
-	-	-	PTN	PLXDC2	NNMT	LRRC18	LGALS2	LOC100507463
-	-	-	PTPN22	POSTN	NOVA1	LST1	LILRA6	LOC100653340
-	-	-	PTPN7	PPAP2B	NOX4	LTB	LILRB1	LOC285628
-	-	-	PTPRC	PPAPDC1A	NPR2	LY9	LILRB2	LOC285972
-	-	-	PVRIG	PRDM5	NRP2	LYZ	LOC100129518	LOC401093
-	-	-	PYHIN1	PRICKLE1	NTSE	MAL	LOC100505501	LOC439949
-	-	-	RAB33A	PRICKLE2	NUAK1	MAP4K1	LOC100505812	LOC643733
-	-	-	RAC2	PRKD1	ODZ4	MBNL1	LOC100506776	LOC79015
-	-	-	RASA2	PRKG1	OLFML1	MCOLN2	LOC100507286	LOC96610
-	-	-	RASSF5	PRNP	OLFML2B	MEI1	LOC100507463	LPXN
-	-	-	RNASE6	PRRX1	OLFML3	MGC40069	LOC100507616	LRCH4
-	-	-	RTKN2	PRTFDC1	P4HA3	MGC42157	LOC100653340	LRMP
-	-	-	RUNX3	PTPN21	PALLD	MIAT	LOC285628	LRRC18
-	-	-	SAMD3	PTPRD	PAM	MIR155	LOC285972	LST1
-	-	-	SAMSN1	PTPRG	PARVA	MIR3929	LOC401093	LTA
-	-	-	SASH3	PTRF	PCDH7	MS4A1	LOC439949	LTB
-	-	-	SCML4	RAB23	PCSK5	MYO1G	LPXN	LY9
-	-	-	SEL1L3	RARRES2	PDE1A	MYO1F	LRMP	MAL
-	-	-	SELL	RASAL2	PDE3A	MZB1	LRRC18	MAP4K1

SNAI2_Regulon	TWIST1_Regulon	TBX5_Regulon	ZNF831_Regulon	PRRX1_Regulon	BNC2_Regulon	IKZF3_Regulon	SP140_Regulon	SCML4_Regulon
-	-	-	SEMA4D	RASGRF2	PDE4DIP	NABP1	LST1	MBNL1
-	-	-	SH2D1A	RBMS2	PDGFC	NAPS8	LTA	MCOLN2
-	-	-	SH2D2A	RBMS3	PDGFRA	NCF1C	LTB	MEI1
-	-	-	SIRPG	RECK	PDGFRB	NCLN	LY9	MGC12916
-	-	-	SLA	RFTN2	PDGFR1	NDUFA2	LYN	MGC40069
-	-	-	SLA2	RHOJ	PDLM2	NEDD9	LYZ	MGC42157
-	-	-	SLAMF1	RNF144A	PDPN	NFATC2	MAP4K1	MIAT
-	-	-	SLAMF6	RNF146	PEL1	NKG7	MCOLN2	MIR155
-	-	-	SLAMF7	ROR1	PELO	NLR3	MEI1	MIR3929
-	-	-	SLC25A53	RRAGA	PHLDB1	NOL3	MGC12916	MOG
-	-	-	SMCHD1	RUNX1T1	PLXDC2	P2RX5	MGC40069	MPEG1
-	-	-	SNX20	RUNX2	PLXNC1	P2RY10	MGC42157	MS4A1
-	-	-	SP140	S100A6	POSTN	P2RY8	MIAT	MYCBP2
-	-	-	SPOCK2	SAMD4A	PPAPDC1A	PAG1	MIR155	MYO1G
-	-	-	SRGN	SERPINF1	PRDM5	PARP15	MIR3929	NABP1
-	-	-	STBIA1	SFRP2	PRDM6	PARP8	MLKL	NCKAP1L
-	-	-	STAMBPL1	SGCD	PRELP	PARVG	MNDA	NCR1
-	-	-	STAP1	SH3PYD2A	PRICKLE1	PATL2	MPEG1	NEDD9
-	-	-	STAT4	SH3RF3	PRICKLE2	PAX5	MS4A1	NEO1
-	-	-	STK17B	SHOX2	PRKD1	PDE7A	MS4A6A	NFATC2
-	-	-	STK4	SLC4A4	PRKG1	PIK3CD	MYO1G	NIN
-	-	-	TAGAP	SLIT3	PRND	PIK3CG	NABP1	NKG7
-	-	-	TARP	SNAI2	PROS1	PIK3R5	NCKAP1L	NLR3
-	-	-	TBC1D10C	SNED1	PRRX1	PIM2	NCLN	NLRP7
-	-	-	TBX21	SNORD114-3	PTGER3	PLA2G2D	NEDD9	NUB1
-	-	-	TESPA1	SNX18	PTPN21	PLAC8	NFATC2	OXNAD1
-	-	-	THEMIS	SNX29	PTPRD	PLCXD2	NKG7	P2RX5
-	-	-	TIGIT	SPARCL1	PTPRG	PLEK	NLR3	P2RY10
-	-	-	TLR10	SPOCK1	PTRF	PNOC	OASL	P2RY8
-	-	-	TMC8	SPON1	PURG	POU2AF1	P2RX5	PAG1
-	-	-	TNFAIP3	SRPX	RAI2	POU2F2	P2RY10	PARP15
-	-	-	TNFAIP8	SSC5D	RARRES2	PPP1R16B	P2RY8	PARVG
-	-	-	TNFRSF17	SSPN	RASGRF2	PPP2R5C	PAG1	PATL2
-	-	-	TNFRSF8	STAB1	RBMS2	PRF1	PARP14	PBRM1
-	-	-	TOX	STGC3	RBMS3	PRKCB	PARP15	PCBP1
-	-	-	TRABD2A	STON1	RCAN2	PRKCQ	PARP8	PCED1B
-	-	-	TRAC	SYNE1	RECK	PRO0471	PARVG	PDE3B
-	-	-	TRAF1	SYT11	RFTN2	PSTPIP1	PATL2	PIK3CD
-	-	-	TRAF3IP3	TAGLN	RNF144A	PTCSC1	PDCD1LG2	PIK3CG
-	-	-	TRAT1	TBC1D28	RORB	PTK2B	PIK3AP1	PIK3R5
-	-	-	TRAV8-3	TBX18	RP56KA2	PTPN22	PIK3CD	PIM2
-	-	-	TRBC1	TBX5	RRAGA	PTPN7	PIK3CG	PLA2G2D
-	-	-	TXK	TBX5-AS1	RUNX1T1	PTPRC	PIK3R5	PLAC8
-	-	-	UBASH3A	TCEAL7	RUNX2	PTPRCAP	PIM2	PLCG2
-	-	-	VAV1	TCF12	SALL1	PVRIG	PLA2G2D	PLCXD2
-	-	-	WIPF1	TCF4	SATB2	PYHIN1	PLAC8	PLEK
-	-	-	XCL1	TENC1	SEC23A	RAB33A	PLCG2	PML
-	-	-	YME1L1	TGFB2	Sep.11	RAC2	PLEK	POU2AF1
-	-	-	ZAP70	THBS2	SERPINE1	RASAL3	PNOC	POU2F2
-	-	-	ZBED2	THSD7A	SERPINF1	RASGRP1	POU2AF1	PPP1R16B
-	-	-	ZBP1	THY1	SFRP2	RASSF5	POU2F2	PRF1
-	-	-	ZC3H12D	TIE1	SFRP4	RCS1	PPP1R16B	PRKCB
-	-	-	ZNF831	TIMP2	SGCD	RGL4	PPP2R5C	PRKCQ
-	-	-	-	TLL1	SH3D19	RHOE	PRF1	PSTPIP1
-	-	-	-	TMEM200A	SH3PYD2A	RHOH	PRKCB	PTCSC1
-	-	-	-	TMEM204	SHOX2	RUNX3	PRKCQ	PTGER4
-	-	-	-	TMEM37	SLIT2	SAMD3	PRO1768	PTPN22
-	-	-	-	TMEM47	SLIT2-IT1	SAMD9L	PSMB9	PTPN6
-	-	-	-	TNS1	SLIT3	SAMSN1	PSTPIP1	PTPN7
-	-	-	-	TPM2	SNED1	SASH3	PTCSC1	PTPRC
-	-	-	-	TPST1	SNX29	SEL1L3	PTPN22	PTPRCAP
-	-	-	-	TRPC1	SPARCL1	SELL	PTPN7	PVRIG
-	-	-	-	TSEN54	SPATS2L	SEMA4D	PTPRC	PYHIN1
-	-	-	-	TSN3	SPOCK1	SH2D1A	PVRIG	PZP
-	-	-	-	TSPAN18	SPON1	SH2D2A	PYHIN1	RAB37
-	-	-	-	TUBB6	SRPX	SIRPG	RAC2	RAC2
-	-	-	-	TWIST2	SRPX2	SIT1	RASSF5	RASAL3
-	-	-	-	VCAN	SRXN1	SLA	RCS1	RASGRP2
-	-	-	-	VGLL3	SSC5D	SLA2	RUNX3	RASSF5
-	-	-	-	VIM	ST6GAL2	SLAMF1	SAMD3	RCS1
-	-	-	-	VSTM4	STON1	SLAMF6	SAMD9L	RGL4
-	-	-	-	WISP1	SULF1	SLAMF7	SAMHD1	RNF166
-	-	-	-	WWTR1	SYNE1	SLC15A2	SAMSN1	RPS15
-	-	-	-	ZCCHC24	SYNPO	SLC25A53	SASH3	RUNX3
-	-	-	-	ZEB1	SYT11	SMCHD1	SEL1L3	SAMD3
-	-	-	-	ZEB2	TAGLN	SNX20	SELL	SAMD9L
-	-	-	-	ZFH4	TBX15	SPB1	SELPLG	SAMHD1
-	-	-	-	ZFH4-AS1	TBX18	SPOCK2	SEMA4D	SAMSN1
-	-	-	-	ZFPM2	TBX2	STAMBPL1	SERPINF9	SASH3
-	-	-	-	ZNF423	TBX5	STAP1	SH2D1A	SCML4
-	-	-	-	ZNF521	TBX5-AS1	STAT4	SH2D2A	SEL1L3
-	-	-	-	ZNF781	TCF4	STK17A	SIRPG	SELL
-	-	-	-	-	TGFB11	STK17B	SLA	SELP
-	-	-	-	-	THBS2	STK4	SLA2	SELPLG
-	-	-	-	-	THSD7A	SYNRG	SLAMF1	SEMA4D
-	-	-	-	-	THY1	TAGAP	SLAMF6	SH2D1A

Supplementary Table 1: Regulons of the transcription factors inferred via the ARACNE algorithm in the GSE32646 “clinical response” dataset.

SYMBOL		
ARHGAP9	GNLY	NKG7
BCL11B	GZMA	P2RX5
BTLA	GZMK	P2RY10
CCL5	ICOS	P2RY8
CCR2	IFNG	PLAC8
CD2	IGHM	POU2AF1
CD247	IGKC	PPP1R16B
CD28	IL18RAP	PRF1
CD3D	IL21R	PTPRC
CD48	IL2RB	PVRIG
CD52	IL2RG	RUNX3
CD69	IL7R	SAMD3
CD8A	INPP5D	SELL
CD96	ITK	SPOCK2
CORO1A	KLRD1	STAT4
CRTAM	LCK	TBC1D10C
CTLA4	LTB	TRAC
CXCR6	LY9	TRAT1
EOMES	MAP4K1	XCL1
FCRL3	MS4A1	ZBED2

Supplementary Table 2: 60 genes from the “immune response transcription network” with the greater up-regulation in the Wicha-bCSC/CC dataset.