

UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM
BIOINFORMÁTICA

SHEYLA TREFFLICH

**ORIGEM EVOLUTIVA DE UNIDADES REGULATÓRIAS DE
EUCARIOTOS: QUEM SURTIU PRIMEIRO, REGULADORES OU
REGULADOS?**

BELO HORIZONTE

2022

SHEYLA TREFFLICH

**ORIGEM EVOLUTIVA DE UNIDADES REGULATÓRIAS DE
EUCARIOTOS: QUEM SURTIU PRIMEIRO, REGULADORES OU
REGULADOS?**

Tese apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para obtenção do grau de "Doutora em Bioinformática".

Orientador: Prof. Dr. Mauro A. A. Castro

Co-orientador: Prof. Dr. José Miguel Ortega

BELO HORIZONTE

2022

043

Trefflich, Sheyla.

Origem evolutiva de unidades regulatórias de eucariotos: quem surgiu primeiro, reguladores ou regulados? [manuscrito] / Sheyla Trefflich. -2022. 105 f. : il. ; 29,5 cm.

Orientador Prof. Dr. Mauro A. A. Castro. Coorientador Prof. Dr. José Miguel Ortega.

Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Regulon. 3. Eucariotos. 4. Ativação Transcricional. I. Castro, Mauro Antônio Alves. II. Ortega, José Miguel. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004



UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

PARECER Nº 29/2021
PROCESSO Nº 23072.250215/2021-64

FOLHA DE APROVAÇÃO

"ORIGEM EVOLUTIVA DE UNIDADES REGULATÓRIAS DE EUCARIOTOS: QUEM SURTIU PRIMEIRO, REGULADORES OU REGULADOS?"

Sheyla Trefflich

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Mauro Antônio Alves Castro - Orientador
Universidade Federal do Paraná

Prof. José Miguel Ortega - Coorientador
Universidade Federal de Minas Gerais

Profa. Glória Regina Franco
Universidade Federal de Minas Gerais

Prof. Rodrigo Juliani Siqueira Dalmolin
Universidade Federal do Rio Grande do Norte

Profa Daniela Fiori Gradia
Universidade Federal do Paraná

Prof. Fábio Fernandes da Rocha Vicente
Universidade Tecnológica Federal do Paraná

Belo Horizonte, 06 de outubro de 2021.



Documento assinado eletronicamente por **Fábio Fernandes da Rocha Vicente, Usuário Externo**, em 06/10/2021, às 12:43, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Mauro Antonio Alves Castro, Usuário Externo**, em 06/10/2021, às 13:14, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Rodrigo Juliani Siqueira Dalmolin, Usuário Externo**, em 06/10/2021, às 14:53, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Daniela Fiori Gradia, Usuário Externo**, em 06/10/2021, às 15:50, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Jose Miguel Ortega, Servidor(a)**, em 13/10/2021, às 17:37, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gloria Regina Franco, Professora do Magistério Superior**, em 18/11/2021, às 15:37, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0983822** e o código CRC **25EE6B98**.

Dedico este trabalho aos meus pais, ao meu esposo e ao meu filho.

”segue o teu destino
rega as tuas plantas
ama as tuas rosas
o resto é a sombra
de árvores alheias”

Fernando Pessoa

AGRADECIMENTOS

Ao Prof. Dr. Mauro A. A. Castro pela orientação acessível, pela oportunidade do trabalho conjunto e pelo amplo diálogo. O cuidado minucioso com o trabalho que desenvolve é exemplar.

Ao Programa de Pós Graduação em Bioinformática da Universidade Federal do Paraná que foi a minha segunda casa desde o início do meu mestrado.

Ao laboratório de Biologia de Sistemas da UFPR, onde esse trabalho foi desenvolvido, em especial aos alunos: Kelin G. de Oliveira, Luis E. A. Rizzardi, Jéssica Magno, Jean S. S. Resende.

Aos professores Dieval Guizelini, Roberto T. Raitz e Maria Berenice R. Steffens (*in memoriam*) e Jeroniza Marchaukoski, pelo bom relacionamento e por todos os ensinamentos. À secretaria Suzana Gobetti pela eficiência, pelo auxílio com os assuntos acadêmicos, pelo apoio e pela amizade sincera.

Aos alunos que passaram pela Bioinformática com os quais eu tive a oportunidade de conviver nesses últimos 7 anos, , entre eles: Diônata Augusto, Guilherme T. Ribas, Bruno Nichio e Nilson A. R. Coimbra. Agradecimento especial para Aniele Langowski, Rodrigo Langowski e Mariane G. Kulik pela sintonia, cumplicidade, atenção, auxílio, boas conversas e a amizade que permanece para além do laboratório.

Ao PPG em Bioinformática da UFMG pela oportunidade em desenvolver o meu doutorado. A CAPES por fomentar essa pesquisa. Ao Prof. Dr. Miguel Ortega pela coorientação e acolhimento em seu laboratório no meu período em Belo Horizonte. Aos queridos e eficientes secretários Sheila C. M. M. Santana e Tiago S. Ramos sempre disponíveis e hospitaleiros durante o meu período em Belo Horizonte, além de serem extremamente acessíveis e eficientes na resolução de todas as demandas acadêmicas que tive nesses 5 anos de doutorado. Aos amigos que a UFMG me presenteou, Wylerson Nogueira, Edson M. de Andrade, Ana Paula Abreu e Carolina Guimarães Rosa.

À minha família, agradeço pelo apoio, dedicação e o afeto incondicionais. Agradecimento especial à minha mãe Yara, à minha irmã Kelly e ao meu cunhado Paulo por terem cuidado com tanto carinho do meu bebê nessa reta final de escrita de tese.

Ao meu esposo Anderson P. Scorsato, meu companheiro de caminhada, meu ombro amigo, o apoio integral em todos os meus desafios e agora pai do meu filho. E ao Erick por ter entendido esse tempo recente longe da mamãe, seu olhar e seu sorriso são o maior combustível e a mais bela demonstração de amor que a vida pode me presentear.

RESUMO

Palavras-chave: Regulons, Redes regulatórias, Elementos regulatórios, Característica ancestral.

A investigação funcional de genes, RNAs e proteínas em larga escala demanda novas estratégias computacionais para extração de informação relevante. Em procariotos, genes relacionados a um mesmo efeito biológico estão usualmente localizados em operons. Em eucariotos não existe a mesma organização, sendo em geral o processo de transcrição independente para cada gene. Apesar disso, nestes organismos existem unidades funcionais controladas por fatores de transcrição, denominadas regulons. As relações regulatórias de um regulon podem surgir ao longo do desenvolvimento de um organismo, em determinado tecido, ou no curso do processo evolutivo. Descrever a formação de regulons pode contribuir com o entendimento de como eles atuam em organismos atuais, tais como o controle transcricional em câncer. Neste estudo investigamos o padrão de origem evolutiva de unidades regulatórias de eucariotos para responder a seguinte pergunta: quem surgiu primeiro, reguladores ou regulados? Testamos 3 hipóteses concorrentes para explicar a formação de regulons: i) fatores de transcrição e seus alvos surgiram de maneira independente; ii) fatores de transcrição surgiram previamente aos seus alvos; iii) fatores de transcrição surgiram posteriormente ao surgimento de seus alvos. Inferimos regulons em câncer de mama e desenvolvemos um método para estimar a distância evolutiva entre reguladores e regulados, inferindo o enraizamento evolutivo em uma dada árvore de espécies. Foram avaliadas 307 regulons, das quais 76 (24,7%) tinham reguladores enraizados junto com seus alvos, 137 (44,3%) tinham reguladores enraizados antes de seus alvos e 94 (30,6%) tinham reguladores enraizados depois de seus alvos. Esses resultados sugerem cenários evolutivos que são consistentes com as três hipóteses apresentadas neste estudo. Em seguida, avaliamos a significância dessas observações e descobrimos que a distribuição geral das raízes evolutivas inferidas dos reguladores precede as raízes evolutivas dos alvos (p -valor = $1e-6$, teste de Wilcoxon-Mann-Whitney). Estes resultados sugerem que regulons não somente são organizados ao redor dos reguladores, como foram formados à partir dos reguladores. Além disso, a identificação de diferentes histórias evolutivas oferece a oportunidade de explicarmos aspectos funcionais tais como o padrão de enraizamento evolutivo em uma rede regulatória. Observamos 4 padrões marcantes: (i) em geral, regulons enraizados em pontos evolutivos próximos se agrupam; (ii) a maioria dos regulons tem enraizamento evolutivo em (LCA) de organismos unicelulares; (iii) regulons associados ao desenvolvimento do câncer de mama são mais recentes; (iv) a maioria dos regulons relacionados a tumores estrogênio positivo e estrogênio negativo estão enraizados em LCA de metazoários. Estes resultados são consistentes com um dos principais aspectos do câncer, em que só é possível observar desorganização tecidual em organismos que possuem um processo de diferenciação celular capaz de formar tecidos.

ABSTRACT

Key words: Regulons, Regulatory network, Regulatory element, Ancestral character.

High-throughput functional analysis of genes, RNAs and proteins demands new computational strategies in order to extract relevant regulatory information. In prokaryotes, genes related to the same biological functions are usually located in operons. In eukaryotes, however, the transcriptional process is organized in a different way, as each gene is independently transcribed. Despite these differences, functional units controlled by transcription factors are present in eukaryotes, which are called regulons, and consist of a set of genes whose activation or repression are under the control of the same transcription factor. The regulatory relations of a regulon can emerge along the development of an organism, in a certain tissue, or even during the evolution process. Describing the formation of regulons may contribute to the understanding of how they act in the extant organisms, such as transcriptional control in cancer. In this research we investigated the evolutionary patterns of eukaryotic regulatory units in order to address the following question: who came first, regulators or regulated targets? Three hypotheses were tested: i) transcription factors and its targets appeared independently; ii) transcription factors came prior to their targets; iii) transcription factors came after their targets. We reconstructed regulons from gene expression data and developed a method to estimate the evolutionary distance between regulators and targets, inferring the point of emergence in a given species tree. A total of 307 regulatory units were evaluated, of which 76 (24.7%) had regulators rooted along with their targets, 137 (44.3%) had regulators rooted before their targets, and 94 (30.6%) had regulators rooted after their targets. These results suggest evolutionary scenarios that are consistent with the three hypotheses stated in this study. We then assessed the significance of these observations and found that the overall distribution of the inferred evolutionary roots of the regulators precedes the evolutionary roots of the targets (p -value = $1e-6$, Wilcoxon-Mann-Whitney test). In addition, the identification of different evolutionary scenarios offers the opportunity to explain functional aspects found in the inferred regulons for breast cancer. Using a metric that estimates the functional similarity between regulons, regulatory units were clustered according to a regulatory network, and onto this regulatory network we mapped the evolutionary roots. Four important patterns were observed: (I) in general, regulons rooted at near evolutionary distances cluster to each other in the regulatory network; (ii) most of the regulons are rooted at the LCA of unicellular organisms; (iii) regulons associated with the development of breast cancer are more recent; (iv) most of the regulons related with positive/negative estrogen tumors are rooted at the LCA of metazoans. These results are consistent with one of the main aspects of cancer, in which tissue disarrangement is only possible in organisms able to form tissues.

LISTA DE FIGURAS

FIGURA 1	– Desenho esquemático de mama normal e ductos tumorais	17
FIGURA 2	– Desenho esquemático da classificação histopatológica para tumores de mama.	18
FIGURA 3	– Classificação histológica de Nottingham.	19
FIGURA 4	– Características das células tumorais	21
FIGURA 5	– Modelo da cascata metastásica em Câncer de Mama	22
FIGURA 6	– Grafo representando regulon	23
FIGURA 7	– Relação de ortologia	24
FIGURA 8	– Sítio de Ligação do Fator de Transcrição	25
FIGURA 9	– Desenho esquemático da reconstrução de redes regulatórias transcricionais	31
FIGURA 10	– Pacote geneplast.data.string.v91	32
FIGURA 11	– Possíveis cenários de enraizamento evolutivo de genes ortólogos e parálogos em uma determinada árvore de espécies.	33
FIGURA 12	– Árvore de espécies utilizada no estudo	37
FIGURA 13	– Árvore de espécies com enraizamento evolutivo	38
FIGURA 14	– Plot de violino para amostras normais	53
FIGURA 15	– Fluxograma do estudo	55
FIGURA 16	– Rede Regulatória Transcricional	57
FIGURA 17	– Cenário evolutivo hipotético para a formação da Rede Regulatória Transcricional	59
FIGURA 18	– Rede Regulatória Transcricional	60
FIGURA 19	– Comparação entre o ponto de enraizamento evolutivo de Tfs, alvos e TcoFs	62
FIGURA 20	– Comparação entre os três grupos de regulons	63
FIGURA 21	– Comparação entre os três grupos de regulons num box plot detalhado	63
FIGURA 22	– Abundância, Diversidade e Plasticidade	64
FIGURA 23	– Grafo de três regulons com informação do ponto de enraizamento evolutivo	65
FIGURA 24	– Representação da árvore de enraizamento evolutivo com uma Tree and Leaf.	67

LISTA DE SIGLAS

LCA	Last Common Ancestor
TCGA	The Cancer Genome Atlas
TRN	redes regulatórias transcricionais
DNA	Deoxyribonucleic Acid
SNP	Polimorfismo de Nucleotídeo Único
ER+	Receptores de Estrogênio Positiva
ER-	Receptores de Estrogênio Negativo
TF	Fator de Transcrição
LCA	Last Common Ancestor
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MI	Informação Mútua
RTN	Reconstruction of Transcriptional networks
lncRNA	Long Non Coding RNA

SUMÁRIO

1	INTRODUÇÃO	14
1.1	BIOLOGIA DE SISTEMAS E SUAS CONTRIBUIÇÕES	15
1.2	PROGRESSÃO DO CÂNCER DE MAMA	16
1.2.1	Subtipos de Câncer de Mama	17
1.3	PROGRESSÃO TUMORAL	20
1.4	MUTAÇÃO E SELEÇÃO CLONAL	22
1.5	UNIDADES REGULATÓRIAS TRANSCRICIONAIS E EVOLUÇÃO	23
1.6	REGULONS E REDES REGULATÓRIAS TRANSCRICIONAIS	25
2	OBJETIVOS	28
2.1	OBJETIVOS GERAIS	28
2.2	OBJETIVOS ESPECÍFICOS	28
3	JUSTIFICATIVA	29
4	MATERIAL E MÉTODOS	30
4.1	AMBIENTE DE PROGRAMAÇÃO	30
4.2	DADOS DE EXPRESSÃO GÊNICA	30
4.3	REDES REGULATÓRIAS EM CÂNCER DE MAMA	31
4.4	DADOS DE ORTOLOGIA/ANOTAÇÃO PARA TRN	32
4.5	INFERÊNCIA DE ENRAIZAMENTO EVOLUTIVO	33
4.5.1	Representação de Redes em grafos	33
4.5.1.1	Mapear informações de enraizamento em redes regulatórias	34
4.5.2	TreeAndLeaf	36
4.6	ANÁLISE EVOLUTIVA	39
4.6.1	Explorando Raízes Evolutivas de Regulons	39
4.7	FLUXO DO ESTUDO	53
5	RESULTADOS E DISCUSSÃO	56
6	CONCLUSÃO	69
7	PERSPECTIVAS	71
	REFERÊNCIAS	72
	Anexo A – LISTA DE PRODUÇÕES	80
	Anexo B – ARTIGO PUBLICADO - (Trefflich <i>et al.</i>, 2019)	81
	Anexo C – VINHETA DO PACOTE GENEPLAST	87
	Anexo D – ARTIGO PUBLICADO - (Corces <i>et al.</i>, 2018)	98
	Anexo E – ARTIGO PUBLICADO - (Chagas <i>et al.</i>, 2019)	101
	Anexo F – ARTIGO PUBLICADO - (Mathias <i>et al.</i>, 2021)	103
	Anexo G – ARTIGO SUBMETIDO - (Cardoso <i>et al.</i>, 2021)	105
	Anexo H – CAPÍTULO DE LIVRO PUBLICADO - (Cruz <i>et al.</i>, 2017)	107

1 INTRODUÇÃO

A Biologia de Sistemas é uma área do conhecimento que vem crescendo com o desenvolvimento de novas tecnologias, permitindo a análise e interpretação de dados em larga escala (Aggarwal e Lee, 2003). O termo "biologia de sistemas" foi criado quando o estudo da biologia demandou esforços multidisciplinares no entendimento dos cenários biológicos, ganhando força com o Projeto Genoma Humano (Likic *et al.*, 2010; Westerhoff e Palsson, 2004). Ao longo do tempo a biologia de sistemas tem permitido a criação de novas áreas que geram contribuição científica focadas em torno do entendimento fundamental dos sistemas biológicos por meio das interações moleculares (Likic *et al.*, 2010)

Com a crescente aquisição de dados biológicos decorrente do desenvolvimento de novas tecnologias, entre eles dados genéticos, estruturais e transcricionais houve uma impulsão no desenvolvimento de ferramentas capazes de realizar análises automáticas, mais rápidas e com maior eficácia (Gaasterland e Sensen, 1996; Fleischmann *et al.*, 1999). Existem diversas abordagens em biologia de sistemas baseadas em modelagens matemáticas e computacionais, porém o desenvolvimento e a manutenção de ferramentas é tarefa desafiadora, como o que ocorre com diversos repositórios especializados na manutenção de dados biológicos, ou repositórios de ferramentas de análise, simulação e visualização (Likic *et al.*, 2010).

O estudo de doenças complexas requer a constante criação de ferramentas capazes de investigar e interpretar padrões em dados ortogonais compostos com várias camadas informativas (Ryan *et al.*, 2013), como os dados provenientes do sequenciamento de tumores de populações que compartilham um mesmo fenótipo. As pesquisas do câncer têm sido muito beneficiadas com o surgimento de bancos de dados relacionados a grandes consórcios internacionais, especializados na coleção e manutenção de dados com grande acurácia. Por exemplo o The Cancer Genome Atlas (TCGA), a maior coleção de fenótipos clínicos e moleculares de mais de 10.000 pacientes com 33 tipos tumorais (Weinstein *et al.*, 2013). Existem também repositórios especializados em apenas um tipo tumoral, como o METABRIC, utilizado nesta abordagem, que contém apenas dados de pacientes com câncer de mama (Curtis *et al.*, 2012)

O câncer de mama é um tipo de câncer de alta prevalência e os estudos multidimensionais, tem contribuído continuamente para uma abordagem mais personalizada

com aumento na eficácia do tratamento e melhora na sobrevivência dos pacientes (Torre *et al.*, 2015). A criação de métodos analíticos que considerem os aspectos evolutivos do câncer, somado às abordagens dos estudos regulatórios já existentes podem enriquecer o entendimento dos mecanismos de instalação, desenvolvimento da doença e resistência terapêutica que alguns pacientes manifestam.

1.1 BIOLOGIA DE SISTEMAS E SUAS CONTRIBUIÇÕES

O desenvolvimento de tecnologias de sequenciamento foi um marco no estudo da biologia, demandando o desenvolvimento de novas ferramentas computacionais para auxiliar na análise e interpretação de dados biológicos em larga escala (Aggarwal e Lee, 2003). A biologia de sistemas auxilia no entendimento sistemático do genoma funcional, contribuindo, por exemplo, na descoberta de biomarcadores, na classificação de doenças, na descoberta de alvos terapêuticos, desenvolvimento de novas drogas, entre outros (Chen *et al.*, 2009).

A biologia de sistemas desenvolve e utiliza métodos computacionais para o tratamento e análise de dados biológicos na tentativa de elucidar novos mecanismos moleculares (Hillmer, 2015). Um dos desafios da biologia de sistemas é construir maneiras claras de integração e visualização de dados multidimensionais (Gehlenborg *et al.*, 2010). A biologia de sistemas pode auxiliar na construção de modelos *in silico* para estudos regulatórios em organismos complexos, interrogando o efeito combinado de múltiplos fatores que contribuem para um determinado fenótipo. Esse tipo de abordagem de estudo traz uma visão mais completa, quando consideramos que ainda que as moléculas tenham suas funções específicas, elas não atuam sozinhas demandando uma investigação mais aprofundada que considere as particularidades interacionais entre genes, proteínas e demais elementos celulares.

Dados coletados a partir de experimentos moleculares são utilizados de maneira análoga à engenharia reversa, na inferência de hipóteses em biologia de sistemas (Karczewski e Snyder, 2018). Assim, dados de transcriptoma, por exemplo, funcionam como pistas que orientam o entendimento dos processos que ocorrem no núcleo celular, no instante em que o transcriptoma foi feito. Se considerarmos que as células podem estar envolvidas em diferentes etapas do ciclo celular, sejam normais ou tumorais, entre outros fatores, compreendemos o nível de complexidade envolvido nesse tipo de estudo. A integração de dados ortogonais estreitam o campo de busca, por exemplo, elucidando relações entre elementos causadores de uma determinada desordem (Ryan *et al.*, 2013).

O estudo de doenças complexas, como o câncer, por exemplo, requer sucessivos avanços tanto na coleta quanto na interpretação destes dados, em um processo contínuo que conta com a contribuição de diferentes áreas do conhecimento. É necessário considerar que quando utilizamos dados provenientes de coleções de grandes consórcios internos, houve

um esforço intenso na padronização de processos relacionados com a extração, transporte, processamento, armazenamento e manutenção das amostras para que os dados utilizados em análises subsequentes sejam minimamente impactadas por fatores externos.

Nesta tese apresentamos novos métodos computacionais na área de biologia de sistemas, aplicando esses métodos na interpretação de redes regulatórias transcricionais (TRN) do câncer de mama.

1.2 PROGRESSÃO DO CÂNCER DE MAMA

O câncer de mama é o câncer de maior prevalência em mulheres, com exceção do câncer de pele não melanoma (Waks e Winer, 2019), sendo também a principal causa de morte por câncer em mulheres ao redor do mundo (Torre *et al.*, 2015). É uma doença que tem ocorrência multifatorial embora ocorram no mundo todo, em cada país ou continente a incidência, a mortalidade, a taxa de sobrevivência variam consideravelmente (Salamat *et al.*, 2018; Zendehdel *et al.*, 2018; N.Hortobagyi *et al.*, 2005). As técnicas de rastreamento da população são estratégias que reduziriam eficientemente o número de novos casos, mas essas técnicas possuem um elevado custo o que acaba sendo uma desvantagem nesse tipo de abordagem, estando disponível apenas para países que possuem boa condição socioeconômica (Momenimovahed e Salehiniya, 2019). Fatores como demora na conclusão do diagnóstico, bem como início de estratégias terapêuticas são outros fatores que aumentam o índice de mortalidade e diminuem a sobrevivência dos pacientes (Abdulrahman e Rahman, 2012).

Seu desenvolvimento pode estar relacionado à etnia, questões de ordem ambiental, estilo de vida, herança genética entre outros novos fatores que vão sendo relacionados a ocorrência da doença ao longo do tempo (Li *et al.*, 2017). Assim, a sobrevivência, a incidência e a taxa de mortalidade em diferentes países e regiões do mundo são bastante variadas (Momenimovahed e Salehiniya, 2019; Hortobagyi *et al.*, 2005). Com o processo de envelhecimento da população dos em desenvolvimento também há um aumento expressivo na ocorrência do câncer de mama, assim como dos outros tipos tumorais (N.Hortobagyi *et al.*, 2005). Importante ressaltar que homens também são acometidos pelo câncer de mama, porém apresentam uma incidência muito menor quando comparados às mulheres (Longo e Giordano, 2018; Giordano *et al.*, 2002). E quando ocorre em homens a prevalência é maior em idosos que sofreram desequilíbrio hormonal e histórico familiar (Abdelwahab, 2017).

Sob uma abordagem anatômica resumida, a mama é um órgão composto tanto por tecido glandular (responsável pela produção lactífera) quanto por tecido adiposo (com função estrutural). A parcela glandular forma os lóbulos que se unem em lobos (**Fig. 1a,b**). Os lobos escoam a produção lactífera através de interconexões de pequenos ductos, que se unem em ductos maiores (**Fig. 1c**), e por fim conduzirão o leite ao meio externo (Zucca-Matthes *et al.*,

2004). Enquanto em homens a mama apresenta uma conformação atrofiada, em mulheres a mama se desenvolve durante a puberdade para que sua funcionalidade seja alcançada durante a maternidade, através da produção de leite (Bernardes, 2010). A mama tem funcionamento regulado pela variação dos hormônios sexuais que tem papel preponderante no desenvolvimento tumoral (Thomas, 1984).

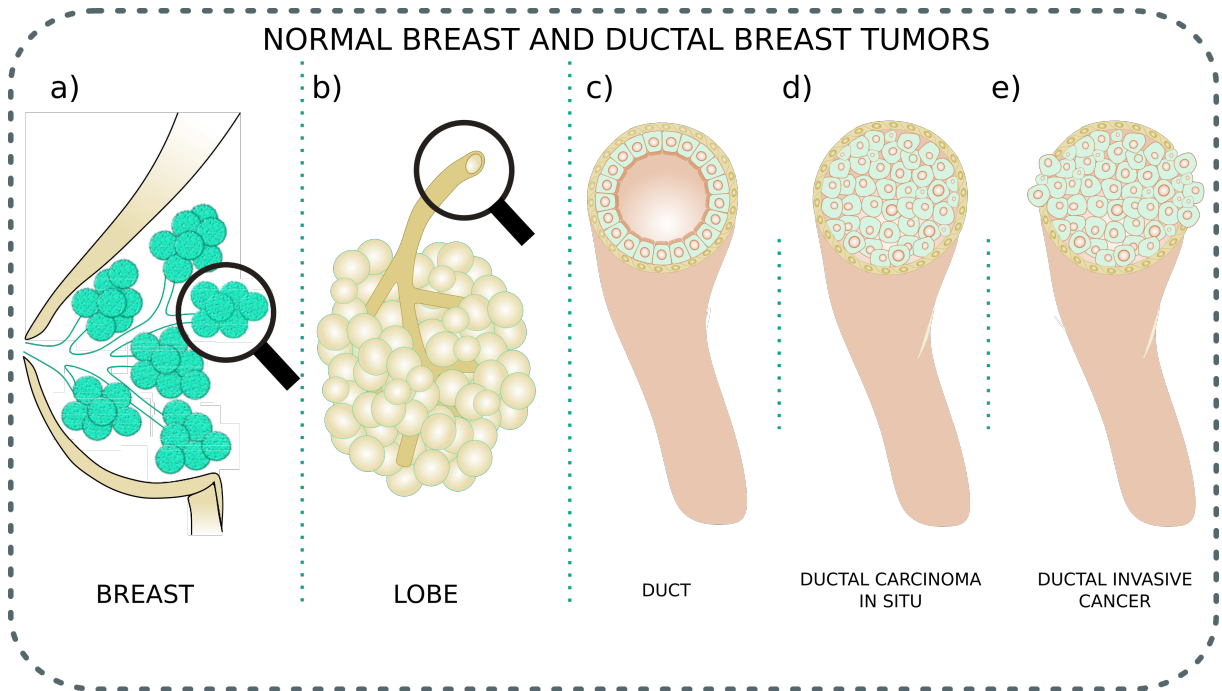


Figura 1. Desenho esquemático de mama normal e ductos tumorais. a) mama normal, b) lobo normal ampliado, c) ducto normal, d) ducto com desenvolvimento tumoral *in situ*, e) ducto com desenvolvimento tumoral invasivo (arte autoral, 2018).

No câncer de mama, o processo de desenvolvimento tumoral passa por estágios definidos, com características que podem distinguir estes estágios. A iniciação do estágio pré-maligno ocorre pelo desenvolvimento de hiperplasia ductal atípica, seguindo de estágio pré-invasivo, ou carcinoma ductal *in situ* (**Fig. 1d**), podendo alcançar um potencial letal com tumor ductal invasivo (**Fig. 1e**).

1.2.1 SUBTIPOS DE CÂNCER DE MAMA

O câncer de mama é uma doença de alta heterogeneidade, com subtipos, características biológicas e padrões clínicos diferentes. A estratificação em subtipos auxilia a escolha de tratamentos com menor toxicidade e com maior efetividade ao paciente. A correta classificação tumoral está intimamente relacionada com a sobrevivência dos pacientes, conferindo melhora no prognóstico. Entre as características utilizadas para a subtipagem de tumores de mama, podemos citar:

- Tamanho do tumor;

- Acometimento linfonoidal;
- Grau histológico;
- Receptores de estrógeno;
- Receptores de progesterona;
- Idade do paciente (Blows *et al.*, 2010; Yersal, 2014)

Pacientes com perfis de subtipagem histopatológica parecidos podem desenvolver características clínicas diferentes, ou mesmo divergente aceitação à terapia. Por essa razão características biológicas do tumor também são frequentemente utilizadas na classificação de subtipos, a exemplo dos perfis de expressão gênica, contribuindo para um melhor seguimento do paciente (Eroles *et al.*, 2012; Rakha e Ellis, 2011).

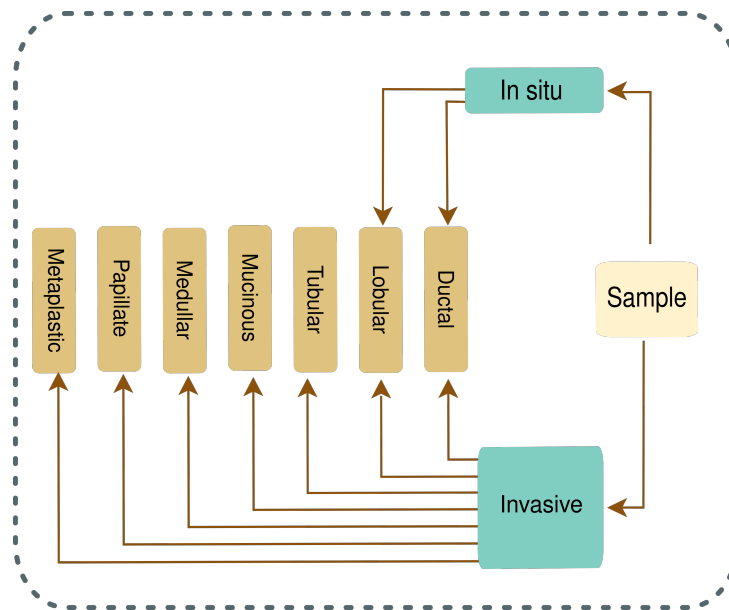


Figura 2. Desenho esquemático da classificação histopatológica para tumores de mama. As amostras tumorais se dividem em duas categorias principais: in situ e invasivo. Posteriormente se determina em qual tipo celular o tumor ocorre (arte autoral, 2018).

Em primeira análise é estabelecido a localização do tumor e o tecido de origem, se está localizado (in situ) ou infiltrado em tecidos adjacentes (invasivo). O tumor mamário pode ter origem em diversos tecidos, mas o mais comum é observarmos tumores ductais e lobulares de origem mamária (American Cancer Society, 2018) (**Fig. 2**). Apesar da classificação histopatológica ainda ser o ponto inicial de investigação e subtipagem, muitas vezes um tumor apresenta características mistas e por essa razão são utilizadas classificações complementares. O Sistema de Gradeamento de Nottingham (**Fig. 3**) é utilizado para estabelecer o grau de desenvolvimento tumoral, ou Índice de Prognóstico de Nottingham (NPI).

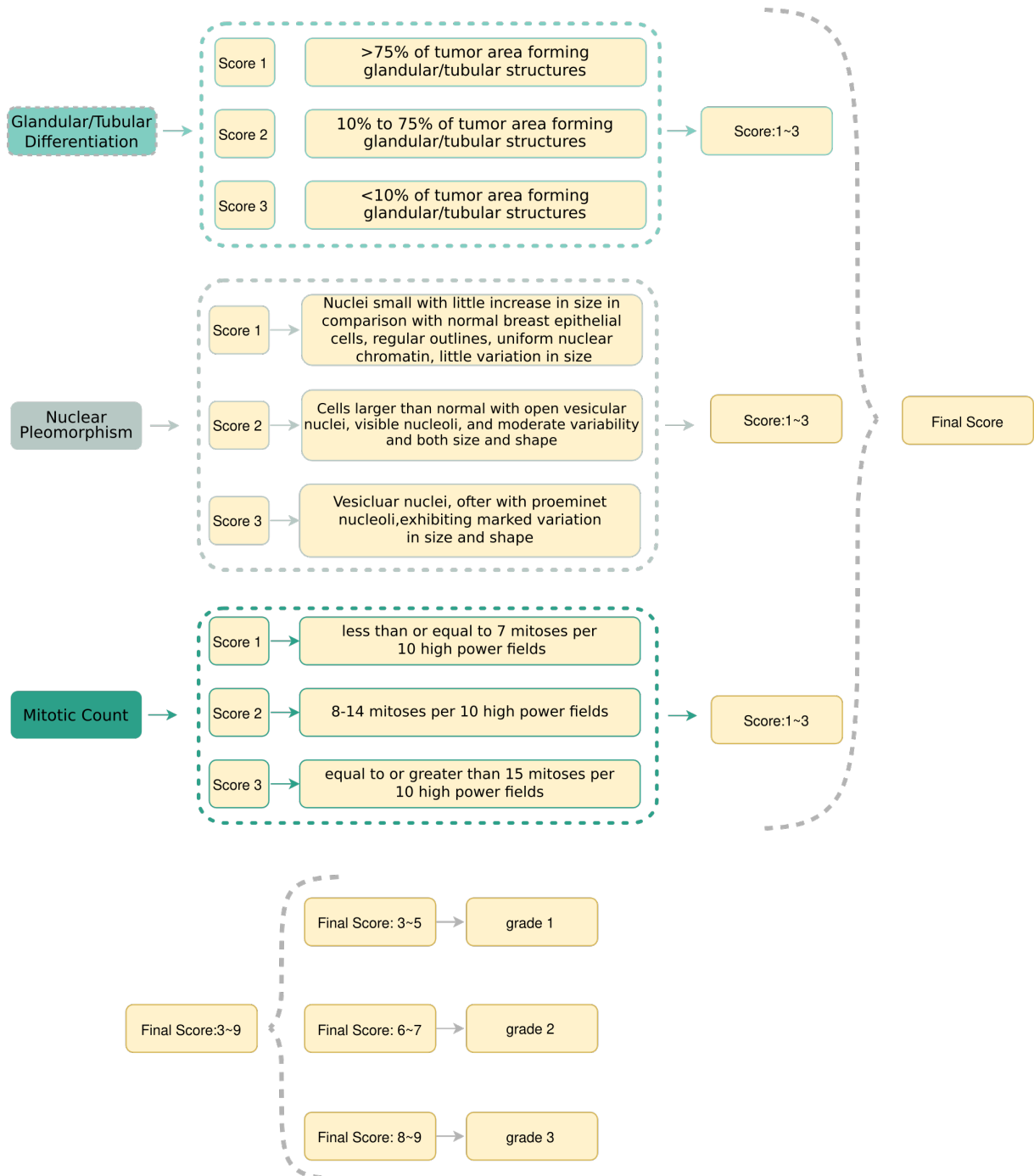


Figura 3. Classificação histológica de Nottingham. Quadro esquemático indicando critérios de pontuação para obtenção do Índice de Prognóstico de Nottingham adaptado de (Rakha e Ellis, 2011)

Para a construção deste índice são considerados: a percentagem do tumor com diferenciação glandular/tubular, características do pleomorfismo nuclear e atividade mitótica de divisão celular. Quanto menor o NPI melhor o prognóstico do paciente. Além disso, a classificação imunohistoquímica de tumores de mama define a qual das três classes seguintes um tumor pertence: receptor de estrógeno (ER) positivo, receptor de progesterona (PR) positivo, e HER2 positivo. Essa classificação tumoral é utilizada em conjunto com a classificação histopatológica e compõem o sistema de classificação mais utilizado em pacientes com câncer de mama (Dai *et al.*, 2015).

Em relação às características biológicas do tumor, estratégias que usam perfis de expressão gênica estabelecem assinaturas para os subtipos tumorais, subdividindo os tumores de mama em 4 categorias: Luminal A, Luminal B, HER2-Enriched, e Basal-Like. Estas categorias são definidas como subtipos intrínsecos pelo fato dos genes escolhidos para defini-las refletirem propriedades intrínsecas dos tumores, a qual se mantém consistente mesmo na comparação entre indivíduos de diferentes etnias (Perou e Borresen-Dale, 2011). Além disso, o estudo PAM50 (Bernard *et al.*, 2009) aprimorou a subtipagem molecular com a construção de uma assinatura derivada da expressão de 50 genes, englobando as mesmas quatro categorias anteriores e adicionando uma: Normal-like.

1.3 PROGRESSÃO TUMORAL

Dependendo do subtipo tumoral, pacientes com tumores primários são tratados com cirurgia, seguido por radioterapia e terapias direcionadas (Petri, 2020). O desenvolvimento de terapias anti-metastásicas é de grande importância quando consideramos que 30% das pacientes com câncer de mama acabam vindo a óbito em decorrência de metástases acometidas às vezes décadas após o tratamento tumoral inicial (Steeg, 2016).

O desenvolvimento do câncer pode ter início com danos no material genético de células normais que, por erros não corrigidos na divisão celular, podem diferenciar-se das células de origem, passando por um processo de transformação tumoral (Thomas, 1984). Em sucessivas divisões, estas células podem acumular grande quantidade de alterações no DNA e ficam menos susceptíveis aos mecanismos de correção.

As células tumorais possuem características bastante marcantes como a manutenção da proliferação celular, alteração no metabolismo energético, resposta anormal ao controle dos sinais de crescimento celular, indução de angiogênese, adesão e diferenciação (**Fig. 4**) (Hanahan e Weinberg, 2011; Trigos *et al.*, 2017). Essas características são complementares quando o desenvolvimento tumoral tem início, auxiliam na manutenção, crescimento e na ocorrência de metástases.

Ao sustentar um crescimento desordenado, as células tumorais originam massas

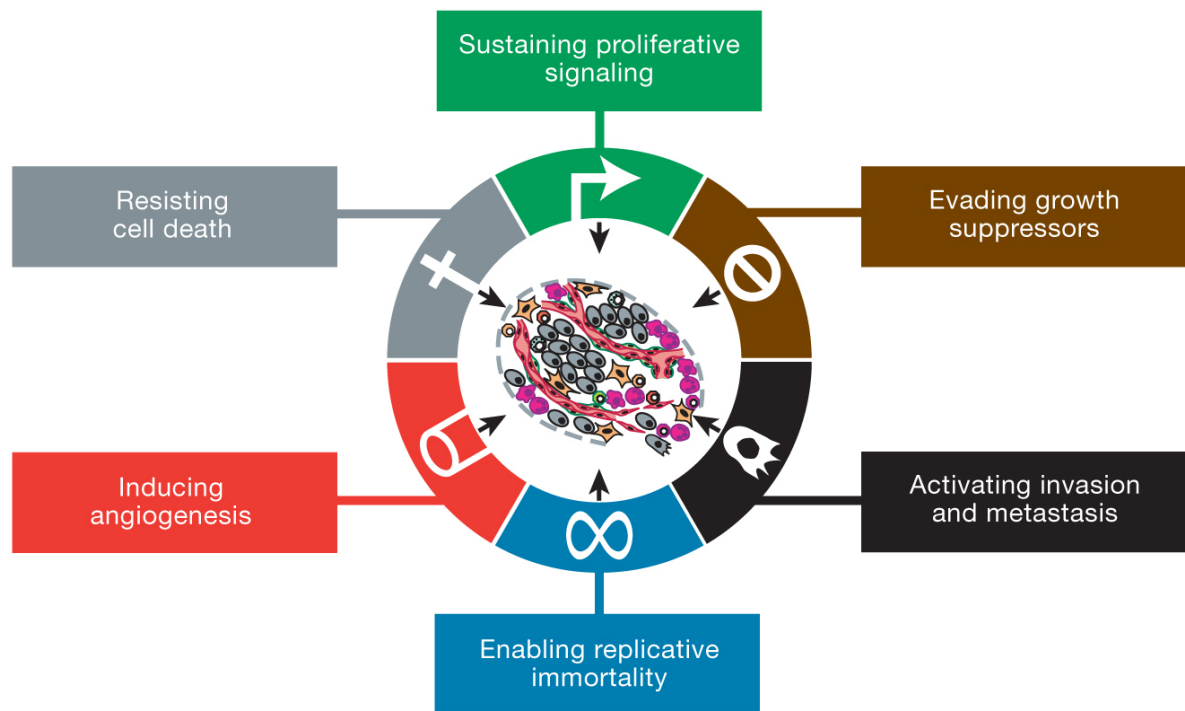


Figura 4. Características das células tumorais. As seis principais características das células tumorais, que conferem instalação e manutenção do tumor, permitindo a proliferação das células e a iniciação dos mecanismos metastásicos (Hanahan e Weinberg, 2011)

celulares amorfas que crescem dentro de um órgão, interferindo no seu correto funcionamento, podendo levar os órgãos atingidos à falência funcional. A massa de células tumorais desenvolve mecanismos próprios para manter a sua sobrevivência, criando um microambiente interno protetor das estratégias de defesa do organismo. Existe uma organização celular na formação tumoral, com eficiente interação de atividade entre múltiplos tipos celulares na configuração do microambiente tumoral (Hanahan e Weinberg, 2011). Apesar da constante criação de novas estratégias terapêuticas, os tumores são formados por células que podem apresentar resistência às terapias aplicadas (Arozarena e Wellbrock, 2017).

A massa tumoral pode ter ainda a capacidade de invadir tecidos adjacentes e extravasar para outras localidades próximas ou distantes do ponto de origem, num processo denominado metástase. O processo metastásico, quando diagnosticado, indica um mau prognóstico (Lee *et al.*, 2010). Na análise comparativa entre o tumor primário e os tumores em metástase se observa a ocorrência de grande diversidade genética, sugerindo que apenas algumas variantes genotípicas pertencentes ao tumor primário terão potencial para desenvolver tumores metastásicos (Varela *et al.*, 2017).

Quando visto sob uma abordagem evolutiva, o câncer é impulsionado por um processo de desdiferenciação de células somáticas que escapam dos mecanismos de controle replicativo originando células tumorais (Crespi e Summers, 2005). Os mecanismos evolutivos estão

presentes na heterogeneidade dos tumores e agem por meio de seleção clonal em vários tipos de cânceres, fato que pode explicar porque alguns pacientes acabam desenvolvendo resistência às estratégias terapêuticas (Lacina *et al.*, 2019).

1.4 MUTAÇÃO E SELEÇÃO CLONAL

A formação do microambiente tumoral é decorrente da proliferação celular e tem papel fundamental no desenvolvimento e instalação da doença. Ao longo do desenvolvimento da doença, no interior do tumor se originam diversos mecanismos que conferem ao conjunto de células cancerosas a capacidade de seguir se instalando no meio onde está, buscar por mais espaço e proliferar (Wang *et al.*, 2017).

Mutações oncogênicas favorecem o desenvolvimento tumoral e conferem capacidade de fixação das células (Hanahan e Weinberg, 2011). A ocorrência de mutações oncogênicas em estágio inicial estimula que novas mutações ocorram, e por essa razão podem ser chamadas de *mutações mutadoras*, as quais podem gerar instabilidade genômica. Em um cenário inicial, este processo leva à perda de *fitness* celular, com aumento de apoptose e diminuição de proliferação, o que estaria intimamente relacionado à seleção clonal negativa de células tumorais. Em contrapartida, quando clones aumentam o *fitness* celular pelo acúmulo de mutações, estes adquirem a capacidade de expansão, num processo denominado seleção clonal positiva (Beckman e Loeb, 2005; Arneth, 2018). Ao longo do tempo as células cancerosas entram numa cascata metastásica, onde ocorre a seleção de células cada vez mais agressivas que se desprendem do tumor inicial, migram para diferentes regiões do organismo e dão origem a novas massas tumorais (**Fig. 5**) (Klein, 2010).

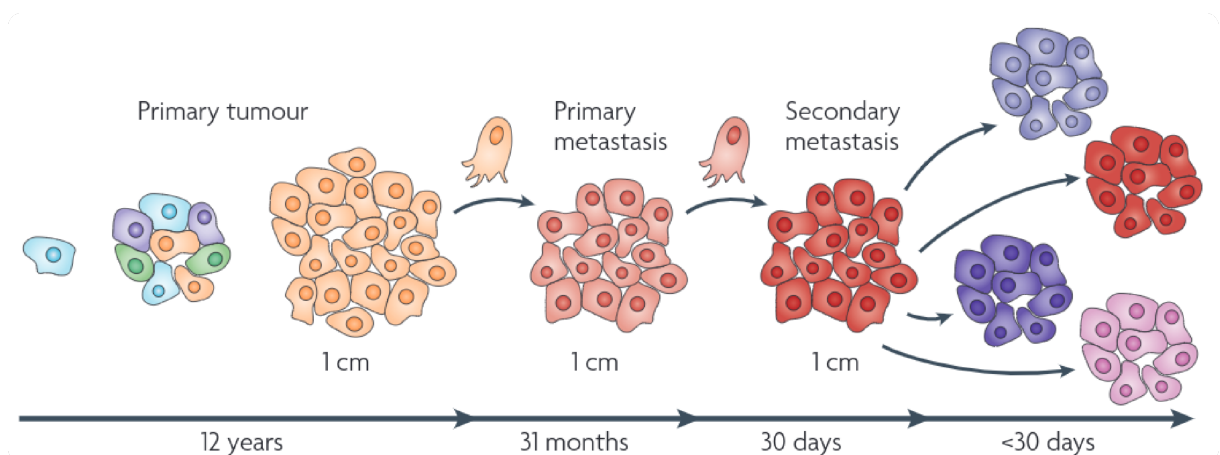


Figura 5. Modelo da cascata metastásica em Câncer de Mama. Seleção de células agressivas selecionadas durante a progressão tumoral e o início da disseminação é iniciada até que ocorra uma sequência final de metástases capazes de levar o paciente a óbito. (Klein, 2010)

O estudo do desenvolvimento tumoral pode auxiliar não só no diagnóstico mas também

na escolha do tratamento a ser utilizado (Ma *et al.*, 2003). Por exemplo, já foram identificadas 36 unidades regulatórias (regulons) enriquecidas com polimorfismos de nucleotídeo único (SNP) associados ao risco de desenvolvimento do câncer de mama (Castro *et al.*, 2015). Estas unidades regulatórias estão subdivididas em dois grupos, Receptores de Estrogênio Positivo (ER+) e Receptores de Estrogênio Negativo (ER-) segundo o padrão de resposta hormonal que é manifestado. Dependendo de quais regulons estão ativos no curso da progressão tumoral, é possível que características genéticas favoreçam o desenvolvimento de tumores responsivos ao tratamento hormonal (Campbell *et al.*, 2018).

1.5 UNIDADES REGULATÓRIAS TRANSCRICIONAIS E EVOLUÇÃO

Um fator de transcrição (TF) juntamente com seus genes alvo formam um regulon, e essa relação pode ser ilustrada na forma de um grafo conforme mostrado na **Figura 6**. O grafo apresentado ilustra um regulon, onde os nós representam elementos do regulon e as arestas representam interações entre os elementos (Carter, 2005). As relações regulatórias entre os elementos que formam um regulon podem surgir em um determinado tecido, ao longo do desenvolvimento de um organismo, ou mesmo serem entendidas em um contexto mais amplo, no curso do processo evolutivo.

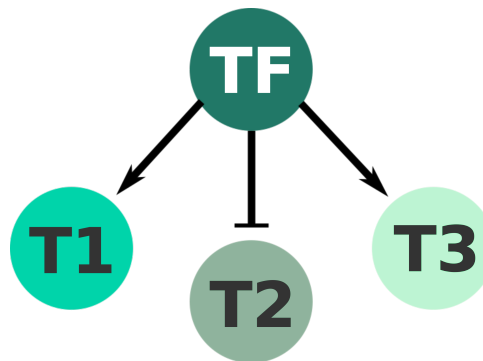


Figura 6. Grafo representando regulon. Ilustração da relação regulatória entre o Fator de Transcrição (TF) e seus genes alvo (T1, T2, T3). Em T1 e T3 ocorre a indução na taxa de expressão gênica e em T2 ocorre a repressão na taxa de expressão gênica.

Durante a história de uma espécie, a vida de um organismo ou o crescimento de uma célula, ocorre a aquisição de características diferenciadas em resposta a diversos fatores. Se essas características são favoráveis elas permanecem e conferem um maior poder adaptativo (Lacina *et al.*, 2019). Em organismos multicelulares, as células são capazes de manifestar alterações adaptativas reversíveis em resposta ao estresse, tais como hipertrofia, inibição da apoptose, displasia e aumento da taxa de divisão celular (Arneth, 2018; Beckman e Loeb, 2005). A metilação e alteração de histonas são exemplos de efeitos epigenéticos decorrentes de processos adaptativos que também estão relacionados ao surgimento de doenças complexas, como o câncer por exemplo (Jaenisch e Bird, 2003).

Além disso, é razoável supor que os genes que compõem um regulon estejam, em algum grau, associados à mesma história evolutiva. É possível que alguns regulons estejam enriquecidos com genes que surgiram em contextos adaptativos semelhantes, ou decorrentes das mesmas pressões seletivas, ou que simplesmente co-ocorrem em unidades regulatórias que foram sendo moldadas no curso da evolução. Apesar de não termos meios para rastrear a história evolutiva das associações entre os genes que integram um regulon, é possível reconstruir as relações de ortologia dos genes que o compõem (Castro *et al.*, 2008).

Por definição, ortólogos são genes que pertencem a espécies diferentes, derivados por especiação, mas que partilham um ancestral comum (Fig. 7) (Gabaldón e Koonin, 2013; Koonin, 2005). Genes ortólogos podem ou não adquirir novas funções ao longo do tempo (Sonnhammer e Koonin, 2002). O estudo da relação entre genes ortólogos é indispensável no entendimento da relação evolutiva entre espécies, bem como para estabelecer a diferença entre genomas de diferentes organismos (Gabaldón e Koonin, 2013).

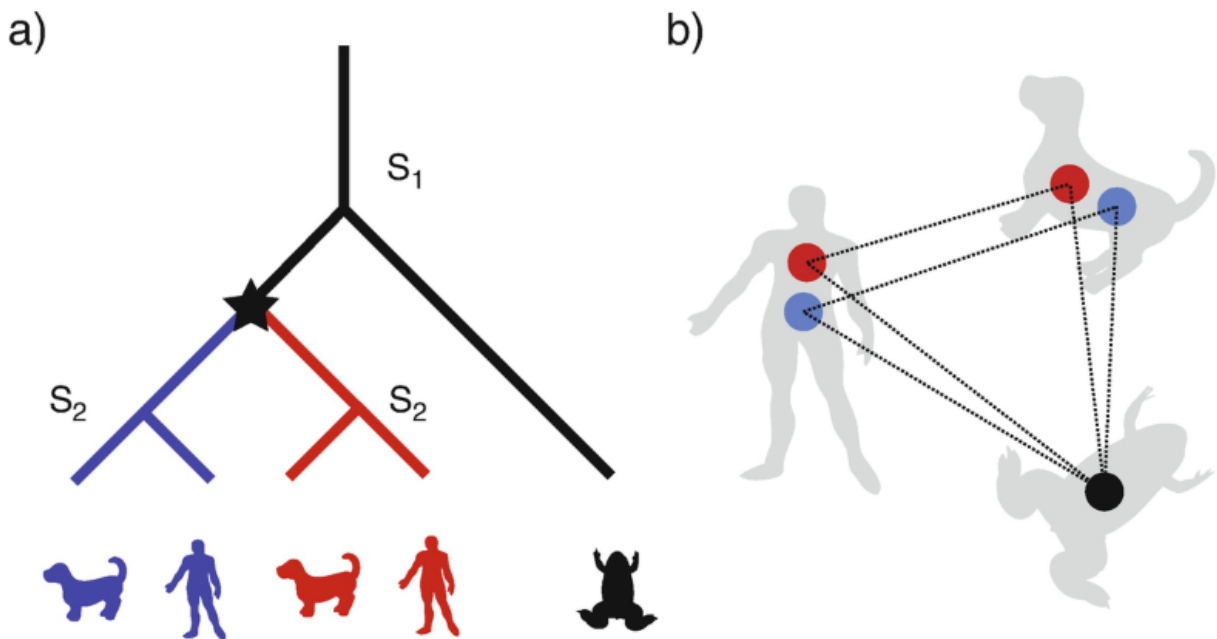


Figura 7. Relação de ortologia. a) Cladograma com dois eventos de especiação que ocorrem em S₁ e S₂, e um evento de duplicação sinalizado pela estrela. b) Grafo representativo da família de ortólogos presentes no cladograma que se originam em sapo e são compartilhados por humanos e caninos, cada um contendo duas cópias do ortólogo. O gene de sapo é ortólogo de todos os demais genes. Os dois genes encontrados em humanos são parálogos entre si, assim como os genes encontrados em caninos. A dupla de genes humanos é coortólogo da dupla de genes encontrados em caninos. (Altenhoff *et al.*, 2019)

Dada uma árvore de espécies, é possível inferir o ponto de surgimento de um gene nesta árvore, representado idealmente pelo último ancestral comum (LCA) de um grupo monofilético em que todos os organismos atuais possuem ao menos uma cópia do ortólogo (Dalmolin e Castro, 2015). Atualmente existem ferramentas computacionais que permitem avaliar as relações de ortologia de um grande número de genes de um organismo (Dalmolin e Castro, 2015), mas não está claro se é possível identificar padrões evolutivos entre grupos de genes que formam as

unidades regulatórias.

A diferença dos perfis regulatórios entre espécies está relacionada, em algum grau, com a distribuição dos sítios de ligação ao DNA de elementos regulatórios, e com o ganho ou perda de regiões cis-reguladoras (Cheng *et al.*, 2014; Arnold *et al.*, 2014; Baker *et al.*, 2012). Na próxima seção será apresentado de que forma a distribuição destes sítios de ligação de elementos regulatórios pode resultar em ganho ou perda de funcionalidade em redes regulatórias (Nocedal *et al.*, 2017).

1.6 REGULONS E REDES REGULATÓRIAS TRANSCRICIONAIS

Existem cerca de 20.000 genes codificantes no genoma humano (Salzberg, 2018) e apesar de todos serem potencialmente informativos quando buscamos entender algum aspecto do funcionamento celular, ainda é desafiador estudar o efeito combinado de genes devido ao grande número de hipóteses a serem testadas.

Estratégias de estudo que utilizam redes regulatórias transcricionais são interessantes de serem exploradas por reduzem o espaço de busca ao focar em gargalos regulatórios (Lefebvre *et al.*, 2012). Ou seja, uma vez que redes regulatórias são centradas em *hubs* transcricionais, podemos estudar por meio destes reguladores o efeito combinado de vários genes (Wang *et al.*, 2009).

Em organismos eucariotos o processo de transcrição gênica pode ser controlado por unidades funcionais reguladas por fatores de transcrição (Touchon e Rocha, 2016). TFs são moduladores da expressão gênica, envolvidos nos mecanismos de regulação pré-transcricional e estão ativos em todos os organismos eucariotos (Yusuf *et al.*, 2012; Shi *et al.*, 2019).

O processo de transcrição de um gene pode ser controlado pelo funcionamento integrado de vários elementos cis-reguladores, tanto de ação proximal, quanto mais distal da região de início da transcrição (Spitz e Furlong, 2012). TFs são um desses elementos e se ligam às regiões cis-reguladoras do gene alvo, proximais ou distais, ativando ou reprimindo sua expressão (**Fig.8**) (Calhoun *et al.*, 2002; Shi *et al.*, 2019).



Figura 8. Sítio de Ligação do Fator de Transcrição. Local de posicionamento do sítio de ligação do Fator de Transcrição de um determinado gene alvo.

Com ampla gama de funcionalidades, os TFs estão envolvidos nos mais variados processos celulares e em diferentes estágios do desenvolvimento de um organismo. Assim,

TFs desempenham papel importante também no desenvolvimento tumoral (Yusuf *et al.*, 2012).

Existem cerca de 1600 TFs anotados no genoma humano e para cada um deles é possível construir um regulon (Lambert *et al.*, 2018). As unidades regulatórias em eucariotos são usualmente inferidas por algoritmos que avaliam a expressão do regulador e de seus potenciais alvos em um conjunto de amostras.

A interação entre os elementos que formam um regulon depende tanto da sequência específica do DNA para a ligação do TF, quanto de fatores epigenéticos, como a acessibilidade da cromatina (Corces *et al.*, 2018). Os TFs interagem com os nucleossomos (unidade estrutural da cromatina) que podem inibir a sua ligação por oclusão dos sítios de ligação (Coux *et al.*, 2020). A cromatina pode mudar a sua conformação física por ação dos remodeladores da cromatina, que facilitam o acesso às regiões de ligação por onde os fatores de transcrição interagem com as sequências de DNA (Tsankov *et al.*, 2015; Coux *et al.*, 2020).

A ação dos TFs é mediada por moléculas modeladoras que interagem com os nucleossomos (unidade estrutural da cromatina) (Zhu *et al.*, 2018). Essas moléculas atuam na conformação física da cromatina, ocluindo ou expondo regiões de sequências que são reconhecidas pelos TFs (Brahma e Henikoff, 2020). Dessa maneira, os fatores de transcrição se ligariam a regiões específicas do DNA quando o nucleossomo muda a sua conformação transitoriamente, e a partir daí os TFs tem o potencial de remodelar a cromatina, desenrolando regiões do DNA que precisam estar acessíveis para a sua interação, orquestrando a transcrição (Tsankov *et al.*, 2015; Brahma e Henikoff, 2020).

Todos os tecidos desempenham um conjunto de funções básicas correlatas, mas é por meio de um programa próprio de expressão gênica que as células manifestam características relacionadas às especialidades do tecido ao qual pertencem (Sonawane *et al.*, 2017; Ko *et al.*, 2017). Em cada tipo celular, a cromatina tem uma conformação topológica particular, tornando os sítios de ligação mais ou menos acessíveis (Schmitt *et al.*, 2016). Assim, a interação de um TF com seus alvos é específica do tecido, e com maior especificidade nos tecidos que expõem os sítios de ligação (Sonawane *et al.*, 2017).

Os fatores de transcrição são altamente conservados e reconhecem as mesmas sequências de DNA em espécies filogeneticamente distantes (Nitta *et al.*, 2015; Kreft *et al.*, 2017). A diversidade regulatória entre espécies se dá pela perda ou ganho de regiões cis-reguladoras (Baker *et al.*, 2012). Entre espécies também se observa uma variação na distribuição de sítios de ligação dos fatores de transcrição ao longo do genoma (Cheng *et al.*, 2014; Arnold *et al.*, 2014). A incorporação de novas funcionalidades em uma rede regulatória transcricional pode ocorrer com a mudança na distribuição dos sítios de ligação dos TFs, sem que seja necessário modificar o motivo de ligação ou a estrutura do TF (Nocedal *et al.*, 2017).

Uma vez que reguladores e alvos podem sofrer diferentes pressões seletivas ao longo

do curso evolutivo (Rogers e Bulyk, 2018), é particularmente interessante considerar quando cada elemento que forma um regulon funcionando em *Homo sapiens* surgiu ao longo do processo evolutivo. Neste estudo, propomos uma estrutura geral para testar em regulons de risco para o desenvolvimento do câncer de mama, quem surgiu primeiro, reguladores ou regulados.

2 OBJETIVOS

2.1 OBJETIVOS GERAIS

Estabelecer em que ponto do processo evolutivo surgiram os fatores de transcrição de risco para o câncer de mama, bem como seus regulons. A descrição do processo de formação de regulons pode auxiliar no entendimento da maneira como, ao longo da evolução, uma rede regulatória estabeleceu a dinâmica de funcionamento que observamos nos organismos atuais.

2.2 OBJETIVOS ESPECÍFICOS

Testar três hipóteses concorrentes para explicar a formação de regulons:

1. fatores de transcrição e seus alvos surgiram de maneira independente;
2. fatores de transcrição surgiram previamente aos alvos que eles regulam;
3. fatores de transcrição surgiram posteriormente ao surgimento de seus alvos.

Descrever possíveis cenários evolutivos para o surgimento de regulons, e investigar se funcionalidades atuais podem ser explicadas pelo padrão de surgimento.

3 JUSTIFICATIVA

Organismos eucariotos possuem unidades funcionais controladas por fatores de transcrição, denominadas regulons, formadas por um conjunto de genes que estão sujeitos à ativação ou repressão em resposta a ação de um determinado fator de transcrição (Culjkovic *et al.*, 2007; Keene, 2007; Margolin *et al.*, 2006). As relações regulatórias entre estes elementos podem surgir ao longo do desenvolvimento de um organismo, em determinado tecido, ou mesmo serem entendidas em um contexto mais amplo, no curso do processo evolutivo. Por exemplo, tanto fatores de transcrição como seus alvos podem ser estudados por uma abordagem de ortologia. Entender o surgimento dos elementos que formam um regulon pode elucidar a maneira como atuam na regulação gênica de células de organismos atuais, tais como o controle transcricional em câncer (Greenman *et al.*, 2012). Neste estudo foi investigado o enraizamento evolutivo de unidades regulatórias associadas ao risco de desenvolvimento do câncer de mama.

4 MATERIAL E MÉTODOS

4.1 AMBIENTE DE PROGRAMAÇÃO

Todas as análises foram realizadas em linguagem de programação *R*, no ambiente de desenvolvimento integrado *RStudio* (RStudio Team, 2015), e a implementação de pacotes atendeu as especificações do projeto R/Bioconductor (Huber *et al.*, 2015). Esse repositório possui um conjunto de ferramentas utilizadas em análises e estudos do genoma em larga escala. A linguagem *R* é uma linguagem de programação de alto nível, que permite a criação de protótipos de maneira rápida, com forte comprometimento em qualidade e reprodutibilidade de resultados. Os pacotes depositados no Bioconductor possuem documentação tutorial para uso, bem como explicações sobre o funcionamento das métricas e funções oferecidas.

4.2 DADOS DE EXPRESSÃO GÊNICA

Neste trabalho foram utilizados dados de expressão gênica descritos no estudo METABRIC (Curtis *et al.*, 2012) que podem ser acessados no repositório Bioconductor através do pacote *Fletcher2013b* (Fletcher *et al.*, 2013).

A base de dados é formada pelo conjunto de dados de amostras de pacientes com câncer de mama incluindo: dados moleculares, dados de expressão gênica de tumores primários (extraídos por técnica de microarranjo) e informações clínicas dos pacientes. As amostras foram coletadas de pacientes do Reino Unido e do Canadá (Curtis *et al.*, 2012). A coleção tumoral contou com mais de 2.000 amostras de tumores de mama recém-congelados. Os tratamentos administrados a essas pacientes foram homogêneos em relação aos agrupamentos clinicamente relevantes, nenhuma paciente portadora de tumores HER2 receberam trastuzumabe. Quase todos os pacientes portadores de tumores responsivos a estrogênio ER+ e/ou linfonodo negativos não receberam quimioterapia. Pacientes ER-negativos e LN-positivos receberam tratamento quimioterápico. A coorte conta com uma coleção de 997 tumores que foram analisados como grupo de descoberta. Um grupo contendo 995 tumores foi utilizado para validação das análises executadas com a primeira coorte (Curtis *et al.*, 2012).

As pacientes pertencentes a essa coorte foram acompanhadas ao longo de quinze

anos. Assim, essa base de dados tráz informações relacionadas ao tempo de sobrevida de cada paciente e ainda qual o o tipo de abordagem terapêutica foi utilizada para cada caso.

4.3 REDES REGULATÓRIAS EM CÂNCER DE MAMA

As redes regulatórias transcricionais utilizadas nesse trabalho, foram previamente reconstruídas a partir dos dados de expressão gênica, utilizando o pacote *RTN* (Fletcher *et al.*, 2013). A ferramenta disponibiliza um conjunto de métodos que juntamente com algoritmo ARACNe (Margolin *et al.*, 2006) analisa os regulons e reconstrói redes regulatórias transcricionais (Margolin *et al.*, 2006; Janky *et al.*, 2014; Aibar *et al.*, 2017).

Nas análises, a partir do mapeamento por informação mútua (MI) (Fletcher *et al.*, 2013) o pacote infere as relações regulatórias entre todos os fatores de transcrição conhecidos e todos os potenciais genes alvo listados nos dados de expressão (Fig. 9) (Castro *et al.*, 2015). As redes regulatórias transcricionais reconstruídas estão disponíveis no pacote *Fletcher2013b* (Fletcher *et al.*, 2013).

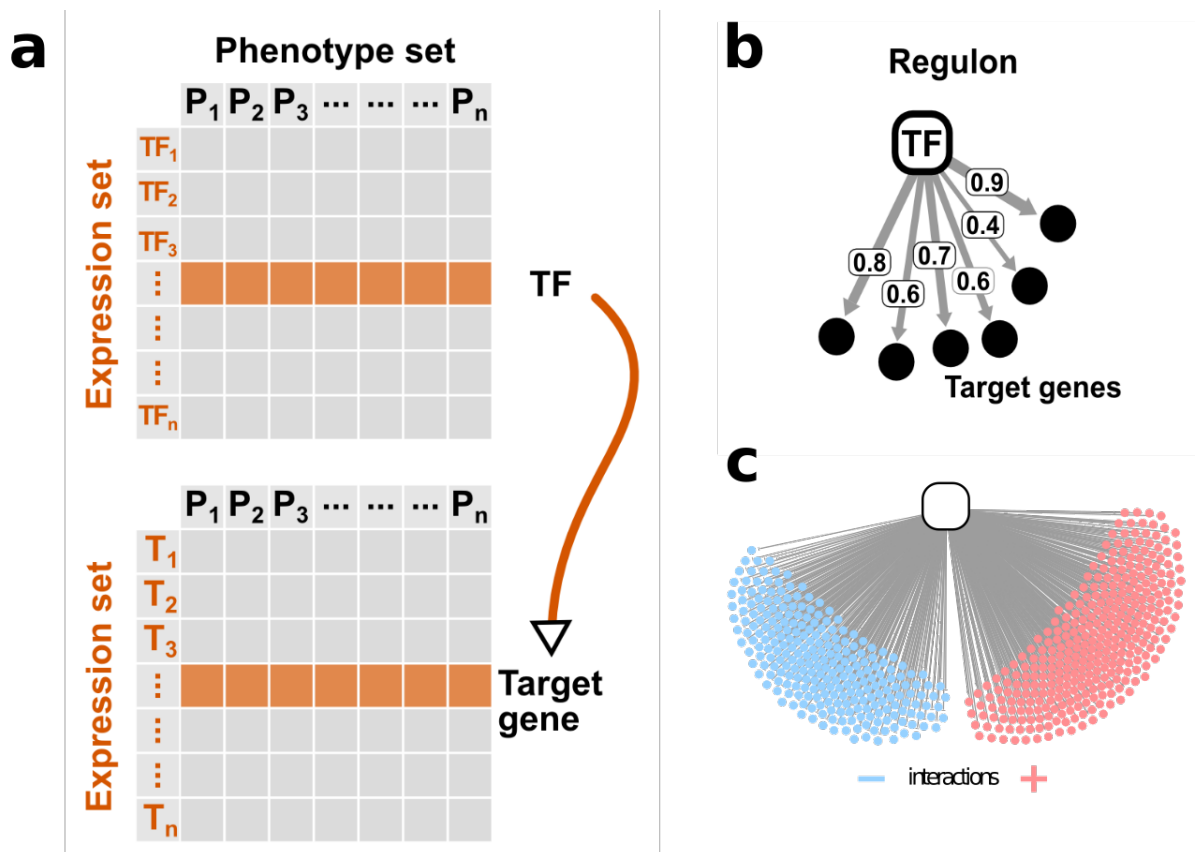


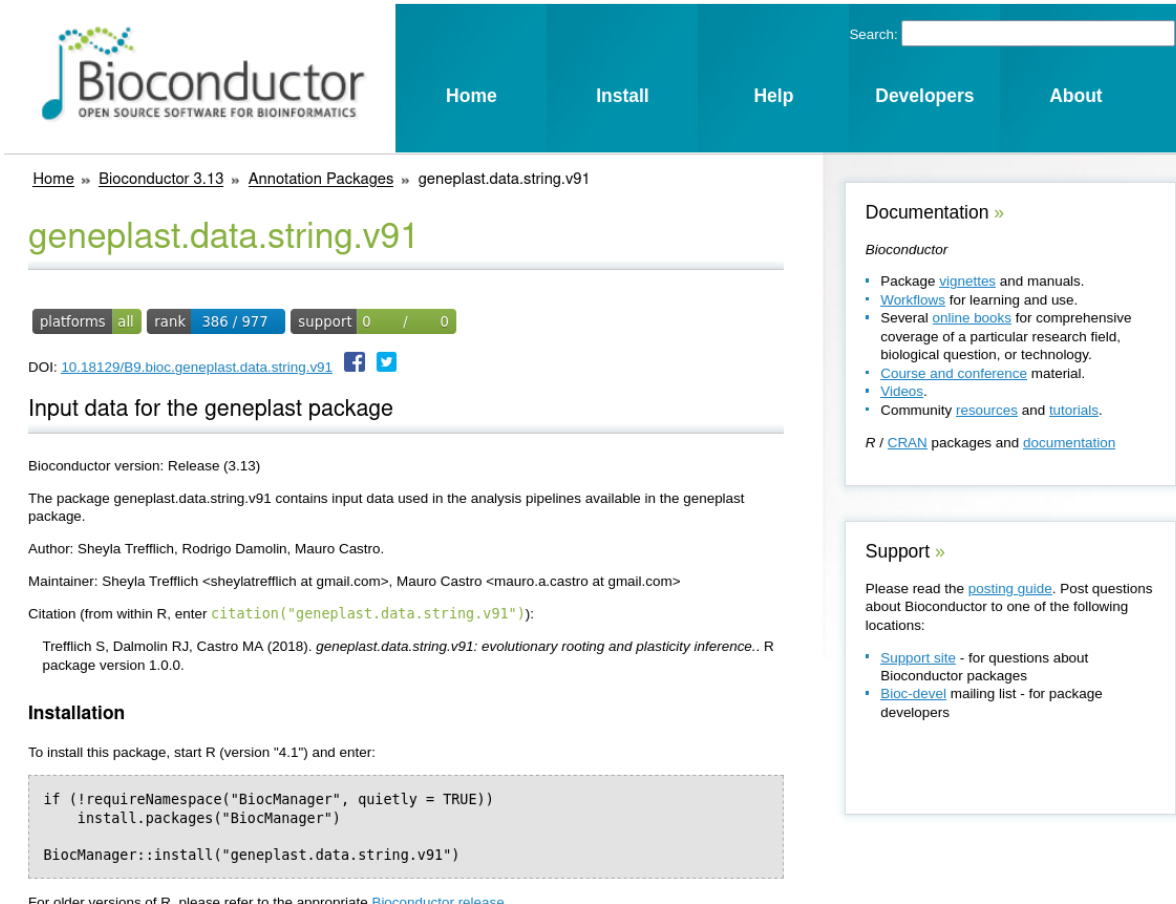
Figura 9. Desenho esquemático da reconstrução de redes regulatórias transcricionais. a) análise das matrizes de expressão gênica dos TFs e do conjunto de genes, para inferência de quais são os alvos de cada regulador. b) TF e genes alvo de um regulon. c) Regulon representando a nuvem de alvos regulados positivamente em rosa, ou seja, os genes que sofrerão a indução da expressão gênica; e negativamente em azul, ou seja, os genes que sofrerão a repressão da expressão gênica. (Chagas, 2017)

4.4 DADOS DE ORTOLOGIA/ANOTAÇÃO PARA TRN

Os dados de ortologia utilizados nas análises descritas nessa abordagem são provenientes do *STRINGdb*, um pacote *R* que contém uma coleção de informações de mais de 24 milhões de proteínas em 5090 organismos (Szkarczyk *et al.*, 2019). Esse pacote pode ser utilizado na atribuição de anotação de ortologia em um conjunto de genes.

Os dados de anotação de ortologia foram pré-processados para a extração apenas das proteínas encontradas em *Homo sapiens*. Posteriormente apenas as proteínas pertencentes à matriz de expressão gênica, da população portadora de câncer de mama, foram mantidas para o cruzamento com as redes regulatórias transcricionais.

A partir deste pré-processamento desenvolvemos o pacote *geneplast.data.string.v91* (Fig. 10) (Trefflich *et al.*, 2018), um pacote de dados *R* depositado no repositório Bioconductor que pode ser utilizado como *input* para análises de ortologia em redes regulatórias transcricionais.



The screenshot shows the Bioconductor website interface for the package *geneplast.data.string.v91*. The page includes the Bioconductor logo, navigation links (Home, Install, Help, Developers, About), and a search bar. The main content area displays the package name, version (3.13), and a navigation breadcrumb: Home » Bioconductor 3.13 » Annotation Packages » geneplast.data.string.v91. Below this, there are statistics for platforms (all), rank (386 / 977), and support (0 / 0). The DOI is 10.18129/B9.bioc.geneplast.data.string.v91. The page is titled "Input data for the geneplast package" and provides details about the package version (Release 3.13), its purpose (input data for analysis pipelines), author (Sheyla Trefflich, Rodrigo Damolin, Mauro Castro), maintainer (Sheyla Trefflich <sheylatrefflich@gmail.com>, Mauro Castro <mauro.a.castro@gmail.com>), and citation (Trefflich S, Dalmolin RJ, Castro MA (2018). *geneplast.data.string.v91: evolutionary rooting and plasticity inference.*. R package version 1.0.0.). An "Installation" section provides instructions to start R (version "4.1") and enter the following code:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("geneplast.data.string.v91")
```

 For older versions of R, it refers to the appropriate [Bioconductor release](#). On the right side, there are sections for "Documentation" and "Support". The "Documentation" section lists links for vignettes, workflows, online books, course and conference material, videos, and community resources and tutorials. The "Support" section includes a posting guide and links to a support site and a mailing list for package developers.

Figura 10. Pacote *geneplast.data.string.v91*. Representação da página do repositório Bioconductor, onde a ferramenta está depositada. Esse pacote contém um conjunto de dados de ortologia, harmonizado para ser utilizado em análises de redes regulatórias transcricionais juntamente com os dados de expressão gênica presentes no pacote *Fletcher2013b*

4.5 INFERÊNCIA DE ENRAIZAMENTO EVOLUTIVO

Para inferir o ponto de enraizamento evolutivo de um gene utilizamos o pacote *geneplast* (Dalmolin e Castro, 2015), uma ferramenta que avalia a distribuição de um grupo de genes ortólogos, em uma dada árvore de espécies (**Fig. 11**).

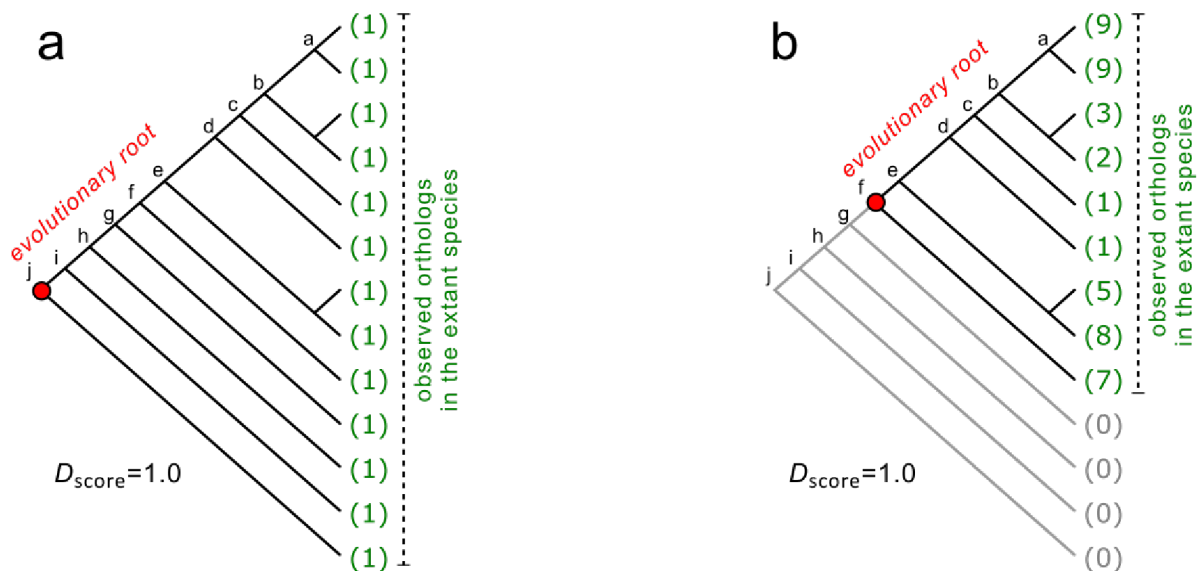


Figura 11. Possíveis cenários de enraizamento evolutivo de genes ortólogos e parálogos em uma determinada árvore de espécies. a, b) círculos vermelhos indicam as raízes evolutivas que melhor explicam os ortólogos observados nesta árvore de espécies. (Dalmolin e Castro, 2015)

A partir da inferência do ponto de enraizamento evolutivo dos genes pertencentes a rede regulatória transcricional de câncer de mama, é atribuído um valor numérico para cada gene definido como *Root*. Esse valor indica o posicionamento na árvore de espécies onde se encontra o LCA que melhor representa o ponto de enraizamento evolutivo do ortólogo.

4.5.1 REPRESENTAÇÃO DE REDES EM GRAFOS

Após a etapa de inferência do enraizamento evolutivo executada pelo pacote *geneplast* (Dalmolin e Castro, 2015) construímos um fluxo de análise que permite visualizar a ilustração dos regulons com a camada informativa relativa ao ponto de enraizamento evolutivo de cada um dos elementos formadores da unidade regulatória utilizando a ferramenta de visualização RedeR (Castro *et al.*, 2012).

RedeR (Castro *et al.*, 2012) é uma ferramenta R, depositada no repositório Bioconductor que permite a visualização de redes em grafos. No funcionamento dos sistemas biológicos o processo de regulação ocorre por meio da ação de genes ou proteínas formando uma rede e a representação por grafos facilita o estudo dos cenários regulatórios.

Desenvolvemos nessa abordagem a representação das redes em grafos com a adição da informação de ortologia dos genes componentes das redes regulatórias transcricionais.

O plot permite a análise do ponto de enraizamento evolutivo utilizando uma régua de gradiente de coloração contendo todos os pontos de enraizamento evolutivo pertencentes à árvore de espécies utilizada no estudo. Num contexto visual torna-se fácil observar se o TF enraiza antes ou depois do ponto de enraizamento do conjunto de alvos que ele regula. Assim observamos, por exemplo, o ponto de enraizamento evolutivo de alvos que têm a regulação compartilhada por mais de um TF.

O pacote *RedeR* (Castro *et al.*, 2012) também permite o agrupamento de alvos que são induzidos ou reprimidos em espaços visuais distintos no grafo, e juntamente com o fluxo de análise criado, permite a análise do enraizamento evolutivo nesses dois grupos de alvos. Importante ressaltar que os plots produzidos pelo pacote *RedeR* (Castro *et al.*, 2012) devem se destinar à análise de poucos regulons, pois dependendo da quantidade de alvos que cada regulon possui, a ilustração excede a capacidade visual de análise.

A fim de oferecer uma documentação detalhada, o fluxo da abordagem desenvolvido neste trabalho utilizando o *RedeR* foi implementado e figura como contribuição na vinheta do pacote *geneplast* (disposta integralmente em **Anexo D** - Case study: *Map rooting information on regulatory networks*). As etapas para reprodução desse estudo de caso seguem abaixo.

4.5.1.1 MAPEAR INFORMAÇÕES DE ENRAIZAMENTO EM REDES REGULATÓRIAS

Este exemplo visa mostrar a raiz evolutiva dos regulons (Fletcher *et al.* 2013). A ideia foi mapear a aparência de cada regulon (e os genes-alvo correspondentes) em uma árvore de espécies. As próximas etapas mostram como transferir informações evolutivas de enraizamento do *geneplast* para um modelo de gráfico. Nota: para fazer este trabalho, a anotação do gene disponível na rede regulatória de entrada precisa corresponder à anotação disponível nos dados do *geneplast* (neste caso, os IDs do gene ENTREZ são usados como chave combinatória entre os conjuntos de dados).

1 - Carregue um objeto de classe TNI e pacotes necessarios. Caso nao possua algum dos pacotes solicitados sera necessario realizar a instalacao [do](#) mesmo. O objeto `rtnilst` fornece regulons disponiveis no pacote de dados `Fletcher2013b` calculado a partir de dados de cancer de mama (Fletcher *et al.* 2013) que foram computados pelo pacote `RTN`.

```
library(RTN)
```

```
library(Fletcher2013b)
```

```
library(RedeR)
library(igraph)
library(RColorBrewer)
data("rtnilst")
```

2 - Extraia dois regulons de rtnilst em um objeto igraph. E possível extrair mais regulons, porém dependendo da quantidade de alvos de cada regulon, a resolução da imagem para análise visual pode ficar comprometida.

```
regs <- c("FOXM1", "PTTG1")
g <- tni.graph(rtnilst, gtype = "rmap", tfs = regs)
```

3 - Mapeie as informações de enraizamento no objeto igraph.

```
g <- ogr2igraph(ogr, cogdata, g, idkey = "ENTREZ")
```

4 - Ajuste as cores para as informações de enraizamento. Aqui utilizamos a paleta com tons vermelhos e azuis, porém esses atributos podem ser adaptados conforme a demanda do usuário.

```
pal <- brewer.pal(9, "RdYlBu")
color_col <- colorRampPalette(pal)(25) #set a color for each root!
g <- att.setv(g=g, from="Root", to="nodeColor", cols=color_col,
  na.col = "grey80", breaks = seq(1,25))
```

5 - Ajustes estéticos para alguns atributos do gráfico. Aqui alteramos o tamanho da letra, o tamanho do nó, a cor das arestas e o contorno dos nós

```
idx <- V(g)$SYMBOL %in% regs
V(g)$nodeFontSize[idx] <- 30
V(g)$nodeFontSize[!idx] <- 1
E(g)$edgeColor <- "grey80"
V(g)$nodeLineColor <- "grey80"
```

6 - Envie o objeto igraph para a interface RedeR. nesse passo abra uma janela onde ser plotado o gráfico. importante ressaltar que quanto maior os regulons escolhidos, mais tempo levar para o grafo aparecer na tela.

```

rdp <- RedPort()
callD(rdp)
resetD(rdp)
addGraph( rdp, g, layout=NULL)
addLegend.color(rdp, colvec=g$legNodeColor$scale, size=15,
  labvec=g$legNodeColor$legend, title="Roots represented in Fig4")
relax(rdp, 15, 100, 20, 50, 10, 100, 10, 2, ps=TRUE)

```

As etapas descritas acima plotam um gráfico similar ao que pode ser visualizado na **Figura 20** no tópico **Resultados e Discussão**. Porém na **Figura 20** plotamos o grafo para três regulons.

4.5.2 TREEANDLEAF

No fluxo de análise utilizamos o plot de dendrograma do tipo TreeAndLeaf, com a adição da informação de ortologia, e para isso utilizamos a ferramenta de mesmo nome, criada pelo grupo de estudos e depositada no repositório R/Bioconductor. A ilustração, que pode ser vista na **Figura 24a**, já tinha sido publicada anteriormente no trabalho de Castro *et al.* (2015), porém só a partir da abordagem aqui desenvolvida que a rede hierárquica adquiriu a camada de informação de ortologia.

Esse tipo de grafo se ocupa de evidenciar as folhas pertencentes a um determinado dendrograma binário. Essa representação gráfica tem muitas vantagens sobre os plots de dendrogramas convencionais, pois permitem o empilhamento de mais de uma camada informativa. Nesse caso específico o tamanho dos nodos se refere a quantidade de alvos que o fator de transcrição regula, ou seja, o tamanho do regulon; a coloração se refere a mediana do ponto de enraizamento evolutivo de todo o conjunto de elementos que formam o regulon. Outros atributos podem ser levados em consideração na representação do grafo, de acordo com necessidade de cada abordagem, já que a ferramenta permite a adição de mais camadas informativas.

O objetivo da criação dessa imagem foi cluserizar os regulons baseado na correlação da expressão gênica dos alvos compartilhados. O estudo de sua publicação (Castro *et al.*, 2015) conseguiu estabelecer a correlação entre regulons que conferem risco para o desenvolvimento do câncer de mama (Receptores de Estrogênio Positivo e Receptores de Estrogênio Negativo, agrupados nos dois círculos grandes pontilhados posicionados no topo e na base da imagem, respectivamente). Assim, cada nodo da árvore representa um regulon, ou seja, o conjunto formado entre Fator de Transcrição e o conjunto de genes alvo. A proximidade dos nodos está relacionada com a correlação de alvos compartilhados.

REF: Homo sapiens (9606)



Figura 12. Árvore de espécies utilizada no estudo Organismo de referência *H. sapiens* localizado no topo do cladograma.

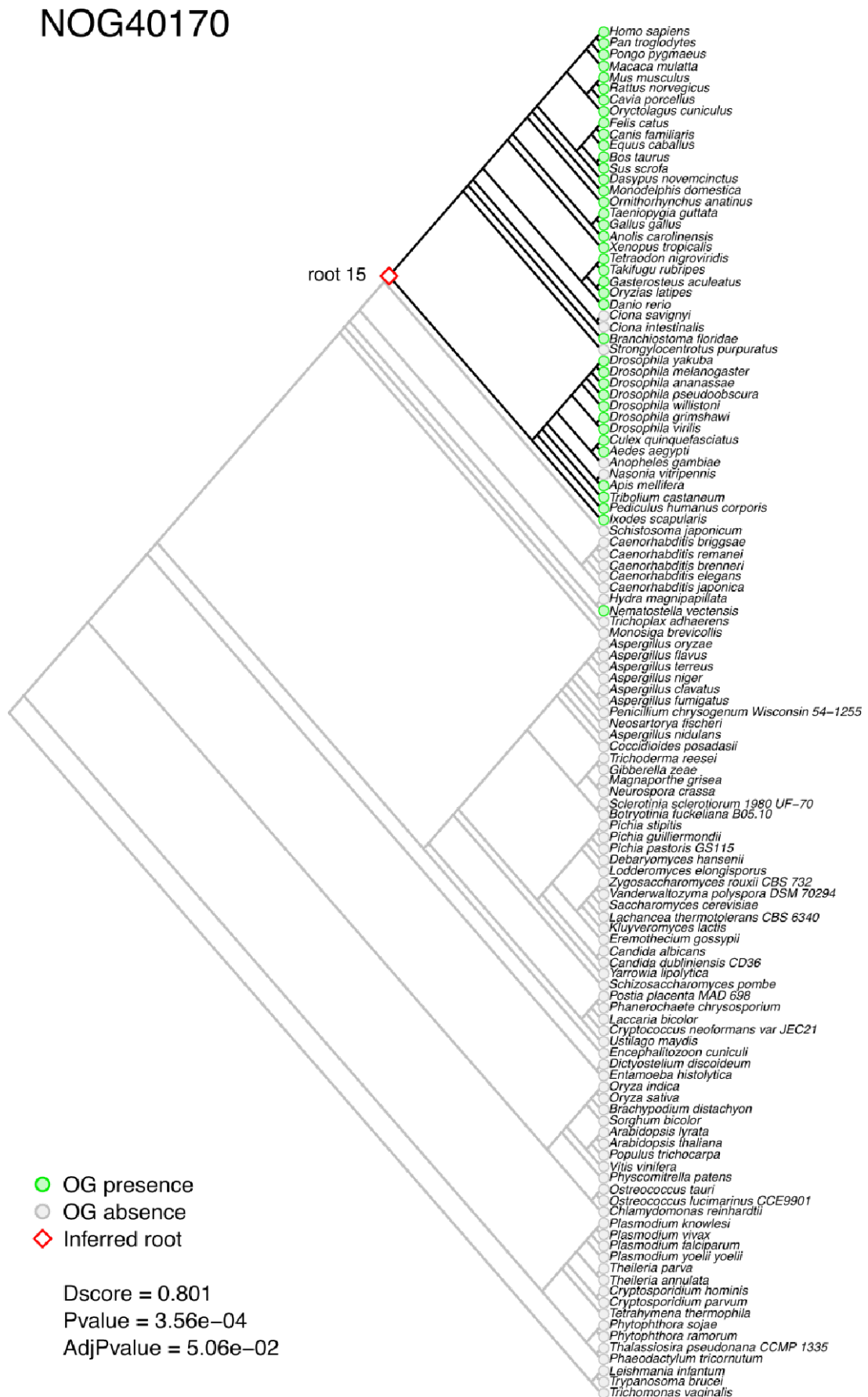


Figura 13. Árvore de espécies com enraizamento evolutivo Árvore com marcação em verde para presença de um determinado gene nos organismos. Em vermelho a localização da inferência do ponto de enraizamento evolutivo do gene em questão.

4.6 ANÁLISE EVOLUTIVA

A partir do valor numérico atribuído ao ponto de enraizamento evolutivo de genes e fatores de transcrição, calculamos o distanciamento evolutivo entre o ponto de surgimento dos elementos que formam os regulons presentes na TRN.

Nesse estudo *H. sapiens* é a espécie de referência e por essa razão aparece no topo da árvore utilizada (**Fig. 12**). Por exemplo, genes humanos cujos ortólogos estão presentes em todas as espécies listadas, aparecem enraizados na base da árvore, portanto longe do LCA de *H. sapiens*. Genes humanos encontrados em poucas espécies terão o LCA enraizado mais próximo a essas espécies e provavelmente mais próximos do LCA de *H. sapiens*. Na **Figura 13** exemplificamos a árvore de espécies com a presença ou ausência de um determinado ortólogo nas espécies contidas na árvore e onde se infere o ponto de enraizamento evolutivo. Entre os dois extremos encontramos diversos outros pontos de enraizamento, totalizando 24 LCAs na árvore de espécies utilizada neste estudo.

Dessa maneira estabelecemos a distância entre os pontos de enraizamento de um regulador (PE_R) e a média dos pontos de enraizamento de seus alvos (\overline{PE}_A) para todas as unidades regulatórias ($DE = PE_R - \overline{PE}_A$) que formam a TRN.

Por análise de permutação, geramos distâncias evolutivas aleatórias a fim de estimarmos um P nominal para cada DE observada. A distância evolutiva obtida entre um dado regulador e seus alvos foi comparada com distância evolutiva obtida para 10.000 regulons aleatórios, permitindo estimar se a DE observada pode ou não ser explicada pelo acaso.

4.6.1 EXPLORANDO RAÍZES EVOLUTIVAS DE REGULONS

Fletcher et al. (2013) reconstruíram regulons para 809 fatores de transcrição (TFs) usando dados transcriptômicos de microarranjo de tecido mamário, tanto de câncer quanto de amostras normais (Curtis et al. 2012). Nosso objetivo aqui é avaliar a raiz evolutiva dos regulons reconstruídos por Fletcher et al. (2013) usando o pacote geneplast.

1. Instalação de pacotes e conjuntos de dados. Certifique-se de instalar todos os pacotes necessários. Instalar e carregar os pacotes de dados geneplast.data.string.v91 e Fletcher2013b disponibilizará todos os dados necessários para este estudo de caso.

```
#-- Call packages
```

```
library(geneplast)
library(geneplast.data.string.v91)
library(RTN)
```



```
library(Fletcher2013b)
library(ggplot2)
library(ggpubr)
library(plyr)
```

2. Inferindo raízes evolutivas Essa análise determinará a raiz evolutiva de um gene com base na distribuição de seus ortólogos em uma dada árvore de espécies. Precisaremos de dois objetos de dados, `cogdata` e `phyloTree`, ambos carregados com a chamada `gpdata-string-v91`. O `cogdata` é um objeto `data.frame` que lista grupos ortólogos (OGs) previstos para 121 espécies eucarióticas, enquanto o `phyloTree` é um objeto de árvore filogenética da classe `phylo`. A função `groot.preprocess` irá verificar os dados de entrada e construir um objeto da classe `OGR`, que será usado nas etapas subsequentes do pipeline de análise.

```
#-- Carregar dados de ortologia do pacote
  'geneplast.data.string.v91'

data(gpdata_string_v91)

#-- Crie um objeto da classe 'OGR' para uma referencia 'spid'

ogr <- groot.preprocess(cogdata=cogdata, phyloTree=phyloTree,
  spid="9606")
```

A função `groot` tem o objetivo de encontrar a raiz evolutiva de um gene em uma árvore filogenética. O método infere a probabilidade de que tal característica estivesse presente no Último Ancestral Comum (LCA) de uma determinada linhagem. A função `groot` avalia a presença e ausência de ortólogos nas espécies existentes da árvore filogenética para construir uma distribuição de probabilidade, que é usada para identificar padrões de herança vertical. O parâmetro `spid = 9606` define *Homo sapiens* como a espécie de referência, que define a linhagem ancestral avaliada na consulta (ou seja, cada ortólogo da espécie de referência terá suas raízes em um ancestral da espécie de referência).

```
#-- Execute a funcao 'groot' para inferir as raizes evolutivas

ogr <- groot(ogr, nPermutations=1000, verbose=TRUE)
```

3. Análise evolutiva de regulons gerados a partir de amostras de câncer de mama

- Mapeamento de anotação de raiz para gene. Nesta seção, mapearemos as raízes evolutivas inferidas (disponíveis no objeto `ogr`) para genes anotados nos regulons

reconstruídos por Fletcher et al. (2013) de amostras de câncer de mama (disponíveis no objeto `rtnilst`). Para um resumo dos regulons no objeto `rtnilst`, recomendamos o uso da função `tni.regulon.summary`, que mostra que existem 809 elementos regulatórios (TFs) e 14131 alvos.

```

#-- Carregue os regulons

data("rtnilst")
tni.regulon.summary(rtnilst)

## This regulatory network comprised of 809 regulons.
## -- DPI-filtered network:
## regulatoryElements    Targets          Edges
##           809           14131          47012
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
##    0.0  10.0  37.0  58.1  80.0 523.0
## -- Reference network:
## regulatoryElements    Targets          Edges
##           809           14131          617672
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
##     0    43   449   764  1245  4148
## ---

```

Vamos transformar o `rtnilst` em um objeto gráfico usando a função `tni.graph`. O gráfico resultante será avaliado pela função `ogr2igraph`, que mapeará a anotação `root-to-gene`; os resultados estarão disponíveis no `data frame` `roots-df` para análise subsequente.

```

#-- Colocar regulons em um objeto 'igraph'

#-- Nota: pequenos regulons (n<15 alvos) sao gerados nesta
      etapa.

graph <- tni.graph(rtnilst, gtype = "rmap")

#-- Mapeie o objeto 'ogr' para o objeto 'igraph'

graph <- ogr2igraph(ogr, cogdata, graph, idkey = "ENTREZ")

#-- Faca um data frame com as raizes do gene

```

```

roots_df <- data.frame(COGID = V(graph)$COGID,
                      SYMBOL = V(graph)$SYMBOL,
                      ENTREZ = V(graph)$ENTREZ,
                      Root = V(graph)$Root,
                      TRN_element =
                        c("Target", "TF") [V(graph)$tfs+1],
                      stringsAsFactors = FALSE)

```

Observe que algum nível de anotação ausente é esperado, pois nem todos os ids de genes listados nos cogdata podem estar disponíveis no objeto gráfico. Além disso, pequenos regulons (n com menos de 15 alvos) são removidos pela função `tni.graph`. Como uma etapa final de pré-processamento, removeremos os genes enraizados na base da árvore filogenética, para os quais as previsões não podem discriminar as raízes ancestrais anteriores. Aqui, 307 TFs e 6308 alvos foram retidos.

```

#-- Remover NAs de anotacoes ausentes

roots_df <- roots_df[complete.cases(roots_df),]

#-- Remova genes enraizados na base da arvore filogenetica

roots_df <- roots_df[roots_df$Root<max(roots_df$Root),]
rownames(roots_df) <- 1:nrow(roots_df)

#-- Verifique TF e contagens de destino

table(roots_df$TRN_element)

## Target  TF
## 6308   307

```

- Comparando reguladores e metas. Uma rede regulatória transcricional (TRN) é formada por reguladores (TFs) e genes alvo. O data frame `roots-df` lista as raízes evolutivas inferidas para cada elemento TRN, incluindo se o elemento TRN é anotado como TF ou alvo.
-

```

head(roots_df)

##      COGID SYMBOL ENTREZ Root TRN_element
## 1 KOG3119 CEBPG 1054 19      TF
## 2 KOG4217 NR4A2 4929 17      TF

```

```
## 3 KOG0493 EN1 2019 17 TF
## 4 NOG80479 TP53 7157 20 TF
## 5 KOG3740 GATAD2A 54815 19 TF
## 6 COG5150 DR1 1810 23 TF
```

```
tail(roots_df)
```

```
##          COGID SYMBOL ENTREZ Root TRN_element
## 6610 COG5640  F11 2160 19 Target
## 6611 KOG1418 KCNK18 338567 24 Target
## 6612 NOG39443 TMEM220 388335 14 Target
## 6613 NOG43522 C1orf170 84808 7 Target
## 6614 NOG127335 C16orf96 342346 6 Target
## 6615 NOG27843 PANX3 116337 13 Target
```

Por exemplo, o gene **CEBPG** é colocado na raiz 19, enquanto o gene **PANX3** é colocado na raiz 13, indicando que a raiz evolutiva inferida para **CEBPG** é mais ancestral do que a raiz evolutiva inferida para **PANX3**. Observe que as raízes evolutivas são enumeradas do nó mais recente ao mais ancestral da árvore filogenética. Além disso, como o objetivo da análise é encontrar a raiz dos ortólogos das espécies de referência, a enumeração da raiz está relacionada à linhagem ancestral das espécies de referência.

Aqui iremos comparar a distribuição das raízes evolutivas inferidas para TFs e genes alvo usando o teste de Wilcoxon-Mann-Whitney e, em seguida, gerar gráficos de violino que podem ser visualizados na **Figura 18b**.

```
#-- Avalie a distribuicao de raiz por TRN_element

wilcox.test(Root ~ TRN_element, data=roots_df)

## Wilcoxon rank sum test with continuity correction
## data: Root by TRN_element
## W = 812534, p-value = 1.6e-06
## alternative hypothesis: true location shift is not equal
## to 0

#-- Definir raizes para exibir no eixo y

roots <- c(4,8,11,13,19,21,25)

#-- Defina uma funcao de resumo para exibir dispersao
```

```
dentro dos violinos

data_summary <- function(x) {
  y <- mean(x); ymin <- y-sd(x); ymax <- y+sd(x)
  return(c(y=y, ymin=ymin, ymax=ymax))
}

#-- Gerar graficos de violino mostrando a distribuicao raiz
por TRN_element

p <- ggplot(roots_df, aes(x=TRN_element, y=Root)) +
  geom_violin(aes(fill=TRN_element), adjust=2,
             show.legend=F) +
  scale_y_continuous(breaks=roots,
                    labels=paste("root", roots)) +
  scale_fill_manual(values=c("#c7eae5", "#dfc27d")) +
  labs(x="TRN elements", y="Root distribution") +
  scale_x_discrete(limits=c("TF", "Target"),
                  labels=c("TFs", "Targets")) +
  theme_classic() +
  theme(text=element_text(size=20)) +
  stat_summary(fun.data = data_summary)
p + stat_compare_means(method="wilcox.test",
                      comparisons =list(c("TF", "Target")),
                      label = "p.signif")
```

Em seguida, calculamos a distância das raízes entre um TF e seus alvos e geramos um gráfico de setores e um boxplot que reproduz os cenários evolutivos discutidos na **Figura 18**.

```
#-- Obtenha raizes para TFs

idx <- roots_df$TRN_element=="TF"
tfroots <- roots_df$Root[idx]
names(tfroots) <- roots_df$SYMBOL[idx]

#-- Obtenha raizes para genes alvo

regulonlist <- tni.get(rtnilst, what = "regulons", idkey =
  "ENTREZ")[names(tfroots)]
targetroots <- lapply(regulonlist, function(reg){
```

```

    roots_df$Root[roots_df$ENTREZ%in%reg]
  })

#-- Calcular distancias das raizes entre um TF e seus alvos

rootdist <- sapply(names(targetroots), function(reg){
  targetroots[[reg]]-tfroots[reg]
})

#-- Calcular distancias das raizes medianas e classifique
objetos relacionados

rootdist_med <- sort(unlist(lapply(rootdist, median)),
  decreasing = T)
rootdist <- rootdist[names(rootdist_med)]
tfroots <- tfroots[names(rootdist_med)]
targetroots <- targetroots[names(rootdist_med)]
regulonlist <- regulonlist[names(rootdist_med)]

#-- Definir grupos de regulon com base nas distancias raiz
medianas

regulon_grouplist <- -sign(rootdist_med)+2
regulon_groupnames <- c("group_a", "group_b", "group_c")
regulon_groupcolors = c("#98d1f2", "grey", "#1c92d5")
names(regulon_groupcolors) <- regulon_groupnames

#-- Gerar um grafico de setores mostrando os regulons
agrupados com base na distancia media entre a raiz de um
TF e seus alvos' roots

n <- as.numeric(table(regulon_grouplist))
pie(n, labels = paste(n, "regulons"), col =
  regulon_groupcolors,
  border="white", cex=1.5, clockwise = TRUE, init.angle=0)
labs <- c("TF-target genes rooted before the TF (group-a)",
  "TF-target genes rooted with the TF (group-b)",
  "TF-target genes rooted after the TF (group-c)")
legend("bottomleft", fill = regulon_groupcolors, bty = "n",
  legend = labs)

```

Regulons agrupados com base na distância média entre a raiz de um TF e as raízes de seus alvos.

```

#-- Gerar um boxplot mostrando regulons individuais

#-- classificacao pela distancia mediana ate a raiz do TF

plot.new()
par(usr=c(c(0,length(rootdist)),range(rootdist)))
boxplot(rootdist, horizontal= F, outline=FALSE, las=2,
        axes=FALSE, add=T,
        pars = list(boxwex = 0.6,
                    boxcol=regulon_groupcolors[regulon_grouplist],
                    whiskcol=regulon_groupcolors[regulon_grouplist]),
        pch="|", lty=1, lwd=0.75,
        col = regulon_groupcolors[regulon_grouplist])
abline(h=0, lmitre=5, col="#E69F00", lwd=3, lt=2)
par(mgp=c(2,0.1,0))
axis(side=1, cex.axis=1.2, padj=0.5, hadj=0.5, las=1,
     lwd=1.5, tcl= -0.2)
par(mgp=c(2.5,1.2,0.5))
axis(side=2, cex.axis=1.2, padj=0.5, hadj=0.5, las=1,
     lwd=1.5, tcl= -0.2)
legend("topright", legend = labs, fill =
      regulon_groupcolors, bty = "n")
title(xlab = "Regulons sorted by the median distance to TF
      root", ylab = "Distance to TF root")

```

- Comparando fatores de transcrição e co-fatores de transcrição.

Co-fatores de transcrição (TcoFs) são determinantes críticos das atividades dos TFs. Os TcoFs não se ligam diretamente ao DNA, mas influenciam a regulação da transcrição, formando complexos de proteínas com os TFs. A seguir, compararemos essas duas classes de reguladores avaliando a distribuição das raízes evolutivas inferidas para TFs e TcoFs. Para executar os snippets subsequentes, exigiremos a lista de TcoFs humanos disponíveis no banco de dados TcoF-DB (Schmeier et al. 2016) (faça o download do arquivo ‘TcoF-DB.xlsx’ conforme indicado abaixo).

```

#-- Baixar o arquivo 'TcoF-DB.xlsx' de
#-- https://tools.sschmeier.com/tcof/browse
#-- /?type=tcof&species=human&class=all
#-- e entao carregue-o com a funcao 'read_excel'

```

```

library(readxl)
TcoF_DB <- read_excel("TcoF-DB.xlsx")

#-- Selecionar TcoFs de alta confianca de acordo com o
    banco de dados TcoF

TcoF_DB <- TcoF_DB[TcoF_DB$Type=="TcoF: class HC",]

#-- Mapeamento de 'TcoF_DB' em 'roots_df'

roots_df_TcoF_DB <- roots_df
roots_df_TcoF_DB$TRN_element <- NA
roots_df_TcoF_DB$TRN_element[roots_df$SYMBOL %in%
    TcoF_DB$Symbol] <- "TcoF"
roots_df_TcoF_DB$TRN_element[roots_df$TRN_element%in%"TF"]
    <- "TF"
roots_df_TcoF_DB <-
    roots_df_TcoF_DB[!is.na(roots_df_TcoF_DB$TRN_element),]
table(roots_df_TcoF_DB$TRN_element)
## TcoF TF
## 146 307

#-- Avaliar a distribuicao da raiz para o TRN_element

wilcox.test(Root ~ TRN_element, data=roots_df_TcoF_DB)

## Wilcoxon rank sum test with continuity correction
## data: Root by TRN_element
## W = 22226, p-value = 0.884
## alternative hypothesis: true location shift is not equal
    to 0

#-- Gerar graficos de violino mostrando a distribuicao da
    raiz por TRN_element

p <- ggplot(roots_df_TcoF_DB, aes(x=TRN_element, y=Root)) +
    geom_violin(aes(fill=TRN_element), adjust=2,
        show.legend=F) +
    scale_y_continuous(breaks=roots,
        labels=paste("root",roots)) +

```



```

scale_fill_manual(values=c("#c7eae5", "#dfc27d")) +
labs(x="TRN elements", y="Root distribution") +
scale_x_discrete(limits=c("TF", "TcoF"),
  labels=c("TFs", "TcoFs")) +
theme_classic() +
theme(text=element_text(size=20)) +
stat_summary(fun.data = data_summary)
p + stat_compare_means(method="wilcox.test",
  comparisons =list(c("TF", "TcoF")),
  label = "p.signif")

```

- Explorando abundância, diversidade e plasticidade.

Nesta seção, mostramos como calcular a abundância, diversidade e plasticidade do OG e, em seguida, mapear essas três métricas para regulons (consulte Castro et al. (2008) e Dalmolin et al. (2011) para uma descrição detalhada). Resumidamente, a métrica de abundância representa o número de ortólogos dividido pelo número de espécies anotadas em um determinado OG; abundância = 1 indica uma relação um-para-um entre o número de ortólogos e espécies, enquanto abundância maior que 1 indica que o número de ortólogos excede o número de espécies. Um grande valor de abundância sugere um grande número de parálogos anotados no OG. A métrica de diversidade representa a distribuição de ortólogos e parálogos em uma árvore de determinada espécie; alta diversidade representa uma distribuição homogênea (por exemplo, um ortólogo em cada espécie), enquanto a baixa diversidade indica que os genes ortólogos estão concentrados em poucas espécies (por exemplo, em um único ramo da árvore de espécies). A plasticidade é a combinação de abundância e diversidade em uma única métrica. A baixa plasticidade é observada em OGs de baixa abundância e alta diversidade (por exemplo, poucos ortólogos distribuídos em muitas espécies), enquanto a alta plasticidade é observada em OGs de alta abundância e baixa diversidade (por exemplo, muitos ortólogos concentrados em poucas espécies).

```

#-- Calculo de abundancia, diversidade e plasticidade de OG

ogp <- gplast.preprocess(cogdata=cogdata,
  sspids=phyloTree$tip.label)
ogp <- gplast(ogp)
gpres <- gplast.get(ogp, what="results")
head(gpres)

##      abundance diversity plasticity

```

```

## COG0001 1.3871 0.6889 0.4150
## COG0002 1.1346 0.8110 0.2386
## COG0003 1.3243 0.9506 0.1739
## COG0004 4.1753 0.8880 0.5654
## COG0005 2.5455 0.9283 0.4182
## COG0006 4.3167 0.9769 0.5298

#-- Mapear a abundancia, diversidade e plasticidade de OGS
    para o data frame 'roots_df'

idx <- match(roots_df$COGID, rownames(gpres))
roots_df$Abundance <- gpres$abundance[idx]
roots_df$Diversity <- gpres$diversity[idx]
roots_df$Plasticity <- gpres$plasticity[idx]

#-- Em seguida, mapear a abundancia, diversidade e
    plasticidade de OGS para regulons

stats_df <- lapply(regulonlist, function(reg){
  temp <- roots_df[roots_df$ENTREZ%in%reg,]
  apply(temp[, c("Abundance", "Diversity", "Plasticity")], 2,
    mean)
})
stats_df <- ldply(stats_df, .id="Regulon",
  stringsAsFactors=FALSE)
stats_df$regulon_groups <-
  regulon_grouplist[stats_df$Regulon]
stats_df$regulon_groups <-
  regulon_groupnames[stats_df$regulon_groups]

#-- Avaliar a abundancia de OGS por grupos de regulon

p <- ggplot(stats_df, aes(x=regulon_groups, y=Abundance,
  fill=regulon_groups)) +
  geom_boxplot(show.legend=F) +
  scale_y_continuous(limits = c(0,60)) +
  scale_x_discrete(limits=c("group_a", "group_c")) +
  scale_fill_manual(values=regulon_groupcolors[c("group_a", "group_c")])
  +
  labs(x="Regulon groups", y="OG's abundance") +
  theme(panel.grid = element_blank()) +

```

```

theme(text=element_text(size=20),
      axis.line.x=element_blank())
p + stat_compare_means(method="wilcox.test",
                      comparisons =list(c("group_a", "group_c")),
                      label = "p.signif")

#-- Avaliar a diversidade de OGs por grupos de regulon

p <- ggplot(stats_df, aes(x=regulon_groups, y=Diversity,
                        fill=regulon_groups)) +
  geom_boxplot(show.legend=F) +
  scale_y_continuous(limits = c(0.5,1)) +
  scale_x_discrete(limits=c("group_a", "group_c")) +
  scale_fill_manual(values=regulon_groupcolors[c("group_a", "group_c")])
  +
  labs(x="Regulon groups", y="OG's diversity") +
  theme(panel.grid = element_blank()) +
  theme(text=element_text(size=20),
        axis.line.x=element_blank())
p + stat_compare_means(method="wilcox.test",
                      comparisons =list(c("group_a", "group_c")),
                      label = "p.signif")

#-- Avaliar a plasticidade de OGs por grupos de regulon

p <- ggplot(stats_df, aes(x=regulon_groups, y=Plasticity,
                        fill=regulon_groups)) +
  geom_boxplot(show.legend=F) +
  scale_y_continuous(limits = c(0,1)) +
  scale_x_discrete(limits=c("group_a", "group_c")) +
  scale_fill_manual(values=regulon_groupcolors[c("group_a", "group_c")])
  +
  labs(x="Regulon groups", y="OG's plasticity") +
  theme(panel.grid = element_blank()) +
  theme(text=element_text(size=20),
        axis.line.x=element_blank())
p + stat_compare_means(method="wilcox.test",
                      comparisons =list(c("group_a", "group_c")),
                      label = "p.signif")

```

4. Análise evolutiva de regulons gerados a partir de amostras de tecido mamário normal

Os regulons são construídos com base na expressão de genes componentes de uma coorte. Grandes coortes de amostras de tumor normalmente contêm vários subtipos moleculares e, normalmente, fornecem boa variabilidade de expressão para a construção de regulons. Em contraste, os conjuntos de amostras que são mais homogêneos podem ser mais desafiadores para explorar com regulons, e este pode ser o caso com conjuntos de amostras normais não cancerosas. Apesar desse desafio, Fletcher et al. (2013) geraram regulons usando amostras de tecido mamário normal, a fim de observar diferenças regulatórias entre células cancerosas e normais. Aqui, executaremos a mesma análise evolutiva descrita com amostras de tecido mamário tumoral, mas agora usando regulons gerados a partir de amostras de tecido mamário normal. Nos próximos passos, mostramos como reproduzir os resultados anteriores usando uma rede regulatória transcricional diferente.

- Mapeamento de anotação das raízes para genes

```

data("rtniNormals")
graph_normals <- tni.graph(rtniNormals, gtype = "rmap")
graph_normals <- ogr2igraph(ogr, cogdata, graph_normals,
  idkey = "ENTREZ")
roots_df_normals <- data.frame(COGID =
  V(graph_normals)$COGID,
  SYMBOL = V(graph_normals)$SYMBOL,
  ENTREZ = V(graph_normals)$ENTREZ,
  Root = V(graph_normals)$Root,
  TRN_element =
    c("Target", "TF") [V(graph_normals)$tfs+1])
roots_df_normals <-
  roots_df_normals[complete.cases(roots_df_normals),]
roots_df_normals <-
  roots_df_normals[roots_df_normals$Root < max(roots_df_normals$Root),]
rownames(roots_df_normals) <- 1:nrow(roots_df_normals)
table(roots_df_normals$TRN_element)

## Target  TF
## 2818   130

```

- Comparando reguladores e alvos

```

#-- Avaliar a distribuio raiz por TRN_element

wilcox.test(Root ~ TRN_element, data=roots_df_normals)

```

```

## Wilcoxon rank sum test with continuity correction
## data: Root by TRN_element
## W = 152522, p-value = 0.001148
## alternative hypothesis: true location shift is not equal
  to 0

#-- Set roots to display in y-axis
roots <- c(4,8,11,13,19,21,25)

#-- Set a summary function to display dispersion within the
  violins
data_summary <- function(x) {
  y <- mean(x); ymin <- y-sd(x); ymax <- y+sd(x)
  return(c(y=y,ymin=ymin,ymax=ymax))
}

#-- Gerar graficos de violino mostrando a distribuio raiz
  por TRN_element

p <- ggplot(roots_df_normals, aes(x=TRN_element, y=Root)) +
  geom_violin(aes(fill=TRN_element), adjust=2,
    show.legend=F) +
  scale_y_continuous(breaks=roots,
    labels=paste("root",roots)) +
  scale_fill_manual(values=c("#c7eae5","#dfc27d")) +
  labs(x="TRN elements", y="Root distribution") +
  scale_x_discrete(limits=c("TF", "Target"),
    labels=c("TFs", "Targets")) +
  theme_classic() +
  theme(text=element_text(size=20)) +
  stat_summary(fun.data = data_summary)
p + stat_compare_means(method="wilcox.test",
  comparisons =list(c("TF", "Target")),
  label = "p.signif")

```

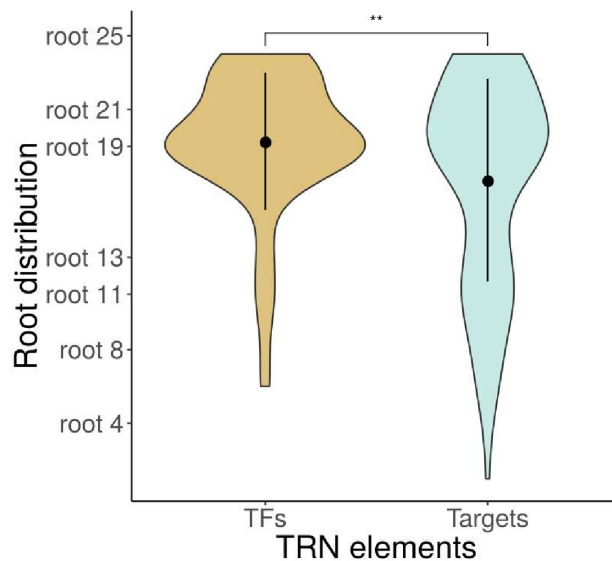


Figura 14. Plot de violino para amostras normais. Em marrom se encontra o grupo dos Fatores de Transcrição e em azul o grupo dos alvos

4.7 FLUXO DO ESTUDO

O fluxograma completo contendo todas as etapas da abordagem desenvolvida neste trabalho pode ser consultado na **Figura 15**. O fluxo de análise requer dois tipos de dados de entrada: dados de expressão e dados de ortologia. Os dados de expressão gênica provenientes do estudo METABRIC (Curtis *et al.*, 2012) estão organizados no pacote *Fletcher2013b* (Fletcher *et al.*, 2014), o qual também disponibiliza as redes regulatórias reconstruídas pelo pacote *RTN* (Fletcher *et al.*, 2013). Apenas com esses dados já é possível plotar grafos de alguns regulons utilizando o pacote *RedeR* (Castro *et al.*, 2012), porém sem qualquer informação evolutiva. Nessa primeira etapa, os regulons reconstruídos indicam quais são os fatores de transcrição e qual o conjunto de genes alvo. Os dados de anotação de ortologia foram obtidos através do pacote *STRINGdb* (Franceschini, 2013) pré-processados, filtrados e harmonizados com as informações de anotação dos dados de expressão dos genes pertencentes a Rede Regulatória Transcricional.

As informações de anotação de ortologia foram mapeadas nas redes regulatórias transcricionais reconstruídas para inferência do enraizamento evolutivo. A partir daí foi possível conhecer o valor numérico do ponto de enraizamento do LCA de cada gene constituinte da Rede Regulatória Transcricional, numa dada árvore de espécies. Nessa etapa do fluxo de estudo os regulons podem ser plotados utilizando o pacote *RedeR*. Os valores de enraizamento evolutivo para cada gene da unidade regulatória se estabelecem como uma camada adicional na visualização, por meio de uma escala de cores atribuídas aos nós do grafo. Nesse tipo de representação gráfica, fica claro, pela coloração de nós se o ponto de enraizamento evolutivo da maioria dos genes alvo ocorreu previamente ou posteriormente ao ponto de enraizamento

evolutivo do elemento regulador (TF).

Após a análise evolutiva estabelecemos três etapas finais:

- Comparação entre reguladores e conjunto de genes alvo;
- Comparação entre Fatores de Transcrição e Cofatores de Transcrição;
- Análise de abundância, diversidade e plasticidade.

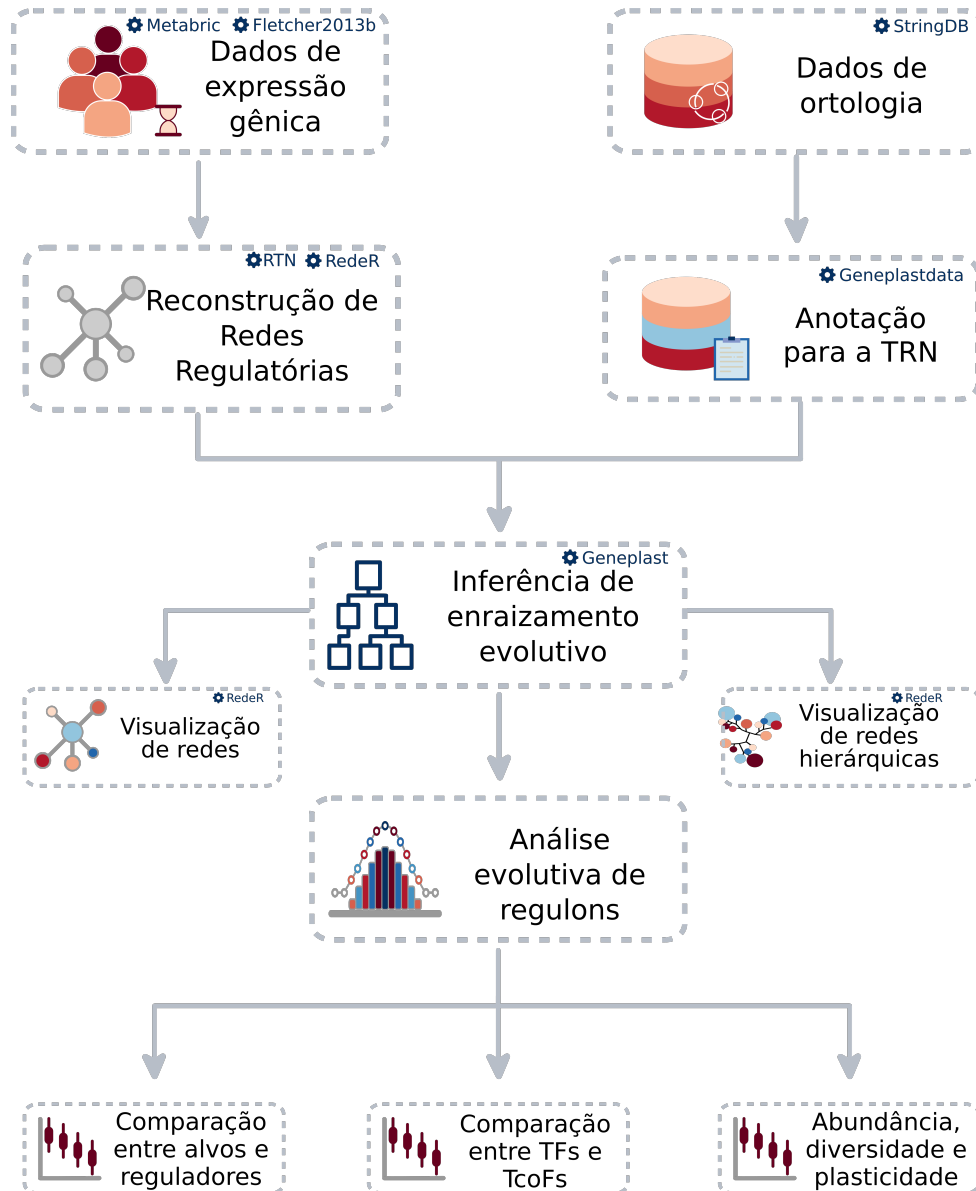


Figura 15. Fluxograma do estudo. Cada etapa da abordagem está representada no quadro, na ordem do *script* de análise. As engrenagens presentes em algumas etapas do fluxo descrevem as ferramentas utilizadas. A inferência do enraizamento evolutivo é precedida do mapeamento das redes regulatórias transcricionais, formadas a partir dos dados de expressão gênica, nos dados de anotação de ortologia. A abordagem permite a visualização das redes e segue com a análise evolutiva que se expande em três braços analíticos: comparação entre TFs - alvos; comparação entre TFs - TcoFs; análise de abundância, diversidade e plasticidade.

5 RESULTADOS E DISCUSSÃO

A abordagem criada neste trabalho se desenvolveu através dos dados de expressão gênica, provenientes do estudo Metabric (Curtis *et al.*, 2012). Esses dados foram anteriormente utilizados para a reconstrução das redes regulatórias transcricionais que foram inferidas com a ferramenta RTN (Castro *et al.*, 2015). As redes reconstruídas estavam prontas, e estão disponíveis na ferramenta Fletcher 2013b (Fletcher *et al.*, 2013).

Nosso objetivo neste trabalho foi atribuir a informação relativa ao ponto de surgimento evolutivo para cada um dos genes que já formavam a Rede Regulatória Transcricional de Câncer de Mama. Assim, com a utilização da ferramenta StringDB (Szkarczyk *et al.*, 2019), organizamos os dados de ortologia, para que permanecêssemos apenas com as informações de genes pertencentes à *H. sapiens*, por ser essa a espécie de estudo, já que as redes regulatórias transcricionais foram inferidas a partir dos dados de expressão gênica de pacientes com câncer de mama. A base de dados do StringDB traz informações de genes de muitos organismos e trabalhar com o banco completo tornaria as análises inviáveis. Assim, na sequência, filtramos os dados para obter apenas as informações de ortologia para os genes que faziam parte das redes regulatórias transcricionais. Houve perda de informação durante a análise por existirem genes componentes das redes regulatórias transcricionais que não possuíam informação de ortologia na base de dados. Vale ressaltar que essa é uma condicionante para análises que utilizam essa abordagem, só podemos analisar sob essa ótica evolutiva componentes de redes regulatórias transcricionais que possuam dados de ortologia. Com os dados de anotação de ortologia dos genes de interesse obtidos, foi possível inferir o ponto de enraizamento evolutivo para cada gene pertencente a rede regulatória transcricional.

De maneira hipotética criamos uma ilustração do processo de formação evolutiva dos regulons para possibilitar um melhor entendimento da abordagem criada neste trabalho. Considerando a estrutura de uma rede regulatória transcricional conforme mostrado na **Figura 16** podemos observar que ela é formada por regulons centrados em fatores de transcrição que possuem alvos regulatórios onde vão exercer papel de indução ou repressão no processo de transcrição gênica. A inferência dos regulons pode ser estimada utilizando algoritmos que avaliam a expressão do regulador (fator de transcrição) e seus potenciais genes alvo num conjunto de amostras, como faz o algoritmo ARACNe, o método mais utilizado (Janky *et al.*,

2014; Margolin *et al.*, 2006; Aibar *et al.*, 2017). Para a compreensão do cenário regulatório é importante considerar que os fatores de transcrição podem compartilhar poder regulatório de genes com outros fatores de transcrição. Vale ressaltar também que o estudo do cenário regulatório envolve a relação entre diversos elementos regulatórios em diversas camadas dos processos celulares, e nesse trabalho focamos apenas na atuação regulatória exercida pelos Fatores de Transcrição sob o conjunto de genes que ele regula.

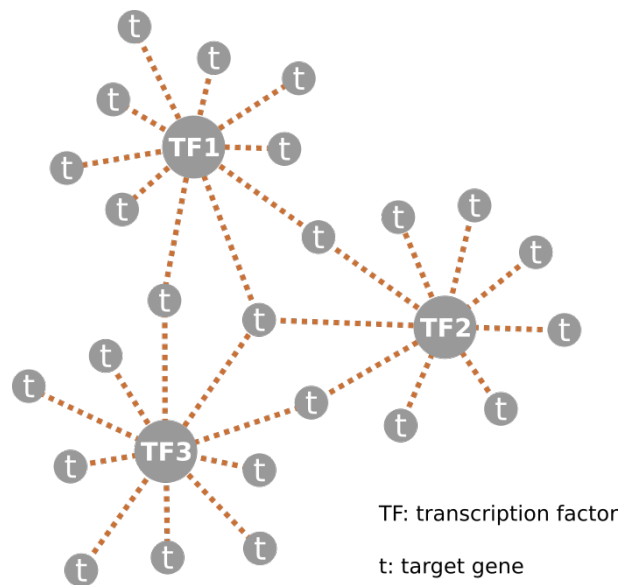


Figura 16. Rede Regulatória Transcricional. Rede utilizada na criação do cenário hipotético de herança vertical de genes anotados em uma árvore de espécies. A rede possui três regulons centrados em seus respectivos Fatores de Transcrição (TF1, TF2 e TF3) e o conjunto de genes alvo (t).

Considerando que os genes são herdados por transferência vertical através de ancestrais e seus descendentes, cada um dos genes formadores da rede regulatória transcricional enraizou em um ponto da árvore de espécies e podemos relacionar esse ponto com o tempo em que cada um desses genes surgiu. Assim é possível supor que a rede regulatória de *H. sapiens* foi sendo moldada ao longo do processo evolutivo, com o surgimento de cada um dos genes num determinado momento da evolução. Assim, é razoável supor que o conjunto de genes que formam um regulon estejam associados a uma mesma história evolutiva. Essa afirmação conduziu essa abordagem de maneira a compreendermos que os regulons são compostos por diversos genes, porém a relação regulatória entre esses elementos da maneira como observamos atualmente em *H. sapiens* só se estabelece ao longo do processo evolutivo na medida em que surgem os fatores de transcrição.

Na **Figura 17** ilustramos uma árvore de espécies que hipoteticamente conta a história evolutiva de todos os genes componentes da rede regulatória transcricional. Nessa figura representamos uma árvore de espécies com o organismo de referência do estudo sp01 localizado no topo da árvore, construída para esse exemplo com um total de 15 espécies e 10 LCAs. Nas **Figuras 17B-D** podemos observar o processo de formação dos regulons da rede regulatória

transcricional em cada etapa do processo evolutivo. De acordo com o ponto de enraizamento marcado pelos LCAs, o surgimento dos genes mais ancestrais da TRN ocorreu no LCA9 (círculo verde na árvore de espécies **Figura 17A**), e a relação regulatória entre eles está representada na **Figura 17D**. Posteriormente surgiram os genes enraizados em LCA6 (círculo vermelho na árvore de espécies **Figura 17A**), compondo novas relações mostradas na **Figura 17C**. Por fim surgiram os genes enraizados no LCA 3 (círculo azul na árvore de espécies **Figura 17A**), completando o cenário regulatório da rede regulatória transcricional na **Figura 17B** relativo à rede presente na **Figura 16**.

A **Figura 17** ilustra que para testar as hipóteses levantadas neste trabalho foi necessário reconstruir os cenários evolutivos para cada um dos genes pertencentes aos regulons de uma rede regulatória transcricional (Castro *et al.*, 2008). A construção de modelos que descrevem a origem dos regulons depende de um mapeamento sistemático dos genes numa dada árvore de espécies, inferido por análise de ortologia. A interação entre os elementos formadores das unidades regulatórias foi sendo construída ao longo do processo evolutivo. A relação entre elementos reguladores e regulados exerce uma forte influência sobre o perfil de expressão gênica observado em conjuntos de células (Chen *et al.*, 2020). Ainda que tenham sofrido pressão seletiva, os genes surgiram em pontos diferentes da evolução dos organismos (Koonin, 2005) e, podem atuar em grupo no direcionamento de um determinado fenótipo.

As relações de ortologia entre genes de diferentes espécies podem ser sistematicamente preditas por comparação entre genomas. Existem vários métodos para prever genes ortólogos e métodos baseados em grafos oferecem melhor velocidade e precisão para gerar grandes escala de anotação de ortologia (Trachana *et al.*, 2011). Para exemplificar o enraizamento evolutivo utilizando análise de ortologia, consideramos um grupo hipotético que compreende 6 genes ortólogos anotados em 3 espécies (**Fig. 18A**). É possível que, no processo de anotação, algumas relações de ortologia não estejam bem resolvidas, ilustradas como os genes parálogos YA1 e YA2, decorrentes de um processo de duplicação gênica. As **Figuras 18B-D** são uma representação adaptada do conceito de ortologia e paralogia de Koonin (2005).

Todos esses genes pertencem à mesma família ancestral, formada pelos genes ancestrais X e Y (**Fig. 18B**), herdados por três espécies existentes (spa, spb e spc) mas com diferentes histórias evolutivas. Os genes Xa, Xb e Xc mostram uma relação de ortologia entre si de um para um decorrente de transferência vertical do gene ancestral X(**Fig. 18C**). O gene X está presente nas três espécies (spa, spb, spc) com uma única cópia, enquanto o gene ancestral Y sofreu um evento de duplicação na spa (**Fig. 18D**). Os genes Ya1 e Ya2 são parálogos entre si e juntos são co-ortólogos do gene Yb. Um evento de deleção é representado pela ausência do gene ancestral Y na espécie spc. Cenários evolutivos mais complexos podem surgir quando eventos de duplicação e deleção ocorrem em ramos internos de uma árvore filogenética, mas em todos os casos apresentados nesta ilustração os genes podem ser anotados como pertencentes

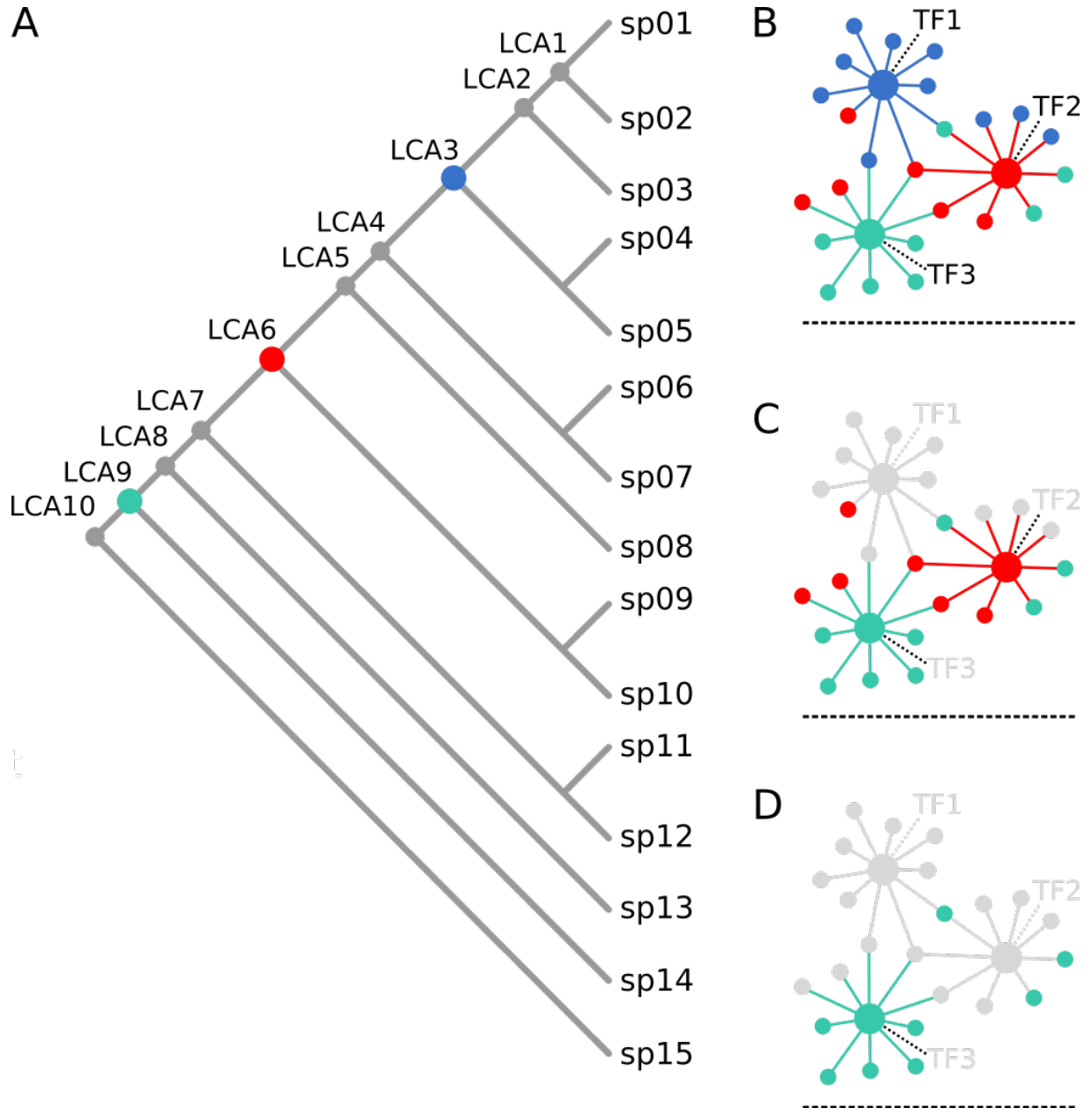


Figura 17. Cenário evolutivo hipotético para a formação da Rede Regulatória Transcricional. Em (a) a árvore de espécies que conta a história evolutiva dos genes formadores da rede regulatória transcricional, possuindo 15 espécies e 10 LCAs. No LCA9, LCA6 e LCA3 marcados com círculos verde, vermelho e azul respectivamente, são os pontos de enraizamento evolutivo dos grupos de genes formadores da TRN. Em (d) observamos o surgimento dos primeiros genes enraizados no LCA9, em (c) observamos o surgimento dos genes enraizados no LCA6 e em (b) observamos o surgimento do último grupamento de genes que completam a formação da TRN, enraizados no LCA3.

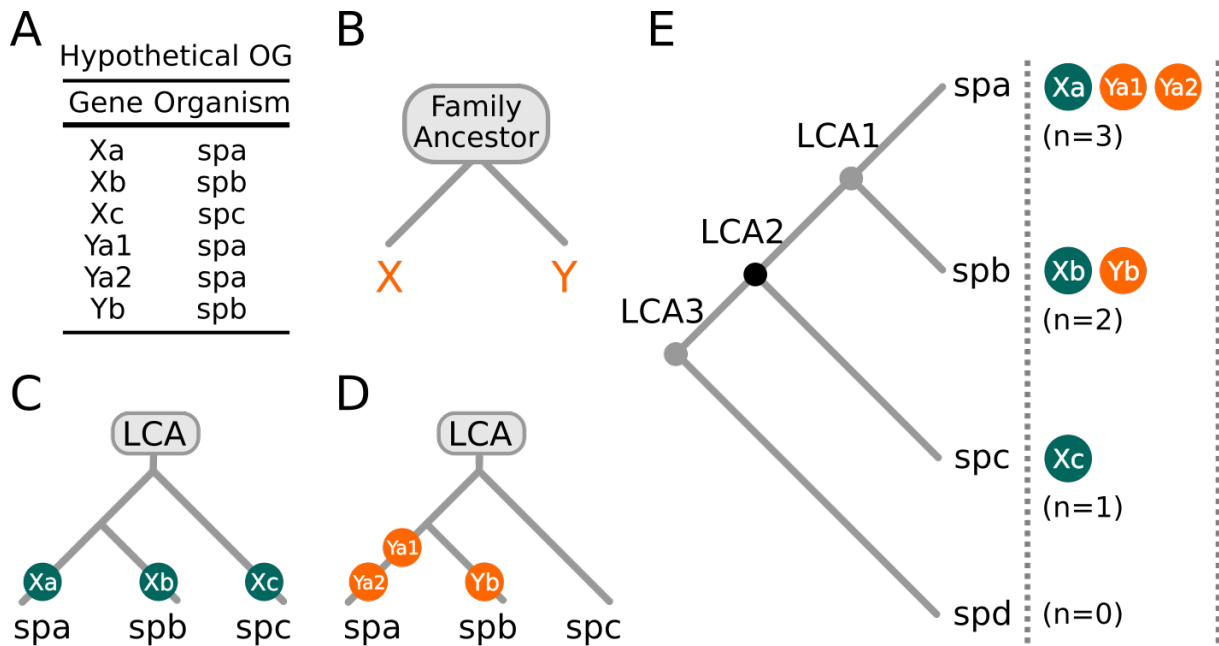


Figura 18. Rede Regulatória Transcricional. Rede utilizada na criação do cenário hipotético de herança vertical de genes anotados em uma árvore de espécies. A rede possui três regulons centralizados em seus respectivos Fatores de Transcrição (TF1, TF2 e TF3) e o conjunto de genes alvo (t).

ao mesmo grupo de ortólogos.

Com base na análise de distribuição de ortólogos em uma dada árvore de espécies é possível inferir a origem evolutiva dos genes de uma espécie. Definimos o organismo spa como referência na **Figura 18E** que lista todos os LCAs (LCA1, LCA2 e LCA3) e todos os genes anotados em cada espécie. Os ortólogos da spa estão enraizados no LCA2, uma vez que todos os descendentes de LCA2 têm pelo menos uma cópia da sequência ancestral. Esse cenário propõe que os genes Xa, Ya1 e Ya2 evoluíram de uma única sequência ancestral, a partir do LCA das espécies comparadas, por meio de uma série de processos de especiação e eventos de duplicação. É possível observar que a ausência do gene ancestral Y na espécie spc não altera a inferência da raiz evolutiva dos genes de spa, e a ausência dos genes ancestrais X e Y nas espécies spd indicam que o LCA3 não está relacionado com a história evolutiva do grupo de ortólogos.

A comparação entre genomas é uma das maneiras utilizadas para prever a relação de ortologia entre genes de diferentes espécies (Gabaldón e Koonin, 2013). O problema de inferir a raiz evolutiva de um gene em uma determinada árvore filogenética implica em encontrar o cenário de transferência vertical mais consistente para um conjunto de ortólogos, onde eventuais inconsistências devido a eventos como transferência horizontal, duplicação e eventos de deleção possam ser resolvidos com uma função que penalize os desvios do padrão filético (Mirkin *et al.*, 2003; Jacox *et al.*, 2016; Dondi *et al.*, 2019; Chan e Robin, 2019).

Sob uma perspectiva evolutiva na construção de regulons eucarióticos, a interação

de um TF com um gene alvo pode ser resumida por um modelo regulatório contendo três elementos (Shi *et al.*, 2019): o fator de transcrição; a região cis-reguladora contendo o sítio de ligação ao TF; o gene alvo. Podemos considerar modelos regulatórios mais complexos, por exemplo: quando a atividade do TF é modulada por um ou mais cofatores (Stampfel *et al.*, 2015); quando os locais de ligação do TF estão presentes em diferentes posições da região reguladora (Grossman *et al.*, 2018); quando fatores epigenéticos influenciam na acessibilidade dos locais de ligação (Wilson e Filipp, 2018); quando vários graus de afinidade de ligação são necessários para interações cooperativas (Jolma *et al.*, 2015) e especificidade de *enhancer* (Arozarena e Wellbrock, 2017). Em todos os casos, a ausência do fator de transcrição ou do gene alvo em um genoma implica na ausência de interação. Isso também é válido para a autorregulação, onde o TF se liga ao seu próprio promotor e ativa ou reprime o processo de transcrição (Ngondo e Carbon, 2014) (ou seja, como regulador e alvo são a mesma entidade na auto-regulação, a ausência do gene TF implica em ausência de interação). Por exemplo, o ganho de novas sequências cis reguladoras pode levar a reorganização da estrutura da rede transcricional (Baker *et al.*, 2012), mas não pode criar interações para alvos que não estão anotados no genoma.

Hormônios esteróides e receptores nucleares ilustram esse problema. Na ausência de um ligante, que função teria um novo receptor? E sem um receptor, quais pressões seletivas orientariam a evolução de um novo ligante? (Thornton, 2001). Embora essas questões possam oferecer uma estrutura básica para explorar processos evolutivos, elas também podem trazer percepções sobre os sistemas regulatórios observados em espécies existentes, como propor cenários evolutivos sobre a construção de unidades regulatórias: 1) O TF e seus alvos apareceram juntos? 2) O TF apareceu antes dos alvos que regula? 3) O TF apareceu após seus alvos?

A **Figura 20** indica que podemos encontrar regulons para cada um desses cenários, mas o cenário mais prevalente é aquele em que o TF surge antes de seus alvos. Na **Figura 19** mostramos a distribuição das raízes evolutivas de regulons construídos por Fletcher *et al.* (2013) em amostras de câncer de mama em humanos. Inferimos a raiz evolutiva para cada elemento da rede regulatória transcricional, formada por 307 TFs e 6308 alvos, em uma árvore de espécies contendo 121 organismos eucariotos (**Fig. 19A**). A **Figura 19B** mostra a distribuição geral das raízes evolutivas inferidas para TFs e alvos e indica que as raízes evolutivas dos TFs são anteriores às raízes evolutivas dos alvos (p -valor = $1e-6$, teste de Wilcoxon-Mann-Whitney).

De acordo com a distribuição geral das raízes evolutivas observamos um acúmulo de genes enraizados no *root 25* que é o LCA mais ancestral nessa árvore de espécies. Esse achado corrobora com a literatura que considera o genoma humano uma quimera de genes herdados de eubactérias e arqueobactérias, ou seja, originados em organismos pré-eucarióticos que permaneceram ao longo de mais de 2 bilhões de anos de evolução ocupando papel central no

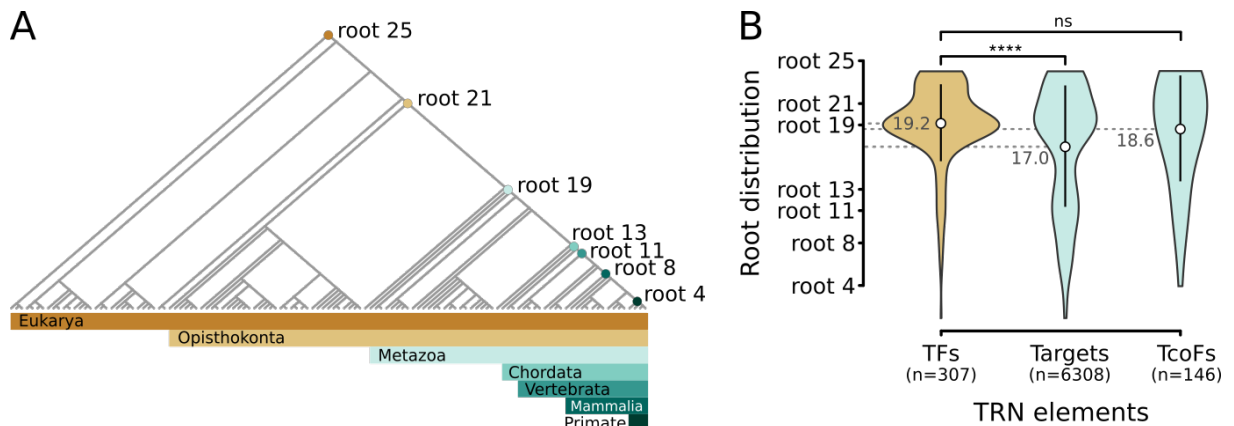


Figura 19. Comparação entre o ponto de enraizamento evolutivo de Tfs, alvos e TcoFs. (A) Uma árvore de espécies com os LCAs enumerados do mais recente ao nó mais ancestral do *Homo sapiens*. (B) Distribuição das raízes evolutivas inferidas para TFs e genes-alvo. Co-fatores de transcrição (TcoFs) também estão incluídos em a análise.

funcionamento do organismo, e a perda ou modificação desses genes resultaria em letalidade (Alvarez-Ponce e McInerney, 2011). DeMendoza (2019) afirma que apesar dos TFs serem encontrados em espécies pré-eucarióticas, muitos TFs emergiram na raiz dos eucariotos e que esses elementos acompanharam as mudanças no ambiente da cromatina nuclear e demais modificações químicas como o surgimento dos nucleossomos, por exemplo. A modificação de histonas, o controle da acessibilidade da cromatina e demais processos que começaram a ocorrer no DNA com o surgimento dos eucariotos dependeram do papel crucial na regulação do genoma executado pelos fatores de transcrição (DeMendoza 2019).

A **Figura 19B** também mostra a distribuição das raízes evolutivas inferidas para outra classe de reguladores, os cofatores de transcrição (TcoFs). Utilizando a mesma abordagem, não foi detectado uma mudança significativa na distribuição de raízes evolutivas entre TFs e TcoFs (p-valor = 0,884, teste de Wilcoxon-Mann-Whitney), ou seja, durante a evolução esses elementos surgiram no mesmo ponto de enraizamento evolutivo. Os cofatores de transcrição foram avaliados quanto a distribuição das raízes evolutivas pois influenciam no processo de regulação da transcrição pela formação de complexos de proteínas com TFs (Schmeier *et al.*, 2016).

A distribuição geral das raízes evolutivas (**Fig. 19B**) mostra uma grande concentração de TFs no *root 19* onde se inicia o surgimento dos organismos metazoários enraizados de acordo com a árvore de espécies. Os metazoários são os primeiros organismos multicelulares, ou seja, foram nesses organismos que as células começaram a ter característica de especialização em diferentes funções e assim foram se dividindo em tecidos. De acordo com Degnan *et al.* (2009) o surgimento e a expansão dos TFs em metazoários contribuiu com o sucesso evolutivo da multicelularidade animal (Degnan *et al.*, 2009).

A **Figura 20** estende a abordagem para regulons individuais, agrupando-os com base

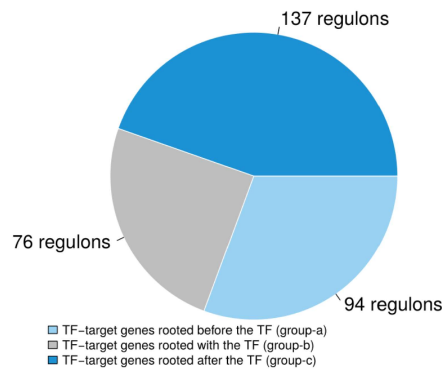


Figura 20. Comparação entre os três grupos de regulons. Os regulons se dividem em três grupos com base na distância média entre as raízes evolutivas inferidas para um TF e seus genes-alvo.

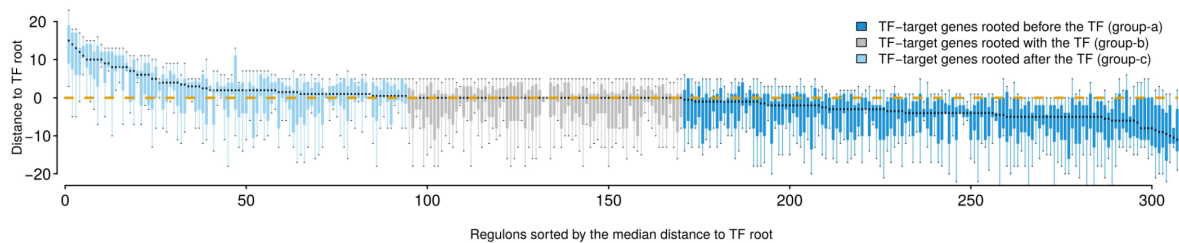


Figura 21. Comparação entre os três grupos de regulons num box plot detalhado. Box plot mostrando regulons individuais. **** valor de $p = 1e-6$ (teste de Wilcoxon-Mann-Whitney); ns = não significativo.

na distância entre as raízes evolutivas dos TFs e alvos: Grupo-a) regulons onde genes alvo são enraizados antes do TF (30,6% dos regulons); Grupo-b) regulons onde genes alvo são enraizados com o TF (24,7% dos regulons); Grupo-c) regulons onde genes alvo são enraizados após o TF (44,3% dos regulons). Esses resultados mostram que encontramos regulons para cada um dos cenários propostos, mas o mais prevalente é aquele em que o TF surge antes de seus alvos. Uma distribuição detalhada das raízes evolutivas inferidas para regulons individuais é fornecida na **Figura 21**. Koster *et al.* (2015) afirmam que a resposta gênica à um programa de expressão é uma propriedade compartilhada por todas as espécies e com o aumento no tamanho e na complexidade dos genomas, os organismos criaram mecanismos capazes de garantir a dinâmica transcricional (Koster *et al.*, 2015).

Posteriormente mapeamos a abundância, diversidade e plasticidade entre os regulons do Grupo-a e Grupo-c, por serem os grupos em que houve diferença do ponto de enraizamento evolutivo de TFs e alvos, mostradas na **Figura 22**. A abundância mapeada para regulons cujos genes alvo estão enraizados antes do TF (Grupo-a) é a mesma daquela mapeada para regulons cujos genes alvo estão enraizados após o TF (Grupo-c) (**Fig. 22a**), sugerindo que o número de ortólogos por espécie é semelhante entre os dois grupos. Em contraste, a diversidade do

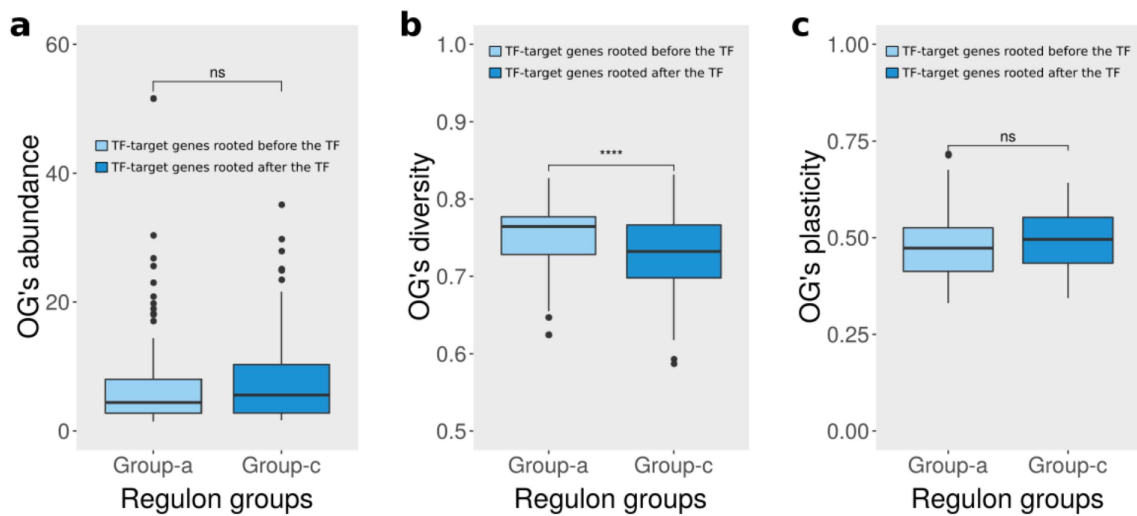


Figura 22. Abundância, Diversidade e Plasticidade. (a) Análise de abundância, (b) Análise de diversidade e (c) análise de plasticidade mapeados para regulons baseados no agrupamento de distância entre enraizamento evolutivo de TFs e alvos. ****P-value = $5.1e-5$ (Teste de Wilcoxon-Mann-Whitney), ns = não significante.

grupo de ortólogos mapeada para o Grupo-a é maior em comparação com o Grupo-c (P-valor = $5,1e-5$; teste de Wilcoxon-Mann-Whitney) (Fig. 22b). Como a diversidade estima a dispersão dos genes ortólogos na árvore da espécie, esses resultados sugerem que os regulons do Grupo-a têm ortólogos mais uniformemente distribuídos, o que geralmente é observado para grupos de ortólogos enraizados na base da árvore filogenética (Castro et al. 2008). Não detectamos nenhuma diferença entre o Grupo-a e o Grupo-c usando os escores de plasticidade mapeados para regulons (Fig. 22c).

As redes regulatórias transcricionais podem ser plotadas em grafos utilizando o pacote RedeR (Castro *et al.*, 2012). A análise visual auxilia no entendimento de como pode ser o cenário regulatório entre poucos regulons. Com o desenvolvimento dessa abordagem, se tornou possível plotar o grafo dos regulons com a informação de ortologia para cada um dos genes contidos nos regulons escolhidos para a ilustração (Fig. 23). Nesse grafo utilizamos uma escala de coloração onde estão posicionados todos os pontos de enraizamento evolutivo, cada um com uma cor, possibilitando a análise visual de ortologia dos regulons. Os pontos de enraizamento evolutivo totalizam 25 e são decorrentes da árvore de espécies utilizada nessa análise (Fig. 12). Nesse exemplo plotamos os regulons de PTTG1, FOXM1 e GATA3. Consideramos na escala que o número 1 se refere ao topo da árvore de espécies, portanto ao ponto de enraizamento evolutivo de *Homo sapiens*, o organismo de estudo, e que o número 25 se refere a base da árvore de espécies ao organismo eucarioto mais ancestral *Trichomonas vaginalis*. Observamos o fator de transcrição PTTG1 tem ponto de enraizamento evolutivo posterior ao ponto de enraizamento de seus alvos, já o TF FOXM1 e GATA3 tem pontos de surgimento evolutivo próximo aos

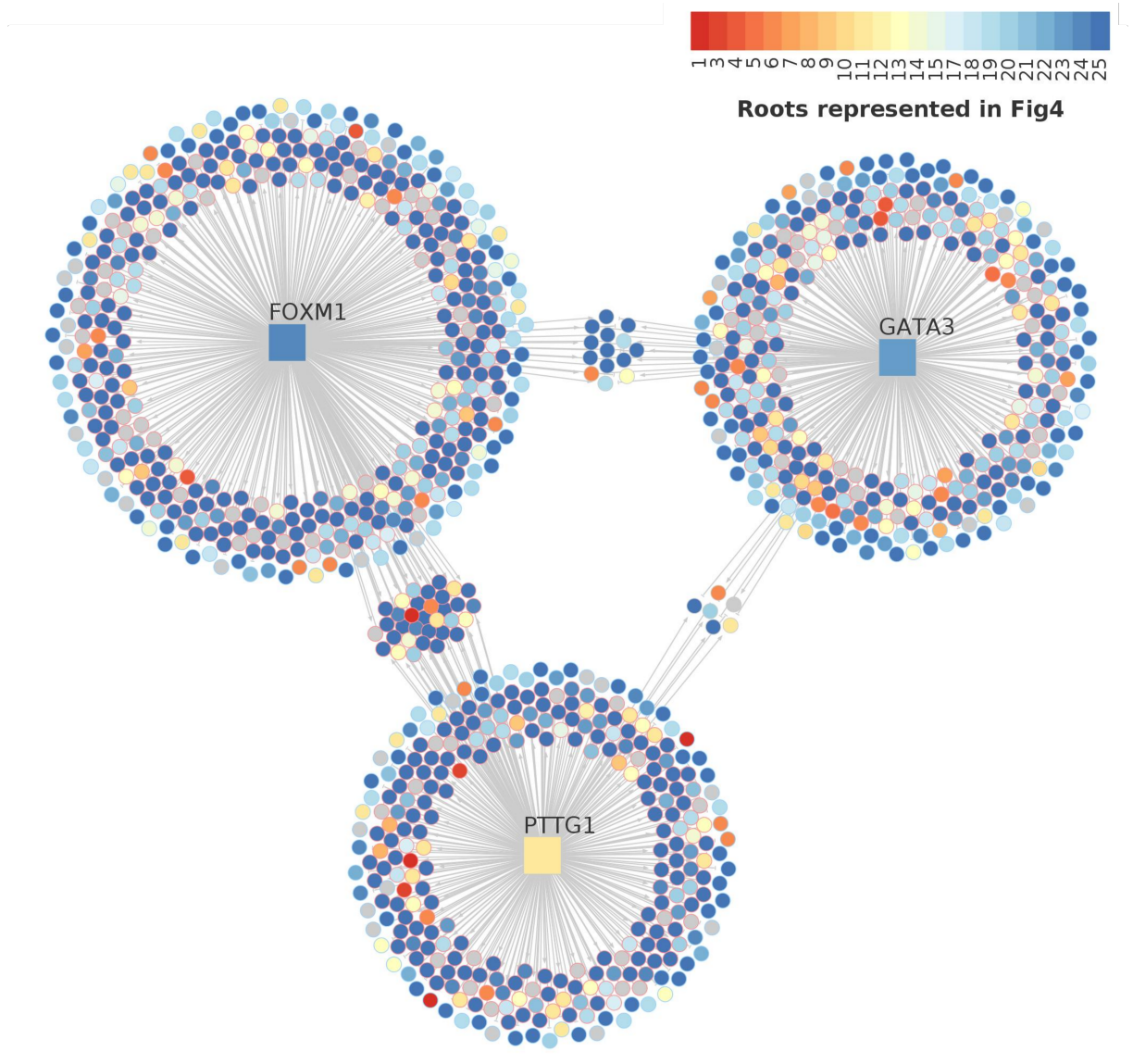


Figura 23. Grafo de três regulons com informação do ponto de enraizamento evolutivo. Os Fatores de Transcrição PTTG1, FOXM1 e GATA3 estão representados por quadrados e o conjunto de genes alvo que eles regulam são representados por círculos. As cores dos elementos geométricos estão relacionadas ao valor do *root* disposto na régua com gradiente de coloração, localizada no canto superior direito. Círculos cinzas representam genes para os quais não havia informação de ortologia e portanto não pudemos estabelecer o ponto de enraizamento evolutivo para esses genes.

alvos que regulam, sugerindo um cosurgimento desses fatores de transcrição e os alvos que eles regulam. Esse tipo de ilustração auxilia a análise visual isolada de poucos regulons, direcionados a responder a pergunta central deste trabalho: Quem surgiu primeiro, reguladores ou regulados?

O artigo de Castro *et al.* (2015) ilustra regulons clusterizados de acordo com a correlação da expressão gênica de alvos compartilhados. Utilizando esse dendrograma aplicamos a abordagem desenvolvida nesta tese, colorindo os nodos de acordo com a mediana do enraizamento evolutivo dos alvos de cada regulon (**Fig. 24b**). Nessa ilustração, cada nodo do grafo está relacionado a um regulon, ou seja, o conjunto formado por um TF e todos os alvos por ele regulados. O tamanho dos nodos está relacionado ao número de alvos que cada regulon possui. O posicionamento dos nodos obedece a correlação de alvos compartilhados, ou seja, nodos próximos possuem alta correlação entre seus alvos. Numa escala de cores correspondente aos grupos coloridos na árvore de espécies mostrada na **Figura 24a** está o posicionamento do enraizamento evolutivo. É possível observar dois *clusters*, um no topo e outro na base do dendrograma, delimitados com círculos pontilhados, que representam regulons associados ao risco do desenvolvimento de câncer de mama ER+ e ER- respectivamente.

Decorrente dessa análise observamos alguns padrões marcantes. 1) Em geral, regulons que possuem a mediana de enraizamento evolutivo do seu conjunto alvos em pontos evolutivos próximos também possuem alta correlação na expressão de alvos compartilhados e, portanto, se agrupam; 2) A maioria dos regulons tem enraizamento evolutivo em LCA de organismos unicelulares, demonstradas no grafo por uma coloração mais esverdeada; 3) Regulons associados ao desenvolvimento do câncer de mama são mais recentes, aqueles contidos nos dois círculos pontilhados no topo e na base do grafo; 4) A maioria dos regulons relacionados a tumores estrogênio positivo e estrogênio negativo estão enraizados em LCA de metazoários.

De acordo com Siddiqui *et al.* (2020) durante a formação tumoral a acidificação do pH, hipóxia e baixa oferta de nutrientes, entre outros fatores, induzem as células a buscar um estilo de vida atavístico, ou seja, manifestar características de organismos ancestrais em busca de sobrevivência adquirindo comportamentos de organismos unicelulares (Siddiqui *et al.*, 2020). Ou seja, o ambiente celular hostil permite a mudança no padrão de expressão de genes para que haja a mudança adaptativa da célula no meio onde ela se encontra. Dentre essas características adaptativas destacamos o potencial de motilidade muito utilizado pelas células tumorais na capacidade de evasão, desenvolvendo uma forma primitiva de memória associativa que as permite responder eficientemente às mudanças ambientais que ocorrem durante a migração (De la Fuente e López, 2020). Esses resultados são consistentes com um dos principais aspectos do câncer, em que só é possível observar desorganização tecidual em organismos que possuem um processo de diferenciação celular capaz de formar tecidos e ainda que as células tumorais

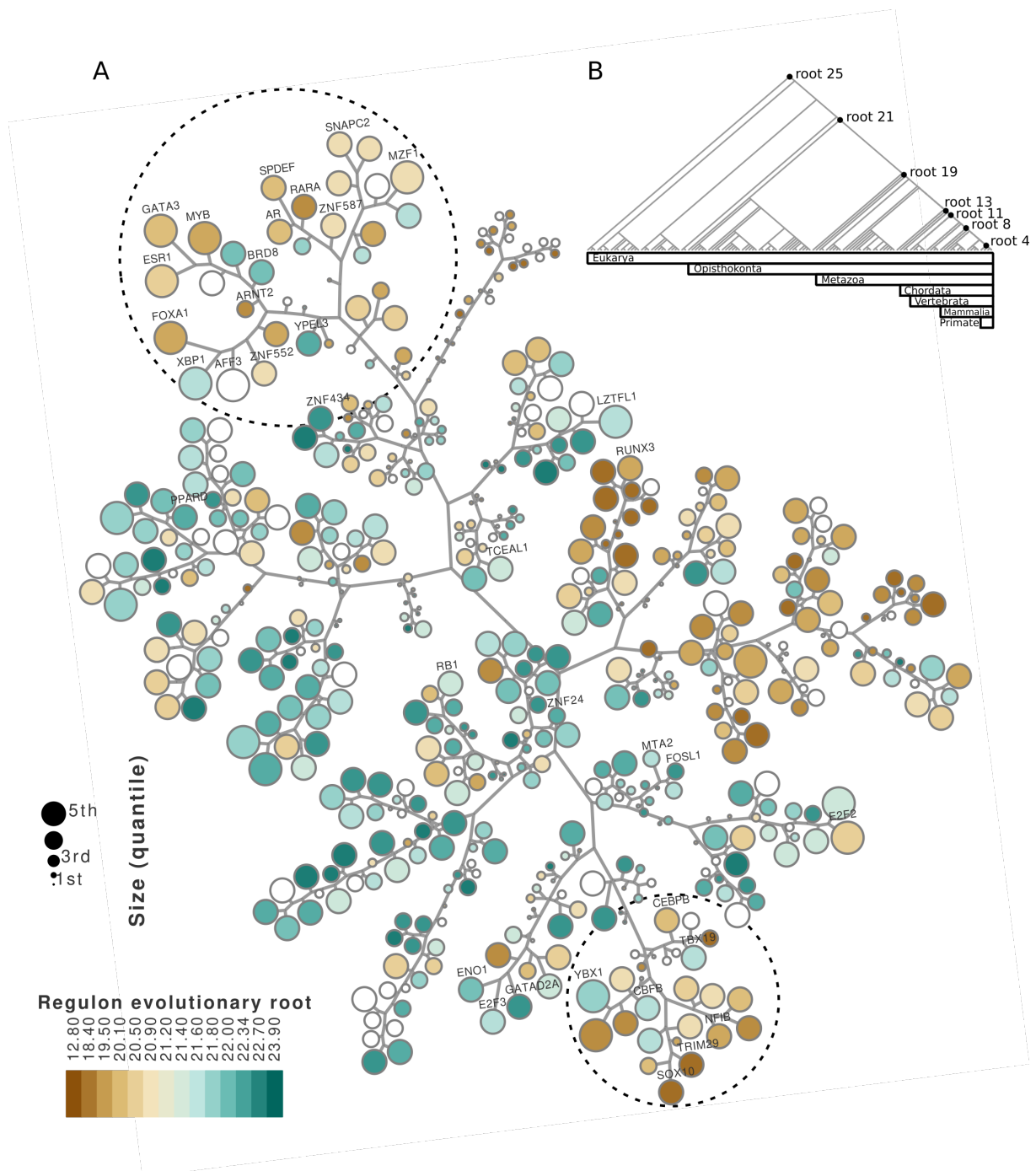


Figura 24. Representação da árvore de enraizamento evolutivo com uma Tree and Leaf. (a) Tree and Leaf da clusterização dos regulons de acordo com a função que cada regulon exerce, nos círculos pontilhados se encontram os dois clusters com os Fatores de Transcrição de risco para o desenvolvimento do câncer de mama. O tamanho dos nodos representa a quantidade de alvos dos regulons, as cores são relacionadas à média do enraizamento dos alvos dos reguladores. (b) Visualização dos últimos ancestrais comuns, de acordo com o posicionamento de cada raiz.

manifestam um comportamento de desdiferenciação celular se assemelhando às características de organismos mais primitivos(Friedmann-Morvinski e Verma, 2014; Yuan *et al.*, 2020).

6 CONCLUSÃO

Os regulons configuram o maior sistema regulatório de controle da expressão gênica em resposta aos estímulos ambientais (Chiang e Schellhorn, 2010). Conjuntos de genes com funcionalidades geralmente relacionadas à manutenção celular são conservados em alguns regulons (Erill *et al.*, 2007), porém a modificação dos regulons e a manutenção de genes não essenciais configuram mecanismos importantes na adaptação das espécies (Chiang e Schellhorn, 2010). Em todas as espécies os regulons são elementos regulatórios com característica de alta plasticidade (Lozada-Chávez *et al.*, 2006). As pressões seletivas as quais os genes estão submetidos durante o processo evolutivo permitem mudanças específicas da linhagem que compõe um regulon, mas também trabalham na manutenção de genes essenciais como componentes desses regulons (Chiang e Schellhorn, 2010). Em organismos complexos as pressões seletivas podem divergir favorecendo a aptidão do organismo ou a aptidão da célula, pois ao mesmo tempo que uma célula de rápida proliferação é selecionada naturalmente em organismos unicelulares, em organismos complexos ela pode favorecer a instalação de um tumor e desfavorecer a vida do organismo (Castro *et al.*, 2008).

Nesta abordagem, desenvolvemos uma estrutura de análise de hipóteses sobre a reconfiguração de regulons, explorando a presença e ausência dos elementos que formam uma rede regulatória transcricional ao longo da evolução de uma linhagem ancestral. A distância entre as raízes evolutivas de um fator de transcrição e seus alvos é uma informação adicional no estudo das redes regulatórias transcricionais. Uma vez que as interações entre TFs e dos alvos que ele regula conferem especificidade de tecido para uma rede regulatória transcricional, prevemos que esta abordagem contribua com estudos que exploram a reorganização dos regulons em diferentes condições. O entendimento do cenário regulatório envolve a utilização de diversas técnicas analíticas considerando o seu elevado grau de complexidade e a utilização de camadas informativas adicionais na sua análise auxiliam na clareza desses estudos.

As produções diretas derivadas da abordagem central desse trabalho estão disponíveis nos **ANEXOS B e C** dessa tese. Dentre eles destacamos a publicação principal, "Which came first, the transcriptional regulator or its target genes? An evolutionary perspective into the construction of eukaryotic regulons" de Trefflich *et al.*, 2019 onde conta a maior parte dos resultados que compõem essa tese. Porém, o desenvolvimento do estudo envolvendo análises

evolutivas das redes regulatórias transcricionais resultou na participação colaborativa em outras publicações apresentadas nos **ANEXOS D-H** deste documento.

7 PERSPECTIVAS

Apresentamos as perspectivas futuras decorrentes dessa tese:

- A partir do trabalho desenvolvido por Castro *et al.* (2015), é necessário o amadurecimento dos resultados desta tese no entendimento evolutivo da *tree and leaf* combinado com análises funcionais.
- O desenvolvimento de um método *in silico* para prospecção de alvos imunológicos assistido por análise regulatória. Mais especificamente, o desenvolvimento de uma nova abordagem computacional para detecção de assinaturas imunológicas no infiltrado linfocitário tumoral, auxiliando a prospecção de alvos imunológicos com uso de redes regulatórias. Esse projeto configura uma proposta de Pós-doc a ser submetida.

REFERÊNCIAS

- ABDELWAHAB, Y. J. Male breast cancer: Epidemiology and risk factors. **Semin Oncol.**, v. 44, n. 4, p. 267–272, 2017.
- ABDULRAHMAN, G. O.; RAHMAN, G. A. Epidemiology of breast cancer in europe and africa. **Journal of Cancer Epidemiology**, v. 2012, p. 1–5, 2012.
- AGGARWAL, K.; LEE, K. H. Functional genomics and proteomics as a foundation for systems biology. **Briefings in Functional Genomics and Proteomics**, v. 2, n. 3, p. 175–184, 2003.
- AIBAR, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. **Nature Methods**, v. 14, n. 11, p. 1083–1086, 2017.
- ALTENHOFF, A. M.; GLOVER, N.; DESSIMOZ, C. Inferring Orthology and Paralogy. **Methods Mol Biol.**, v. 1910, p. 149–175, 2019.
- ALVAREZ-PONCE, D.; MCINERNEY, J. O. The human genome retains relics of its prokaryotic ancestry: Human genes of archaeobacterial and eubacterial origin exhibit remarkable differences. **Genome Biology and Evolution**, v. 3, n. 1, p. 782–790, 2011.
- ARNETH, B. Comparison of Burnet’s clonal selection theory with tumor cell-clone development. **Theranostics**, v. 8, n. 12, p. 3392–3399, 2018.
- ARNOLD, C.; GERLACH D.AND SPIES, D.; MATTS, J. A. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. **Nature Genetics**, v. 46, n. 7, p. 685–692, 2014.
- AROZARENA, I.; WELLBROCK, C. Overcoming resistance to BRAF inhibitors. **Annals of Translational Medicine**, v. 5, n. 19, p. 1–12, 2017.
- BAKER, C. *et al.* Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. **Cell**, v. 151, n. 1, p. 80–95, 2012.
- BECKMAN, R. A.; LOEB, L. A. Negative clonal selection in tumor evolution. **Genetics**, v. 171, n. 4, p. 2123–2131, 2005.
- BERNARD, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. **Journal of Clinical Oncology**, v. 27, n. 8, p. 1160–1167, 2009. ISSN 0732183X.
- BERNARDES, A. Anatomia da mama feminina. p. 167–174, 2010.
- BLOWS, F. M. *et al.* Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: A collaborative analysis of data for 10,159 cases from 12 studies. **PLoS Medicine**, v. 7, n. 5, 2010. ISSN 15491277.
- BRAHMA, S.; HENIKOFF, S. Epigenome Regulation by Dynamic Nucleosome Unwrapping. **Trends in Biochemical Sciences**, v. 45, n. 1, p. 13–26, 2020.

- CALHOUN, V. C.; STATHOPOULOS, A.; LEVINE, M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila* Antennapedia complex. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 14, p. 9243–9247, 2002.
- CAMPBELL, T. M. *et al.* Era binding by transcription factors NFIB and YBX1 enables FGFR2 signaling to modulate estrogen responsiveness in breast cancer. **Cancer Research**, v. 78, n. 2, p. 410–421, 2018.
- CARDOSO, M. *et al.* Novel lincRNAs co-expression networks identifies linc00504 with oncogenic role in luminal a breast cancer cells. **XXX**, xx, n. x, p. 1–15, 2021.
- CARTER, G. W. Inferring network interactions within a cell. **Briefings in Bioinformatics**, v. 6, n. 4, p. 380–389, 2005.
- CASTRO, M. A. *et al.* Evolutionary origins of human apoptosis and genome-stability gene networks. **Nucleic Acids Research**, v. 36, n. 19, p. 6269–6283, 2008.
- CASTRO, M. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. **Nature Genetics**, v. 48, n. 1, p. 12–21, 2015.
- CASTRO, M. A. *et al.* RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. **Genome Biology**, v. 13, n. 4, 2012.
- CHAGAS, V. S. **RTNduals: Ferramenta para análise de co-regulação entre regulons e inferência de dual**, Dissertação (Mestrado em Bioinformática) - Pós-graduação em Bioinformática, Universidade Fedearl do Paraná. Curitiba, p. 1–68, 2017.
- CHAGAS, V. S. *et al.* RTNduals: An R/Bioconductor package for analysis of co-regulation and inference of dual regulons. **Bioinformatics**, v. 35, n. 24, p. 5357–5358, 2019.
- CHAN, Y. ban; ROBIN, C. Reconciliation of a gene network and species tree. **Journal of Theoretical Biology**, v. 472, p. 54–66, 2019.
- CHEN, C. H. *et al.* Determinants of transcription factor regulatory range. **Nature Communications**, v. 11, n. 1, p. 1–15, 2020.
- CHEN, X.; JORGENSON, E.; CHEUNG, S. T. New tools for functional genomic analysis. **Drug Discovery Today**, v. 14, n. 15-16, p. 754–760, 2009.
- CHENG, Y. *et al.* Principles of regulatory information conservation between mouse and human. **Nature**, v. 515, n. 7527, p. 371–375, 2014.
- CHIANG, S. M.; SCHELLHORN, H. E. Evolution of the RpoS regulon: Origin of RpoS and the conservation of RpoS-dependent regulation in bacteria. **Journal of Molecular Evolution**, v. 70, n. 6, p. 557–571, 2010.
- CORCES, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. **Science**, v. 362, n. 6413, p. 1–13, 2018.
- COUX, R. X.; OWENS, N. D.; NAVARRO, P. Chromatin accessibility and transcription factor binding through the perspective of mitosis. **Transcription**, v. 11, n. 5, p. 236–240, 2020.

- CRESPI, B.; SUMMERS, K. Evolutionary biology of cancer. **Trends in Ecology and Evolution**, v. 20, n. 10, p. 545–552, 2005.
- CRUZ, M. *et al.* Protein Function Prediction. **Functional Genomics**, 2017.
- CULJKOVIC, B.; TOPISIROVIC, I.; BORDEN, K. L. B. Controlling Gene Expression through RNA Regulons The Role of the Eukaryotic Translation Initiation Factor eIF4E. **Cell Cycle**, v. 61, n. 1, p. 65–69, 2007.
- CURTIS, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. **Nature**, v. 486, n. 7403, p. 346–352, 2012.
- DAI, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. **Am J Cancer Res**, v. 5, n. 10, p. 2929–2943, 2015. ISSN 2156-6976. Disponível em: <www.ajcr.us>.
- DALMOLIN, R. J.; CASTRO, M. **Geneplast: evolutionary rooting and plasticity inference. R package version 1.0.0.** 2015.
- De la Fuente, I. M.; LÓPEZ, J. I. Cell motility and cancer. **Cancers**, v. 12, n. 8, p. 1–15, 2020.
- DEGNAN, B. M. *et al.* Early evolution of metazoan transcription factors. **Current Opinion in Genetics and Development**, v. 19, n. 6, p. 591–599, 2009.
- DONDI, R.; LAFOND, M.; SCORNAVACCA, C. Reconciling multiple genes trees via segmental duplications and losses. **Algorithms for Molecular Biology**, v. 14, n. 1, p. 1–19, 2019.
- ERILL, I.; CAMPOY, S.; BARBÉ, J. Aeons of distress: An evolutionary perspective on the bacterial SOS response. **FEMS Microbiology Reviews**, v. 31, n. 6, p. 637–656, 2007.
- ERILES, P. *et al.* Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. **Cancer Treatment Reviews**, Elsevier Ltd, v. 38, n. 6, p. 698–707, 2012. ISSN 03057372. Disponível em: <<http://dx.doi.org/10.1016/j.ctrv.2011.11.005>>.
- FLEISCHMANN, W. *et al.* A novel method for automatic functional annotation of proteins. **Bioinformatics**, v. 15, p. 228–33, 1999.
- FLETCHER, M. N. C. *et al.* Master regulators of FGFR2 signalling and breast cancer risk. **Nature Communications**, 2013.
- FLETCHER, M. N. C. *et al.* Vignette for Fletcher2013b : master regulators of FGFR2 signalling and breast cancer risk . 2014.
- FRANCESCHINI, A. STRINGdb Package Vignette. n. July, p. 1–13, 2013.
- FRIEDMANN-MORVINSKI, D.; VERMA, I. M. Dedifferentiation and reprogramming: Origins of cancer stem cells. **EMBO Reports**, v. 15, n. 3, p. 244–253, 2014.
- GAASTERLAND, T.; SENSEN, C. Fully automated genome analysis that reflects user needs and preferences: A detailed introduction to the magpie system architecture. **Biochimie**, v. 78, p. 302–10, 1996.

- GABALDÓN, T.; KOONIN, E. V. Functional and evolutionary implications of gene orthology. **Nature Reviews Genetics**, v. 14, n. 5, p. 360–366, 2013.
- GEHLENBORG, N. *et al.* Visualization of omics data for systems biology. **Nature Methods**, v. 7, n. 3, p. S56–S68, 2010.
- GIORDANO, S. H.; BUZDAR, A. U.; HORTOBAGYI, G. N. Breast cancer in men. **Ann Intern Med.**, v. 138, n. 8, p. 678–87, 2002.
- GREENMAN, C. D. *et al.* Estimation of rearrangement phylogeny for cancer genomes. **Genome research**, v. 22, p. 346–361, 2012.
- GROSSMAN, S. R. *et al.* Positional specificity of different transcription factor classes within enhancers. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 30, p. E7222–E7230, 2018.
- HANAHAN, D.; WEINBERG, R. A. A. Hallmarks of Cancer: The Next Generation Douglasle. **Cell**, v. 144, n. 5, p. 646–674, 2011.
- HILLMER, R. A. Systems Biology for Biologists. **PLoS Pathogens**, v. 11, n. 5, p. 1–7, 2015.
- HORTOBAGYI, G. N. *et al.* The Global Breast Cancer Burden: Variations in Epidemiology and Survival. **Clinical Breast Cancer**, v. 6, n. 5, p. 391–401, 2005.
- HUBER, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, v. 12, n. 2, p. 115–121, 2015.
- JACOX, E. *et al.* EcceTERA: Comprehensive gene tree-species tree reconciliation using parsimony. **Bioinformatics**, v. 32, n. 13, p. 2056–2058, 2016.
- JAENISCH, R.; BIRD, A. review Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. **Nature Genetics**, v. 33, n. 3S, p. 245–254, 2003.
- JANKY, R. *et al.* iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. **PLoS Computational Biology**, v. 10, n. 7, 2014.
- JOLMA, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. **Nature**, Nature Publishing Group, v. 527, n. 7578, p. 384–388, 2015.
- KARCZEWSKI, K. J.; SNYDER, M. P. Integrative omics for health and disease. **Nature Reviews Genetics**, 2018.
- KEENE, J. D. RNA regulons: coordination of post-transcriptional events. **Nature Reviews Genetics**, v. 8, n. 7, p. 533–543, 2007.
- KLEIN, C. A. Parallel progression of tumour and metastases. **Nature Reviews Cancer**, v. 10, n. 2, p. 156, 2010.
- KO, J. Y.; OH, S.; YOO, K. H. Functional enhancers as master regulators of Tissue-Specific gene regulation and cancer development. **Molecules and Cells**, v. 40, n. 3, p. 169–177, 2017.
- KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. **Annual Review of Genetics**, v. 39, p. 309–338, 2005.

- KOSTER, M. J.; SNEL, B.; TIMMERS, H. T. M. Genesis of chromatin and transcription dynamics in the origin of species. **Cell**, v. 161, n. 4, p. 724–736, 2015.
- KREFT, L. *et al.* ConTra v3: A tool to identify transcription factor binding sites across species, update 2017. **Nucleic Acids Research**, v. 45, n. W1, p. W490–W494, 2017.
- LACINA, L. *et al.* Evolution of cancer progression in the context of Darwinism. **Anticancer Research**, v. 39, n. 1, p. 1–16, 2019.
- LAMBERT, S. A. *et al.* The Human Transcription Factors. **Cell**, v. 172, n. 4, p. 650–665, 2018.
- LEE, S. K. *et al.* Characteristics of metastasis in the breast from extramammary malignancies. **Journal of Surgical Oncology**, v. 101, n. 2, p. 137–140, 2010.
- LEFEBVRE, C.; RIECKHOF, G.; CALIFANO, A. Reverse-engineering human regulatory networks. **Wiley Interdisciplinary Reviews: Systems Biology and Medicine**, v. 4, n. 4, p. 311–325, 2012.
- LI, L. *et al.* Association between oral contraceptive use as a risk factor and triple-negative breast cancer: A systematic review and meta-analysis. **Molecular and Clinical Oncology**, v. 7, n. 1, p. 76–80, 2017. ISSN 2049-9450.
- LIKIC, V. A. *et al.* Systems Biology: The Next Frontier for Bioinformatics. **Advances in Bioinformatics**, v. 20, p. 1–10, 2010.
- LONGO, D. L.; GIORDANO, S. H. Breast Cancer in Men. **N Engl J Med**, v. 378, p. 2311–2331, 2018.
- LOZADA-CHÁVEZ, I.; JANGA, S. C.; COLLADO-VIDES, J. Bacterial regulatory networks are extremely flexible in evolution. **Nucleic Acids Research**, v. 34, n. 12, p. 3434–3445, 2006.
- MA, X.-J. *et al.* Gene expression profiles of human breast cancer progression. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 10, p. 5974–5979, 2003.
- MARGOLIN, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. **BMC Bioinformatics**, v. 7, n. Suppl 1, p. S7, 2006.
- MATHIAS, C. *et al.* Novel lncRNAs co-expression networks identifies linc00504 with oncogenic role in luminal a breast cancer cells. **International Journal of Molecular Sciences**, v. 22, n. 5, p. 1–15, 2021.
- MIRKIN, B. G. *et al.* Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. **BMC Evolutionary Biology**, v. 3, p. 1–34, 2003.
- MOMENIMOVAHED, Z.; SALEHINIYA, H. Epidemiological characteristics of and risk factors for breast cancer in the world. **Breast Cancer: Targets and Therapy**, v. 11, p. 151–164, 2019.
- NGONDO, R. P.; CARBON, P. Transcription factor abundance controlled by an auto-regulatory mechanism involving a transcription start site switch. **Nucleic Acids Research**, v. 42, n. 4, p. 2171–2184, 2014.

- N.HORTOBAGYI, G. *et al.* The global breast cancer burden: variations in epidemiology and survival. **Clin Breast Cancer**, v. 6, n. 5, p. 391–401, 2005.
- NITTA, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. **eLife**, v. 2015, n. 4, p. 1–20, 2015.
- NOCEDAL, I.; MANCERA, E.; JOHNSON, A. D. Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. **eLife**, v. 6, p. 1–20, 2017.
- PEROU, C. M.; BORRESEN-DALE, A. L. Systems biology and genomics of breast cancer. **Cold Spring Harbor Perspectives in Biology**, v. 3, n. 2, p. 1–17, 2011. ISSN 19430264.
- PETRI, C. M. K. B. J. Regulation of breast cancer metastasis signaling by mirnas. **Cancer Metastasis Rev.**, v. 39, n. 3, p. 837–886, 2020.
- RAKHA, E. A.; ELLIS, I. O. Modern classification of breast cancer: Should we stick with morphology or convert to molecular profile characteristics. **Advances in Anatomic Pathology**, v. 18, n. 4, p. 255–267, 2011. ISSN 10724109.
- ROGERS, J. M.; BULYK, M. L. Diversification of transcription factor–DNA interactions and the evolution of gene regulatory networks. **Wiley Interdisciplinary Reviews: Systems Biology and Medicine**, v. 10, n. 5, p. 1–12, 2018.
- RStudio Team. **RStudio: Integrated Development Environment for R**. Boston, MA: [s.n.], 2015.
- RYAN, C. J. *et al.* High-resolution network biology: Connecting sequence with function. **Nature Reviews Genetics**, v. 14, n. 12, p. 865–879, 2013.
- SALAMAT, F. *et al.* Subtypes of benign breast disease as a risk factor of breast cancer: A systematic review and meta analyses. **Iranian Journal of Medical Sciences**, v. 43, n. 4, p. 355–364, 2018.
- SALZBERG, S. L. Open questions: How many genes do we have? **BMC Biology**, v. 16, n. 1, p. 10–12, 2018.
- SCHMEIER, S. *et al.* TcoF-DB v2: Update of the database of human and mouse transcription co-factors and transcription factor interactions. **Nucleic Acids Research**, v. 45, n. D1, p. D145–D150, 2016.
- SCHMITT, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. **Cell Reports**, v. 17, n. 8, p. 2042–2059, 2016.
- SHI, W.; FORNES, O.; WASSERMAN, W. W. Gene expression models based on transcription factor binding events confer insight into functional cis-regulatory variants. **Bioinformatics**, v. 35, n. 15, p. 2610–2617, 2019.
- SIDDIQUI, S. *et al.* Cell cannibalism in oral cancer: A sign of aggressiveness, de-evolution, and retroversion of multicellularity. **J Cancer Res Ther.**, v. 15, n. 3, p. 631–637, 2020.
- SONAWANE, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. **Cell Reports**, v. 21, n. 4, p. 1077–1088, 2017.

- SONNHAMMER, E.; KOONIN, E. V. Orthology, paralogy and proposed classification for paralog subtypes. **Trends Genet**, v. 18, n. 12, p. 619–620, 2002.
- SPITZ, F.; FURLONG, E. E. Transcription factors: From enhancer binding to developmental control. **Nature Reviews Genetics**, v. 13, n. 9, p. 613–626, 2012.
- STAMPFEL, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. **Nature**, v. 528, n. 7580, p. 147–151, 2015.
- STEEG, P. S. Targeting metastasis. **Nat Rev Cancer**, v. 16, n. 4, p. 201–18, 2016.
- SZKLARCZYK, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. **Nucleic Acids Research**, v. 47, n. D1, p. D607–D613, 2019.
- THOMAS, D. B. Do hormones cause breast cancer? **Cancer**, v. 53, n. 3 S, p. 595–604, 1984.
- THORNTON, J. W. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 10, p. 5671–5676, 2001.
- TORRE, L. A. *et al.* Global Cancer Statistics, 2012. **CA: a cancer journal of clinicians**, v. 65, n. 2, p. 87–108, 2015.
- TOUCHON, M.; ROCHA, E. P. Coevolution of the organization and structure of prokaryotic genomes. **Cold Spring Harbor Perspectives in Biology**, v. 8, n. 1, p. 1–18, 2016.
- TRACHANA, K. *et al.* Orthology prediction methods: A quality assessment using curated protein families. **BioEssays**, v. 33, n. 10, p. 769–780, 2011.
- TREFFLICH, S.; DALMOLIN, R. J.; CASTRO, M. A. A. **geneplast.data.string.v91: evolutionary rooting and plasticity inference.. R package version 1.0.0**. 2018.
- TREFFLICH, S. *et al.* Which came first, the transcriptional regulator or its target genes? An evolutionary perspective into the construction of eukaryotic regulons. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1863, n. 6, 2019.
- TRIGOS, A. S. *et al.* Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 114, n. 24, p. 6406–6411, 2017.
- TSANKOV, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. **Nature**, v. 518, n. 7539, p. 344–349, 2015.
- VARELA, I.; MENENDEZ, P.; SANJUAN-PLA, A. Oncotarget 66742 www.impactjournals.com/oncotarget Intratumoral heterogeneity and clonal evolution in blood malignancies and solid tumors. **Oncotarget**, v. 8, n. 39, p. 66742–66746, 2017.
- WAKS, A. G.; WINER, E. P. Breast Cancer Treatment: A Review. **JAMA**, v. 321, n. 3, p. 288–300, 2019.
- WANG, K. *et al.* Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. **Nature Biotechnology**, v. 27, n. 9, p. 829–837, 2009.

- WANG, M. *et al.* Role of tumor microenvironment in tumorigenesis. **Journal of Cancer**, v. 8, n. 5, p. 761–773, 2017.
- WEINSTEIN, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. **Nature Genetics**, v. 10, p. 1113–20, 2013.
- WESTERHOFF, H. V.; PALSSON, B. O. The evolution of molecular biology into systems biology. **Nature Biotechnology**, v. 22, p. 1249–1252, 2004.
- WILSON, S.; FILIPP, F. V. A network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer. **npj Systems Biology and Applications**, v. 4, n. 1, p. 1–10, 2018.
- YERSAL, O. Biological subtypes of breast cancer: Prognostic and therapeutic implications. **World Journal of Clinical Oncology**, v. 5, n. 3, p. 412, 2014. ISSN 2218-4333. Disponível em: <<http://www.wjgnet.com/2218-4333/full/v5/i3/412.htm>>.
- YUAN, S.; NORGDARD, R. J.; STANGER, B. Z. Cellular Plasticity in Cancer. **Cancer Discov.**, v. 9, n. 7, p. 837–851, 2020.
- YUSUF, D. *et al.* Genome Biology. **Genome Biology**, v. 13, n. 3, p. R24, 2012.
- ZENDEHDEL, M. *et al.* Subtypes of benign breast disease as a risk factor for breast cancer: A systematic review and meta-analysis protocol. **Med Sci.**, v. 43, n. 1, p. 1–8, 2018.
- ZHU, F. *et al.* The interaction landscape between transcription factors and the nucleosome. **Nature**, v. 562, n. 7725, p. 76–81, 2018.
- ZUCCA-MATTHES, G.; URBAN, C.; VALLEJO, A. Anatomy of the nipple and breast ducts revisited. **Cancer**, v. 101, n. 9, p. 1947–1957, 2004.

ANEXO A – LISTA DE PRODUÇÕES

- 1.(Artigo Publicado) - Trefflich *et al.*, 2019 (**Anexo B**)
- 2.(Vinheta do pacote Geneplast) - Dalmolin *et al.*, 2018 (**Anexo C**)
- 3.(Artigo Publicado) - Corces *et al.*, 2018 (**Anexo D**)
- 4.(Artigo Publicado) - Chagas *et al.*, 2019 (**Anexo E**)
- 5.(Artigo Publicado) - Mathias *et al.*, 2021 (**Anexo F**)
- 6.(Artigo Submetido) - Cardoso *et al.* 2021 (**Anexo G**)
- 7.(Capítulo de Livro Publicado) - Cruz *et al.*, 2017 (**Anexo H**)

ANEXO B – ARTIGO PUBLICADO - (Trefflich *et al.*, 2019)

O artigo intitulado: “Which came first, the transcriptional regulator or its targets genes? An evolutionary perspective into the construction of eukaryotic regulons” (Trefflich *et al.*, 2019), publicado na revista BBA – Gene Regulatory Mechanisms, constitui a principal publicação desta tese. O artigo foi aceito em 30 de novembro de 2019 e publicado em 09 de dezembro de 2019.

Neste artigo apresentamos uma nova abordagem que pode auxiliar na formulação de hipóteses testáveis sobre a construção de redes regulatórias transcricionais, explorando a presença e ausência de elementos regulatórios em linhagens ancestrais.



Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Which came first, the transcriptional regulator or its target genes? An evolutionary perspective into the construction of eukaryotic regulons[☆]



Sheyla Trefflich^{a,b}, Rodrigo J.S. Dalmolin^c, José Miguel Ortega^a, Mauro A.A. Castro^{b,*}

^a Graduate Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil

^b Bioinformatics and Systems Biology Laboratory, Federal University of Paraná, Curitiba 81520-260, Brazil

^c Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal 59078-400, Brazil

ARTICLE INFO

Keywords:

Regulons

Regulatory network

Regulatory element

Ancestral character

ABSTRACT

Eukaryotic regulons are regulatory units formed by a set of genes under the control of the same transcription factor (TF). Despite the functional plasticity, TFs are highly conserved and recognize the same DNA sequences in different organisms. One of the main factors that confer regulatory specificity is the distribution of the binding sites of the TFs along the genome, allowing the configuration of different transcriptional regulatory networks (TRNs) from the same regulator. A similar scenario occurs between tissues of the same organism, where a TRN can be rewired by epigenetic factors, modulating the accessibility of the TF to its binding sites. In this article we discuss concepts that can help to formulate testable hypotheses about the construction of regulons, exploring the presence and absence of the elements that form a TRN throughout the evolution of an ancestral lineage.

This article is part of a Special Issue entitled: Transcriptional Profiles and Regulatory Gene Networks edited by Dr. Federico Manuel Giorgi and Dr. Shaun Mahony.

1. Introduction

There are about 20,000 protein-coding genes annotated in the human genome [1], and although all are potentially informative when we seek to understand some aspect of cell function, it is still a challenge to study the combined effect of genes due to the large number of hypotheses to be tested. Strategies that reconstruct transcriptional regulatory networks (TRNs) can reduce the search space by focusing on regulatory bottlenecks [2], which provides opportunities to study regulation in the context of network nodes that are central control points of cell function, also known as regulatory hubs [3]. TRNs are centered on transcription factors (TFs), regulators that recognize specific DNA sequences and guide the expression of the genome.

TRNs can be reconstructed from gene expression data generated for all genes in a set of samples [4]. The reconstruction of these networks takes into account prior knowledge about the molecular mechanisms that regulate the gene expression and the organization of genes in the genome. For example, in prokaryotic organisms, genes related to the same biological effect are usually co-localized in the genome, forming polycistronic units or operons. When starting the transcription process, all genes of an operon are expressed together [5], with multiple

coordinated operons forming a regulon [6]. In eukaryotes, the transcription process is usually independent for each gene and the term *regulon* refers to the set of genes controlled by the same regulator [7], which binds to specific cis-regulatory regions, resulting either in activation or repression of target gene expression [8] (Fig. 1A). There are about 1600 TFs annotated in the human genome [9] and for each of them it is possible to construct a regulon.

The structure of an eukaryotic regulatory network is illustrated in Fig. 1B, outlined with three regulons. The inference of regulons can be estimated by algorithms that evaluate the expression of a regulator and its potential targets in a set of samples, with the ARACNe algorithm being one of the most used methods [10–12]. The regulatory interactions between the elements that form a regulon depend on both the TF binding to specific DNA sequences and epigenetic factors, such as the chromatin accessibility [13]. TF bindings are context-dependent and are capable of remodeling the chromatin, dynamically guiding cell differentiation in mammals [14]. The interactions between a TF and its targets are tissue-specific, with greater tissue specificity than the expression of the regulator [15].

Despite the functional plasticity, TFs are highly conserved and recognize the same DNA sequences in different species [16,17].

[☆] This article is part of a Special Issue entitled: Transcriptional Profiles and Regulatory Gene Networks edited by Dr. Federico Manuel Giorgi and Dr. Shaun Mahony.

* Corresponding author.

E-mail address: mauro.castro@ufpr.br (M.A.A. Castro).

<https://doi.org/10.1016/j.bbagrm.2019.194472>

Received 31 May 2019; Received in revised form 6 November 2019; Accepted 30 November 2019

Available online 09 December 2019

1874-9399/ © 2019 Elsevier B.V. All rights reserved.

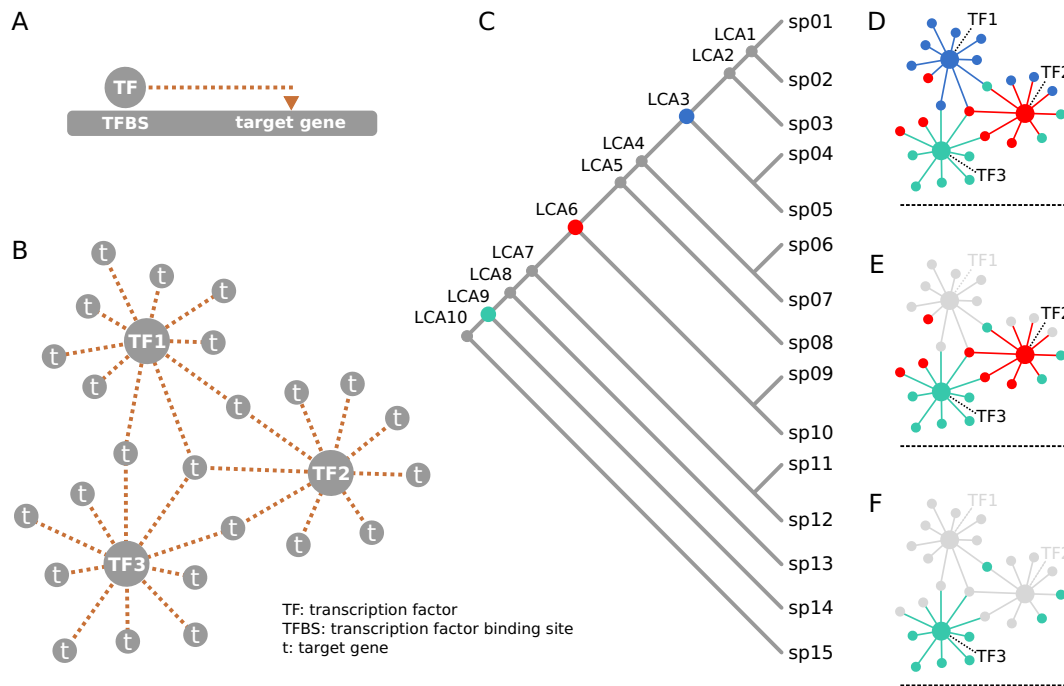


Fig. 1. A hypothetical scenario of vertical inheritance for genes annotated in three regulons. (A) Diagram illustrating a transcription factor (TF)–target gene interaction. (B) Diagram illustrating a TRN containing three regulons. (C) A species tree whose leaves represent 15 extant species (sp1, sp2, ..., sp15) and the internal nodes the Last Common Ancestor (LCA) of a monophyletic group. Colored circles represent rooting points mapped to the orthologous genes annotated in sp1. (D) Orthologous genes rooted at LCA3 (blue) or earlier. (E) Orthologous genes rooted at LCA6 (red) or earlier. (F) Orthologous genes rooted at LCA9 (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Regulatory differences between species seem to be related to the distribution of TF-binding sites along the genome, especially distal to the promoter regions of orthologous genes [18,19]. Gain and loss of cis-regulatory sequences are important factors in the reorganization of the structure of transcriptional networks, contributing to the regulatory diversity observed between species [20]. In addition, changes in the distribution of conserved cis-regulatory sequences may explain the incorporation of new functions into evolving regulatory networks [21]. Since regulators and targets may undergo different selective pressures [22], it is particularly interesting to consider when each element forming a regulon arose throughout the evolutionary process. In this paper we propose a general framework to test the following question: Which came first, the transcriptional regulator or its target genes?

2. Mapping regulons in a species tree

It is reasonable to assume that the set of genes belonging to a regulon is to some extent associated with the same evolutionary history, co-occurring in regulatory units shaped in the course of evolution. To test this hypothesis we need to reconstruct evolutionary scenarios for each of the genes annotated in a regulon [23]. For example, Figs. 1C–F illustrate hypothetical evolutionary scenarios for the genes annotated in the three regulons outlined in Fig. 1B.

Consider that in these scenarios the genes are inherited by vertical transfer, from the ancestors to their descendants. Fig. 1C represents a given species tree whose leaves indicate 15 extant species (sp1, sp2, ..., sp15) and the internal nodes the Last Common Ancestor (LCA) of a monophyletic group (*i.e.* all descendants of an ancestral lineage). In addition, all genes in the regulons belong to the sp1 species (*i.e.* the regulons are inferred for sp1); therefore, the tree is organized with sp1 at the top in order to simplify the observation of its LCAs (LCA1, LCA2, ..., LCA10). The colored circles represent rooting points mapped to the sp1 genes. For example, eleven genes (blue) are rooted at LCA3, including TF1 (Fig. 1D), and all the other genes are rooted at earlier points of the species tree. Seven genes (red) are rooted at LCA6,

including TF2 (Fig. 1E), and nine genes (green) are rooted at LCA9, including TF3 (Fig. 1F). The construction of models that describe the origins of a regulon requires the systematic mapping of genes in a given species tree, which can be inferred by orthology analysis, as discussed next.

3. Mapping orthologous genes in a species tree

Orthology relations between genes from different species can be systematically predicted by comparing genomes. There are several methods to predict orthologous genes, with graph-based methods offering the best trade-off between speed and accuracy to generate large-scale orthology annotation [24]. To exemplify the evolutionary rooting using orthology analysis, consider a hypothetical orthologous group (OG) consisting of 6 orthologs annotated in 3 species (Fig. 2A). In the annotation process it is possible that some orthology relations are not well resolved, here exemplified by the YA1 and YA2 paralogous genes. Fig. 2B–D present diagrams adapted from Koonin (2005) [25] illustrating orthology and paralogy concepts in the evolutionary history of the genes listed in Fig. 2A.

All these genes belong to the same ancestral family, formed by the X and Y ancestral genes (Fig. 2B), inherited by three extant species (spa, spb and spc), but with different evolutionary histories. The Xa, Xb and Xc genes are orthologous to each other, showing a one-to-one orthologous relationship, originated by vertical descent from the X ancestral gene (Fig. 2C). The X gene is present as a single copy in the three species, while the Y ancestral gene underwent a duplication event on the ancestral branch of the spa species (Fig. 2D). The Ya1 and Ya2 genes are paralogous to each other, and co-orthologs of the Yb gene. A deletion event is represented by the absence of the Y ancestral gene in the spc species. More complex evolutionary scenarios may arise when duplication and deletion events occur in internal branches of a phylogenetic tree, but in all cases the genes may be annotated to the same orthologous group.

Based on the analysis of the distribution of orthologs in a given

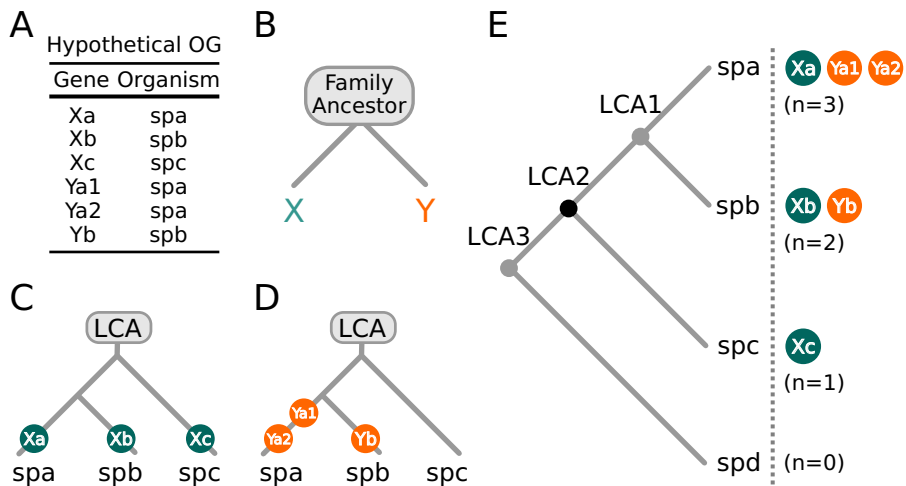


Fig. 2. Schematic representation of the evolutionary root of a hypothetical OG. (A) Orthologous genes annotated in the OG. (B) All orthologous genes belong to the same ancestral family. (C, D) Ancestral genes (X and Y) were inherited by three extant species (spa, spb and spc), but with different evolutionary histories. (E) Distribution of the orthologous genes in the species tree, with the LCAs enumerated from the most recent to the most ancestral node of the spa species. In this scenario, genes Xa, Ya1 and Ya2 (from spa) are rooted in the LCA2.

species tree, it is possible to infer the evolutionary origin of the genes of an extant species. For example, taking the spa species as reference, Fig. 2E lists all LCAs (LCA1, LCA2, and LCA3) and all genes annotated in each species. The spa orthologs are rooted in the LCA2, since all LCA2 descendants have at least one copy of the ancestral sequence. This scenario proposes that the Xa, Ya1 and Ya2 genes evolved from a single ancestral sequence, from the LCA of the compared species, through a series of speciation processes and duplication events. Note that the absence of the Y ancestral gene in the spc species does not change the inference of the evolutionary root of the spa's genes, and the absence of both X and Y ancestral genes in the spd species indicates that the LCA3 is unrelated to the OG's evolutionary history.

The problem of finding the evolutionary root of a gene in a given phylogenetic tree has already been formulated and involves finding the most consistent vertical descent scenario for a set of orthologs, where eventual inconsistencies due to horizontal transfer, duplication and deletion events may be resolved with a function that penalizes deviations from a phyletic pattern [26–29], as discussed next.

4. An evolutionary perspective into the construction of eukaryotic regulons

The interaction of a TF with a target gene can be summarized by a regulatory model with three elements [8]: i) the transcription factor; ii) the cis-regulatory region containing the TF-binding site; and iii) the target gene (Fig. 1A). We may consider more complex regulatory models, for example, when TF's activity is modulated by one or more co-factors [30], when TF binding sites are present in different positions of the regulatory region [31], when epigenetic factors influence the accessibility of the binding sites [32], or when varying degrees of binding affinity are required for cooperative interactions [33] and enhancer specificity [34]. In all cases the absence of either the transcription factor or the target gene in a genome implies no interaction. This proposition might be also valid for auto-regulation, in which the TF binds to its own promoter and either activates or represses transcription [35] (i.e. as regulator and target are the same entity in the auto-regulation, the absence of the TF gene implies no interaction as well). For example, gaining new cis-regulatory sequences may lead to the reorganization of the transcriptional network structure [20], but it can not create interactions for targets that are not annotated in the genome.

Steroid hormones and nuclear receptors illustrate this problem. In the absence of a ligand, what function serves a new receptor? And without a receptor, which selective pressures guide the evolution of a new ligand? [36]. Although these questions may offer only a basic framework for exploring evolutionary processes, they can also bring

insights into regulatory systems observed in extant species, such as proposing evolutionary scenarios about the construction of a regulatory unit: 1) Did the TF and its targets appear together? 2) Did the TF appear before the targets it regulates? 3) Did TF appear after its targets?

Fig. 3 indicates that we can find regulons for each of these scenarios, but the most prevalent scenario is the one in which the TF arises previously to its targets. This figure shows the distribution of the evolutionary roots of regulons constructed by Fletcher et al. [4] in human breast cancer samples. Each regulon consists of a TF and its targets, here summarized in a TRN containing 307 TFs and 6308 targets. The geneplast R package [37] was used to infer the evolutionary roots by orthology analysis; for each element of the TRN an evolutionary root was inferred in a species tree containing 121 eukaryotes (Fig. 3A; a detailed tree, with the name of each species, is available in the documentation of the geneplast package). Fig. 3B shows the overall distribution of the inferred evolutionary roots for TFs and targets and indicates that the evolutionary roots of the TFs are prior to the evolutionary roots of the targets (p-value = $1e-6$, Wilcoxon-Mann-Whitney test). Fig. 3B also shows the distribution of evolutionary roots inferred for another class of regulators, the transcription co-factors (TcoFs), which influence the transcriptional regulation by forming protein complexes with TFs [38]. Using the same approach, we did not detect a significant shift in the distribution of evolutionary roots between TFs and TcoFs (p-value = 0.884, Wilcoxon-Mann-Whitney test).

Fig. 3C extends the approach to individual regulons, grouping regulons based on the distance between the evolutionary roots of TFs and targets: Group-a) TF-target genes rooted before the TF (30.6% of the regulons), Group-b) TF-target genes rooted with the TF (24.7% of the regulons), and Group-c) TF-target genes rooted after the TF (44.3% of the regulons). A detailed distribution of the evolutionary roots inferred for individual regulons is provided in Fig. 3D, and an R script that describes all steps to reproduce these observations is available in the Bioconductor repository (Supporting material). In the Supporting material we also reproduce the main observations in a different TRN, using regulons generated from normal breast tissue samples, and expand the proposed analysis framework to explore different metrics derived from OG information.

5. Concluding remarks

In this paper we illustrate a framework to formulate testable hypotheses about the reconfiguration of regulons, exploring the presence and absence of the elements that form a TRN along the evolution of an ancestral lineage. Since interactions between TFs and targets confer tissue specificity to a TRN, we anticipate that this framework will contribute to studies exploring the reorganization of regulons between

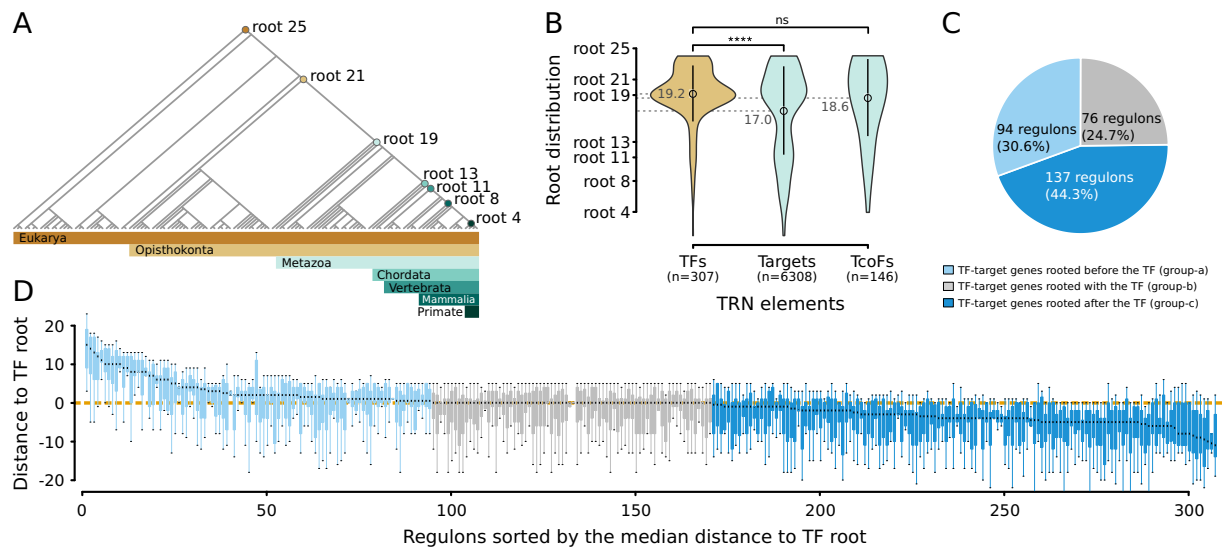


Fig. 3. Evolutionary roots of TFs and target genes inferred for regulons from a human TRN. (A) A species tree with the LCAs enumerated from the most recent to the most ancestral node of *Homo sapiens*. (B) Distribution of the evolutionary roots inferred for TFs and target genes. Transcription co-factors (TcoFs) are also included in the analysis. (C) Regulons split into three groups based on the median distance between the evolutionary roots inferred for a TF and its target genes. (D) Same as in C, but showing individual regulons. ****p-value = 1e-6 (Wilcoxon-Mann-Whitney test); ns = not significant. To reproduce these plots, please see the supporting R script available at the Bioconductor repository ([Supporting material](#)).

different conditions. The distance between the evolutionary roots of a TF and its targets can be used as additional information to study the plasticity of regulatory networks.

Transparency document

The [Transparency document](#) associated with this article can be found, in online version.

Supporting material

R script to reproduce observations from [Fig. 3](#) (<http://bioconductor.org/packages/geneplast/>).

Acknowledgment

This work was supported by the National Council for Scientific and Technological Development (CNPq), grant no. 407090/2016-9, and the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil.

Declaration of competing interest

The authors declare that there are no competing interests associated with this manuscript.

References

- [1] S.L. Salzberg, Open questions: how many genes do we have? *BMC Biol.* 16 (1) (2018) 10–12, <https://doi.org/10.1186/s12915-018-0564-x>.
- [2] C. Lefebvre, G. Rieckhof, A. Califano, Reverse-engineering human regulatory networks, *Wiley Interdiscip. Rev. Syst. Biol. Med.* 4 (4) (2012) 311–325, <https://doi.org/10.1002/wsbm.1159>.
- [3] K. Wang, M. Saito, B.C. Bisikirska, M.J. Alvarez, W.K. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A.A. Margolin, U. Klein, R. Dalla-Favera, A. Califano, Genome-wide identification of post-translational modulators of transcription factor activity in human B cells, *Nat. Biotechnol.* 27 (9) (2009) 829–837, <https://doi.org/10.1038/nbt.1563>.
- [4] M.N. Fletcher, M.A. Castro, X. Wang, I. De Santiago, M. O'Reilly, F. Chin, O.M. Rueda, C. Caldas, B.A. Ponder, F. Markowitz, K.B. Meyer, Master regulators of FGFR2 signalling and breast cancer risk, *Nat. Commun.* 4 (2013) 2464, <https://doi.org/10.1038/ncomms3464>.
- [5] F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins, *J. Mol. Biol.* 3 (1961) 318–356, [https://doi.org/10.1016/S0022-2836\(61\)80072-7](https://doi.org/10.1016/S0022-2836(61)80072-7).
- [6] B. Liu, C. Zhou, G. Li, H. Zhang, E. Zeng, Q. Liu, Q. Ma, Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses, *Sci. Rep.* 6 (2016) 23030, <https://doi.org/10.1038/srep23030>.
- [7] M. Touchon, E.P. Rocha, Coevolution of the organization and structure of prokaryotic genomes, *Cold Spring Harb. Perspect. Biol.* 8 (1) (2016) a018168, <https://doi.org/10.1101/cshperspect.a018168>.
- [8] W. Shi, O. Fornes, W.W. Wasserman, Gene expression models based on transcription factor binding events confer insight into functional cis-regulatory variants, *Bioinformatics* (2019), <https://doi.org/10.1093/bioinformatics/bty992> (Epub ahead of print).
- [9] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T.R. Hughes, M.T. Weirauch, The human transcription factors, *Cell* 172 (4) (2018) 650–665, <https://doi.org/10.1016/j.cell.2018.01.029>.
- [10] A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, A. Califano, Reverse engineering cellular networks, *Nat. Protoc.* 1 (2) (2006) 662–671, <https://doi.org/10.1038/nprot.2006.106>.
- [11] R. Janky, A. Verfaillie, H. Imrichová, B. Van de Sande, L. Standaert, V. Christiaens, G. Hulselmans, K. Herten, M. Naval Sanchez, D. Potier, D. Svetlichnyy, Z. Kalender Atak, M. Fiers, J.C. Marine, S. Aerts, iRegulon: from a gene list to a gene regulatory network using large motif and track collections, *PLoS Comput. Biol.* 10 (7) (2014) e1003731, <https://doi.org/10.1371/journal.pcbi.1003731>.
- [12] S. Aibar, C.B. González-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z.K. Atak, J. Wouters, S. Aerts, SCENIC: single-cell regulatory network inference and clustering, *Nat. Methods* 14 (11) (2017) 1083–1086, <https://doi.org/10.1038/nmeth.4463>.
- [13] M.R. Corces, J.M. Granja, S. Shams, B.H. Louie, J.A. Seoane, W. Zhou, T.C. Silva, C. Groeneveld, C.K. Wong, S.W. Cho, A.T. Satpathy, M.R. Mumbach, K.A. Hoadley, A.G. Robertson, N.C. Sheffield, I. Felau, M.A.A. Castro, B.P. Berman, L.M. Staudt, J.C. Zenklusen, P.W. Laird, C. Curtis, W.J. Greenleaf, H.Y. Chang, The chromatin accessibility landscape of primary human cancers, *Science* 362 (6413) (2018) eaav1898, <https://doi.org/10.1126/science.aav1898>.
- [14] A.M. Tsankov, H. Gu, V. Akopian, M.J. Ziller, J. Donaghey, I. Amit, A. Gnirke, A. Meissner, Transcription factor binding dynamics during human ES cell differentiation, *Nature* 518 (7539) (2015) 344–349, <https://doi.org/10.1038/nature14233>.
- [15] A.R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J.N. Paulson, C.M. Lopes-Ramos, D.L. DeMeo, J. Quackenbush, K. Glass, M.L. Kuijjer, Understanding tissue-specific gene regulation, *Cell Rep.* 21 (4) (2017) 1077–1088, <https://doi.org/10.1016/j.celrep.2017.10.001>.
- [16] K.R. Nitta, A. Jolma, Y. Yin, E. Morgunova, T. Kivioja, J. Akhtar, K. Hens, J. Toivonen, B. Deplancke, E.E.M. Furlong, J. Taipale, Conservation of transcription factor binding specificities across 600 million years of bilateria evolution, *eLife* 4 (2015) e04837, <https://doi.org/10.7554/eLife.04837.001>.
- [17] L. Krefl, A. Soete, P. Hulpiau, A. Botzki, Y. Saeys, P.D. Bleser, ConTra v3: a tool to identify transcription factor binding sites across species, *Nucleic Acids Res.* 45 (W1) (2017) W490–W494, <https://doi.org/10.1093/nar/gkx376>.
- [18] Y. Cheng, Z. Ma, B.H. Kim, W. Wu, P. Cayting, A.P. Boyle, V. Sundaram, X. Xing, N. Dogan, J. Li, G. Euskirchen, S. Lin, Y. Lin, A. Visel, T. Kawli, X. Yang, D. Patacisi, C.A. Keller, B. Giardine, A. Kundaje, T. Wang, L.A. Pennacchio, Z. Weng, R.C. Hardison, M.P. Snyder, Principles of regulatory information conservation

- between mouse and human, *Nature* 515 (7527) (2014) 371–375, <https://doi.org/10.1038/nature13985>.
- [19] C.D. Arnold, D. Gerlach, D. Spies, J.A. Matts, Y.A. Sytnikova, M. Pagani, N.C. Lau, A. Stark, Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during cis-regulatory evolution, *Nat. Genet.* 46 (7) (2014) 685–692, <https://doi.org/10.1038/ng.3009>.
- [20] C.R. Baker, L.N. Booth, T.R. Sorrells, A.D. Johnson, Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification, *Cell* 151 (1) (2012) 80–95, <https://doi.org/10.1016/j.cell.2012.08.018>.
- [21] I. Nocedal, E. Mancera, A.D. Johnson, Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator, *Elife* 6 (2017) e23250, <https://doi.org/10.7554/eLife.23250>.
- [22] J.M. Rogers, M.L. Bulyk, Diversification of transcription factor–DNA interactions and the evolution of gene regulatory networks, *Wiley Interdiscip. Rev. Syst. Biol. Med.* 10 (2018) e1423, <https://doi.org/10.1002/wsbm.1423>.
- [23] M.A. Castro, R.J. Dalmolin, J.C. Moreira, J.C. Mombach, R.M. de Almeida, Evolutionary origins of human apoptosis and genome-stability gene networks, *Nucleic Acids Res.* 36 (19) (2008) 6269–6283, <https://doi.org/10.1093/nar/gkn636>.
- [24] K. Trachana, T.A. Larsson, S. Powell, W.H. Chen, T. Doerks, J. Muller, P. Bork, Orthology prediction methods: a quality assessment using curated protein families, *BioEssays* 33 (10) (2011) 769–780, <https://doi.org/10.1002/bies.201100062>.
- [25] E.V. Koonin, Orthologs, Paralogs, and evolutionary genomics, *Annu. Rev. Genet.* 39 (2005) 309–338, <https://doi.org/10.1146/annurev.genet.39.073003.114725>.
- [26] B.G. Mirkin, T.I. Fenner, M.Y. Galperin, E.V. Koonin, Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evol. Biol.* 3 (2003) 2, <https://doi.org/10.1186/1471-2148-3-2>.
- [27] E. Jacox, C. Chauve, G.J. Szollosi, Y. Ponty, C. Scornavacca, ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony, *Bioinformatics* 32 (13) (2016) 2056–2058, <https://doi.org/10.1093/bioinformatics/btw105>.
- [28] R. Dondi, M. Lafond, C. Scornavacca, Reconciling multiple genes trees via segmental duplications and losses, *Algorithms for Molecular Biology* 14 (2019) 7, <https://doi.org/10.1186/s13015-019-0139-6>.
- [29] Y.B. Chan, C. Robin, Reconciliation of a gene network and species tree, *J. Theor. Biol.* 472 (2019) 54–66, <https://doi.org/10.1016/j.jtbi.2019.04.001>.
- [30] G. Stampfel, T. Kazmar, O. Frank, S. Wienerroither, F. Reiter, A. Stark, Transcriptional regulators form diverse groups with context-dependent regulatory functions, *Nature* 528 (7580) (2015) 147–151, <https://doi.org/10.1038/nature15545>.
- [31] S.R. Grossman, J. Engreitz, J.P. Ray, T.H. Nguyen, N. Hacohen, E.S. Lander, Positional specificity of different transcription factor classes within enhancers, *Proceedings of the National Academy of Sciences of USA* 115 (30) (2018) E7222–E7230, <https://doi.org/10.1073/pnas.1804663115>.
- [32] S. Wilson, F.V. Filipp, A network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer, *NPJ Systems Biology and Applications* 4 (24) (2018), <https://doi.org/10.1038/s41540-018-0061-4>.
- [33] A. Jolma, Y. Yin, K.R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, J. Taipale, DNA-dependent formation of transcription factor pairs alters their binding specificity, *Nature* 527 (7578) (2015) 384–388, <https://doi.org/10.1038/nature15518>.
- [34] A. Zandvakili, I. Campbell, L.M. Gutzwiller, M.T. Weirauch, B. Gebelein, Degenerate Pax2 and senseless binding motifs improve detection of low-affinity sites required for enhancer specificity, *PLoS Genet.* 14 (4) (2018) e1007289, <https://doi.org/10.1371/journal.pgen.1007289>.
- [35] R.P. Ngondo, P. Carbon, Transcription factor abundance controlled by an autoregulatory mechanism involving a transcription start site switch, *Nucleic Acids Res.* 42 (4) (2014) 2171–2184, <https://doi.org/10.1093/nar/gkt1136>.
- [36] J.W. Thornton, Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions, *Proceedings of the National Academy of Sciences of USA* 98 (10) (2001) 5671–5676, <https://doi.org/10.1073/pnas.091553298>.
- [37] R.J. Dalmolin, M.A.A. Castro, Geneplast: Evolutionary Rooting and Plasticity Inference, R Package Version 1.10.3, (2015), <https://doi.org/10.18129/B9.bioc.geneplast>.
- [38] S. Schmeier, T. Alam, M. Essack, V.B. Bajic, TcoF-DB v2: update of the database of human and mouse transcription co-factors and transcription factor interactions, *Nucleic Acids Res.* 45 (D1) (2016) D145–D150, <https://doi.org/10.1093/nar/gkw1007>.

ANEXO C – VINHETA DO PACOTE GENEPLAST

A vinheta do pacote geneplast (doi:10.18129/B9.bioc.geneplast) disponibiliza o código fonte para reprodução dos resultados apresentados nesta tese. Esta vinheta apresenta exemplos e documentação necessária para instalação, uso do pacote e estudos de caso. Um destes estudos de caso demonstra o enraizamento evolutivo de redes regulatórias. A vinheta do pacote geneplast pode ser acessada em <https://www.bioconductor.org/packages/geneplast/>

Geneplast: evolutionary rooting and plasticity analysis of orthologous groups.

Rodrigo JS Dalmolin, Sheyla Trefflich, Diego AA Morais, Mauro AA Castro.

18 August 2018

Abstract

Geneplast is designed for evolutionary and plasticity analysis based on orthologous groups distribution in a given species tree. It uses Shannon information theory and orthologs abundance to estimate the Evolutionary Plasticity Index. Additionally, it implements the Bridge algorithm to determine the evolutionary root of a given gene based on its orthologs distribution

Overview

Geneplast is designed for evolutionary and plasticity analysis based on the distribuion of orthologous groups in a given species tree. It uses Shannon information theory to estimate the Evolutionary Plasticity Index (EPI) (Dalmolin et al. (2011), Castro et al. (2008)).

Figure 1 shows a toy example to illustrate the analysis. The observed itens in **Figure 1a** are distributed evenly among the different species (i.e. high diversity), while **Figure 1b** shows the opposite case. The diversity is given by the normalized Shannon's diversity and represents the distribution of orthologous and paralogous genes in a set of species. High diversity represents an homogeneous distribution among the evaluated species, while low diversity indicates that few species concentrate most of the observed orthologous genes.

The *EPI* characterizes the evolutionary history of a given orthologous group (OG). It accesses the distribution of orthologs and paralogs and is defined as,

$$EPI = 1 - \frac{H\alpha}{\sqrt{D\alpha}}, (1)$$

where *Dalpha* represents the OG abundance and *Halpha* the OG diversity. Low values of *Dalpha* combined with high values for *Halpha* indicates an orthologous group of low plasticity, that is, few OG members distributed over many species. It also indicates that the OG might have experienced few modifications (i.e. duplication and deletion episodes) during the evolution. Note that $0 \leq Halpha \leq 1$ and $Dalpha \geq 1$. As a result, $0 \leq EPI \leq 1$. For further information about the *EPI*, please see (Dalmolin et al. 2011).

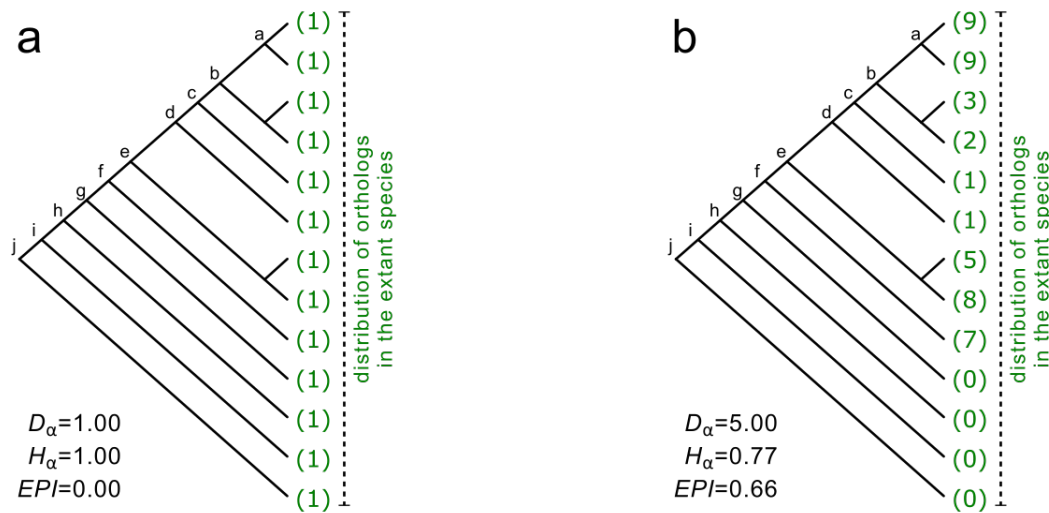


Figure 1. **Toy examples showing the distribution of orthologous and paralogous genes in a given species tree.** (a) OG of low abundance *Dalpha*, high diversity (*Halpha*) and consequently low plasticity (*PI*). In this hypothetical case, the OG comprises orthologous genes observed in all species, without apparent deletion or duplication episodes. (b) in this example the OG is observed in many species, but not all, with many paralogs in some of them. Green numbers represents the number of orthologous genes in each species.

geneplast also implements a new algorithm called *Bridge* in order to interrogate the evolutionary root of a given gene based on the distribution of its orthologs. The *Bridge* algorithm assesses the probability that an ortholog of a given gene is present in each last common ancestor (LCA) of a given species (in a given species tree). As a result, this approach infers the evolutionary root representing the gene emergence. The method is designed to deal with large scale queries in order to interrogate, for example, all genes annotated in a network (please refer to (Castro et al. 2008) for a case study illustrating the advantages of using this approach).

To illustrate the rooting inference consider the evolutionary scenarios presented in **Figure 2** for the same hypothetical OGs. These OGs comprise a number of orthologous genes distributed among 13 species, and the pattern of presence or absence is indicated by green and grey colours, respectively. Observe that at least one ortholog is present in all extant species in **Figure 2a**. To explain this common genetic trait, one possible evolutionary scenario could assume that the ortholog was present in the LCA of all species and was genetically transmitted up to the descendants. For this case, the evolutionary root might be placed at the bottom of the species tree (i.e. node *g*). The same reasoning can be done in **Figure 2b**, but with the evolutionary root placed at node *d*. The **geneplast** rooting pipeline is designed to infer the most consistent rooting scenario for the observed orthologs in a given species tree. The pipeline provides a consistency score called *Dscore* which estimates the stability of the inferred root, as well as an associated empirical *p-value* computed by permutation analysis.

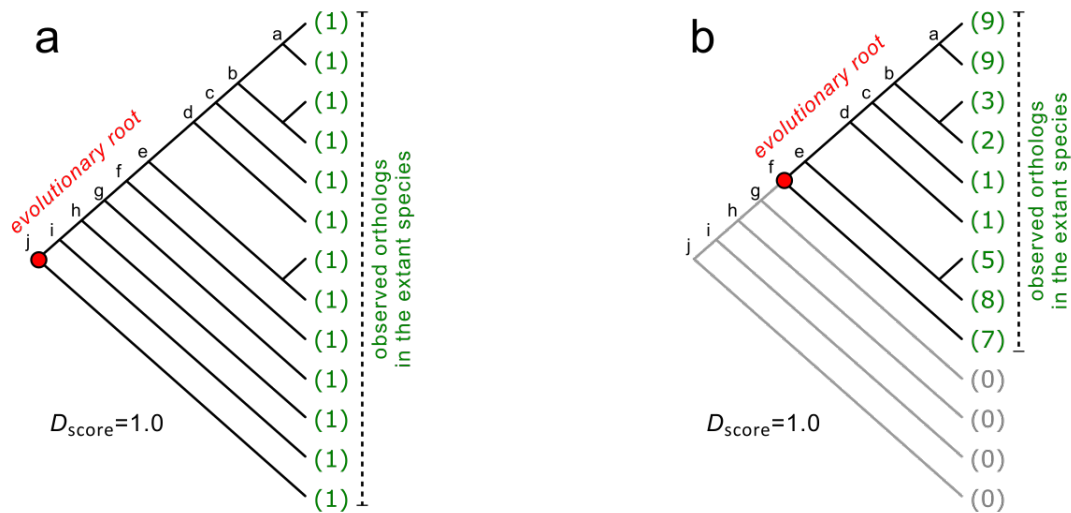


Figure 2. Possible evolutionary rooting scenarios for the same toy examples depicted in Figure 1. (a, b) Red circles indicate the evolutionary roots that best explain the observed orthologs in this species tree.

Quick start

The orthology data required to run **geneplast** is available in the `gpdata.gs` dataset. This dataset includes four objects containing information about Clusters of Orthologous Groups derived from the **STRING database**, release 9.1. **geneplast** can also be used with other sources of orthology information, provided that the input is set according to the `gpdata.gs` data structure (*note: in order to reduce the processing time this example uses a subset of the STRING database*).

```
library(geneplast)
data(gpdata.gs)
```

Evolutionary plasticity inference

The first step is to create an OGP object by running the `gplast.preprocess` function. This example uses 121 eukaryotic species from the *STRING* database and all OGs mapped to the genome stability gene network (Castro et al. 2008). Next, the `gplast` function perform the plasticity analysis and the `gplast.get` returns the results:

1 - Create an object of class OGP.

```
ogp <- gplast.preprocess(cogdata=cogdata, sspids=sspids, cogids=cogids, verbose=FALSE)
```

2 - Run the `gplast` function.

```
ogp <- gplast(ogp, verbose=FALSE)
```

3 - Get results.

```
res <- gplast.get(ogp, what="results")
head(res)
```

```
##          abundance diversity plasticity
## KOG0011    1.7328    0.9532    0.2759
## KOG0028    3.1466    0.9207    0.4809
```

```
## KOG0034    4.1121    0.9216    0.5455
## KOG0037    2.8252    0.9116    0.4577
## KOG0045    7.3534    0.8965    0.6694
## KOG0192   26.9286    0.8284    0.8404
```

The results are returned in a 3-column `data.frame` with OG ids (cogids) identified in `row.names`. Columns are named as *abundance*, *diversity*, and *plasticity*.

The metric *abundance* simply indicates the ratio of orthologs and paralogs by species. For example, KOG0011 comprises 201 genes distributed in 116 eukaryotic species, with a resulting abundance of 1.7328. Abundance of 1 indicates an one-to-one orthology relationship, while high abundance denotes many duplication episodes on the OG's evolutionary history. Diversity is obtained applying normalized Shannon entropy on orthologous distribution and Plasticity is obtained by *EPI* index, as described equation (1).

Evolutionary rooting inference

The rooting analysis starts with an OGR object by running the `groot.preprocess` function. This example uses all *OGs* mapped to the genome stability gene network using *H. sapiens* as reference species (Castro et al. 2008) and is set to perform 100 permutations for demonstration purposes (for a full analysis, please set `Permutations>=1000`). Next, the `groot` function performs the rooting analysis and the results are retrieved by `groot.get`, which returns a `data.frame` listing the root of each OG evaluated by the `groot` method. The pipeline also returns the inconsistency score, which estimates the stability of the rooting analysis, as well as the associated empirical *p-value*. Additionally, the `groot.plot` function allows the visualization of the inferred root for a given OG (e.g. **Figure 3**) and the LCAs for the reference species (**Figure 4**).

1 - Create an object of class OGR.

```
ogr <- groot.preprocess(cogdata=cogdata, phyloTree=phyloTree,
                       spid="9606", cogids=cogids, verbose=FALSE)
```

2 - Run the groot function.

```
set.seed(1)
ogr <- groot(ogr, nPermutations=100, verbose=FALSE)
```

3 - Get results.

```
res <- groot.get(ogr, what="results")
head(res)
```

```
##           Root Dscore   Pvalue AdjPvalue
## NOG251516    3  0.67 2.49e-10  3.54e-08
## NOG80202     4  1.00 1.46e-09  2.07e-07
## NOG72146     6  0.82 2.99e-05  4.24e-03
## NOG44788     6  0.56 1.61e-04  2.28e-02
## NOG39906     7  1.00 8.30e-09  1.18e-06
## NOG45364     9  0.83 1.94e-07  2.76e-05
```

4 - Check the inferred root of a given OG

```
groot.plot(ogr, whichOG="NOG40170")
```

```
## PDF file 'groot_NOG40170_9606LCAs.pdf' has been generated!
```

5 - Visualization of the LCAs for the reference species in the analysis (i.e. *H. sapiens*)

```
groot.plot(ogr, plot.lcas = TRUE)
```

```
## PDF file 'groot_9606LCAs.pdf' has been generated!
```

NOG40170



Figure 3. Inferred evolutionary rooting scenario for NOG40170. Monophyletic groups are ordered to show all branches of the tree below the queried species in the analysis.

REF: Homo sapiens (9606)



Figure 4. Visualization of the LCAs for the reference species in the analysis.

Case studies

High-throughput rooting inference

This example shows how to assess all *OGs* annotated for *H. sapiens*.

1 - Load orthology data from the **geneplast.data.string.v91** package (*currently available under request*).

```
library(geneplast.data.string.v91)
data(gpdata_string_v91)
```

2 - Create an object of class 'OGR' for a reference 'spid'.

```
ogr <- groot.preprocess(cogdata=cogdata, phyloTree=phyloTree, spid="9606")
```

3 - Run the `groot` function and infer the evolutionary roots. *Note: this step should take a long processing time due to the large number of OGs in the input data (also, `nPermutations` argument is set to 100 for demonstration purpose only).*

```
ogr <- groot(ogr, nPermutations=100, verbose=TRUE)
```

Map rooting information on PPI networks

This example aims to show the evolutionary root of a protein-protein interaction (PPI) network, mapping the appearance of each gene in a given species tree. The next steps show how to transfer evolutionary rooting information from `geneplast` to a graph model. *Note: to make this work the gene annotation available from the input PPI network needs to match the annotation available from the `geneplast` data (in this case, `ENTREZ` gene IDs are used to match the datasets).*

1 - Load a PPI network and required packages. The `igraph` object called 'ppi.gs' provides PPI information for apoptosis and genome-stability genes (Castro et al. 2008).

```
library(RedeR)
library(igraph)
library(RColorBrewer)
data(ppi.gs)
```

2 - Map rooting information on the `igraph` object.

```
g <- ogr2igraph(ogr, cogdata, ppi.gs, idkey = "ENTREZ")
```

3 - Adjust colors for rooting information.

```
pal <- brewer.pal(9, "RdYlBu")
color_col <- colorRampPalette(pal)(25) #set a color for each root!
g <- att.setv(g=g, from="Root", to="nodeColor", cols=color_col,
             na.col = "grey80", breaks = seq(1,25))
```

4 - Aesthetic adjusts for some graph attributes.

```
g <- att.setv(g = g, from = "SYMBOL", to = "nodeAlias")
E(g)$edgeColor <- "grey80"
V(g)$nodeLineColor <- "grey80"
```

5 - Send the `igraph` object to `RedeR` interface.

```
rdp <- RedPort()
callD(rdp)
resetD(rdp)
addGraph(rdp, g)
addLegend.color(rdp, colvec=g$legNodeColor$scale, size=15,
               labvec=g$legNodeColor$legend, title="Roots represented in Fig4")
```

6 - Get apoptosis and genome-stability sub-networks.

```
g1 <- induced_subgraph(g=g, V(g)$name[V(g)$Apoptosis==1])
g2 <- induced_subgraph(g=g, V(g)$name[V(g)$GenomeStability==1])
```

7 - Group apoptosis and genome-stability genes into containers.

```

myTheme <- list(nestFontSize=25, zoom=80, isNest=TRUE, gscale=65, theme=2)
addGraph(rdp, g1, gcoord=c(25, 50), theme = c(myTheme, nestAlias="Apoptosis"))
addGraph(rdp, g2, gcoord=c(75, 50), theme = c(myTheme, nestAlias="Genome Stability"))
relax(rdp, p1=50, p2=50, p3=50, p4=50, p5= 50, ps = TRUE)

```

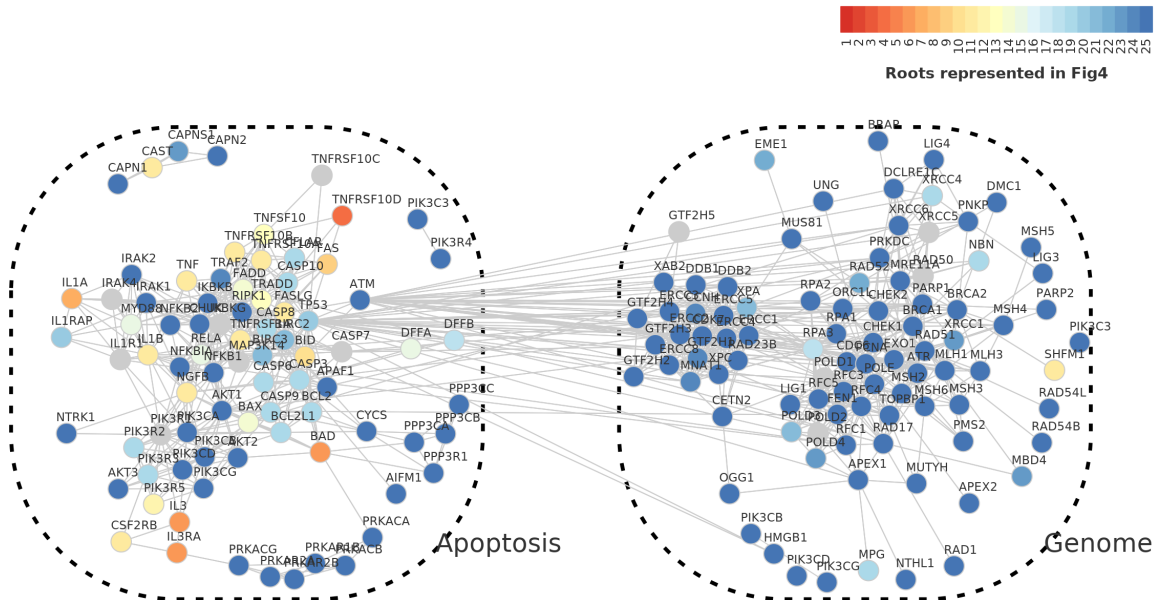


Figure 5. Inferred evolutionary roots of a protein-protein interaction network.

Map rooting information on regulatory networks

This example aims to show the evolutionary root of regulons (Fletcher et al. 2013). The idea is to map the appearance of each regulon (and the corresponding target genes) in a species tree. The next steps show how to transfer evolutionary rooting information from **geneplast** to a graph model. *Note: to make this work the gene annotation available from the input regulatory network needs to match the annotation available from the geneplast data (in this case, ENTREZ gene IDs are used to match the datasets).*

1 - Load a **TNI** class object and required packages. The **rtni1st** object provides regulons available from the **Fletcher2013b** data package computed from breast cancer data (Fletcher et al. 2013).

```

library(RTN)
library(Fletcher2013b)
library(RedeR)
library(igraph)
library(RColorBrewer)
data("rtni1st")

```

2 - Extract two regulons from **rtni1st** into an **igraph** object.

```

regs <- c("FOXM1", "PTTG1")
g <- rtni.graph(rtni1st, gtype = "rmap", tfs = regs)

```

3 - Map rooting information on the **igraph** object.


```
g <- ogr2igraph(ogr, cogdata, g, idkey = "ENTREZ")
```

4 - Adjust colors for rooting information.

```
pal <- brewer.pal(9, "RdYlBu")
color_col <- colorRampPalette(pal)(25) #set a color for each root!
g <- att.setv(g=g, from="Root", to="nodeColor", cols=color_col,
             na.col = "grey80", breaks = seq(1,25))
```

5 - Aesthetic adjusts for some graph attributes.

```
idx <- V(g)$SYMBOL %in% regs
V(g)$nodeFontSize[idx] <- 30
V(g)$nodeFontSize[!idx] <- 1
E(g)$edgeColor <- "grey80"
V(g)$nodeLineColor <- "grey80"
```

6 - Send the igraph object to **RedeR** interface.

```
rdp <- RedPort()
callD(rdp)
resetD(rdp)
addGraph(rdp, g, layout=NULL)
addLegend.color(rdp, colvec=g$legNodeColor$scale, size=15,
               labvec=g$legNodeColor$legend, title="Roots represented in Fig4")
relax(rdp, 15, 100, 20, 50, 10, 100, 10, 2, ps=TRUE)
```

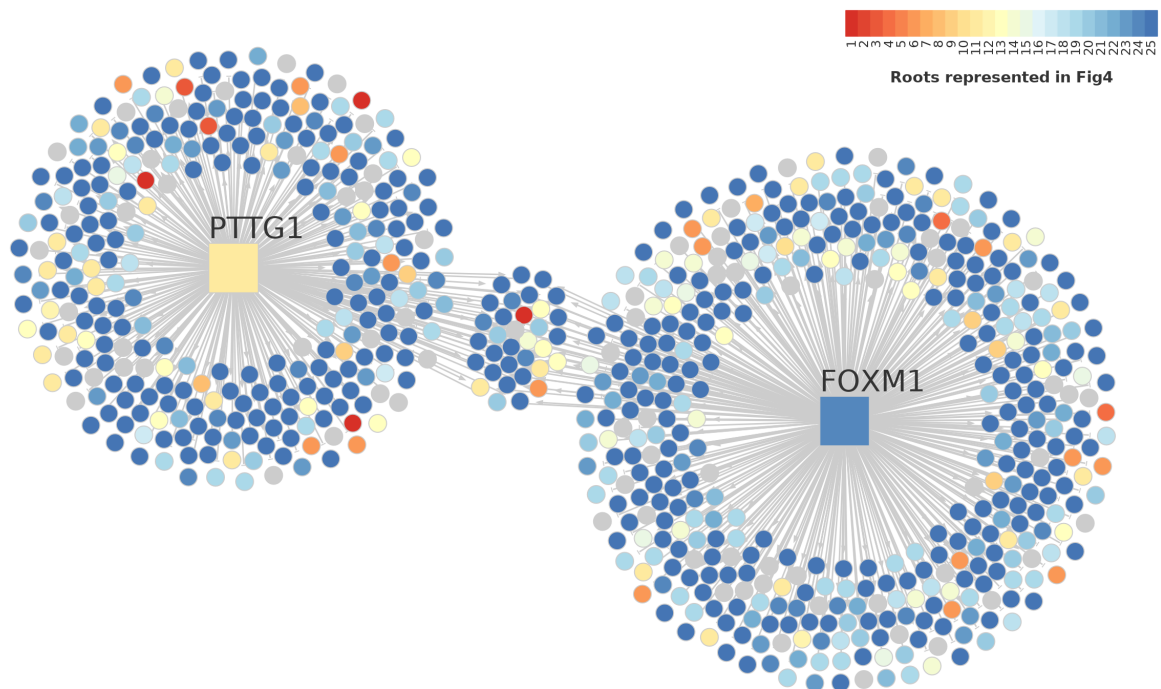


Figure 6. Inferred evolutionary roots of two regulators (FOXM1 and PTTG1) and the corresponding targets.

Session information

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.1 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
## [1] LC_CTYPE=en_GB.UTF-8      LC_NUMERIC=C
## [3] LC_TIME=en_GB.UTF-8       LC_COLLATE=en_GB.UTF-8
## [5] LC_MONETARY=en_GB.UTF-8   LC_MESSAGES=en_GB.UTF-8
## [7] LC_PAPER=en_GB.UTF-8      LC_NAME=C
## [9] LC_ADDRESS=C              LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] geneplast_1.6.2
##
## loaded via a namespace (and not attached):
## [1] Rcpp_0.12.18  lattice_0.20-35 ape_5.1      snow_0.4-2
## [5] digest_0.6.15 rprojroot_1.3-2 grid_3.5.1   nlme_3.1-137
## [9] backports_1.1.2 magrittr_1.5   evaluate_0.11 stringi_1.2.4
## [13] rmarkdown_1.10 BiocStyle_2.8.2 tools_3.5.1  stringr_1.3.1
## [17] igraph_1.2.2  parallel_3.5.1 yaml_2.2.0   compiler_3.5.1
## [21] pkgconfig_2.0.2 htmltools_0.3.6 knitr_1.20
```

References

- Castro, Mauro AA, Rodrigo JS Dalmolin, Jose CF Moreira, Jose CM Mombach, and Rita MC de Almeida. 2008. "Evolutionary Origins of Human Apoptosis and Genome-Stability Gene Networks." *Nucleic Acids Research* 36 (19): 6269–83. doi:10.1093/nar/gkn636.
- Dalmolin, Rodrigo JS, Mauro AA Castro, Jose Rybarczyk-Filho, Luis Souza, Rita MC de Almeida, and Jose CF Moreira. 2011. "Evolutionary Plasticity Determination by Orthologous Groups Distribution." *Biology Direct* 6 (1): 22. doi:10.1186/1745-6150-6-22.
- Fletcher, Michael, Mauro Castro, Suet-Feung Chin, Oscar Rueda, Xin Wang, Carlos Caldas, Bruce Ponder, Florian Markowetz, and Kerstin Meyer. 2013. "Master Regulators of Fgfr2 Signalling and Breast Cancer Risk." *Nature Communications* 4: 2464. doi:10.1038/ncomms3464.

ANEXO D – ARTIGO PUBLICADO - (Corces *et al.*, 2018)

The chromatin accessibility landscape of primary human cancers. Corces *et al.*, SCIENCE, v. 362, n.6413 p. 420-434, 2018. DOI:10.1126/science.aav1898

O artigo é um estudo sobre a acessibilidade da cromatina combinando diversas abordagens, entre elas a reconstrução de redes regulatórias na análise da predisposição ao desenvolvimento do câncer.

A participação da doutoranda nesse trabalho é derivada de diversas atualizações efetuadas no pacote RTN, a ferramenta utilizada para reconstrução de redes regulatórias utilizada no *paper* em questão, realizadas em decorrência da criação da abordagem central dessa tese.

RESEARCH ARTICLE SUMMARY

CANCER

The chromatin accessibility landscape of primary human cancers

M. Ryan Corces*, Jeffrey M. Granja*, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, Clarice Groeneveld, Christopher K. Wong, Seung Woo Cho, Ansuman T. Satpathy, Maxwell R. Mumbach, Katherine A. Hoadley, A. Gordon Robertson, Nathan C. Sheffield, Ina Felau, Mauro A. A. Castro, Benjamin P. Berman, Louis M. Staudt, Jean C. Zenklusen, Peter W. Laird, Christina Curtis, The Cancer Genome Atlas Analysis Network, William J. Greenleaf†, Howard Y. Chang†

INTRODUCTION: Cancer is one of the leading causes of death worldwide. Although the 2% of the human genome that encodes proteins has been extensively studied, much remains to be learned about the noncoding genome and gene regulation in cancer. Genes are turned on and off in the proper cell types and cell states by transcription factor (TF) proteins acting on DNA regulatory elements that are scattered over the vast noncoding genome and exert long-range influences. The Cancer Genome Atlas (TCGA) is a global consortium that aims to accelerate the understanding of the molecular basis of cancer. TCGA has systematically collected DNA mutation, methyl-

ation, RNA expression, and other comprehensive datasets from primary human cancer tissue. TCGA has served as an invaluable resource for the identification of genomic aberrations, altered transcriptional networks, and cancer subtypes. Nonetheless, the gene regulatory landscapes of these tumors have largely been inferred through indirect means.

RATIONALE: A hallmark of active DNA regulatory elements is chromatin accessibility. Eukaryotic genomes are compacted in chromatin, a complex of DNA and proteins, and only the active regulatory elements are accessible by the cell's machinery such as TFs. The assay for

transposase-accessible chromatin using sequencing (ATAC-seq) quantifies DNA accessibility through the use of transposase enzymes that insert sequencing adapters at these accessible chromatin sites. ATAC-seq enables the genome-wide profiling of TF binding events that orchestrate gene expression programs and give a cell its identity.

RESULTS: We generated high-quality ATAC-seq data in 410 tumor samples from TCGA, identifying diverse regulatory landscapes across 23 cancer types. These chromatin accessibility profiles identify cancer- and tissue-specific DNA regulatory elements that enable classification of

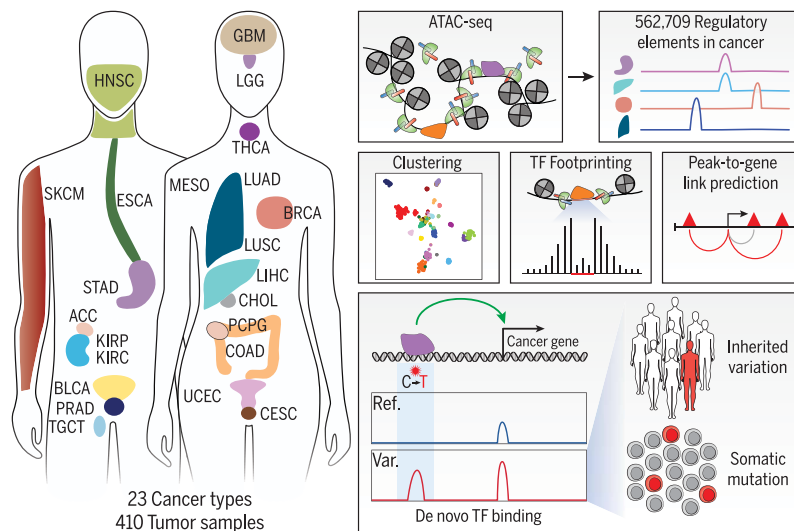
ON OUR WEBSITE

Read the full article at <http://dx.doi.org/10.1126/science.aav1898>

tumor subtypes with newly recognized prognostic importance. We identify distinct TF activities in cancer based on differences in the inferred patterns of TF-DNA interaction and gene

expression. Genome-wide correlation of gene expression and chromatin accessibility predicts tens of thousands of putative interactions between distal regulatory elements and gene promoters, including key oncogenes and targets in cancer immunotherapy, such as *MYC*, *SRC*, *BCL2*, and *PDL1*. Moreover, these regulatory interactions inform known genetic risk loci linked to cancer predisposition, nominating biochemical mechanisms and target genes for many cancer-linked genetic variants. Lastly, integration with mutation profiling by whole-genome sequencing identifies cancer-relevant noncoding mutations that are associated with altered gene expression. A single-base mutation located 12 kilobases upstream of the *FGD4* gene, a regulator of the actin cytoskeleton, generates a putative de novo binding site for an NKX TF and is associated with an increase in chromatin accessibility and a concomitant increase in *FGD4* gene expression.

CONCLUSION: The accessible genome of primary human cancers provides a wealth of information on the susceptibility, mechanisms, prognosis, and potential therapeutic strategies of diverse cancer types. Prediction of interactions between DNA regulatory elements and gene promoters sets the stage for future integrative gene regulatory network analyses. The discovery of hundreds of noncoding somatic mutations that exhibit allele-specific regulatory effects suggests a pervasive mechanism for cancer cells to manipulate gene expression and increase cellular fitness. These data may serve as a foundational resource for the cancer research community. ■



Cancer gene regulatory landscape. Chromatin accessibility profiling of 23 human cancer types (left) in 410 tumor samples from TCGA revealed 562,709 DNA regulatory elements. The activity of these DNA elements organized cancer subtypes, identified TF proteins and regulatory elements controlling cancer gene expression, and suggested molecular mechanisms for cancer-associated inherited variants and somatic mutations in the noncoding genome. See main article for abbreviations of cancer types. Ref., reference; Var., variant.

The list of author affiliations is available in the full article online. *These authors contributed equally to this work.

†Corresponding author. Email: howchang@stanford.edu (H.Y.C.); wjg@stanford.edu (W.J.G.)

Cite this article as M. R. Corces et al., *Science* 362, eaav1898 (2018). DOI: 10.1126/science.aav1898

The Cancer Genome Atlas Analysis Network Collaborators List

Rehan Akbani¹⁹, Christopher C. Benz²⁰, Evan A. Boyle²¹, Bradley M. Broom¹⁹, Andrew D. Cherniack^{22,23}, Brian Craft²⁴, John A. Demchok²⁵, Ashley S. Doane²⁶, Olivier Elemento²⁶, Martin L. Ferguson²⁵, Mary J. Goldman²⁴, D. Neil Hayes²⁷, Jing He²⁸, Toshinori Hinoue²⁹, Marcin Imielinski²⁶, Steven J.M. Jones³⁰, Anab Kemal²⁵, Theo A. Knijnenburg³¹, Anil Korkut¹⁹, De-Chen Lin³², Yuexin Liu¹⁹, Michael K.A. Mensah²⁵, Gordon B. Mills³³, Vincent P. Reuter³⁴, Andre Schultz¹⁹, Hui Shen²⁹, Jason P. Smith³⁴, Roy Tarnuzzer²⁵, Sheyla Trefflich³⁵, Zhining Wang²⁵, John N. Weinstein¹⁹, Lindsay C. Westlake^{22,23}, Jin Xu²⁸, Liming Yang²⁵, Christina Yau^{20,36}, Yang Zhao²⁸, Jingchun Zhu²⁴

¹⁹Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX 77030, USA.

²⁰Buck Institute for Research on Aging, Novato, CA 94945, USA.

²¹Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.

²²Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.

²³Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA 02215, USA

²⁴Genomics Institute, University of California, Santa Cruz, CA 95064, USA.

²⁵National Cancer Institute, Bethesda, MD 20892, USA.

²⁶Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA.

²⁷Department of Genetics and Genomics, University of Tennessee Health Science Center, Memphis, TN 38117, USA.

²⁸Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA.

²⁹Van Andel Research Institute, Grand Rapids, MI 49503, USA.

³⁰Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, V5Z 4S6, Canada.

³¹Institute for Systems Biology, Seattle, WA 98109, USA.

³²Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.

³³Oregon Health Sciences University, Knight Cancer Institute, Portland, OR 97239, USA.

³⁴Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA.

³⁵Graduate Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, 31270-901, Brazil.

³⁶Department of Surgery, University of California, San Francisco, CA 94115, USA.

ANEXO E – ARTIGO PUBLICADO - (Chagas *et al.*, 2019)

RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons. Chagas *et al.*, *BIOINFORMATICS*, v. 35, n. 24, p. 5357-5358, 2019. DOI: 10.1093/bioinformatics/btz534

O artigo introduz o pacote RTNduals (depositado no repositório Bioconductor) uma ferramenta que identifica uma dupla de regulons com um número de alvos compartilhados estatisticamente significativo. Esse é um pacote que utiliza as redes regulatórias transcricionais para identificar associações co-reguladoras entre regulons e que se estendeu a partir do pacote RTN.

A participação da doutoranda nesse trabalho é derivada de diversas atualizações efetuadas no pacote RTN, a ferramenta utilizada para reconstrução de redes regulatórias utilizada no *paper* em questão, efetuadas em decorrência da criação da abordagem central dessa tese. A participação também decorre da criação da ilustração presente nessa publicação.

Systems biology

***RTNduals*: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons**

Vinicius S. Chagas^{1,†}, Clarice S. Groeneveld^{1,†}, Kelin G. Oliveira^{1,2},
Sheyla Trefflich³, Rodrigo C. de Almeida⁴, Bruce A. J. Ponder⁵,
Kerstin B. Meyer^{5,6}, Steven J. M. Jones⁷, A. Gordon Robertson^{7,*,*‡} and
Mauro A. A. Castro^{1,*,*‡}

¹Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba 81520-260, Brazil, ²Department of Clinical Sciences, Section of Oncology and Pathology, Lund University, Lund 221 85, Sweden, ³Graduate Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil, ⁴Department of Biomedical Data Sciences, Molecular Epidemiology, Leiden University Medical Center, Leiden 2300 RC, The Netherlands, ⁵Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, UK, ⁶Wellcome Sanger Institute, Hinxton CB10 1SA, UK and ⁷Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver V5Z 4S6, Canada

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Janet Kelso

Received on September 9, 2018; revised on April 1, 2019; editorial decision on June 21, 2019; accepted on June 26, 2019

Abstract

Motivation: Transcription factors (TFs) are key regulators of gene expression, and can activate or repress multiple target genes, forming regulatory units, or regulons. Understanding downstream effects of these regulators includes evaluating how TFs cooperate or compete within regulatory networks. Here we present *RTNduals*, an R/Bioconductor package that implements a general method for analyzing pairs of regulons.

Results: *RTNduals* identifies a dual regulon when the number of targets shared between a pair of regulators is statistically significant. The package extends the *RTN* (Reconstruction of Transcriptional Networks) package, and uses *RTN* transcriptional networks to identify significant co-regulatory associations between regulons. The [Supplementary Information](#) reports two case studies for TFs using the METABRIC and TCGA breast cancer cohorts.

Availability and implementation: *RTNduals* is written in the R language, and is available from the Bioconductor project at <http://bioconductor.org/packages/RTNduals/>.

Contact: grobertson@bcgsc.ca or mauro.castro@ufpr.br

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Gene regulation in eukaryotes integrates a large number of interconnected regulatory influences. Some of the major contributors in gene regulation are transcription factors (TFs): proteins that can act as

activators or repressors of gene expression, typically by binding to regulatory DNA regions and recruiting the transcriptional apparatus (Yamaguchi *et al.*, 2017). TFs are widely used in methods that reconstruct transcriptional networks, and algorithms that reconstruct

ANEXO F – ARTIGO PUBLICADO - (Mathias *et al.*, 2021)

Novel lncRNAs Co-Expression Networks Identifies LINC00504 with Oncogenic Role in Luminal A Breast Cancer Cells. Mathias *et al.*, INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES, v. 22, n. 2420, p. 1-15, 2021. DOI:10.3390/ijms22052420

O artigo descreve o estudo do potencial regulatório de redes de co-expressão com regulons de RNAs longos não codificantes (lncRNA) em dados de pacientes com câncer de mama.

A participação da doutoranda nesse trabalho é derivada de análises efetuadas em conjunto utilizando o pacote RTN.



Article

Novel lncRNAs Co-Expression Networks Identifies LINC00504 with Oncogenic Role in Luminal A Breast Cancer Cells

Carolina Mathias ¹, Clarice S. Groeneveld ^{2,3}, Sheyla Trefflich ⁴, Erika P. Zambalde ¹, Rubens S. Lima ⁵, Cícero A. Urban ⁵, Karin B. Prado ¹, Enilze M. S. F. Ribeiro ¹, Mauro A. A. Castro ⁴ , Daniela F. Gradia ¹ and Jaqueline C. de Oliveira ^{1,*}

¹ Post-Graduation Program in Genetics, Department of Genetics, Federal University of Parana, Curitiba 81530-900, PR, Brazil; carol.mathias1@hotmail.com (C.M.); erikazambaldi@gmail.com (E.P.Z.); kbraun@ufpr.br (K.B.P.); enilzeribeiro@gmail.com (E.M.S.F.R.); danielagrada@gmail.com (D.F.G.)

² Cartes d'Identité des Tumeurs Program, Ligue Nationale Contre le Cancer, 75013 Paris, France; clari.groeneveld@gmail.com

³ Oncologie Moléculaire, Institut Curie, CNRS, UMR144, Equipe Labellisée Ligue Contre le Cancer, 75005 Paris, France

⁴ Bioinformatics and Systems Biology Laboratory, Polytechnic Center, Federal University of Parana (UFPR), Curitiba 81520-260, PR, Brazil; sheylatrefflich@gmail.com (S.T.); mauro.a.castro@gmail.com (M.A.A.C.)

⁵ Breast Disease Center, Hospital Nossa Senhora das Graças, Curitiba 80810040, PR, Brazil; rsilveiralima@uol.com.br (R.S.L.); cicerourban@hotmail.com (C.A.U.)

* Correspondence: jaqueline.carvalho@ufpr.br



Citation: Mathias, C.; Groeneveld, C.S.; Trefflich, S.; Zambalde, E.P.; Lima, R.S.; Urban, C.A.; Prado, K.B.; Ribeiro, E.M.S.F.; Castro, M.A.A.; Gradia, D.F.; et al. Novel lncRNAs Co-Expression Networks Identifies LINC00504 with Oncogenic Role in Luminal A Breast Cancer Cells. *Int. J. Mol. Sci.* **2021**, *22*, 2420. <https://doi.org/10.3390/ijms22052420>

Received: 29 December 2020

Accepted: 25 January 2021

Published: 28 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Long non-coding RNAs (lncRNAs) are functional transcripts with more than 200 nucleotides. These molecules exhibit great regulatory capacity and may act at different levels of gene expression regulation. Despite this regulatory versatility, the biology of these molecules is still poorly understood. Computational approaches are being increasingly used to elucidate biological mechanisms in which these lncRNAs may be involved. Co-expression networks can serve as great allies in elucidating the possible regulatory contexts in which these molecules are involved. Herein, we propose the use of the pipeline deposited in the RTN package to build lncRNAs co-expression networks using TCGA breast cancer (BC) cohort data. Worldwide, BC is the most common cancer in women and has great molecular heterogeneity. We identified an enriched co-expression network for the validation of relevant cell processes in the context of BC, including LINC00504. This lncRNA has increased expression in luminal subtype A samples, and is associated with prognosis in basal-like subtype. Silencing this lncRNA in luminal A cell lines resulted in decreased cell viability and colony formation. These results highlight the relevance of the proposed method for the identification of lncRNAs in specific biological contexts.

Keywords: LINC00504; breast cancer; co-expression; lncRNA; luminal A

1. Introduction

Non-coding RNAs are a big class of transcripts that can be classified according to their size, comprising small RNAs <200 nucleotides (nt) and long non-coding RNAs (lncRNA) >200 nt [1]. lncRNA molecules are usually transcribed by RNA polymerase II, capped, and polyadenylated with some being also spliced. lncRNAs present high tissue specificity and great regulatory versatility, acting at different levels of gene expression regulation [2,3]. lncRNAs have already been analyzed in several human diseases, including cancer, with varying regulatory activity as either oncogenic or tumor suppressor, whose activity can modulate all hallmarks of cancer [4]. For example, the lncRNA HOTAIR can promote tumor growth and metastasis in several cancer types, such as breast, hepatocellular, lung and gastric cancer. [5]. One lncRNA known to act as a tumor suppressor, regulating p53, is the Maternally expressed gene 3 (MEG3). Several studies have shown down-regulation of this lncRNA in human cancers, such as lung, breast, gastric and colorectal [6].

ANEXO G – ARTIGO SUBMETIDO - (Cardoso *et al.*, 2021)

TreeAndLeaf: an R/Bioconductor package for representing graphs and trees with focus on the leaves. Cardoso *et al.*, 2021.

O artigo ainda está em fase de submissão e introduz a ferramenta TreeAndLeaf (depositada no repositório Bioconductor) que combina algoritmos na organização de árvores binárias, possibilitando a representação combinada de várias camadas de informações em folhas de dendrograma.

A participação da doutoranda nesse trabalho é decorrente da elaboração conjunta das ilustrações presentes no artigo e de dois dos estudos de caso que são exemplos para a utilização da ferramenta.

TreeAndLeaf : R/Bioconductor package for representing graphs and trees with focus on the leaves

Wilson A. Cardoso^{1,†}, Luis E. A. Miranda^{1,†}, Leonardo M. de F. Costa^{1,†},
Gustavo A. S. Pereira¹, Sheyla Trefflich^{1,2}, Diego Arthur de Aguiar de Moraes³, Rodrigo
J. S. Damasceno³, Bruce A. J. Pereira⁴, Karoline B. Meyer⁵, Mauro A. A. Castro^{1,*}

¹Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba, 81520-260, Brazil.

²Bioinformatics Department, Federal University of Minas Gerais, Belo Horizonte, 31270-901, Brazil.

³Bioinformatics Infrastructure Environment, Federal University of Rio Grande, Rio Grande, 96201-900, Brazil.

⁴Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre for Health Research and Biomedicine, 477 Williamstown Road, Cambridge, CB2 0RQ, United Kingdom.

⁵Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, United Kingdom.

[†]The authors wish to be known that, in their opinion, the first three authors should be considered as Joint First Authors.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Dendrogram is a classical diagram for visualizing binary trees. Although efficient to represent hierarchical relations, it provides limited space for displaying information on the leaf elements, especially for large trees.

Results: Here we present TreeAndLeaf, an R/Bioconductor package that implements a hybrid layout strategy to represent tree diagrams with focus on the leaves. The TreeAndLeaf package combines force-directed graph and tree layout algorithms using a single visualization system, allowing projection of multiple layers of information onto a graph-tree diagram. The **Supplementary Information** provides two case studies that use breast cancer data from epidemiological and experimental studies.

Availability: TreeAndLeaf is written in the R language, and is available from the Bioconductor project at <http://bioconductor.org/packages/TreeAndLeaf/> (version >= 1.4.1)

Contact: mauro.castro@ufpr.br

Supplementary information: Supplementary data are available at [www.biorxiv.org](https://www.biorxiv.org/content/10.1101/2018.08.14.244444) online.

1 Introduction

A dendrogram is a classical diagram used to represent hierarchical relations. It is a type of tree, containing an origin (root), edges (branches), inner and outer nodes, as illustrated in Figure 1A. The root is where the branches and nodes come from, indicating the direction to the outer nodes or leaves. The dendrogram is also a binary tree, in which the number of edges connected to inner nodes is not greater than three (Lapointe and Legendre, 1995).

Dendrograms are effectively used to represent trees and clustering structures. However, a dendrogram layout optimizes some aesthetic

qualities (Rusu and Santiago, 2008), using most of the diagram space to arrange branches and inner nodes, a tradeoff that results in limited space to arrange the leaves. For large dendrograms, the outer nodes are often squeezed into small slots. Therefore, a dendrogram may not provide the best layout when the user needs to visualize the information contained in the leaves.

The TreeAndLeaf package aims to improve the visualization of dendrogram leaves by combining force-directed graph and tree layout algorithms, a hybrid strategy that refines the space distribution of the outer nodes, shifting the focus of the visualization to the leaves. We have originally developed this approach for integrative network analysis (Castro et al., 2016; Campbell et al., 2017) and we anticipate a broader applicability of the TreeAndLeaf package due to the widespread use of dendrograms.

ANEXO H – CAPÍTULO DE LIVRO PUBLICADO - (Cruz *et al.*, 2017)

Protein Function Prediction foi redigido para figurar como um capítulo do livro Functional Genomics. A participação da doutoranda nesse trabalho é decorrente da criação conceitual das figuras que ilustram o processo de anotação funcional de proteínas:

- Fluxograma de orientação no estabelecimento e entendimento da homologia entre sequências de proteínas;
- Fluxograma de demonstração do processo de anotação de uma sequência de proteína, um proteoma ou um metagenoma por comparação com um banco de dados existente;
- Estudo de caso que inclui a utilização de ferramentas relacionadas com: busca de sequência de proteínas, alinhamento, anotação, análise de família de proteínas, análise de ontologia, predição de hélices transmembrana, interação proteína-proteína.



[Functional Genomics](#) pp 55-75 | [Cite as](#)

Protein Function Prediction

Authors

[Authors and affiliations](#)

Leonardo Magalhães Cruz , Sheyla Trefflich, Vinicius Almir Weiss, Mauro Antônio Alves Castro

Protocol

First Online: 13 September 2017

3
Readers

1.2k
Downloads

Part of the [Methods in Molecular Biology](#) book series (MIMB, volume 1654)

Abstract

Protein function is a concept that can have different interpretations in different biological contexts, and the number and diversity of novel proteins identified by large-scale “omics” technologies poses increasingly new challenges. In this review we explore current strategies used to predict protein function focused on high-throughput sequence analysis, as for example, inference based on sequence similarity, sequence composition, structure, and protein–protein interaction. Various prediction strategies are discussed together with illustrative workflows highlighting the use of some benchmark tools and knowledge bases in the field.