UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
DEPARTAMENTO DE BIOLOGIA MOLECULAR E BIOTECNOLOGIA
TRABALHO DE CONCLUSÃO EM BIOINFORMÁTICA

ITAMAR JOSÉ GUIMARÃES NUNES

GENE EXPRESSION ANALYSIS PLATFORM (GEAP): UMA PLATAFORMA FLEXÍVEL E INTUITIVA PARA ANÁLISES DE TRANSCRIPTOMA

Porto Alegre
Junho de 2018

ITAMAR JOSÉ GUIMARÃES NUNES

GENE EXPRESSION ANALYSIS PLATFORM (GEAP): UMA PLATAFORMA FLEXÍVEL E INTUITIVA PARA ANÁLISES DE TRANSCRIPTOMA

Trabalho de Conclusão apresentado à Comissão de Graduação do Curso de Biotecnologia da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de bacharel em Bioinformática.

ORIENTADORA: Profa. Dra. Mariana Recamonde Mendoza
CO-ORIENTADOR: Dr. Bruno César Feltes

Porto Alegre
Junho de 2018

ITAMAR JOSÉ GUIMARÃES NUNES

GENE EXPRESSION ANALYSIS PLATFORM (GEAP): UMA PLATAFORMA FLEXÍVEL E INTUITIVA PARA ANÁLISES DE TRANSCRIPTOMA

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Bioinformática e aprovado em sua forma final pela orientadora e pela Banca Examinadora.

orientadora:_____
Profa. Dra. Mariana Recamonde Mendoza
Doutora pelo Instituto de Informática/UFRGS – Porto Alegre, RS

Banca Examinadora:

Profa. Dra. Ursula Matte, UFRGS
Doutora pelo Departamento de Genética – Porto Alegre, RS

Dr. Tiago Falcon Lopes, HCPA
Doutor pela Universidade de São Paulo/USP – Ribeirão Preto, SP

Porto Alegre
Junho de 2018

RESUMO

Atualmente, diversas técnicas em biologia molecular estão disponíveis com o propósito de auxiliar na observação de processos biológicos. Dentre essas técnicas, estão as análises de transcriptoma, que possibilitam quantificar a expressão de praticamente todos os genes em uma amostra sob uma determinada condição. Ao analisar transcriptomas, é possível explorar um grande volume de informações sobre fenômenos biológicos, inclusive de diversas doenças. Felizmente, muitos dados de transcriptoma estão disponíveis no banco de dados do *Gene Expression Omnibus* (GEO). Porém, analisar estes dados não é uma tarefa simples, sendo necessário um software específico para o Kit experimental utilizado na análise ou, alternativamente, conhecimento na linguagem de programação R. Os softwares de transcriptoma costumam suportar apenas arquivos do próprio fabricante, e podem não oferecer opções flexíveis para manipular seus dados. Por outro lado, aprender uma linguagem de programação não é trivial, e mesmo usuários com maior experiência podem ter dificuldades ao lidar com vários formatos de arquivos do GEO, considerando as milhares de plataformas presentes. Além disso, diminuir os requisitos de informática nesse tipo de análise pode acelerar a eficiência de obter seus resultados. Neste sentido, o programa *Gene Expression Analysis Console* (**GEAP**) foi desenvolvido com o propósito de analisar transcriptomas de forma visual e intuitiva, tendo maior foco em dados de microarranjo. Através do **GEAP**, o usuário pode: (i) baixar dados de séries e amostras do GEO sabendo apenas seu código de série (*Geo Series*, ou GSE) ou de amostra (*Geo Sample*, ou GSM), sejam estes dados brutos ou previamente tratados; (ii) ler arquivos de transcriptoma pertencentes a milhares de plataformas atualmente disponíveis; (iii) permitir a criação de seu próprio conjunto de dados customizado, sendo uma ferramenta eficaz para lidar com formatos de arquivo de tabela; (iv) preparar e tratar estatisticamente os dados com algoritmos documentados pela literatura; (v) verificar a qualidade dos dados de forma visual e detalhada através de diagramas representativos; (vi) comparar a diferença de expressão entre as amostras e automatizar este processo seguindo metodologias bem consolidadas; (vii) visualizar os resultados de expressão diferencial por meio de tabelas otimizadas e diferentes tipos de gráficos que respondem à interação do usuário; e (viii) filtrar e organizar os resultados de forma personalizada, permitindo encontrar genes significantes. Por fim, este software foi desenvolvido tentando reunir os melhores atributos dos programas de microarranjo e da programação em R, com a finalidade de que qualquer usuário possa ter fácil acesso a análises de expressão gênica por transcriptoma, ao mesmo tempo permitindo uma forte flexibilidade de manipulação dos dados, que até então era alcançada apenas com o R.

Palavras-Chave: Genes, Expressão, Transcriptoma, Bioinformática, R.

ABSTRACT

Currently, a number of procedures in molecular biology are available in order to help the exploring of biological processes. Among these procedures, there is the transcriptome analysis, which allows to measure gene expression of virtually all genes contained in a set of samples under one specific condition. Through transcriptomic analyses, a large volume of information concerning several biological phenomena can be explored, including numerous diseases. Fortunately, many transcriptomic data are currently available in Gene Expression Omnibus (GEO) database. However, analyzing these data is not a trivial task, since it needs specific software for the experimental kit used in analysis, or alternatively, mastering the knowledge of R programming language. Usually, the transcriptome analysis softwares only support files from their own manufacturer, and also can lack flexible options for manipulating data. On the other side, it is not trivial to learn a programming language, and even experienced users can undergo hardship when dealing with the variety of file formats from GEO if we take into account the thousands of available platforms. Moreover, decreasing informatics requirements can bring more efficiency in getting results from the analyses. In this sense, a new software named *Gene Expression Analysis Console* (**GEAP**) was developed in order to analysise transcriptomes in a visual and intuitive manner, mainly focusing on microarray data. By using **GEAP**, the user can: (i) download series and samples from GEO by only knowing the GSE/GSM code, both for RAW data or values previously treated by the author; (ii) read transcriptomic files from thousands of available platforms; (iii) allow creating your own custom microarray data set, being as an effective tool for datasets formatted as tables; (iv) prepare the data and perform statistic treatment of expression values with well-documented algorithms from literature; (v) visually check the quality of your data with details through representative diagrams; (vi) compare the expression difference between samples and automate this process with at least six comparison methods, in addition to five options of statistic parameters for results correction; (vii) view the differentially expressed results in optimized tables and up to four dynamic charts that respond to user interaction; and (viii) filter and order results with customizable options, helping to find relevant genes. Lastly, this software was developed with as an attempt of putting together the best attributes from microarray programs and from R programming. Its aim is that any user could easily access transcriptome analysis, at the same time providing enough flexibility for data manipulation, which until now only has been reached with R.

Keywords: Gene, Expression, Transcriptomics, Bioinformatics, R.

SUMÁRIO

LISTA DE ABREVIATURAS

CDF	Chip Definition File
CEL	Cell File Format
DE	Diferencialmente Expresso (Sonda ou Gene)
FDR	False Discovery Rate
FTP	File Transfer Protocol
GDS	GEO Data Set
GEAP	Gene Expression Analysis Platform
GEO	Gene Expression Omnibus
GPL	GEO Platform
GSE	GEO Series
GSM	GEO Sample
GZip	Extensão GNU GZip
log_2FC	Logaritmo na base 2 de Fold-Change
MAS	MicroArray Suite (Affymetrix)
mRNA	RNA mensageiro
NCBI	National Center for Biotechnology Information
PNG	Extensão Portable Network Graphics
R	Linguagem de programação R
RAM	Random-access memory
RMA	Robust Multi-array Average
SNP	Single Nucleotide Polymorphism
SOFT	Simple Omnibus Format in Text
TAR	Extensão Tape Archive
TIFF	Extensão Tagged Image File Format
TXT	Extensão arquivo de texto
ValCTRL	Valor resumido para grupo controle
ValEXP	Valor resumido para grupo experimental

Parte I

Princípios básicos de Transcriptoma

1 INTRODUÇÃO

Durante o Projeto Genoma Humano, desde a década de 90 até o início do século XXI, acreditava-se que mapear todos os genes do genoma humano levaria à tona o pleno conhecimento do nosso organismo e, que consequentemente, nos daria a resposta para todas ou boa parte das doenças. Hoje, estima-se que menos que 2% do nosso genoma é realmente codificado para proteínas (CONSORTIUM et al., 2004; PENNISI, 2007), sendo sugeridos aproximadamente 19 mil genes codificantes (EZKURDIA et al., 2014). Destes genes, grande parte expressa de forma diferente dependendo do tipo celular, da condição ou do período de tempo em que a célula contendo o genoma se encontra (RALSTON; SHAW, 2008).

A expressão gênica ocorre quando uma sequência nucleotídica é utilizada como molde para a transcrição de um RNA. Em geral, a maioria é transcrita para RNA mensageiro (mRNA), o qual é traduzido para uma proteína no citoplasma (ALBERTS et al., 2013). Embora diversas moléculas atuem em funções biológicas, as proteínas possuem um papel fundamental entre as atividades bioquímicas do organismo e, em boa parte dos casos, é sua quantidade que regula um determinado processo biológico. Portanto, quantificar o mRNA que está sendo produzido por transcrição pode oferecer uma indicação do quanto um gene está sendo expresso até seu produto final — a proteína. Neste sentido, como um esforço para mensurar a expressão de todos os genes de um genoma dentro de condições específicas, atualmente utiliza-se diversas técnicas para quantificação do mRNA, dentre elas o **microarranjo** (derivado de *DNA Microarray*). Com esta técnica, é possível usar dezenas ou centenas de milhares de combinações de sequências como referência para medir a expressão de cada gene em um único experimento (SCHENA et al., 1995). Essas combinações se chamam **sondas** (*probes*); e em linhas gerais, cada sonda representa uma sequência complementar a de um gene específico, a fim de ”marcar”a presença desta sequência em uma amostra. A utilização das sondas será discutida com melhor profundidade nas seções seguintes.

Com o microarranjo, uma quantidade enorme de informações pode ser obtida a partir de poucas amostras, as quais descrevem o perfil de expressão de todos os genes em uma condição específica. Isso permite, por exemplo, que se observe quais genes estão expressos em uma amostra de tecido em condição patológica (e.g. câncer, inflamação crônica, infecções por patógenos etc), e quais diferenças este perfil possui para um tecido em estado não-patológico (WANG et al., 2005). O conjunto de informações de transcritos quantificados para amostras em condições específicas, seja por microarranjo ou outra técnica semelhante, é chamado **transcriptoma**.

Da mesma forma, isso gera um volume de dados extenso associado a um estudo. Com o propósito de garantir a reprodutibilidade deste tipo de estudo, a partir de 2003, revistas como *Physiological Genomics* e *American Physiological Society* passaram a adotar

uma norma que exige a todos os autores utilizando dados de transcriptoma por microarranjo para que submetam estes ao servidor on-line do *National Center for Biotechnology Information* (NCBI), como um pré-requisito para publicação dos resultados (VENTURA, 2005). O domínio em que se encontram os dados de expressão gênica no NCBI é o *Gene Expression Omnibus* (GEO)[1], originalmente criado para submissão, depósito e recuperação de dados de hibridização genômica e expressão gênica de alto rendimento (EDGAR; DOMRACHEV; LASH, 2002).

É importante constatar que existem diferentes tipos de técnicas para analisar transcriptomas. Até então, o microarranjo é o mais comum entre as técnicas, sendo categorizado pelo GEO dentro do grupo *Expression profiling by array*. Outros exemplos são o miRNome (categoria *Non-coding RNA profiling by high throughput sequencing*), onde se observa o perfil de expressão de microRNAs, e mais recentemente o RNA-Seq (categoria *Expression profiling by high throughput sequencing*), uma técnica mais complexa que também identifica *splicing* alternativo entre as sequências de RNA. Neste trabalho, o foco será *DNA microarrays*, que até então é uma das técnicas mais simples e computacionalmente baratas para análise de expressão gênica na bioinformática, além de possuir uma quantidade enorme de dados disponíveis em relação às outras modalidades de transcriptoma. Portanto, quando o termo 'transcriptoma' for mencionado aqui, estaremos nos referindo implicitamente ao conjunto de transcritos analisados por microarranjo.

Na prática, quando se deseja analisar os dados de microarranjo depositados no GEO, existem algumas opções:

1. Acessar o estudo pelo GEO2R, do próprio GEO;
2. Utilizar um programa disponibilizado pela plataforma que forneceu o kit para análise transcriptômica;
3. Entender programação e fazer uso da linguagem R, a qual possui uma grande diversidade de pacotes estatísticos para estudos em biologia molecular e bioinformática;
4. Enviar para um especialista ou laboratório dedicado que entenda sobre o assunto.

Naturalmente, os quatro casos possuem suas próprias limitações. O GEO2R funciona exclusivamente para séries de transcriptomas que estejam acuradas, que compõem uma baixa parcela dos estudos disponíveis. Já os programas fornecidos pelas plataformas são restritos aos formatos de dados gerados pelos kits destas, e muitas vezes carecem de abordagem estatísticas importantes — apenas citando exemplos: avaliação de qualidade, customização da normalização de valores de expressão ou correção do valor-p dos resultados não são sempre encontrados entre estes programas. E ainda que encontrados, um programa pode não ter os mesmos padrões estatísticos dos demais, não suportar como input um grande número de amostras ou não disponibilizar todas as possibilidades de análise, sendo bastante limitados.

O cenário se altera quando o usuário tem conhecimento de programação e estatística, pois a plataforma R é dotada de uma infinidade de opções de ferramentas que auxiliam qualquer análise de expressão gênica cujas metodologias já estão bem descritas pela literatura (GENTLEMAN et al., 2006). O problema é que a quantidade de usuários com conhecimento em ambas as áreas é escasso e, geralmente, estes mesmos usuários podem não ter conhecimento em biologia em si, o pode se tornar um empecilho em casos onde

[1] URL: `https://www.ncbi.nlm.nih.gov/geo/`

tal conhecimento é necessário. Um exemplo prático é quando se está estudando câncer: espera-se que as amostras apresentem uma variabilidade discrepante entre si (HANAUER et al., 2007), o que aos olhos de alguém sem entendimento no assunto poderia parecer que foi um erro no algoritmo ou na etapa experimental que causou as anormalidades. Além de tudo, em termos de eficiência, o pesquisador poderia ser se dedicar ao estudo biológico em questão, e não necessariamente ter a obrigação de saber programar para isso, ainda que este seja um conhecimento valioso para outros propósitos. Programação exige tempo, esforço e a frustração é quase inevitável para quem nunca teve contato com este assunto. Por fim, outra forma de resolver a questão de como analisar os dados seria solicitar esta função a terceiros, o que pode custar recursos, além de não ser interessante para quem deseja manter a privacidade de seu estudo.

Neste sentido, torna-se necessária uma ferramenta que permita analisar os dados de microarranjo de forma intuitiva, direta e confiável. Foi com esse propósito que foi desenvolvida a ferramenta *Gene Expression Analysis Console* (**GEAP**), um ambiente construído especialmente — porém não exclusivamente — para análise transcriptômica por microarranjo. O objetivo desta ferramenta é viabilizar a análise de dados de transcriptoma, proporcionando as seguintes funcionalidades:

- Apresentar uma interface visual, intuitiva e pronta para uso;
- Isentar o domínio de programação como um pré-requisito para analisar transcriptomas;
- Permitir a leitura de qualquer tipo de dados de microarranjo oferecido pelo banco de dados do GEO, bem como definir um método padronizado de obtenção dos resultados a partir destes dados;
- Oferecer múltiplas opções de tratamento estatístico conforme a exigência do usuário;
- Permitir a fácil manipulação dos dados, a fim de flexibilizar a comparação entre amostras de acordo com a preferência do usuário;
- Apresentar os dados de forma visual e consistente, auxiliando a interpretação dos resultados biológicos.

Adicionalmente, é importante que os métodos utilizados para qualquer aplicação biológica já estejam descritos em outros estudos, o que é o caso da maioria dos pacotes na linguagem R desenvolvidos para este propósito. Após publicados, esses pacotes se tornam disponíveis no servidor on-line do *Bioconductor*[2] em código aberto (GENTLEMAN et al., 2004), e muitos dos que citaremos adiante se tornaram referências padrão para análise de expressão gênica. Por sinal, uma das principais funcionalidades do **GEAP** é justamente utilizar o R em segundo plano durante essas análises, combinando as vantagens do R com as da linguagem compilada em C# para outras funções que são menos eficientes ou impossíveis no R — especialmente se tratando de representação gráfica eficiente e de multiprocessamento, que são limitações clássicas do R por este não ter sido desenvolvido com tal intuito.

Apesar de o **GEAP** ter como objetivo facilitar os passos das análises de transcriptomas, é indispensável que qualquer usuário possua o conhecimento prévio do assunto, e por este motivo, a próxima seção será dedicada à explicação da técnica de microarranjo em si. A

[2]URL: `https://www.bioconductor.org/`

seguir, na seção 3, as limitações da técnica e a motivação de desenvolver o **GEAP** serão discutidas. A partir da seção 4, o conteúdo do programa será apresentado, bem como a lógica de seu funcionamento. A conclusão do tema principal se sucederá na seção 5. Finalmente, na seção 6, perspectivas de como melhorar a ferramenta serão apontadas para suas futuras versões.

2 MICROARRANJO

Para se dar início a uma análise de microarranjo, a fim de identificar o perfil de expressão gênica de uma amostra biológica de interesse, é essencial entender a lógica por trás desta técnica. Portanto, será feita uma breve introdução de como e por que se utiliza microarranjos.

2.1 Fundamentos e procedimentos experimentais da técnica

Genes são representados como fragmentos do DNA que refletem em uma manifestação biológica. A maioria das funções bioquímicas no organismo, de fato, são realizadas por intermédio de proteínas, mas a estrutura e função destas proteínas geralmente são determinadas pelas sequências nucleotídicas dos genes que as codificam (DARNELL et al., 1990; ALBERTS et al., 2013). Um gene codificante, por ser espacialmente restrito à molécula de DNA, requer que uma sequência complementar a sua seja sintetizada e conduzida para o ribossomo, sendo que essa sequência refletirá na produção da proteína-alvo. A molécula intermediária que contém a sequência complementar do gene codificante no DNA e que será enviada ao ribossomo é chamada RNA mensageiro (mRNA). Pode-se assumir que a quantidade de mRNA sendo produzido é proporcional à produção da respectiva proteína, o que influencia nas funções biológicas que essa proteína atua[1] (KOUSSOUNADIS et al., 2015; LIU; BEYER; AEBERSOLD, 2016). Contudo, DNA e RNA são moléculas pequenas e invisíveis pela microscopia clássica, além apresentarem muitas combinações possíveis, tornando difíceis de serem diretamente quantificadas.

A técnica de microarranjo, por sua vez, tem como objetivo quantificar o mRNA produzido em uma amostra, e para isso utiliza moléculas (sondas) que se combinam com cada variação possível de mRNA (SCHENA et al., 1995). As sondas só se complementam ao mRNA correspondente se este está presente na amostra, e emitem uma luz ao se combinarem, sendo que a quantidade de luz emitida é proporcional ao quanto de mRNA está presente na amostra (Figura 2.1), assim como à qualidade do pareamento. Como só um tipo de luz é emitido, para identificar qual combinação de mRNA está sendo quantificada, cada sonda deve ser colocada separadamente na amostra. Portanto, a solução contendo a amostra é distribuída uniformemente entre dezenas ou centenas de milhares de poços com tamanho microscópico situados sobre uma placa minúscula (também conhecida como *Biochip*), onde cada poço contém uma sonda específica para cada gene (SCHENA et al., 1995). Em alguns casos, pode haver mais de um poço de sonda para cada variação de mRNA a fim de evitar ambiguidades ou falsos positivos — algo inerente de toda técnica

[1] Mais precisamente, quantidades maiores de mRNA nem sempre aumentam a quantidade de proteína, já o contrário necessariamente causa diminuição da produção proteica.

de biologia molecular. Isso dependerá da **plataforma**, que é o tipo de kit fornecido pela empresa que produziu as sondas. Algumas plataformas possuem mais variantes de sondas e outras menos, mas todas têm o mesmo objetivo de cobrir todos os genes possíveis para o organismo sendo estudado.

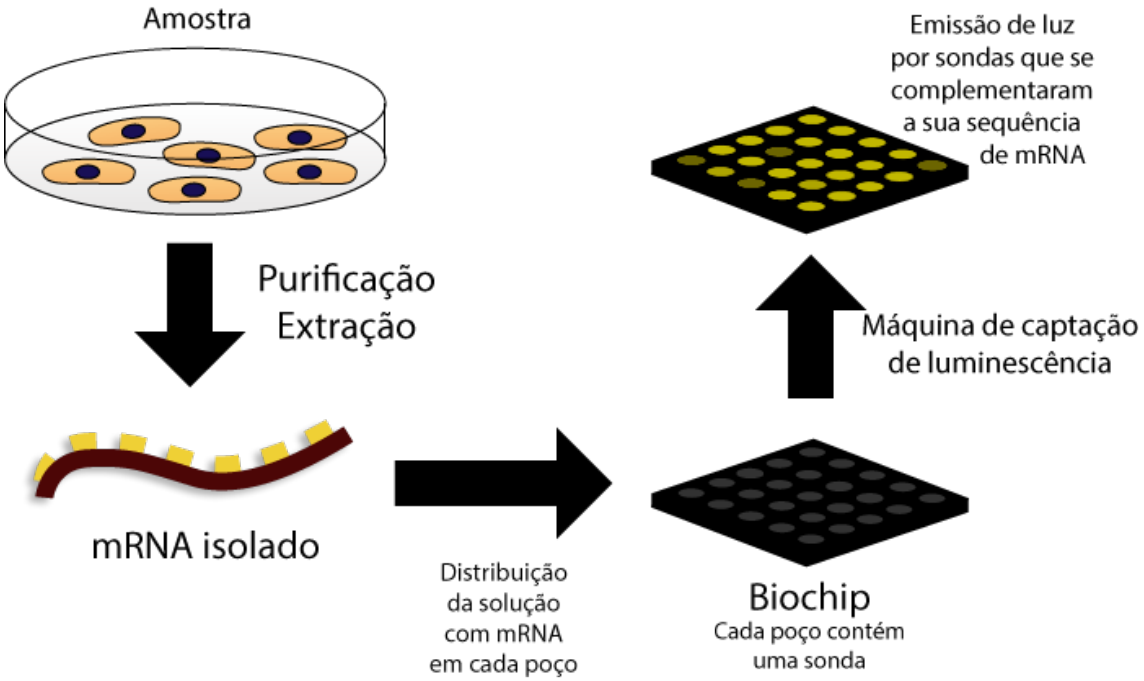

Figura 2.1: Diagrama resumido da técnica de microarranjo.

Vale ressaltar que, embora uma boa parcela das plataformas utilize apenas uma cor de iluminação, chamada canal único (ou *single-channel*), existem muitas plataformas que utilizam canal duplo (*dual-channel*), onde sondas que produzem uma iluminação verde para um conjunto de amostras são sobrepostas com sondas de iluminação vermelha para outro conjunto de amostras. No caso do *dual-channel*, utiliza-se duas amostras – uma para cada grupo – para cada *Biochip*. A luz emitida pelo *dual-channel* é amarela quando as sondas verdes e vermelhas hibridizam-se com o mRNA na mesma proporção, ou enviesada para verde ou vermelho dependendo de quanto uma condição apresenta maior concentração de mRNA hibridizado do que a outra. Embora as diferenças entre canal único ou duplo já tenham sido debatidas (BLALOCK, 2003; SÎRBU et al., 2012), a mesma lógica é aplicada para ambos os casos.

Independentemente da quantidade de canais, a obtenção dos dados é a mesma: a iluminação é captada por uma máquina e convertida em valores numéricos que posteriormente são processados por um computador. Os valores são agrupados entre dois conjuntos de amostras, um experimental (e.g. tecido patológico) e outro controle (e.g. tecido normal), e são comparados para cada sonda. Se um valor estiver maior no grupo experimental, então havia mais mRNA relativo a um certo gene na condição experimental, e vice-versa para o grupo controle. Vale destacar que estes valores não são absolutos, e a menos que todas as etapas de bancada tenham sido feitas de forma idêntica, desde a purificação de cada amostra até a diluição e distribuição entre os poços, as amostras só podem ser comparadas entre si dentro de um mesmo procedimento experimental. Por ser improvável que múltiplos procedimentos sejam iguais em seu decorrer prático, não é recomendado misturar amostras de diferentes experimentos.

Após realizado o estudo, ainda existe uma grande quantidade de informações de expressão gênica que foi obtida para um ou mais casos de estudo, e o autor pode optar em disponibilizar seus dados gratuitamente pelo GEO. Esse é o passo mais importante, pois novos estudos podem ser realizados utilizando dados que já existem. Como lidar com esses dados costuma ser uma tarefa para o profissional em bioinformática encarregado pela análise de microarranjo.

2.2 Dados de microarranjo

2.2.1 Componentes básicos

Uma análise completa de microarranjo é composta por três elementos básicos: sondas, amostras e anotações. Se organizássemos todas essas informações em uma tabela, o resultado seria algo semelhante ao Quadro 2.1.

Cada plataforma possui identificadores únicos de sonda, e a maioria dos arquivos utiliza esses identificadores para relacionar seus dados. Em uma analogia a bancos de dados relacionais, as sondas seriam a principal chave primária entre as tabelas de microarranjo, por isso certos programas as tratam como nomes de linhas de uma tabela.

As amostras, por outro lado, são colunas compostas por valores numéricos reais. Cada valor numérico tem relação com uma sonda, pois como foi explicado na seção anterior, tais valores são a quantidade de luz emitida em um micropoço do *Biochip*. Isso também significa que os valores não podem ser negativos, pois não existe emissão negativa de luz.

Em uma visão mais abstrata, apenas sondas e amostras seriam o suficiente para análises de transcriptoma pela razão de que os algoritmos foram programados para lidar apenas com valores e identificadores. Nesta lógica, contudo, o pesquisador teria em mãos não mais do que códigos e números. Deste modo, é necessário um contexto para cada sonda, o que entre os dados é chamado de **anotação** (*annotation*). A anotação é um conjunto de colunas que não faz parte das amostras e que serve para descrever as características de uma sonda — portanto, uma tabela de anotação tem os códigos de sonda como seus nomes de linha e uma característica de sonda para cada coluna. Comumente, a parte mais importante da anotação seria o símbolo ou nome do gene, o qual atribui à sonda seu

Quadro 2.1: Representação de uma tabela fictícia com os dados de microarranjo necessários para uma análise completa. A primeira coluna representa o código de cada sonda. A segunda coluna demostra os nomes dos genes associados à sonda da primeira coluna, destacando que a quarta linha exemplifica um caso de ambiguidade entre dois possíveis genes que a sonda relaciona. As colunas três em diante, com iniciais "GSM", representam amostras e seus valores de intensidade de luz emitida por sonda, e as reticências ilustram como se houvesse um total de 99 amostras na tabela (número arbitrário). Cada valor em uma coluna de amostra é relacionado com a sonda da mesma linha, identificada na primeira coluna (ID).

ID	Gene Symbol	GSM01	GSM02	...	GSM99
SONDA001	abc-51	863.11	450.40	...	28.77
SONDA002	xyz-1	686.84	345.78	...	119.87
SONDA003	abr-9	839.83	254.66	...	315.45
SONDA004	fgh-2 /// xyz-1	988.18	562.25	...	184.97

contexto biológico, mas outras informações como sequência nucleotídica ou ontologias gênicas também podem estar disponíveis dependendo da plataforma em questão.

2.2.2 Componentes do GEO

Dito acima, uma tabela contendo os três elementos básicos é o suficiente para uma análise transcriptômica. Onde encontrar a tabela será critério do usuário, podendo ser obtida diretamente a partir de dados provenientes de uma análise experimental ou minerada a partir de um banco de dados. Atualmente, o banco de dados com maior número de dados de microarranjo depositados é o GEO do NCBI, citado anteriormente. Neste servidor, o usuário tem acesso a diversos transcriptomas resultantes de estudos envolvendo áreas como biomedicina, microbiologia, botânica, dentre outras.

Os dados no GEO são agrupados em quatro categorias, sendo elas:

- Amostra (*GEO Sample*, ou GSM);
- Conjunto de dados (*GEO Dataset*, ou GDS);
- Série (*GEO Series*, ou GSE);
- Plataforma (*GEO Platform*, ou GPL).

De acordo com a documentação oficial do GEO[2], as siglas de três dígitos (GSM, GDS, GSE e GPL), acompanhadas por um número, são utilizadas como identificador para cada tabela de dados. Uma tabela do GEO, além de seu conteúdo, é acompanhada por uma tabela de duas colunas contendo sua descrição (e.g. título, identificador, data de publicação etc). Essa tabela de informações também inclui códigos de outras tabelas relacionadas a esta.

Um **GSM** é uma tabela de amostra, e sua composição principal é uma coluna de identificação de sonda e outra com os valores emitidos pela sonda. Não é incomum GSMs incluírem outras colunas que caracterizam estes valores, como valor-p, que qualifica a acurácia do sinal obtido. Cada GSM está relacionado a um GPL, que é a plataforma que descreve suas sondas. A sua tabela de informações indica este GPL associado, mas também pode apresentar características da amostra, como por exemplo, a idade do paciente ou o estágio da doença em que o tecido da amostra foi retirado durante um estudo biomédico.

GDS e **GSE** são conjuntos de dados amostrais, ou de modo simplificado, um grupo de GSMs. Também podem ser representados como o resultado de um estudo ou um conjunto de experimentos realizados pelos mesmos autores, e é comum que sua tabela de informações contenha o título e sumário do estudo. A diferença entre GDS e GSE é que o GDS contém dados acurados que podem ser observados pela página do GEO e analisados com a ferramenta GEO2R no próprio servidor. Além disso, todo GDS pode ser acessado como se fosse um GSE, mas nem todos os GSEs são acurados — na verdade, apenas uma parcela é qualificada pelo GEO para se tornar GDS.

Um **GPL** é uma tabela de anotações de sonda para uma plataforma específica. Cada plataforma define quais atributos de sonda estarão disponíveis, embora seja esperado que qualquer tabela GPL para microarranjo contenha uma coluna com os símbolos dos genes (*Gene Symbol*) que associa a sonda a um ou mais genes. Nota-se que uma mesma sonda

[2]URL: `https://www.ncbi.nlm.nih.gov/geo/info/`

pode apontar para múltiplos genes em certos casos, o que trata-se de uma limitação da técnica e da plataforma em questão.

Os quatro tipos de dados são armazenados em arquivos no servidor do GEO. Os formatos de arquivo mais comuns são HTML, para consulta no próprio site, e SOFT, que é uma versão textual das tabelas de conteúdo e informações. Entretanto, nem sempre os formatos do GEO são os desejados, pois os valores amostrais costumam ser previamente tratados pelo autor, e o GPL pode não conter todas as informações provenientes pela plataforma. Portanto, o GEO também disponibiliza um servidor FTP para o *download* dos dados brutos (formato *RAW*), que são os arquivos originais obtidos pela plataforma. Isso permite que o usuário realize por conta o tratamento estatístico das amostras, além de recuperar informações do GPL que podem não estar disponíveis no formato padrão do GEO.

2.2.3 Plataformas e fabricantes

Analisar um microarranjo a partir de seus dados brutos geralmente é a opção mais interessante, principalmente em casos onde o autor não realizou um tratamento estatístico correto sobre os dados ou surgiram técnicas mais recentes e qualificadas para lidar com os valores do arranjo. Apesar disso, analisar os dados brutos pode ser um desafio até mesmo para o usuário programador, e isso se deve à diversidade de plataformas disponíveis, pois cada GSE possui um formato definido de arquivo que varia conforme o fabricante. Alguns fabricantes destas plataformas mantém um formato de arquivo padrão para todos os GSEs de um mesmo GPL, enquanto que outros apresentam uma forte heterogeneidade entre os arquivos no contexto de uma mesma plataforma. A Tabela 2.2 exemplifica alguns dos tipos de dados gerados por cada fabricante. Três fabricantes são responsáveis pela maior parte das plataformas disponíveis no NCBI: **Affymetrix**[3], **Illumina**[4] e **Agilent**[5].

Tabela 2.2: Tipos de dados de microarranjo associados às três fabricantes citadas neste capítulo. Dados da Affymetrix e da Agilent costumam estar presentes em um mesmo formato para todos os casos de uso, enquanto que os arquivos da Illumina podem ser fornecidos em um ou mais formatos descritos nessa tabela. Outros tipos de arquivos não listados aqui podem ser consultados na Tabela 4.1, apresentada adiante (Seção 4.3).

Nome	Extensão	Fabricante
Cell Intensity File	CEL (.cel)	Affymetrix
Agilent Raw Data	Texto (.txt)	Agilent
Bead Level Data	Texto (.txt)	Illumina
Chip Definition File Package	Pacote R (.tar)	Affymetrix
Manifest File	BGX (.bgx) Texto (.txt)	Illumina
Intensity Data	IDAT (.idat)	Illumina

A Affymetrix, até então, assume a liderança em quantidade de séries de microarranjo depositadas no GEO (DU; KIBBE; LIN, 2007; GOHLMANN; TALLOEN, 2009). Seus

[3]URL: `https://www.affymetrix.com/analysis`
[4]URL: `https://www.illumina.com/`
[5]URL: `http://www.agilent.com.br/`

arquivos de dados brutos consistem em amostras no formato CEL. Um arquivo CEL informa a posição espacial bidimensional de cada micropoço sobre o *Biochip* e a intensidade de luz nesta posição, mas não contém os nomes de sonda. Para isso, torna-se necessário um arquivo adicional chamado CDF, que varia conforme o GPL da amostra e é oferecido pelo servidor do *Bioconductor*. O CDF é um arquivo de anotação a nível de sonda (*probe-level annotation*) carregado como um pacote no R, e sua função é associar cada posição X e Y informada no CEL com seu respectivo identificador de sonda, bem como a iluminação captada.

A Illumina é uma fabricante presente em diversas outras áreas na aplicação de técnicas de biologia molecular, incluindo o sequenciamento de genomas (DU; KIBBE; LIN, 2007). Suas plataformas de micrarranjo, porém, são as mais heterogêneas dentre as três empresas. Em um mesmo GPL, é possível observar dados brutos de GSEs com formatos de arquivo completamente distintos, o que torna esta a fabricante com as plataformas mais complexas de serem analisadas no trio. A funcionalidade dos pacotes do R para analisar suas plataformas também é escassa e mais suscetível a erros em comparação com a das fabricantes concorrentes.

A Agilent Technologies, fundada como uma ramificação da Hewlett-Packard (HP), tem forte atuação no setor de química analítica. Seus dados brutos são padronizados e costumam incluir anotações de sonda nos próprios arquivos de amostra (WOLBER et al., 2006). Dependendo da plataforma, um arquivo GPL de sufixo "*old annotations*"contendo anotações antigas das sondas ou fotografias do Biochip em formato TIFF também podem estar disponíveis.

2.3 Passos de uma análise de Microarranjo

Sabendo os tipos de arquivos utilizados, um fluxo lógico de diferentes etapas são seguidas para analisar dados de microarranjo. Para que a análise tenha sucesso, os seguintes itens devem ser satisfeitos nessa ordem:

1. Minerar os dados a partir dos bancos de dados disponíveis;
2. Ler os dados minerados;
3. Se aplicável, tratar as amostras estatisticamente, incluindo correção de fundo e normalização;
4. Verificar a qualidade dos arranjos;
5. Se presentes, remover amostras com artefatos;
6. Comparar diferença de expressão entre grupos amostrais;
7. Filtrar resultados estatisticamente significantes;
8. Visualizar resultados e atribuir a estes um contexto biológico.

O primeiro item ocorre na etapa de **obtenção de dados**. O segundo e o terceiro item formam uma etapa definida como **pré-análise**, onde cada tipo de arquivo é lido, processado, unificado e convertido em um formato padrão para qualquer arranjo. Tendo estes dados formatados adequadamente, é na etapa de **análise de qualidade** que o quarto e o

quinto item são realizados, embora este último possa ser apenas a exclusão das amostras problemáticas durante a seleção na etapa seguinte. Avaliar a qualidade pode ser um passo opcional, porém a falta desta etapa pode acarretar em resultados incongruentes e errôneos caso artefatos estiverem presentes no arranjo. Após a análise de qualidade, o sexto item é definido pela **análise de expressão diferencial**, onde amostras são selecionadas para os grupos experimento e controle e estes são comparados entre si. Por último, os itens sete e oito se resumem em avaliar e apresentar os **resultados**, o que também requer que as anotações estejam definidas para as sondas no intuito de acessar seu contexto biológico. Todas essas etapas estão resumidas na Figura 2.2 e serão discutidas com maior profundidade a seguir.

Figura 2.2: Fluxograma representando a sequência de procedimentos mais comumente utilizada pelos usuários durante uma análise de transcriptoma.

2.3.1 Obtenção dos dados

Como sugerido em Componentes do GEO (Seção 2.2.2), onde obter os dados fica a critério do usuário, sendo o GEO a maior fonte de dados de microarranjo nos tempos atuais. O domínio *GEO Datasets*[6] é um dos locais recomendados para buscas de dados de transcriptoma, pois pode ser feito textualmente caso o usuário esteja a procura da descrição de um estudo ou por um assunto específico. Os resultados da busca apresentarão registros GSE e GSM, e dentro de cada registro há *links* para baixar os arquivos em formato padrão GEO. No mesmo registro, o usuário também pode fazer o *download* dos dados brutos, contanto que estes estejam disponíveis na tabela *Supplementary file* da mesma página.

No R, existe um pacote chamado *GEOquery* (DAVIS S, 2007), o realiza a obtenção dos arquivos do GEO utilizando apenas o código de registro do arranjo como argumento de função. Arquivos de anotação em nível de sonda, como CDF da Affymetrix, também podem ser obtidos com o pacote *BiocInstaller* (DAN TENENBAUM, 2018). O *download*

[6]URL: `https://www.ncbi.nlm.nih.gov/gds`

do GPL com as anotações para cada sonda pode ser feito neste passo, mas sua inclusão nos arranjos pode se efetuar em qualquer etapa entre a pré-análise até a expressão difencial.

2.3.2 Pré-análise

A leitura dos dados baixados pode ser feita por um programa oferecido pela fabricante da plataforma ou, alternativamente, utilizando o R. Se for uma plataforma da Affymetrix, o programa só aceitará arquivos CEL, e a anotação em CDF será baixada automaticamente quando ausente. Os softwares de outras fabricantes se comportarão de forma análoga.

Já no R, existem pacotes para cada plataforma ou fabricante. Para Affymetrix, usa-se o *affy* (GAUTIER et al., 2004) ou o *oligo* (CARVALHO; IRIZARRY, 2010). Para Illumina, existem os pacotes *lumi* (DU; KIBBE; LIN, 2008), *illuminaio* (SMITH et al., 2013) e *beadarray* (DUNNING et al., 2007). Para Agilent, há o pacote *agilp* (CHAIN, 2012). Outra alternativa para as plataformas do Illumina e da Agilent é o pacote *limma* (SMYTH, 2005), que também possui funções de leitura para os dados destes e de outros fabricantes.

Após os dados terem sido lidos, caso o usuário esteja utilizando dados brutos, um passo de tratamento de arranjos deve ser realizado. Esse passo existe para tentar corrigir erros que podem ter ocorrido durante a etapa experimental. Para todas plataformas, existem dois tipos de erro: iluminação de fundo e diferença escalar. A iluminação de fundo pode ocorrer quando a luz emitida por um poço é sobreposta com outros micropoços vizinhos ou quando há algum erro na aparelhagem que cause desvios na absorção de luz, e o tratamento utilizado é chamado **correção de fundo** (EDWARDS, 2003). Diferenças escalares, por outro lado, são variações na escala dos valores de uma amostra para outra quando estes deveriam estar no mesmo nível. Tais diferenças podem ser causadas por diversos fatores como contaminação ou mal manuseio do aparelho, e o tratamento utilizado neste caso é chamado **normalização** (STAFFORD, 2008). Se o microarranjo for da Affymetrix, há também uma correção chamada *Probe Perfect Match*, pois as plataformas desta fabricante também utilizam sondas chamadas *Mismatch*, que diferem do mRNA complementar em apenas uma base, sendo critério do usuário incluí-las ou não entre os valores de expressão (LAIRD, 2010).

2.3.3 Análise de qualidade

Os softwares de cada plataforma não costumam disponibilizar uma análise de qualidade completa, sendo esta etapa destinada a usuários do R. Dentre os pacotes utilizados para avaliar a qualidade das séries, o mais completo é o *arrayQualityMetrics* (KAUFFMANN; GENTLEMAN; HUBER, 2008a).

Neste passo, três principais avaliações são feitas com respeito aos valores de expressão das amostras. A primeira avaliação compara a distância entre os valores de cada amostra com as demais amostras. A segunda avaliação verifica a distribuição dos valores de expressão entre as sondas para uma mesma amostra. A terceira e última avaliação compara a variação das mesmas sondas entre diferentes amostras. Se em qualquer uma das três avaliações houver variações bruscas, a amostra será atribuída como "discrepante"(*outlier*) para a etapa testada. Valores discrepantes podem indicar defeitos durante o procedimento experimental do microarranjo, mas também são esperados em amostras onde há muita variabilidade genética ou epigenética envolvida, como certos casos envolvendo câncer, por exemplo.

Se ocorrer o caso de uma amostra discrepante não ser esperada, a melhor escolha poderia ser remover a coluna desta amostra ou apenas não levá-la em conta durante análises posteriores. Retirar a coluna também permite que o usuário realize a análise novamente, dessa vez sem a amostra problemática.

2.3.4 Análise de expressão diferencial

O principal objetivo do microarranjo é comparar o perfil de expressão de uma amostra biológica em um estado em relação a outro estado. Em ensaios biomédicos, por exemplo, geralmente amostras de tecido em um estado saudável são incluídas no grupo "controle", enquanto que aqueles tecidos em estado patológico são inclusos no grupo "experimento". Matematicamente, os grupos são idênticos, mas é mais aconselhável que essa lógica de distribuição de amostras se mantenha por questões de interpretação dos resultados.

Em ambos os grupos, as sondas têm seus valores resumidos a um único número para cada sonda (EFRON et al., 2001). Esse resumo se dá por diferentes meios, como as médias dos valores, por exemplo. Já o *limma*, que é o principal pacote do R utilizado para essa etapa de análise, utiliza um ajuste envolvendo regressão linear para chegar a um resmo dos valores. No final, um valor experimento (ValEXP) e um valor controle (ValCTRL) é obtido para cada sonda, e um valor denominado logaritmo de base 2 de *Fold-Change* ($\log_2 FC$) é calculado utilizando esses valores através da seguinte fórmula:

$$\log_2 FC = \log_2 \left(\frac{ValEXP}{ValCTRL} \right)$$

O $\log_2 FC$ de uma sonda representa a proporção do valor de expressão entre o grupo experimental para o controle, sendo positivo se a sonda estiver mais expressa na condição experimental e negativo se este for o caso na condição controle. Genes com maior expressão no grupo experimental são denominados **super-expressos**, enquanto que os mais expressos no grupo controle são chamados **sub-expressos**. Nota-se que ValEXP ou ValCTRL não pode ter valor zero, pois não existe logaritmo de zero. Alguns softwares, como os da Affymetrix apresentam os valores em *Fold-Change* (sem o logaritmo), resultando em valores somente positivos — contudo, mesmo nestes casos o ValCTRL é o denominador da divisão e não pode ser zero, caso contrário retornará valores infinitos (IRIZARRY et al., 2003).

Desde que pelo menos três amostras tenham sido utilizadas para a análise de expressão diferencial, um valor-p pode ser calculado para verificar a plausibilidade do valor de $\log_2 FC$. Em certos casos, o valor-p não é o suficiente para se ter certeza quanto ao resultado obtido, e existem algoritmos como o *False Discovery Rate* (FDR) (BENJAMINI; HOCHBERG, 1995) que ajustam o valor-p dos resultados. Enquanto o FDR é o único algoritmo entre muitos dos softwares das fabricantes de transcriptoma, o R disponibiliza diferentes opções de correção de valor-p além deste. Uma opção diferente pode ser interessante para quando o FDR limita demasiadamente o número de resultados significantes ao passo que sua falta causa um excesso de genes diferencialmente expressos.

2.3.5 Filtragem e Visualização dos resultados

Em geral, o que se busca de resultados em uma análise de micrarranjo são genes **diferencialmente expressos** (DE) entre condições experimento e controle. Os genes que entram nessa categoria devem satisfazer algum critério de filtragem definido pelo usuário

— por exemplo, todos os genes com $\log_2 FC$ maior que $1,00$ ou inferior a $-1,00$ e contendo um valor-p corrigido menor que $0,05$. Isso porque há pouco sentido em considerar todas as dezenas de milhares de genes como resultados significativos. Tanto o R quanto os programas das fabricantes possuem um sistema de filtragem de resultados (IRIZARRY et al., 2003), ainda que o R, por ser uma linguagem de programação, permita personalizar melhor estes filtros.

Além disso, é conveniente ter os resultados resumidos em diagramas informativos. Gráficos de distribuição são algo em comum entre os softwares de microarranjo e o R, mas apenas o R apresenta opções ilimitadas de visualização. Um diagrama característico do R é o gráfico de vulcão (*Volcano Plot*), apresentando o $\log_2 FC$ no eixo horizontal e o valor-p corrigido no eixo vertical para cada sonda (CUI; CHURCHILL, 2003). Este tipo de visualização é importante para avaliar a significância dos resultados, onde uma alta ocorrência de pontos localizados na parte superior do vulcão pode indicar baixa certeza – ou alta, dependendo do referencial e escala utilizados – em relação aos resultados.

Fundamentalmente, este capítulo resumiu os passos de uma análise padrão de microarranjo. Entretanto, existem limitações em cada uma dessas etapas, que serão o alicerce da motivação para produzir o presente trabalho.

3 MOTIVAÇÕES E OBJETIVOS

3.1 Limitações das análises

Uma dificuldade crucial em procedimentos de análise no ramo da bioinformática é a falta de padronização das técnicas aplicadas. No caso do microarranjo, desafios relacionados à variabilidade de arquivos são vistos logo nas etapas iniciais de pré-análise, pois ainda não há programa capaz de processar todos os formatos possíveis. Mesmo com conhecimento de programação, o usuário deve lidar com diversos pacotes no R que são suscetíveis a erros e nem sempre adequadamente documentados, e outras linguagens semelhantes não apresentam um esforço equivalente para manipular dados com expressão gênica. Certos programas executáveis costumam satisfazer os critérios visuais e de usabilidade para uma ferramenta intuitiva de transcriptoma, mas não apresentam tantas opções de leitura e manipulação estatística dos dados quanto uma linguagem de programação, o que é fundamental para propósitos científicos. Etapas como análise de qualidade, por exemplo, estão ausentes em muitos destes softwares, apesar de clara a sua relevância.

Mesmo no R, existem limitações. Os graus de conhecimento variam entre os adeptos à linguagem, sendo que os programadores mais avançados compõem um nicho particularmente limitado. Isso implica que a maior parcela dos usuários do R depende de roteiros pré-definidos para cada plataforma de microarranjo, podendo possuir dificuldades em adaptar sua análise a algoritmos com maior complexidade quando indispensável. Além da questão técnica, os gráficos obtidos para os resultados são estáticos e não possuem qualquer forma de interação direta ou confortável. Para cada ocasião onde se deseja explorar dados e resultados, um conjunto de códigos em linha de comando previamente construído pelo usuário deve ser satisfeito.

Deste modo, existe a necessidade de um programa que reúna o melhor dos dois mundos, apresentando a interface visual, intuitiva e interativa dos programas de microarranjo e ao mesmo tempo a flexibilidade do R em manipular dados.

3.2 Objetivos específicos

O **GEAP** foi desenvolvido para buscar amenizar as limitações descritas acima. O lema do GEAP é claro e direto: ter uma interface fácil, flexível e pronta para usar. Esse tipo de interface aumenta a eficiência do pesquisador para obtenção de resultados, e da mesma forma para o progresso científico.

Mais especificamente, a ferramenta **GEAP** foi desenvolvida após cumprir os seguintes objetivos, na ordem:

1. Reunir e compreender o maior número de formatos de arquivos de microarranjo presentes no GEO em tempo hábil;

2. Explorar os métodos de tratamento estatístico empregados sobre estes arquivos, quando aplicável;

3. Desenvolver algoritmos que permitam identificar, ler e processar os tipos de arquivos encontrados, levando em conta diversos casos de uso e colocando-os em um formato comum e padrão;

4. Montar funções que executem análises de qualidade e expressão diferencial utilizando os dados processados;

5. Construir uma interface gráfica que permita ao usuário realizar todas essas funções de forma visual e intuitiva, isentando-se da necessidade de entender programação;

6. Enriquecer a interface com funcionalidades que auxiliem as análises, seja aumentando sua eficiência, introduzindo maior flexibilidade de manipular os dados, facilitando a obtenção dos resultados ou aprimorando o conforto de seu acesso.

Em outras palavras, foram estudados padrões de arquivos e fluxos lógicos de como analisar transcriptomas no R. Após, estes conceitos foram trazidos para uma interface gráfica, procurando tornar as análises o mais simples e amigável possível, mas também oferecendo espaço para mais opções avançadas quando houver necessidade.

Parte II

O programa GEAP

4 MECANISMO E INTERFACE DE USUÁRIO

Como constatado no último capítulo, tanto os softwares disponíveis para analisar microarranjo quanto a linguagem R têm suas próprias limitações. O **GEAP** foi desenvolvido para buscar amenizar essas limitações, concatenando os benefícios de um programa executável com as funcionalidades do R. Neste capítulo, será descrita a interface do **GEAP**, bem como os procedimentos de análise transcriptômica empregados no programa.

4.1 Especificações de desenvolvimento

O **GEAP** foi programado em duas linguagens: C# (versão 7) e R (versão 3.3.2). Da mesma forma, o tempo de execução da ferramenta é mediado por ambas as linguagens em duas camadas: uma interface gráfica, promovida pelo programa compilado em C#, e uma linha de comando do R, sendo executada em plano de fundo a partir de sua versão portabilizada. Enquanto uma janela é apresentada para exploração da ferramenta, existem pontos específicos onde comandos são enviados e recebidos do R conforme necessário, havendo trocas de comunicação em tempo real entre ambas as linguagens.

Todas as funções associadas a visualização são realizadas com o algoritmo compilado em C#. Além disso, certos métodos desenvolvidos em C# substituíram funções originalmente cumpridas pelo R, como o *download* dos dados, a verificação de arquivos, a seleção de amostras, a construção de gráficos e a filtragem dos resultados. Essa substituição foi vantajosa em termos de performance, pois o C# é uma linguagem compilada e permite múltiplas unidades de processamento simultaneamente — em constraste com o R, que é uma linguagem interpretada, onde um comando por vez é compilado e executado. Por outro lado, o R foi utilizado durante a fase de execução dos algoritmos para leitura e processamento de arquivos de série e amostra, para controle de qualidade de arranjos e para expressão diferencial. Neste caso, os pacotes utilizados nesses procedimentos possuem suporte consolidado e estão sendo atualizados eventualmente, tornando vantajoso mantê-los como parte da ferramenta principal. Coerentemente, os pacotes do R que apresentaram erros não foram selecionados para montagem das funções.

4.2 Inicialização

Até a data desta publicação, o **GEAP** só está disponível para o sistema operacional Windows. Não é necessário instalar para se ter acesso à ferramenta, apenas extrair os arquivos em qualquer pasta de interesse. Para o programa ser executado no computador, ele deve seguir os requisitos mínimos do sistema, que são:

- Sistema operacional Windows 7, 8 ou 10;
- Microsoft .NET Framework 4.5.2 ou superior, caso não esteja instalado por padrão;
- Memória RAM de 4GB ou mais;
- Processador de 2.4 GHz (recomenda-se uma capacidade maior de rotação ou múltiplos núcleos).

Estes requisitos podem aumentar conforme o volume dos dados utilizados. Maiores quantidades de dados inevitavelmente exigirão maior custo computacional e maior alocação de memória RAM.

Durante a inicialização, um logotipo do **GEAP** surge e permanece na tela enquanto os componentes do programa são carregados. Como constatado na seção anterior, um ambiente do R também é executado em plano de fundo e persistirá enquanto o **GEAP** estiver aberto. Após a inicialização do programa, surge uma janela com o menu principal (Figura 4.1) composto por seis botões, onde cada botão conduz o usuário a uma seção. Os botões são:

- Iniciar Análise: Leva o usuário diretamente para a etapa de obtenção dos dados e pré-análise;
- Biblioteca: Redireciona à sessão de biblioteca, a qual armazena dados de arranjos que o usuário utilizou anteriormente;
- Projetos: Redireciona à sessão dos projetos, onde o usuário pode armazenar informações de análises feitas anteriormente e consultá-las em qualquer momento;
- Configurações avançadas: Permite o usuário ajustar opções definidas globalmente no programa;
- Utilitários: Ferramentas que não necessariamente fazem parte da análise principal, mas que podem auxiliar as análises de microarranjo;
- Documentação: Apresenta uma descrição organizada de cada seção do programa, assim como referências bibliográficas de cada algoritmo aplicado ou pacote utilizado.

4.3 Pré-análise

Ao pressionar o botão **Iniciar Análise**, o usuário é redirecionado para um sub-menu contendo três botões (Figura 4.2). Cada botão corresponde a uma forma de como o usuário deseja carregar os dados. Um único arquivo de série, múltiplos arquivos de amostra ou a criação de uma tabela customizada são as três formas disponíveis para carregar arquivos de transcriptoma.

Além dos três principais botões na tela principal, também nota-se uma barra no canto inferior contendo botões com os nomes das páginas navegadas. Os nomes de cada botão

Figura 4.1: Tela inicial do **GEAP**.

são resumidos por questões de espaço, mas o nome completo pode ser verificado ao posicionar o mouse sobre o botão. Cada botão retorna o usuário até a seção navegada conforme o nome explicitado. Deve-se ter cautela, porém, que o contrário não pode ser feito — isto é, não é permitido ir diretamente à tela de onde retornou por através deste menu, certas vezes tendo de repetir os passos realizados para a ida. Além da barra mencionada, há uma outra barra mais abaixo com um ícone de "i"dentro de um círculo, que apresenta um texto informativo ao posicionar o mouse sobre certos elementos da janela.

Figura 4.2: Modalidades de obtenção de dados para a pré-análise.

4.3.1 Série Completa (GSE)

Nesse menu (Figura 4.3), o usuário obterá os dados levando em conta a série inteira de um experimento — isto é, o conjunto completo com todas as amostras.

Figura 4.3: Tela de pré-análise para obtenção de série completa.

Para qualquer tipo de dados, existem três métodos de obtenção: *Download* da *Web*, arquivo existente no computador ou arquivo da biblioteca. No primeiro caso, o usuário pode fazer o Download de uma série a partir de seu código no NCBI, contanto que seja um código válido para microarranjo e os dados estejam disponíveis. No segundo caso, pode-se carregar um único arquivo existente no computador. No terceiro caso, um arquivo salvo na biblioteca é diretamente carregado a partir desta. Em termos de velocidade, esta é a opção mais rápida, pois os dados já estão devidamente organizados para leitura pelo programa.

Se a opção de dados brutos não estiver marcada, os arquivos que serão obtidos estarão no formato TXT (.txt), SOFT (.soft) ou em suas variantes comprimidas em GZip (.txt.gz e .soft.gz, respectivamente). Marcando a opção de obter dados brutos, os tipos TAR (.tar) comum ou comprimido (.tar.gz) também passam a ser aceitos para leitura. A extensão TAR é a utilizada tanto pelo GEO para comprimir dados brutos quanto pelo *Bioconductor* para comprimir os pacotes do R.

As modalidades *Download* da *Web*, arquivo existente e arquivo da biblioteca também são válidas para obter a plataforma GPL e os dados brutos. Quando os dados brutos são de uma plataforma da Affymetrix, os arquivos CDF são automaticamente obtidos junto com os demais — a menos que a opção de *Download* pela *Web* não tenha sido marcada, caso o usuário queira carregar o arquivo CDF localmente do computador. Neste caso, uma janela de diálogo solicitando o endereço do arquivo é aberta para o usuário indicar o arquivo em questão.

Neste exemplo, analisaremos a série de código GSE2600, derivada de um estudo sobre infecção de células de linhagem NB4 por *Anaplasma phagocytophilum*. Esta série possui seis amostras, sendo três de células NB4 não-infectadas (GSM49939, GSM49940 e GSM49941) e outras três de células NB4 infectadas (GSM49942, GSM49943, GSM49944). A escolha para este exemplo foi arbitrária, e os mesmos passos funcionarão com outras

séries como GSE39252 da Affymetrix, GSE80080 da Illumina ou GSE51612 da Agilent. Tendo selecionado os parâmetros corretos, inicia-se a pré-análise pelo botão “Carregar”, agora em cor verde. Se o usuário optou por obter os dados pela Web, uma janela de diálogo aparecerá para o usuário confirmar se foi o arranjo correto (Figura 4.4).

Se a opção de dados brutos não tivesse sido selecionada, o programa carregaria a matriz de texto com valores tratados estatisticamente pelo autor do experimento, e o procedimento passaria diretamente para a próxima seção.

Neste exemplo, os dados brutos foram optados a serem obtidos pela Web. Após a confirmação da série desejada, uma janela de carregamento é aberta (Figura 4.5).

Figura 4.4: Diálogo de confirmação se o código GSE inserido corresponde à serie desejada.

Figura 4.5: Janela de carregamento. A barra azul representa o progresso de Download dos dados, enquanto a barra verde indica o progresso do processamento dos dados.

Em plano de fundo, o programa está fazendo o Download dos dados. Após baixados,

todos os arquivos comprimidos são extraídos e cada arquivo é identificado pelo programa. Todos os arquivos são verificados por seu formato e compatibilidade, e apenas os compatíveis são agrupados por nomes de sonda. Se houverem arquivos corrompidos, o erro será avisado ao usuário, porém não serão levados em conta na pré-análise. Um resumo dos arquivos que o programa suporta está descrito na Tabela 4.1.

Tabela 4.1: Tipos de dados utilizados. Até a versão BETA concluída, apenas os arquivos CEL, SOFT e texto são suportados.

Nome	Extensão	Tipo	Fornecedor
Cell Intensity File	CEL (.cel)	Amostra	Affymetrix
Agilent Raw Data	Texto (.txt)	Amostra	Agilent
Arquivo de Imagem	TIFF (.tif)	Amostra	Agilent Illumina
GenePix Results	GPR (.gpr)	Amostra	GenePix
Bead Level Data	Texto (.txt)	Amostra	Illumina
Pair Report	PAIR (.pair)	Amostra	NimbleGen
Chip Definition File Package	Pacote R (.tar)	Anotação	Bioconductor
GenePix Array List	GAL (.gal)	Anotação	GenePix
Simple Omnibus Format in Text	SOFT (.soft) Texto (.txt)	Amostra, Plataforma ou Série	GEO
Arquivos Suplementares	Texto (.txt)	Complemento	GEO
Manifest File	BGX (.bgx) Texto (.txt)	Anotação	Illumina
Matriz customizada	Qualquer	Série ou Anotação	Usuário
Matriz de Série	Texto (.txt)	Série	GEO
Intensity Data	IDAT (.idat)	Série	Illumina

Com todos os arquivos verificados, uma janela aparecerá com os grupos amostrais disponíveis, bem como as anotações carregadas, se houver (Figura 4.6). Às vezes, um GSE pode ser composto em múltiplas séries ou ser associado a múltiplos GPLs, e esta janela se torna útil nestes casos por listar todas as de plataforma em seu canto superior.

Nota: Se um GSE for obtido localmente e os dados brutos forem optados, é possível que não se tenha informações a respeito de seu GPL, portanto o programa irá perguntar o código do GPL se este tiver que ser obtido pela Web.

4.3.2 Amostras (GSM)

O usuário também pode optar por carregar cada amostra de forma separada, ou mesmo carregar múltiplos arquivos de série. A lógica por trás dessa modalidade (Figura 4.7), é a mesma da demonstrada na série completa: os dados podem ser obtidos pela Web, por arquivos locais ou pela biblioteca. A diferença é que apenas amostras específicas são obtidas, e múltiplos arquivos de amostra são levados em conta em vez de um único arquivo de série. Outra diferença é que, por serem múltiplos arquivos envolvidos, pode-se obter os arquivos por mais de uma forma (i.e. pela Web, localmente e pela biblioteca ao

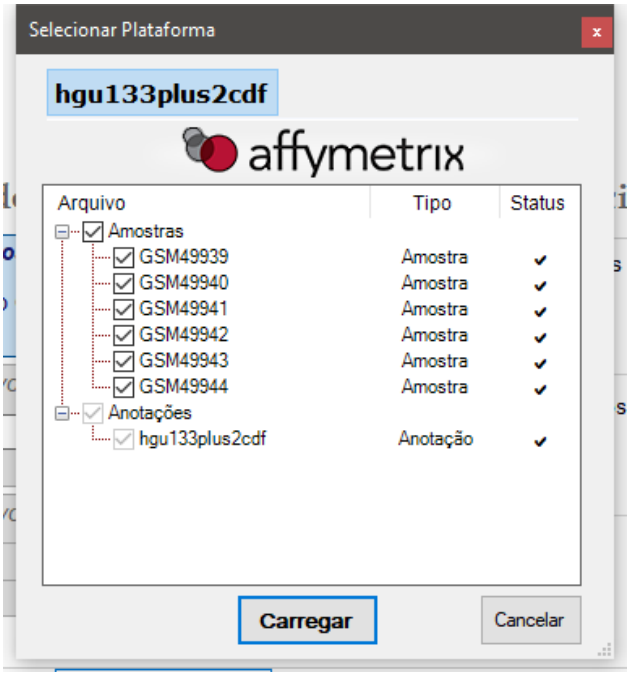

Figura 4.6: Amostras carregadas juntamente com a plataforma e o pacote de anotação em nível de sonda. O nome do grupo está de acordo com o pacote da plataforma, informado no cabeçalho da janela.

mesmo tempo). Os arquivos serão agrupados de acordo com suas plataformas ou sondas específicas, a mesma janela da Figura 4.6 aparecerá para o usuário selecionar o grupo desejado.

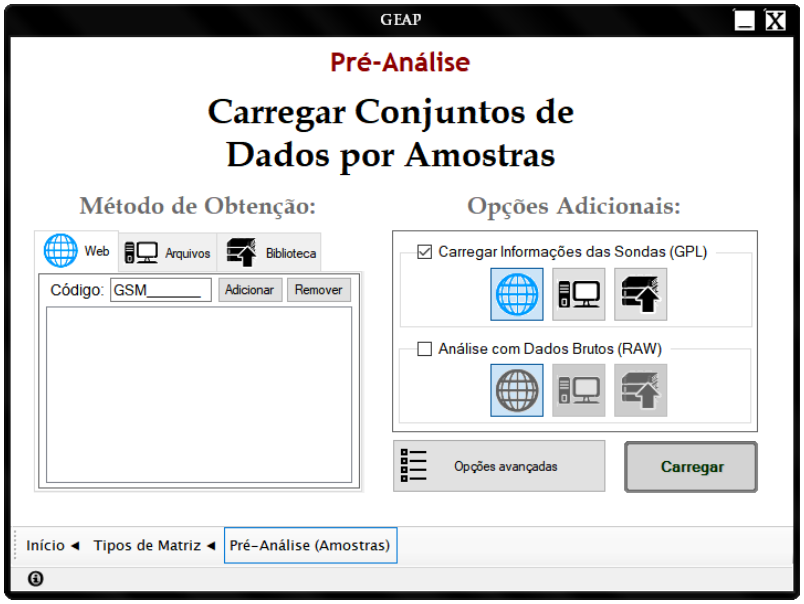

Figura 4.7: Tela de pré-análise para obtenção de amostras separadas.

4.3.3 Tabela Customizada

Se os dados do microarranjo não existirem no NCBI, não houver suporte do arquivo de arranjo no programa ou o usuário simplesmente tenha interesse em inserir todos os

dados manualmente, uma tabela customizada torna-se uma opção interessante. Nessa modalidade, o usuário pode criar, desde o zero, uma tabela customizada com seus valores de expressão e atributos de sonda próprios (Figura 4.8).

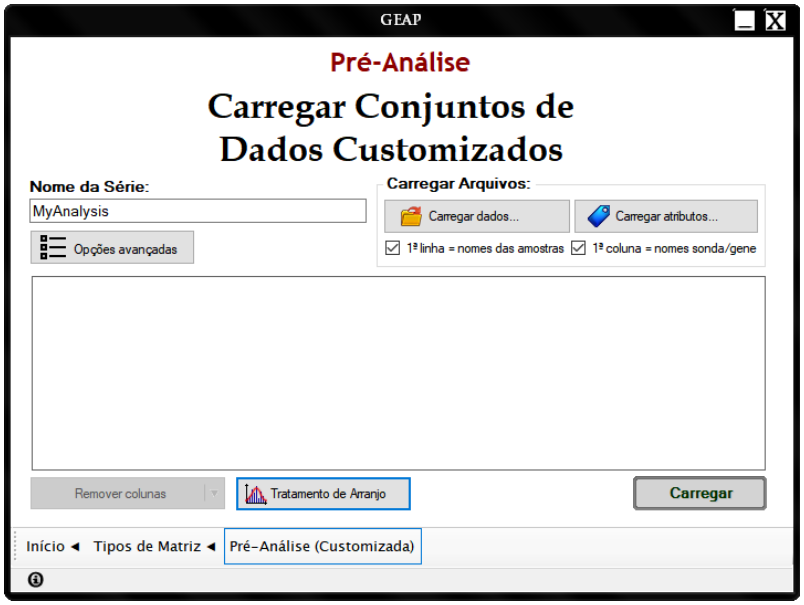

Figura 4.8: Janela de carregamento. A barra azul representa o progresso de Download dos dados, enquanto a barra verde indica o progresso do processamento dos dados.

No painel, o usuário primeiro indica um nome para sua nova série. Após, carrega uma tabela de valores de expressão ao clicar no botão **Carregar dados...**. A tabela de entrada é um arquivo de texto com um número fixo de linhas e colunas. Se a tabela possui um cabeçalho com os nomes de coluna, marcar a opção "1ª linha = nomes das amostras". Se a primeira coluna estiver representando os nomes de sonda, marcar "1ª coluna = nomes sonda/gene". Não são permitidas sondas repetidas. São necessárias pelo menos duas colunas de valores numéricos de expressão para a pré-análise.

A opção **Carregar atributos...** tem a mesma função que a Carregar dados, exceto que leva em conta todas as colunas como atributos de sonda, sendo numéricas ou não, e não requer que haja o mesmo número de linhas já presente na tabela. No botão de tratamento de arranjo, pode-se definir o tratamento estatístico que será feito nos valores de expressão durante a pré-análise. As especificações de cada parâmetro de tratamento serão discutidas na seção seguinte

Após preenchida, a tabela customizada se parecerá com a Figura 4.9.

4.4 Tratamento de Amostras

Se os arranjos selecionados pelo usuário forem dados brutos ou uma tabela customizada, torna-se necessário tratar os dados antes de utilizá-los em análises posteriores. No GEAP, para cada plataforma, existem diferentes métodos de tratamento com diversos parâmetros, cada um correspondendo a um algoritmo no R. As plataformas e seus respectivos métodos são:

- Affymetrix
  - Método Expresso — Pacote *affy* (GAUTIER et al., 2004):

Figura 4.9: Tabela customizada carregada com valores de expressão (em verde) e atributos (em roxo).

    - Correção de fundo com as opções Affymetrix Microarray Suite (MAS), Análise Robutos de Múltiplos Arranjos (RMA) ou nenhuma;
    - Normalização com as opções Quantis (*Quantiles*), Loess, Spline Cúbico (*QSpline*) e Conjunto invariante (*Invariante set*);
    - Correção de *Match* de sondas com as opções Apenar Perfect-Match (*PMOnly*), Subtrair sinais Mismatch (*SubtractMM*) ou Affymetrix Microarray Suite (*MAS*);
    - Forma de apresentação dos valores finais, sendo as opções Diferença média de expressão (*AvgDiff*), Remoção de Outliers por Li & Wong (LiWong) e Polimento de medianas de Tukey (*MedianPolish*).
  - Método GC-RMA — Pacote *gcrma* (WU et al., 2012):
    - Correção ótica de fundo;
    - Normalização de valores pelo método dos Quantis (*Quantiles*).
  - Método Plier — Pacote *plier* (INC; MILLER, 2018):
    - Normalização de valores pelo método dos Quantis (*Quantiles*);
    - Correção de pares de sondas com as opções Apenar Perfect-Match (*PMOnly*), União Perfect-Match com Mismatch (*Together*), Separação Perfect-Match de Mismatch (*SeparateMM*) ou Somente Mismatch (*MMOnly*).
- Agilent — Pacote *limma* (SMYTH, 2005):
  - Correção de fundo
    - Modelo de Convolução Normal+Exponencial (*normexp*)
    - Subtrair com o Fundo (*subtract*)
    - Diferença do Fundo Maior ou Igual a 0.5 (*half*)
    - Diferença do Fundo Corrigida em Mínimo Positivo (*minimum*)
    - Mínimo de Fundo Entre Vizinhos (*movingmin*)

        * Edwards (*edwards*)
        * Nenhuma correção
    - Se a Correção de fundo for NormExp, este algoritmo possui os sub-métodos:
        * Saddle
        * MLE
        * RMA
        * RMA75
    - Normalização:
        * Escala (*scale*)
        * Quantis (*quantile*)
        * Regressão Polinomial Local Cíclica (*cyclicloess*)
        * Nenhuma
    - Se a Correção de fundo for CyclicLoess, o algoritmo possui os sub-métodos:
        * Rápido (*fast*)
        * Affymetrix (*affy*)
        * Pares (*pairs*)

- Illumina e tabelas customizadas — Pacote *limma*:
    - Normalização
        * Quantis (*quantile*)
        * Estabilização de Variância (*vsn*)
        * Classificação Invariante (*rankInvariant*)
        * Splines Quadráticos (*qspline*)
        * Splines Robustos (*rsn*)
        * Diferença pela Mediana (*median*)
        * Nenhuma
    - Tranformação de Matriz
        * Logaritmo de Base 2 (*log2*)
        * Correção Normal+Exponencial e Quantis (*neqc*)
        * Estabilização de Variância (*vst*)
        * Nenhuma

Neste exemplo, como o arranjo obtido pertence a uma plataforma Affymetrix, os parâmetros de tratamento correspondem a esta fabricante (Figura 4.10).

4.5 Visão geral dos arranjos

Qualquer que seja a modalidade de pré-análise que foi concluída, todas as modalidades convergem a um formato comum de dados de arranjo, o qual é possível de ser processado pelas análises posteriores. Neste passo, uma janela com a visão geral do arranjo é demonstrada (Figura 4.11). Caso as informações das tabelas do GSE, GPL ou GSMs tenham sido disponíveis durante a obtenção dos dados, essas informações poderão

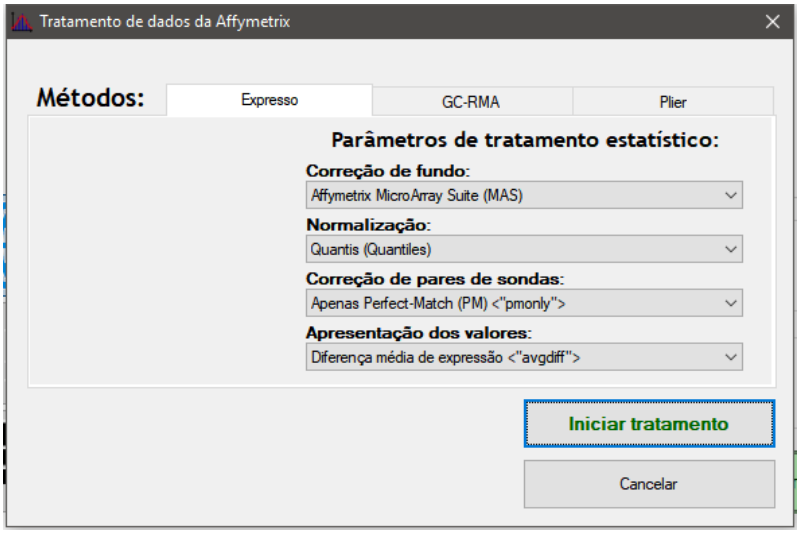

Figura 4.10: Janela de tratamento estatísticos para plataformas Affymetrix.

ser consultadas em sua própria aba (Figura 4.12). Se o usuário obteve as informações localmente sem nenhum acesso à internet e estes dados não possuem informações dos conjuntos de dados, as informações são omitidas na visão geral. Se for preferência do usuário, as tabelas de informações disponíveis podem ser exportadas, e o arranjo final também pode ser salvo na biblioteca.

Figura 4.11: Visão geral do arranjo carregado. À esquerda, uma forma resumida dos dados é apresentada ao usuário. À direita, localizam-se botões que conduzem o usuário a análises posteriores, onde uma ilustração ao lado indica o pacote do R utilizado para sua função.

A partir deste menu, o usuário pode prosseguir para a análise de qualidade ou pular diretamente para a análise de expressão diferencial.

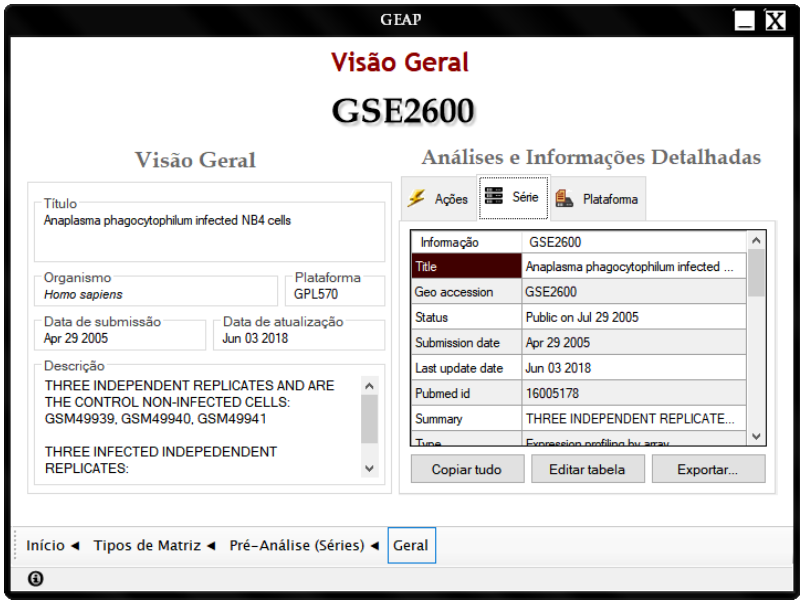

Figura 4.12: Tabela de informações da série carregada.

4.5.1 Análise de Qualidade

A análise de qualidade, como discutida anteriormente, é essencial para encontrar amostras que não são adequadas para serem incluídas nas comparações de expressão diferencial. Isso permite que o usuário elimine amostras que podem enviesar totalmente os valores, caso discrepâncias não sejam esperadas ou quando existe algum artefato entre as amostras. Na aba "ações", ao clicar no botão de iniciar análise no menu **Análise de Qualidade**, uma janela é aberta (Figura 4.13).

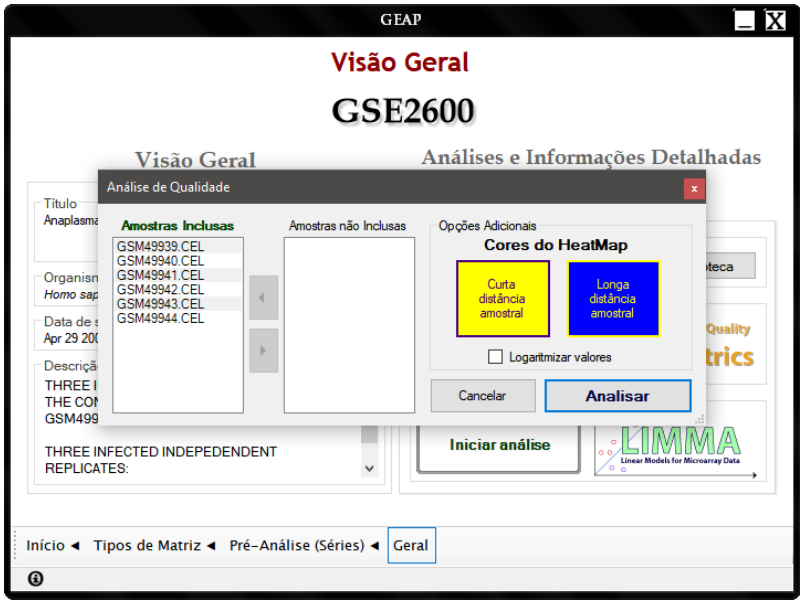

Figura 4.13: Caixa de diálogo para análise de qualidade.

Nela, o usuário seleciona quais amostras quer avaliar, bem como personalizar as cores dos gráficos que representam as distâncias dos valores amostrais. O padrão é amarelo para valores distantes e azul para valores mais próximos. Como ilustrado ao lado do botão,

durante essa análise, o algoritmo do pacote *arrayQualityMetrics* é aplicado (KAUFFMANN; GENTLEMAN; HUBER, 2008b). Após o processamento da análise, uma tela de resultado é demonstrada como na Figura 4.14. Se houver amostras com valores discrepantes em qualquer uma das etapas, um número é destacado para esta amostra na tabela resultante indicando em qual etapa ocorreu a discrepância. Neste exemplo, a amostra seis possui uma discrepância na segunda etapa de verificação (Figura 4.14). Ao trocar para a aba da seção dois, essa discrepância é representada na segunda imagem da esquerda para direita (Figura 4.15). Para visualizar esse gráfico mais de perto, clique sobre ele. Uma janela de visualização de imagem aparecerá (Figura 4.16), e o usuário poderá ajustar o *zoom* conforme desejado.

Com isso, torna-se mais fácil identificar a necessidade de eliminar ou não certas amostras da análise durante o passo da análise de expressão diferencial.

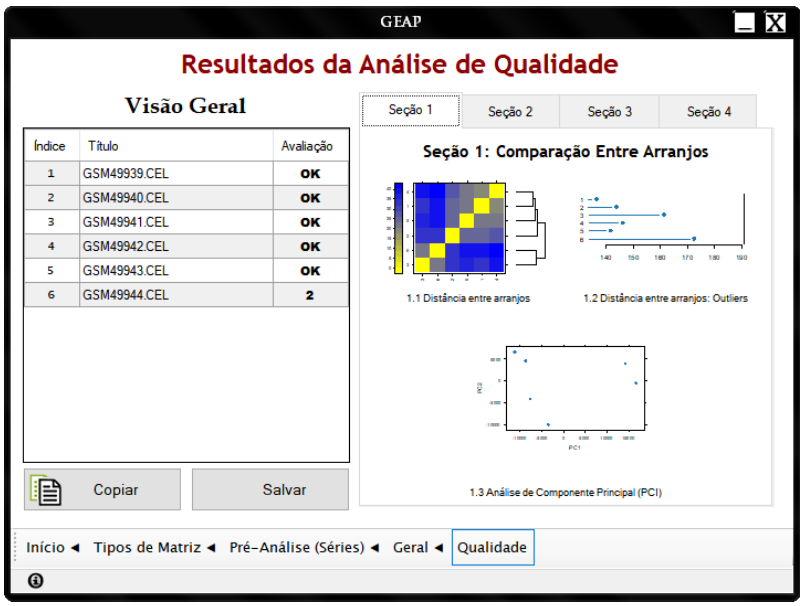

Figura 4.14: Resultados da análise de qualidade. Nota-se que a sexta amostra demostrou uma discrepância de valores na segunda etapa da análise.

4.6 Análise de Expressão Diferencial

Nesta etapa, amostras serão selecionadas para serem comparadas entre si por expressão diferencial, conforme discutido nas seções anteriores. No GEAP, são disponíveis três modalidades para realizar as comparações:

1. Experimento X Controle;
2. Entre Múltiplos Grupos
3. Sequencialmente Entre Diferentes Etapas

Embora, por princípio, todas as comparações sejam feitas entre grupo experimental e grupo controle, as modalidades segunda e a terceira permitem automatizar comparações feitas consecutivamente, o que muitas vezes é a realidade das análises de expressão diferencial. Isso porque muitas séries disponibilizam amostras em três ou mais condições

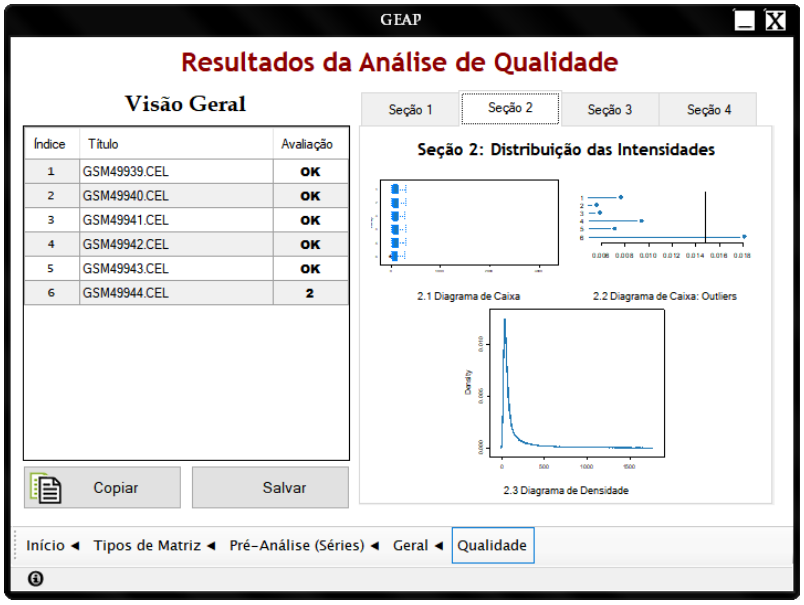

Figura 4.15: Segunda seção entre os resultados da análise de qualidade. O segundo gráfico (canto superior direito) indica a ocorrência de discrepância da sexta amostra.

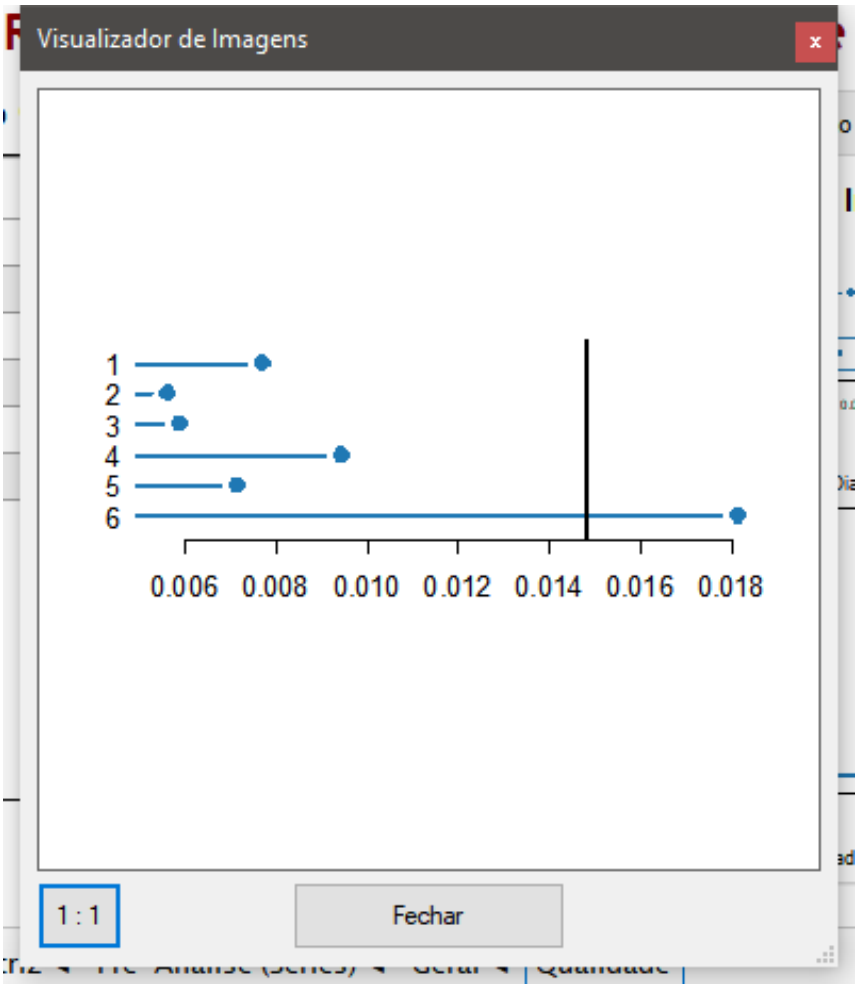

Figura 4.16: Diagrama demonstrando discrepância encontrada para a sexta amostra.

— por exemplo, quatro amostras de câncer hepático no estágio inicial, três no estágio avançado e cinco de tecido de fígado saudável[1]. Neste último exemplo, as amostras poderiam ser comparadas sequencialmente entre cada condição pelo terceiro modo de comparação, ou combinatorialmente entre todas as condições pela segunda modalidade comparativa.

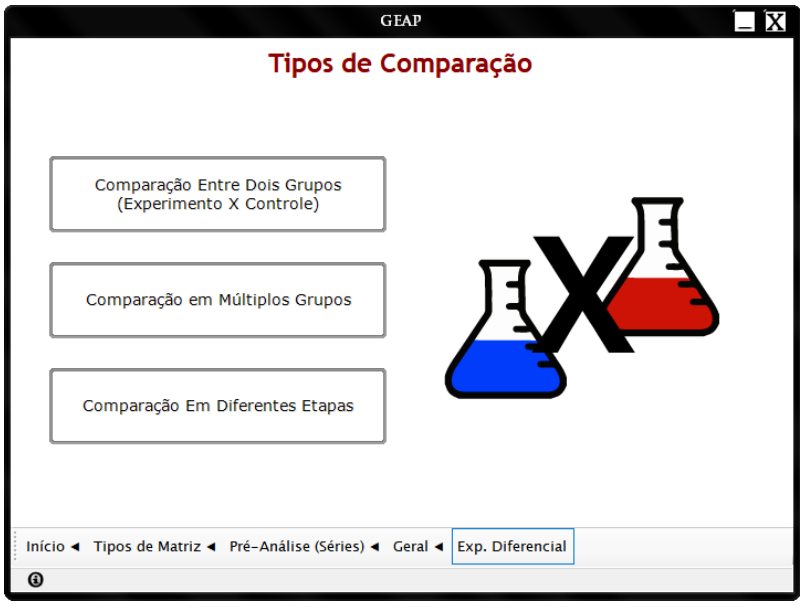

Figura 4.17: Tipos de análise de expressão diferencial.

4.6.1 Comparação Entre Dois Grupos (Experimento X Controle)

Essa é a modalidade básica da análise de expressão diferencial. Ao ser exibido menu (Figura 4.18), observa-se três colunas distintas: “Experimento”, “Amostras”e “Controle”. Inicialmente, apenas a coluna Amostras está preenchida, contendo todas as amostras processadas durante a pré-análise.

Para inserir cada amostra nos grupos experimental ou controle, basta o usuário selecionar suas amostras de interesse e clicar nos botões laterais à tabela. O botão ficará acesso quando um ou mais itens forem selecionados, e as setas estampadas em cada botão apontam a direção para qual lista a amostra selecionada será transferida. Botões com setas duplas executam a transferência de todos os itens de uma lista para outra. Neste exemplo, distribuiremos as amostras de câncer para o grupo experimental e as saudáveis para o grupo controle (Figura 4.19), não incluindo a sexta amostra por ter apresentado valores discrepantes anteriormente, pela análise de qualidade.

No canto lateral direito (Figuras 4.18 e 4.19), há três blocos de controle: Informações das amostras, anotações de sondas e opções avançadas. Se as informações das amostras foram obtidas durante as etapas iniciais, suas informações de título e código GSM são exibidas na janela correspondente a Informações das Amostras quando um item da coluna é selecionado. Se o GPL foi obtido ou dados brutos com anotações forem fornecidos, a opção de incluir nomes de sondas é habilitada. Se houver alguma anotação com o nome “Gene Symbol”, essa anotação é marcada por padrão entre as informações das

[1]Este exemplo é arbitrário e não tem nenhuma relação com as amostras da série GSE2600, que está sendo utilizada como exemplo nas figuras.

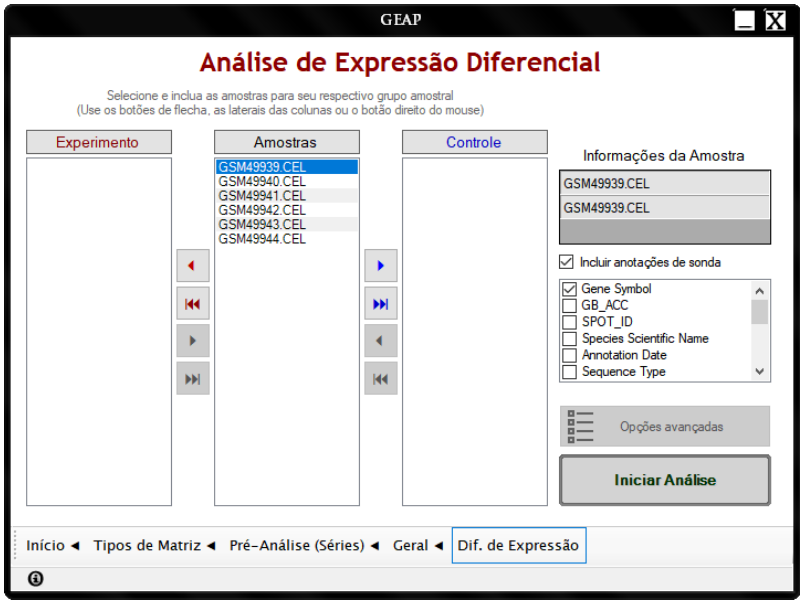

Figura 4.18: Comparação de expressão diferencial entre dois grupos.

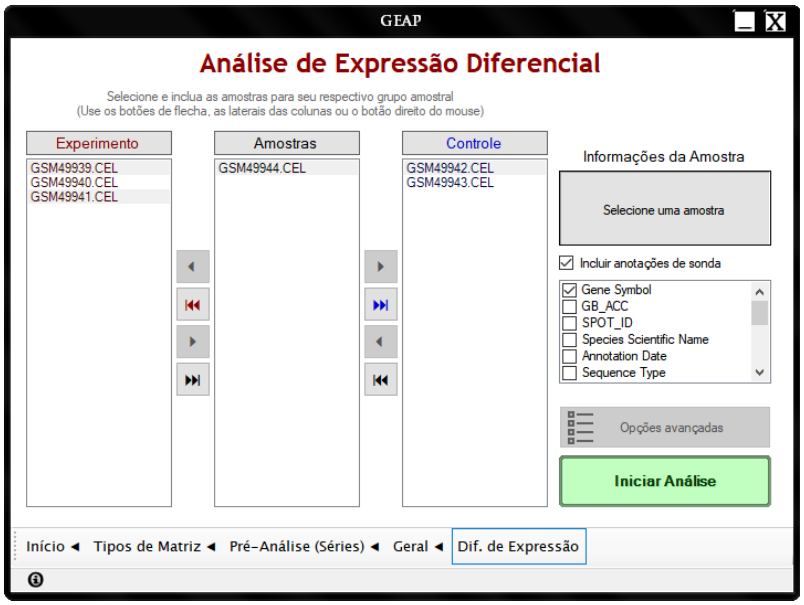

Figura 4.19: Amostras sendo selecionadas para comparação de expressão diferencial entre dois grupos.

amostras. Todas as anotações marcadas pelo usuário aparecerão juntamente às sondas entre os resultados.

Por último, existem as opções avançadas, onde o método de ajuste de valor-p é escolhido entre FDR (BENJAMINI; HOCHBERG, 1995), Benjamini-Yekutieli (BENJAMINI; YEKUTIELI, 2001), Hochberg (HOCHBERG, 1988), Holm (HOLM, 1979) ou Hommel (HOMMEL, 1988), sendo o FDR a opção padrão.

Finalmente, tendo no mínimo uma amostra para cada grupo, inicia-se a comparação clicando no botão Iniciar Análise. Nota-se, porém, que o valor-p só pode ser calculado quando existe pelo menos duas amostras em um dos grupos comparativos.

4.6.2 Comparação Entre Múltiplos Grupos

Neste caso, em vez de apenas um par de conjuntos de amostras ser comparado, diversos pares são comparados entre si. Ao acessar a tela desta modalidade, observa-se que, além das três colunas já vistas no modo Experimento X Controle, existe uma coluna adicional rotulada “Grupos”(Figura 4.20). Há também o botão “Modo de Comparação”, onde o usuário pode opinar entre as seguintes opções:

- Por Grupo: O par Experimento e Controle de cada grupo é comparado;
- Primeiro Controle: Para cada grupo, compara seu conjunto Experimento com o conjunto Controle do primeiro grupo;
- Lateral: Além do modo de comparação Por Grupo, compara Experimento X Experimento e Controle X Controle de um grupo com todos os demais grupos;
- Cruzado: Além do modo de comparação Por Grupo, compara Experimento de um grupo com controle de todos os demais grupos;
- Combinatório: Combina os modos Lateral e Cruzado – isto é, compara todas as combinações possíveis de conjuntos de amostras.

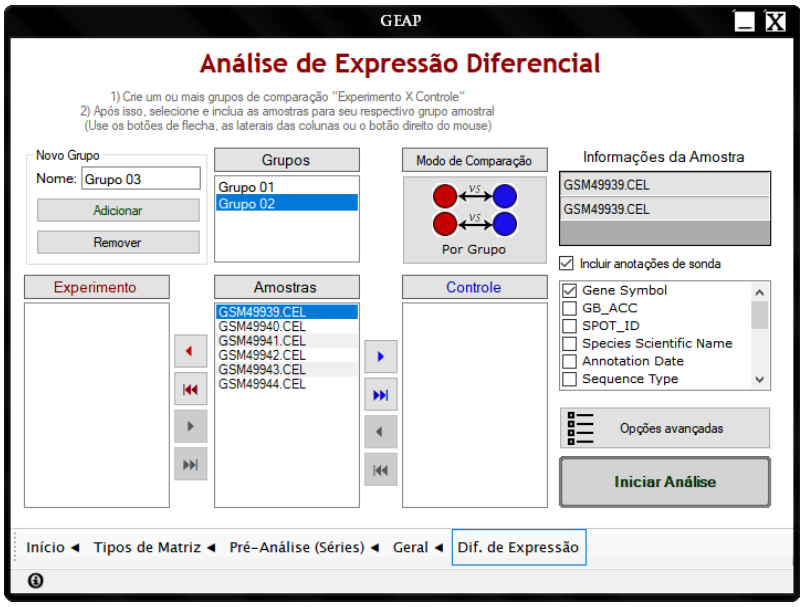

Figura 4.20: Comparação de expressão diferencial entre múltiplos grupos.

O primeiro passo na modalidade de comparação múltipla é adicionar grupos de pares de amostras. Cada grupo possui um nome, o qual deve ser único. Logo após a tela ser carregada, um novo grupo chamado “Grupo 01”é previamente adicionado, mas este pode ser removido pelo botão Remover caso for indesejado. Quando um grupo é selecionado, as colunas Experimento e Controle são preenchidas com as amostras selecionadas para este grupo. A exceção é quando a opção de modo de comparação está marcada como Primeiro Controle, onde o conjunto Controle permanece o mesmo e apenas o conjunto Experimento varia. Diferentemente da primeira modalidade de análise de expressão diferencial, as amostras não são transferidas para cada grupo, apenas adicionadas.

Os demais controles na região laterial direita se preservam em todas as modalidades de expressão diferencial. Exceto que, desta vez, para iniciar a análise, é necessário que todos os grupos contenham pelo menos uma amostra nos conjuntos Experimento e Controle.

4.6.3 Comparação Sequencial Entre Diferentes Etapas

A mesma lógica da comparação entre múltiplos grupos se aplica nessa modalidade. A única diferença é que todas as comparações são feitas sequencialmente, isto é, um conjunto de amostras com seu próximo. Por isso, o programa denomina cada conjunto de amostras como uma "Etapa"(Figura 4.21). Por exemplo, quando há dez etapas, as comparações feitas são Etapa 2 (experimento) X Etapa 1 (controle), Etapa 3 (experimento) X Etapa 2 (controle) e assim por diante até Etapa 10 (experimento) X Etapa 9 (controle). As etapas também devem ter nomes únicos, e uma primeira etapa chamada "Etapa 01"é adicionada previamente toda vez que o usuário acessa a tela. Da mesma forma como se adiciona grupos durante a comparação entre múltiplos grupos, neste modo se adiciona etapas, onde cada nome de etapa é definido pelo usuário. No exemplo atual, serão distribuídas as amostras em três etapas (Figura 4.22), adicionando duas amostras igualmente em cada etapa.

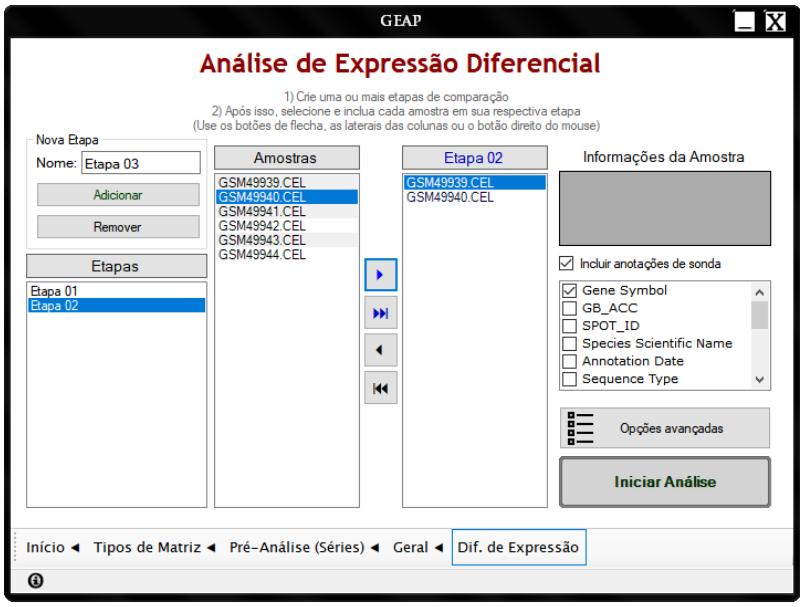

Figura 4.21: Comparação de expressão diferencial entre diferentes etapas.

4.7 Resultados

Após o término do processamento da análise de expressão diferencial, o usuário é conduzido a uma tela descrevendo os resultados obtidos. O conteúdo da janela possui algumas diferenças entre quando a análise é feita pela modalidade Experimento X Controle ou por uma das outras duas.

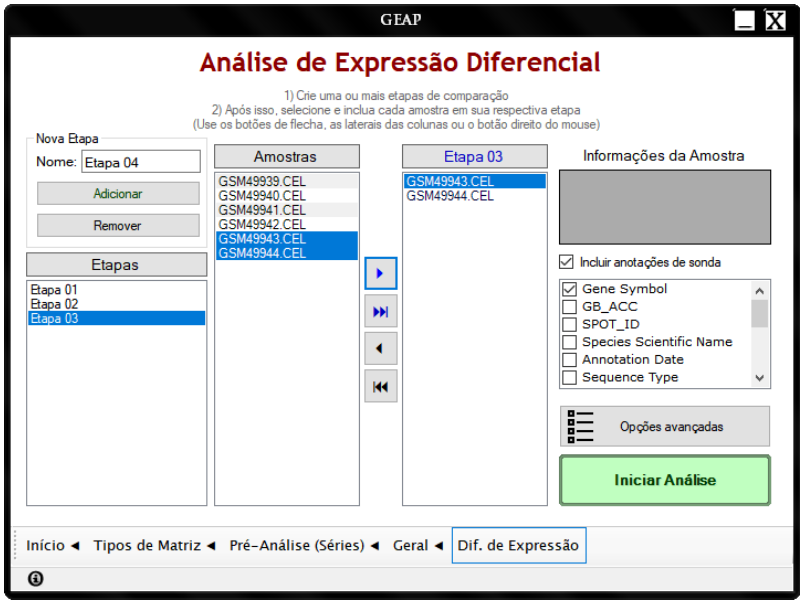

Figura 4.22: Amostras distribuídas entre diferentes etapas para comparação.

4.7.1 Resultados de uma análise Experimento X Controle

Na metade da esquerda da tela, aparece a lista das sondas que demonstraram os maiores e os menores valores de log_2FC após a comparação (Figura 4.23). Se o nome do gene associado a esta sonda estiver presente — isto é, a coluna Gene Symbol ter sido selecionada antes da análise — e o campo não estiver em branco, este nome será exibido no lugar do código da sonda. Abaixo, há dois botões: um acessa a tabela completa dos resultados obtidos, bem como os gráficos interativos, e o outro apresenta um breve relatório de todas as análises realizadas até pelo usuário, assim parte dos resultados obtidos.

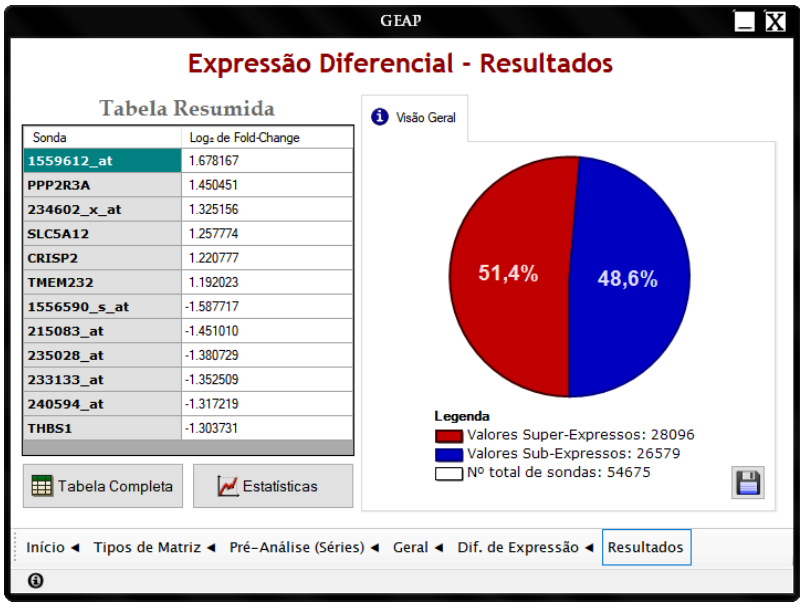

Figura 4.23: Resultados de uma análise na modalidade Experimento X Controle.

Na metade direita, um gráfico de pizza demonstra a proporção entre sondas super-expressas e sub-expressas (Figura 4.23), em porcentagem. O mesmo gráfico é exibido na janela da tabela completa.

Ao clicar no botão Tabela Completa, uma nova janela é aberta. Na metade esquerda, uma tabela contendo todas as linhas e colunas de resultados pode ser navegada e manipulada (Figura 4.24). Os botões no canto superior esquerdo são para ordenar a tabela de acordo com os valores de uma coluna de dados e para fixar ou desafixar a primeira coluna da tabela. Nota-se que as colunas podem ser reordenadas ao clicar e arrastar seu cabeçalho, portanto qualquer coluna que for colocada na primeira ordem pode ser fixada.

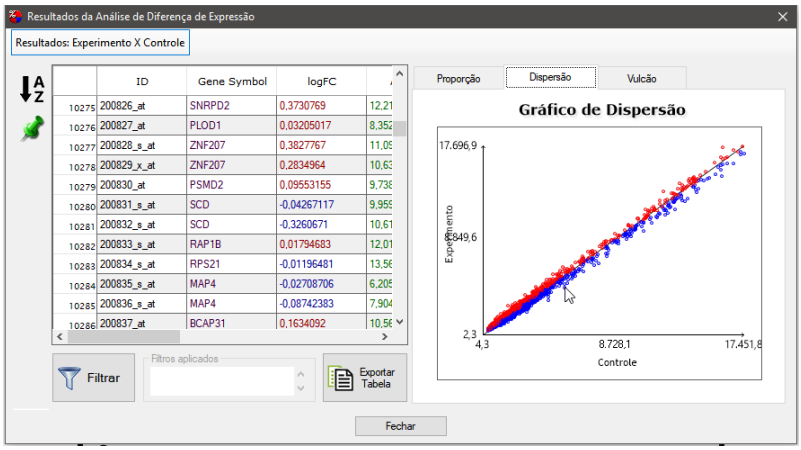

Figura 4.24: Janela com a tabela completa dos resultados e um gráfico de dispersão. Na primeira coluna da tabela (ID), os itens estão em preto. As colunas de anotação têm seu texto em roxo, e as colunas numéricas apresentam seu texto em verde. A coluna "logFC"é exceção, onde cada célula apresenta um texto vermelho para valores maiores que zero e azul para menores que zero. O propósito das cores vermelho e azul é o mesmo para os gráficos de dispersão e vulcão, denotando genes super e sub-expressos, respectivamente.

Abaixo da tabela, um botão "Filtrar"e outro "Exportar Tabela"estão separados por uma lista de filtros aplicados. Através do botão de Filtrar, uma janela surge com opções padrões e customizadas de filtro (Figura 4.25). A personalização de filtros no **GEAP** é um dos pontos diferenciais do programa, pois oferece uma flexibilidade excepcional na obtenção de resultados. Além de ser capaz de filtrar os valores de log_2FC e valor-p, o programa pode aplicar qualquer tipo de filtro para valores numéricos ou textuais na seção "Filtros personalizados", que filtrarão todos os dados que satisfaçam estes filtros. Colunas numéricas possuem filtros de comparação entre números (menor, igual, maior, entre outros), enquanto que colunas textuais filtram valores que começam, iniciam ou contêm um determinado texto. Para cada filtro personalizado, o usuário deve preencher as seguintes colunas:

1. Lógico (A partir da segunda coluna): A cada filtro, aplica um teste lógico "E"ou "OU", onde "E"requer que o filtro anterior seja verdadeiro para o presente ser validado;

2. Coluna: Nome da coluna em que se deseja aplicar o filtro;

3. Condição: Parâmetro condicional para aplicar o filtro, apresentando opções distintas entre colunas numéricas e textuais;

4. Valor: Número ou texto a satisfazer a condição. Em colunas numéricas, apenas números, sinais e pontuações são aceitas;

5. Absoluto: Em colunas numéricas, verifica se seu valor absoluto satisfaz o valor sugerido. Em colunas textuais, diferencia letras maiúsculas de minúsculas;

6. Deletar: Remove a linha com o filtro.

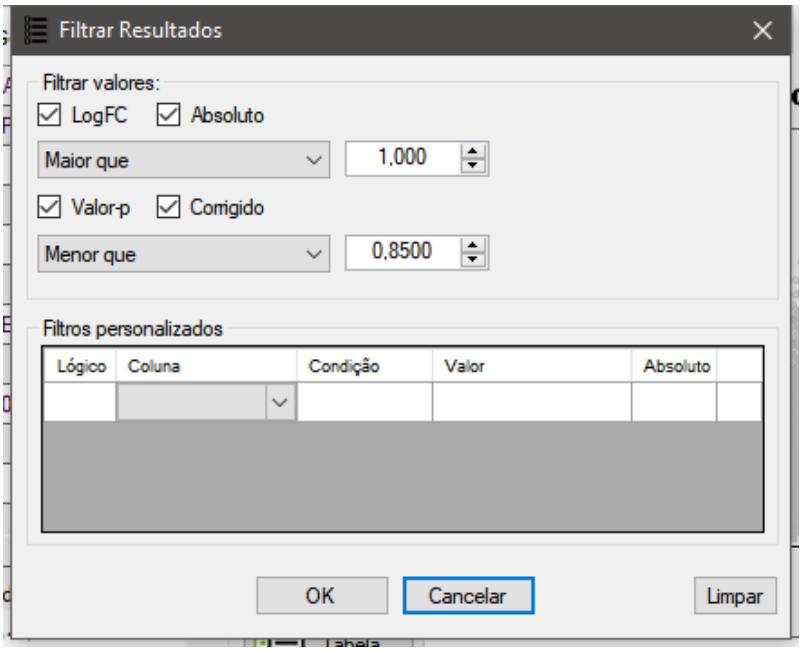

Figura 4.25: Tabela de filtros padrões (log_2FC e valor-p) e personalizados (grade na metade inferior da janela).

No canto extremo há o botão "Limpar". Este botão remove todos os filtros definidos e restaura os resultados ao seu estado original.

Ao sair da janela de filtro, sua confirmação causa a atualização da lista Filtros Aplicados, que apresenta todos os filtros que estão sendo utilizados pelo usuário. O botão no lado direito dessa lista exporta a tabela para um arquivo de texto conforme a tabela estiver organizada ou filtrada no momento.

Na metade direita da janela da Tabela Completa, existem três tipos de gráficos: Proporção, Distribuição e Vulcão (Figura 4.24). O gráfico de proporção é do tipo Pizza, sendo o mesmo visto no resumo dos resultados (Figura 4.23). A diferença é que as proporções se alteram conforme novos filtros são aplicados nos resultados, produzindo uma parcela cinzenta com o número de sondas que não foram filtradas. O mesmo efeito ocorre nos dois demais gráficos, onde pontos cinzentos representam sondas não filtradas.

O gráfico de distribuição representa o resumo dos valores de expressão para os grupos experimento (eixo Y) e controle (eixo X) (Figura 4.24). O gráfico de vulcão, como explicado na seção 2.3.5, apresenta o log_2FC e o valor-p corrigido nos eixos X e Y, respectivamente (Figura 4.26). Uma novidade que o **GEAP** oferece é que o usuário pode consultar o nome dos genes ou da sonda arrastando o mouse sobre os pontos de ambos os gráficos, e clicar em qualquer um destes pontos causa a seleção da sonda correspondente na tabela. Isso permite uma rápida consulta dos genes de modo visual. Além disso, múltiplos pontos podem ser selecionados ao arrastar o clique, embora a seleção simultânea de grandes quantidades de pontos possa apresentar um pequeno atraso. Se um filtro for aplicado, como o da Figura 4.25, os pontos não-filtrados serão parcialmente ocultos e apenas os coloridos permanecerão selecionáveis (Figura 4.27).

Quando desejado, as imagens dos gráficos podem ser copiadas ou salvas abrindo um menu com o botão direito do mouse sobre cada imagem. Para salvar, estão disponíveis os formatos mais comuns de imagem, incluindo JPEG, PNG e TIFF.

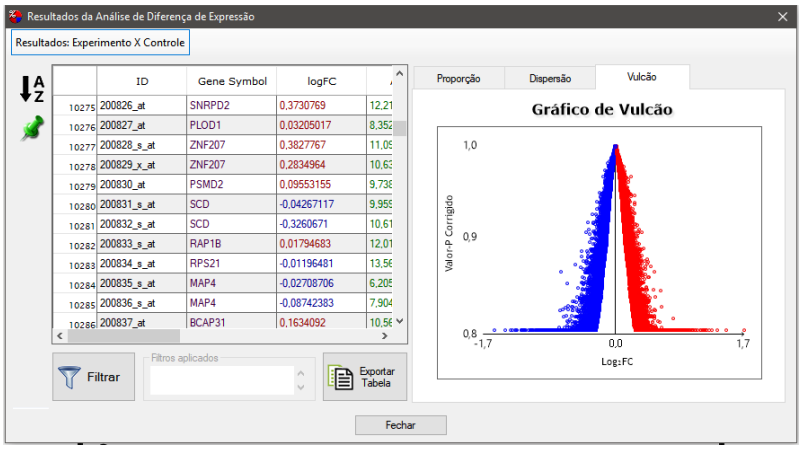

Figura 4.26: Gráfico de vulcão para os resultados.

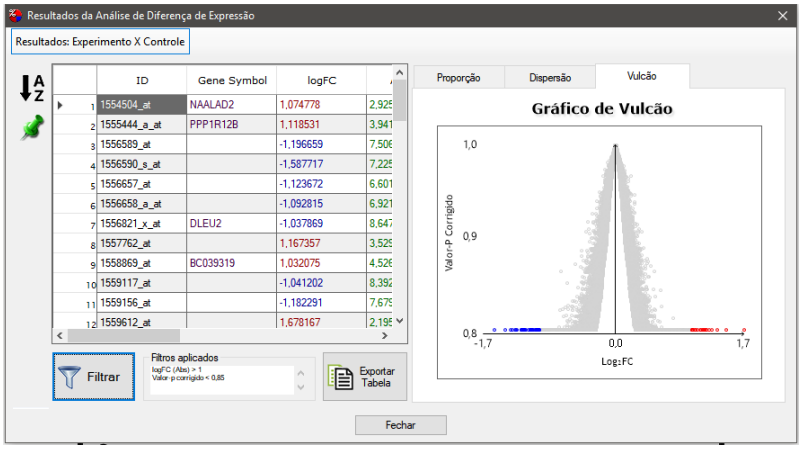

Figura 4.27: Gráfico de vulcão após um filtro ser aplicado. Observa-se que o valor-p das sondas é tão alto que mesmo um filtro próximo de 1, como o 0,85 aplicado, não consegue filtrar a maior parte das sondas presentes.

4.7.2 Resultados de uma análise entre múltiplos grupos ou diferentes etapas

Tomando como exemplo a análise iniciada na Seção 4.6.3, sendo três etapas, apenas duas comparações foram realizadas. A tela de resultados para comparações múltiplas e entre etapas possuem algumas diferenças em relação à modalidade Experimento X Controle, como pode ser observado na Figura 4.28. A tabela resumida na metade esquerda apresenta as comparações feitas e o número de valores com expressão diferencial. Na metade direita, há um gráfico ilustrando as proporções ao longo das etapas. Para resultados de comparações múltiplas, a única distinção é que o gráfico é representado em barras separadas, pois as comparações podem não ter relação uma com a outra. Uma outra diferença está na Tabela Completa (Figura 4.29), onde a região do cabeçalho contém abas selecionáveis, uma para cada resultado de comparação. Nessa tabela, encontra-se

o mesmo gráfico disponibilizado no resumo dos resultados, com a diferença de que sua porção central se torna cinzenta ao aplicar um filtro. Nota-se que a aplicação dos filtros é realizada para todos os resultados de uma só vez.

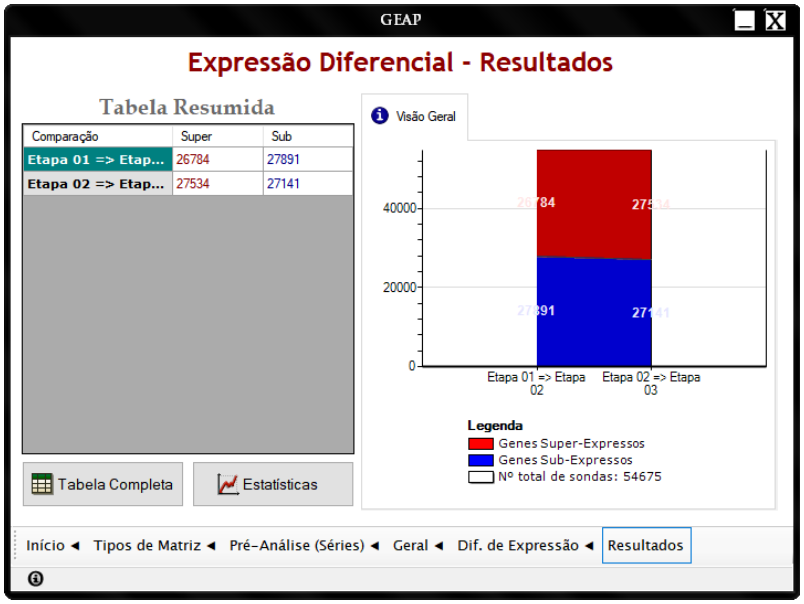

Figura 4.28: Resultados de uma análises na modalidade Entre Diferentes Etapas.

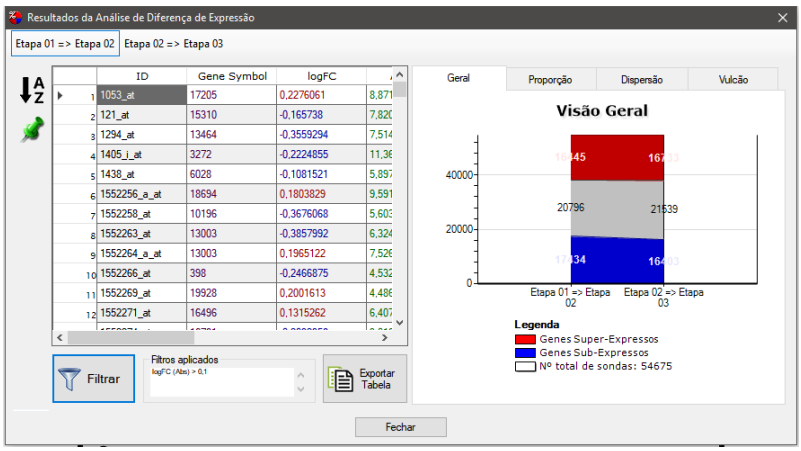

Figura 4.29: Tabela com resultados de uma análise de expressão diferencial entre diferentes etapas. Um filtro foi aplicado, tornando a parcela central do gráfico cinzenta, que se traduz na proporção de sondas não-filtradas.

5 CONCLUSÃO

No presente trabalho, foi desenvolvida uma plataforma de análise de transcriptomas por microarranjo, incluindo uma interface gráfica e a integração de funcionalidades que, até então, eram acessadas apenas pelo R. A ferramenta apresentou suporte para identificação, leitura e processamento estatístico de diversos tipos de arquivos, incluindo dados brutos gerados pelas fabricantes dos Kits de microarranjo. Não somente isso, como também permitiu a realização de análises de qualidade e expressão diferencial de forma visual, ainda integrando funcionalidades avançadas para comparações entre amostras. Em adição ao que já existe até recentemente entre os programas de transcriptoma, o **GEAP** também apresentou um sistema de obtenção de arquivos de microarranjo por *download*, criação de tabelas cutomizadas, interação com múltiplos gráficos que respondem aos comandos do usuário e filtros avançados para melhor refinamento dos genes DE obtidos.

Enquanto o ambiente visual cumpre o objetivo de uma aplicação padrão para transcriptomas, a flexibilidade das funções de seleção de amostras, correção estatística e filtragem de resultados satisfazem necessidades que, usualmente, apenas poderiam ser supridas por linhas de comando no R. Esta última observação é reforçada tendo em vista que o programa também utiliza a linguagem R como parte de seu funcionamento, sendo uma combinação eficaz com C#, que possui vantagens em termos de visualização e performance.

Todas as funcionalidades obtidas pelo **GEAP** poderiam ampliar o acesso às análises transcriptômicas para estudos no ramo da bioinformática e da saúde, bem como contribuir para novas produções científicas. Enquanto isso, mais melhorias poderão ser buscadas para as próximas versões deste software, sendo algumas delas discutidas no próximo e último capítulo.

6 UM PASSO À FRENTE

6.1 Melhorando o GEAP com o TypeChecker

Durante a etapa de pré-análise no **GEAP**, cada tipo de arquivo incluído pelo usuário é identificado e validado pelo programa. Essa identificação ocorre através de um conjunto de instruções que verificam as partes do arquivo dependendo de sua extensão e de seu conteúdo, aceitando apenas arquivos que sigam um formato predefinido pelo programa.

O problema é que, só no repositório do GEO, existe uma infinidade de tipos de arquivo possíveis para dados de transcriptoma, enquanto que o número de tipos diferentes de arquivos suportados pelo **GEAP** ainda é escasso, não cobrindo todas as possibilidades de leitura de arquivos de arranjo para certas plataformas. Como discutido anteriormente na seção 2.2.3, plataformas como Illumina são particularmente heterogêneas entre seus dados, ao mesmo tempo que podem conter dados valiosos obtidos em seus estudos. A primeira opção para confrontar essa dificuldade seria criando tabelas customizadas no próprio programa. Porém, nada garante que os dados estarão formatados de forma coerente com uma tabela de dados, e o usuário será obrigado a estudar os tipos de dados que está lidando, sendo que certos dados podem não possuir alguma documentação que descreva a seu respeito.

Uma solução de longo prazo seria montar as próprias instruções toda vez que um novo formato de arquivo for encontrado, e da mesma forma maneira compartilhar tais instruções com outros usuários pessoalmente ou como parte do próprio programa. Com esse objetivo, uma expansão para o **GEAP** foi desenvolvida com o nome de **TypeChecker** (Figura 6.1).

O TypeChecker, em linhas breves, é uma interface visual que permite montar novas instruções para leitura e pré-análise de arquivos no **GEAP**. As instruções são construídas utilizando blocos de Ação, Condição e Loop (iteração), não sendo necessário o conhecimento em qualquer linguagem de programação. Além das instruções, o programa oferece liberdade para personalizar o tratamento das amostras validadas. Caso o usuário tenha algum conhecimento em R, também poderá acessar comandos por esta linguagem antes, durante e após o tratamento. Adicionalmente, há templates prontos do R para usuários que não estejam familiarizados com a linguagem.

Após terminar as novas instruções, o arquivo é salvo, compilado e o usuário pode mover a compilação resultante para a pasta `Data/typecheck` a partir do diretório do **GEAP**. Com o arquivo presente na pasta informada, o **GEAP** processa e utiliza esse formato nas próximas verificações de arquivos com a extensão indicada pelo usuário. O arquivo pode também ser compartilhado, e futuramente o **GEAP** poderá providenciar um banco de dados próprio para inserir dados criados pelo usuário. Com isso, por fim, os

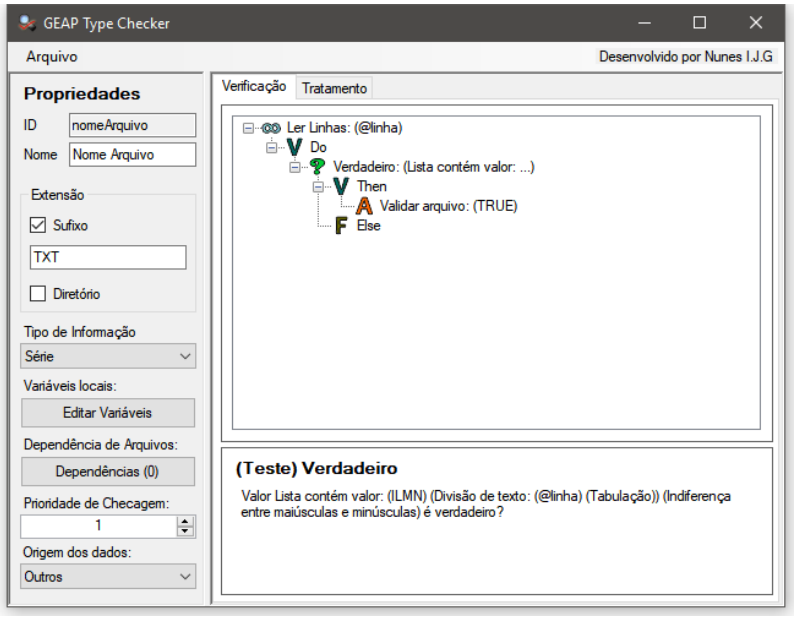

Figura 6.1: Tela principal do TypeChecker exemplificando a montagem de uma instrução.

usuários também poderão contribuir para melhorar o programa, o que facilitaria o trabalho do próprio pesquisador e de outras pessoas a longo prazo.

6.2 Perspectivas

6.2.1 Navegação e busca por arranjos através do programa

Uma das limitações do programa é que o usuário precisa primeiramente navegar pela Web dentro do domínio do GEO para saber o código do transcriptoma o qual está buscando. O banco de dados do GEO Datasets é vasto e seus tipos de dados podem variar bastante. Essa grande quantidade enorme de dados e opções para o usuário pode dificultar o início da procura pelos arranjos corretos.

Futuramente, poderia ser implementado um pequeno navegador para fins de pesquisa por dados de microarranjo. Tal navegador filtraria os resultados para colocá-los de modo simples ao usuário, e ignorando dados que não são de microarranjo ou que não teriam como ser suportados de forma alguma.

6.2.2 Mais contextualização biológica

Nem todas as plataformas oferecem um contexto biológico para seus genes em seu GPL. Certos ramos, como a Biologia de Sistemas, requerem dados adicionais a respeito dos genes DE obtidos nas análises, incluindo os processos biológicos em que atuam. Uma melhoria plausível seria a inclusão de análises de ontologias gênicas, onde se atribui anotações de processos observados na literatura para cada gene. Os dados de anotação são fornecidos pelo site *Gene Ontology Consortium*[1], e a análise pode ser feita pelo R utilizando pacotes como *GOexpress* (RUE-ALBRECHT et al., 2016).

[1]URL: `http://www.geneontology.org/`

6.2.3 Mais suporte a formatos de dados

Existem pacotes do R que estão se adaptando a novos formatos de dados gerados pelas fabricantes de transcriptoma. Embora o TypeChecker seja uma opção a longo prazo para dados armazenados em texto, ele ainda permanece ineficiente quando os dados estão compilados ou já possuem um suporte sólido dentro do R. O pacote *oligo*, por exemplo, permite processar dados tanto de microarranjos como de polimorfismos de nucloetídeo único (SNP), e até então suporta arquivos CEL da Affymetrix e XYS da Nimblegen (CARVALHO; IRIZARRY, 2010). Incluir esse tipo de pacote no **GEAP** enriquecerá sua capacidade de lidar com a variabilidade de dados e o manterá atualizado para plataformas mais recentes.

6.2.4 Análises de metilação e acetilação

Até então, apenas dados de perfil de expressão por transcriptoma são válidos para o programa. Contudo, a técnica de microarranjo sozinha não responde todas as perguntas sobre atividade biológica, e cada vez mais tem-se observado causas epigenéticas em processos biológicos (LAIRD, 2010). O GEO possui uma série de dados para perfis de metilação e acetilação, e tornar possível lidar com dados de padrões epigenéticos adicionaria uma funcionalidade significante ao programa.

6.2.5 Compatibilidade com outros sistemas operacionais

Um número significante de usuários opera em sistemas operacionais Linux e Mac OS. Visto que um dos propósitos do programa é oferecer acesso a análises transcriptômicas para qualquer usuário, disponibilizar o programa para estes ambientes expandiria sua acessibilidade. Ainda que o C# tenha nascido como uma linguagem exclusiva para Windows, uma série de esforços foi feita para introduzir a linguagem e sua interface aos outros sistemais operacionais. Atualmente, projetos como Mono[2] e MonoDevelop[3] já tornam isso possível, embora com diferenças sutis no aspecto visual da aplicação. De todo modo, tais projetos permitirão que o **GEAP** também possa ser desenvolvido e introduzido para Linux e Mac.

Por fim, espera-se que a maior parte das ideias presentes neste capítulo possam ser colocadas em prática para a próxima versão do **GEAP**.

[2]URL: `https://www.mono-project.com/`
[3]URL: `https://www.monodevelop.com/`

REFERÊNCIAS BIBLIOGRÁFICAS

ALBERTS, B.; BRAY, D.; HOPKIN, K.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Essential cell biology**. [S.l.]: Garland Science, 2013.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the royal statistical society. Series B (Methodological)**, [S.l.], p.289–300, 1995.

BENJAMINI, Y.; YEKUTIELI, D. The control of the false discovery rate in multiple testing under dependency. **Annals of statistics**, [S.l.], p.1165–1188, 2001.

BLALOCK, E. M. **A beginner's guide to microarrays**. [S.l.]: Springer Science & Business Media, 2003.

CARVALHO, B. S.; IRIZARRY, R. A. A framework for oligonucleotide microarray preprocessing. **Bioinformatics**, [S.l.], v.26, n.19, p.2363–2367, 2010.

CHAIN, B. **agilp**: agilent expression array processing package. [S.l.]: Bioconductor, 2012.

CONSORTIUM, I. H. G. S. et al. Finishing the euchromatic sequence of the human genome. **Nature**, [S.l.], v.431, n.7011, p.931, 2004.

CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cDNA microarray experiments. **Genome biology**, [S.l.], v.4, n.4, p.210, 2003.

DAN TENENBAUM, B. T. **BiocInstaller**: install/update bioconductor, cran, and github packages. [S.l.]: Bioconductor, 2018.

DARNELL, J. E.; LODISH, H. F.; BALTIMORE, D. et al. **Molecular cell biology**. [S.l.]: Scientific American Books New York, 1990. v.2.

DAVIS S, M. P. **GEOquery**: a bridge between the gene expression omnibus (GEO) and bioconductor. [S.l.]: Bioinformatics, 2007. 1846–1847p.

DU, P.; KIBBE, W. A.; LIN, S. M. nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays. **Biology Direct**, [S.l.], v.2, n.1, p.16, 2007.

DU, P.; KIBBE, W. A.; LIN, S. M. lumi: a pipeline for processing Illumina microarray. **Bioinformatics**, [S.l.], v.24, n.13, p.1547–1548, 2008.

DUNNING, M. J.; SMITH, M. L.; RITCHIE, M. E.; TAVARÉ, S. beadarray: r classes and methods for Illumina bead-based data. **Bioinformatics**, [S.l.], v.23, n.16, p.2183–2184, 2007.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. **Nucleic Acids Research**, [S.l.], v.30, n.1, p.207–210, 2002.

EDWARDS, D. Non-linear normalization and background correction in one-channel cDNA microarray studies. **Bioinformatics**, [S.l.], v.19, n.7, p.825–833, 2003.

EFRON, B.; TIBSHIRANI, R.; STOREY, J. D.; TUSHER, V. Empirical Bayes analysis of a microarray experiment. **Journal of the American Statistical Association**, [S.l.], v.96, n.456, p.1151–1160, 2001.

EZKURDIA, I.; JUAN, D.; RODRIGUEZ, J. M.; FRANKISH, A.; DIEKHANS, M.; HARROW, J.; VAZQUEZ, J.; VALENCIA, A.; TRESS, M. L. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. **Human molecular genetics**, [S.l.], v.23, n.22, p.5866–5878, 2014.

GAUTIER, L.; COPE, L.; BOLSTAD, B. M.; IRIZARRY, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. **Bioinformatics**, [S.l.], v.20, n.3, p.307–315, 2004.

GENTLEMAN, R. C.; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology**, [S.l.], v.5, n.10, p.R80, 2004.

GENTLEMAN, R.; CAREY, V.; HUBER, W.; IRIZARRY, R.; DUDOIT, S. **Bioinformatics and computational biology solutions using R and Bioconductor**. [S.l.]: Springer Science & Business Media, 2006.

GOHLMANN, H.; TALLOEN, W. **Gene expression studies using Affymetrix microarrays**. [S.l.]: CRC Press, 2009.

HANAUER, D. A.; RHODES, D. R.; SINHA-KUMAR, C.; CHINNAIYAN, A. M. Bioinformatics approaches in the study of cancer. **Current Molecular Medicine**, [S.l.], v.7, n.1, p.133–141, 2007.

HOCHBERG, Y. A sharper Bonferroni procedure for multiple tests of significance. **Biometrika**, [S.l.], v.75, n.4, p.800–802, 1988.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian journal of Statistics**, [S.l.], p.65–70, 1979.

HOMMEL, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. **Biometrika**, [S.l.], v.75, n.2, p.383–386, 1988.

INC, A.; MILLER, C. J. plier: implements the affymetrix plier algorithm. **R package version**, [S.l.], 2018.

IRIZARRY, R. A.; BOLSTAD, B. M.; COLLIN, F.; COPE, L. M.; HOBBS, B.; SPEED, T. P. Summaries of Affymetrix GeneChip probe level data. **Nucleic acids research**, [S.l.], v.31, n.4, p.e15–e15, 2003.

KAUFFMANN, A.; GENTLEMAN, R.; HUBER, W. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. **Bioinformatics**, [S.l.], v.25, n.3, p.415–6, 2008.

KAUFFMANN, A.; GENTLEMAN, R.; HUBER, W. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. **Bioinformatics**, [S.l.], v.25, n.3, p.415–6, 2008.

KOUSSOUNADIS, A.; LANGDON, S. P.; UM, I. H.; HARRISON, D. J.; SMITH, V. A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. **Scientific reports**, [S.l.], v.5, p.10775, 2015.

LAIRD, P. W. Principles and challenges of genome-wide DNA methylation analysis. **Nature Reviews Genetics**, [S.l.], v.11, n.3, p.191, 2010.

LIU, Y.; BEYER, A.; AEBERSOLD, R. On the dependency of cellular protein levels on mRNA abundance. **Cell**, [S.l.], v.165, n.3, p.535–550, 2016.

PENNISI, E. DNA study forces rethink of what it means to be a gene. **Science**, [S.l.], v.316, n.5831, p.1556–1557, 2007.

RALSTON, A.; SHAW, K. Gene expression regulates cell differentiation. **Nature Education**, [S.l.], v.1, n.1, p.127, 2008.

RUE-ALBRECHT, K.; MCGETTIGAN, P. A.; HERNÁNDEZ, B.; NALPAS, N. C.; MAGEE, D. A.; PARNELL, A. C.; GORDON, S. V.; MACHUGH, D. E. GOexpress: an r/bioconductor package for the identification and visualisation of robust gene ontology signatures through supervised learning of gene expression data. **BMC Bioinformatics**, [S.l.], v.17, n.1, p.126, 2016.

SCHENA, M.; SHALON, D.; DAVIS, R. W.; BROWN, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, [S.l.], v.270, n.5235, p.467–470, 1995.

SÎRBU, A.; KERR, G.; CRANE, M.; RUSKIN, H. J. RNA-Seq vs dual-and single-channel microarray data: sensitivity analysis for differential expression and clustering. **PLOS one**, [S.l.], v.7, n.12, p.e50986, 2012.

SMITH, M. L.; BAGGERLY, K. A.; BENGTSSON, H.; RITCHIE, M. E.; HANSEN, K. D. illuminaio: an open source idat parsing tool for illumina microarrays. **F1000Research**, [S.l.], v.2, 2013.

SMYTH, G. K. Limma: linear models for microarray data. In: **Bioinformatics and computational biology solutions using R and Bioconductor**. [S.l.]: Springer, 2005. p.397–420.

STAFFORD, P. **Methods in microarray normalization**. [S.l.]: CRC Press, 2008.

VENTURA, B. **Mandatory submission of microarray data to public repositories**: how is it working? [S.l.]: Am Physiological Soc, 2005.

WANG, Y.; TETKO, I. V.; HALL, M. A.; FRANK, E.; FACIUS, A.; MAYER, K. F.; MEWES, H. W. Gene selection from microarray data for cancer classification – a machine learning approach. **Computational Biology and Chemistry**, [S.l.], v.29, n.1, p.37–46, 2005.

WOLBER, P. K.; COLLINS, P. J.; LUCAS, A. B.; DE WITTE, A.; SHANNON, K. W. [2] The Agilent In Situ-Synthesized Microarray Platform. **Methods in enzymology**, [S.l.], v.410, p.28–57, 2006.

WU, J.; IRIZARRY, R.; MACDONALD, J.; GENTRY, J. Gcrma: background adjustment using sequence information. **R package version**, [S.l.], v.2200, 2012.

7 ADENDOS

OMICS A Journal of Integrative Biology
Volume 18, Number 6, 2014

DOI: 10.1089/omi.2013.0144

Fetal Alcohol Syndrome, Chemo-Biology and OMICS: Ethanol Effects on Vitamin Metabolism During Neurodevelopment as Measured by Systems Biology Analysis

Bruno César Feltes, Joice de Faria Poloni, Itamar José Guimarães Nunes, and Diego Bonatto

Abstract

Fetal alcohol syndrome (FAS) is a prenatal disease characterized by fetal morphological and neurological abnormalities originating from exposure to alcohol. Although FAS is a well-described pathology, the molecular mechanisms underlying its progression are virtually unknown. Moreover, alcohol abuse can affect vitamin metabolism and absorption, although how alcohol impairs such biochemical pathways remains to be elucidated. We employed a variety of systems chemo-biology tools to understand the interplay between ethanol metabolism and vitamins during mouse neurodevelopment. For this purpose, we designed interactomes and employed transcriptomic data analysis approaches to study the neural tissue of *Mus musculus* exposed to ethanol prenatally and postnatally, simulating conditions that could lead to FAS development at different life stages. Our results showed that FAS can promote early changes in neurotransmitter release and glutamate equilibrium, as well as an abnormal calcium influx that can lead to neuroinflammation and impaired neurodifferentiation, both extensively connected with vitamin action and metabolism. Genes related to retinoic acid, niacin, vitamin D, and folate metabolism were underexpressed during neurodevelopment and appear to contribute to neuroinflammation progression and impaired synapsis. Our results also indicate that genes coding for tubulin, tubulin-associated proteins, synapse plasticity proteins, and proteins related to neurodifferentiation are extensively affected by ethanol exposure. Finally, we developed a molecular model of how ethanol can affect vitamin metabolism and impair neurodevelopment.

Introduction

The maternal consumption of alcohol, especially during the initial 3–6 weeks of brain development, can lead to abnormal fetal nervous system changes during pregnancy, resulting in fetal alcohol syndrome (FAS) and fetal alcohol syndrome disorders (FASD) (de la Monte and Kril, 2014; Jaurena et al., 2011; O'Leary, 2004; Wentzel and Eriksson, 2009; Zhou et al., 2011). Although alcohol abstinence is recommended during pregnancy, more than 20% of pregnant women worldwide continue to abuse alcohol (van der Wulp et al., 2013). In these cases, a wide range of abnormal neurological outcomes can arise from FAS, including excessive neuron apoptosis (Genetta et al., 2007; Maffi et al., 2008), the risk of neuronal disorders (RNDs), and brain malformations during early embryonic development that affect neural crest and neural tube development (Wentzel and Eriksson, 2009; Zhou et al., 2011). In addition, alcohol abuse induces neuronal changes that affect both prenatal and postnatal life, including learning and cognitive impairments in young adults (O'Leary, 2004). Although FAS is an extensively studied pathology, the molecular pathways underlying its effects remain to be elucidated.

One of the many pathways affected by alcohol consumption is vitamin metabolism. Vitamin supplementation is necessary for fetal development, and specific vitamins play pivotal roles in the control of embryonic neurodevelopment (Table 1). For example, vitamins A and B_9 are related to neural tube closure and development (Table 1). In addition, reduced intake of vitamins has also been related to brain malformations or changes in neurodifferentiation patterns (Table 1). Moreover, alcohol consumption is already known to decrease the serum levels and absorption of the active forms of vitamin A (retinoic acid; RA), vitamin B_1 (thiamine; TM), vitamin B_9 (folic acid; FA), and vitamin E (α-tocopherol; α-TC) (Bjorneboe et al., 1987; Goez et al., 2011; Hewitt

Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, Departamento de Biologia Molecular e Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.

TABLE 1. MAJOR VITAMINS PRESENT IN THE INTERACTOME FOR *M. MUSCULUS* AND THEIR ROLE DURING NEURODEVELOPMENT OR PROPER NEURAL TISSUE FUNCTION THROUGHOUT EMBRYONIC BRAIN DEVELOPMENT OR NEURONAL *IN VITRO* LINEAGES

Vitamin	*Role in the neural tissue*
Vitamin A (RA)	RA is related to the control of the hindbrain and forebrain development and neural tube differentiation (Jiang et al., 2012; Rhinn and Dolle, 2012). It also regulates neuronal patterning along the anterior-posterior axis (Shearer et al., 2012).
Vitamin C (AC)	Vitamin C is essential for hippocampal development and for hippocampal postnatal function in guinea pigs (Tveden-Nyborg et al., 2012). Vitamin C also exerts anti-oxidant effects, preventing neurotoxic insults in the brain, such as ROS production (Tveden-Nyborg and Lykkesfeldt, 2009). Another study showed that ascorbic acid is highly present throughout midbrain development during human pregnancy (Adlard et al., 1974). Moreover, ascorbic acid was able to induce differentiation of CNS precursor cells into neurons and astrocytes (Lee et al., 2003)
Vitamin D [25-(OH) D_3]/[1,25-(OH)$_2D_3$]	Vitamin D induces neurite formation (Eyles et al., 2011). Additionally, 25-hydroxyvitamin D_3 upregulates nerve growth factor (NGF), which is essential for survival and growth of hippocampal and forebrain neurons (Eyles et al., 2011). Vitamin D deficiency is also correlated with decreased apoptosis and diminished cortex thickness (Eyles et al., 2013; Harms et al., 2011). The vitamin D receptor (VDR) is also present in the hippocampus (Eyles et al., 2013).
Vitamin K (PQN/MQN)	Promotes survival of cultured rat embryo CNS neurons (Nakajima et al., 1993). Pregnant women treated with vitamin K antagonist (warfarin) presented fetuses with abnormal dilatation of cerebral ventricle, microcephaly and mental retardation (Tsaioun, 1999). Showed neuroprotective role against oxidative stress (Josey et al., 2013).
Vitamin E (α-TC)	Described as playing an essential role in early brain formation, where tocopherol transporter protein (TTP) was present in the hindbrain and forebrain in zebrafish (Miller et al., 2012).
Vitamin B_{12} (CBL)	CBL is related to the process of myelination (Black, 2008), and deficiency in vitamin B_{12} is related to neural tube defects (Kirsch et al., 2013; van de Rest et al., 2012; Veena et al., 2010). Although not confirmed, a study shows that vitamin B_{12} deficiency might play a role in the developing brain and may change the normal cognitive status later in life (Bhate et al., 2008).
Vitamin B_9 (FA)	FA supplementation reduces the risk of neural tube defects in human embryos (Kirsch et al., 2013; Leung et al., 2013; Ross, 2010). It is also essential for fetal spine and cranial formation (Morse, 2012). Deficiency in FA is also related to inhibited neural rosette differentiation in monkey embryonic stem cells (Chen et al., 2012b).
Vitamin B_6 (PDX)	Pyridoxine was related to increased survival of neuronal cells *in vitro* by stimulation neurotransmitter release (Danielyan et al., 2011). Additionally, a study in rats shows that PDX deficiency caused diminished hippocampal weight and electrical activity, most likely due to poor myelination (Krishna and Ramakrishna, 2004). The catalytic form of vitamin B6 (pyridoxal phosphate) is also found in multiple parts of the brain and has its highest concentration in the olfactory tubercle (Ebadi, 1981).
Vitamin B_5 (PA)	Inactivation of panthotenate kinases, which phosphorylates PA, is related to neurodegeneration diseases during childhood **(Garcia et al., 2012).** Nevertheless, no studies have been performed to elucidate the role of PA alone during brain formation throughout embryogenesis.
Vitamin B_3 (NC)	Newborn mice injected with an antagonist of niacin showed damage in the central nervous system (CNS), and motor neurons as well as dorsal horn cells in the spinal cord showed signs of chromatolysis (Aikawa and Suzuki, 1986). However, no studies have been performed to elucidate the role of NC alone during brain formation throughout embryogenesis.
Vitamin B_2 (RBF)	RBF deficiency reduced the levels of important components of the myelin membrane in adult rats (Ogunleye and Odutuga, 1989). However, no studies have been performed to elucidate the role of RBF alone during brain formation throughout embryogenesis.
Vitamin B_1 (TM)	TM deficiency in rats caused abnormal growth of the hippocampus (Ba et al., 1996), and appears to affect myelinogenesis, axonal growth and synapsis formation (Ba, 2005). TM deffciency was also related to thalamus degeneration in alcoholics (Qin and Crews, 2013)
Vitamin H (BT)	Errors in biotin metabolism can cause enlargement of cerebral ventricles (Yokoi et al., 2009).

AC, Ascorbic acid; BT, Biotin; CBL, Cobalamin; 25-(OH) D_3, 25-hydroxyvitamin D_3; 1,25-(OH)$_2D_3$, 1,25 hydroxyvitamin D_3; FA, Folic acid (folate); MQN, Menaquinone; NC, Niacin; PA, Panthotenic acid; PDX, Pyridoxine; PQN, Phylloquinone; RA, Retinoic acid; RBF, Riboflavin; α-TC, α-Tocopherol; TM, Thiamine.

et al., 2011; Qin and Crews, 2013; Singleton and Martin, 2001). However, knowledge concerning the mechanisms through which alcohol affects the vitamin levels and metabolism at the molecular level is still scarce. Furthermore, because the effects of FAS are mainly on the nervous system and because all vitamins appear to play pivotal roles in brain formation (Table 1), it is crucial to understand the interplay between vitamins' biochemical pathways and alcohol during neurogenesis.

Using systems chemo-biology tools, we investigated different interactomic databases to elucidate the interplay between ethanol and different vitamins in the model organism *Mus musculus*. In addition, we compared transcriptomic data from available experimental studies that simulated maternal alcohol abuse and the effects of ethanol in the nervous system of the fetuses of *M. musculus*. Transcriptomic data originating from the adult *M. musculus* brain exposed to ethanol was also investigated to elucidate the main biological processes and changes in mRNA expression from ethanol over the short and long terms. Finally, the results gathered from systems chemo-biology analyses were used to develop interaction models for ethanol and vitamin metabolism, as well as to identify the gene expression changes caused by ethanol exposure in the nervous systems at different developmental stages and adulthood.

Materials and Methods

Interactome data mining and the design of chemo-biology networks

To design chemo-biology interatomic networks and to elucidate the interplay among neurodevelopment, vitamins, and ethanol, the metasearch engines STITCH 3.1 (http://stitch.embl.de) and STRING 9.05 (http://string-db.org) (Jensen et al., 2009; Snel et al., 2000) were used. All major active forms of vitamins (Table 1) commonly employed in commercial vitamin supplementation, as well as ethanol, were used as the initial seeds for network prospecting in STITCH. The STITCH software allows visualization of the physical connections between different biomolecules and chemical compounds, whereas STRING shows only biomolecules interactions (Kuhn et al., 2012). Each connection (edge) among biomolecules possesses a degree of confidence between 0.0 and 1.0 (with 1.0 indicating the highest confidence). The parameters used to prospect the networks for *M. musculus* in STITCH and STRING software were as follows: all prediction methods enabled, excluding text mining; 95 to 100 interactions (for each vitamin subnetwork for the ethanol subnetwork), resulting in a 2213-node network. In addition, a new network was developed to construct an interactome network for the microarray data for *M. musculus*, resulting in a network of 7395 nodes (Fig. 1); degree of confidence, medium (0.400); and a network depth equal to 1. The results gathered using these search engines were analyzed with Cytoscape 2.8.2 (Shannon et al., 2003) and Cytoscape 3.0. In addition, the GeneCards [http://www.genecards.org/] (Rebhan et al., 1997; Safran et al., 2010), KEGG [http://www.genome.jp/kegg/] (Carbon et al., 2009; Kanehisa and Goto, 2000), AmiGO 1.8 [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi] (Carbon et al., 2009), Reactome [http://www.reactome.org] (Jupe et al., 2012), BioCyc [http://biocyc.org/] (Caspi et al., 2010), and QuickGO [http://www.ebi.ac.uk/QuickGO/] (Binns et al., 2009) search engines were also employed, using their default parameters.

Gene expression data for the interatomic networks

We evaluated the transcriptomic data gathered from the matrix file GSE43324 (available at Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo]) as follows: pregnant C57BL/6J mice treated with saline, where fetuses were euthanized at embryonic day 16 (E.16) and whole brains were removed (termed control group ''b''), compared with pregnant C57BL/6J mice prenatally treated with intraperitoneal injections of ethanol (2.5 g/kg of ethanol in saline) during gestational days 14 and 16 (acute ethanol exposition; group ''a''), followed by embryo euthanasia at E.16, as described by Janus and Singh (2013). A mean value of expression for each gene was generated for both groups ''a'' and ''b''. In addition, a transcriptional analysis, derived from matrix file GSE34469, was performed using pregnant *M. musculus* treated with saline, where the adult offspring were euthanized at postnatal day 70 (termed control group ''b''), and compared to pregnant *M. musculus* treated with ethanol injections (2.5 g/kg of ethanol in saline) twice on gestational days 8 and 11, where the adult offspring were sacrificed at postnatal day 70 (ethanol group ''a''), as described by Janus et al. (2012). The same transcriptomic study compared pregnant *M. musculus* treated with saline, where the adults were sacrificed at postnatal day 70 and the whole brains were removed (termed control group ''b''), and pregnant *M. musculus* treated with ethanol injections (2.5 g/kg of ethanol in saline) twice on gestational days 14 and 16, where the adult offspring were sacrificed at postnatal day 70 (ethanol group ''a'').

Finally, data were gathered from another study derived from the matrix file GSE34549. Here, we compared *M. musculus* treated with 0.15 M saline alone as the control, where the adults were sacrificed at day postnatal 60 and the whole brains were removed (termed control group ''b''), with *M. musculus* treated with ethanol injections (2.5 g/kg of ethanol in 0.15 M saline) twice on days 4 and 7, the adult mouse was sacrificed at postnatal day 60 (ethanol group ''a''), as described by Kleiber (2012).

Additionally to the average expression values (already calculated with log 2) of the datasets, we applied Equation 1 (Castro et al., 2009), to value the relative expression, and the gathered data were overlaid in interactomes-derived clusters.

$$Z = \frac{a}{(a+b)} \qquad \text{(Eq.1)}$$

where a corresponds to the ethanol treated samples, and b indicates the control group. In this sense, the genes were considered overexpressed when $0.55 < Z \leq 1.00$; genes were considered underexpressed when $0.00 < Z < 0.45$ and, at last, genes were considered nondifferentially expressed when $0.45 < Z < 0.55$.

Different Venn Diagrams were created using the online tool Data Overlapping and Area-Proportional Venn Diagram [http://apps.bioinforx.com/bxaf6/tools/app_overlap.php] to visualize the number of over- and underexpressed genes shared among the networks.

Modular analysis of the main interactome network

The MME-Network (Fig. 1C) was analyzed in terms of the major clusters or module composition using the program

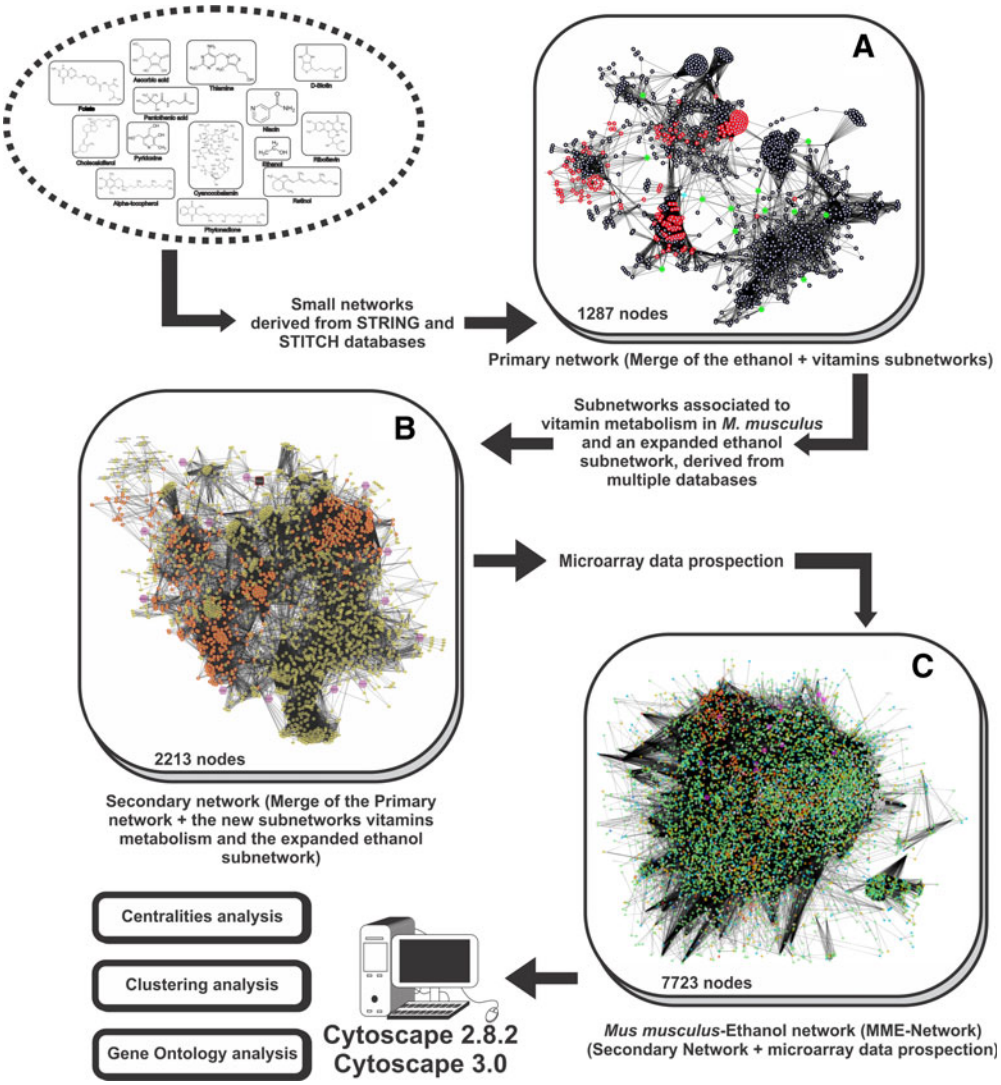

FIG. 1. Experimental systems chemo-biology workflow. In **(A)**, the primary network is composed of 1287 nodes (14 vitamins, ethanol, and 1262 biomolecules). The nodes linked to the ethanol subnetwork are shown with *red borders*. The vitamins were selected and used as initial inputs for searching different small subnetworks that were merged into a single large interactome. **(B)** Secondary network composed of 2213 nodes (14 vitamins, ethanol, and 2198 biomolecules). The data for nodes associated with vitamin and ethanol metabolism were gathered from multiple databases and merged with the primary network. **(C)** *Mus musculus*-Ethanol-Network (MME-Network), composed of 7723 nodes (14 vitamins, ethanol, and 7708 biomolecules). The data gathered from the microarrays were collected from the GEO database and were then entered into STRING and merged with the secondary network. The MME-Network was further analyzed with Cytoscape 2.8.2 and 3.0.

Molecular Complex Detection (MCODE) (Bader and Hogue, 2003). MCODE is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein, and isolates the dense regions according to parameters selected by the researcher (Bader and Hogue, 2003). The parameters for MCODE cluster finding were as follows: loops included; degree cutoff, 3; expansion of a cluster by one neighbor shell allowed (fluff option enabled); deletion of a single connected node from clusters (haircut option enabled); node density cutoff, 0.1; node score cutoff, 0.2; kcore, 2; and maximum network depth, 100. Each cluster generates a value of ''cliquishness'' (C*i*), which is the degree of connection in a given group of proteins. Thus, the higher the C*i* value, the more connected the cluster (Bader and Hogue, 2003).

Centrality analysis of the major resulting network

Centrality analysis was performed for the ''secondary network'' (Fig. 1B) using the program CentiScaPe 1.2 (Scardoni et al., 2009). In this analysis, the CentiScaPe algorithm evaluates each network node according to the node degree, betweenness, and closeness to establish the most ''central'' nodes within the network. Thus, the most topologically relevant node for a determined biochemical pathway or module can be obtained and further analyzed. In general terms, the closeness

analysis (1) indicates the probability that any node in our network is relevant to another protein/chemical compound in a signaling network or its associated network (Scardoni et al., 2009), as determined using Equation 2:

$$Clo(v) = \frac{1}{\sum w \in v^{dist(v,w)}} \quad \text{(Eq.2)}$$

where the closeness value of node v ($Clo(v)$) is determined by computing and totaling the shortest paths among node v and all other nodes (w; $dist(v,w)$) found within a network (1). The average closeness (Clo) score was obtained by calculating the sum of different closeness scores (Clo_i) divided by the total number of nodes analyzed ($N(v)$) (Equation 3).

$$\langle Clo \rangle = \frac{\sum_i Clo_i}{N_{(v)}} \quad \text{(Eq.3)}$$

The higher the closeness value compared to the average closeness score, the higher the relevance of a specific node for the other nodes within the network/module. In turn, the betweenness indicates the number of the shortest paths that go through each node (Equation 4) (Newman, 2005; Scardoni et al., 2009):

$$Bet(v) = \sum_{s \neq v \neq w \in V} \frac{\sigma_{sw}(v)}{\sigma_{sw}} \quad \text{(Eq.4)}$$

where σ_{sw} total number of the shortest paths from node s to node w, and $\sigma_{sw}(v)$ is the number of those paths that pass through the node. The average betweenness score (Bet) of the network was calculated using Equation 5, where the sum of different betweenness scores (Bet_i) is divided by the total number of nodes analyzed ($N(v)$):)].

$$\langle Bet \rangle = \frac{\sum_i Bet_i}{N_{(v)}} \quad \text{(Eq.5)}$$

Thus, nodes with high betweenness scores compared to the average betweenness score of the network are responsible for controlling the flow of information through the network topology. The higher a node's betweenness score, the higher the probability that the node connects different modules or biological processes, such nodes are called bottleneck nodes.

Finally, the node degree ($Deg(v)$) is a measure that indicates the number of connections (Ei) that involve a specific node (v) (Equation 6):

$$Deg(v) = \sum E_i \quad \text{(Eq.6)}$$

The average node degree of a network (Deg) is given by Equation 7, where the sum of different node degree scores (Bet_i) is divided by the total number of nodes ($N(v)$) present in the network:

$$\langle Deg \rangle = \frac{\sum_i Deg_i}{N_{(v)}} \quad \text{(Eq.7)}$$

Nodes with a high node degree are called hubs (Scardoni et al., 2009) and have key regulatory functions in the cell.

Gene ontology analyses of the major resulting network

The modules generated by MCODE were further studied by focusing on major biology-associated processes using the Biological Network Gene Ontology (BiNGO) 2.44 Cytoscape 2.8.3 plugin (Maere et al., 2005), available at http://www.cytoscape.org/plugins2.php#IO_PLUGINS. The degree of functional enrichment for a given cluster and category was quantitatively assessed (p value) using a hypergeometric distribution. BiNGO provides p values assessed by functional themes that are overrepresented on a given set of genes (e.g., clusters) (Maere et al., 2005). Multiple test correction was also assessed by applying the false discovery rate (FDR) algorithm (Benjamini and Hochberg, 1995), which was fully implemented in BiNGO software at a significance level of $p < 0.05$. The most statistically relevant processes were taken into account when developing the interaction model.

Results

Design of the interactome networks, topological analysis and transcriptomic data for Mus musculus

Systems chemo-biology tools allow prospecting of new drug targets and interaction between chemical compounds and biological networks (Chandra and Padiadpu, 2013; Csermely et al., 2013; Schneider and Klabunde, 2013). Our group has successfully employed systems chemo-biology to discover potential new anti-tumor drugs for gastric cancer (Rosado et al., 2011) and to understand the molecular pathways underlying fetal malformations associated with tobacco abuse during pregnancy (Feltes et al., 2013).

We first prospected small networks related to (i) the main active forms of each vitamin (Table 1), named the ''primary network'' (Fig. 1A), and (ii) metabolic-associated pathways for each vitamin and for ethanol in the STITCH and STRING databases for *M. musculus*, named the ''secondary network'' (Fig. 1B). Once gathered, these small networks were merged with the transcriptomic data in one large network named the *M. musculus*-ethanol network (MME-Network) (Fig. 1C).

The large MME-Network (Fig. 1C) was overlaid with four different transcriptomic datasets related to mouse offspring exposed to ethanol (Supplementary Table S1; Supplementary Material is available online at www.liebertpub.com). For this purpose, we used the public transcriptomic data available in the GEO database regarding *M. musculus* females exposed to the same concentration of ethanol (2.5 g ethanol/kg) during pregnancy and postnatal stage to simulate acute ethanol exposure. We also evaluated the late-life transcriptomic effects of ethanol exposure in the litters of pregnant females, which is necessary for understanding the cognitive and learning impairments observed in young adults with FAS (O'Leary, 2004). Thus, under- and over-upregulated genes were selected for transcriptome landscape analysis by overlaying these data on the following networks: (i) Prenatally-Exposed-Network (PE-Network; Fig. 2A), where the fetuses were exposed to ethanol during development (E.14 and E.16) and euthanized before birth (E.16), as described in the transcriptomics series GSE43324; (ii) Postnatal-Exposed MME-Network (PSE-Network; Fig. 2B), with pups exposed to ethanol (postnatal days 4 and 7) and euthanized at adult day 60 as indicated in GSE34549; (iii) Early Gestation-Exposed-Postnatal-Network (EGEP-Network; Fig. 2C), referent to

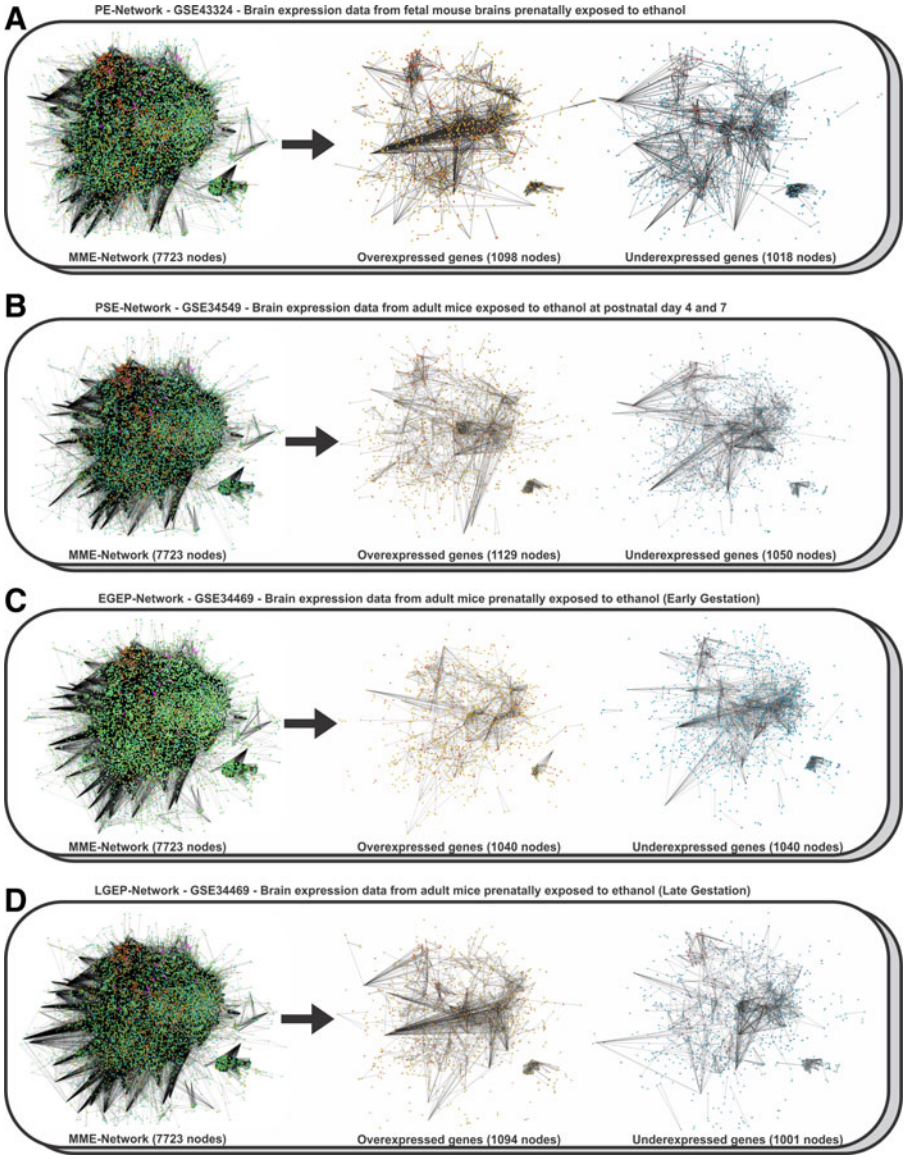

FIG. 2. Network landscape analysis of microarray data from *Mus musculus* embryos exposed to ethanol (MME-Network). In **(A)**, Prenatally Exposed (PE-) Network, derived from the transcriptomic analysis of prenatally ethanol-exposed embryonic brains from mice euthanized at E.16. **(B)** Postnatal-Exposed Network (PSE-Network), derived from a transcriptomic analysis of brains of postnatal ethanol-exposed mice, euthanized at day 70. **(C)** Early Gestation-Exposed-Postnatal-Network (EGEP-Network), derived from a transcriptomic analysis from the brains of prenatally ethanol-exposed embryos (E.4 and E.7), euthanized at postnatal day 60. **(D)** Late Gestation-Exposed-Postnatal-Network (LGEP-Network), derived from a transcriptomic analysis from the brain of prenatally ethanol-exposed embryos (E.14 and E.16), euthanized at postnatal day 60.

transcriptomics series GSE34469 where the fetuses were exposed to ethanol during development (E.8 and E.11) and euthanized at adult day 70; and (iv) Late Gestation-Exposed-Postnatal-Network (LGEP-Network) (Fig. 2D), referent to the series GSE34469, in which the fetuses were exposed to ethanol during development (E.14 and E.16) and euthanized at adult day 70.

Once the networks were overlaid with transcriptomic data, we select all those genes whose expression were similar in all treatment conditions (Fig. 3), allowing us to further analyze what genes could be commonly associated with acute and chronic ethanol exposure. In this sense, 19 genes were identified that were underexpressed in the PE-, LGEP- and PSE-Networks (Fig. 3A). Interestingly, these same 19 genes were present in the EGEP-, LGEP- and PSE-Networks (Fig. 3A).

Next, we generated another set of diagrams for overexpressed genes (Figs. 3B–E). The data show only five overexpressed genes that are shared among all transcriptomic

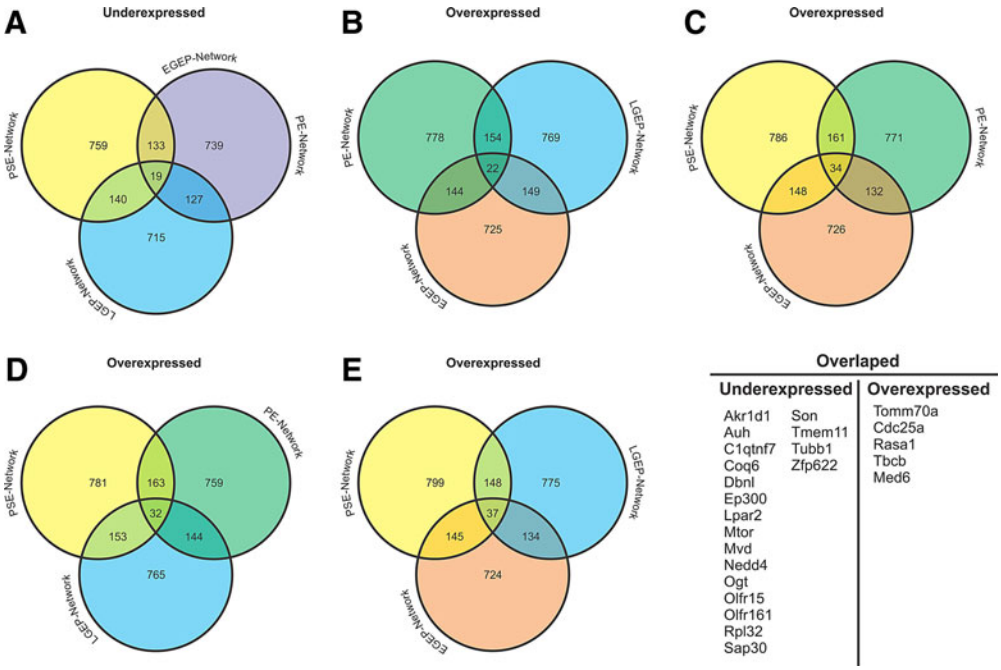

FIG. 3. Overlaps between the under- and overexpressed genes in all transcriptomic datasets. The *green circle* represents the Prenatally-Exposed (PE-Network, mice prenatally exposed to ethanol during E.14 and E.16 and euthanized at E.16). The *yellow circle* indicates the Postnatally-Exposed Network (PSE-Network, in which the dams were exposed to ethanol at postnatal days 4 and 7, and the offspring were euthanized at adult day 60). By its turns, the *blue circle* refers to the Late Gestation-Exposed-Postnatal-Network (LGEP-Network, in which the fetuses were exposed to ethanol during development at E.14 and E.16 and euthanized at adult day 70). The *orange circle* represents the Early Gestation-Exposed-Postnatal-Network (EGEP-Network, in which fetuses were exposed to ethanol at E.8 and E.11 and euthanized at adult day 70). Finally, the *purple circle* represents a fusion of the EGEP- and PE-Networks because they displayed the same underexpressed genes in the overlaps. **(A)** Overlap of the underexpressed genes, which showed 19 genes in common (displayed in the table on the *right side* of the figure); **(B)** Venn diagrams of the overexpressed genes of the PE-, EGEP-, and LGEP-Networks, sharing 22 genes; **(C)** Overexpressed genes overlapping among the PE-, PSE-, and EGEP-Networks, which showed 34 shared nodes; **(D)** Venn diagram showing the overlap between the overexpressed genes among the PE-, PSE-, and LGEP-Networks, with 32 shared genes; **(E)** Overlaps between the overexpressed genes in the LGEP-, PSE-, and EGEP-Network, revealing 37 shared nodes. The common nodes among the Venn diagrams of **B–E** are also listed in the table on the *right side* of the figure.

series (Fig. 3B–E). The relationship of these genes and their probable roles during pregnancy, neurogenesis, and vitamin metabolism will be discussed further. Nevertheless, the fact they are present in different ethanol exposure experiments in individuals of different ages suggests that they are closely related with the long-term effects of ethanol in brain development and physiology. In addition, we evaluated the ''secondary network'' (Fig. 1B) for the most topologically relevant nodes.

In a scale-free biological network, the most topologically relevant nodes are the hub-bottlenecks (HBs) (Yu et al., 2007) because they combine the bottleneck function (nodes that connect different clusters within a network and, consequently, display a betweenness score above the network average) and the property of hubs (nodes with a number of connections above the average node degree value of the network). Thus, HBs are critical nodes in a biological network (Yu et al., 2007). In our analysis, we observed 349 HB nodes in the ''secondary network'' of *M. musculus* (Fig. 1B). Of the 349 HBs in the *M. musculus* secondary network, 174 (49.8%) were connected to the ethanol subnetwork (Fig. 4A).

To understand how ethanol interacts with vitamins and the different proteins studied, we evaluated each transcriptomic network for the presence of modules or clusters, which allowed us to discover major biochemical pathways related to ethanol-vitamin metabolism. We found 15 modules above our cutoff score. Once the modules were obtained, a gene ontology (GO) analysis was performed. Biological processes that are important for neurodevelopment and neurobiological functions as well for vitamin metabolism that were present in each cluster were listed (Supplementary Table S2). The GOs relevant for vitamin, alcohol, and neurological function in Supplementary Table S2 are green. In addition, processes that were related to inflammation are blue, and the GOs related to amino acid metabolism are purple. Likewise, we performed additional GO analyses for the selected over- and underexpressed genes of each transcriptomic set. The main observed GOs were inflammation, synapses and neurotransmitter release,

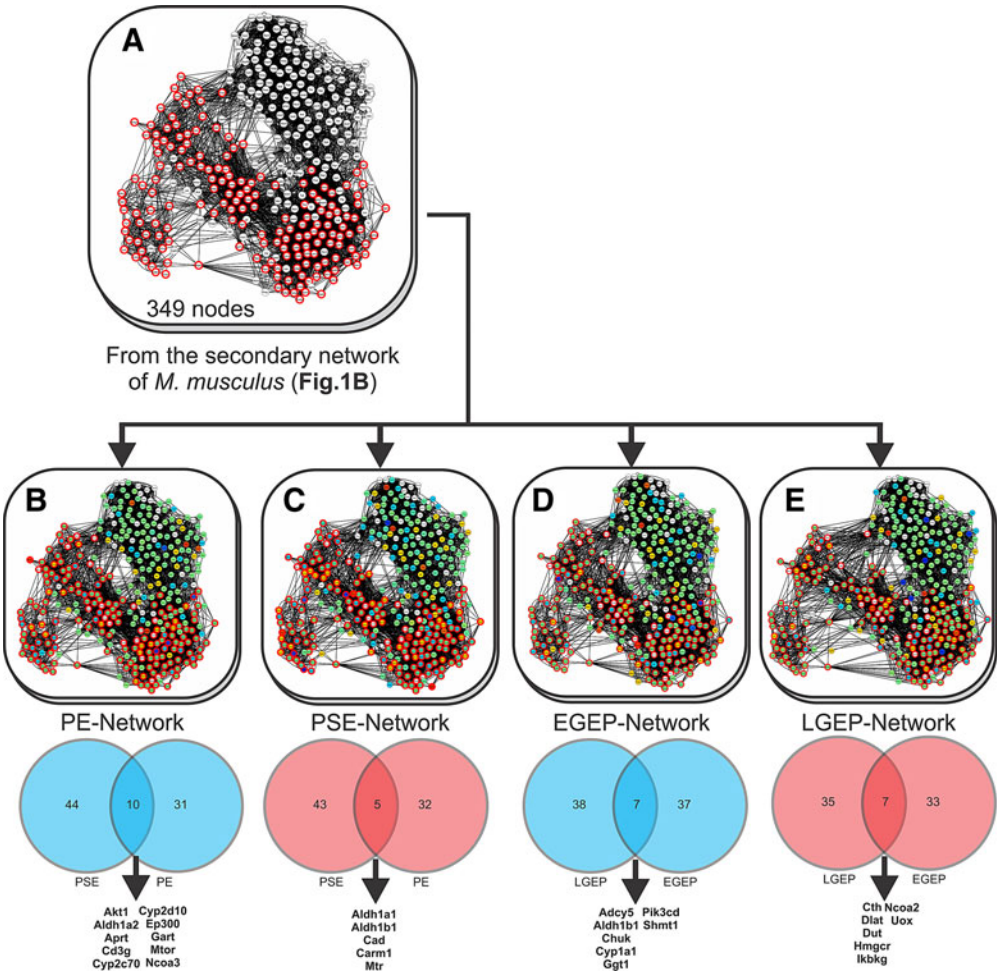

FIG. 4. Subnetworks derived from hubs-bottleneck (HB) analysis. Nodes colored with *red borders* are those found in the *M. musculus*-Ethanol Network (MME-Network). **(A)** *M. musculus* HBs; **(B)** HB displaying the expression data from the Prenatally Exposed Network (PE-Network); **(C)** HBs displaying the expression data from Postnatal-Exposed Network (PSE-Network); **(D)** HBs displaying the expression data from the Early Gestation-Exposed-Postnatal-Network (EGEP-Network); **(E)** HBs displaying the expression data from the Late Gestation-Exposed-Postnatal-Network (LGEP-Network). The *Venn diagrams* below each network display the overlaps between the indicated networks.

glutamate synthesis and metabolism, calcium ion signaling, and homeostasis and neurodifferentiation, and the summary of the GO information gathered in each network for over- and underexpressed genes is found in Table 2 (Fig. 3; for full data of the over- and underexpressed genes GO, see Supplementary Tables S3–6). Our analysis excluded GOs that were not associated with significant bioprocesses due to a lack of data or that were too general (e.g., the regulation of a biological process, regulation of transcription, or metabolism of organic substances). In addition, processes that were repeated among the GOs of over- and underexpressed genes were deleted. As expected, in the overexpressed GOs of different networks, the bioprocess of alcohol metabolism and processes associated with neuron physiology and function were highly expressed because the transcriptomic data were gathered from murine neural tissues (Supplementary Tables S3–S6).

In addition, the modularity and GO combined analysis for the MME-Network reveled that all modules, with exception of cluster 9 (Supplementary Table S2), were associated with neurodevelopment, alcohol metabolism, and/or vitamin metabolism, indicating that those bioprocess are closely related. Because these clusters are defined by highly dense, interconnected regions, the fact that they show close relationships with those processes may be useful for understanding how FAS affects neurodevelopment through vitamin metabolism.

Centralities analysis and overlaps among the ethanol-exposed groups

We also compared the under- and overexpressed genes in the centrality results of the PE-Network versus PSE-Network analyses (Fig. 4B and 4C, respectively) to understand the main differences between alcohol abuse in the developing organism and in the adult individual. We also evaluated the results of the EGEP-Network versus the LGEP-Network analyses (Fig. 4D and 4E) to observe the changes in HB status

Table 2. Major GO Terms Referent to Over- and Underexpressed Genes in the Interactomes for Each Transcriptomic Sets

Network (expression)	*GO*	*Corr* p-*value*	*x*	*Proteins*
PE-Network (overexpressed)	Negative regulation of apoptosis	4.6×10^{-13}	47	BMI1|SNCB|XIAP|PAFAH2|PRDX3|ITSN1|ADORA1|WT1| PCGF2|BDNF|CASP3|ATG5|LHX3|DNAJC5|MYC|HELLS| CD27|IHH|SPP1|FN1|ZC3HC1|MSH2|IL7|RXFP2|GRIN1| SPHK1|NR4A2|PROKR1|LIG4|DAPK1|ATF5|NME5|MNAT1|NOTCH1|EYA1| GNAQ|SFRP2|HIPK2|CX3CR1|SIX1|MTR| CFDP1|BMP7
	Vasculature development	4.1×10^{-7}	31	FGFR2|CAV1|NRP1|FGF9|LEPR|MMP2|WT1|CDH5| CTNNB1|SEMA5A|ATG5|APOE|TDGF1|GATAD2A|RHOB|NOS3|PLXND1|IHH| FN1|KLF5|SMAD5|SPHK1|EFNB2| GALT|FZD5|MNAT1|NOTCH1|PROK1|JUN|NTRK2|ZFPM2
	Aging	6.3×10^{-7}	14	MSH6|GNAO1|MSH2|POU1F1|GHRHR|NCAM1|CDKN2A| CYP27B1|APOE|MTR|MNT|SLC18A2|INPP5D|HAP1
	Negative regulation of cell communication	9.5×10^{-7}	28	HCRT|CAV1|FGFR3|GRIK1|FGF9|MBIP|ADORA1|GPC3| TDGF1|SKIL|INPP5D|AXIN1|IHH|PTPRC|AVP|PTPRF|PRKCD|CISH|SIGIRR| GRB10|CCND1|NOTCH1|BMPER|SFRP2|AVPR1A|BMP7|DRD1A|GRB14
	Negative regulation of neuron apoptosis	1.06×10^{-5}	12	BDNF|SNCB|XIAP|MSH2|HIPK2|SIX1|GRIN1|NR4A2| DNAJC5|PRDX3|LIG4|ITSN1
PE-Network (underexpressed)	Post-translational protein modification	9.8×10^{-10}	76	CDK19|CDC14B|STK35|PTPN22|RPS6KB2|LPAR2|LATS2| BTK|AKT1|GPX1|SIN3B|PRMT1|CRY2|PLOD1|SH2D1B1| PRKACA|FGF2|MAP2K7|EGFR|IRAK2|SRPK2|CAMK1G|PTPRM|PHKG2|CDK8| SOCS7|ARL6|PRKCQ|PPP1CA|EP300|PIAS4|PDGFRB|PIAS2|FBXO15| EIF2AK2|NSD1|UBE2T| MAP3K11|RAB3B|SRM|ERBB3|BRSK2|MAPKAPK3| TRIB3|KIT|EPHB3|CD74|GCKR|VRK1|MAP3K3|C1QTNF2|PKD1| PPP3CA|TCF3|PIK3R1|PTPN18|FLT4|TGFBR1|PTPRA|CS| PDE6G|EPHA1|RPS6KA1|GCK|PLK1|NEDD4|RNF2|NTRK1|PRKAR1A|GRK5| MTOR|MAPK8IP1|MERTK|IKBKB| BMPR1B|OPN4
	Calcium ion homeostasis	9.2×10^{-9}	25	GNA13|CCL2|PTGER3|PMCH|IL6ST|PIK3CB|HC|GRIK2| TRHR|PTH1R|NMB|PPOX|KCNA5|NPY1R|ITGB3|CSRP3| BAK1|HRH3|GCK|PLCG2|RYR1|TBXA2R|EPOR|BANK1| IL2
	Retinoid metabolic process	2.6×10^{-7}	15	EBP|CYP11A1|MVD|CYP11B1|AMACR|LSS|CPN2| SC4MOL|CYP17A1|INSIG2|AKR1C6|INSIG1|BMPR1B| FGF2|AKR1D1

(*continued*)

TABLE 2. (CONTINUED)

Network (expression)	*GO*	*Corr* p-*value*	*x*	*Proteins*
	c-AMP- mediated signaling	1.9×10^{-6}	15	P2RY12|GNA13|ADRB3|NPB|ADRB1|PTGER3|ADCY8| S1PR4|ADCY5|PTH1R|LHCGR|HTR4|RAPGEF4|FSHR| OPRD1
	Cognition	5.9×10^{-4}	80	GLRA1|ADCY8|OLFR1254|OLFR703|OLFR295|UCHL1| RPE65|OLFR808|NR2E1|OLFR1016|GPX1|OLFR1469| OLFR692|OLFR1054|OLFR554|OLFR1427|OLFR228|PLCB2|GJE1|OLFR1058| OLFR1152|OLFR461|OLFR836|WNT10B| MYO6|OLFR460|OLFR1247|OLFR167|ESR2|OLFR1348| OLFR161|AAAS|OLFR399|OLFR59|OLFR11|OLFR96| OLFR15|OLFR90|OLFR392|OLFR1046|OLFR1045|OLFR16|OLFR1104|GJA10| OLFR1234|C3|OLFR1085|OLFR1442| OLFR829|OLFR1500|PPT1|KIT|OLFR137|OLFR729| OLFR437|OLFR821|OLFR578|OLFR381|ACE|HRH3| OLFR1176|OLFR720|OLFR1094|OLFR502|OLFR1226| OLFR282|OLFR1451|OLFR716|NPY1R|DBH|PDE6G| OLFR1458|OLFR2|OLFR449|OLFR993|OLFR1490|OLFR376|OLFR73|OPN4| OLFR1122
EGEP-Network (overexpressed)	Negative regulation of cell death	7.5×10^{-6}	30	RBP4|TSPO|CAV1|NR2E3|PDX1|IL15|PTTG1|SLFN3|GLI3| ADORA1|TGFB2|SLFN2|LIF|BDNF|GPC3|CDKN2B|HSF1| GATA3|RARA|ITCH|BMP2|WNT10B|JARID2|RALBP1| SMAD3|GJB6|CTH|PLA2G2A|GLMN|WNT11
EGEP-Network (underexpressed)	Positive regulation of axonogenesis	7.4×10^{-6}	9	NTRK3|APC2|TIAM1|PLXNB1|ADNP|PAFAH1B1|NEFL| DSCAM|NGF
	Positive regulation of cell communication	4.6×10^{-5}	25	IL6|FKBP8|UTS2|PPARD|CARD9|ERBB4|CD3E|TAC1| ITGA2|JAG1|DGKI|MBD2|FURIN|NCAM1|ACVR1B| CDKN2A|MYD88|CD36|AGT|EXOC4|ADAM17|IL1B|NMU|CHUK|GHR
	Response to axon injury	4.1×10^{-4}	5	LAMB2|BCL2|BAX|NEFL|MMP2
LGEP-Network (overexpressed)	Positive regulation of apoptosis	1.3×10^{-13}	42	USP7|CDK5R1|TLR4|RRM2B|NR3C1|ZBTB16|LPAR1| PMAIP1|MMP2|IL10|ALDH1A2|NOD1|ALDH1A3|TICAM1|PCSK9|DIABLO| INPP5D|FAS|TRAF6|CASP2|MAP2K7| MAP2K6|CCAR1|COL18A1|PRKCA|TXNIP|IL2RA|PTPRF| GRIN1|BRCA2|IDO1|ATM|CIDEC|NOTCH2|NOTCH1| ADRB2|PSEN1|EEF1E1|ENDOG|PDE5A|WNT11|LRP5
	TNF-mediated signaling pathway	1.9×10^{-4}	5	TRAF2|TNFRSF11A|TNFSF11|KRT18|FAS

(*continued*)

TABLE 2. (CONTINUED)

Network (expression)	*GO*	*Corr* p-*value*	*x*	*Proteins*
LGEP-Network (underexpression)	Positive regulation of cell communication	4×10^{-9}	32	DCC|FGF18|FKBP8|CAV1|FGF9|CSF1|FGF10|LPAR2|EIF2A|ITGB3|TLR6|ITSN1|SRC|PHIP|ACVR1B|CD44|IFNG|GATA4|RBCK1|IL1A|CHUK|IL4|BMP4|DIXDC1|KL|CENPJ|KITL|WNT7B|CCR2|JAK2|MTOR|GHSR
	Lamellipodium assembly	7.2×10^{-4}	5	NCK1|SH2B1|CPB2|NCKAP1|FGD4
PSE-Network (overexpressed)	Negative regulation of apoptosis	2.3×10^{-10}	39	XRCC5|STIL|FGFR1|XIAP|SNCA|ELK1|NFKB1|BDKRB2|ADORA1|PHIP|BDNF|PTK2|CD44|ATG5|BCL2|AGT|PPP2CB|VNN1|NKX2-5|ERCC2|APC|BMP4|EEF1A2|SKP2|GIF|PROKR1|ESR2|TAX1BP1|DAPK1|RAD51|EYA1|TNFSF13B|IGBP1|MTR|CFDP1|TRP73|APIP|WNT7A|NGF
	Proteolysis	1.1×10^{-4}	41	C2|MASP1|CNDP2|UBE3A|MMP8|ENPEP|MMP2|PSMB4|CYLD|CUL7|PPP2CB|USP34|CUL1|CAPN7|SEC11C|UFD1L|FBXO2|SKP2|CAPN2|FURIN|AFG3L1|PSMB8|PSMB9|FOLH1|BLMH|CUL4A|TMPRSS11E|CLPP|PRCP|CTSC|ADAM12|TBL1X|CTSH|PMPCA|PMPCB|NCLN|PLAU
	Tachykinin receptor signaling pathway	2.1×10^{-4}	4	UQCRC2|METAP1|USP8|UQCRC1|APTACR2|TACR1|TAC1|TAC2
PSE-Network (underexpressed)	Axonogenesis	5.2×10^{-7}	24	FGFR2|ENAH|CDK5R1|WNT3A|UCHL1|KIF5C|DPYSL5|RTN4R|STXBP1|PIP5K1C|SLIT1|CXCL12|CTNNA2|NRCAM|ROBO1|CXCR4|MNX1|RIT1|SEMA3A|BMPR1B|BOC|APBB1|GAP43|KALRN
	Cytosolic ion calcium homeostasis	1.04×10^{-5}	16	CALCR|MCHR1|RXFP3|EDN1|PTH1R|NMB|CXCR3|ITPR3|EDNRA|GCK|AGTR1A|RYR1|TGM2|UTS2R|GLP1R|CACNA1A
	Regulation of c-AMP biosynthetic process	9×10^{-5}	13	CALCR|ADCY2|ADCYAP1R1|EDN1|PTH1R|TIMP2|EDNRA|S1PR3|HTR1B|S1PR4|HTR7|PTH|GLP1R
	Synaptic transmission	7.6×10^{-4}	21	GJD2|MYO5A|STX1A|HTT|GABRA6|MAOB|PPYR1|CLSTN1|STXBP1|SNAPIN|NTSR2|CTNNA2|CTNNB1|HTR1B|CAMK4|HTR7|HRG|VAMP2|TPR|SNAP25|CACNA1A

x, number of proteins associated with a given GO in the network.

TABLE 3. LIST OF OVERLAPPING NODES FOUND IN THE UNDER- AND OVEREXPRESSED GENES OF ALL FOUR NETWORKS (FIG. 3) AND AMONG THE HB NETWORKS (FIG. 4)

Protein	*Identity*	*Role in neurodevelopment and/or neurological function*	*Expression*
Adcy5	Adenylate cyclase	NDL	Underexpressed
Akr1d1	Aldo-Keto reductase	NDL	Underexpressed
Akt1	Kinase	Involved in neuronal differentiation (Park et al., 2012).	Underexpressed
Aldh1a2	Aldehyde dehydrogenase	Involved in the patterning of the CNS and neural tube (Marei et al., 2012; Strate et al., 2009). Could also be related to hindbrain defects in *Xenopus laevis* (Vitobello et al., 2011).	Underexpressed
Aldh1b1	Aldehyde dehydrogenase	Downregulated by ethanol during early nerulation (Zhou et al., 2011).	Underexpressed (EGEP-LGEP) Overexpressed (PE-PSE)
Aprt	Aphosphoribosyl-transferase	Aprt expression increases in course of neuron maturation in cell cultures (Brosh et al., 1990)	Underexpressed
Auh	Enoyl-CoA hydratase	Role in neural survival through its action on AU-rich elements (ARE) (Kurimoto et al., 2009).	Underexpressed
C1qtnf7	C1q and TNF related protein	NDL	Underexpressed
Cd3g	T-Cell surface glycoprotein	NDL	Underexpressed
Chuk (IκBα)	Serine/threonine kinase	Expression of this protein blocks self-renewal and induces neurodifferentiation (Khoshnan and Patterson, 2012).	Underexpressed
Coq6	Monooxygenase	NDL	Underexpressed
Cyp1a1	Cytochrome P450 family	Related to xenobiotic metabolism in the brain, where this protein was found with high activity in glial cells (Kapoor et al., 2006) and also abundant in the cerebral cortex and cerebellum (Iba et al., 2003).	Underexpressed (EGEP-LGEP) Overexpresed (PSE)
Cyp2c70	Cytochrome P450 family	NDL	Underexpressed
Cyp2d10	Cytochrome P450 family	NDL	Underexpressed
Dbn1	Actin-binding adapter protein	Plays a role in spine formation and synaptogenesis (Park et al., 2009)	Underexpressed
Ep300	Histone acetyltransferase	Expressed in multiple regions of the brain, including hippocampus, cerebral and cerebellar cortices and medulla oblongata (Tan et al., 2009)	Underexpressed
Gart	Phosphoribosyl-glycinamide Formyltransferase	Polymorphism in this gene was related to mouse neural tube defects (Pangilinan et al., 2012). This protein is also related to prenatal cerebellar development (Brodsky et al., 1997)	Underexpressed
Ggt1	Gamma-glutamyl transpeptidase	NDL	Underexpressed
Lpar2	Lysophosphatidic acid receptor	LPA has been implicated in neurogenesis of the CNS, targeting neural progenitors, neurons, astrocytes, microglia, oligodendrocytes and Schwann cells (Goldshmit et al., 2010)	Underexpressed
Mtor	Serine/Threonine kinase	Involved in synaptic plasticity, neuron survival and repair against brain injuries (Russo et al., 2012)	Underexpressed
Mvd	Mevalonate pyrophosphate decarboxylase	NDL	Underexpressed

(*continued*)

TABLE 3. (CONTINUED)

Protein	*Identity*	*Role in neurodevelopment and/or neurological function*	*Expression*
Ncoa3 (Src3)	Histone acetyltransferase	Expressed at high levels in the hippocampus (Tetel, 2009).	Underexpressed
Nedd4	E3 ubiquitin-protein ligase	NDL	Underexpressed
Ogt	O-Linked N-acetylglucosamine transferase	Cellular nutrient sensor, which may play a role in placental protection and in neurodevelopment by protecting the brain from insults such as nutrient deficiency (Howerton et al., 2013)	Underexpressed
Olfr15	Olfactory receptor	NDL	Underexpressed
Olfr161	Olfactory receptor	NDL	Underexpressed
Pik3cd	Kinase	NDL	Underexpressed
Rpl32	Ribosomal protein	NDL	Underexpressed
Sap30	Sin3A-associated protein	NDL	Underexpressed
Shmt1	Serine hydroxymethyl-transferase	Related to prepulse inhibition in mice (Maekawa et al., 2010). Lack of Shmt1 also results in neural tube defects in mice (Beaudin et al., 2011)	
Tmen11 (PMI)	Transmembrane protein	Involved in *Drosophila melanogaster* synapse formation and lifespan (Macchi et al., 2013)	Underexpressed
Tubb1	Tubulin, Beta 1 Class VI	Protein restricted to regions of the peripheral and central nervous system during early-differentiating neurons in zebrafish (Oehlmann et al., 2004).	Underexpressed
Zfp622	Zinc finger protein	NDL	Underexpressed
Son	Splicing cofactor	Expression of Son was related to neurogenesis during embryogenesis and postnatal brain (Ahn et al., 2011).	Underexpressed
Aldh1a1	Aldehyde dehydrogenase	ALDH1A1 expression appears in more differentiated parts of the developing brain such as cerebellar vermis or fetal white matter (Adam et al., 2012)	Overexpressed
Cad	Trifunctional protein (carbamyl phosphate synthetase, aspartate transcarbamylase and, dihydroorotase).	Cad protein was observed to be elevated during rat and hamster prenatal brain formation and hamster early postnatal brain development (Cammer and Downing, 1991). The authors argue that Cad is related to pyrimidine synthesis in astrocytes and in the grey matter.	Overexpressed
Carm1	Methyltransferase	Expression of this protein, inhibits HuD, a protein that is important for neurodifferentiation, synaptogenesis and learning and memory (Lim and Alkon, 2012).	Overexpressed
Cdc25a	Phosphatase	NDL	Overexpressed
Cth	Broad substrate specificity (deaminase, dehydratase, lyase, desulfhydrase)	Polymorphisms in this gene were related to autism (Bowers et al., 2011).	Overexpressed
Dlat	Pyruvate dehydrogenase	NDL	Overexpressed
Dut (DUTPase)	Nucleotido-hydrolase	Expressed in the prenatal rat brain, DUTPase might is responsible to maintain the low frequency of dUMP incorporation into DNA (Focher et al., 1990).	Overexpressed

(*continued*)

TABLE 3. (CONTINUED)

Protein	*Identity*	*Role in neurodevelopment and/or neurological function*	*Expression*
Hmgcr	Transmembrane glycoprotein	Overexpression of this protein, in combination to underexpression of ABCA1 (not present in our networks) is related to increased risk of Alzheimer's disease (Rodriguez-Rodriguez et al., 2009).	Overexpressed
Ikbkg	Kinase	NDL	Overexpressed
Med6	Transcription factor	NDL	Overexpressed
Mtr	Methyltransferase	Uses cobalamin (CBL) as co-factor, where deficiency of CBL causes dramatic degrease of Mtr (Gueant et al., 2013). In the same article is discussed that CBL deficiency during mice is associated with impaired memory.	Overexpressed
Ncoa2	Histone acetyltransferase	Overall, Ncoa2 expression is not detectable in the brain but is expressed in the anterior pituitary (Meijer et al., 2000) and in high levels in the dentate gyrus during adult stages but low on prenatal stages (Schmidt et al., 2007)	Overexpressed
Rasa1	GTPase-activating protein	NDL	Overexpressed
Tbcb	Tubulin folding cofactor B	Overexpression of TBCB results in abnormalities in the growth cone morphology, later causing neuronal degeneration (Lopez-Fanarraga et al., 2007)	Overexpressed
Tomm70a (KIAA0719)	Translocase	Regulated by thyroid hormone, which can lead to brain malformations when at abnormal levels (Alvarez-Dolado et al., 1999).	Overexpressed
Uox	Urate oxidase	Uox expression is correlated to diminished neuroprotective effects of urate in astrocytes and neurons (Cipriani et al., 2012). Its expression is also related to exacerbate the lesions caused by 6-hydroxydopmaine in dopaminergic neurons (Chen et al., 2013).	Overexpressed

The full description of the most relevant targets and proteins are discussed along the study.
NDL, No Direct Link.

in adult individuals exposed to ethanol at different stages of embryonic development (Fig. 4).

The centrality analysis of the secondary network (Fig. 1C), the overlaps among the under- and overexpressed genes of the PE-, EGEP-, LGEP- and PSE-Networks, and the overlaps of the expression of the HB subnetworks (Fig. 4B-E) resulted in a list of 51 potential targets involved in FAS progression (Table 3). Our results for the under- and overexpressed genes are listed in Table 3.

Among the selected targets is AU-rich hydrolase (AUH) (Table 3), a protein that binds to AU-rich elements (ARE) in RNAs (Kurimoto et al., 2009). *AUH* mRNA lacks ARE and is upregulated in the mouse brain when mood stabilizers (e.g., lithium carbonate and valproic acid) are administered. Interestingly, these drugs upregulate the expression of ARE-containing mRNAs, such as the apoptotic inducer BCL2 (Kurimoto et al., 2009). This correlation indicates that AUH may promote neuron survival against apoptosis. Because AUH is downregulated in our networks, it becomes an important target in understanding FAS-induced neuronal damage. Moreover, debrin 1 (Dbn1) was also found among the underexpressed genes in the prospected networks (Table 3). Dbn1 is related to the formation of neuronal gap junctions in the mesencephalic trigeminal nucleus, which is crucial for synapse function through receiving inputs from nerve terminals and neurotransmitters (Park et al., 2009). Synapse plasticity, and protection against brain injury (Russo et al., 2012) is also related to mTor expression, which was reduced in the four networks (Fig. 2A-D). Changes in the mTor pathway have also been linked to neurological diseases such

as Alzheimer's and Parkinson's diseases (Russo et al., 2012). mTor has been linked to synaptic plasticity, both in long term potentiation in the hippocampus and by coordinating protein synthesis (Russo et al., 2012). The co-activator-associated arginine methyltransferase (Carm1) was also among the overexpressed genes in the overlaps between the HB subnetworks of the PE- and PSE-Networks (Fig. 4; Table 3). Carm1 was found to be responsible for the inhibition of HuD, a protein that is related to synaptogenesis, learning, memory, and neurodifferentiation (Lim and Alkon, 2012). This finding indicated that some of the highly topologically relevant proteins affected by ethanol are associated with synaptic plasticity, consistent with FAS-associated learning and memory impairments.

Proteins that belong to the tubulin family, such as Tub11, or that affect tubulin mechanisms, such as Tcbc and Son, were among our potential targets. Both Son, a splicing cofactor linked to mitotic spindle assemble (Ahn et al., 2011), and Tub11, a tubulin Beta 1 class VI, were among our underexpressed genes. Both proteins appear to be present during neurogenesis in embryonic development and are related to the differentiation of different brain regions (Ahn et al., 2011; Oehlmann et al., 2004). We also identified Tbcb among the overexpressed genes in our networks (Table 3). Tbcb is a tubulin-folding cofactor, primarily associated with axonogenesis, which can cause brain malformations when overexpressed (Lopez-Fanarraga et al., 2007). These results show that ethanol exposure induced the deregulation of tubulin and tubulin-associated proteins during neurodifferentiation.

The protein Gart was also among the overlaps of HB subnetworks in the underexpressed datasets for the PE- and PSE-Networks (Fig. 4). This protein appears to be expressed at higher levels in the prenatal cerebellum as compared to adults (Brodsky et al., 1997). More interestingly, human neuroblastoma cells treated with 6-hydroxy-dopamine, which mimic the effects of Parkinson's disease, showed Gart to be downregulated (Noelker et al., 2012). The authors also propose that downregulation of Gart could lead to neuron apoptosis. This is consistent with our data, which shows that in the PE- and PSE-Networks the process of negative regulation of apoptosis and neuron apoptosis to be overexpressed (Table 2). These observations indicate that Gart downregulation by ethanol during the early stages of development could lead to learning and memory impairments in adults.

The coactivator Ncoa3 (SCR3) is also present in the overlaps among the downregulated gene datasets in the HB subnetwork of the PE- and PSE-Networks (Fig. 4). Ncoa3 is expressed in the hippocampus and is related to retinoic acid (RA) signaling (Kashyap and Gudas, 2010). Retinoic acid (RA) is a vitamin A derivative involved in neural differentiation and neurogenesis (Chen et al., 2012) (Table 1). The presence of RA mediated the release of coactivators, leading to the transcriptional activation of retinoic acid receptors (RARs) (Kashyap and Gudas, 2010). In addition, Ncoa3 is a downstream mediator of vitamin D (VD) signaling (Ahn et al., 2009). These results indicate that ethanol is able to reduce VD and RA signaling through the downregulation of Ncoa3 in the prenatal brain, affecting brain regions such as the hippocampus. Moreover, Shmt1is another gene downregulated in the same network overlaps that is also related to vitamin metabolism and is a serine hydroxymethyltransferase involved in folate metabolism (Beaudin et al., 2011). The authors note that *Shmt1*$^{(+/-)}$ mice showed impairment in neural tube closure and that Shmt1 expression is also coordinated by RA, indicating that this protein could have an important role in ethanol-mediated brain defects.

Discussion

Interpolation between ethanol and vitamin metabolism during neurodevelopment

Considering the data gathered from systems biology analyses, we performed a more detailed evaluation of the major vitamin-related pathways and neurodevelopment processes affected by the ethanol in the following sections.

Retinol signalization, folic acid metabolism, synapsis induction, and circadian rhythm are affected by ethanol during embryogenesis. The GO analysis of cluster 3 (Supplementary Table S2) indicated the presence of proteins related to circadian rhythm and the folic acid, retinoid, and neurotransmitter metabolic processes. It should be noted that RA synthesis is induced upon the loss of synaptic activity and decreased dendritic calcium levels (Chen et al., 2012). In the prenatal ethanol exposure network (PE-network; Fig. 2A), we found that retinoid metabolic processes and Ca^{2+} ion homeostasis are underexpressed (Table 2). This is interesting because both calcium ion homeostasis and RA metabolism genes were underexpressed, showing that the induction of RA to overcome synaptic loss might not be possible in ethanol-exposed fetuses. Moreover, RARs have also been found to be essential for improved learning and memory in the adult brain and have even alleviated memory deficits in a transgenic mice model for Alzheimer's disease (Nomoto et al., 2012). Cognitive and learning abilities have already been found to be affected in young adults displaying FAS (O'Leary, 2004), and the impairment of RA metabolism could be an explanation that has not yet been examined. Moreover, RARα is abundantly found in the cortex and hippocampus (Nomoto et al., 2012). Our data indicated that RARα is downregulated in the EGEP-Network (Chen et al., 2012) (Supplementary Table S1), showing that the changes in synaptic plasticity and learning behaviors that depend on RA occur specifically during early development.

To corroborate the idea that ethanol affects RA, we observed that ALDH1A2 (RALDH2) gene, which codes for an aldehyde dehydrogenase and is responsible for the synthesis of RA from retinal (Strate et al., 2009), was found to be underexpressed in both the PE-Network and the PSE-Network (Fig. 4). This is interesting because ALDH1B1, another aldehyde dehydrogenase, is downregulated in ethanol-exposed embryos during nerulation (Zhou et al., 2011), a finding corroborated in our systems chemo-biology analysis, as ALDH1B1 was downregulated in both the EGEP-Network and the LGEP-Network (Fig. 4).

Folic acid (FA) metabolism was also associated with Cluster 3 (Supplementary Table S2). It is already known that ethanol affects folic acid absorption in guinea pigs (Hewitt et al., 2011) and that FA has multiple roles in neural tissue (Table 1). Remarkably, FA is able to differentiate neurospheres into multiple neural cell types and promote synaptic connections in Pax3-deficient mice (Ichi et al., 2012).

In the transcriptomic datasets analyzed, only the dihydrofolate reductase (DHFR) gene, which codes for a key

enzyme in folate metabolism (Cluster 3: Supplementary Table S2), was found to be underexpressed in fetuses exposed to ethanol in the final phase of development (LGEP-Network; Fig. 2D). One study shows that FA deficiency decreases neural progenitor cell proliferation in the mouse forebrain during late gestation (Craciunescu et al., 2004). Therefore, ethanol may also affect neuronal proliferation through FA metabolism, mainly through DHFR deregulation.

Another important process found within Cluster 3 is circadian rhythm (Supplementary Table S2). Ethanol consumption and abuse affect are known to affect sleep cycles and melatonin secretion (Brager et al., 2010; Roehrs and Roth, 2001). Interestingly, our data analyses (PE-Network; Fig. 2A) indicate that calcium ion homeostasis is downregulated (Table 2). Circadian rhythm is controlled by melatonin secretion, which consequently has a central role in inducing neurodevelopment during embryogenesis by inducing calcium ion signaling (de Faria Poloni et al., 2011). In pregnant women, the abuse of ethanol could skew proper melatonin secretion and calcium ion signaling and subsequently change sleep induction and neurodevelopment. We found RA, thiamine (TM), α-tocopherol (α-TC) and phytonadione (phylloquinone; PQN) in cluster 3.

Ethanol negatively affects vitamin D metabolism and leads to its degradation. Another interesting cluster (Cluster 5) presented several GOs related to neurodevelopment and RA and VD metabolism (Supplementary Table S2). Among all of the genes/proteins belonging to this cluster, CYP2R1 (Figs. 2A and 2C) was found to be underexpressed in both murine fetuses exposed to ethanol (PE-Network; Fig. 2A) and also in murine adults that were exposed to ethanol during development (EGEP-Network; Fig. 2C).

CYP2R1 is a VD hydroxylase that converts vitamin D_3 into the first active ligand (25-hydroxy vitamin D_3—$25OHD_3$) for the vitamin D receptor (VDR) (Eyles et al., 2013). VDR forms heterodimers with retinoid X receptors (RXR) to initiate transcription during the differentiation of different tissues (Eyles et al., 2013). It has been reported that VD deficiency is correlated with decreased intracellular calcium levels in rat cortex (Baksi and Hughes, 1982), which infers that ethanol affects calcium ion homeostasis in the PE-Network and is critical for neurodevelopment. VDr expression is also observed in differentiating fields in rodent brains and in proliferating cells in the lateral ventricle (Eyles et al., 2013).

Interestingly, VD deficiency is associated with low induction of neurogenesis and the loss of apoptosis, generating larger brains due to abnormal proliferation (Eyles et al., 2013). This statement corroborates our GO results that indicate an increase in the negative regulation of apoptosis in EGEP-Network (Table 2). Thus, by affecting VD metabolism, VDR-dependent transcription could also be affected by ethanol, not only through target-protein recruitment but also through the formation of heterodimers with RXR, culminating in the loss of the signaling pathways associated with vitamin A.

Another gene directly related to VD metabolism is CYP27B1. The gene product converts $25\text{-}OHD_3$ into 24,25-hydroxy vitamin D_3 and 1,25-hydroxy vitamin D_3 (Eyles et al., 2013). Cyp27b1 was found to be underexpressed in the murine pups exposed to ethanol (PSE-Network; Fig. 2B).

Another important gene found among the overexpressed genes in the LGEP-Network (Fig. 2D) is CDK11B (CDK11p58) (Supplementary Table S1). Remarkably, CDK11p58 promotes the inhibition of VDR through ubiquitin-proteasome-mediated degradation (Chi et al., 2009), indicating that ethanol could interfere with VD action by promoting the degradation of VDR. Consistent with that hypothesis, protein ubiquitination was present in Cluster 5, proteolysis involved in cellular protein catabolic process was present in Cluster 15 (Supplementary Table S2), and the proteolytic genes were among those overexpressed in the transcriptomic sets of the PE- and PSE-Networks.

Interplay between ethanol exposure, vitamin deficiency, and neuroinflammation. A major result of this systems chemo-biology analysis is that ethanol has been directly connected to the positive induction of inflammation (Clusters 3, 5, 6, 8, 10–11, and 14; Supplementary Table S2), especially in the overexpression of genes in murine adult individuals exposed to ethanol during embryogenesis (EGEP-Network; Fig. 2C; Supplementary Table S4). This result indicates that the induction of inflammation occurs during early gestation and extends through development into adult. Consistent with these data, VDR deletion was observed to reduce the activity of IκBα protein, a potent inhibitor of the inflammatory-associated transcriptional factor NFκB, (Wu et al., 2010). It is important to note that inflammatory insults during pregnancy have already been correlated with Alzheimer's and Parkinson's diseases (Miller and O'Callaghan, 2008). Indeed, the activation of NFκB in adults leads to amyloid-β accumulation due to the synthesis of leukotriene D4 in cortical neurons (Wang et al., 2013).

In addition, IκBα promotes neurodifferentiation by blocking self-renewal and by indirectly reducing the levels of repressor element silencing transcription factor (REST), an inhibitor of neurogenesis (Khoshnan and Patterson, 2012).

Supporting the finding that vitamin metabolism could be correlated with neuroprotection against inflammation promoted by IκBα, proteins associated with the process of fat-soluble vitamin metabolism are underexpressed in the EGEP-Network (Supplementary Table S4).

Corroborating the hypothesis that ethanol can promote neuroinflammation through VDR downregulation, IκBα was underexpressed in the overlaps between the PE- and PSE-Networks (Fig. 4; Table 3). VD and AT, as well as RA, are fat-soluble vitamins and may protect neurons against inflammation.

Vitamin deficiency driven by ethanol exposure leads to altered glutamate uptake. Glutamate has a major role as an excitatory neurotransmitter in the mammalian brain and is responsible for multiple aspects of neural activity, such as cognition and memory, which are both affected by FAS and FASD (O'Leary, 2004; Ruediger and Bolz, 2007). However, the overstimulation of glutamate can be responsible for brain injury and neuron apoptosis (Lu et al., 2013). The data gathered in this study showed that fetuses exposed to ethanol (EGEP-Network; Fig. 2C) early in development display an overstimulation of glutamine metabolism, which is a normal component controlling of glutamate levels in the central nervous system through the glutamine-glutamate cycle (Supplementary Table S4).

Glutamate also activates G-protein-coupled metabotropic receptors that can exert their effects through the cyclic

adenosine monophosphate (cAMP) pathway (Ruediger and Bolz, 2007), which is related to neurodevelopment and melatonin regulation (de Faria Poloni et al., 2011). Interestingly, the cAMP pathway is downregulated (Table 2) in fetuses exposed to ethanol during development (PE-Network; Fig. 2A) and also in pups postnatally exposed to ethanol (PSE-Networks; Fig. 2B). This indicates that ethanol exposure in the brain might have a negative effect on the cAMP pathway, thereby resulting in defects in G-protein-coupled receptor activity. To corroborate with our hypothesis, it was already observed that glutamate levels are abnormally high during a mice model of FAS (Karl et al., 1995).

Ascorbic acid (AC) is released into the extracellular space to protect neurons exposed to cytotoxic concentrations of glutamate (Lane and Lawen, 2013). AC is a water-soluble vitamin, and water-soluble vitamin metabolic processes were among the GOs observed in cluster 12 (Supplementary Table S2).

The results from fetuses exposed to ethanol during the late phase of development (LGEP-Network; Fig. 2D) indicated that the gene coding for nicotinamide-nucleotide adenylyltransferase (NMNAT2), an enzyme predominantly expressed in the brain and related to NADP biosynthesis, is underexpressed. This result is interesting, as it seems that nicotinamide [niacin (NC)] deficiency is already correlated to neuronal damage (Table 1). It should be pointed the NMNAT2 was among the overexpressed genes in the EGEP-Network, indicating that the ethanol-induced deficiency in NC may be more aggressive in later stages of development.

Conclusion

In summary, in our chemo-systems biology analysis, we prospected interactomes and combined the topological, GO and transcriptomic analyses of four ethanol-exposed groups of mice at different ages. The results gathered from this work helped to elucidate FAS development and its interaction with vitamin metabolism (Fig. 5). Ethanol appears to impair biological processes such as (i) the circadian cycle; (ii) calcium ion homeostasis; (iii) the glutamine pathway; (iv) the cAMP pathway; (v) inflammation; (vi) neuron differentiation; and (vii) synapse formation and plasticity. These processes appear to be closely related to vitamin metabolism, particularly for RA, VD, NC, and FA. Because ethanol is already correlated with vitamin deficiency, and vitamins are crucial for brain development, understanding the relationship between ethanol and vitamins appears to be essential for preventing

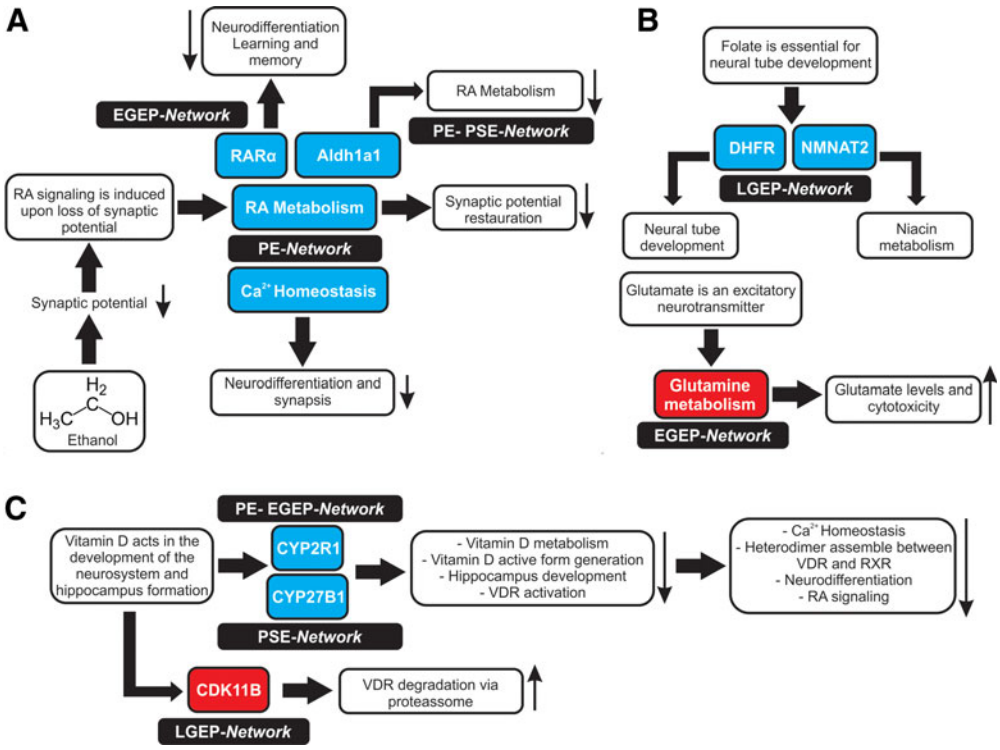

FIG. 5. Summary of the main findings of how ethanol may affect neurodevelopment in a FAS condition. The *blue rectangles* indicate a bioprocess in which the associated genes are found underegulated in the transcriptomic data and the *red rectangles* indicate those which are overexpressed. The *black rectangles* indicate the network where the bioprocess was observed. **(A)** summary of the action of ethanol in Ca^{2+} homeostasis and RA metabolism, resulting in the loss of synaptic potential and neurodifferentiation. **(B)** Summary of the observed negative effects of ethanol in folate and niacin metabolism. It also shows that ethanol leads to increased glutamate cytotoxicity. **(C)** action of ethanol in vitamin D metabolism. In this scenario, ethanol may provoke vitamin D receptor (VDR) degradation via proteasome. The model also indicates that vitamin D metabolism is negatively regulated by ethanol abuse, leading to loss of neurodifferentiation, and culminating in a negative regulation of both vitamin D and vitamin A signaling pathways.

the development of FAS and related outcomes. The targets selected in this work by data crossing and HB analysis among the different ethanol-exposed mice also generated important targets to be reviewed for FAS prevention and treatment because none of them had previously been correlated with FAS or FASD. Tubulin and tubulin-associated proteins, synapse plasticity proteins, and the proteins related to neurodifferentiation are of particular interest.

Acknowledgments

We would like to thank Msc. Kendi Nishino Miyamoto for helping with the statistical analyzes. This work was supported by research grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; grant no. 301149/2012-7), the Programa Institutos Nacionais de Ciência e Tecnologia (INCT de Processos Redox em Biomedicina-REDOXOMA; grant no. 573530/2008-4), Fundação de Amparo a Pesquisa do Rio Grande do Sul FAPERGS (PRONEM grant no. 11/2072-2), and the Programa Binacional de Terapia Celular–Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (PROBITEC-CAPES; grant no. 004/12).

Author Disclosure Statement

The authors declare that they have no conflicts of interest.

References

Ahn EY, Dekelver RC, Lo MC, et al. (2011). SON controls cell-cycle progression by coordinated regulation of RNA splicing. Mol Cell 42, 185–198.

Ahn J, Albanes D, Berndt SI, et al. (2009). Vitamin D-related genes, serum vitamin D concentrations and prostate cancer risk. Carcinogenesis 30, 769–776.

Bader GD, and Hogue CW. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform 4, 2.

Baksi SN, and Hughes MJ. (1982). Chronic vitamin D deficiency in the weanling rat alters catecholamine metabolism in the cortex. Brain Res 242, 387–390.

Beaudin AE, Abarinov EV, Noden DM, et al. (2011). Shmt1 and de novo thymidylate biosynthesis underlie folate-responsive neural tube defects in mice. Am J Clin Nutrition 93, 789–798.

Benjamini Y, and Hochberg Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. J Roy Stat Soc B Met 57, 289–300.

Binns D, Dimmer E, Huntley R, Barrell D, O'donovan C, and Apweiler R. (2009). QuickGO: A web-based tool for Gene Ontology searching. Bioinformatics 25, 3045–3046.

Bjorneboe GE, Bjorneboe A, Hagen BF, Morland J, and Drevon CA. (1987). Reduced hepatic alpha-tocopherol content after long-term administration of ethanol to rats. Biochim Biophys Acta 918, 236–241.

Brager AJ, Ruby CL, Prosser RA, and Glass JD. (2010). Chronic ethanol disrupts circadian photic entrainment and daily locomotor activity in the mouse. Alcoholism, Clin Exper Res34, 1266–1273.

Brodsky G, Barnes T, Bleskan J, Becker L, Cox M, and Patterson D. (1997). The human GARS-AIRS-GART gene encodes two proteins which are differentially expressed during human brain development and temporally overexpressed in cerebellum of individuals with Down syndrome. Human Mol Genet 6, 2043–2050.

Carbon S, Ireland A, Mungall CJ, et al. (2009). AmiGO: Online access to ontology and annotation data. Bioinformatics 25, 288–289.

Caspi R, Altman T, Dale JM, et al. (2010). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38, D473–479.

Castro MA, Filho JL, Dalmolin RJ, et al. (2009). ViaComplex: Software for landscape analysis of gene expression networks in genomic context. Bioinformatics 25, 1468–1469.

Chandra N, and Padiadpu J. (2013). Network approaches to drug discovery. Expert Opin Drug Disc 8, 7–20.

Chen L, Lau AG, and Sarti F. (2014). Synaptic retinoic acid signaling and homeostatic synaptic plasticity. Neuropharmacology 78, 3–12.

Chi Y, Hong Y, Zong H, et al. (2009). CDK11p58 represses vitamin D receptor-mediated transcriptional activation through promoting its ubiquitin-proteasome degradation. Biochem Biophys Res Commun 386, 493–498.

Craciunescu CN, Brown EC, Mar MH, Albright CD, Nadeau MR, and Zeisel SH. (2004). Folic acid deficiency during late gestation decreases progenitor cell proliferation and increases apoptosis in fetal mouse brain. J Nutrition 134, 162–166.

Csermely P, Korcsmaros T, Kiss HJ, London G, and Nussinov R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacol Therapeut 138, 333–408.

De Faria Poloni J, Feltes BC, and Bonatto D. (2011). Melatonin as a central molecule connecting neural development and calcium signaling. Funct Integrative Genom 11, 383–388.

De La Monte SM, and Kril JJ. (2014). Human alcohol-related neuropathology. Acta Neuropathol 127, 71–90.

Eyles DW, Burne TH, and Mcgrath JJ. (2013). Vitamin D, effects on brain development, adult brain function and the links between low levels of vitamin D and neuropsychiatric disease. Frontiers Neuroendocrinol 34, 47–64.

Feltes BC, Poloni Jde F, Notari DL, and Bonatto D. (2013). Toxicological effects of the different substances in tobacco smoke on human embryonic development by a systems chemo-biology approach. PloS One 8, e61743.

Genetta T, Lee BH, and Sola A. (2007). Low doses of ethanol and hypoxia administered together act synergistically to promote the death of cortical neurons. J Neurosci Res 85, 131–138.

Goez HR, Scott O, and Hasal S. (2011). Fetal exposure to alcohol, developmental brain anomaly, and vitamin A deficiency: A case report. J Child Neurol 26, 231–234.

Hewitt AJ, Knuff AL, Jefkins MJ, Collier CP, Reynolds JN, and Brien JF. (2011). Chronic ethanol exposure and folic acid supplementation: Fetal growth and folate status in the maternal and fetal guinea pig. Reprod Toxicol 31, 500–506.

Ichi S, Nakazaki H, Boshnjaku V, et al. (2012). Fetal neural tube stem cells from Pax3 mutant mice proliferate, differentiate, and form synaptic connections when stimulated with folic acid. Stem Cells Develop 21, 321–330.

Jaurena MB, Carri NG, Battiato NL, and Rovasio RA. (2011). Trophic and proliferative perturbations of in vivo/in vitro cephalic neural crest cells after ethanol exposure are prevented by neurotrophin 3. Neurotoxicol Teratol 33, 422–430.

Jensen LJ, Kuhn M, Stark M, et al. (2009). STRING 8—A global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37, D412–416.

Jupe S, Akkerman JW, Soranzo N, and Ouwehand WH. (2012). Reactome—A curated knowledgebase of biological pathways: megakaryocytes and platelets. J Thrombosis Haemostasis 10, 2399–2402.

Kanehisa M, and Goto S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27–30.

Karl PI, Kwun R, Slonim A, and Fisher SE. (1995). Ethanol elevates fetal serum glutamate levels in the rat. Alcohol Clin Exper Res 19, 177–181.

Kashyap V, and Gudas LJ. (2010). Epigenetic regulatory mechanisms distinguish retinoic acid-mediated transcriptional responses in stem cells and fibroblasts. J Biol Chem 285, 14534–14548.

Khoshnan A, and Patterson PH. (2012). Elevated IKKalpha accelerates the differentiation of human neuronal progenitor cells and induces MeCP2-dependent BDNF expression. PloS One 7, e41794.

Kuhn M, Szklarczyk D, Franceschini A, Von Mering C, Jensen LJ, and Bork P. (2012). STITCH 3: Zooming in on protein–chemical interactions. Nucleic Acids Res 40, D876–880.

Kurimoto K, Kuwasako K, Sandercock AM, et al. (2009). AU-rich RNA-binding induces changes in the quaternary structure of AUH. Proteins 75, 360–372.

Lane DJ, and Lawen A. (2013). The glutamate aspartate transporter (GLAST) mediates L-glutamate-stimulated ascorbate-release via swelling-activated anion channels in cultured neonatal rodent astrocytes. Cell Biochem Biophys 65, 107–119.

Lim CS, and Alkon DL. (2012). Protein kinase C stimulates HuD-mediated mRNA stability and protein expression of neurotrophic factors and enhances dendritic maturation of hippocampal neurons in culture. Hippocampus 22, 2303–2319.

Lopez-Fanarraga M, Carranza G, Bellido J, Kortazar D, Villegas JC, and Zabala JC. (2007). Tubulin cofactor B plays a role in the neuronal growth cone. J Neurochem 100, 1680–1687.

Lu XC, Dave JR, Chen Z, Cao Y, Liao Z, and Tortella FC. (2013). Nefiracetam attenuates post-ischemic nonconvulsive seizures in rats and protects neuronal cell death induced by veratridine and glutamate. Life Sci 92, 1055–1063.

Maere S, Heymans K, and Kuiper M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21, 3448–3449.

Maffi SK, Rathinam ML, Cherian PP, et al. (2008). Glutathione content as a potential mediator of the vulnerability of cultured fetal cortical neurons to ethanol-induced apoptosis. J Neurosci Res 86, 1064–1076.

Miller DB, and O'Callaghan JP. (2008). Do early-life insults contribute to the late-life development of Parkinson and Alzheimer diseases? Metab Clin Exper 57, S44–49.

Newman MEJ. (2005). A measure of betweenness centrality based on random walks. Soc Networks 27, 39–54.

Noelker C, Schwake M, Balzer-Geldsetzer M, et al. (2012). Differentially expressed gene profile in the 6-hydroxy-dopamine-induced cell culture model of Parkinson's disease. Neurosci Lett 507, 10–15.

Nomoto M, Takeda Y, Uchida S, et al. (2012). Dysfunction of the RAR/RXR signaling pathway in the forebrain impairs hippocampal memory and synaptic plasticity. Mol Brain 5, 8.

O'Leary CM. (2004). Fetal alcohol syndrome: Diagnosis, epidemiology, and developmental outcomes. J Paed Child Health 40, 2–7.

Oehlmann VD, Berger S, Sterner C, and Korsching SI. (2004). Zebrafish beta tubulin 1 expression is limited to the nervous system throughout development, and in the adult brain is restricted to a subset of proliferative regions. Gene Expression Patt 4, 191–198.

Park H, Yamada K, Kojo A, Sato S, Onozuka M, and Yamamoto T. (2009). Drebrin (developmentally regulated brain protein) is associated with axo-somatic synapses and neuronal gap junctions in rat mesencephalic trigeminal nucleus. Neurosci Lett 461, 95–99.

Qin L, and Crews FT. (2013). Focal thalamic degeneration from ethanol and thiamine deficiency is associated with neuroimmune gene induction, microglial activation, and lack of monocarboxylic acid transporters. Alcohol Clin Exper Res 38, 357–371.

Rebhan M, Chalifa-Caspi V, Prilusky J, and Lancet D. (1997). GeneCards: Integrating information about genes, proteins and diseases. Trends Genetics 13, 163.

Roehrs T, and Roth T. (2001). Sleep, sleepiness, sleep disorders and alcohol use and abuse. Sleep Med Rev 5, 287–297.

Rosado JO, Henriques JP, and Bonatto D. (2011). A systems pharmacology analysis of major chemotherapy combination regimens used in gastric cancer treatment: Predicting potential new protein targets and drugs. Curr Cancer Drug Targets 11, 849–869.

Ruediger T, and Bolz J. (2007). Neurotransmitters and the development of neuronal circuits. Adv Exper Med Biol 621, 104–115.

Russo E, Citraro R, Constanti A, and De Sarro G. (2012). The mTOR signaling pathway in the brain: Focus on epilepsy and epileptogenesis. Mol Neurobiol 46, 662–681.

Safran M, Dalah I, Alexander J, et al. (2010). GeneCards Version 3: The human gene integrator. Database J Biol Databases Curation 2010, baq020.

Scardoni G, Petterlini M, and Laudanna C. (2009). Analyzing biological network parameters with CentiScaPe. Bioinformatics 25, 2857–2859.

Schneider HC, and Klabunde T. (2013). Understanding drugs and diseases by systems biology? Bioorg Med Chem Lett 23, 1168–1176.

Shannon P, Markiel A, Ozier O, et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498–2504.

Singleton CK, and Martin PR. (2001). Molecular mechanisms of thiamine utilization. Curr Mol Med 1, 197–207.

Snel B, Lehmann G, Bork P, and Huynen MA. (2000). STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28, 3442–3444.

Strate I, Min TH, Iliev D, and Pera EM. (2009). Retinol dehydrogenase 10 is a feedback regulator of retinoic acid signalling during axis formation and patterning of the central nervous system. Development 136, 461–472.

Van Der Wulp NY, Hoving C, and De Vries H. (2013). A qualitative investigation of alcohol use advice during pregnancy: Experiences of Dutch midwives, pregnant women and their partners. Midwifery 29, e89–98.

Wang XY, Tang SS, Hu M, et al. (2013). Leukotriene D4 induces amyloid-beta generation via CysLT(1)R-mediated NF-kappaB pathways in primary neurons. Neurochem Intl 62, 340–347.

Wentzel P, and Eriksson UJ. (2009). Altered gene expression in neural crest cells exposed to ethanol in vitro. Brain Res 1305, S50–60.

Wu S, Xia Y, Liu X, and Sun J. (2010). Vitamin D receptor deletion leads to reduced level of IkappaBalpha protein through protein translation, protein–protein interaction, and post-translational modification. Intl J Biochem Cell Biol 42, 329–336.

Yu H, Kim PM, Sprecher E, Trifonov V, and Gerstein M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3, e59.

Zhou FC, Zhao Q, Liu Y, et al. (2011). Alteration of gene expression by alcohol exposure at early neurulation. BMC Genomics 12, 124.

Address correspondence to:
Diego Bonatto, PhD
Centro de Biotecnologia da UFRGS, Sala 219
Departamento de Biologia Molecular e Biotecnologia
Universidade Federal do Rio Grande do Sul (UFRGS)
Avenida Bento Gonçalves 9500, Prédio 43421
Porto Alegre 91509-900
Rio Grande do Sul
Brazil

E-mail: diegobonatto@gmail.com

Envelhecimento Molecular: Uma Série de Teorias para um Único Mecanismo

32
CAPÍTULO

Itamar José Guimarães Nunes
Diego Bonatto

O QUE É ENVELHECIMENTO?

Sabemos que todos os indivíduos da espécie humana envelhecem com o passar do tempo, e isso não é válido apenas para o *Homo sapiens*, mas para a maior parte dos seres vivos que temos conhecimento. Apesar disso, ainda pouco se sabe como funciona – ou mesmo o que é em si – o processo de envelhecimento em um organismo, sobretudo em nível molecular. Strehler, em 1977, tenta explicar tal fenômeno atribuindo-o quatro principais características[1]: (i) universalidade, pois todo processo associado ao envelhecimento em um determinado organismo ocorre em todos os demais indivíduos de sua espécie; (ii) intrinsecalidade, onde fatores externos (por exemplo, tabaco, poluição, entre outros) podem contribuir para a progressão do envelhecimento, mas não são a causa estrita de sua ocorrência; (iii) progressividade, pelo fato de que os fenótipos associados à idade não se manifestam de forma súbita, mas de forma gradual e cumulativa – embora estes fenótipos possam causar predisposição a eventos súbitos comuns em idosos como acidentes cerebrovasculares ou infarto agudo do miocárdio; e (iv) nocividade, na qual o organismo envelhecido vai perdendo suas capacidades fisiológicas e aumenta as chances do mesmo desenvolver certas doenças, o que está relacionado à elevada taxa de mortalidade de indivíduos idosos.

Além destes padrões, pode-se atribuir ao envelhecimento a característica de "multifatorialidade", pois este envolve mecanismos moleculares extremamente complexos que possuem causas ainda desconhecidas[2]. Desta maneira, torna-se refutável a hipótese de que apenas um único processo biológico contribui para todas as mudanças ocorrentes pela idade do organismo. Apesar disso, observou-se, a nível molecular, uma série de padrões que deram origem a hipóteses e teorias que tentam explicar o envelhecimento como consequência de um acúmulo de erros ou danos fisiológicos irreparáveis[3]. Entre essas teorias, estão incluídas as teorias de "Uso e desgaste", "Confiabilidade", "Telômeros", "Estresse oxidativo", "Dano de DNA", "*Inflammaging*" e "*DevAge*", nas quais estão resumidas na Tabela 32.1 e serão discutidas a seguir.

TEORIAS MOLECULARES E CELULARES

Teorias do uso e desgaste e da confiabilidade

O conceito filosófico e científico de envelhecimento vem sido determinado e reestruturado há mais de dois milênios, mesmo antes de Aristóteles[4]. A ideia que se tinha era puramente intuitiva: achava-se que envelhecimento era consequência do acúmulo de pequenos

Tabela 32.1 – Principais teorias citadas no capítulo que tentam explicar, de diferentes formas, como funciona o processo de envelhecimento

Nome (denominação original, em inglês)	Inferência	Principal(is) contribuinte(s) para o envelhecimento
Teoria do uso e desgaste (*Wear and tear theory*)	O organismo, por si só, tende a se deteriorar com o tempo, assim como ocorre em objetos inanimados	O desgaste frente às experiências com o ambiente externo
Teoria da confiabilidade (*Reliability theory*)	O organismo possui um número limitado de vias redundantes que se reduz conforme o avanço da idade, tornando o sistema vulnerável ao colapso de poucos mecanismos	Redundâncias (múltiplas rotas moleculares para uma mesma função)
Teoria dos telômeros (*Telomere theory*)	O encurtamento dos telômeros restringe as células a um número limitado de replicações	Ausência de telomerase e senescência
Teoria do estresse oxidativo (*Free-radical theory*)	O acúmulo de espécies reativas de oxigênio aumenta com a idade danifica as células gradualmente	Radicais livres derivados de oxigênio
Hipótese da restrição calórica (*Caloric restriction*)	Diminuir o consumo de calorias sem comprometer a quantidade de nutrientes pode melhorar a expectativa de vida	Ácidos graxos e glicose em excesso
Teoria do dano de DNA (*DNA damage theory*)	O acúmulo de danos irreparáveis no DNA geram mutações deletérias e declínio funcional às células	Ultravioleta, radicais livres e ineficiência do mecanismo de reparação
Hipótese do Inflammaging (*Inflammaging hypothesis*)	A manifestação de respostas inflamatórias aumenta com a idade, podendo induzir uma série de doenças degenerativas	Citocinas pró-inflamatórias e Imunossenescência
Hipótese do DevAge (*Developmental hypothesis*)	Os mecanismos moleculares necessários para o desenvolvimento embrionário podem ser os mesmos que contribuem para doenças associadas com o envelhecimento	Padrões transcrição por modificações epigenéticas

danos causados por eventos estocásticos que ocorriam ao longo da vida do organismo. Assim como a sola de um sapato pode se tornar suja e gasta conforme seu uso, ou uma lâmina pode enferrujar com sua exposição à umidade – ou mesmo ficar sem capacidade de corte se muito manuseada para tal – acreditava-se que o corpo também se deteriorava com o tempo, desgastando-se progressivamente até colapsar (falecer). É com essa observação que se definiu o que se denomina "Teoria do desgaste" (Figura 32.1).

Todavia, essa teoria acabou sendo, em parte, refutada com o progresso do conhecimento científico. Com ela, não se podia explicar por que diferentes organismos de um mesmo clado (tais como os mamíferos) e mesmo ambiente possuem expectativas de vida tão diferentes entre si. Além disso, o deterioramento de objetos como o sapato gasto e a lâmina cega é irreversível, sendo característico para objetos inanimados. Seres vivos, em contrapartida, possuem mecanismos de reparo e de renovação celular que podem reverter esses danos. Uma forte comprovação é

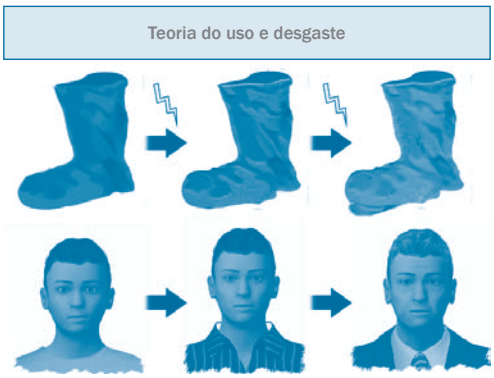

Figura 32.1 – Representação da teoria que retrata o "uso e desgaste" como causa do envelhecimento. Pela lógica deste princípio, assim como uma bota de couro pode se tornar desgastada por causa de seu uso recorrente em diversos ambientes, o ser humano também pode sofrer um "desgaste" semelhante, no qual pode ser responsável pelo fenótipo característico dos indivíduos idosos.

que exceções como as hidras de água doce[5], as medusas do gênero *Turritopsis*[6], algumas leveduras que se dividem simetricamente[7] e determinadas linhagens de células cancerígenas, quando em condições ideais de cres-

cimento, são potencialmente imortais (em termos biológicos).

Neste sentido, tentou-se estabelecer uma outra teoria que pudesse explicar o envelhecimento de forma geral, mas com um embasamento mais matemático e menos especulativo. Uma ideia recente, chamada "Teoria da confiabilidade", indica que um organismo é um sistema bioquímico que possui múltiplas vias necessárias para sua manutenção[8] (Figura 32.2). Dentre essas vias, existem rotas que são idênticas ("redundantes") para uma mesma função, onde a integridade de apenas uma dessas vias pode ser suficiente para manter o organismo vivo. Isto é, a falha total do sistema (por exemplo, o colapso de um mecanismo bioquímico) ocorre apenas se todas suas rotas moleculares redundantes forem obstruídas. Porém, com o passar do tempo, o organismo tende a acumular "falhas" aleatórias que obstruem irreversivelmente cada uma dessas vias redundantes, ao ponto que o indivíduo, já idoso, acaba dependendo de um único mecanismo intacto e sem redundâncias[8]. Como essa única via restante é um elemento fisiologicamente importante e insubstituível, se induzida a um novo erro, causará a falha completa do sistema – o que pode ser traduzido fisiologicamente por doenças associadas ao envelhecimento ou até mesmo o falecimento do indivíduo. E é por esse motivo que se observa uma taxa de mortalidade exponencialmente maior em indivíduos mais idosos.

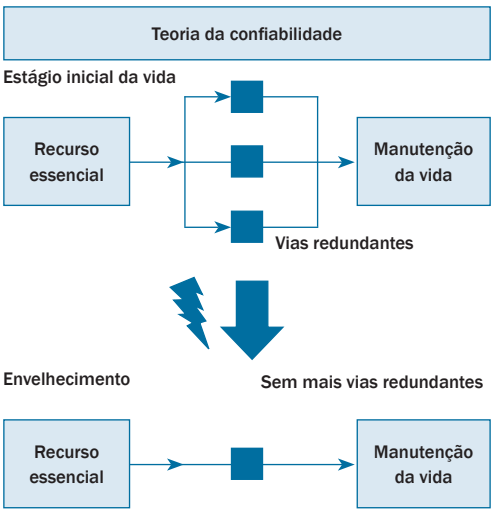

Figura 32.2 – Esquema que exemplifica um sistema de redundâncias e a sua degradação pelo envelhecimento. Diferentes mecanismos de mesma função (quadrados em verde) utilizam os recursos essenciais (por exemplo, nutrientes) para a manutenção fisiológica do organismo. Mesmo se a maior parte dessas vias forem perdidas, sobrando apenas uma, o sistema continuará funcionando, embora dependendo de apenas uma via. Desta forma, a confiabilidade na integridade do sistema diminui, pois se obstruíssemos esta única via restante, todo o sistema entraria em colapso. Assim, essa teoria sugere que, durante o envelhecimento, os sistemas de manutenção das funções fisiológicas possuem menos mecanismos redundantes.

Por fim, essas determinações dão margem a uma série de perguntas, como: que tipo de mudanças são acumuladas durante toda a vida e resultam em defeitos sistêmicos que culminam em um fenótipo envelhecido? E o que poderia causar essas mudanças? Para tentar ajudar a responder essas questões, algumas hipóteses foram formuladas em termos moleculares do envelhecimento.

Senescência celular e a teoria dos telômeros

Sem dúvida, qualquer estudante de biologia molecular já ouviu falar ou estará ciente, em algum momento, da existência dos telômeros e a sua importância para a célula. Um telômero nada mais é que uma sequência repetitiva de nucleotídeos (TTAGGG, em vertebrados) que se localiza na extremidade dos cromossomos de eucariotos[9,10]. Para estes organismos, a sua função principal está relacionada com o ciclo celular, permitindo que o genoma se replique integralmente e protegendo os cromossomos contra sua degradação ou sua fusão com outras moléculas de DNA. Durante a replicação, como as enzimas envolvidas não são capazes de duplicar toda a fita de DNA até sua extremidade, cada divisão celular acarreta no encurtamento do cromossomo a partir de suas pontas. Neste caso, os telômeros, que estão nessas pontas, são degradados, assim protegendo todos os genes que poderiam ser importantes para a manutenção da célula. Assim, após seu encurtamento, uma enzima denominada "telomerase" pode reconstituir a cadeia telomérica[9,11].

Essa função do telômero oferece um benefício momentâneo para a sobrevivência da célula, mas tem um ponto fraco a longo prazo. A grande maioria das células do nosso organismo não produz a telomerase, e esta acaba sendo

exclusiva para células germinativas, células-tronco embrionárias e alguns outros tipos de células-tronco e leucócitos[9]. Devido a isso, após um certo número de replicações, a maioria das células acaba perdendo completamente seus telômeros e herdando cromossomos defeituosos. Isso repercute em um processo chamado "senescência celular", que é a incapacidade destas células de se dividirem novamente – fazendo muitas delas também sofrerem apoptose[9].

Neste sentido, a "Teoria dos telômeros" pressupõe que o envelhecimento é causado principalmente pela replicação limitada e pela senescência das células somáticas devido ao encurtamento telomérico[12] (Figura 32.3). Curiosamente, pode-se dizer – de forma especulativa – que essa postulagem também é fiel à teoria da confiabilidade, pois presume que cada repetição de telômeros forma um mecanismo redundante para a integridade cromossomal, colapsando o sistema apenas após sua degradação completa.

Como suporte à teoria, observou-se em diversos estudos que, de fato, indivíduos com cadeias de telômeros maiores tendem a possuir uma maior expectativa de vida, enquanto os que manifestam doenças que causam encurtamento de telômeros (como disqueratose congênita) têm maior taxa de mortalidade com a idade[13]. As células do tipo fibroblastos são "imortalizadas" por métodos que induzem a atividade da telomerase, de forma que a mesma se prolifera indefinidamente em meios de cultura ideais[9,11]. Todavia, apesar de ser uma abordagem científica muito popular, controlar a manutenção telomérica como um método anti-envelhecimento ainda é um dos maiores desafios para os pesquisadores de gerontologia. Não são todas as células que são imortalizadas apenas com a presença de telomerase, e a eficácia deste método geralmente só pode ser obtida se os mecanismos de parada do ciclo celular forem inativados[9]. O problema é que tais mecanismos evitam a progressão de tumores, e seu silenciamento pode aumentar significativamente o risco de câncer. É por isso que certos tipos de células de tumores são por vezes considerados "imortais".

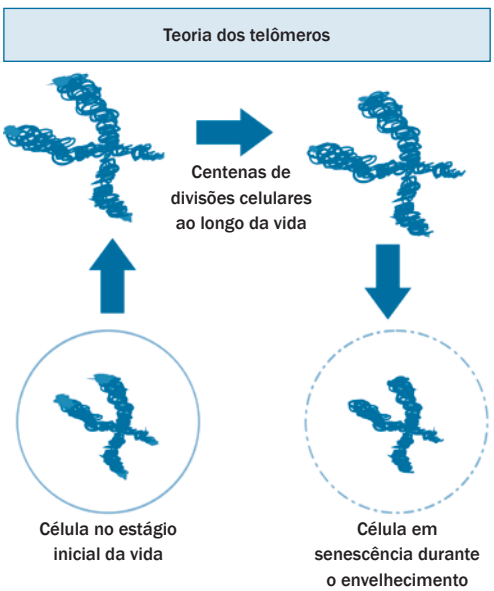

Figura 32.3 – Representação resumida da teoria dos telômeros. A cada divisão celular, os cromossomos (emaranhados em preto) não conseguem replicar totalmente o DNA e acabam perdendo parte dos telômeros (em vermelho) na sua extremidade. Após um certo número de divisões celulares, porém, esses telômeros se esgotam e o material genético importante da célula acaba sendo comprometido. Neste ponto, a célula entra em senescência e não realiza mais replicações.

Contudo, embora a senescência esteja associada com o avanço da idade, ainda não se pode determinar o encurtamento dos telômeros como a causa única do envelhecimento. Um exemplo que rejeita a universalidade desta teoria foi visto em um experimento com camundongos, onde a produção de telomerase foi induzida por terapia gênica sem promover câncer[14]. Como resultado deste estudo, o tempo de vida dos roedores aumentou em até 24%, mas não preveniu o envelhecimento por completo. Por isso, é bastante provável que outros mecanismos associados com o envelhecimento também estejam envolvidos.

Teoria do estresse oxidativo e suas derivações

Na década de 1950, Gerschman e Harman observaram, em seus respectivos estudos, que as espécies reativas de oxigênio (ou "ERO") geravam efeitos fisiológicos nocivos às células, sendo semelhantes a uma exposição contínua à radiação por raios X[15,16]. ERO são moléculas altamente reativas que contêm oxigênio em sua composição, podendo reagir com enzimas, lipídios, moléculas de DNA, entre outros componentes celulares[17]. Dentro desse grupo temos os "radicais livres", nos

quais são formados pela redução de oxigênio molecular (adição de um elétron em O_2), a qual dá origem a moléculas como ânion superóxido ($O_2^{\bullet -}$) e radicais hidroxila ($OH^{\bullet}$). Outros tipos de ROS incluem o peróxido de hidrogênio (H_2O_2), que é formado pela catálise de $O_2^{\bullet -}$ pela enzima superóxido dismutase (SOD) – neste caso, não ele é tão reativo quanto os radicais livres. Harman, sabendo que estas moléculas altamente reativas são produzidas normalmente no organismo, conectou isso com a ideia já postulada de que a irradiação em seres vivos, na qual causa liberação de ERO, é responsável por um maior índice de mutações e câncer e, talvez, pelo próprio envelhecimento[16]. Isso deu origem à "Teoria dos radicais livres" (do inglês, *Free--radical theory of aging*), na qual indica que o envelhecimento pode ser definido por danos crônicos e irreparáveis causados pelo estresse oxidativo de radicais livres (Figura 32.4). Posteriormente, houve a inclusão de ERO não radicalares como H_2O_2[18], portanto podemos

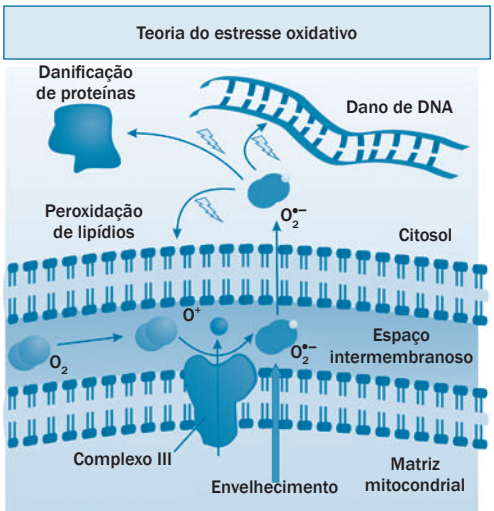

Figura 32.4 – Esquema de formação do radical ânion superóxido ($O_2^{\bullet -}$) durante a fosforilação oxidativa. Apesar da representação, apenas uma pequena parte do oxigênio molecular se torna $O_2^{\bullet -}$ durante a respiração celular, e os radicais que se formam, ainda assim, podem ser convertidos para peróxido de hidrogênio (H_2O_2). Contudo, a quantidade de espécies reativas de oxigênio (ROS) se acumula durante o envelhecimento – possivelmente ao ponto de os mecanismos antioxidantes não conseguirem controlar os níveis de ROS. Com o aumento do estresse oxidativo, vários componentes celulares podem ser prejudicados, o que acarreta no declínio do funcionamento da célula ou mesmo em morte celular por apoptose.

também, designar essa teoria simplesmente por "Teoria do estresse oxidativo".

Esse postulado foi um dos mais influentes e debatidos no campo de estudo do envelhecimento desde as últimas décadas, abrindo portas para uma série de novas teorias. Posteriormente, outros trabalhos indicaram que a mitocôndria poderia ser uma organela-chave para a causa dos danos celulares associados ao envelhecimento, pois é uma fonte importante de ERO intracelular[19]. Sabe-se que certos radicais livres, como $O_2^{\bullet -}$, podem ser gerados pela própria respiração celular na mitocôndria – mais precisamente por fosforilação oxidativa[10] – e estima-se que o DNA mitocondrial é 10 a 20 vezes mais suscetível a danos por oxidação do que o DNA nuclear[9,20]. Além disso, o avanço da idade do organismo também está relacionado com maiores taxas de degradação e de mutações no DNA mitocondrial, o que resulta no declínio funcional da mitocôndria. Essa observação foi o alicerce para a ideia de que a mitocôndria poderia ser a principal contribuinte para o acúmulo de estresse oxidativo durante o envelhecimento, e essa suposição é conhecida como "Teoria do envelhecimento mitocondrial"[19,21].

Em resposta a isso, procurou-se diferentes formas de equilibrar os níveis de ERO intracelular, principalmente de origem mitocondrial. Infelizmente, observou-se que a suplementação com agentes antioxidantes não acarretou efeitos positivos para a expectativa de vida[22]. Porém, foi visto que uma dieta de baixas calorias (incluindo um menor consumo de ácidos graxos ou glicose) e boa qualidade nutricional ofereceu resultados positivos em favor dos métodos anti-envelhecimento. Essa dieta, denominada "restrição calórica" (RC), resultou em um maior tempo médio de vida para organismos-modelo como *Saccharomyces cerevisiae*, *Caenorhabditis elegans* e *Drosophila melanogaster*[23]. Assim, supôs-se que a ingestão calórica estivesse correlacionada com a atividade oxidativa da mitocôndria. Entretanto, os mecanismos que explicam os efeitos da RC ainda permanecem muito debatidos e pouco entendidos, e existem trabalhos que sugerem resultados opostos entre si – tanto defendendo o aumento quanto a diminuição na produção de ERO pela mitocôndria em resposta à RC[22]. Para tornar ainda mais confuso, viu-se que diminuir ou inibir a ativi-

dade respiratória mitocondrial por intervenções genéticas – mimetizando a RC – pode gerar efeitos distintos dependendo do clado do organismo[22]. Tais desregulações estendem o tempo de vida de *C. elegans*, porém causam certas patologias relacionadas com o envelhecimento em vertebrados.

Deste modo, apesar das muitas validações a favor da teoria do estresse oxidativo, também existem uma série de dúvidas e controvérsias a respeito do efeito de ERO sobre o envelhecimento. Quando se confirmou que o envelhecimento pode estar vinculado diretamente com maior dano por estresse oxidativo e menor função mitocondrial[24], esperava-se que, ao aumentar a atividade antioxidante (e.g. por meio de dietas ou alterações genéticas), a expectativa de vida aumentasse – pois tal experimento havia sido efetivo em invertebrados como *D. melanogaster*. Contudo, para vertebrados, o resultado não foi tão satisfatório: em camundongos, por exemplo, um conjunto de trabalhos recentes apresentou pouca ou nenhuma alteração significante na idade máxima com dietas antioxidantes[24,25]. Por esse motivo, ainda é muito cedo para estabelecer se ERO são a causa ou a consequência do processo de envelhecimento, ainda que, para o melhor entendimento deste, a teoria do estresse oxidativo seja promissora.

Teoria do dano de DNA

O nosso genoma contém genes responsáveis pela geração de proteínas que atuam amplamente na manutenção das funções fisiológicas do nosso organismo. Uma pequena alteração no nosso material genético e podemos desenvolver desde nenhuma mudança fenotípica aparente até um quadro patológico letal. O problema é que vários fatores como estresse oxidativo, radiação ionizante ou UV, calor excessivo e certos fármacos podem resultar no dano químico do DNA. Isto inclui a formação de lesões, substituição de nucleotídeos, dímeros de pirimidina, entre outras consequências[26]. Para evitar isso, existem uma série de mecanismos de reparo e conservação genética que evitam certas mudanças no DNA das células – afinal, é mais provável que uma alteração química em um gene seja nociva do que ofereça algum benefício fisiológico. Se a mudança for pequena, esta é reconhecida por enzimas responsáveis pelos mecanismos de reparação de danos por excisão de nucleotídeos (NER, do inglês *nucleotide excision repair*) ou de base (BER, do inglês *base excision repair*) ou por mal pareamento de bases (MMR, do inglês *DNA mismatch repair*). Se a mudança for suficientemente brusca, a célula pode entrar em morte celular programada (apoptose), impedindo que essa danificação se perpetue em suas replicações posteriores como uma mutação ou mesmo como a causa de um câncer[27].

Contudo, sabemos que esses mecanismos não são perfeitos. Um exemplo disso é que, caso ocorra um dano muito severo que quebre de ambas as fitas do DNA em uma mesma região, alguns mecanismos de reparação podem acabar não sendo eficientes na hora de restaurarem a sequência nucleotídica da molécula por não terem um referencial com a sequência correta – podendo gerar alterações na sequência que não necessariamente serão eliminadas[27,28]. Outro exemplo é que os mecanismos de reparo de DNA mitocondrial (mtDNA) podem não ser tão eficientes quanto os de DNA nuclear para certos danos, como dímeros de pirimidina[26]. Isso significa que, durante a vida inteira, o nosso material genético pode acumular diversos danos nos quais não necessariamente serão reconhecidos pelos nossos mecanismos de reparação, e isso pode afetar drasticamente a expressão gênica. A teoria que associa esses danos com o detrimento fisiológico e a progressão do envelhecimento é denominada como "Teoria do dano de DNA"[3] (Figura 32.5).

Uma das evidências que apoiam essa teoria é que a ocorrência de alguma mutação nos genes envolvidos na via NER – que causam a perda de função das enzimas desta via – resulta em doenças que diminuem drasticamente a expectativa de vida do indivíduo, como Xerodermia pigmentosa (XP) e síndrome de Cockaine[29]. Além de serem responsáveis por um envelhecimento prematuro, esses quadros patológicos também aumentam drasticamente a propensão de morte celular, distúrbios neurológicos e manifestação de diferentes tipos de câncer. Nos EUA, estima-se que 30% dos pacientes com XP acabam falecendo antes dos 32 anos[29].

Curiosamente, a teoria do dano de DNA parece contribuir para a teoria do estresse oxidativo. Alguns estudos com ratos indi-

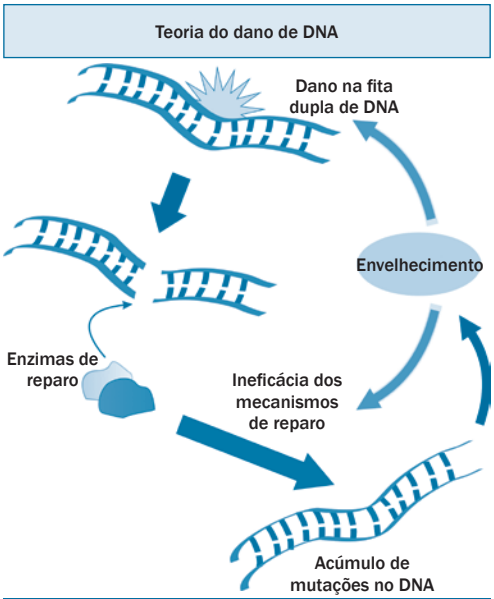

Figura 32.5 – Uma das representações gráficas da teoria do dano de DNA, na qual designa o envelhecimento como um mecanismo resultante de um acúmulo de mutações ocorrentes pela vida inteira. Como demonstrado, esse acúmulo pode ser resultante possibilidade de erros ou de ineficácia do sistema de reparação de DNA. Os danos que contribuem para tais mutações variam, podendo ser radiação UV, estresse oxidativo, compostos mutagênicos, entre outros.

caram que a quantidade de danos de DNA nuclear e mitocondrial causados por ERO é significativamente maior no fígado, nos rins e no intestino durante o envelhecimento normal[30]. Em humanos, esse mesmo tipo de dano também aumenta no cérebro com a idade e parece ser mais abundante no DNA mitocondrial – o que pode dar suporte à teoria do envelhecimento mitocondrial.

Doenças associadas ao envelhecimento e a hipótese do *Inflammaging*

Embora popularmente seja comum ouvirmos a expressão "morrer de velho" como causa do falecimento de uma pessoa idosa, em termos de idade, sempre existe uma ou mais doenças envolvidas com as condições que levaram o organismo a óbito. Na verdade, o envelhecimento nada mais é do que o contribuinte universal para a maioria das patologias associadas à idade, mas cada distúrbio fisiológico tem sua própria causa e característica. Portanto, para se tentar entender o envelhecimento e formular diferentes hipóteses que procuram prever o seu funcionamento, o primeiro passo é buscar padrões moleculares encontrados em cada uma dessas doenças.

Um dos padrões observados na maioria das doenças envolvidas com a idade, senão em todas, é a presença de algum elemento associado ao sistema imunológico ou à inflamação[31–33]. Há menos de duas décadas, viu-se que a maioria desses elementos são principalmente proteínas ou peptídeos que promovem a sinalização ativadora da cascata inflamatória – também conhecidos como "citocinas pró-inflamatórias". Estas moléculas são secretadas por leucócitos durante uma resposta imunológica – por exemplo, no caso de uma infecção gerada pelo corte com uma faca enferrujada – e têm a função de propagar a cascata inflamatória no tecido local, entrando em contato com outras células vizinhas[34]. Com o estímulo das citocinas, estas células podem ainda produzir mais agentes pró-inflamatórios.

Em uma condição denominada "inflamação aguda", agentes anti-inflamatórios são ativados após a remoção da infecção para promover o retorno da homeostase no tecido inflamado. Contudo, existem condições patológicas ou infecciosas em que, por causas obscuras, essa cascata imunológica acaba persistindo de forma gradual e cumulativa – gerando o que é conhecido como "inflamação crônica"[34]. Em vários quadros patológicos, como doença de Alzheimer, Mal de Parkinson, aterosclerose, artrite reumatoide, diabetes tipo II, sarcopenia, entre outras doenças, foi evidenciado um aumento gradual e cumulativo na produção e circulação sanguínea de citocinas pró-inflamatórias conforme o agravamento da doença[33]. Essas moléculas sinalizadoras também podem elevar o risco de desenvolvimento de câncer, pois estimulam a proliferação e a capacidade invasiva das células, assim como a angiogênese[33]. É importante ressaltar que todas essas doenças citadas acima têm alguma relação com o envelhecimento e são muito recorrentes em idosos[31–33].

Curiosamente, mesmo em idosos saudáveis, é visto que a quantidade de citocinas pró-inflamatórias na circulação sanguínea é maior se compararmos com indivíduos mais jovens. Além disso, foi evidenciada uma correlação entre a maior taxa de mortalidade

com o avanço da idade e um aumento sutil da inflamação crônica[31]. Sabendo desta associação, uma nova teoria denominada "Teoria do *Inflammaging*" (do inglês, *"Inflammation hypothesis of aging"*) foi proposta defendendo que o envelhecimento pode ser causado pelo aumento gradual da manifestação de respostas imunológicas ou inflamatórias crônicas, nas quais aumentam o risco de doenças e mortalidade associadas à idade (Figura 32.6). Seu principal fundamento é que a circulação das citocinas pró-inflamatórias aumenta no sangue durante o envelhecimento normal e possui ainda mais contraste em doenças da idade. Além disso, após dez anos de análises com amostras de sangue, um estudo levou a crer que duas citocinas pró-inflamatórias, IL-6 e TNFRSF1A, pudessem ser as proteínas com maior associação à mortalidade humana causada por qualquer tipo de causa patológica, contribuindo fortemente com essa teoria[35].

Entretanto, ainda não se sabe exatamente o que causa essa mudança no comportamento do sistema imunológico. Em uma revisão recente, um dos autores da teoria do *Inflammaging* propõe diferentes possibilidades para a manifestação de inflamação crônica[31]. Entre essas hipóteses, temos: (i) o acúmulo de "lixo celular" (i.e. macromoléculas, organelas ou células danificadas) devido à ineficiência de sua eliminação ou à maior recorrência de danos químicos, pois parte desse material pode

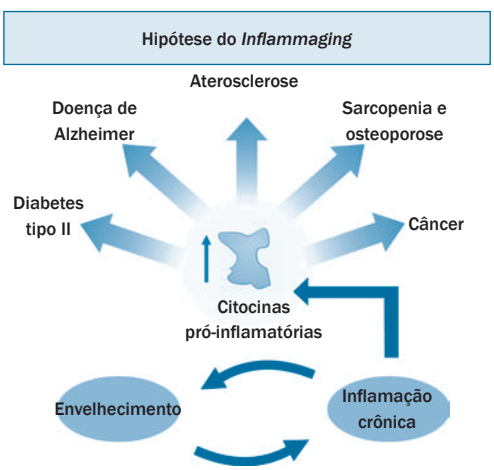

Figura 32.6 – Esquema exemplificando doenças degenerativas que estão associadas com ambos envelhecimento e inflamação crônica, se acordo com a hipótese do *inflammaging*.

mimetizar componentes bacterianos e ativar o sistema imunológico; (ii) a perda da capacidade de controlar a flora microbiana do intestino, na qual secreta constituintes microbianos que acabam escoando para a corrente sanguínea e induzindo inflamação; (iii) danos persistentes nas mitocôndrias, que ao serem lisadas, por terem características procariotas, liberam DNA mitocondrial e peptídeos que são reconhecidos pela célula como patógenos e ativam vias de resposta inflamatória; (iv) a senescência celular e a falta de um mecanismo de eliminação de células senescentes, pois estas se acumulam durante o envelhecimento e acabam secretando grandes quantidades de citocinas pró-inflamatórias; e (v) o declínio do sistema imunológico adaptativo, também conhecido como "imunossenescência", que pode ser gerado por uma mudança intrínseca da função dos leucócitos devido à exposição contínua aos certos antígenos ou outros agentes estressores[31,32].

Neste último cenário, temos dois tipos de células imunológicas: as células T imaturas ("virgens"), que têm a função de reconhecer novos tipos de antígenos – os quais promovem a maturação e a diferenciação destas células; e as células T de memória, que já maturaram e atuam defendendo o organismo especificamente contra os antígenos que reconheceram. A imunossenescência, neste sentido, ocorre pela exposição contínua a poucos tipos de antígenos durante a vida toda, o que culmina às células T virgens a se especializarem a este grupo pequeno de estressores. Assim, propõe-se que, durante o envelhecimento, boa parte das células T virgens acabaram sendo diferenciadas para células T de memória – que além de mais reativas à cascata inflamatória, não são eficazes no reconhecimento de novos patógenos. Isso também pode ser entendido como uma "ocupação excessiva do espaço imunológico"[32].

Neste contexto, o sistema imunitário pode contribuir significativamente para o entendimento do envelhecimento. Em contrapartida, o *inflammaging* ainda é uma hipótese nova e não completamente consolidada, visto que ainda não se sabe tudo sobre o próprio sistema imunológico em si. Por exemplo, não se pode aplicar a mesma ideia de imunossenescência, que envolve células T, para o envelhecimento de invertebrados ou de outros organismos

que não possuem uma imunidade adaptativa consolidada[34,36]. Além disso, mesmo em mamíferos, os mecanismos inflamatórios são indispensáveis para defesa do organismo a patógenos e toxinas e, como será visto na seção a seguir, exercem um papel importante durante o desenvolvimento embrionário. Por estas e outras razões, ainda que promissora, a postulagem do *inflammaging* ainda permanece como uma hipótese a ser investigada.

Antagonismo pleiotrópico e a hipótese do DevAge

Diferente de uma máquina ou um sistema de computadores, que são compostos por um conjunto de peças conhecidas e desenvolvidas para suas respectivas funções, o nosso organismo funciona por meio de mecanismos extremamente complexos e muitas vezes imprevisíveis. Um exemplo de sua complexidade é que muitas das proteínas que o compõem podem ser "pleiotrópicas" – isto é, participam de mais de um processo biológico ou podem ser responsáveis por mais de um fenótipo.

Muitas vezes, essas múltiplas funções podem servir para um mesmo fim. A citocina IL-1, por exemplo, é responsável por promover diferentes mecanismos de resposta inflamatória com o único objetivo de eliminar patógenos – o que inclui a ativação de leucócitos, produção de coaguladores sanguíneos e até mesmo indução de febre[34]. Mas, além desse sinergismo, as proteínas ainda podem possuir diferentes funções com efeitos contrários entre si, em uma característica conhecida como "pleiotropia antagônica". Geralmente, esse caráter é aplicado a proteínas que possuem consequências benéficas ao organismo em uma condição, mas nocivas em outra. Um exemplo deste antagonismo está na proteína p53. Embora esta seja responsável pela supressão do ciclo celular na presença de danos de DNA, sendo essencial para a prevenção de câncer, a p53 também pode reprimir a proliferação de células-tronco, impedindo a regeneração dos tecidos[37].

A noção de antagonismo pleiotrópico foi originada a partir da "Teoria da pleiotropia antagônica" postulada por Willians em 1957[38]. Para ele, o envelhecimento pode ter uma consequência evolutiva: muitos genes são selecionados e conservados para as próximas gerações porque eles proporcionam vantagens durante o período reprodutivo do indivíduo. Acontece que, após esse período de maturidade sexual, a pressão da seleção natural diminui – provavelmente por causa da menor taxa reprodutiva e maior taxa de mortalidade durante toda evolução. Por isso, foi visto que muitos dos genes que conferem vantagens durante o período de desenvolvimento e maturidade sexual, antagonicamente, podem ser responsáveis por fenótipos ou patologias associados com o avanço da idade pós-reprodutiva.

Espécies reativas de oxigênio e citocinas pró-inflamatórias, por exemplo, fazem parte deste cenário de contrariedade biológica. Mais recentemente, associou-se o antagonismo pleiotrópico às teorias de estresse oxidativo e do *inflammaging*, além de outros fatores como a epigenética. Isso porque o oxigênio, responsável pela produção ERO, também é essencial para regular os mecanismos epigenéticos e de ciclo celular durante o desenvolvimento embrionário[39]. Citocinas pró-inflamatórias, embora prejudiciais durante a inflamação crônica, também são necessárias para a formação do nicho de células-tronco durante o desenvolvimento embrionário[39]. Esta e outras evidências serviram de fundamento para a formação de uma nova teoria denominada "Teoria do DevAge" (do inglês, *Development hypothesis of aging*), na qual prevê que os mesmos mecanismos responsáveis pelo envelhecimento podem ser os mesmos necessários para o desenvolvimento embrionário e pós-embrionário. Além disso, ela também propõe que o superestímulo ou inibição de certos mecanismos essenciais durante o desenvolvimento fetal (como metabolismo de glicose, níveis de oxigênio molecular, entre outros) pode afetar nocivamente o crescimento do feto e culminar em quadros patológicos associados com o envelhecimento[39] (Figura 32.7).

Estima-se que pelo menos quatro fatores biológicos do desenvolvimento contribuam para a determinação do fenótipo durante o envelhecimento[39,40], que são: (i) oxigênio molecular, no qual é necessário para a respiração e a diferenciação tecidual, além de estar envolvido em uma remodelagem da cromatina que se perpetua até o envelhecimento e pode contribuir para certas doenças se presente em grandes

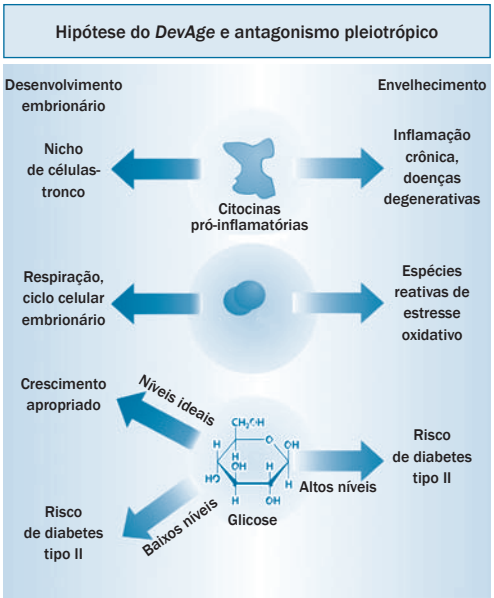

Figura 32.7 – Representação da hipótese do *DevAge* que exemplifica sua associação com a teoria da pleiotropia antagônica. Embora tenha-se determinado, em outras teorias, que agentes oxidantes e pró-inflamatórios são nocivos para o organismo adulto e contribuem para o envelhecimento, estes demonstram papéis cruciais durante o desenvolvimento embrionário. Além disso, se suas funções forem, de alguma forma, obstruídas ou alteradas em sua quantidade, é possível que doenças associadas com o envelhecimento se manifestem em estágios posteriores[39]. Adicionalmente, no caso da glicose, temos um curioso antagonismo, pois tanto baixos níveis de glicose em um embrião como altos níveis da mesma em um indivíduo adulto podem aumentar a propensão de diabetes tipo II (o tipo não dependente de insulina)[39].

quantidades; (ii) citocinas pró-inflamatórias, que podem contribuir, por exemplo, para a fixação do trofoblasto – cujo processo, se em falta, pode restringir o crescimento do feto e acarretar a este um envelhecimento prematuro[40,41]; (iii) glicose, que é responsável por uma série de etapas no desenvolvimento embrionário, mas que em quantidades muito diferentes do ideal, pode causar danificações severas nos órgãos do feto e propensão a diabetes tipo II; e (iv) modificações epigenéticas, onde padrões de metilação (adição de metila de histonas) podem mudar conforme a nutrição ou a ocorrência de estresses no útero durante o desenvolvimento embrionário, sendo associados com uma série de doenças durante o envelhecimento[39].

Adicionalmente, teorias como a do dano de DNA também podem estar associadas com antagonismo pleiotrópico e DevAge. Uma especulação que exemplifica essa ligação indica que danificações no DNA durante o período de desenvolvimento embrionário podem induzir padrões de expressão que acarretam ao maior acúmulo de danos no material genético durante o período pós-natal[42]. Outra especulação parte do fato de que existem uma série de mecanismos de reparo para cada tipo de dano de DNA, onde cada um requer um certo tempo para sua função. Neste caso, tal hipótese infere que as vias de reparo que consomem menos tempo (i.e. não

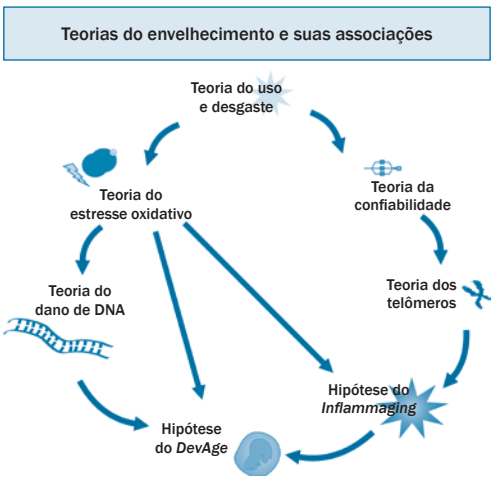

Figura 32.8 – Modelo representativo das possíveis conexões entre as teorias descritas na Tabela 32.1. Como discutido neste capítulo, a teoria da confiabilidade pode ser uma explicação mais matemática e molecularmente plausível da teoria do desgaste. Ela pode ser base para outras teorias, como a dos telômeros, como discutido no capítulo. A teoria do estresse oxidativo também se encaixa perfeitamente na teoria do desgaste, pois o acúmulo de radicais livres é responsável pela degradação (ou "desgaste") dos processos fisiológicos do organismo. Espécies reativas de oxigênio (ERO) também têm relação direta com a teoria do dano de DNA porque radicais livres são responsáveis pela danificação do material genético. Além disso, o oxigênio tem relação com DevAge pelos seus benefícios durante o desenvolvimento embrionário e seus efeitos nocivos durante o envelhecimento. ERO, por sua vez, também possui relação com *inflammaging* – um exemplo disso é que eles são liberados por células T citotóxicas para eliminar os patógenos durante uma inflamação[34]. Por fim, uma outra teoria que pode ter relação com *inflammaging* é a dos telômeros, pois como mencionado anteriormente, células senescentes – incluindo as com os telômeros desintegrados – tendem a liberar mais citocinas pró-inflamatórias. Todas essas ligações indicam que o envelhecimento não é um processo sozinho, mas um conjunto de processos que resultam em um mesmo fim.

necessitam de replicação cromossômica) são priorizadas na fase inicial da vida, pois permitem um desenvolvimento e maturação mais rápidos e, possivelmente, oferecem uma vantagem competitiva[43]. O problema é que essas vias são mais propensas a erros e podem causar um acúmulo de danos que pode trazer consequências nocivas às fases posteriores da idade[43]. Entretanto, ambas as hipóteses ainda necessitam ser exploradas.

Em resumo, a hipótese do DevAge não se restringe a tentar entender os mecanismos que danificam gradualmente o funcionamento do organismo e que induzem um fenótipo de envelhecido. Mais do que isso, ela tem como objetivo conectar elementos de teorias já existentes para explicar por que um processo que é benéfico no início da vida pode ser nocivo em estágios posteriores.

CONCLUSÃO

O envelhecimento é um processo complexo, caracterizado pela interconectividade de seus componentes moleculares. Visto que nem mesmo temos conhecimento de todas as possibilidades bioquímicas de um organismo normal, não é surpresa que saibamos ainda menos sobre como e por que um organismo envelhece.

Assim, é improvável que apenas uma das teorias moleculares explica, de forma satisfatória, todas as facetas do envelhecimento. Pelo contrário, pode ser mais plausível considerar que todas elas estão parcialmente corretas e que apenas a integração destes mecanismos usando ferramentas como a Biologia de Sistemas e outras áreas integrativas do conhecimento poderiam fornecer um panorama geral dos mecanismos de envelhecimento. Torna-se urgente que, cada vez mais, visões integrativas de mecanismos moleculares sejam aplicadas para a compreensão do envelhecimento e as suas consequencias fisiológicas.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Bernard Louis Strehler. Time, Cells and Ageing. New York, USA: Academic Press; 1977.
2. Olson CB. A review of why and how we age: a defense of multifactorial aging. Mech Ageing Dev. 1987;41(1-2):1–28.
3. Yin D, Chen K. The essential mechanisms of aging: Irreparable damage accumulation of biochemical side-reactions. Exp Gerontol. 2005;40:455–65.
4. Grant R. Concepts of aging: an historical review. Perspect Biol Med. 1963;
5. Martínez DE. Mortality patterns suggest lack of senescence in hydra. Exp Gerontol. 1998;33(3):217–25.
6. Miglietta MP, Piraino S, Kubota S, Schuchert P. Species in the genus Turritopsis (Cnidaria, Hydrozoa): A molecular evaluation. J Zool Syst Evol Res. 2007;45(1):11–9.
7. Coelho M, Dereli A, Haese A, Kühn S, Malinovska L, Desantis ME, et al. Fission yeast does not age under favorable conditions, but does so after stress. Curr Biol. 2013;23(19):1844–52.
8. Gavrilov LA, Gavrilova NS. The reliability theory of aging and longevity. J Theor Biol. 2001;213(4):527–45.
9. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Artmed, editor. Biochemistry and Molecular Biology Education. 2008.
10. Voet D, Judith G. Voet. Biochemistry, 4th Edition. 2010.
11. Holt SE, Shay JW, Wright WE. Refining the telomere-telomerase hypothesis of aging and cancer. Nat Biotechnol. 1996;14(7):836–9.
12. Harley CB, Futcher AB, Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature. 1990;345(6274):458–60.
13. Cawthon RM, Smith KR, O'Brien E, Sivatchenko A, Kerber RA. Association between telomere length in blood and mortality in people aged 60 years or older. Lancet. 2003;361(9355):393–5.
14. Bernardes de Jesus B, Vera E, Schneeberger K, Tejera AM, Ayuso E, Bosch F, et al. Telomerase gene therapy in adult and old mice delays aging and increases longevity without increasing cancer. EMBO Mol Med. 2012;4(8):691–704.
15. Gerschman R, Gilbert DL, S W Nye, Dwyer P, Fenn WO. Oxygen poisoning and x-irradiation: a mechanism in common. Science. 1954;119(3097):623–6.
16. Harman D. Aging: a theory based on free radical and radiation chemistry. J Gerontol. 1956;11(3):298–300.
17. Sharma P, Jha AB, Dubey RS, Pessarakli M. Reactive Oxygen Species, Oxidative Damage, and Antioxidative Defense Mechanism in Plants under Stressful Conditions. Journal of Botany. 2012. p. 1–26.
18. Liochev SI. Reactive oxygen species and the free radical theory of aging. Free Radic Biol Med. 2013;60:1–4.
19. Viña J, Borras C, Abdelaziz KM, Garcia-Valles R, Gomez-Cabrera MC. The free radical theory of aging revisited: the cell signaling disruption

theory of aging. Antioxid Redox Signal [Internet]. 2013;19(8):779–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23841595
20. Nagakawa Y, Williams GM, Zheng Q, Tsuchida A, Aoki T, Montgomery RA, et al. Oxidative mitochondrial DNA damage and deletion in hepatocytes of rejecting liver allografts in rats: Role of TNF-α. Hepatology. 2005;42(1):208–15.
21. Miquel J, Economos AC, Fleming J, Johnson JE. Mitochondrial role in cell aging. Exp Gerontol. 1980;15(6):575–91.
22. Kowaltowski AJ. Caloric restriction and redox state: Does this diet increase or decrease oxidant production? Redox Report. 2011. p. 237–41.
23. Ristow M, Schmeisser S. Extending life span by increasing oxidative stress. Free Radical Biology and Medicine. 2011. p. 327–36.
24. Lapointe J, Hekimi S. When a theory of aging ages badly. Cellular and Molecular Life Sciences. 2010. p. 1–8.
25. Vitale G, Salvioli S, Franceschi C. Oxidative stress and the ageing endocrine system. Nat Rev Endocrinol. 2013;9(4):228–40.
26. Bohr VA. Repair of oxidative DNA damage in nuclear and mitochondrial DNA, and some changes with aging in mammalian cells. Free Radic Biol Med. 2002;32(9):804–12.
27. Bernstein C, Prasad AR, Nfonsam V, Harris Bernstein. DNA damage, DNA repair and cancer. Chen C, editor. New Research Directions in DNA Repair. 2013.
28. Dueva R, Iliakis G. Alternative pathways of non-homologous end joining (NHEJ) in genomic instability and cancer. Transl Cancer Res. 2013;2(3):163–77.
29. DiGiovanna JJ, Kraemer KH. Shining a light on xeroderma pigmentosum. J Invest Dermatol. 2012;132(3 Pt 2):785–96.
30. Mecocci P, MacGarvey U, Kaufman AE, Koontz D, Shoffner JM, Wallace DC, et al. Oxidative damage to mitochondrial DNA shows marked age-dependent increases in human brain. Ann Neurol. 1993;34(4):609–16.
31. Franceschi C, Campisi J. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. Journals Gerontol - Ser A Biol Sci Med Sci. 2014;69(SUPPL. 1).
32. De Martinis M, Franceschi C, Monti D, Ginaldi L. Inflammation markers predicting frailty and mortality in the elderly. Exp Mol Pathol. 2006;80:219–27.
33. Olivieri F, Rippo MR, Monsurrò V, Salvioli S, Capri M, Procopio AD, et al. MicroRNAs linking inflamm-aging, cellular senescence and cancer. Ageing Research Reviews. 2013. p. 1056–68.
34. Kindt TJ, Goldsby RA, Osborne BA. Kuby Immunology [Internet]. Kuby Immunology. 2007. 574 p. Available from: http://books.google.com/books?id=oOsFf2WfE5wC&pgis=1
35. Varadhan R, Yao W, Matteini A, Beamer BA, Xue QL, Yang H, et al. Simple biologically informed infammatory index of two serum cytokines predicts 10 year all-cause mortality in older adults. Journals Gerontol - Ser A Biol Sci Med Sci. 2014;69 A(2):165–73.
36. Arala-Chaves M, Sequeira T. Is there any kind of adaptive immunity in invertebrates? Aquaculture. 2000;191(1-3):247–58.
37. Rodier F, Campisi J, Bhaumik D. Two faces of p53: Aging and tumor suppression. Nucleic Acids Res. 2007;35(22):7475–84.
38. Williams GC. Pleiotropy, Natural Selection, and the Evolution of Senescence. Evolution (N Y). 1957;11(4):398–411.
39. Feltes B, Poloni J de F, Bonatto D. Development and Aging: Two Opposite but Complementary Phenomena. In: AI Y, SM J, editors. Aging and Health - A Systems Biology Perspective. 40th ed. Interdisciplinary Topics in Gerontology; 2015.
40. Feltes BC, de Faria Poloni J, Bonatto D. The developmental aging and origins of health and disease hypotheses explained by different protein networks. Biogerontology. 2011;12:293–308.
41. Leonard S, Murrant C, Tayade C, van den Heuvel M, Watering R, Croy BA. Mechanisms Regulating Immune Cell Contributions to Spiral Artery Modification - Facts and Hypotheses - A Review. Placenta. 2006. p. 40–6.
42. Fernandez-Capetillo O. Intrauterine programming of ageing. EMBO Rep. 2010;11(1):32–6.
43. Engels WR, Johnson-Schlitz D, Flores C, White L, Preston CR. A third link connecting aging with double strand break repair. Cell Cycle. 2007. p. 131–5.