

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
INSTITUTO DE BIOCÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO CIÊNCIAS BIOLÓGICAS - GENÉTICA

Giordano Bruno Sanches Seco

**Análise Comparativa de Métodos de Determinação de
Vias Biológicas Alteradas em Dados Toxicogenômicos
de estudos *in vitro* e *in vivo***

Botucatu, Junho de 2018

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO CIÊNCIAS BIOLÓGICAS - GENÉTICA

Giordano Bruno Sanches Seco

**Análise Comparativa de Métodos de Determinação de
Vias Biológicas Alteradas em Dados Toxicogenômicos
de estudos *in vitro* e *in vivo***

Dissertação apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Mestre no Programa de Pós-Graduação em Ciências Biológicas - Genética.

Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Junho de 2018.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Seco, Giordano Bruno Sanches.

Análise comparativa de métodos de determinação de vias biológicas alteradas em dados toxicogenômicos de estudos in vitro e in vivo / Giordano Bruno Sanches Seco. - Botucatu, 2018

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: José Luiz Rybarczyk Filho

Capes: 20204000

1. Agentes antineoplásicos. 2. Bioinformática. 3. Microarranjos de DNA. 4. Análise de microarranjo. 5. Toxicogenética.

Palavras-chave: Antineoplásicos; Bioinformática; Microarranjo; Toxicogenômica.

Agradecimentos

Agradeço aos processos CNPq 134467/2016-7, CNPq 458810/2013-4 e CNPq 473789/2013-2 cujo auxílio financeiro possibilitou a execução desse trabalho.

Agradeço ao Prof. Dr. José Luiz Rybarczyk Filho, cuja orientação foi de fundamental importância para solucionar os diversos desafios que apareceram durante a realização deste trabalho.

Agradeço a Dr(a) Agnes Alessandra Sekijima Takeda, cujo suporte fornecido na interpretação da parte biológica dos dados e na montagem de figuras permitiu a confecção dos resultados e discussões.

Agradeço aos colegas de laboratório André Luiz Molan, Carlos Alberto Oliveira de Biagi Junior e José Rafael Pilan cujo suporte tanto em nível técnico quanto pessoal foi essencial para execução das tarefas necessárias para a confecção deste trabalho.

Agradeço a todas as demais pessoas (professores, funcionários e amigos) que de alguma forma, direta ou não, tenham contribuído para a realização deste trabalho.

Agradeço imensamente a minha família, em especial ao meu pai Adalberto Alexandre Godoy Sanches Seco e mãe Márcia Regina Ogava Seco por todo suporte pessoal fornecido durante a realização deste trabalho.

Por fim agradeço a Deus pela vida.

Resumo

A toxicogenômica é uma área de estudo que se utiliza de dados provindos de tecnologias das ciências ômicas e de metodologias de análise de dados da bioinformática para caracterizar perfis biológicos de compostos a fim realizar um estudo sobre seu comportamento no sistema biológico. Um dos problemas na análise de dados de expressão gênica é que existem diversas metodologias disponíveis, cada uma aborda os dados brutos de maneira diferente e consequentemente fornecem resultados diferentes. Porém estes resultados não estão necessariamente errados, por essa razão abordar dados toxicogenômicos de diversas maneiras pode ajudar pesquisadores a tecer hipóteses sobre os mecanismos das drogas. Isso pode implicar em redução de custos na fase pré-clínica do desenvolvimento de medicamentos e numa melhor qualidade de tratamento para os pacientes. A abordagem mais comum na análise de dados de expressão gênica é o enriquecimento funcional dos genes que se mostrem diferencialmente expressos. Entretanto é possível que informações relevantes possam ser obtidas utilizando métodos com outras considerações. Para avaliar esta hipótese o presente trabalho tem por objetivo analisar dados de expressão gênica disponíveis em bases de dados on-line referentes a 3 drogas antineoplásicas testadas em fígado de *Rattus norvegicus* (estudos *in vivo* e *in vitro*) e *Homo sapiens* (estudo *in vivo* apenas) para avaliação de ontologias alteradas em cada droga. Duas metodologias de bioinformática foram utilizadas para análise desses dados. A tradicional análise de genes diferencialmente expressos (EFDEGs) e a análise de grupos de genes funcionalmente associados determinados pela entropia de shannon (AGFAGs). Para cada um dos três métodos foram testados três normalizações: RMA, MAS5, GCRMA. Os resultados sugerem que os 3 tipos de estudos tem pouca similaridade entre si e a pouca que existe se limita, aos principais mecanismos das drogas antineoplásicas como ciclo celular, apoptose, dano e reparo de DNA. Portanto não se pode afirmar ainda que um tipo de estudo possa substituir o outro. Também determinamos que resultados diferentes fornecidos pelas 2 metodologias para um mesmo caso podem não estar necessariamente errados e algumas vezes podem ser complementares. A meta-análise desse trabalho não só mostrou similaridade e complementariedade entre os métodos como também encontrou lncRNAs potencialmente responsáveis com a regulação pós-transcricional em hepatócitos humanos. Isso evidencia a importância da prática de se utilizar dados previamente obtidos em análises exploratórias para confecção de hipóteses. Assim o uso de ambas as metodologias pode ter valor como ferramenta para análises exploratórias em dados biológicos disponíveis.

Abstract

Toxicogenomics is an area that uses data generated by *omics technologies* and *bioinformatics* to characterize biological profiles of chemical compounds in order to analyze the biological system's behavior in response to said compounds. A common problem in the analysis of gene expression data is the large amount of methodologies available, each one approaches the raw data differently and consequently provide different results. But these results are not necessarily wrong, for that reason approaching toxicogenomic data in several ways may help researchers to hypothesize about a drug's mechanisms. This may imply cost reduction in the pre-clinical phase of drug development and in a better quality of treatment for patients. The most common approach in the analysis of gene expression data is the functional enrichment of differentially expressed genes (EFDEGs). However, it is possible that relevant information can be obtained using methods with other considerations. To evaluate this hypothesis this work aims to analyze gene expression data available in the OPEN TG-GATEs online database about 3 antineoplastic drugs tested in the liver of *Rattus norvergicus* (*in vivo* and *in vitro*) and *Homo sapiens* (*in vivo* only) for the evaluation of altered ontologies in each drug. Two bioinformatics methodologies were used to analyze these data. The traditional analysis of differentially expressed genes (EFDEGs) and the analysis of groups of functionally associated genes determined by shannon entropy (AGFAGs). For each of the two methods, three normalizations were tested: RMA, MAS5, GCRMA. Results suggest that the 3 types of studies have little similarity to each other and the few that exist are limited to the main mechanisms of antineoplastic drugs such as cell cycle, apoptosis and DNA damage and repair. Therefore, it still cannot be said that one type of study can replace the other. We also found out that the different results provided by the two methodologies for the same case may not necessarily be wrong and sometimes may be complementary. The meta-analysis done in this work not only showed similarity and complementarity between methods but also found differentially expressed lncRNAs potentially responsible for post-transcriptional regulation in human hepatocytes. This evidences the importance of using previously obtained data in exploratory analyzes to make hypotheses. Thus the use of both methodologies may have value as a tool for exploratory analyzes.

Lista de Figuras

1.1	Relação hierarquica das quatro frentes das ciências ômicas	p. 15
1.2	<i>Workflow</i> descrevendo um experimento geral de Toxicogenômica	p. 18
1.3	Ilustração das etapas de um experimento de microarranjo de dois canais (duas cores)	p. 29
1.4	<i>Workflow</i> de um experimento de toxicogenômica envolvendo o uso de dados de expressão gênica	p. 30
1.5	Cronologia dos experimentos <i>in vivo</i> no <i>Rattus norvegicus</i> . Adaptado de (TAKASHIMA et al., 2006)	p. 38
3.1	<i>Workflow</i> da pesquisa	p. 50
3.2	Exemplo de uma segmento da ontologia BP	p. 52
4.1	Boxplots das 3 normalizações para os chips da ciclofosfamida e estudo <i>Homo sapiens in vitro</i>	p. 62
4.2	Distribuição d GDEs das 3 drogas no método EFDEGs	p. 64
4.3	Gráficos de linhas para ilustrar os GDEs das 3 drogas no método EFDEGs com <i>cutoff</i> de 0,001	p. 65
4.4	Distribuição de BPs das 3 drogas no método EFDEGs	p. 68
4.5	Gráficos de linhas para ilustrar os GDEs das 3 drogas no método EFDEGs com <i>cutoff</i> de 0,001	p. 69
4.6	Distribuição de GDEs da ciclofosfamida no método EFDEGs	p. 71
4.7	Gráficos de linhas para ilustrar os GDEs da ciclofosfamida no método EF- DEGs com <i>cutoff</i> de 0,001	p. 72
4.8	Distribuição de BPs da ciclofosfamida no método EFDEGs	p. 74

4.9	Gráficos de linhas para ilustrar os BPs da ciclofosfamida no método EF-DEGs com <i>cutoff</i> de 0,001	p. 75
4.10	Expressão dos genes anotados na ontologia GO:0002933	p. 78
4.11	Expressão dos genes anotados na ontologia GO:0009822	p. 79
4.12	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Homo sapiens in vitro</i> na ciclofosfamida com uso do STRING	p. 82
4.13	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Rattus norvegicus in vitro</i> na ciclofosfamida com uso do STRING	p. 86
4.14	Rede PPI derivada do universo de GDEs do estudo <i>Rattus norvegicus in vitro</i> na ciclofosfamida com uso do STRING	p. 89
4.15	Gráficos de linhas para ilustrar os GDEs do etopósido no método EFDEGs com <i>cutoff</i> de 0,001	p. 90
4.16	Distribuição de GDEs do etopósido no método EFDEGs	p. 91
4.17	Distribuição de BPs do etopósido no método EFDEGs	p. 93
4.18	Gráficos de linhas para ilustrar os BPs do etopósido no método EFDEGs com <i>cutoff</i> de 0,001	p. 94
4.19	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Homo sapiens in vitro</i> no etopósido com uso do STRING	p. 98
4.20	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Homo sapiens in vitro</i> no etopósido com uso do STRING	p. 100
4.21	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Rattus norvegicus in vivo</i> no etopósido com uso do STRING	p. 102
4.22	Distribuição de GDEs da lomustina no método EFDEGs	p. 104
4.23	Gráficos de linhas para ilustrar os GDEs da lomustina no método EFDEGs com <i>cutoff</i> de 0,001	p. 105
4.24	Distribuição de BPs da lomustina no método EFDEGs	p. 106

4.25	Gráficos de linhas para ilustrar os BPs da lomustina no método EFDEGs com <i>cutoff</i> de 0,001	p. 107
4.26	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Homo sapiens in vitro</i> na lomustina com uso do STRING	p. 109
4.27	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Rattus norvegicus in vitro</i> na lomustina com uso do STRING	p. 111
4.28	Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo <i>Rattus norvegicus in vivo</i> na lomustina com uso do STRING	p. 113
4.29	Distribuição de BPs das 3 drogas no método AGFAGs (atividade)	p. 115
4.30	Gráficos de linhas para ilustrar os BPs das 3 drogas no método AGFAGs (atividade) com <i>cutoff</i> de 0,001	p. 116
4.31	Distribuição de BPs das 3 drogas no método AGFAGs (diversidade)	p. 118
4.32	Gráficos de linhas para ilustrar os BPs das 3 drogas no método AGFAGs (diversidade) com <i>cutoff</i> de 0,001	p. 119
4.33	Distribuição de BPs da ciclofosfamida drogas no método AGFAGs (atividade)p.	121
4.34	Gráficos de linhas para ilustrar os BPs da ciclofosfamida no método AGFAGs (atividade) com <i>cutoff</i> de 0,001	p. 122
4.35	Distribuição de BPs da ciclofosfamida drogas no método AGFAGs (diversidade)	p. 123
4.36	Gráficos de linhas para ilustrar os BPs da ciclofosfamida no método AGFAGs (diversidade) com <i>cutoff</i> de 0,001	p. 124
4.37	Distribuição de BPs do etopósido no método AGFAGs (atividade)	p. 132
4.38	Gráficos de linhas para ilustrar os BPs do etopósido no método AGFAGs (atividade) com <i>cutoff</i> de 0,001	p. 133
4.39	Distribuição de BPs do etopósido no método AGFAGs (diversidade)	p. 134
4.40	Gráficos de linhas para ilustrar os BPs do etopósido no método AGFAGs (diversidade) com <i>cutoff</i> de 0,001	p. 135

4.41	Distribuição de BPs da lomustina no método AGFAGs (atividade)	p. 142
4.42	Gráficos de linhas para ilustrar os BPs da lomustina no método AGFAGs (atividade) com <i>cutoff</i> de 0,001	p. 143
4.43	Distribuição de BPs da lomustina no método AGFAGs (diversidade)	p. 144
4.44	Gráficos de linhas para ilustrar os BPs da lomustina no método AGFAGs (diversidade) com <i>cutoff</i> de 0,001	p. 145

Lista de Tabelas

1.1	Resumo do design experimental dos estudos <i>in vivo</i>	p. 40
1.2	Resumo do Design Experimental dos estudos <i>in vitro</i>	p. 41
1.3	Tabela de relação entre drogas antineoplásicas e estudos realizados no OPEN TG-GATEs	p. 44
1.4	Tabela com o protocolo utilizado nos estudos <i>in vivo</i> no OPEN TG-GATEs	p. 44
1.5	Tabela com o protocolo utilizado nos estudos <i>in vivo</i> no OPEN TG-GATEs	p. 45
3.1	Métodos de normalização de dados de microarranjos e suas características .	p. 56
4.1	Processos chave encontrados nos estudos/drogas no EFDEGs	p. 152
4.2	Processos chave encontrados nos estudos/drogas no AGFAGs (atividade) .	p. 152
4.3	Processos chave encontrados nos estudos/drogas no AGFAGs (diversidade)	p. 153

Sumário

Resumo	p. 2
Abstract	p. 3
1 Introdução	p. 12
1.1 Advento das Tecnologias de Alto Desempenho: As ciências Ômicas	p. 12
1.2 Toxicogenômica	p. 16
1.2.1 Definição	p. 16
1.2.2 O Potencial da Toxicogenômica: <i>Design</i> Experimental e Aplicações Comuns	p. 17
1.2.3 Redes Biológicas: A Toxicogenômica nos Dias Atuais	p. 24
1.3 Microarranjo (<i>microarray</i>)	p. 28
1.3.1 Princípios da Técnica	p. 28
1.3.2 <i>Design</i> do Experimento de Expressão Gênica e Análise de Dados: Aplicações, Exemplos, Limitações e Perspectivas	p. 30
1.4 O Projeto Toxicogenômico Japonês Open TG-GATEs e os Dados Seleccionados	p. 36
1.4.1 OPEN TG-GATES: Design Experimental	p. 37
1.4.1.1 Estudos <i>In vivo</i>	p. 37
1.4.1.2 Estudos <i>In vitro</i>	p. 40
1.4.2 Sobre as Drogas	p. 41
1.4.3 Antineoplásicos e o OPEN TG-GATEs	p. 43

1.5	Justificativa	p. 45
2	Objetivos	p. 48
2.1	Objetivos Principais	p. 48
2.2	Objetivos Específicos	p. 48
3	Materiais e Metodos	p. 49
3.1	<i>Workflow</i>	p. 49
3.2	<i>Gene Ontology</i> (GO)	p. 51
3.3	STRING: Um Banco de Dados de Interações Proteína-Proteína	p. 53
3.4	Normalização dos dados	p. 54
3.4.1	<i>Robust Multi-array Average</i> (RMA)	p. 56
3.4.2	<i>GC Robust Multi-array</i> (GCRMA)	p. 57
3.4.3	<i>Microarray Suite 5</i> (MAS5)	p. 57
3.5	Enriquecimento Funcional de Genes Diferencialmente Expressos (EFDEGs)	p. 58
3.6	Análise de Grupos de Genes Funcionalmente Associados (AGFAGs)	p. 59
4	Resultados Discussão	p. 61
4.1	Sobre as análises e normalizações	p. 61
4.2	Método EFDEGs	p. 63
4.2.1	Aspectos Gerais: Análise Numérica e Descritiva	p. 63
4.2.2	Ciclofosfamida	p. 70
4.2.3	Etopósido	p. 90
4.2.4	Lomustina	p. 103
4.3	Método AGFAGs	p. 114

4.3.1	Aspectos Gerais: Análise Numérica e Descritiva	p. 114
4.3.2	Ciclofosfamida	p. 120
4.3.3	Etopósido	p. 131
4.3.4	Lomustina	p. 141
4.4	Convergência dos Resultados	p. 150
5	Conclusão	p. 154
	Referências Bibliográficas	p. 156

1 *Introdução*

1.1 Advento das Tecnologias de Alto Desempenho: As ciências Ômicas

Hoje pesquisadores tem acesso a tecnologias que são capazes de gerar informações sobre diferentes aspectos do material genético com vazão, eficiência e precisão muito maiores do que as tecnologias disponíveis nas décadas de 80 e 90, utilizadas para realização do projeto de sequenciamento do genoma humano e com um custo consideravelmente menor (GALAS; MCCORMACK, 2003) (NRC et al., 2007). O conjunto dessas tecnologias capazes de gerar grandes quantidades de informação sobre os diferentes aspectos do material genético, será aqui referido como “**tecnologias ômicas**”. As **ciências ômicas** são, essencialmente, a aplicação das tecnologias ômicas, ferramentas de **bioinformática** e de métodos de integração e análise de dados de grande porte para realizar um estudo global de um determinado sistema biológico exposto a determinadas condições. Alguns exemplos de **ciências ômicas** são: **genômica**, **proteômica**, **transcriptômica** e **metabolômica**. Esses quatro exemplos são a principal forma de busca de informações que a **toxicogenômica** utiliza para compreender um sistema biológico.

A **genômica** e suas tecnologias relacionadas englobam principalmente o sequenciamento do genoma de organismos, a mapeamento dos genomas (genes) e análises genotípicas. As tecnologias genômicas se desenvolveram muito na década passada (2000-2010), seja na capacidade de sequenciamento de genomas inteiros ou seja na capacidade de realizar análises genotípicas (NRC et al., 2007). Isto possibilita a avaliação de múltiplas variáveis em todo o genoma de diversos tipos de organismos ou mesmo a avaliação de polimorfismos genéticos em grandes populações de indivíduos de uma determinada espécie (GRIFFITHS, 2005). A avaliação da variação do genoma entre indivíduos pode fornecer informações sobre as dife-

rentes reações que humanos tem à agentes químicos e outros fatores ambientais. A avaliação de diferenças entre genomas de indivíduos é composta principalmente pela busca de SNPs (*single nucleotide polymorphisms*), ou seja variações na própria sequência de nucleotídeos no genoma dos indivíduos e portanto na informação genética codificada. Essa avaliação também é realizada pela busca de fatores epigenéticos, que são mudanças reversíveis e herdáveis capazes de alterar a maneira como os genes se expressam sem causar mudança na informação codificada por eles (sequência de nucleotídeos). Alguns fatores epigenéticos comuns são: metilação do DNA, *Imprinting* e modificações das histonas (NRC et al., 2007; HANDE, 2017).

A **transcriptômica** é capaz de determinar o perfil de expressão gênica de um único gene ou de vários (isso depende da técnica a ser utilizada) a partir da análise do transcriptoma, este representa o conjunto de todos os transcritos celulares em um determinado momento. A técnica *Northern Blot* (ALWINE; KEMP; STARK, 1977), por exemplo, é capaz de mensurar a expressão relativa de um único gene entre um caso e um controle (SCHULZE; DOWNWARD, 2001). Hoje em dia existem técnicas capazes de mensurar a expressão relativa de milhares de genes simultaneamente como os *chips* de microarranjo e o RNA-seq (WANG; GERSTEIN; SNYDER, 2009). Todas essas técnicas usam como base a determinação de quantidades relativas de transcritos celulares para determinar quais genes se encontram subexpressos ou superexpressos. A vantagem dessas últimas tecnologias é que elas podem fornecer informações de diversos genes simultaneamente (SCHULZE; DOWNWARD, 2001). Dessa maneira é possível determinar perfis de expressão gênica, dando ao pesquisador a capacidade de determinar genes que estão super ou subexpressos em diversas situações. Assim é possível por meio do uso de bases de dados que relacionam os genes a sua função (proteína que codifica ou ação reguladora), determinar processos biológicos que estão super ou subexpressos. Cabe lembrar que, entretanto, a relação entre quantidade relativa de transcritos celulares e concentração das proteínas nem sempre é linear e isto constitui um desafio para análise do transcriptoma (NRC et al., 2007; HANDE, 2017).

A **Proteômica** estuda o conjunto de todas as proteínas expressas em um determinado momento na célula ou tecido o qual é chamado de **proteoma**. As proteínas são as efetoras diretas da informação codificada no genoma e por isso ter acesso a sua distribuição na célula/tecido pode fornecer informações que não podem ser obtidas por vias da genômica ou

transcriptômica. Ter acesso a essas informações permite o pesquisador criar bases de dados que associam sequências do genoma com sua respectiva proteína e suas concentrações, assim como associações entre duas ou mais proteínas. O proteoma não é imutável ou fixo como o genoma, as concentrações das proteínas estão sempre variando como resposta aos estímulos que o ambiente fornece ao organismo. Não existe nenhuma tecnologia análoga ao PCR (*polimerase chain reaction*), que permite amplificação do genoma e transcriptoma, para aplicar ao proteoma. As proteínas devem ser analisadas em suas concentrações nativas, que podem abranger até 6 ordens de magnitude (de 1 a 10^6) e isto constitui um desafio considerável para análise do proteoma (NRC et al., 2007; HANDE, 2017).

Outros desafios incluem modificações constantes do proteoma (danos), causados por agentes químicos intermediários formados por substâncias tóxicas e estresse oxidativo endógeno. Devido a essas razões as tecnologias proteômicas envolvem elaboradas análises seriais de poucos produtos. Isso é o oposto do que ocorre com tecnologias paralelas, como *chips* de microarranjo, que permitem avaliação de milhares de produtos. As técnicas mais utilizadas para determinação do proteoma, atualmente, constituem um conjunto de diversas variações e adaptações da espectrometria de massa (*Mass Spectrometry* (MS)) (AEBERSOLD; MANN, 2003).

A **metabolômica** estuda o conjunto de todos os metabólitos do organismo em um determinado momento o qual é chamado de **metaboloma**. Metabólitos intermediáveis refletem a interferências de proteínas em vias bioquímicas e portanto representam o estado biológico de uma célula ou tecido de uma maneira análoga ao proteoma. O metaboloma possui uma grande diversidade química, ao contrário do genoma, do transcriptoma e do proteoma. Alguns exemplos de tipos de metabólitos são: pequenos peptídeos, lipídios, precursores de ácidos nucleicos, diversos produtos de degradação de agentes químicos intermediários em processos de biosíntese e de catabolismo e metabólitos de compostos exógenos (derivados da dieta e intervenções médicas, por exemplo) (NRC et al., 2007; HANDE, 2017). Essa diversidade química do metaboloma constitui um desafio para sua análise pois dificulta sua avaliação a partir de uma única técnica.

A metabolômica também sofre de algumas dificuldades presentes na proteômica como o fato de o metaboloma não ser imutável ou fixo como o genoma e que as abundâncias nativas dos metabólitos podem ter muitas ordens de magnitude (NRC et al., 2007; HANDE, 2017). Algumas técnicas utilizadas para avaliação do metaboloma são diversas variações da

ressonância magnética nuclear (NMR) e da espectrometria de massa (ERNST et al., 1987; WU et al., 2008).

A figura 1.1 abaixo demonstra as quatro frentes das ciências ômicas e seus objetos de estudo em uma maneira hierárquica, considerando a direção do fluxo de informação genética.

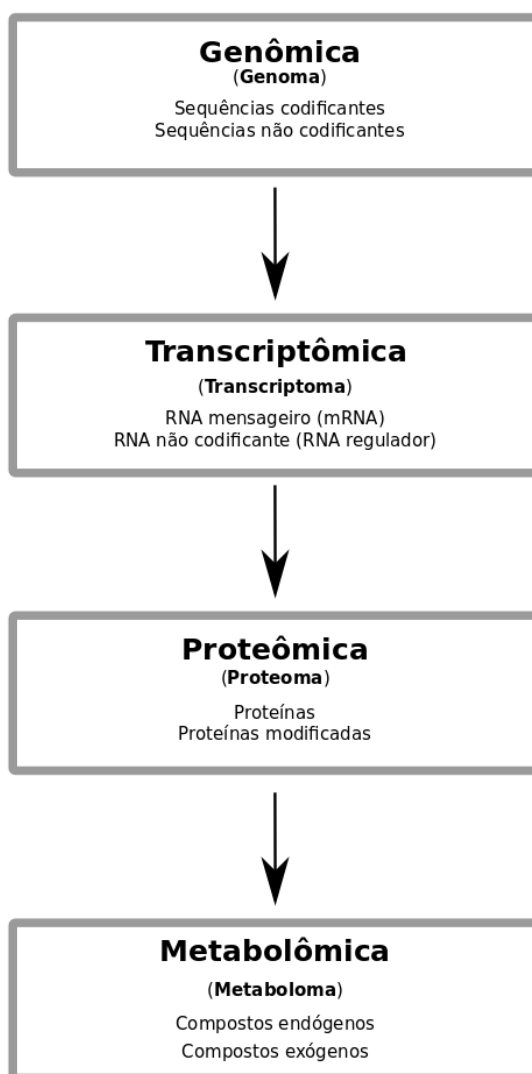


Figura 1.1: Relação hierárquica das quatro frentes das ciências ômicas. Adaptado de (NRC et al., 2007; HANDE, 2017).

1.2 Toxicogenômica

1.2.1 Definição

A **toxicogenômica** combina a toxicologia com as ciências ômicas para obter e analisar informações referentes a efeitos adversos em células ou tecidos de várias espécies de organismos expostos a uma determinada molécula. O comitê responsável pelo desenvolvimento do livro *Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment*, publicado em 2007 pela NAP (*National Academies Press*) sob autoria do NRC (*National Research Council*) (NRC et al., 2007; HANDE, 2017), define toxicogenômica como:

“Aplicação de tecnologias genômicas (por exemplo, genética, sequenciamento de genomas, análise de perfis de expressão gênica, proteômica, metabolômica, e abordagens relacionadas) para estudar os efeitos adversos de produtos químicos ambientais e farmacêuticos sobre saúde humana e ambiente.”

O relatório (NRC et al., 2007; HANDE, 2017) ainda descreve em linhas gerais os objetivos da **toxicogenômica**:

*“Toxicogenômica combina toxicologia com tecnologias genômicas capazes de gerar grandes quantidades de informação para integrar dados referentes à alterações tóxicas específicas em genes, proteínas e padrões de expressão de metabólitos com respostas fenotípicas de células, tecidos e organismos. A **toxicogenômica** pode fornecer informações sobre a interação genoma-ambiente e sobre o comportamento das vias metabólicas e redes biológicas às perturbações. As informações obtidas por esta nova abordagem podem ser mais discriminatórias, preditivas, e sensíveis do que as usadas atualmente para avaliar exposições a substâncias tóxicas e para prever seus efeitos na saúde humana.”*

1.2.2 O Potencial da Toxicogenômica: *Design* Experimental e Aplicações Comuns

Com as tecnologias atuais a geração de grandes quantidades de dados é algo usual nas ciências ômicas, algo que não acontecia com as técnicas de uma década atrás. Em vez disso o desafio atual compreende: coleta, manipulação, análise e interpretação eficientes dos dados obtidos. As quatro frentes das ciências ômicas (genômica, transcriptômica, proteômica e metabolômica) examinam aspectos diferentes das respostas celulares ou teciduais. Entretanto as abordagens gerais do *design* experimental de ensaios toxicogenômicos são quase que idênticas, o que varia é: a técnica utilizada para obtenção de dados, a natureza dos dados a serem analisados e como consequência os tipos de conclusões que podem ser retiradas do experimento (NRC et al., 2007; HANDE, 2017).

Neste item abordaremos elementos do *design* experimental geral dos ensaios toxicogenômicos e aplicações potenciais da **toxicogenômica**. Abaixo podemos ver na figura 1.2 um *workflow* descrevendo as etapas em um experimento de **toxicogenômica**.

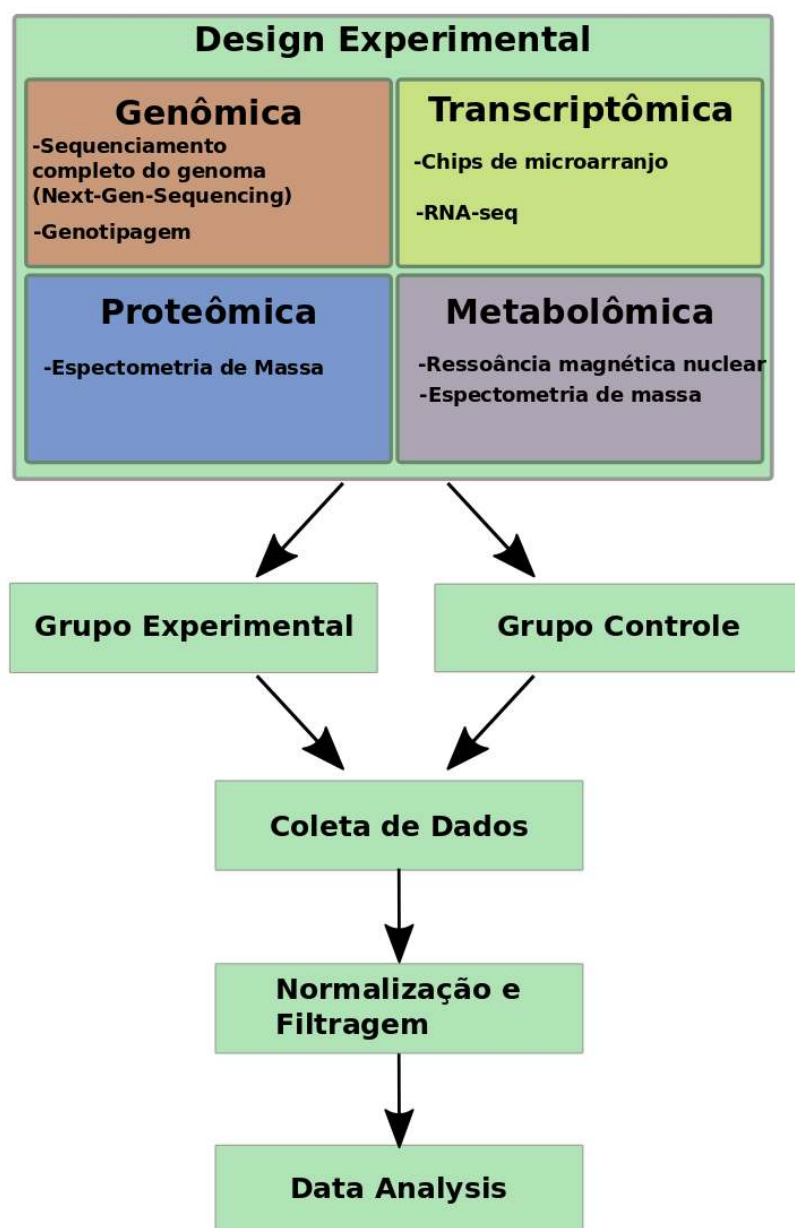


Figura 1.2: *Workflow* descrevendo um experimento geral de toxicogenômica: a primeira parte de qualquer ensaio tóxicogenômico é o *design* experimental dos experimentos responsáveis pela geração dos dados. Para isso é preciso ter em mente qual a hipótese inicial e posteriormente determinar fatores como: qual a técnica ideal para a validação da hipótese, como definir grupo caso e grupo controle, determinar o espaço amostral de cada grupo de modo que as estatísticas possam ser extrapoladas para a população geral com certo nível de confiança e determinar o número de replicatas de cada situação caso-controle. Após a determinação do *design* experimental e de sua execução os dados devem ser coletados e então normalizados e filtrados. Os métodos de normalização utilizados vão depender do tipo de experimento que foi realizado e do tipo de dado em questão. Por último temos a análise dos dados, os métodos são variados e a determinação de um método de análise ideal para busca de informações desejadas é um passo fundamental para obter resultados coerentes. Adaptado de (NRC et al., 2007; HANDE, 2017).

Existem diversas abordagens que podem ser dadas a um estudo toxicogenômico. Abaixo comenta-se sobre quatro abordagens muito comuns: a **avaliação de exposição**, a **avaliação da toxicidade**, a **avaliação da variabilidade humana** e a **construção de redes**.

A **avaliação da exposição** tem o objetivo de utilizar dados toxicogenômicos para mensurar a exposição de um sistema a um certo agente. O grande interesse da **toxicogenômica**, nesse caso, é a determinação de biomarcadores de exposição e de respostas a exposição (NRC et al., 2007; HANDE, 2017). Tipicamente adutos são os principais tipos moleculares que são utilizados como biomarcadores de exposição, como por exemplo adutos da hemoglobina formados perante exposição à benzeno ou produtos de pirólise (ALEXANDER et al., 2002; MEDEIROS; BIRD; WITZ, 1997; SKIPPER et al., 1994), os adutos de DNA produzidos por exposição a cloreto de vinila ou uretano (SKIPPER et al., 1994) e os hidrocarbonetos aromáticos policíclicos formados pela fumaça do cigarro (SHUGART; HOLLAND; RANN, 1983; PERERA; SANTELLA; POIRIER, 1986).

Geralmente os biomarcadores são definidos como moléculas únicas, porém é possível utilizar conjuntos de tipos moleculares como um perfil de exposição a certos produtos. Esse tipo de abordagem é potencialmente mais sensível e específica do que a convencional já que agentes diferentes podem possuir perfis de exposição superpostos em alguns casos. Um exemplo seria a diferenciação do potencial mutagênico de agentes carcinogênicos. O metil metanosulfonato produz a N7-metilguanina como principal aduto de DNA, que também é a principal lesão causada por agentes carcinogênicos como a N-methyl-N'-nitro-N-nitrosuréia. A diferença entre os efeitos carcinogênicos e mutagênicos causados por esses dois compostos, entretanto, é causada por variações em níveis de lesões relativamente menores: o agente altamente mutagênico, O6-metilguanina, é responsável por 0.1% do dano ao DNA causado pelo metil metanosulfonato, mas é responsável por até 7% no caso da N-methyl-N'-nitro-N-nitrosuréia (NRC et al., 2007; PEGG, 2000; MONTESANO; PEGG; MARGISON, 1980).

Embora os adutos sejam os principais biomarcadores de exposição utilizados atualmente, a **toxicogenômica** fornece a possibilidade de utilizar outros aspectos do organismo como marcadores de exposição (NRC et al., 2007; HANDE, 2017). Células e tecidos tem a expressão de seus genes e seu metabolismo alterados durante a exposição à diversos agentes. Um bom exemplo é a indução da expressão de grupos de genes específicos em resposta a choque térmico, hipóxia e estresse osmótico (FINKEL; HOLBROOK, 2000; CHELLAPPAN, 2001; BERRA; GINOUVÈS; POUYSSÉGUR, 2006).

Condições diferentes produzem diferentes padrões de mudanças no perfil de expressão de genes, proteínas e metabólitos. O uso de redes de interação, integrando a complexa relação dos genes a seus transcritos, proteínas e metabólitos era algo impensável para as tecnologias utilizadas no projeto do genoma humano. Entretanto as tecnologias utilizadas hoje em ensaios toxicogenômicos permitem a possibilidade desse tipo de abordagem. O uso da **toxicogenômica** na avaliação de exposição e de risco ainda está em fase preliminar, e seu uso na prática cotidiana de laboratórios ainda não é uma realidade (NRC et al., 2007; HANDE, 2017). Os principais desafios a serem superados são:

- Definição do que constitui um perfil transcriptômico, proteômico ou metabolômico como uma assinatura específica a um agente toxicante e do limiar de exposição mínimo que tais perfis podem identificar.
- Determinação de quais fluidos e tecidos, relativamente mais acessíveis a pesquisadores, podem melhor tipificar a exposição humana à agentes tóxicos especialmente a exposições baixas.
- Determinação de como exposição a múltiplos agentes tóxicos pode afetar os perfis transcriptômicos, proteômicos e metabolômicos e do limiar de exposição mínimo que tais perfis são capazes de detectar.
- Determinação de como o intervalo pós-exposição de agentes tóxicos afeta a habilidade de reconhece-los através de suas assinaturas toxicogenômicas.

A **avaliação de toxicidade** tem o objetivo de rastrear danos causados por agentes tóxicos a sistemas biológicos. Geralmente as abordagens mais tradicionais desse tipo de estudo requerem extensos estudos *in vivo* e *in vitro* para determinação dos riscos de um certo composto. Embora este tipo de estudo seja essencial no desenvolvimento de drogas, o custo elevado e o tempo demorado permacem como obstáculos consideráveis.

A questão crítica sobre novas técnicas toxicogenômicas é se elas podem melhorar o rastreamento de riscos, fazendo testes mais rápidos, mais abrangentes, menos dependentes de animais, mais preditivos e precisos com custos acessíveis (NRC et al., 2007; HANDE, 2017). Para atingir esse fim será necessário a construção de bases de dados toxicogenômicos e toxicológicos extensivas para criação e treinamento de algoritmos classificadores.

Assim será possível a criação de algoritmos que possam prever classes de agentes segundo seu potencial de causar dano. Outros problemas essenciais a serem aderidos nesse tipo de abordagem são: a tradução de resultados de estudos *in vitro* para estudos *in vivo* e extrapolação de resultados entre espécies. Somente com soluções viáveis para esses problemas será possível desenvolver novos testes de avaliação de risco que possam efetivamente incorporar dados toxicogenômicos para aumentar sua eficiência, sensibilidade e reduzir seus custos em relação a testes mais tradicionais (NRC et al., 2007; HANDE, 2017).

A **análise da variabilidade humana** tem o objetivo de determinar como a variação no material genético de indivíduos afeta sua resposta a agentes tóxicos. Existem duas principais fontes de variação no material genético sendo que a primeira é variação da informação codificada nos ácidos nucleicos. Dentro de uma mesma espécie é comum que indivíduos diferentes possuam trocas de bases pontuais em seu código genético, chamadas de SNPs (*Single Nucleotide Polymorphisms*). Outros tipos de alteração da informação codificada são as mutações de inserção e de deleção (GRIFFITHS, 2005). Um outro tipo de variação do material genético é aquela que não altera a informação codificada e sim a maneira como esta se expressa. Nessa situação, os fatores epigenéticos são a principal fonte de variação (NRC et al., 2007; HANDE, 2017).

Esse tipo de abordagem ainda está em estágio inicial, ainda é necessário superar diversos obstáculos para consolidar metodologias que utilizem assinaturas toxicogenômicas para predição dos diferentes efeitos que um agente tóxico podera induzir em indivíduos diferentes (NRC et al., 2007; GRIFFITHS, 2005).

A **análise de mecanismos de ação** tem como objetivo a determinação dos processos biológicos por trás dos efeitos de um agente tóxico. Informações sobre os mecanismos de ação das drogas podem ser úteis na avaliação de risco de novos compostos ou no desenvolvimento de drogas como sugerido em (PETRICON et al., 2002), perfis proteômicos são utilizados para a predição de câncer de prostata.

A aplicação de **toxicogenômica** para o estudo de mecanismos de toxicidade baseia-se na premissa de que a lesão química ou física é mediada por, ou refletida em, mudanças no nível de transcritos, proteínas ou metabólitos. Apesar da aceitação geral desse conceito, dados de estudos de investigação de mecanismos de ação ainda são pouco utilizados na avaliação de risco. As principais razões para isso são a falta de conhecimento dos mecanismos de ação envolvidos com a toxicidade expressa por drogas e falta de dados referentes a exposições a

múltiplos agentes em tempo real (NRC et al., 2007; HANDE, 2017).

Apesar disso já existem relatos que descrevem evidências de possíveis vantagens da aplicação informações mecanísticas para avaliação de risco de compostos como demonstrado em (COUNCIL et al., 2006), onde é realizada uma avaliação de risco de dioxinas para a saúde humana e em (IARC et al., 1987) onde são explorados os efeitos carcinogênicos do Benzo(a)pireno. Existem três principais tipos de estudos mecanísticos realizados:

- **A abordagem contextual** coloca dados toxicogenômicos recém-gerados no contexto do conhecimento biológico e toxicológico disponível, comparando um conjunto de elementos (perfis de transcriptoma, por exemplo) com parâmetros de resposta conhecidos (NRC et al., 2007; HANDE, 2017). Essa abordagem é baseada em um processo de raciocínio dedutivo e é utilizada quando, por exemplo, deseja-se saber quais os efeitos de um composto desconhecido utilizando como ponto de partida perfis de expressão gênica, de outros compostos, que sejam altamente similares e cujos efeitos já são conhecidos. Apesar de ser uma abordagem altamente descritiva (e não preditiva) ela é muito útil pois apresenta baixo viés e não requer conhecimento detalhado das complexas anotações de vias metabólicas e genes. A hipótese por trás dessa abordagem é a de que compostos com toxicidade similares afetam genes similares, como demonstrado em (HAMADEH et al., 2002) .
- **A abordagem exploratória** parte de uma descoberta para uma abordagem orientada por hipóteses, que no final tem o objetivo de avaliar vias específicas. Por esse motivo esse tipo de estudo pode elucidar novos mecanismos de ação em situações específicas (NRC et al., 2007; HANDE, 2017). O processo de raciocínio é, por definição, mais indutivo nesse caso. Por exemplo, um certo gene responsável por uma determinada função biológica é selecionado em um estudo que utiliza a abordagem contextual para comparar dois compostos de toxicidade similar. A abordagem exploratória então partiria da hipótese levantada no estudo anterior, de que este gene estaria relacionado com a expressão da toxicidade em questão, e realizaria uma re-avaliação do sistema biológico devidamente perturbado para determinação de funções biológicas significativamente alteradas. Este tipo de abordagem é realizada geralmente em conjunto com análise de vias biológicas, exemplos de bases de dados dessas vias incluem o *Gene Ontology* (GO) que liga genes a processos biológicos (BP), funções moleculares (MF) e compartimentos celulares (CC) e também o KEGG (*Kyoto Encyclopedia of Genes*

and Genomes), que liga genes a vias metabólicas. Um exemplo de estudo que utiliza essa abordagem seria em (VEZINA; WALKER; OLSON, 2004), onde microarranjos de DNA foram utilizados para identificar perfis de expressão gênica de exposição subcrônica de três ligantes ao receptor aril hidrocarboneto e um não ligante com estrutura similar aos outros. Com base nesse perfis o estudo conseguiu selecionar genes que antes não haviam sido caracterizados como alvos dos ligantes. Em outro estudo de mecanismos de carcinogênese química, Dickinson *et al* (DICKINSON et al., 2004) comparou os efeitos do fármaco genotóxico com cisplatina com estresse osmótico não genotóxico causado pelo cloreto de sódio. Perfis transcriptômicos de células tratadas com cisplatina revelaram aumentos significativos nos transcritos associados a dano no DNA (por exemplo, membros da família GADD45) e reparo adaptativo de células. (por exemplo, fos e homólogos HSP40). Em contraste, em concentrações equitóxicas, o perfil de expressão gênica de células tratadas com cloreto de sódio não indicou mudanças nas transcritos associados ao dano e reparo do DNA.

- **A abordagem de construção de redes** usa padrões de transcrição, co-regulação ou outros meios para construir mapas de interação entre seus nodos. A modelagem do sistema biológico em forma de redes possibilita a dedução de relações que colocam respostas individuais dentro de um quadro mais amplo que considera as complexas interações que existem nesse sistema. Essa abordagem busca modelar o sistema biológico como um todo, integrando dados de diversas procedências para poder descrever as interações do sistema ou até mesmo predizê-las e é frequentemente chamada de “biologia de sistemas” (NRC et al., 2007; HANDE, 2017). Os estudos realizados por Samsom *et al* e colaboradores (JELINSKY; SAMSON, 1999; JELINSKY et al., 2000) são um exemplo desse tipo de abordagem, onde foi realizada análise a nível de “biologia de sistema” de ações de agentes alquilantes em *Sacchromyces cerevisiae*. O grupo produziu diferentes cepas mutantes, através de deleção, para avaliar as funções individuais de genes, proteínas e redes de interatoma (BEGLEY et al., 2002; BEGLEY; SAMSON, 2004; SAID et al., 2004). Esses estudos revelaram novas redes de sinalização de estresse e novas interações de genes envolvidos no reparo do DNA, na renovação de proteínas e na regulação metabólica. O poder deste modelo da *S. cerevisiae* é derivado da integração de perfis de transcriptoma, uma extensa base de dados de interações proteína-proteína e proteína-gene (XENARIOS et al., 2002) e a disponibilidade de uma biblioteca quase compreensiva de mutantes com deleções (GIAEVER

et al., 2002).

Outras aplicações potenciais incluem determinação da **relação dose-resposta**, determinação do “**limiar sem risco**”, estudo da **extrapolação entre espécies**, estudo dos **efeitos no desenvolvimento** e estudo dos **efeitos de diversos agentes químicos** agindo simultaneamente.

1.2.3 Redes Biológicas: A Toxicogenômica nos Dias Atuais

Na última subseção foi discutido o potencial que a toxicogenômica tem de melhorar a avaliação de risco de novos compostos através das abordagens descritas. A maioria dos artigos citados são da década de 2000-2010, onde o principal protocolo de aplicação da toxicogenômica envolvia o uso de dados transcriptômicos (principalmente *chips* de microarranjo).

Afim de se extrair o máximo de informações possíveis dos dados provindos das ciências ômicas e otimizar a avaliação de risco de compostos, o protocolo seguido basicamente envolvia o uso de métodos estatísticos para determinação de listas de genes diferencialmente expressos, listas de funções alteradas (*GO* e *KEGG*) para confecção de hipóteses (NACU et al., 2007). Nos últimos anos, entretanto, a quantidade de dados geradas pelas tecnologias ômicas tem aumentado cada vez mais e além do mais a abordagem descrita acima, embora muito informativa como mostrado em (BUESEN et al., 2017), não foi o suficiente para oferecer toda a otimização da avaliação de risco de compostos esperada inicialmente (SOUZA; KLEINJANS; JENNEN, 2017). Isso acabou dando espaço para o desenvolvimento de um outro tipo de abordagem que é mais utilizada atualmente: a abordagem de construção de redes (já mencionada na última seção).

A razão pela qual essa abordagem se desenvolveu tanto nos últimos anos foi que o uso apenas de listas de anotações funcionais (*GO* e *KEGG*) não consegue modelar todas as complexas interações dentro de um sistema biológico. Isso também significa que essa clássica abordagem não aproveita ao máximo toda a gama de dados gerados através das tecnologias ômicas (genômica, transcriptômica, proteômica e metabolômica), isto é, realizar a integração de toda essa informação.

Dessa maneira as redes biológicas começaram a ser desenvolvidas para maior aproveitamento da informação adquirida (integração) e ao mesmo tempo conseguir modelar o sistema biológico de maneira mais fidedigna em relação a realidade (NACU et al., 2007; SOUZA; KLEINJANS; JENNEN, 2017). Avaliar diversos parâmetros da topologia das redes possibilita a descoberta de informações referentes as interações dos nodos dessas redes, e estes nodos podem representar diversos elementos do sistema biológico: genes, transcritos, proteínas, metabólitos, etc. As interações entre esse nodos podem ter variadas naturezas (geralmente física ou funcional) e é aqui que o grande volume de dados gerados pelas ciências ômicas tem sua importância, na determinação das interações entre os nodos da rede.

As principais redes biológicas que existem no contexto toxicogenômico são:

- Redes de interação proteína-proteína
- Redes de co-regulação (interação DNA-proteína)
- Redes de co-expressão (interação transcrito-transcrito)
- Redes de metabólitos
- Redes de sinalização

Existem basicamente duas maneiras de realizar um estudo com redes biológicas, a primeira é utilizar uma rede de interações já construída (redes de interação proteína-proteína do STRING, por exemplo) e projetar sobre ela dados adicionais (expressão gênica, expressão de proteínas, concentrações, etc). A segunda é a predição da própria rede de interações a partir dos dados gerados, uma revisão dos métodos para predição de redes pode ser encontrada em (NACU et al., 2007; CHARITOU; BRYAN; LYNN, 2016).

Em qualquer caso essas redes são utilizadas para diversas finalidades dependendo dos seus elementos constituintes e do tipo de interação entre eles. Os parâmetros avaliados para análise, entretanto, são os mesmos. Os nodos “*hub*”, por exemplo, são elementos da rede com alta conectividade e são um dos parâmetros mais investigados nessas redes. A razão disso é por que pode se imaginar que um nodo altamente conectado ira inevitavelmente influenciar um grande número de processos. De fato, as redes biológicas de interações moleculares apresentam um decaimento exponencial no que se refere ao número de nodos com elevado número de interações, pois estas redes são mais resistentes a perturbações aleatórias

(ALBERT; JEONG; BARABÁSI, 2000). Estudos mostram que os nodos *hub* são alvos preferencias de patógenos (DYER; MURALI; SOBRAL, 2008; SCHLEKER; TRILLING, 2013) e que o câncer tende a afetar principalmente nodos mais conectados (JONSSON; BATES, 2006), embora isso aparentemente só seja válido para alguns tipos de câncer como sugerido em (SCHAEFER; SERRANO; ANDRADE-NAVARRO, 2015). Os perfis topologicos, nesse caso o universo de nodos *hub*, podem ser utilizados como biomarcadores moleculares para diversas condições, em (KILLICK et al., 2014) é mostrado que as proteínas *hub* são diferentes entre redes de macrófagos bovinos para infecções de *Mycobacterium bovis* virulento e não virulento.

Outro parâmetro muito avaliado nessas redes são os nodos *bottleneck*, nodos com elevado *betweenness*, ou seja, os nodos ligados pela menor distância em uma determinada rede (YU et al., 2007). Nodos mais próximos podem estar envolvidos em processos de comunicação, pois uma menor distância facilita o fluxo de informações (CHARITOU; BRYAN; LYNN, 2016). Os nodos *bottleneck* também são alvos preferenciais de patógenos e outras doenças, de fato muitas vezes, dentro de uma determinada rede, nodos *bottleneck* tendem a ser também os nodos *hub*, como mostrado em (LAWLESS et al., 2014) onde redes construídas com genes diferencialmente expressos de monócitos isolados de leite a 36 horas pós-infecção com *S. uberis* mostravam que os top 20 nodos *hub* também eram os top 20 nodos *bottleneck*.

Outro parâmetro muito utilizado na análise de redes biológicas é a modularidade, a maioria das redes biológicas é modular em natureza, de maneira tal que nodos dentro um mesmo módulo tendem a ser enriquecidos para as mesmas funções biológicas (WATTS; STROGATZ, 1998; HARTWELL et al., 1999; CHARITOU; BRYAN; LYNN, 2016). Proteínas ou genes que atuam na mesma doença ou em doenças com fenótipos parecidos tendem a interagir preferencialmente um com o outro no módulo referente a tal condição (BARABÁSI; GULBAHCE; LOSCALZO, 2011).

Essa abordagem de módulos consegue identificar funções que geralmente não seriam identificadas meramente com a projeção/integração de anotações funcionais com a rede. Portanto utilizar essa abordagem modular em redes associadas a doenças pode então detectar novos genes/proteínas em módulos associados com esta condição, de maneira que as redes possam servir para explicação dos mecanismos moleculares por trás dos fenótipos da doença em questão ou como potenciais biomarcadores desta, de fato a topologia geral desse tal módulo pode ser um biomarcador para a doença (CHARITOU; BRYAN; LYNN, 2016).

Exemplos de uso da avaliação de redes por modularidade são variados, por exemplo, em *O’Roak et al* é mostrado que 39% das mutações *de novo* ou disruptivas em genes associados com o autismo segundo sequenciamentos do exoma estavam altamente interconectados em um módulo relacionado com o remodelamento da cromatina/ β catenina. Em (LAGE et al., 2012), é mostrado como fatores de risco para desenvolvimento de doenças cardíacas congênitas convergem funcionalmente para módulos associados com a regulação do desenvolvimento cardíaco. a análise dos módulos também tem ganho popularidade em estudos genômicos com animais, particularmente na identificação de módulos funcionalmente relevantes em redes de co-expressão (FREEMAN et al., 2012; MALIK et al., 2015).

Naturalmente ainda existem muitos desafios para serem superados nessa abordagem de redes, além das limitações estatísticas, existe um problema fundamental, o interatoma dos organismos não é estático como genoma. Dessa maneira uma rede construída a partir de dados de um experimento vai representar o sistema biológico naquele momento apenas. Entretanto o desenvolvimento de novas tecnologias de alto desempenho para facilitar a geração de dados e de tecnologias que permitam avaliar o sistema biológico em diversas condições (*knock-in* ou *knock-out* de transgenes) e no nível de uma única célula (*single cell seq*) vai permitir a construção de modelagens do sistema biológico cada vez mais próximas da realidade. De fato o próximo desafio da toxicogenômica (além de lidar com uma quantidade ainda maior de dados) é a determinação de modelos ou *design* que possa avaliar o sistema biológico em níveis que incluem: múltiplas gerações, período de vida e célula única (*single cell*) (RUDEN; GURDZIEL; ASCHNER, 2017).

1.3 Microarranjo (*microarray*)

Este trabalho visa utilizar dados de *chips* de microarranjo, portanto esta sessão tem o objetivo de explorar os princípios desta técnica, o experimento toxicogenômico do *chip* de microarranjo em todas as suas etapas e da análise dos dados, também abordaremos os problemas e limitações relacionados.

1.3.1 Princípios da Técnica

Os *chips* de microarranjo são capazes de fornecer informações a respeito das quantidades relativas de transcritos celulares, ou seja, permite a avaliação do transcriptoma de uma célula ou tecido. A vantagem dos *chips* de microarranjo em relação a outras técnicas que também trabalham avaliação de transcritos celulares, como o *Northern Blot* por exemplo, é que o *chip* de microarranjo é capaz de fornecer informação de diversos transcritos simultaneamente. *Chips* de microarranjo são compostos por diversas sondas com oligonucleotídeos, as quais são utilizadas para hibridização com uma biblioteca de cDNA criada a partir de mRNA das amostras. Hibridização é o processo de pareamento de sequências de nucleotídeos complementares, no caso dos *chips* de microarranjo isso se refere ao pareamento de bases entre os oligonucleotídeos da sonda e as sequências do cDNA. Após a hibridização o *chip* é lavado e as sondas com o material hibridizado são utilizadas para determinação de intensidades luminosas, esse sinal luminoso constitui a natureza quantitativa dos dados obtidos em um ensaio de microarranjo e está diretamente relacionado com a quantidade relativa de mRNA presente nas amostras caso e controle. Os métodos de sintetização dos oligonucleotídeos nas superfícies das sondas, o material da superfície e a maneira de se distinguir as intensidades do caso e do controle variam de acordo com o *chip* utilizado (NRC et al., 2007; HANDE, 2017).

Abaixo, na figura 1.3, temos um fluxograma demonstrando as etapas de um experimento de microarranjo.

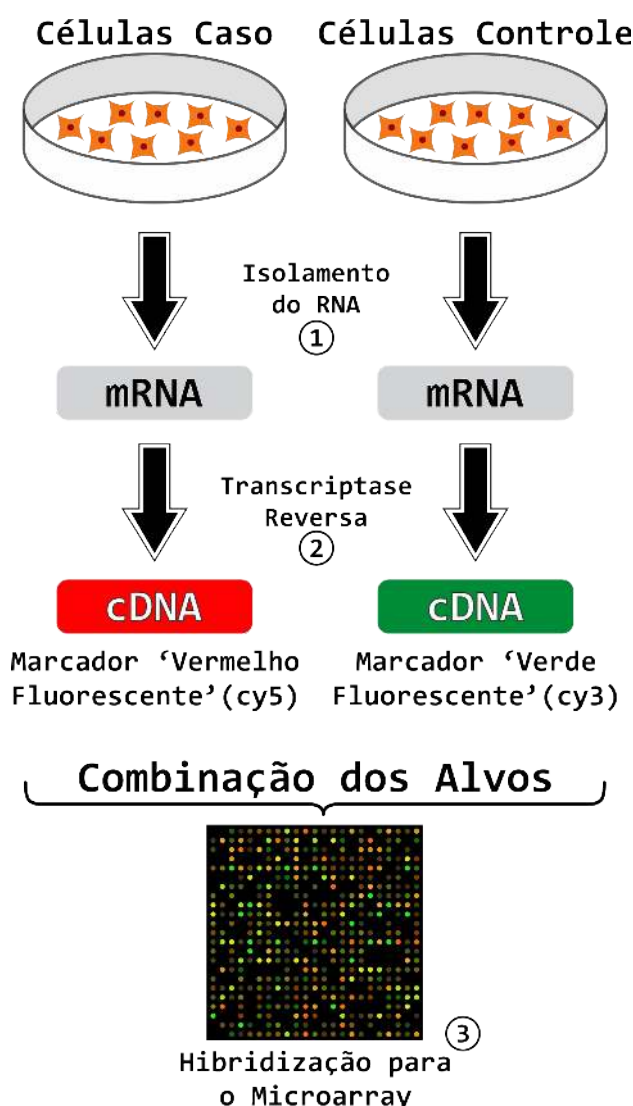


Figura 1.3: Ilustração das etapas de um experimento de microarray de dois canais (duas cores). Aqui duas amostras, uma caso e a outra o controle, tem seu mRNA extraído, isolado e utilizado para construção de uma biblioteca de cDNA (DNA codificante sem íntrons) marcada. Geralmente essa marcação é realizada via uso de fluoróforos, moléculas que emitem ondas eletromagnéticas no comprimento da luz visível (390 nm a 700 nm). Os fluoróforos mais utilizados são o Cy3 emitindo luz verde (570 nm) e o Cy5 que emite luz vermelha (670 nm) (HOEN et al., 2003). O cDNA então é hibridizado com as sondas no *chip* de microarray. O *chip* então é lavado e colocado para leitura num scanner para registrar as intensidades luminosas de cada sonda. Analisando as intensidades luminosas de cada cor é possível determinar o quão sub ou superexpresso um gene esta no caso em relação ao seu controle, já que a intensidade luminosa está diretamente associada a abundância de mRNA em cada situação (caso e controle) (NRC et al., 2007; HANDE, 2017). Existe também o microarray de um canal só, suas etapas são idênticas ao microarray de dois canais, a diferença é que para comparar um caso e um controle são necessários dois chips de microarray. Essa é uma estratégia interessante, adotada pela linha GeneChip® da affymetrix, já que elimina o problema das escalas de intensidades de cores diferentes que existe no microarray de duas cores. Outros problemas atenuados nesta linha incluem hibridização não específica, o chip da affymetrix por exemplo tem para cada sonda milhares de cópias complementares à um tipo de oligonucleotídeo de 25 pb (pares de bases) (essa sonda é chamada de *Perfect Match*), cada sonda PM tem uma sonda *mismatch* (MM) correspondente que contém milhares de cópias do mesmo oligonucleotídeo da sonda PM, porém com a alteração de um nucleotídeo (o décimo terceiro). Assim as sondas MM podem ser utilizadas para avaliar a hibridização não específica de cDNA e servir como uma maneira de quantificar o ruído (DALMA-WEISZHAUSZ et al., 2006). Adaptado de (BIAZI-JUNIOR,).

1.3.2 *Design* do Experimento de Expressão Gênica e Análise de Dados: Aplicações, Exemplos, Limitações e Perspectivas

Os passos para realizar um estudo toxicogenômico qualquer com dados de expressão gênica (seja com chips de microarray ou RNA-seq) são idênticos ao do *workflow* mostrado na figura 1.2. Abaixo, na figura 1.4, podemos ver um outro *workflow* descrevendo os passos necessários para realização de um experimento toxicogenômico com dados de expressão gênica juntamente com algumas possíveis análises.

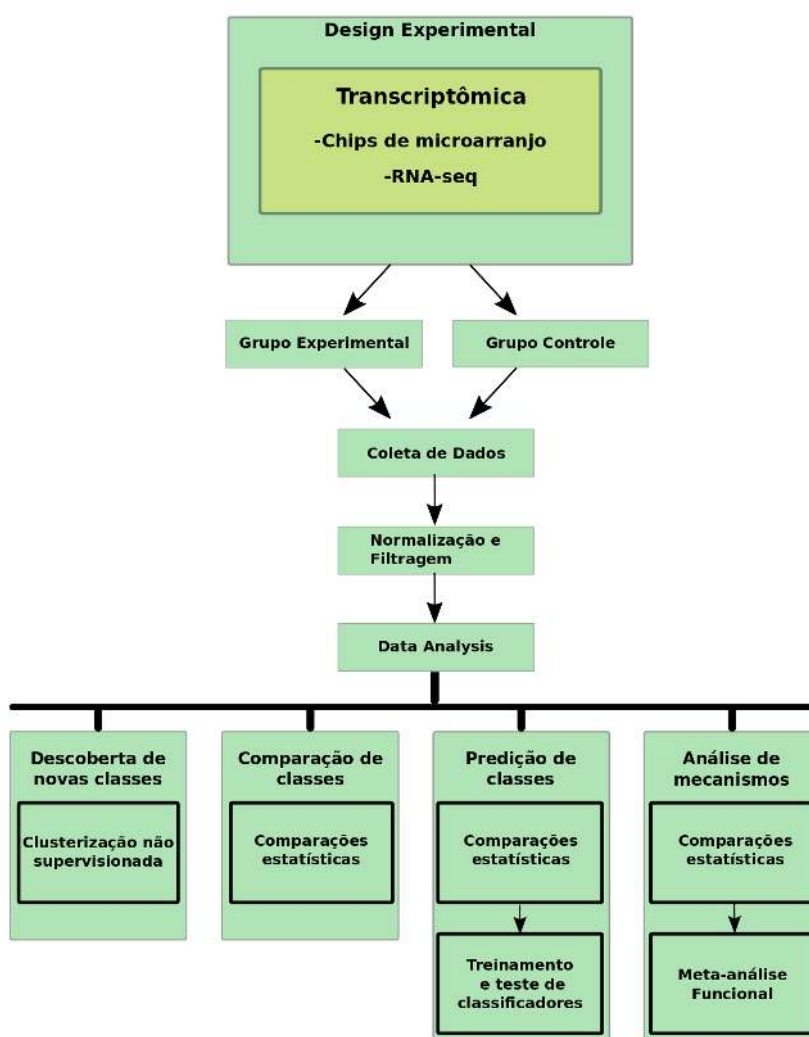


Figura 1.4: *Workflow* de um experimento de **toxicogenômica** envolvendo o uso de dados de expressão gênica (transcriptoma) com atenção especial as abordagens mais comuns para análise dos dados obtidos, depois de devidamente normalizados e filtrados. Adaptado de (NRC et al., 2007; HANDE, 2017).

A **descoberta de classes**, como evidente pelo nome, tem foco em determinar novas classes de agentes tóxicos. Um exemplo simples seria um experimento que administrasse vários fármacos nefrotóxicos nos rins de ratos, individualmente, como feito em (AMIN et al., 2004). Os dados de expressão gênica são então coletados após os animais demonstrarem sinais de falência renal. O uso de descoberta de classes nestes dados é capaz de agrupar (clusterizar) esses fármacos de acordo com diferentes fatores como o tipo de célula afetada ou os mecanismos envolvidos, como consequência seria possível descobrir novos subgrupos de agentes nefrotóxicos (seja em relação a célula alvo ou ao mecanismo de ação).

A descoberta de classes utiliza de métodos de **clusterização não supervisionados** para realizar o agrupamento de perfis de expressão gênica similares (NRC et al., 2007; HANDE, 2017). É importante salientar que esse tipo de clusterização, não utiliza dados categorizados ou classificados como *input*, ou seja, no exemplo do estudo da nefrotoxicidade de fármacos em ratos, estes métodos não levariam em consideração as condições de cada amostra (droga, dose e tempo de amostragem por exemplo). Este tipo de clusterização apenas agrupa perfis de expressão gênica de acordo com alguma medida de similaridade (NRC et al., 2007; HANDE, 2017). Muitos métodos foram desenvolvidos para realizar este tipo de análise: clusterização hierarquica (WEINSTEIN et al., 1997; EISEN et al., 1998; WEN et al., 1998), agrupamento *K-means* (SOUKAS et al., 2000), Mapas de Kohonen (TAMAYO et al., 1999; TÖRÖNEN et al., 1999; WANG et al., 2002; HERRERO; VALENCIA; DOPAZO, 2001), *relevance networks* (BUTTE; KOHANE, 1999), *force-directed layout* (KIM et al., 2001) e análise de componente principal (PCA) (RAYCHAUDHURI; STUART; ALTMAN, 2000).

Fundamentalmente cada um desses métodos utilizam de alguma característica dos dados de entrada e um método de mensurar similaridade desta característica para então agrupar conjuntos que sejam similares em relação a essa determinada característica. Um problema que merece atenção nesse tipo de abordagem é a reprodutibilidade dos resultados, variações técnicas no *design* experimental e na análise dos dados podem gerar variação exarcebada em resultados de amostras em condições similares (NRC et al., 2007; HANDE, 2017). Entretanto alguns relatórios sugerem que o uso de práticas laboratoriais padrões e a análise cuidadosa dos dados pode levar a resultados reprodutíveis de alta qualidade que de fato refletem a biologia do sistema observado (WEIS, 2005; DOBBIN et al., 2005; IRIZARRY et al., 2005; LARKIN et al., 2005).

Por fim pode-se dizer que o objetivo principal da descoberta de classes na **toxico-genômica** é a elucidação de genes e mecanismos de ação envolvidos na resposta celular de sistemas biológicos a um agente tóxico. Entretanto este é um estudo cujo objetivo é a determinação de hipóteses. Experimentos subsequentes geralmente são necessários para a validação dos resultados de um estudo de descoberta de classes.

A **comparação de classes** utiliza comparações estatísticas de dois estados fenotípicos diferentes para determinação de genes ou perfís de expressão gênica que melhor identificam ou separam estes dois estados. Naturalmente aqui são necessários métodos estatísticos supervisionados, ou seja, é necessário conhecimento a priori das classes as quais as amostras pertencem, ao contrário dos métodos de clusterização não supervisionados utilizados na descoberta de classes.

Um exemplo simples de aplicação dessa abordagem: imagine um conjunto de dados de expressão gênica referentes a tratamentos de fármacos hepatotóxicos e cardiotoxicos. Quais genes ou perfís de expressão gênica melhor distinguem essas duas classes de dados? A comparação de classes é uma abordagem para responder esse tipo de pergunta. Muitos testes estatísticos são utilizados para determinar diferenças entre as classes como teste t e Mann-Whitney (para duas classes) e análise de variância “ANOVA” (para três ou mais classes) (DRĂGHICI, 2016).

Estes testes determinam p-valores para os genes, baseado na sua capacidade de diferenciar as classes em questão. Naturalmente existem problemas no que se refere a confiabilidade dos resultados, isto por que cada um dos testes mencionados toma como base diversas suposições sobre os dados, que na realidade podem não ser reais e como consequência erros do tipo I (falso positivo) e do tipo II (falso negativo) podem ser cometidos (DRĂGHICI, 2016).

Problemas comuns envolvem a distribuição dos dados (normal ou não), a homogeneidade da variância dos dados entre os grupos e a questão da dependência ou independência das variáveis. Outro problema nesse tipo de estudo é o problema de testes múltiplos, conforme o número de hipóteses a serem testadas aumenta a chance de um acontecimento raro aparecer (um *outlier*) também aumenta e por extensão as chances de se rejeitar incorretamente a hipótese nula to teste (erros do tipo I) também aumenta.

Existem, na literatura, alguns testes feitos para tentar contornar este problema como em (TUSHER; TIBSHIRANI; CHU, 2001), onde um método (SAM–*Significance Analysis of Microarray*) que utiliza um sistema que atribui um *score* a cada gene baseado na sua expressão relativa e no desvio padrão das medidas das réplicas juntamente com um teste de FDR (*False Discovery Rate*; taxa de descobertas falsas) para determinação de genes diferencialmente expressos é desenvolvido. Outra abordagem utilizada para esse problema é a aplicação da correção de Bonferroni (ARMSTRONG, 2014). Entretanto, deve-se salientar que esse métodos de correção para o problema de múltiplos testes assumem independência entre os grupos, algo que não é real pois genes interagem entre si ou com produtos de outros genes (NRC et al., 2007; HANDE, 2017).

Ainda não existe um teste ideal que pode ser aplicado a qualquer tipo de conjunto de dados sem esperar erros do tipo I e II, toda escolha de testes é fundamentada na determinação de um equilíbrio entre erros do tipo I e tipo II, esta é uma área que ainda está sobre pesquisa constante. Em conclusão, apesar de todos os problemas existentes, os testes estatísticos são uma ferramenta útil para priorizar certos genes para realizar análises subsequentes. Por essa razão o principal objetivo da comparação entre classes é determinar conjuntos de genes que os dados indicam ser úteis na diferenciação dos grupos em questão. Os resultados desta abordagem são úteis como ponto de partida para estudo dos mecanismos de ação ou para classificação de compostos de acordo com seu mecanismo de ação (NRC et al., 2007; HANDE, 2017).

A **predição de classes** tem como objetivo a predição de efeitos biológicos baseado em perfis de expressão gênica. Fenótipos similares possuem padrões de expressão gênica similares, tipicamente esse tipo de estudo necessita de um estudo de comparação de classes como ponto de partida. Um conjunto de dados com classes bem definidas passa por um estudo de comparação de classes para determinação de conjuntos de genes que melhor conseguem distinguir as classes envolvidas. Esses genes são então utilizados como *input* a diversos tipos de algoritmos para determinação de regras que podem prever os possíveis fenótipos de um composto baseado no seu perfil de expressão gênica. Por exemplo, seria possível utilizar uma base de dados de expressão gênica de compostos neurotóxicos para determinação de padrões de expressão gênica altamente correlacionados com neurotoxicidade, assim poderíamos usar essas regras para predição de possível neurotoxicidade em um novo composto (NRC et al., 2007; HANDE, 2017).

Os métodos utilizados para confecção de algoritmos para determinação de regras são variados, tipicamente envolvem o uso de *machine learning* e redes neurais. Alguns desses métodos serão citados juntamente com exemplos de suas aplicações e a literatura pertinente: *weighted voting* (GOLUB et al., 1999), redes neurais artificiais (ELLIS et al., 2002; BLOOM et al., 2004), análise discriminante (NGUYEN; ROCKE, 2002; ORR; SCHERF, 2002; ANTONIADIS; LAMBERT-LACROIX; LEBLANC, 2003; LE et al., 2003), árvores de classificação e regressão (BOULESTEIX; TUTZ; STRIMMER, 2003), máquina de vetores de suporte (SVM) (BROWN et al., 2000; RAMASWAMY et al., 2001) e k-vizinhos mais próximos (THEILHABER et al., 2002).

A **análise de mecanismos** tem como objetivo a determinação de mecanismos de ação associados com certos fenótipos. Embora a descoberta de classes, a comparação de classes e a predição de classes possam fornecer informações sobre o fenótipo dos compostos (ou o fenótipo esperado), elas não necessariamente fornecem informações sobre o principal mecanismo de ação responsável por tal fenótipo (NRC et al., 2007; HANDE, 2017).

A bioinformática tem papel essencial no desenvolvimento de métodos para a integração de dados toxicogenômicos aos seus respectivos fenótipos, algumas bases de dados com esse tipo de abordagem em mente já existem, um exemplo é o já mencionado *Gene Ontology* que associa genes a termos relacionados com seus processos biológicos (BP), funções moleculares (MF) e compartimentos celulares (CC) (ASHBURNER et al., 2000). Outras bases de dados interessantes são o STRING, onde redes de interação proteína-proteína de diversas espécies estão disponíveis para *download* (SZKLARCZYK et al., 2014), e o KEGG onde se encontra uma base de dados sobre vias metabólicas de diversos organismos (KANEHISA; GOTO, 2000).

Outros tipos de informação que podem ser utilizados incluem: produtos gênicos, dados sobre mapeamentos genéticos, informações sobre estrutura e atividade de genes e proteínas, relação dose-resposta de compostos, informações fenotípicas ou clínicas e sequência de genomas. Alguns *softwares* que integram diferentes dados toxicogenômicos já estão disponíveis para uso, alguns exemplos são: MAPPFinder (DONIGER et al., 2003), GOMiner (ZEEBERG et al., 2003) e EASE (HOSACK et al., 2003). Porém é importante salientar que ainda não existe um método universalmente aceito para a integração de dados de expressão gênica, proteínas e metabólitos a vias metabólicas funcionalmente relevantes a seus respectivos fenótipos (NRC et al., 2007; HANDE, 2017).

Esse tipo de abordagem, recentemente, tenta integrar diversos tipos de dados toxicogenômicos em redes de interação (genoma, transcriptoma, proteoma e metaboloma). Modelos recentes e sua literatura pertinente incluem: redes booleanas (AKUTSU; MIYANO; KUHARA, 2000; SAVOIE et al., 2003; SOINOV, 2003), redes booleanas probabilísticas (SHMULEVICH et al., 2002; DATTA et al., 2004; HASHIMOTO et al., 2004) e redes bayesianas (FRIEDMAN et al., 2000; IMOTO et al., 2003; SAVOIE et al., 2003; TAMADA et al., 2003; ZOU; CONZEN, 2004). A base desses modelos de predição de interações numa rede é representar os objetos (genes, proteínas e metabólitos) como nodos e as interações como as arestas que ligam esses nodos. Um conjunto de regras, particular a cada método, é utilizado para determinar o poder de interação entre os nodos e como extensão disso, se uma determinada consequência será induzida a partir dessa interação. Esses modelos ainda necessitam de muito desenvolvimento para deixarem de ser descritivos, em sua maior parte, para se tornarem de fato preditivos (NRC et al., 2007; HANDE, 2017).

Alguns pesquisadores sugerem que o uso de determinadas redes podem de fato ter um bom poder preditivo como em (WIBACK; MAHADEVAN; PALSSON, 2004; FAMILI; MAHADEVAN; PALSSON, 2005), onde o monitoramento do fluxo metabólico e outras ciências ômicas são utilizadas para modelagem de redes de interações. Essa abordagem de se integrar diversos tipos de dados biológicos para predição de interação entre seus objetos e de suas consequências recebe o nome de “biologia de sistemas” e é justamente essa integração de diversos dados sobre variados aspectos que lhe dá o potencial de gerar modelos que representem um sistema biológico mais próximo da realidade (NRC et al., 2007; HANDE, 2017). A tendência de se construir redes de interação para compreensão funcional do sistema biológico é discutida em detalhes em Barábasi *et al* (BARABASI; OLTVAI, 2004).

Isto não significa que informações toxicogenômicas obtidas de outras abordagens (descoberta de classes, comparação de classes e predição de classes) sejam dispensáveis, longe disso, mas ter a capacidade de modelar de maneira fidedigna toda a complexa rede de interações entre genes, transcritos celulares, proteínas, vias metabólicas e integrar tal modelo com dados tóxicogenômicos (perfis de expressão gênica, por exemplo) para compreender toda a sequência de mecanismos envolvidas no desenvolvimento de um determinado fenótipo é um objetivo fundamental da **toxicogenômica**. A principal limitação nesse tipo de abordagem é a capacidade de modelar o sistema de interações biológicas de diversos objetos, sendo que a transdução de sinais bioquímicos é especialmente complexa (NRC et al., 2007; HANDE,

2017).

Algumas iniciativas já tentaram realizar a modelagem do fluxo metabólico em sistemas mais simples como bactérias e leveduras, mas ainda assim somente aproximações grosseiras foram capazes de ser realizadas a respeito das respostas funcionais do sistema, e isso apenas sobre condições extremamente controladas. Por fim pode-se concluir que este tipo de abordagem ainda necessita de muito desenvolvimento, seja na construção de bases de dados de informações toxicogenômicas ou na modelagem do sistema de interações biológicas no organismo, para realizar a predição de fenótipos e dos mecanismos de ação induzidos por agentes tóxicos (NRC et al., 2007; HANDE, 2017).

Cabe lembrar aqui que todas as quatro abordagens explanadas de maneira geral acima, no final, tem como função principal a **busca de hipóteses** e não a validação das mesmas. Os experimentos de bancada continuarão sendo a principal fonte de validação de hipóteses. Entretanto o desenvolvimento de modelos cada vez mais robustos e próximos da realidade irá limitar cada vez mais o número de hipóteses a serem testadas em experimentos de bancada, oque por sua vez irá reduzir custos em pesquisas científicas ou no desenvolvimento de drogas.

1.4 O Projeto Toxicogenômico Japonês Open TG-GATEs e os Dados Selecionados

Este projeto durou 5 anos (2002-2007) e foi realizado pelo governo japonês em conjunto com 15 empresas farmacêuticas, metade dos custos foram fomentados pelo governo e a outra metade pelas empresas. O projeto criou uma base de dados para armazenamento de dados de expressão gênica de chips de microarranjo e dados toxicológicos referentes a 150 substâncias (a maioria são drogas utilizadas em tratamentos) para estudos *in vivo* e *in vitro* para *Rattus norvegicus* e *in vitro* em *Homo sapiens* utilizando como plataforma o *Affymetrix GeneChip®* segundo (UEHARA et al., 2010), entretanto os metadados (<http://toxico.nibiohn.go.jp/english/datalist.html>) da base de dados indicam o uso de até 170 compostos. A finalidade era fornecer a pesquisadores uma base de dados toxicogenômicos e toxicológicos referentes aos efeitos adversos dos compostos testados, para possibilitar a confecção de variados estudos para busca de biomarcadores, determinação de mecanismos de ação e o desenvolvimento de novos métodos de análise. O *Rattus norvegicus* foi escolhido por ser o organismo modelo que mais foi utilizado para avaliações toxicológicas em estudos

pré-clínicos, como consequência existe uma abundância de dados toxicológicos referentes a esse organismo. O fígado foi o alvo principal desse projeto uma vez que a maioria dos efeitos adversos clinicamente relevantes acontecem no fígado, embora os rins também tenham sido órgãos alvos para a construção do OPEN TG-GATES. Outro fator importante é que o fígado possui uma composição celular bastante homogênea, o que minimiza a variância natural causada pela diferença das amostras (UEHARA et al., 2010). Os estudos *in vivo* foram realizados somente no *Rattus norvegicus*, no fígado, nos rins e no plasma com doses únicas e com doses repetidas. Os estudos *in vitro* foram realizados apenas com doses únicas em culturas de hepatócitos para o *Rattus norvegicus* e para o *Homo sapiens*.

1.4.1 OPEN TG-GATES: Design Experimental

1.4.1.1 Estudos *In vivo*

Os experimentos com 170 fármacos testados *in vivo* para a *Rattus norvegicus* foram realizados em laboratórios diferentes. Os animais foram comprados no mesmo instituto (Charles River Japan Inc.; Kanagawa, Japan) com 5 semanas de idade e posteriormente encaminhados a laboratórios diferentes onde passaram por um período de aclimação de uma semana em ambientes/condições controlados: jaulas de aço, iluminação por 12 horas (7:00;19:00), temperatura mantida entre 21°C e 25°C, ventilação com troca de ar 15 vezes por hora, humidade mantida entre 40%-70% e livre acesso a água e comida. Após a aclimação de todos os animais foi realizada sua separação em diversos grupos de 5 animais para administração dos compostos. Entende-se por “grupo” a combinação das informações: composto, tipo de estudo, dose e tempo. Os animais eram divididos em seus grupos de acordo com um algoritmo de estratificação randômica que considerava o peso dos animais (TAKASHIMA et al., 2006). Tanto para o estudo *in vivo* com dose única quanto o de dose repetida foram coletados dados de expressão gênica em quatro doses (incluindo o controle/veículo) e quatro pontos no tempo, outras informações coletadas em ambos estudos *in vivo* incluem: histopatologia dos órgãos de interesse, peso do animal, peso do órgão, consumo de alimentos, hematologia e bioquímica do sangue. Dessa maneira cada caso do tipo composto-estudo-dose continham em si quatro grupos (um para cada ponto no tempo) e um total de 20 animais. Cada dose de cada composto com seu veículo era administrada a 20 animais e a cada ponto no tempo 5 animais eram sacrificados (TAKASHIMA et al., 2006), a figura 1.5 abaixo contém um esquema ilustrando a cronologia dos estudos *in vivo* no *Rattus*

norvegicus. É difícil determinar a quantidade exata de animais que foram utilizados para a construção da base de dados pois entre os estudos *in vivo* realizados, com dosagem única e repetida, existem variações quanto as drogas que foram utilizadas em cada estudo e quanto aos órgãos estudados em cada droga.

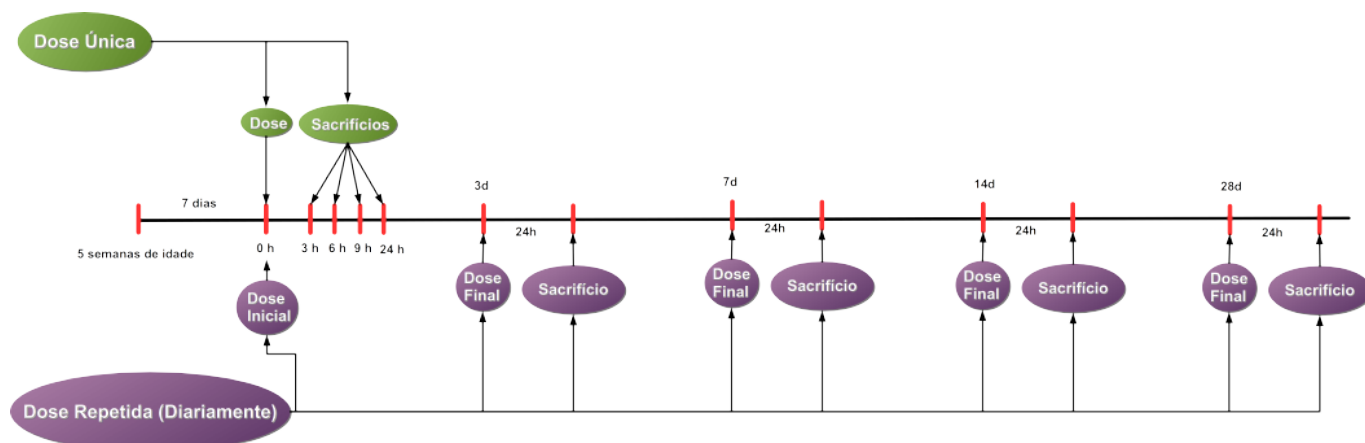


Figura 1.5: Cronologia dos experimentos *in vivo* no *Rattus norvegicus*. Adaptado de (TAKASHIMA et al., 2006)

A administração de compostos era realizada entre as 9:00 e as 11:30 e para evitar efeitos do ciclo circadiano, o sacrifício e necrópsia também eram realizados nesse horário. O peso dos animais era registrado todos os dias, o consumo de alimentos era registrado a cada 4 dias nos estudos *in vivo* com dose repetida. As amostras de sangue para realização da hematologia e da avaliação da bioquímica do sangue foram coletadas no dia do sacrifício (TAKASHIMA et al., 2006). Após a coleta de sangue os animais eram sacrificados por exsanguinação da aorta abdominal sob anestesia com éter e posteriormente amostras dos órgãos de interesse eram coletadas para realização dos estudos de histopatologia e de expressão gênica.

O protocolo de amostragem era baseado na extração de 3 amostras de cada órgão de interesse para cada um dos 5 animais de cada grupo. Para os estudos de histopatologia no fígado, o sítio de amostragem foi escolhido de maneira a evitar proximidade com a veia porta hepática e proximidade com a área parenquimal do lobo límbico onde o tecido é mais consistente. Se a porção central dos fragmentos extraídos apresentasse sinais de danos mediante inspeção visual, então o procedimento de amostragem era repetido em uma localização distal do sítio de amostragem anterior no fígado. No rim foi realizado um procedimento de amostragem parecido, 3 amostras do rim esquerdo foram obtidas sendo que eram extraídos fragmentos com 1 mm de espessura (que inclui a região medular e porção

do córtex) utilizando-se quatro laminais alinhadas contra o eixo longo do rim. Caso o rim esquerdo apresentasse sinais de danos mediante inspeção visual o mesmo procedimento de amostragem era aplicado ao rim direito. Imagens dos tecidos também fazem parte da base de dados OPEN TG-GATEs (TAKASHIMA et al., 2006).

Os experimentos de expressão gênica nos estudos *in vivo* foram realizados em triplicatas com três dos cinco ratos de cada grupo, sendo que os animais escolhidos eram aqueles que estavam mais próximos do peso médio do grupo, utilizando inicialmente como plataforma o *GeneChip*[®] RAE230A (Affymetrix, Santa Clara, California, USA) com 15923 sondas para cobrir todo o genoma do organismo, posteriormente a plataforma foi trocada pelo *GeneChip*[®] RAE2302 (Affymetrix, Santa Clara, California, USA) com 31099 sondas. Para obtenção do RNA foram coletadas alíquotas das amostras dos órgãos (procedimento descrito no último parágrafo) de aproximadamente 30 mg no mesmo dia do sacrifício e armazenadas no *RNAlater*[®] (Ambion, Austin, TX, USA) pela noite subsequente a 4°C e posteriormente congelados a -80°C e enviadas para as instalações do instituto nacional de ciências da saúde para continuação dos experimentos (UEHARA et al., 2008; UEHARA et al., 2010; UEHARA et al., 2011; UEHARA et al., 2013). O RNA total das amostras foi extraído com o *RNeasy kit* da *Bio Robot 3000* (Qiagen, Valencia, CA, USA), a homogeneização das amostras foi realizada com o *Mill Mixer* (Qiagen) com contas de zircônio. A pureza do RNA era confirmada com eletroforese em gel para avaliar se a razão 260/280 nm estava entre 2.0 e 2.2. Posteriormente o isolamento do RNA foi realizado seguindo as instruções do fabricante do kit (TAKASHIMA et al., 2006).

A geração da biblioteca de cDNA (DNA codificante) foi realizada de acordo com as instruções dos fabricantes do equipamento utilizado que incluía:

- O *Superscript Choice System* (Invitrogen, Carlsbad, CA, USA) e o *primer* T7-(dT) 24-oligonucleotide (Affymetrix) para síntese de cDNA a partir do RNA isolado.
- O *cDNA Cleanup Module* (Affymetrix) para purificação das bibliotecas.
- O *BioArray High yield RNA Transcript Labeling Kit* (Enzo Diagnostics, Farmingdale, NY, USA) para síntese de cRNA rotulado com biotina (fluoróforos).

Posteriormente a biblioteca de cRNA era utilizada para hibridização com as sondas nas superfícies dos *chips* (*Affymetrix GeneChip*® RAE2302 nesse caso), os chips eram lavados e colocados em um *scanner* para registrar as intensidades luminosas das sondas. Os dados brutos estão disponíveis para download no sítio <https://dbarchive.biosciencedbc.jp/en/open-tggates/data-2.html> no formato “.CEL”.

Segue abaixo a tabela 1.1 resumindo o protocolo utilizado nos experimentos *in vivo*.

Tabela 1.1: Resumo do design experimental dos estudos *in vivo*

	In Vivo
Animal	<i>Rattus norvegicus</i> (<i>Sprague-Dawley</i> ; 6 semanas de idade)
Administração	Oral ou Intravenosa
Veículo	Metilcelulose 0.5% ou óleo de milho (oral); Solução salina ou de glicose a 5% (intavenosa)
Dose	Controle, Baixa, Média e Alta
Tipos de Dose	Única e Repetida
Sacrifício	3, 6, 9 e 24 horas após administração em dose única; 24 horas após última administração em 3, 7, 14 e 28 dias
Amostras	Fígado, Rim e Plasma
Items	Histopatologia: fígado e rim, Peso corporal, peso dos órgãos, consumo de alimentos, hematologia, bioquímica do sangue, expressão gênica
Microarranjo	Affymetrix Genechip Rat 2302 (fígado e rim)
Replicatas	3
Medicamentos	160 compostos; 160 no fígado e 41 no rim
Arrays	14143 no fígado; 7378 em dose única e 6765 em dose repetida. 3905 nos rins
Sondas/Array	31099

1.4.1.2 Estudos *In vitro*

Os estudos *in vitro* foram realizados utilizando hepatócitos primários do *Rattus norvegicus* e do *Homo sapiens*, sendo que amostras foram colhidas em quatro doses e três pontos no tempo (não foi realizado estudo com doses repetidas nos estudos *in vitro*). Hepatócitos

Tabela 1.2: Resumo do Design Experimental dos estudos *in vitro*

	In Vitro
Animais	Rattus norvegicus (Sprague-Dawley ; 6 semanas de idade); Homo sapiens (células congeladas)
Administração	Diretamente no meio de cultura
Veículo	Meio de cultura; Dimetilsulfóxido 0,5
Dose	Controle, Baixa, Média e Alta
Tipos de Dose	Única
Sacrifício	2 h, 8 h e 24 h
Amostras	Hepatócitos isolados por digestão com collagenases (Rattus norvegicus); Hepatócitos congelados (Homo sapiens)
Items	Expressão gênica e Viabilidade celular (conteúdo de DNA e liberação de LDH)
Microarranjo	Affymetrix Genechip Rat 2302 (Rattus norvegicus); Affymetrix Genechip HG-U133 plus2 (Homo sapiens)
Replicatas	2
Medicamentos	164 compostos; 145 no Rattus norvegicus e 158 no Homo sapiens
Arrays	3370 no Rattus norvegicus; 2605 para Homo sapiens
Sondas/Array	31099 (Rat 2302); 54675 (HG-U133)

humanos congelados foram comprados da Tissue Transformation Technologies, Inc. (Edison, NJ, USA) e da CellzDirect, Inc. (Pittsboro, NC, USA). Os hepatócitos primários dos ratos foram retirados de animais de 5 semanas de idade, comprados no mesmo instituto que os animais usados em estudos *in vivo*, e utilizados para criar as culturas, os protocolos utilizados para a criação das culturas podem ser encontrados em (SUMIDA et al., 2011). Para os estudos *in vivo* foi também realizado testes de viabilidade celular, checando o conteúdo de DNA das culturas e a liberação de LDH, os protocolos utilizados para realização destes testes podem ser encontrados no material suplementar de (TAKASHIMA et al., 2006).

A tabela 1.2 abaixo resume o protocolo seguido para os estudos *in vivo*:

1.4.2 Sobre as Drogas

Foi dito que um total de 170 compostos foram estudados durante a construção do OPEN TG-GATEs, entretanto, ao observar o item 'Medicamentos' das tabelas 1.1 e 1.2 pode-se

facilmente chegar a conclusão de que as drogas não estão igualmente distribuídas entre os tipos de estudos na base dados. Certos casos de estudos comparativos dos dados de expressão gênica podem ser inviáveis, por exemplo, se droga foi estudada com dose única mas não repetida (ou vice-versa).

De qualquer forma existe uma abundância de dados de expressão gênica no OPEN TG-GATEs, que podem ser utilizados, em conjunto com as abordagens toxicogenômicas apresentadas nas últimas seções, para uma diversa gama de objetivos:

- Determinação das vias ontológicas (GO) ou *KEGG Pathways* que mais se encontram alterados.
- Determinação das funções mais alteradas por um composto.
- Correlação entre vias alteradas de maneira que seja possível estudar mecanismos de ação por trás dos efeitos (adversos ou não) do composto.
- Estudo da relação dose-resposta dos compostos.
- Encontrar biomarcadores de toxicidade ou de exposição a estes composto.

Tendo em vista que a maioria dos dados são referentes ao fígado, esses dados de expressão gênica podem ser utilizados principalmente para estudo dos efeitos adversos causados pelos compostos. Existem diversas categorias de drogas entre as 170 utilizadas nessa base de dados, entretanto existe uma categoria de drogas que é amplamente utilizada por médicos e que possui uma série de efeitos adversos severos para o usuário: os antineoplásicos.

1.4.3 Antineoplásicos e o OPEN TG-GATEs

Antineoplásicos são medicamentos utilizados no tratamento de neoplasias (tumores) e existem muitos mecanismos de ação que podem ser explorados nesse tipo de tratamento. Cada droga antineoplásica recebe uma classificação de acordo com o seu mecanismo de ação, alguns exemplos: agentes alquilantes, antimetabólitos, alcalóides vegetais e antibióticos antitumorais. A maioria dos mecanismos envolvidos na atividade antitumoral geralmente está envolvida com o impedimento do processo de replicação de DNA e conseqüentemente da divisão celular, de maneira ciclo-celular específica ou não (ALMEIDA et al., 2005). Essas drogas afetam principalmente células que se dividem rapidamente como as células tumorais, mas afetam também diversos tipos celulares saudáveis e esta é a razão pela qual estas drogas causam tantos efeitos adversos. Os efeitos adversos mais comuns relacionados a essas drogas incluem: fraqueza, ansia de vomito, alopecia, etc.

É de interesse das empresas farmaceuticas, ter a capacidade de melhor compreender os mecanismos por trás dos efeitos adversos exibido por essas drogas para melhor poder controla-los/alivia-los. Assim essas empresas podem fornecer uma melhor qualidade de tratamento para os pacientes, e garantir melhor relação entre pacientes e industrias, algo de suma importância nos dias de hoje, pois a capacidade que os pacientes tem de adquirir informações está cada vez maior e as impresas precisasm estar preparadas para lidar com esse perfil de paciente caso queira garantir seu espaço no mercado. Outra utilidade que as abordagens de toxicogenômica podem fornecer a essas empresas é a possibilidade de redução de custos nas fases pré-clínicas do desenvolvimento de novos fármacos, mediante uma melhor capacidade de predição de efeitos adversos de novos compostos.

Existem estudos que se utilizam de *machine learning* ou redes biológicas para estudo de antineoplásicos e de seus efeitos adversos, como em (DONG et al., 2015), onde dados de expressão gênica de drogas antineoplásicas e dados sobre efeitos adversos retirados da base de dados *Cancer Cell Line Encyclopedia* (CCLE) são utilizados juntamente com um algoritmo do tipo SVM (*support vector machine*) para predição de efeitos adversos a partir de seu perfil de expressão gênica. Outro estudo interessante pode ser encontrado em (BUTTE et al., 2000), onde dados de expressão gênica de 60 linhagens de células cancerígenas e dados de drogas antineoplásicas passam por uma filtragem estatística que mensura a correlação entre os elementos dos dois conjuntos de dados. Posteriormente os dados filtrados são então utilizados para construção de redes de relevância, cuja função é auxiliar a descoberta de

relações funcionais entre expressão gênica e atividade anti-tumoral.

Na base de dados OPEN TG-GATEs existem três antineoplásicos que estão presentes simultaneamente nos estudos *in vitro* e *in vivo* do *Rattus norvegicus* e nos estudos *in vitro* do *Homo sapiens* no fígado: a ciclofosfamida, o etopósido e a lomustina. As tabelas 1.3, 1.4 e 1.5 abaixo fornecem mais informações sobre os estudos dessas drogas.

Tabela 1.3: Tabela de relação entre drogas antineoplásicas e estudos realizados no OPEN TG-GATEs

Composto	In vitro		In vivo - Rato Fígado		In vivo - Rato Rim	
	Humano	Rato	Única	Repetida	Única	Repetida
Ciclofosfamida	X	X	X	X	X	X
Etopósido	X	X	X	X		
Lomustina	X	X	X	X	X	X

Tabela 1.3: O X indica presença da droga no estudo em questão, pode-se notar que o etopósido é a única droga que não possui todos os estudos.

Tabela 1.4: Tabela com o protocolo utilizado nos estudos *in vivo* no OPEN TG-GATEs

Composto	Veículo	Rota	Orgão	Dose Única (mg/Kg)			Dose Repetida (mg/Kg)		
				Baixa	Média	Alta	Baixa	Média	Alta
Ciclofosfamida	0.5% MC	Oral	Fígado	15	50	150	1.5	5	15
	0.5% MC	Oral	Rim	1.5	5	15	1.5	5	15
Etopósido	0.5% MC	Oral	Fígado	10	100	1000	3	10	30
Lomustina	0.5% MC	Oral	Fígado	0.6	2	6	0.6	2	6
	0.5% MC	Oral	Rim	0.6	2	6	0.6	2	6

Tabela 1.4: MC = Metilcelulose; note que o etopósido tem estudos realizados apenas no fígado.

Tabela 1.5: Tabela com o protocolo utilizado nos estudos *in vivo* no OPEN TG-GATEs

Composto	Veículo	Célula	Dose Humanos (mg/Kg)			Dose Ratos (mg/Kg)		
			Baixa	Média	Alta	Baixa	Média	Alta
Ciclofosfamida	Meio de cultura	Hepatócito	8	40	200	80	400	200
Etopósido	DMSO	Hepatócito	14	70	350	X	66	330
Lomustina	DMSO	Hepatócito	4.8	24	120	4.8	24	120

Tabela 1.5: DMSO = Dimetilsulfóxido; note que o etopósido teve sua dose baixa omitida no estudo dos ratos.

Segundo (UEHARA et al., 2008), essas doses foram definidas como baixa média e alta de acordo com a literatura e as modalidades de tratamento já existentes.

1.5 Justificativa

Até o momento foi abordado nessa breve revisão: a importância que o material genético e sua expressão tem para regulação dos diversos processos que ocorrem em um sistema biológico; as ciências ômicas e suas tecnologias, que são responsáveis pelo estudo de quatro aspectos fundamentais de um sistema biológico (genôma, transcriptôma, proteôma e o metabolôma); a definição da **toxicogenômica** como uma área que busca integrar e analisar diversos dados oriundos das ciências ômicas para atingir os mais variados objetivos e dos desafios que existem para suas implementações. Foi definido também os passos necessários para realização de um estudo toxicogenômico qualquer e aplicações comuns da **toxicogenômica**; o *chip* de microarranjo como uma ferramenta essencial para caracterização e estudo do transcriptôma e os princípios por trás da técnica; o *design* experimental de estudos toxicogenômicos que utilizam os chips de microarranjo e quatro principais tipos de análises de resultados que existem para ensaios transcriptômicos.

A análise de mecanismos de ação é um dos principais tipos de abordagens utilizados em ensaios transcriptômicos. Uma maneira comum de se realizar esse tipo de estudo é através do uso de bases de dados que contenham informações que liguem um gene a determinadas funções, o *Gene Ontology* é um bom exemplo disso. Para se determinar mecanismos que estão significativamente alterados na presença de uma droga, por exemplo, geralmente a abordagem mais utilizada é a comparação de perfis de expressão gênica de caso e controle.

A partir do conjunto dos dados de expressão gênica, no entanto, a análise visando à mensuração e caracterização do metabolismo celular não é trivial. Utiliza-se de um estado celular controle para agrupar os genes pelo grau de alteração de sua expressão em relação a este controle (CASTRO et al., 2009). Esse é o processo de determinação de genes diferencialmente expressos (DEGs) e é realizado utilizando de algum método estatístico para determinação da significância estatística das alterações dos níveis de expressão do gene entre tratamento e controle e utilizando um critério de avaliação da magnitude dessa variação. Posteriormente é realizado o enriquecimento funcional dos DEGs, o processo de associar os DEGs a suas funções biológicas específicas e de determinar a sua significância estatística. Assim é possível determinar quais funções biológicas se encontram significativamente alteradas e por extensão, é possível deduzir quais mecanismos de ação estão mais ativos no tratamento em relação ao controle.

Esse processo descrito acima recebe o nome de “**enriquecimento funcional de DEGS**” e é considerado como um tipo de “análise clássica” de dados de expressão gênica. No entanto existe uma limitação especial com esse tipo de abordagem, além das dezenas de outras limitações a cerca do *design* experimental, dos métodos estatísticos utilizados e principalmente da confiabilidade da base de dados utilizada. Essa limitação é o fato de que esse tipo de abordagem não considera genes que não estejam diferencialmente expressos para a determinação de funções biológicas significativamente alteradas.

A hipótese tecida aqui é que talvez a utilização de métodos que determinem funções biológicas significativamente alteradas e que não tenham a limitação descrita acima possam fornecer informações adicionais a respeito dos efeitos que drogas, ou qualquer outra condição, possam ter sobre o organismo.

Uma metodologia estatística que contorna o problema da não utilização dos genes não diferencialmente expressos é o **Viacomplex** (CASTRO et al., 2009), que passará a ser referido como análise de grupos de genes funcionalmente associados (AGFAGs). Esse método

agrupa genes de acordo com sua função, utilizando o *Gene Ontology* para isso, e utiliza a atividade conjunta de todos os genes em cada grupo para determinação de funções significativamente alteradas.

Nenhuma dessas duas metodologias incorpora, necessariamente, o uso de redes, que é a atual tendência no que se refere a análise de dados das ciências ômicas. Para então aderir a essa tendência e considerando a hipótese tecida acima, uma metodologia de redes também será considerada.

A proposta deste trabalho é o desenvolvimento de um *pipeline* de análises que compare os resultados das 2 metodologias: enriquecimento funcional de DEGs (EFDEGs) e análise de grupos de genes funcionalmente associados (AGFAGs). Comparar os resultados de metodologias com diferentes abordagens talvez possa elucidar novos mecanismos de ação das drogas as serem estudadas. Esse tipo de abordagem tem o potencial de gerar redução de custos com o desenvolvimento de novos fármacos (principalmente na fase pré-clínica) e de possibilitar uma melhor qualidade de tratamento para os pacientes.

No caso, as drogas selecionadas para realizar esta meta-análise são os antineoplásicos. Essas drogas foram escolhidas principalmente pela grande quantidade de dados disponíveis e pela sua relevância na medicina. Os antineoplásicos são amplamente utilizados em grande escala e frequentemente são a fonte de diversos efeitos adversos nos pacientes, portanto desenvolver metodologias para melhor compreensão do comportamento dessas drogas é de interesse para centros de pesquisa e para empresas farmacêuticas. Além de determinar as funções biológicas mais alteradas, pretende-se abordar a questão “*in vitro* vs *in vivo*”, para determinar as principais diferenças entre os estudos e até onde pode se esperar que um estudo *in vitro* reproduza um estudo *vivo*. Sendo assim, podemos caracterizar o estudo toxicogenômico proposto como uma análise de mecanismos de ação com abordagem exploratória e de também de construção de redes.

Tendo como condição a disponibilidade de estudos no fígado (para estudo de efeitos adversos), os antineoplásicos da base de dados OPEN TG-GATEs que foram escolhidos para serem analisados são: ***Ciclofosfamida***, ***Etopósido*** e ***Lomustina***.

2 *Objetivos*

2.1 Objetivos Principais

O objetivo principal deste trabalho é realizar a meta-análise funcional de antineoplásicos através da associação de dados de expressão gênica da base de dados online OPEN TG-GATEs com as ontologias do *Gene Ontology*: processos biológicos (BPs). A determinação das ontologias significantemente alteradas será feita considerando 2 metodologias: o enriquecimento funcional de DEGs (EFDEGs) e a análise de GFAGs (AGFAGs). No final pretende-se comparar os resultados das duas metodologias entre si para busca de informações adicionais a respeito do comportamento das drogas e analisar a clássica questão : “*in vivo vs in vitro*”.

2.2 Objetivos Específicos

Questões secundárias:

- Avaliar a convergência entre os resultados dos 2 métodos utilizados.
- Comparar todos os resultados obtidos com a literatura.

3 Materiais e Metodos

3.1 Workflow

Neste projeto, propomos a análise de dados de expressão gênica de fígado de *R. norvegicus* e *Homo sapiens* para um grupo de drogas em três tipos de estudo. Estes dados estão disponíveis no banco de dados do projeto toxicogenômico japonês (OPEN TG-GATEs). Para determinação das ontologias alteradas são propostas 2 metodologias: *i*) enriquecimento funcional de genes diferencialmente expressos (EFDEGs); *ii*) análise de grupos de genes funcionalmente associados (AGFAGs).

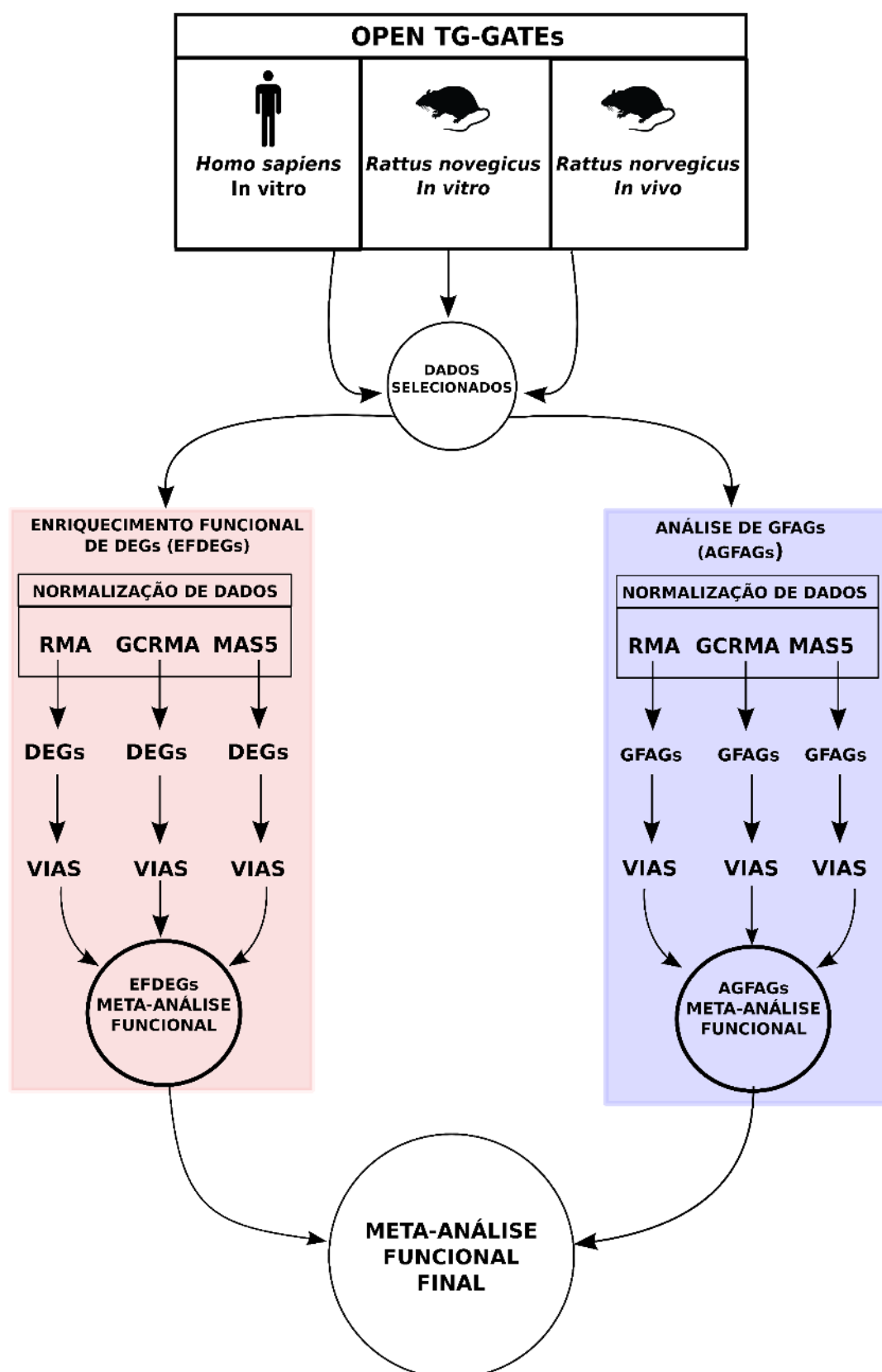


Figura 3.1: *Workflow:* Primeiramente podemos observar a base de dados online OPEN TG-GATEs, que contém dados de expressão gênica de diversas drogas em três tipos de estudo. Posteriormente será realizada uma seleção de dados referentes a quais drogas e seus respectivos casos (tempo e dose) serão utilizados nesse trabalho. Os dados selecionados são a base de tudo o que vem a seguir pois eles são a entrada de cada um dos dois métodos utilizados para determinação de ontologias do *Gene Ontology*. Será realizada para cada caso de cada droga, três tipos de normalização e esses dados normalizados serão utilizados para determinação de ontologias alteradas, o que significa que cada método irá fornecer três resultados para cada droga e caso. Posteriormente será realizada uma meta-análise funcional das ontologias alteradas considerando o resultado das 3 normalizações. Os resultados da meta-análise funcional de cada metodologia serão utilizados para realizar uma meta-análise funcional final.

3.2 *Gene Ontology* (GO)

A ferramenta *Gene Ontology* (GO) (ASHBURNER et al., 2000), está disponível no sítio (<http://www.geneontology.org/>) e tem como objetivo classificar e organizar as descrições dos genes e seus respectivos produtos. O GO iniciou em 1998 com somente três organismos: *Saccharomyces cerevisiae*, *Mus musculus* e *Drosophila melanogaster*, mas atualmente há mais espécies catalogadas, incluindo *Homo sapiens* *Rattus norvegicus* que são de interesse para esse trabalho.

As informações foram organizadas em três classificações independentes: processos biológicos, componentes celulares e funções moleculares. Cada uma tem uma abordagem diferente em relação a associação do gene e seu produto.

- **Função Molecular (MF):** Representa as atividades moleculares ao qual o gene está associado. Descreve, principalmente, ligações moleculares (ligação com enzimas, com oxigênio, ferro, ácidos graxos, etc) e atividade de enzimas específicas (aromatase, quinina 3-monoxigenase, testosterona 6-beta-hidroxilase, etc).
- **Compartimento Celular (CC):** Representa as localizações celulares em que o produto do gene desempenha sua função. Descreve, principalmente, organelas (mitocôndria, nucleoplasma, citoplasma), membranas (membrana celular, nuclear, do retículo endoplasmático) e complexos moleculares (complexo MCM, ciclina B1-CDK1, Ndc80, etc).
- **Processo Biológico (BP):** Representa os objetivos a serem atingidos pelas funções moleculares ao qual o gene está associado (replicação do DNA, processos metabólicos e catabólicos de substâncias, atividade de ontologias enzimáticas como epoxigenases, transporte de moléculas, etc).

As ontologias estão organizadas de maneira hierárquica, assim existem ontologias mãe e filhas sendo que uma ontologia pode ser mãe de mais de uma ontologia filha e estas podem ter mais de uma mãe. Quanta maior for a hierarquia da ontologia mais geral e abrangente ela é, caso contrário a ontologia tende a ser cada vez mais específica. Graficamente as ontologias são representadas em forma de árvores, esses gráficos também são chamados de grafos direcionados acíclicos. Um exemplo dessa representação gráfica de ontologias pode ser observado na figura 3.2.

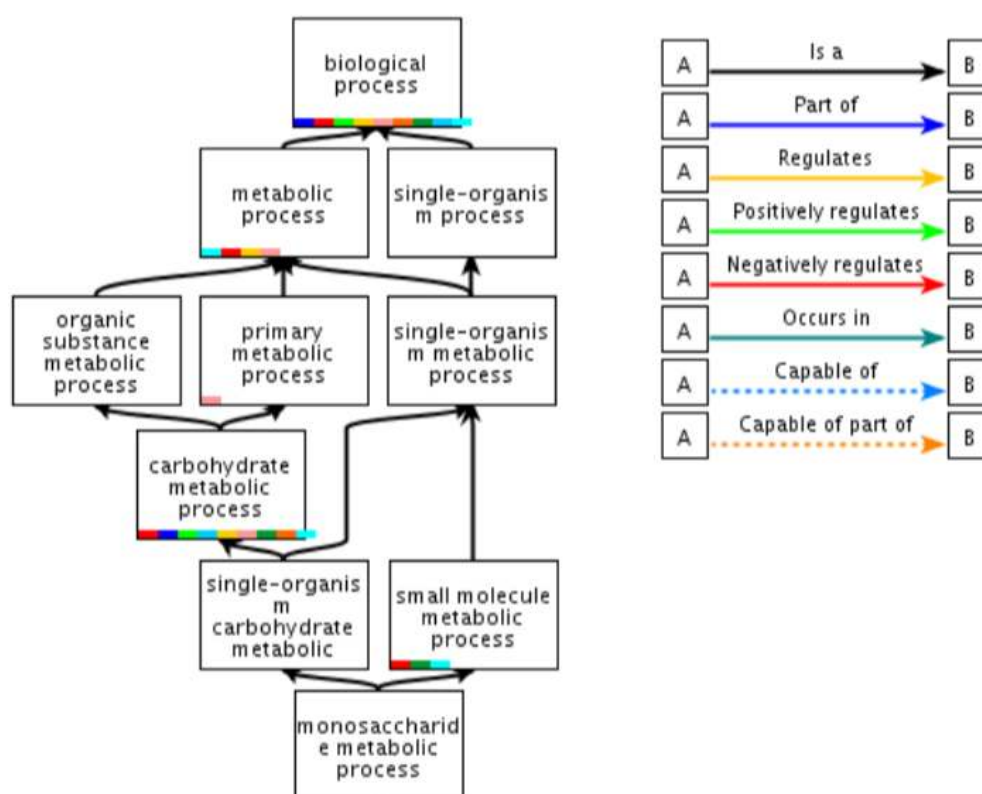


Figura 3.2: Exemplo de um segmento da ontologia BP, as ontologias na parte inferior são as mais específicas enquanto as da parte superior são mais gerais. Neste exemplo a ontologia *biological process* é a de maior hierarquia, sendo a mais geral de todas e não possuindo uma ontologia mãe, mas possui ontologias filhas: *metabolic process* e *single-organism process*. A ontologia *metabolic process* possui mais de uma ontologia filha enquanto a ontologia *single organism process* possui apenas uma. Como já mencionado as ontologias podem ser termais de uma ontologia filha e também podem ter mais de uma ontologia mãe como podemos observar na ontologia *single organism metabolic process*, que é filha das ontologias *metabolic process* e *single organism process*. Retirado de <https://www.ebi.ac.uk/QuickGO/> em maio de 2017.

Neste trabalho as ontologias do *Homo sapiens* e do *Rattus norvegicus* serão utilizadas por meio do pacote “GO.db” para o R, disponível na plataforma bioconductor.

3.3 STRING: Um Banco de Dados de Interações Proteína-Proteína

Redes são a representação de elementos e suas respectivas interações entre si, a natureza dos elementos (também chamados de nodos ou vértices) e das interações (também chamadas de arestas) são variáveis de rede para rede. Essas entidades são tipicamente representadas por matrizes de adjacência, que são matrizes binárias onde as linhas e colunas representam os elementos e os valores 1 e 0 são utilizados para representar a presença ou ausência, respectivamente, de interação entre os elementos. Redes biológicas tem sido cada vez mais utilizadas em análises e visualização de dados massivos produzidos pelos experimentos atuais. Um exemplo de rede biológica são as que associam genes a seus produtos e elementos reguladores. Outro exemplo são redes de interação proteína-proteína, onde seus nodos são as proteínas e as arestas representam interações entre essas proteínas.

O STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) é um banco de dados *online* de interações proteína-proteína. Esse banco de dados é mantido Laboratório Europeu de Biologia Molecular (*European Laboratory for Molecular Biology* - EMBL) desde 2000 (MERING et al., 2005; MERING et al., 2007; JENSEN et al., 2009; SZKLARCZYK et al., 2011; SZKLARCZYK et al., 2014) e pode ser encontrado no sítio “<https://string-db.org/>”. Hoje o STRING encontra-se na versão 10.5 com dados referentes a 2031 espécies, aproximadamente 9,6 milhões de polipeptídeos registrados e mais de 1,3 bilhões de interações. A natureza das interações registradas varia, atualmente existem 7 tipos de interações com a qual esse banco de dados trabalha: *Conserved Neighborhood*, *Co-occurrence*, *Fusion*, *Co-expression*, *Experiments*, *Databases* e *Textmining*.

As interações do tipo *Conserved Neighborhood* associam proteínas cujos genes codificantes (no genoma procarioto) estejam próximos um do outro (distância intergênica de no máximo 300 pares de bases). As interações do tipo *Co-occurrence* servem para indicar proteínas cujos genes codificantes possuam homólogos em outras espécies. As interações do tipo *fusion* indicam proteínas cujos genes codificantes sofram o processo de fusão. As interações do tipo *Co-expression* indicam proteínas cujos genes codificantes apresentam co-expressão. As interações do tipo *Experiments* são interações proteína-proteína retiradas de outros bancos de dados de interações entre proteínas (DIP,BIND,BioGRID). As interações do tipo *Databases* representam interações de proteínas considerando rotas metabólicas,

são retiradas de bancos de dados como KEGG e Reactome. As interações do tipo *textmining* representam interações proteína-proteína encontradas nos resumos de artigos científicos (MERING et al., 2005; MERING et al., 2007; JENSEN et al., 2009; SZKLARCZYK et al., 2011; SZKLARCZYK et al., 2014).

Nesse trabalho as redes de interação proteína-proteína para *Homo sapiens* e *Rattus norvegicus* foram obtidas da base dados STRING, considerando apenas os tipos de interação *Experiments* e *Databases*. Essas interações são retiradas de bancos de dados curados manualmente e portanto possuem um certo grau de confiança. O score mínimo para que as interações sejam aceitas é de 0.8, o uso destes parâmetros restringe ao máximo o número de interações proteína-proteína que sejam falsos positivos. Essas redes serão utilizadas para visualização e interpretação dos resultados no método EFDEGs, onde se trabalha com genes diferencialmente expressos.

3.4 Normalização dos dados

A normalização dos dados de expressão gênica em experimentos de microarranjo é um processo importante para a determinação dos perfis de expressão. Normalmente o que se deseja saber nesse tipo de experimento é o quanto a expressão de genes varia em resposta a diferentes estímulos a que o organismo está sujeito. Entretanto os dados observados no final de um experimento de microarranjo não contém somente a variação intrínseca do material genético, ou seja, a variação que geralmente é a de interesse dos pesquisadores. É comum que outras fontes de variação sejam introduzidas nas diversas fases do experimento de microarranjo (marcação do cDNA, hibridização e escaneamento) (DUDOIT et al., 2002).

Essas fontes de variação (“artefatos”) podem levar a resultados incorretos, por essa razão a remoção do viés estatístico introduzido por esses artefatos através da normalização dos dados é primordial. Normalmente por normalização se quer dizer um procedimento de remoção de variações entre os dados de entrada, para que a maior parte dos valores de uma distribuição fique em torno de uma média, facilitando a visualização da variação intrínseca. Entretanto quando se fala sobre normalização de dados de expressão gênica de microarranjo, existem outras etapas além da que foi descrita acima, o objetivo final da normalização de dados de expressão gênica de microarranjo é a obtenção de valores de expressão que sejam o mais próximo possível da realidade.

Para isso é necessário: considerar a influência do *background* (ruído) nas intensidades das sondas do *chip*, realizar a normalização dos dados corrigidos em relação ao ruído/*background* e utilizar algum modelo estatístico para determinação do valor de expressão final. Nos *chips* da *affymetrix*, cada gene é representado por 16-20 pares de sondas sendo que uma delas é a sonda PM (*Perfect Match*) e a outra a sonda MM (*Mismatch*). Os oligonucleotídeos da sonda MM são idênticos aos da PM exceto por um nucleotídeo da décima terceira posição, a intenção disso é utilizar o sinal emitido pelas sondas MM como ruído devido ao *background* e a hibridização não específica. Uma parte importante das diferenças existentes entre os variados métodos de normalização de microarranjo é justamente o uso das sondas PM e MM para determinação da intensidade de cada sonda.

Neste trabalho as normalizações foram realizadas com o auxílio de três pacotes da plataforma Bioconductor (GENTLEMAN et al., 2004): “*Limma*” (SMYTH, 2005), “*Affy*” (GAUTIER et al., 2004) e “*Affypdnn*” (NIELSEN; GAUTIER; GAUTIER,). Os três métodos de normalização que serão utilizados são: *Robust Multi-array Average* (RMA) (IRIZARRY et al., 2005), *GC Robust Multi-array Average* (GCRMA) (WU et al., 2004), *Microarray Suite 5* (MAS5) (AFFYMETRIX, 2002). A tabela 3.1 abaixo mostra algumas características dos métodos de normalização a serem utilizados

Tabela 3.1: Métodos de normalização de dados de microarranjos e suas características

Método	Correção de Background	Normalização	Expressão medida
RMA	Correção global	<i>Quantile</i>	Modelo linear robusto (Median Polish)
GCRMA	Correção baseada na sequência de cada sonda	<i>Quantile</i>	Modelo linear robusto (Median Polish)
MAS5	Subtração das intensidades sondas MM	<i>Scale factor</i>	Estatística robusta (Tukey-biweight)

3.4.1 *Robust Multi-array Average* (RMA)

Esse método de normalização foi desenvolvido em (IRIZARRY et al., 2005), e utiliza um modelo linear robusto (*Tukey median polish*) para determinar os valores de expressão das sondas baseados em seus valores observados corrigidos por um método de correção de *background* que considera todas as sondas dos *chips* de uma amostra que posteriormente são normalizados com o método *Quantile*. A correção global do *background* usa um modelo que utiliza todas as sondas de todos os *chips* (replicatas) de uma amostra e considera que o sinal medido é a soma de um sinal verdadeiro, cuja distribuição de intensidades decai de maneira exponencial conforme aumenta, e de um sinal ruído cujas intensidades tem uma distribuição normal. A normalização dos dados é feita com o método *Quantile*, o qual foi demonstrado por Bolstad *et al* como sendo um método excelente para a linha *Genechip*® da *affymetrix* (BOLSTAD et al., 2003). O modelo linear utilizado aqui determina apenas os valores de expressão das sondas PM (*Perfect Match*) em escala logarítmica considerando três componentes: a intensidade devido a sonda (o que geralmente é o interesse), a intensidade devido ao próprio *array* e a intensidade devido ao ruído. Neste modelo é assumido que as

sondas MM não trazem informação confiável quanto ao ruído. Em (IRIZARRY et al., 2005) é demonstrado que quanto maior o valor de PM maior é o valor de MM, uma evidência de que talvez as sondas MM não emitam apenas ruído como sinal.

3.4.2 *GC Robust Multi-array* (GCRMA)

Esta é uma adaptação do RMA, o método de normalização dos dados e de determinação da expressão final são iguais aos do RMA. Esse método se diferencia pela abordagem de cálculo do efeito de afinidade das sondas para os valores de expressão final. Este se utiliza de um modelo matemático que considera a sequência de bases nitrogenadas dos oligonucleotídeos presentes nas sondas. Com base nesse efeito de afinidade determina-se um valor corrigido para as intensidades das sondas MM. Esse valor é então subtraído das intensidades das sondas PM para determinação da expressão final, que segue o mesmo modelo linear robusto do RMA. O GCRMA, portanto, assim como o MAS5 considera que as sondas MM de alguma forma podem medir um sinal ruído em relação a sua sonda PM, o que os diferencia é a abordagem matemática (e as hipóteses por trás delas) que os modelos utilizam (JIANG et al., 2008) (WU et al., 2004).

3.4.3 *Microarray Suite 5* (MAS5)

Esse método se diferencia do RMA em duas partes: a correção do *background* utilizada e o modelo de determinação do sinal final. A correção do *background* é feita de maneira espacial, dividindo o *chip* em partes retangulares e calculando o ruído de cada parte a partir das sondas MM, ou seja, ao contrário do RMA esse modelo mantém a hipótese de que as sondas MM representam o erro medido nas sondas PM (IRIZARRY et al., 2005). O sinal final é calculado, em escala logarítmica, a partir do modelo *tukey-biweight*, utilizando valores das sondas PM subtraídos do ruído calculado a partir das sondas MM. Caso a sonda MM tenha um valor maior que a PM, um outro método de correção é utilizado pois isso implicaria em um sinal negativo. A normalização utilizada é a *scale factor*, mais informações sobre este procedimento podem ser encontradas no manual do *affymetrix* (AFFYMETRIX, 2002).

3.5 Enriquecimento Funcional de Genes Diferencialmente Expressos (EFDEGs)

O EFDEGs permite a determinação de ontologias (BP, MF e CC) significativamente alteradas, utilizando como entrada os genes diferencialmente expressos (GDEs) entre um caso e um controle. Serão considerados como diferencialmente expressos os genes que apresentarem q-valor menor que 0.05 e logFC (log-Fold-Change) maior que 1 ou menor que -1 (o que implica em um valor de expressão até 10 vezes maior ou menor do caso em relação ao controle). Os valores dos parâmetros mencionados acima são determinados no processo de normalização dos dados. Os GDEs são então utilizados como entrada no pacote “topGO”, responsável pelo enriquecimento funcional dos DEGs. Esse pacote tem a vantagem de permitir com que seu usuário possa customizar o teste estatístico utilizado para determinação da significância estatística e as correções adicionais (ALEXA; RAHNENFUHRER, 2010).

Nesse trabalho foi utilizado como teste estatístico o teste exato de fisher. Também foi aplicado a correção “*Weight01*” que é uma junção de duas outras correções gráficas presentes no pacote: a “*elim*” e o “*Weight*”. O *Weight* considera a hierarquia das ontologias, essa correção considera apenas as interações locais (termo pai e termos filhos). O *elim* faz uma varredura completa da hierarquia GO de baixo para cima, isso é, dos termos mais específicos aos mais gerais e descarta genes que já tenham sido anotados em ontologias descendentes. Em função dessas correções, tanto o *weight* quanto o *elim* tendem a ser mais conservativos (menos falsos positivos) que a abordagem “classic” do pacote topGO e que não considera a hierarquia dos termos (ALEXA; RAHNENFUHRER, 2010). Devido ao processo de eliminação na correção *elim*, ela tende a ser mais conservativa que a correção *Weight*, mas tende a não detectar positivos verdadeiros em partes elevadas da hierarquia GO (erros do tipo II; falso negativo) (ALEXA; RAHNENFUHRER, 2010). Por sua vez, o *Weight* tende a ser menos conservativo que o *elim* detectando mais positivos verdadeiros, porém cometendo mais erros do tipo I (falso positivo) que o *elim* (ALEXA; RAHNENFUHRER, 2010).

Por essa razão a correção *Weight01* foi escolhida pois esta mantém um equilíbrio entre erros do tipo I e erros do tipo II. Para este método serão consideradas como significativamente alteradas as ontologias que apresentarem um q-valor menor que 0.001 ($FDR < 0.05$).

3.6 Análise de Grupos de Genes Funcionalmente Associados (AGFAGs)

A análise de grupos de genes funcionalmente associados (AGFAGs) foi realizado por meio do pacote *EntropyGenesCluster*, ainda em desenvolvimento, para a linguagem de programação R. A metodologia do pacote pode ser encontrada em (CASTRO et al., 2009), onde é chamada de ViaComplex. A metodologia fornece uma análise estatística da alteração das rotas biológicas de interesse com aplicação da estatística *bootstrap* e do teste FDR (*False Discovery Rate*) para seleção de “Grupos de Genes Funcionalmente Associados” (GFAGs) significativamente alterados.

Primeiramente o usuário deve fornecer dados de expressão gênica, um caso e um controle, o método vai então agrupar os genes presentes nos dados fornecidos em de acordo com suas funções. Esses grupos são chamados de grupos de genes funcionalmente associados (GFAGs) e para determina-los o pacote *EntropyClusterGenes* utiliza as ontologias do GO, que na linguagem R podem ser obtidas com o pacote “GO.db”. Na verdade esses GFAGs são cada uma das ontologias existentes (BP, CC e MF) no *Gene Ontology* e portanto um gene pode estar presente em mais de um GFAG.

Depois de agrupar os genes o pacote define alguns parâmetros para cada GFAG, a diversidade do caso (H_e) e do controle H_c e a atividade do caso (N_e) e do controle (N_c). A atividade é simplesmente a soma dos valores de expressão (do caso ou do controle) de cada gene no grupo e é utilizada para determinar um outro parâmetro chamado atividade relativa (n) que é definido como a razão entre a atividade do caso e a soma da atividade do controle e atividade do caso ($n = N_e / (N_e + N_c)$). Se a atividade relativa for maior que 0.5, isso implica que a expressão do caso é maior que a do experimento, e caso contrário, uma expressão menor. Outro parâmetro definido é a diversidade relativa (h), que é a razão entre a diversidade do caso e a soma da diversidade do controle e do caso ($h = H_e / (H_e + H_c)$). Este último parâmetro tem um comportamento parecido com o da atividade relativa, a expressão do caso é maior se seu valor for superior a 0,5 e inferior caso contrário.

A partir dos valores da atividade relativa e da diversidade relativa o pacote *EntropyGenesCluster* determinará os p-valores dos GFAGs, utilizando um método de reamostragem com substituição (*bootstrap*). O método *bootstrap* utiliza os dados de entrada que contém todos os genes e seus valores de expressão para caso e controle e então calcula valores de atividade relativa e de diversidade relativa de grupos de genes que são definidos aleatoriamente. Esses grupos possuem exatamente o mesmo número de genes que o GFAG cujo p-valor é calculado, e se a atividade relativa ou a diversidade relativa sejam menores a “operação” é considerada um sucesso. O número de operações realizado é definido pelo usuário, naturalmente quanto mais operações são realizadas mais confiáveis serão os p-valores obtidos entretanto isto também aumenta o tempo de execução do pacote. Por fim o p-valor é definido como a razão entre o número de sucessos e o número total de operações, este p-valor é ainda corrigido por um FDR (*False Discovery Rate*) de 0.05, determinando para o GFAG em questão um p-valor ajustado, chamado de q-valor.

4 *Resultados Discussão*

4.1 Sobre as análises e normalizações

As análises dos estarão divididas em duas etapas, a primeira é uma análise numérico-descritiva dos resultados (ontologias e/ou GDEs) encontradas de maneira geral e para cada droga considerando as 3 normalizações (“rma”.“gcrma” e “mas5”). A segunda é uma análise biológica, uma avaliação das ontologias encontradas como significantes. Essa segunda análise é realizada apenas para as 3 drogas considerando uma das normalizações, a normalização “rma”, pois foi a normalização cujos resultados apresentaram a distribuição mais homogênea entre as utilizadas.

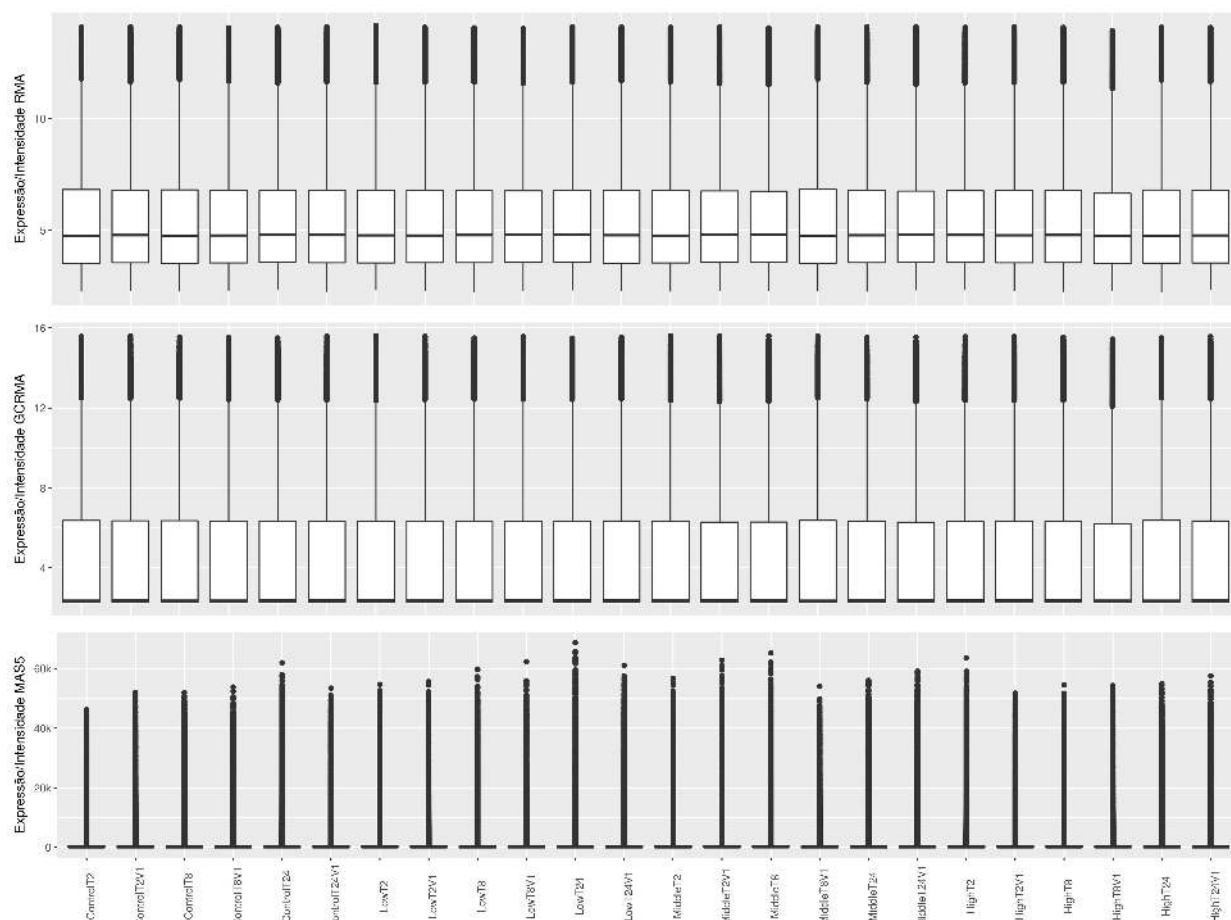


Figura 4.1: Boxplots das 3 normalizações para os chips da ciclofosfamida e estudo *Homo sapiens in vitro*: O eixo vs é comum aos 3 gráficos e contém os casos dose/tempo e o eixo y é diferente para cada gráfico pois reflete a expressão segundo cada tipo de normalização.

Como pode se observar na figura 4.1, a normalização “mas5” apresenta dados centralizados em torno da mediana (assim como os outros dois métodos), mas tem um grande “achatamento” dos dados, o primeiro e o segundo quartil estão especialmente achatados e o número de *outliers* é muito maior do que nas outras normalizações. As normalizações “rma” e “gcrma” possuem boxplots bem mais parecidos entre si, entretanto ainda observa-se que no método “gcrma” existe um achatamento do primeiro quartil dos dados. O método “RMA” claramente possui as distribuições mais homogêneas, e esse comportamento entre as 3 normalizações se repete para todas as drogas em todos os estudos. Por essa razão a normalização “rma” foi selecionada para realizar as análises biológicas.

4.2 Método EFDEGs

4.2.1 Aspectos Gerais: Análise Numérica e Descritiva

Uma análise descritiva (numérica) dos resultados gerais do método em questão pode fornecer *insights* valiosos sobre comportamentos, padrões e tendências segundo diferentes classificadores dos mesmos (“estudos” ou “espécies”) dentro das diversas condições (ex: tempo, dose, normalização) sem mesmo entrar-se no mérito da análise funcional/biológica de fato.

Abaixo na Figura 4.2, observa-se a distribuição de GDEs encontrados para diferentes condições em 3 diferentes valores de *cutoff* para o p-valor. Algumas tendências que podem ser notadas em tal figura são:

- As doses altas (“High”, em vermelho) em geral possuem um maior número de GDEs.
- Para todas as doses, é visível que aparentemente existem “ciclos de ativação” do que se diz respeito ao tempo.
- Esses ciclos de ativação parecem ter um comportamento bem definido nos estudos *in vitro*, sendo que quanto maior o tempo, maior é o número de GDEs encontrados. Um comportamento linear de certa forma, e que se repete para os 3 valores de *cutoff*.
- Nos estudos *in vivo* os ciclos de ativação aparentam não ter um comportamento tão uniforme entre os 3 valores de *cutoff*, embora esteja aparente que o tempo inicial (3 h) seja o que apresenta o maior número de resultados, um comportamento bem diferente dos estudos *in vitro*.

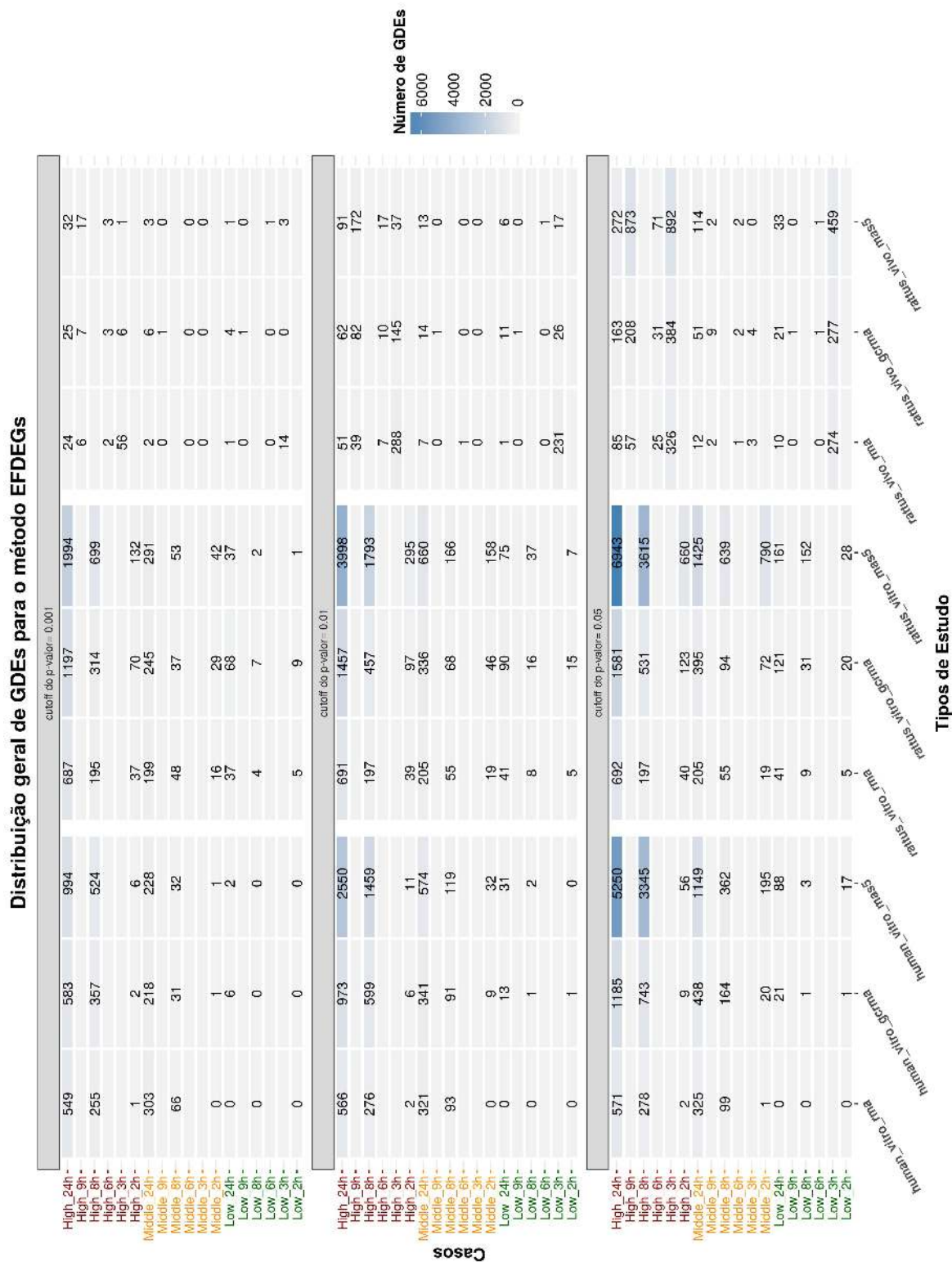


Figura 4.2: GDEs encontrados para as 3 drogas antineoplásicas. Distribuição geral de GDEs encontrados no método EFDEGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que as células vazias indicam a inexistência do experimento em tais condições. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

- Os estudos *in vitro* em geral trazem resultados mais numerosos que os estudos *in vivo*.
- O método de normalização com maior número de resultados foi o “mas5” seguido do “gcrma” e por último, com resultados em menor número, temos o “rma”.

Um gráfico de linhas para cada “coluna” da Figura 4.2, logo abaixo, facilita ainda mais a visualização de alguns comportamentos/tendências mencionados acima.

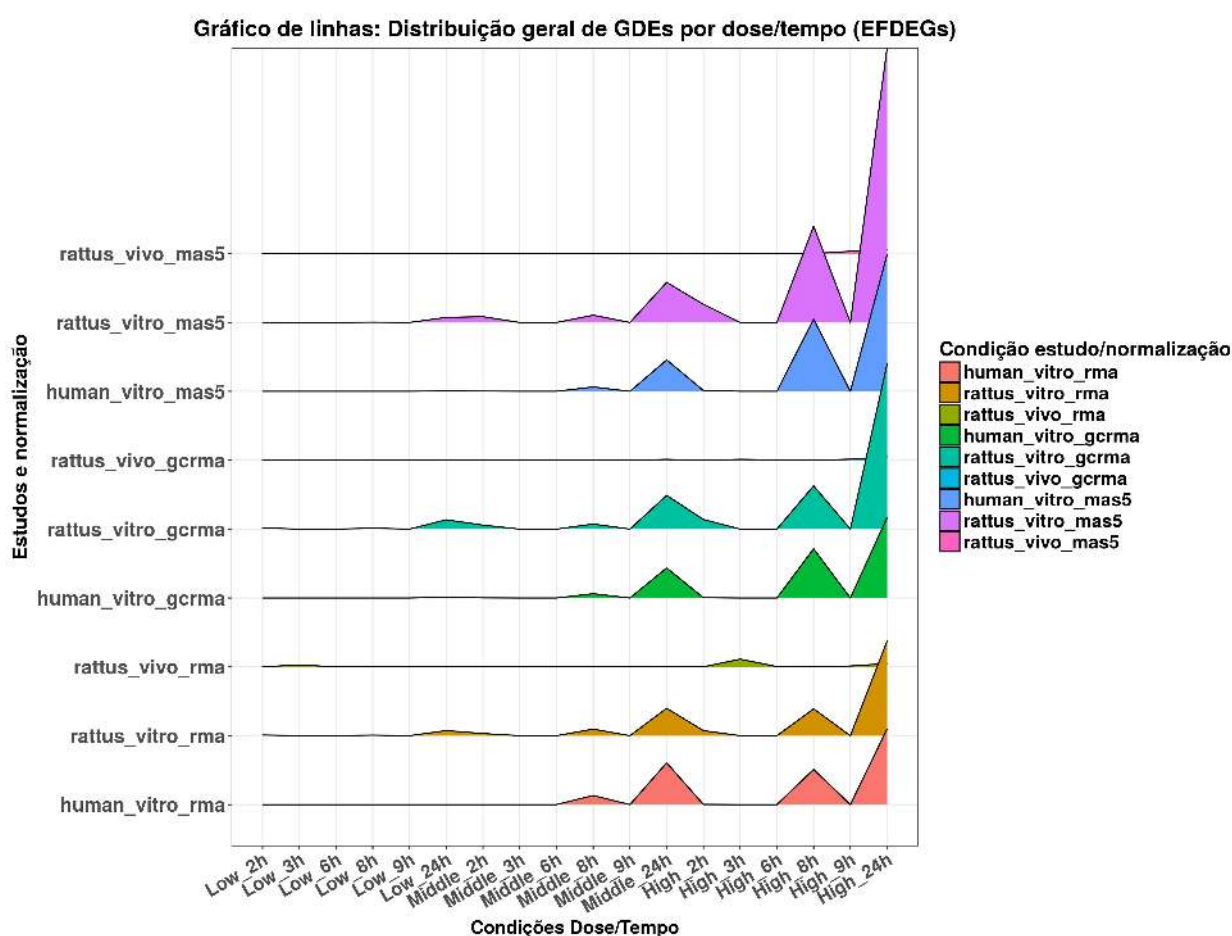


Figura 4.3: Gráfico de linhas para os GDEs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.2, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de GDEs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.2 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Pode-se observar facilmente, na Figura 4.3, que a normalização “mas5” retorna mais resultados entre as 3 utilizadas, seguida do “gcrma” e por fim do “rma”. Isso talvez se deva ao fato de que esta normalização foi desenvolvida pela própria affymetrix no início da ano de 2000, quando *chips* de microarranjo possuíam custos extremamente elevados em relação aos preços atuais e portanto, quando poucos *chips* estavam a disposição dos pesquisadores, assim sendo o “mas5” foi desenvolvido para experimentos com uma única amostra (*single sample*) (AFFYMETRIX, 2002). Os estudos *in vitro* aqui apresentados foram realizados em duplicata e os *in vivo* em triplicata, explicando o porque dos resultados elevados do “mas5”. O “rma” e o “gcrma” foram ambos desenvolvidos para experimentos com mais de 1 amostra (replicatas) (IRIZARRY et al., 2005).

Observa-se também a existência dos “ciclos de ativação” mencionados anteriormente. Nos estudos *in vitro* esses ciclos são bem definidos, a quantidade de GDEs aumenta diretamente em relação ao aumento da dose e do tempo. Inclusive a combinação “*High-24h*” é sempre aquela que retorna mais resultados nos estudos *in vitro*. Os estudos *in vitro* possuem amostragem nos tempos 2 h, 8 h e 24 h portanto não haverão resultados nos tempos 3 h, 6 h e 9 h como pode-se observar. Também se observa que as doses baixas (“*Low*”) praticamente não trouxeram GDEs, com isso já pode-se esperar que as 3 drogas não encontraram muitos GDEs para estudos com dose baixa e que o mesmo comportamento provavelmente será encontrado para os resultados das ontologias.

O comportamento dos estudos *in vivo* é bem diferente dos estudos *in vitro*, sendo que o número de resultados desses estudos foram diminutos em relação aos *in vitro*. Uma tendência interessante que pode-se observar nos estudos *in vivo* é como os tempos iniciais das doses (3 h) das doses “*High*” e “*Low*” aparentam serem mais “ativos” do que os tempos 6 h, 9 h e 24 h (sendo que o tempo de 24 h é o segundo mais ativo). Esse comportamento não se apresenta para a dose *Middle*. Essa diferença radical de comportamento dos resultados pode refletir uma diferença intrínseca entre experimentos *in vivo* e *in vitro* para uma ou mais das 3 drogas sendo analisadas.

Os resultados obtidos com o método EFDEGs são constituídos de GDEs e posteriormente das ontologias associadas aos GDEs. Ou seja, os GDEs são a entrada necessária para obtenção das ontologias, assim sendo pode-se presumir que quanto mais GDEs forem encontrados mais ontologias significantes serão encontradas. Com isso vem a dedução de que talvez uma análise numérico-descritivo das ontologias encontradas tenham um comportamento parecido com a dos GDEs. Para avaliar essa hipótese figuras parecidas com a Figura 4.2 e a Figura 4.3 foram feitas para a ontologia BP (*Biological Process*). A Figura 4.4 abaixo, ilustra o número de BPs significantes segundo 3 diferentes valores de *cutoff* do p-valor.

O número de ontologias significantes encontradas está diretamente relacionado com o número de GDEs encontrados, uma vez que os GDEs são o ponto de partida na determinação de ontologias alteradas. É visível que os comportamentos descritos para os GDEs na Figura 4.2 também são observados na Figura 4.4. A Figura 4.5 facilita a visualização desses comportamentos.

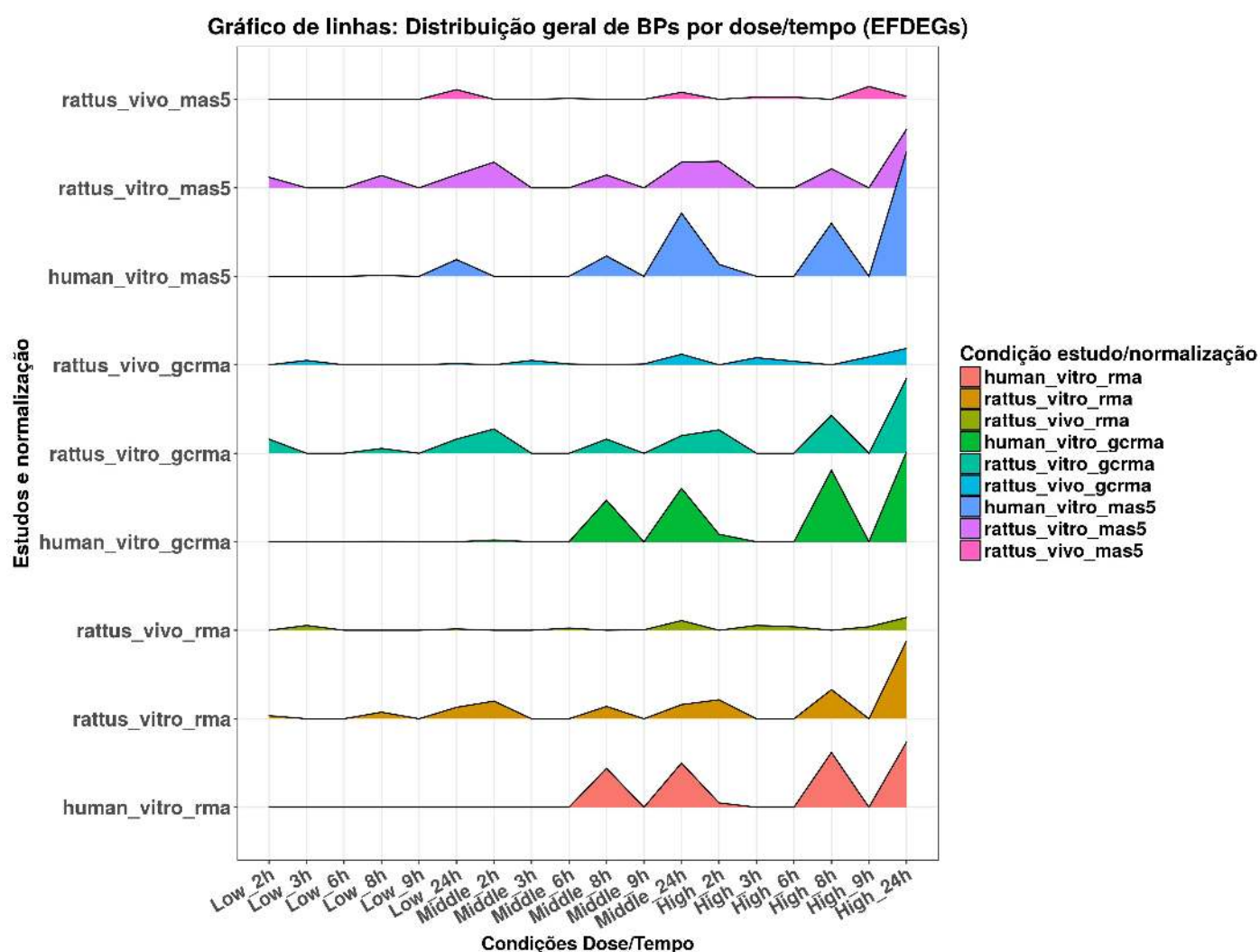


Figura 4.5: Gráfico de linhas para os BPs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.4, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.4 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Algumas das tendências observadas na Figura 4.5 acima são:

- O método de normalização “mas5” é o que trás o maior número de resultados, seguido do “gcrma” e do “rma”.
- Em geral quanto maior a dose maior é o número de BPs encontrados.
- Os estudos *in vitro* em geral fornecem mais resultados do que os *in vivo*.
- Os “ciclos de ativação” também estão presentes e se comportam de maneira similar. Nos estudos *in vitro* a quantidade de BPs aumenta junto com a dose e com o tempo, sendo que o caso *High-24h* é sempre o que retorna mais resultados nos estudos *in vitro*. Nos estudos *in vivo* existe uma certa irregularidade entre tipos de normalização, embora fique aparente que os tempos de 24 h sejam os que retornem mais resultados. Essa irregularidade pode significar que talvez cada droga tenha um comportamento diferente em relação aos estudos *in vivo*.

Alguns *insights* importantes foram derivados das figuras, que traçaram um panorama geral dos resultados obtidos nas 3 drogas. Para complementar ainda mais a análise é necessário observar as tendências nos resultados de cada droga separadamente e posteriormente realizar uma análise de um ponto de vista biológico/funcional.

4.2.2 Ciclofosfamida

Para observar as tendências e comportamentos dos resultados da droga ciclofosfamida no método EFDEGS, será realizado a mesma abordagem análise dos resultados gerais. Abaixo a Figura 4.6 mostra a distribuição de GDEs encontrados em diversos casos.

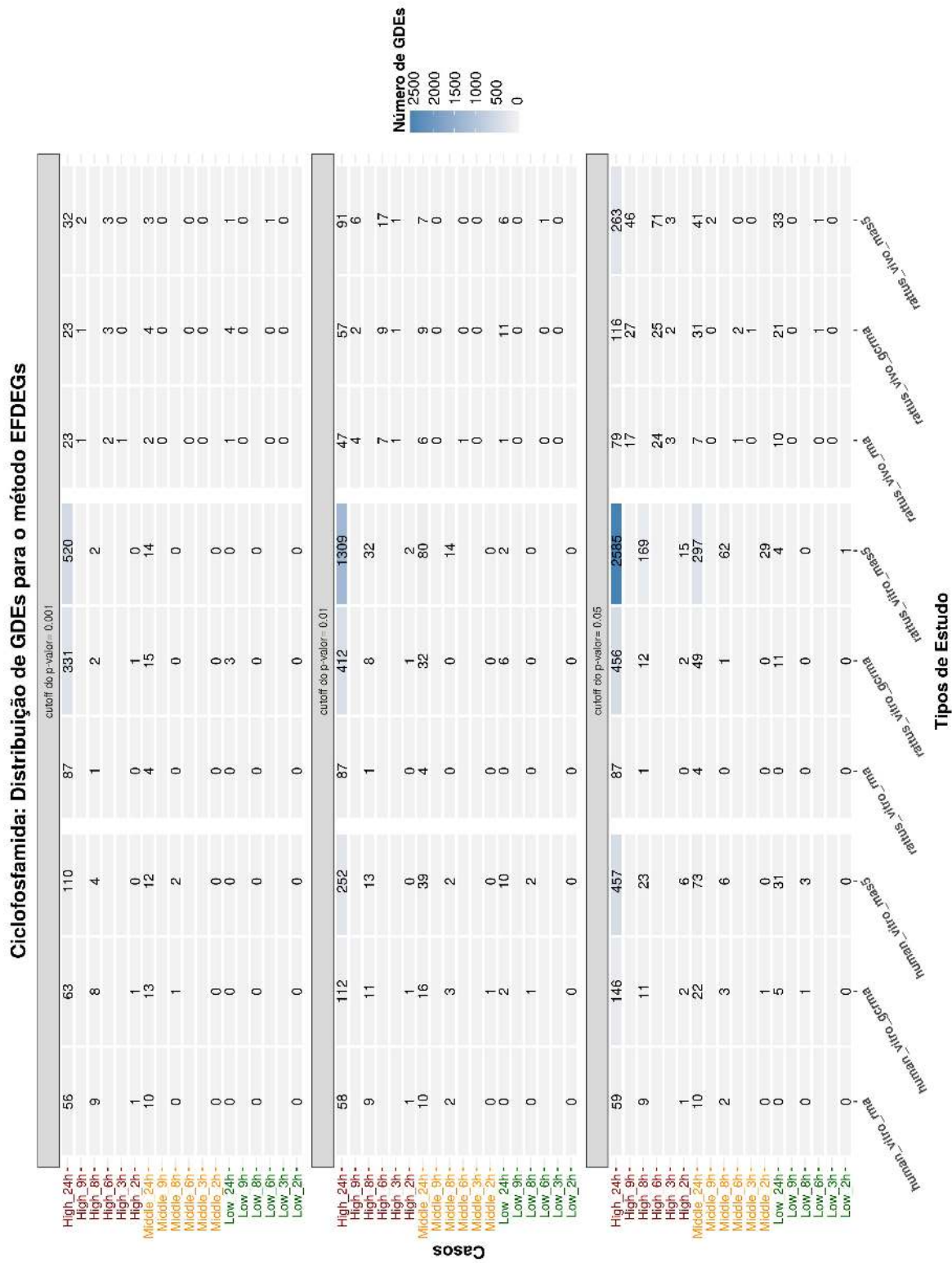


Figura 4.6: GDEs encontrados para a ciclofosfamida no método EFDEGs. Distribuição geral de GDEs encontrados no método EFDEGs para a ciclofosfamida considerando 3 diferentes valores de cutoff de p-valor (0,05, 0,01 e 0,001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que as células vazias indicam a inexistência do experimento em tais condições. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

Para facilitar a visualização das tendências/padrões nos resultados da Figura 4.6 foi traçado um gráfico de linhas na Figura 4.7.

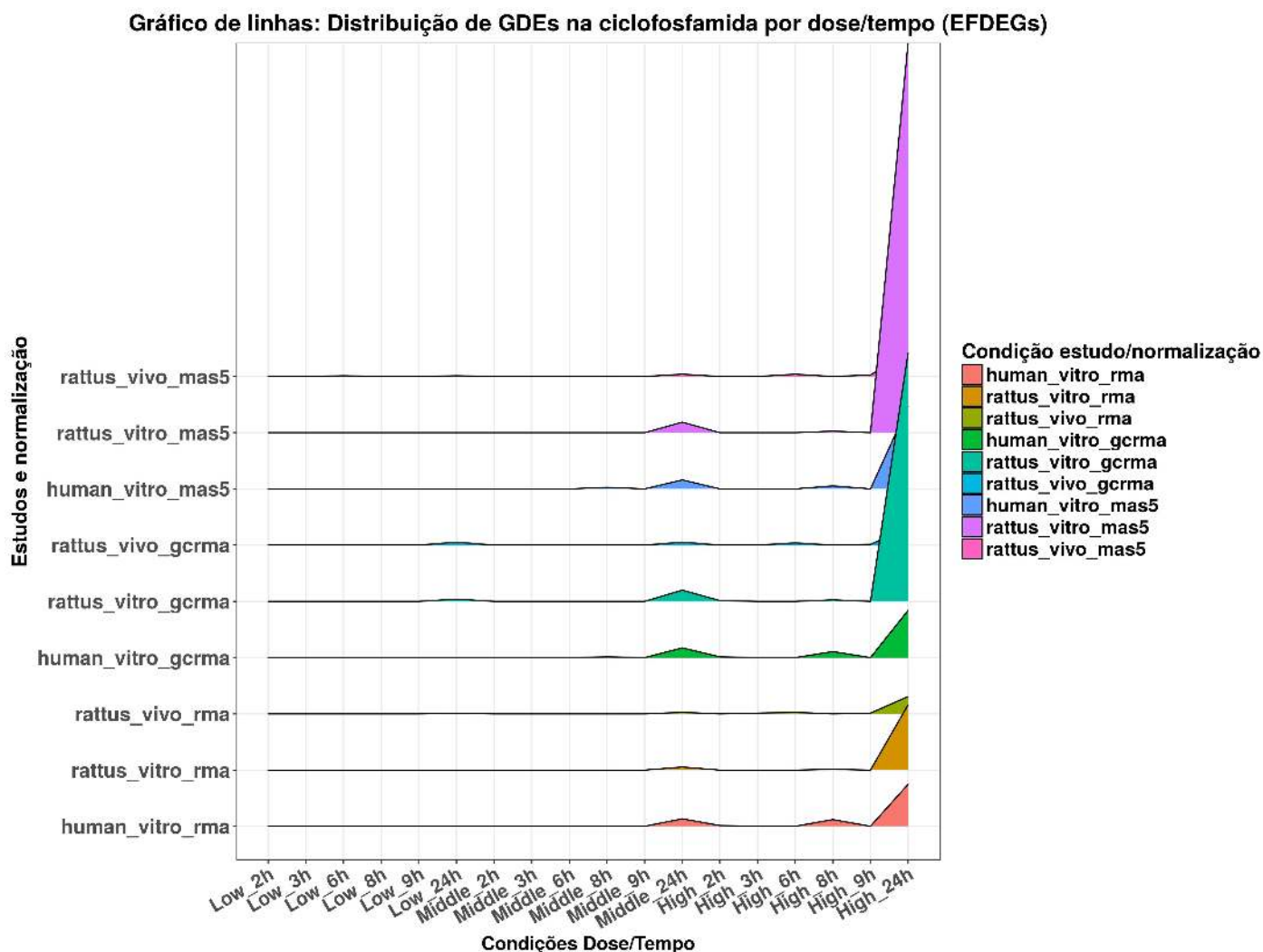


Figura 4.7: Gráfico de linhas para os GDEs da ciclofosfamida no método EFDEGs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.6, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de GDEs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.6 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Alguns dos comportamentos são similares e outros diferentes dos que aparecem na Figura 4.3 para os resultados gerais. Ao considerarmos uma droga específica o comportamento dos resultados é alterado, sumarizando os comportamentos observados abaixo:

- O método de normalização “mas5” continua sendo o que retorna mais resultados, seguido do “gcrma” e do “rma”. A mesma sequência dos resultados gerais, portanto esse comportamento permanece inalterado. No entanto é interessante notar que a “performance” das normalizações “mas5” e “gcrma” estão muito parecidas.
- Quanto maior a dose maior é o número de resultados entre os tempos dos mesmos.
- Os estudos *in vitro* continuam trazendo mais resultados que os estudos *in vivo*.
- Os “ciclos de ativação” dos estudos *in vitro* permacem com o mesmo comportamento apresentado nos resultados gerais (Figura 4.3), isto é, quanto maior a dose e o tempo maior é o número de resultados, sendo que o caso *High-24h* é o que sempre retorna o maior número de resultados.
- Nos estudos *in vivo* os resultados GDEs são bem escassos, no entanto nota-se a presença de picos nos tempos 24 h, principalmente na dose alta (*High*). Existem também picos diminutos em *High-6h*.

O próximo passo é observar as tendências e comportamentos das ontologias e checar se acompanham o comportamento dos GDEs.

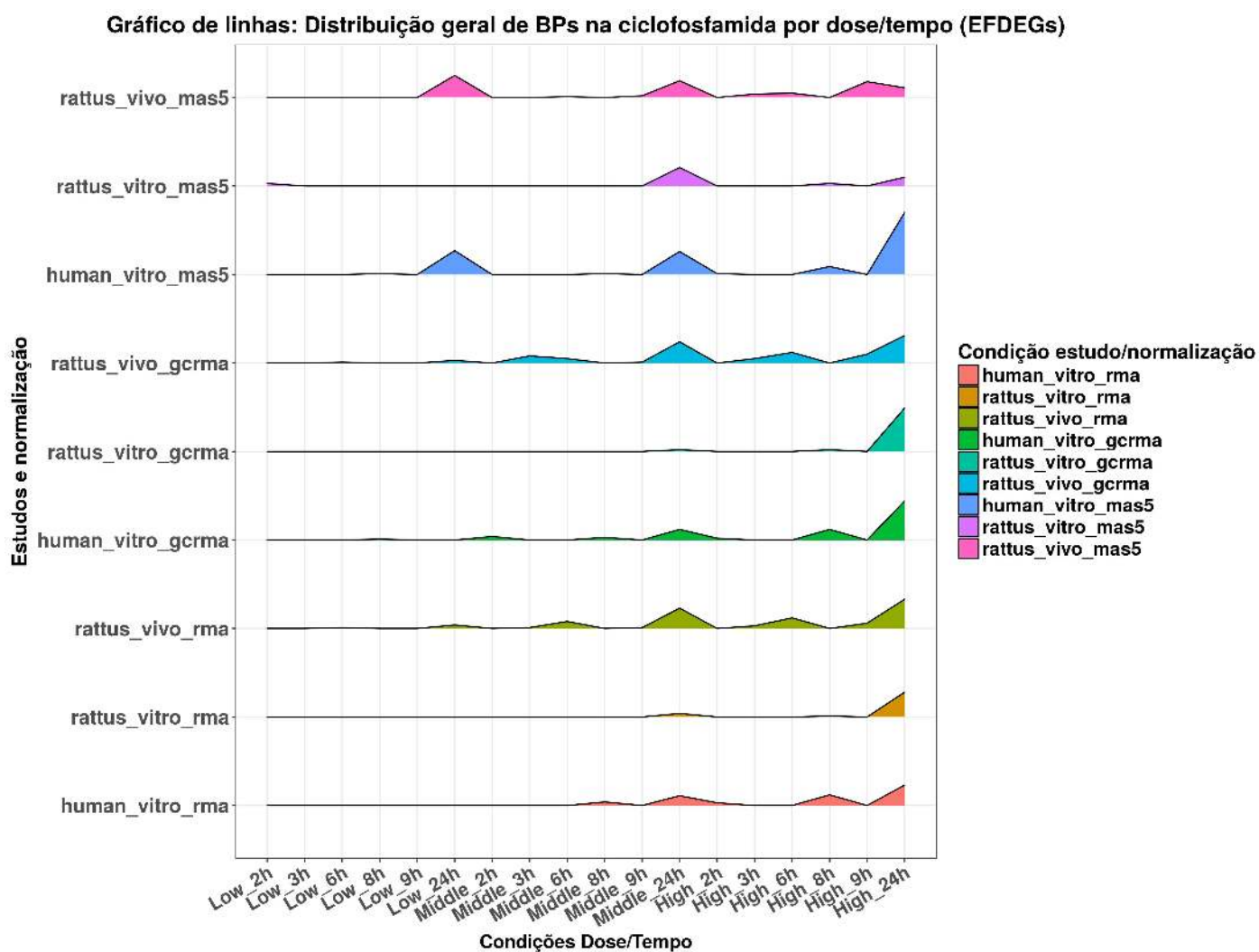


Figura 4.9: Gráfico de linhas para os BPs da ciclofosfamida no método EFDEGs com cutoff de 0.001.

Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.8, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.8 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Observando a Figura 4.8 e a Figura 4.9 pode-se concluir que as tendências para as ontologias BP são similares as tendências aparentes das figuras 4.6 e 4.7, que ilustram os GDEs encontrados. Isso é algo esperado, pois os GDEs são necessários para determinação de ontologias alteradas. Percebe-se que, como esperado, os tempos de 24 h tem os maiores picos. Nota-se também a existência de picos em *High-6h* e *High-9h*.

Somente com essa análise numérico-descritiva, foi possível determinar que a ciclofosfamida tem um comportamento similar para os 3 tipos de estudos, sendo que os resultados aumentam conforme aumenta-se a dose e o tempo. Também fica aparente como os resultados das ontologias estão alinhados com os dos GDEs. Para complementar os resultados, no entanto, faz-se necessário o uso de uma interpretação biológica/funcional dos resultados encontrados. Ou seja, se inicialmente consideramos “o quanto”, agora será considerado “o que” foi encontrado. Isso será realizado nos próximos passos considerando apenas a normalização “rma” e *cutoff* de p-valor de 0.001.

Agora uma breve descrição numérica dos resultados (GDEs/BPs) que foram encontrados nos 3 estudos da ciclofosfamida. O estudo *Homo sapiens in vitro* encontrou entre todas as doses e tempos 76 GDEs e 53 ontologias. O estudo do *Rattus norvegicus in vitro* encontrou 92 GDEs e 33 BPs. Por fim o estudo do *Rattus norvegicus in vivo* encontrou 30 GDEs e 92 BPs.

Para determinar quais são os GDEs e processos biológicos (BPs) mais promissores ou relevantes, foi observado o quão presente tal GDE/BP está nos estudos analisados. Para isso foi analisado a presença ou não de um GDE/BP dentro do universo de uma determinada dose (“High”, “Middle”, “Low”, desconsiderando o tempo). A soma destes universos pode ser entendida como um universo formado por todos resultados não nulos. Não foram encontrados GDEs que estivessem simultaneamente presentes nos 3 estudos e mesmo na comparação *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo* foi encontrado apenas um único GDE em comum, o gene “Car3”. Não foram encontrados BPs que estavam presentes simultaneamente nos 3 tipos de estudo, independente da dose e do tempo. Resta então a análise dos estudos separadamente e determinar quais são os BPs mais comuns em cada um deles.

É interessante destacar que dentro do universo de todos os 76 GDEs encontrados no estudo *Homo sapiens in vitro*, a maioria dos GDEs que estavam subexpressos tinha importância relevante para regulação do ciclo celular mitótico, indicando uma regulação negativa do ciclo celular mitótico, como CCNB1, CENPE e DEPDC1 e também de proteínas da

membrana celular (quimiocinas e seus ligantes em geral), como CXCL1, CXCL6, CCL2 e por fim um gene associado a regulação de canais de cálcio, o RRAD. Os superexpressos estão quase todos relacionados com metabolismo, especialmente da via citocromo P450, alguns estão relacionados com regulação do ciclo celular MCM10, UHRF1 e CDC6 e por fim um gene associado a resposta a danos ao DNA (WEIZMANN, 2018).

Para o estudo *Homo sapiens in vitro* foram encontradas 7 GDEs e 11 BPs que estavam presentes em todos casos não nulos, no caso da ciclofosfamida, isso pode ser entendido como o universo de GDEs/BPs encontrados nas doses “*High*” e “*Middle*” (independente do tempo), visto que a dose *Low* não encontrou nenhum resultado. Trata-se dos GDEs CYP3A5, CYP3A4, CYP3A7-CYP3A51P, CYP2C9, CYP2C8, TSKU, EGOT. As ontologias encontradas são: i) GO:0002933–hidroxilação de lipídios; ii) GO:0009822–processo catabólico de alcalóides; iii) GO:0070989–demetilação oxidativa; iv) GO:0006805–processo metabólico de xenobióticos; v) GO:0042738–processo catabólico de droga exógena; vi) GO:0019373–via epóxigenase P450; vii) GO:0016098–processo metabólico de monoterpénóides; viii) GO:0036378–processo biossintético de calcitriol a partir de calciol; ix) GO:0097267–via omega-hidroxilase P450; x) GO:0006706–processo catabólico de esteróides; xi) GO:0008209–processo metabólico de andrógenos.

Esses resultados tem respaldo na literatura científica, a ontologia GO:0009822, é uma consequência do mecanismo de ação da ciclofosfamida, pois esta é um agente alquilante, que age realizando alquilação do DNA na guanina (HUMANS et al., 2012). Como resultado da degradação/catabolismo da ciclofosfamida no organismo observamos esta ontologia. A ontologia GO:0002933 também tem respaldo na literatura, e de certa forma, é um complemento da outra ontologia, pois o processo de hidroxilação de lipídios é importante na metabolização de xenobióticos. As figuras 4.10 e 4.11 abaixo mostram o comportamento dos genes nas duas ontologias encontradas para o caso *High-24h*.

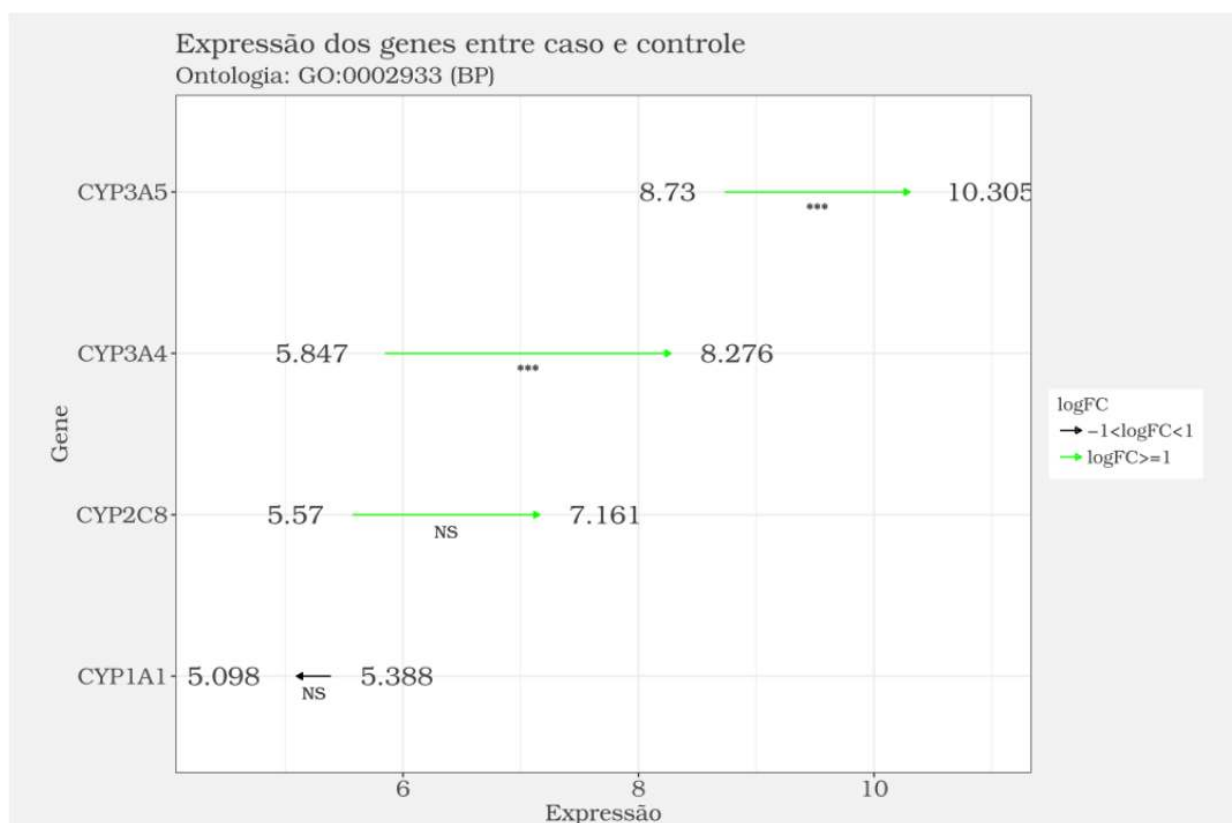


Figura 4.10: Expressão dos genes anotados na ontologia GO:0002933 para o caso *High-24h*. A figura mostra as expressões entre caso e controle dos genes anotados na ontologia **GO:0002933** (hidroxilação de lipídios). O valor na parte traseira das setas corresponde ao valor de expressão no controle e o valor na ponta corresponde a expressão no caso. Genes com um LogFC maior que 1 ou menor que -1 são considerados superexpressos e subexpressos respectivamente. Genes superexpressos possuem setas verdes, subexpressos setas vermelhas e genes que não são considerados como diferencialmente expressos possuem setas pretas. Ps: NS ≥ 0.05 , * < 0.05 , ** < 0.01 , *** < 0.001

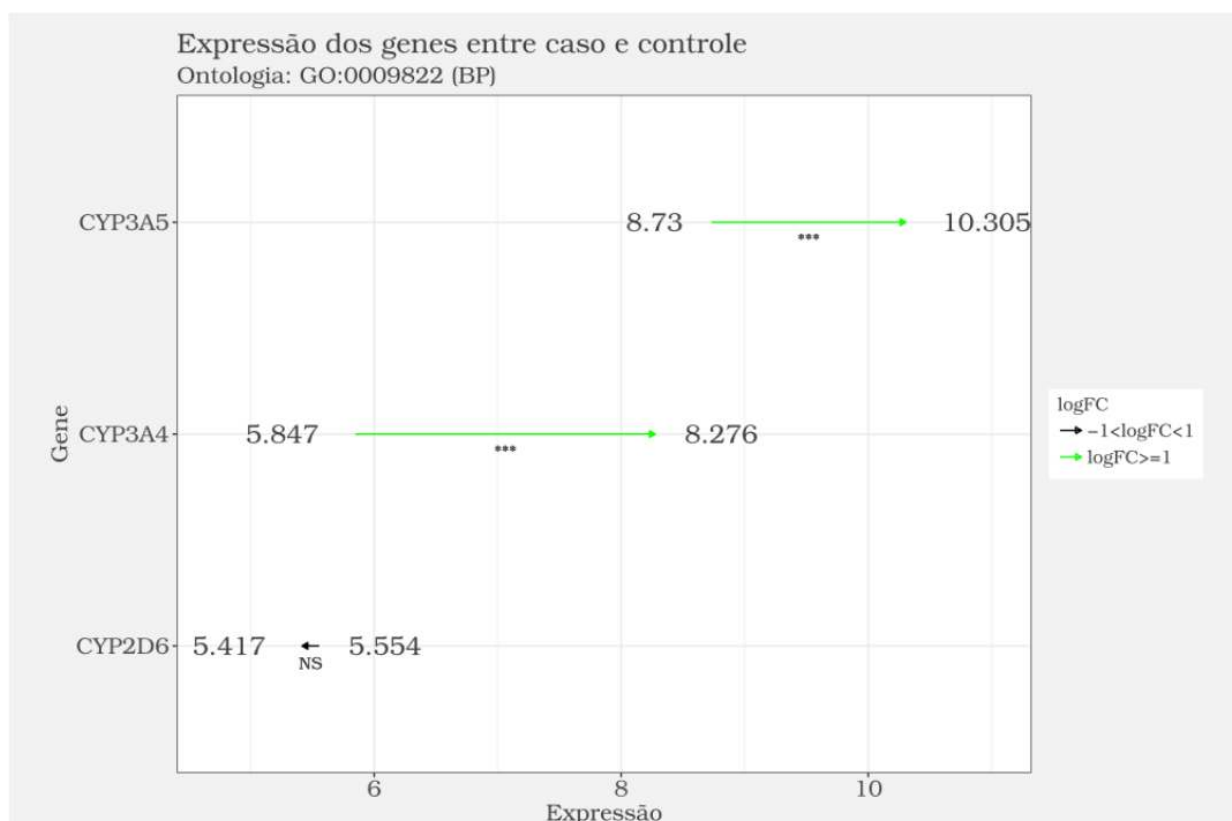


Figura 4.11: Expressão dos genes anotados na ontologia GO:0009822 para o caso *High-24h*. A figura mostra as expressões entre caso e controle dos genes anotados na ontologia GO:0009822 (processo catabólico de alcalóides). O valor na parte traseira das setas corresponde ao valor de expressão no controle e o valor na ponta corresponde a expressão no caso. Genes com um LogFC maior que 1 ou menor que -1 são considerados superexpressos e subexpressos respectivamente. Genes superexpressos possuem setas verdes, subexpressos setas vermelhas e genes que não são considerados como diferencialmente expressos possuem setas pretas. Ps: NS ≥ 0.05 , * < 0.05 , ** < 0.01 , *** < 0.001

Como pode-se observar ambas as ontologias estudadas nos gráficos apresentam GDEs que foram encontrados em todos os casos não nulos, trata-se dos GDEs CYP3A4 e CYP3A5. O gene CYP3A4 codifica uma enzima de mesmo nome que faz parte da superfamília de enzimas citocromo P450, cujos membros são monooxigenases (LEHMANN et al., 1998). A CYP3A4 especificamente é uma heme-proteína, caracterizadas pela presença de um átomo de ferro (Fe^{++}) no centro de um anel orgânico (LEHMANN et al., 1998). Diversos estudos demonstraram que esta enzima cataliza reações no metabolismo de drogas e na síntese de componentes orgânicos como colesterol, lipídios e esteróides (Inclusive a ciclofosfamida é um substrato dessa enzima) ((EKHART et al., 2008),(HUANG; ROY; WAXMAN, 2000)). O gene CYP3A5 também pertence a mesma superfamília de enzimas que o gene CYP3A4 e, portanto, também possui funcionalidade similar (EKHART et al., 2008). O mesmo se aplica a outros GDEs que foram encontrados no universo composto pelas 11 ontologias que foram encontrados em comum para os casos não nulos como: CYP3A7-CYP3A51P, CYP2C9,

CYP2C8, CYP4A11, CYP2B6, CYP1A1 e o POR. Todos esses genes estão associados com o citocromo P450 e estão superexpressos para todos os casos (dose/tempo) em que apareceram na ciclofosfamida no estudo *Homo sapiens in vitro* (rma). Isso é um comportamento que tem respaldo na literatura científica, pois basicamente indica um elevado metabolismo de xenobióticos (EKHART et al., 2008). De uma maneira geral pode-se dizer que da perspectiva das ontologias, o estudo *Homo sapiens in vitro* parece indicar uma elevada taxa de metabolismo, especificamente catabolismo de lípidios e esteróides, e é interessante notar que alguns dos GDEs dentro do universo das 11 ontologias acima também são encontrados dentro do universo de GDEs em comum as casos não nulos, são eles: CYP3A5, CYP3A4, CYP3A7-CYP3A51P, CYP2C9 e CYP2C8.

De uma perspectiva considerando apenas os 7 GDEs encontrados em comum as doses “High” e “Middle”, pode-se notar que também uma elevada quantidade de genes associados ao metabolismo de xenobióticos, e uso do sistema P450, especificamente com catabolismo de lípidios e esteróides, o mesmo tipo de atividade observado com as 11 ontologias. Entretanto é interessante notar a presença de dois GDEs que não estão diretamente envolvidos em tais processos: o EGOT e o TSKU.

O GDE EGOT *Eosinophil Granule Ontogeny Transcript* é um gene que produz um lncRNA (long non-coding RNA), um estudo mostra que a regulação positiva desse transcrito inibe a proliferação de células e promove a apoptose das mesmas no glioma humano (WU et al., 2017). Outro estudo mostra que a regulação negativa desse gene está correlacionada com tumores malignos em câncer de mama (XU et al., 2015). Em nosso estudo o EGOT se apresenta subexpresso em todos os casos (dose/tempo) em que aparece na ciclofosfamida no estudo *Homo sapiens in vitro*. Apesar de ser utilizada como antineoplásico, a ciclofosfamida também apresenta potencial carcinogênico (ZEMBLICKIS et al., 1993) (SCHMÄHL; HABS, 1979). Uma vez que o gene EGOT se encontra subexpresso nas doses média e alta do estudo, pode-se presumir que tais doses podem induzir uma regulação negativa do gene EGOT, e tomando como base os estudos, isso pode ser um biomarcador para chances elevadas de efeitos carcinogênicos induzidos pela droga. Entretanto outra maneira de interpretar a regulação negativa do EGOT, é considerar que o estudo com a ciclofosfamida foi feito em hepatócitos, células que possuem grande capacidade de proliferação mediante danos de diversas origens ao organismo. Como a ciclofosfamida certamente causa tais danos, espera-se que o tecido reaja com reparação, o que explicaria a regulação negativa da gene EGOT.

O gene TSKU (*small leucine rich proteoglycan*) (WEIZMANN, 2018) se apresenta superexpresso em todos os casos que aparece no estudo *Homo sapiens in vitro* (rma) da ciclofosfamida. Um estudo (JANSSEN et al., 2015) mostra que esse gene fica positivamente regulado mediante ativação do PPar α , um fator de transcrição presente no tecido hepático que está envolvido na regulação da expressão de genes associados com: diferenciação celular, desenvolvimento, metabolismo (carboidratos, lipídios, proteínas), e carcinogênese. Embora o gene TSKU esteja pouco caracterizado atualmente, pode-se presumir que sua superexpressão em nossos estudos tem respaldo na literatura científica, os resultados obtidos indicam uma elevada metabolização de xenobióticos.

Para uma visualização mais ampla dos resultados, a figura 4.12 abaixo foi criada, contendo redes de interação proteína-proteína (PPI) montadas no STRING a partir do universo de GDEs encontrados no estudo *Homo sapiens in vitro*, utilizando um score de pelo menos 0.9, com interações do tipo *experiments* e *Databases*. Dados de expressão gênica foram projetados sobre a rede formando no final um uma espécie de *heatmap*. O nível de expressão (Nexp) para cada nodo nas projeções são determinados utilizando dados de expressão de um caso e um controle utilizando a fórmula:

$$N_{exp} = \frac{\text{Caso}}{\text{Caso} + \text{Controle}} \quad (4.1)$$

Dessa maneira genes cuja expressão não se alterou entre caso e controle possuiram um Nexp igual a 0,5. Se a expressão for maior no caso então o Nexp será maior que 0,5 e menor na situação oposta.

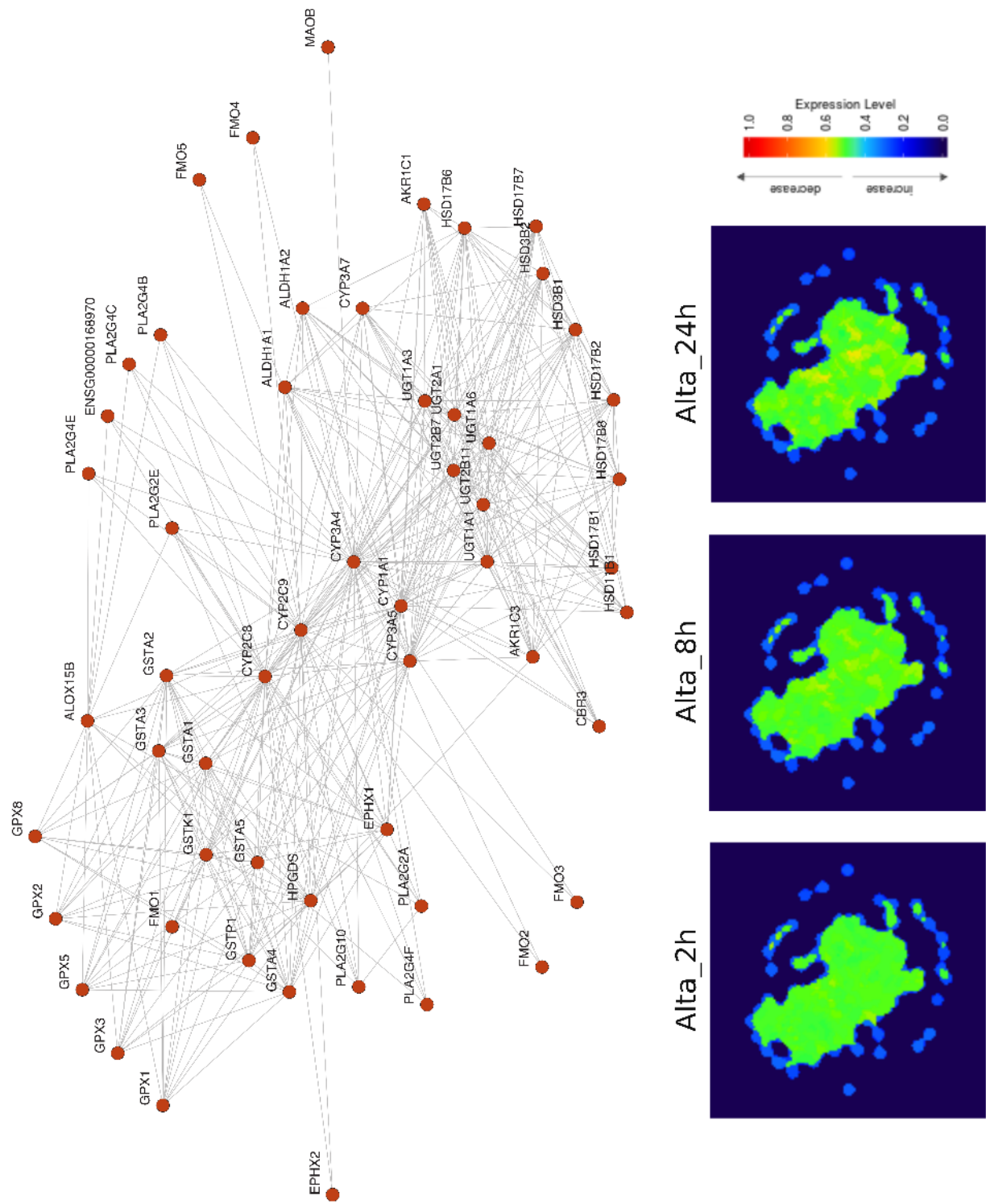


Figura 4.12: Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo *Homo sapiens in vitro* na ciclofosfamida com uso do STRING. Ao lado 3 projeções de dados de expressão gênica sobre a mesma rede, cada projeção foi feita com dados de expressão referentes a tempos diferentes: 2 h, 8 h e 24 h

Podemos observar vários genes associados com o metabolismo de xenobióticos na rede, como os genes CYP, importantes para o sistema citocromo P450, como: CYP3A4, CYP3A5, CYP2C9, CYP1A1 todos altamente conectados na rede e sua expressão aumenta no caso em relação ao controle conforme o tempo aumenta como indica o amarelamento da rede.

Os genes ALDH relacionados com as aldeído-desidrogenases como ALDH1A1 e ALDH1A2 também mostram expressão que se eleva no caso em relação ao controle conforme o tempo aumenta (WEIZMANN, 2018). Uma vez que a ciclofosfamida possui um metabólito chamado aldofosfamida, um tautômero de seu principal metabólito a 4-hidroxíciclofosfamida, o qual é derivada do metabolismo da ciclofosfamida pelo sistema P450 (BODDY; YULE, 2000). Observamos que as enzimas do CYP estão, em geral, superexpressas e com isso pode se deduzir que as quantidades de metabólitos da ciclofosfamida no organismo aumentarão o que por sua vez vai causar um aumento na expressão de genes relacionados com o metabolismo de aldeídos.

Os genes UGT que codificam glucuronosiltransferases, enzimas envolvidas com o metabolismo de moléculas lipofílicas (esteróides, hormônios e diversas drogas) como: UGT1A1, UGT2B11, UGT1A3, UGT2A1, UGT2B7 e UGT2B6, também apresentam tom amarelo nas projeções conforme o tempo aumenta, indicando sua regulação positiva na administração da ciclofosfamida (WEIZMANN, 2018). Estudos recentes sugerem que estas enzimas tem um papel fundamental no metabolismo de segunda fase da ciclofosfamida (VEAL et al., 2016) (PETROS et al., 2005).

A maioria dos genes nas áreas mais amareladas (indicando regulação positiva dos genes) estão relacionados com o metabolismo de xenobióticos, essa foi mesma conclusão que foi derivada da observação das ontologias e dos GDEs em comum aos casos não nulos. Outros genes de importância ao metabolismo observados na rede incluem as monoamina oxigenases, as flavina monooxigenases, glutathione S transferases e glutathione peroxidases. Naturalmente a rede não contém genes que não se traduzem em proteínas, como o TSKU.

No estudo do *Rattus norvegicus in vitro* também não foram encontradas BPs que estavam simultaneamente nas 3 doses, foi encontrada apenas uma única ontologia entre as doses “High” e “Middle”: I) GO:0052695 glucoronidação celular (hepática)

A maior parte dos GDEs subexpressos existentes entre os 92 encontrados no estudo do *Rattus norvegicus in vitro* estão associados com proteínas da membrana celular, com trans-

porte, motilidade celular e citoesqueleto. Isso indica uma regulação negativa do transporte transmembrana (especialmente do influxo de cálcio) e de todos os processos celulares dependentes disso. Em geral as ontologias BP indicam uma regulação negativa da transporte transmembrana (quimiotaxia e das proteínas na membrana, em geral), da migração e da motilidade celular, da secreção de substâncias pelas células. Quanto aos superexpressos a maioria está associado ao metabolismo de drogas especialmente pelas vias CYP, UGT, GST e ALDH. É interessante observar que um dos genes superexpressos, CKAP2, está associado com a estabilização do citoesqueleto, regulação do ciclo celular mitótico e a apoptose celular (WEIZMANN, 2018).

Quanto aos GDEs apenas quatro foram encontrados em comum entre as doses “High” e “Middle”: UGT2B7, ALDH1A7, CCNA2 e UGT2B1.

A glucoronidação é um processo importante para a metabolização de xenobióticos, e se consiste na transferência de uma unidade de ácido glucorônico a um substrato da enzima uridina 5'-difosfo-glucuronosiltransferase (UDP-glucuronosiltransferase, UGT) (KING et al., 2000). Sabe-se que esta enzima é afetada pela metabolização da ciclofosfamida, observação da ontologia GO:0052695 em comum a múltiplas doses no estudo *Rattus norvegicus in vitro* tem respaldo na literatura científica como mostrado em (LEAR; NATION; STUPANS, 1992). Existem 3 genes do *Rattus norvegicus* a essa ontologia, 2 deles são GDEs superexpressos: UGT2B1 e UGT2B7, ambos envolvidos com o processo de glucoronidação (WEIZMANN, 2018). Curiosamente ambos também são encontrados no universo formado pela comparação dos GDEs nas doses “High” e “Middle”. Assim sendo pode-se concluir que da perspectiva das ontologias, o estudo *Rattus norvegicus in vitro* também possui elevada atividade relacionada com metabolismo de xenobióticos, embora por uma via diferente da identificada no *Homo sapiens*.

Na perspectiva dos GDEs em comum a múltiplas doses no estudo *Rattus norvegicus in vitro*, dos quatro encontrados, dois estão envolvidos com o processo de glucoronidação. Os outros dois, ALDH1A7 e CCNA2 serão discutidos afrente.

O ALDH1A7 transcreve uma proteína envolvida com a atividade de enzimas pertencentes a família das aldeído-desidrogenases, este GDE está superexpresso em todas os casos que aparece. Um dos metabólitos da ciclofosfamida é a aldociclofosfamida, esta família de enzimas promove a oxidação de aldeídos, maneira pela qual a aldociclofosfamida é metabolizada (WEIZMANN, 2018).

O CCNA2 é um gene cuja a atividade está envolvida com a transição do ciclo celular entre fases G1/S e G2/M. Este gene sempre aparece superexpresso no estudo *Rattus norvegicus in vitro* (WEIZMANN, 2018).

Uma rede foi montada com os 4 GDEs mencionados acima no STRING, utilizando as interações *Experiments* e *Databases*, com no máximo 50 interações por nodo e um *score* mínimo de 0.90.

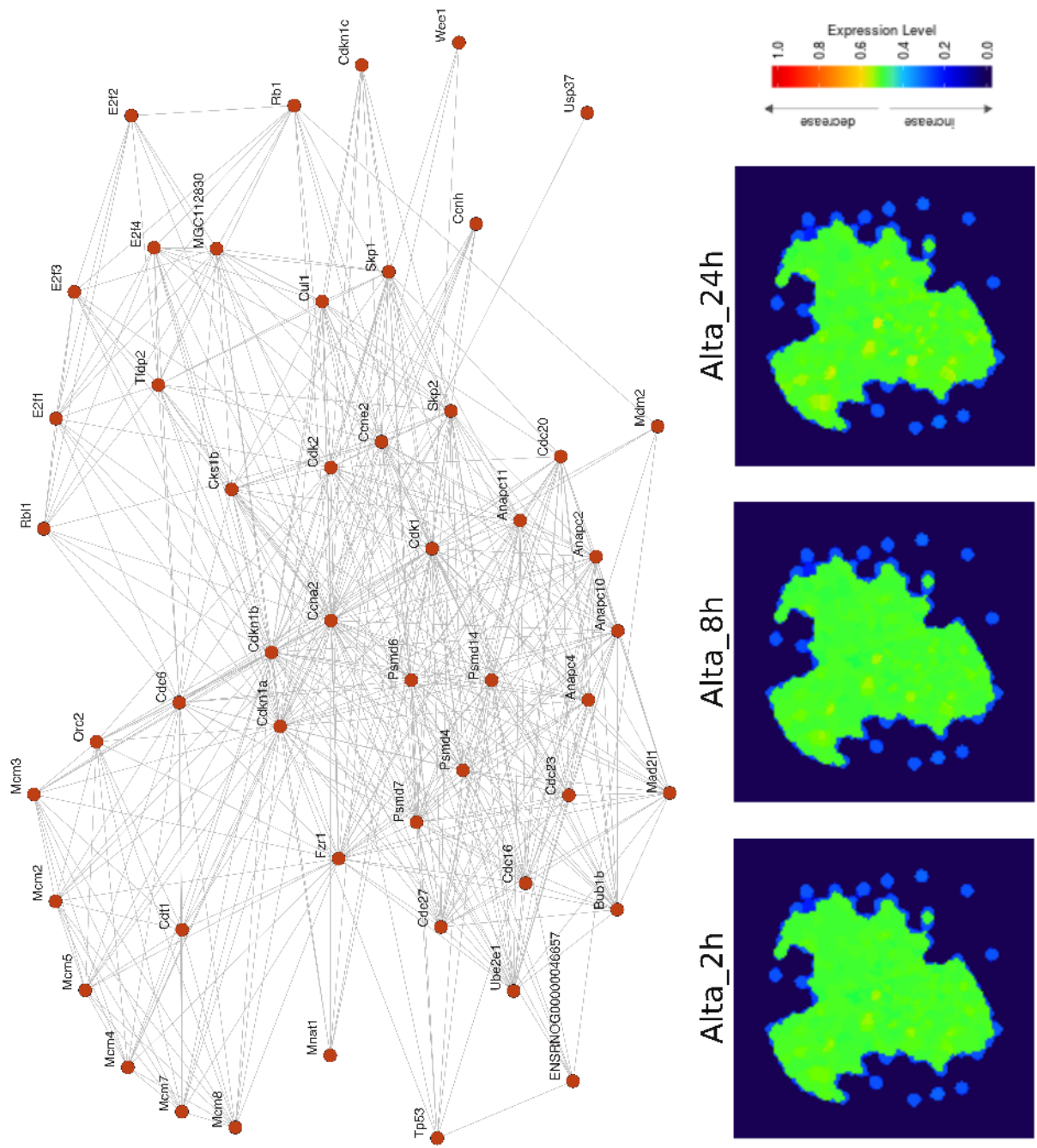


Figura 4.13: Rede GPI derivada dos GDEs em comum aos casos não nulos do estudo *Rattus norvegicus in vitro* na ciclofosfamida com uso do STRING. Ao lado 3 projeções de dados de expressão gênica sobre a mesma rede, cada projeção foi feita com dados de expressão referentes a tempos diferentes: 2 h, 8 h e 24 h

A maior parte dos genes da acima estão associados com a regulação do ciclo celular incluindo quinases dependentes de ciclinas (Cdc), proteínas inibidoras de quinase dependentes de ciclina (Cdkn), as próprias ciclinas (Ccn) e complexo protéico de manutenção de minicromossomos (Mcm) (WEIZMANN, 2018). A maioria desses genes estão localizados em áreas amareladas da rede, indicando sua superexpressão e portanto uma regulação positiva da transição do ciclo celular, posivelmente as células do tecido hepático estão em processo de proliferação, o que obviamente remete-nos ao processo de mitose. Entretanto a presença de alguns outros genes, como , superexpressos indica que o processo de mitose pode estar sendo interrompido, o que se é esperado em função da ciclofosfamida.

Dentro do universo de 30 GDEs no estudo do *Rattus norvegicus in vivo*, a maioria dos genes subexpressos estava associados com metabolismo de lípidios. Genes chaves no metabolismo e biosíntese de ácidos graxos, lípidios, triglicerídeos, como PNPLA3, FABP5, THRSP, SCD, FASN e ACACA sofrem uma regulação negativa massiva no estudo do *Rattus norvegicus in vitro*. Os genes superexpressos estão associados com oxidação do grupo heme, controle do crescimento celular, apoptose e resposta a danos ao DNA. Alguns exemplos incluem: IFIT2, IRF1, HMOX1, PML e STAT3 (WEIZMANN, 2018).

A maioria das 92 ontologias encontradas no estudo do *Rattus norvegicus in vivo* está associada com processos que foram também observados no universo de GDEs: síntese e metabolismo de ácidos graxos, lípidios e triglicerídeos. Quanto as ontologias foram encontradas 2 em comum as 3 doses, 7 entre “*High*” e “*Middle*”, 4 entre “*Middle*” e “*Low*” e 2 entre “*High*” e “*Low*”.

Quanto aos GDEs, foi encontrado apenas 1, entre as doses “*High*” e “*Middle*”: o GDE Pnpla3.

Lista de ontologias em comum as 3 doses: I) GO:0036153–remodelamento da cadeia de acil triglicerídeos; ii) GO:0010477–resposta a dióxido de enxofre.

Lista de ontologias em comum as doses “*High*” e “*Middle*”: i) GO:0036153–remodelamento da cadeia de acil triglicerídeos; ii) GO:0010477–resposta a dióxido de enxofre; iii) GO:0000320–reentrada no ciclo celular mitótico; iv) GO:0033197–resposta a vitamina E; v) GO:0007595–lactação; vi) GO:0032764–regulação negativa da produção de citocinas de mastócitos; vii) GO:0006788–oxidação heme.

Lista de ontologias em comum as doses “*High*” e “*Low*”: i) GO:0036153–remodelamento da cadeia de acil triglicerídeos; ii) GO:0010477–resposta a dióxido de enxofre.

Lista de ontologias em comum as doses “*Middle*” e “*Low*”: i) GO:0036153–remodelamento da cadeia de acil triglicerídeos; ii) GO:0010477–resposta a dióxido de enxofre; iii) GO:0019433–processo catabólico de triglicerídeos; iv) GO:0019432–processo biossintético de triglicerídeos.

A maioria das ontologias tem relação com o metabolismo de triglicerídeos, inclusive o GDE Pnpla3, está presente na maioria deles. Este GDE codifica uma triacilglicerol lipase (proteína) que atua na hidrólise de triglicerídeos em adipócitos (WEIZMANN, 2018). Este gene se encontra subexpresso no estudo *Rattus norvegicus in vivo*.

Devido ao número menor de GDEs encontrados no estudo *Rattus norvegicus in vivo*, foi construída uma rede no utilizando todo o universo de GDEs do estudo. Nos outros dois estudos anteriores as redes foram construídas com GDEs que foram considerados mais relevantes. Os parâmetros para a construção da rede no STRING permanecem idênticos.

4.2.3 Etopósido

Abaixo segue as análises dos resultados para o etopósido no método EFDEGs, seguindo a mesma abordagem utilizada na ciclofosfamida. As figuras 4.16 e 4.15 demonstram a distribuição de GDEs encontrados para o etopósido.

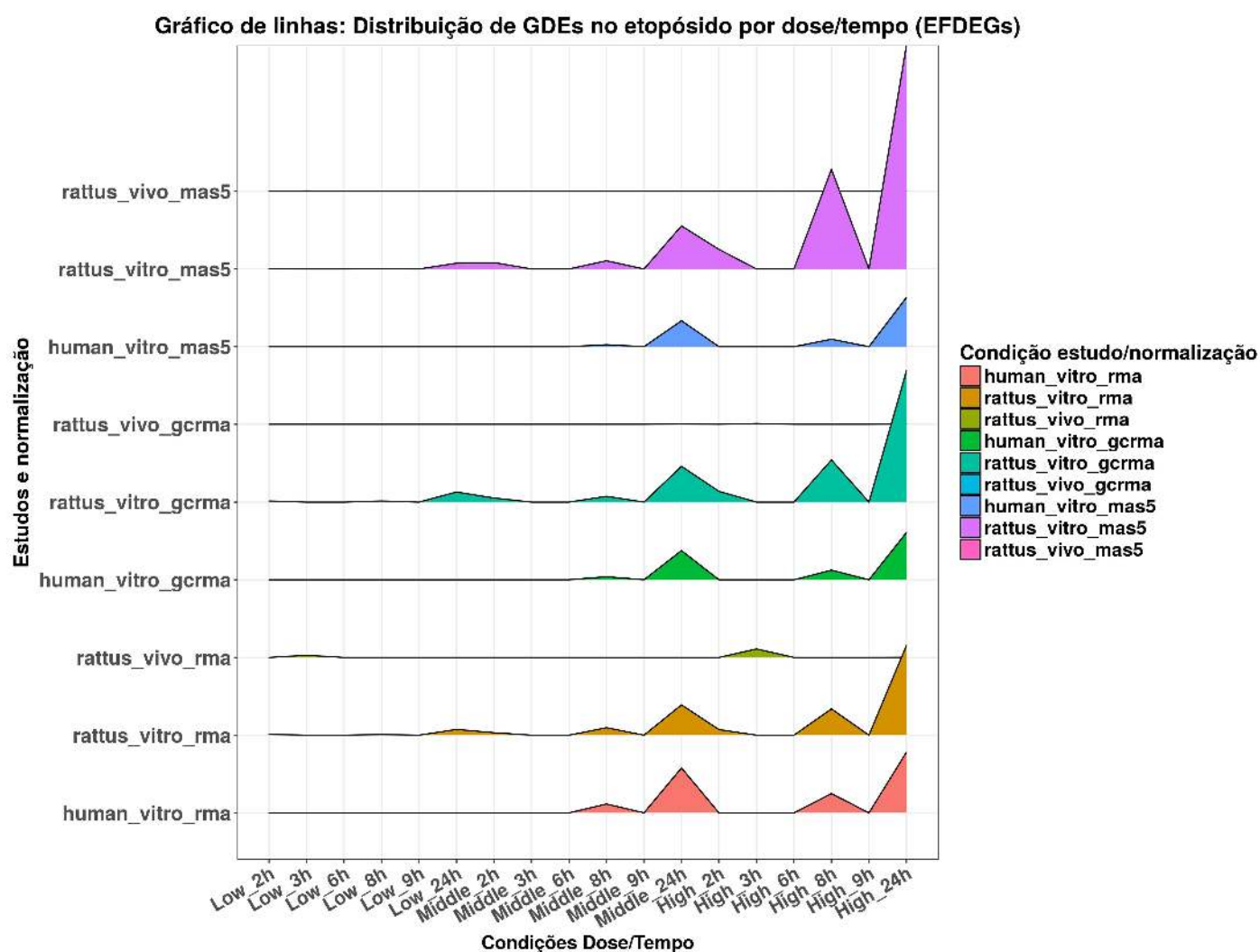


Figura 4.15: Gráfico de linhas para os GDEs da etopósido no método EFDEGs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.16, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de GDEs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.16 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Etopósido: Distribuição de GDEs para o método EFDEGs

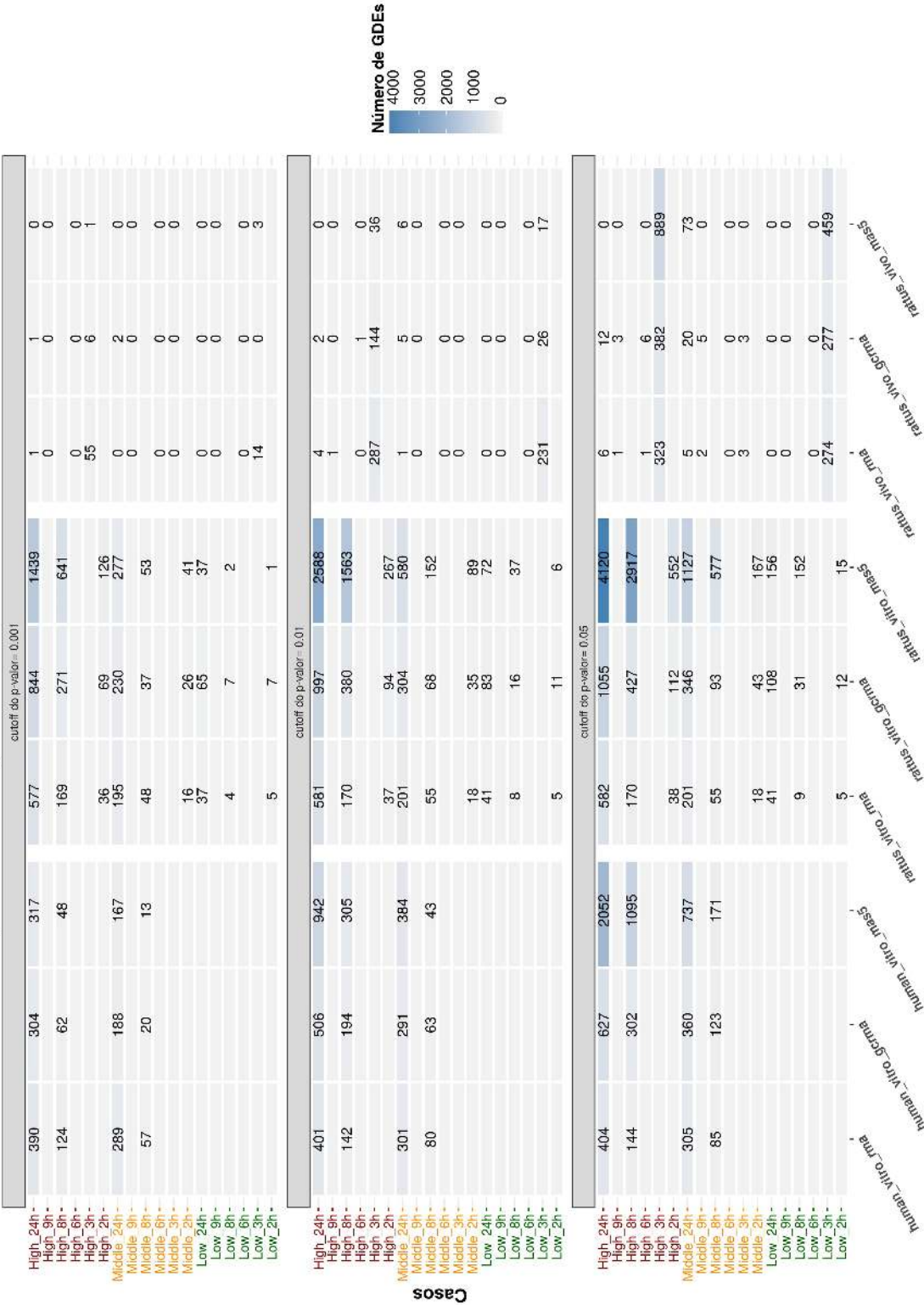


Figura 4.16: GDEs encontrados para a etopósido no método EFDEGs. Distribuição geral de GDEs encontrados no método EFDEGs para o etopósido considerando 3 diferentes valores de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que as células vazias indicam a inexistência do experimento em tais condições. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

Assim como na ciclofosfamida, as figuras 4.16 e 4.15 deixam claro que algumas das tendências são iguais as dos resultados gerais nas figuras 4.2 e 4.3, porém outras são diferentes. Abaixo segue uma sumarização das tendências observadas segundo as figuras 4.16 e 4.15:

- O método de normalização “mas5” retorna o maior número de resultados, seguido do “gcrma” e do “rma”. A mesma sequência dos resultados gerais, portanto esse comportamento permanece inalterado.
- Quanto maior a dose maior é o número de resultados entre os tempos dos mesmos.
- Os estudos *in vitro* retornam mais resultados que os estudos *in vivo*.
- Os “ciclos de ativação” dos estudos *in vitro* permacem com o mesmo comportamento apresentado nos resultados gerais (Figura 4.3), isto é, quanto maior a dose e o tempo maior é o número de resultados, sendo que o caso *High-24h* é o que sempre retorna o maior número de resultados.
- Os “ciclos de ativação” continuam presentes porém estão diferentes para os estudos *in vivo*. Note que no tempo de 24 h praticamente não são encontrados resultados, ao contrário do que acontece nos resultados gerais segundo as figuras 4.2 e 4.3 ou mesmo na ciclofosfamida segundo as figuras 4.8 e 4.9. Isso mostra que o comportamento observado nos resultados gerais, onde os estudos *in vivo* apresentavam picos elevados no tempo de 3 h para as doses “*Low*” e “*High*” provavelmente se devem a um comportamento intrínseco ao etopósido. Esses resultados sugerem que, mesmo de um ponto de vista puramente quantitativo, a maneira de agir do etopósido é muito diferente entre estudos *in vivo* e *in vitro*. Será interessante realizar uma análise biológica/funcional desses resultados posteriormente.

O próximo passo é observar as tendências e comportamentos das ontologias e checar se acompanham o comportamento dos GDEs.

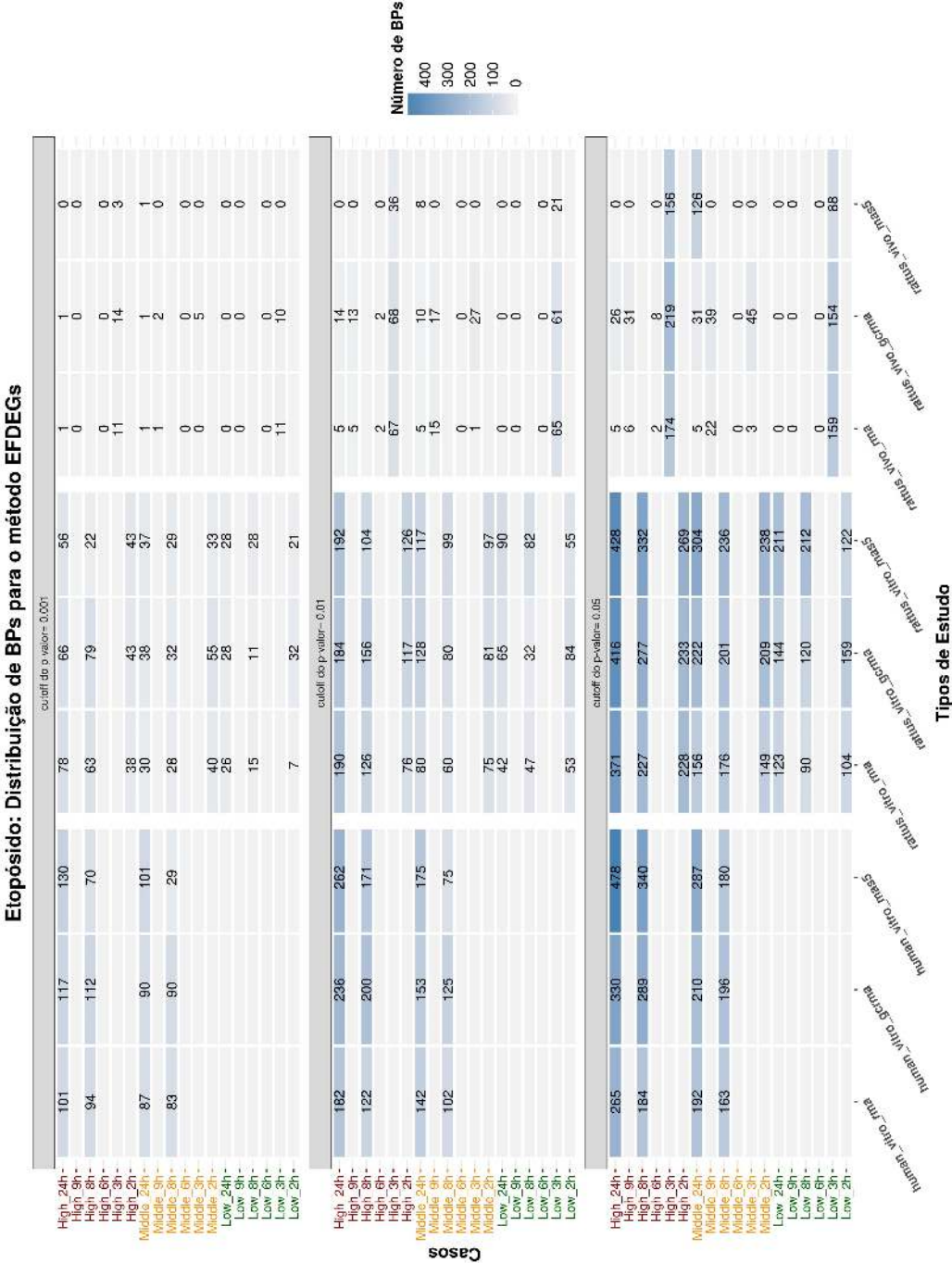


Figura 4.17: BPs encontrados para o etopósido no método EFDEGs. Distribuição geral de BPs encontrados no método EFDEGs para o etopósido considerando 3 valores diferentes de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que as células vazias indicam a inexistência do experimento em tais condições. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de BPs encontrados nas diferentes situações. As significâncias das ontologias são determinadas com uso do pacote “topGO” da plataforma *bioconductor*, utilizando uma lista de GDEs como entrada, um algoritmo de correção gráfica para determinação de pesos para as ontologias chamado “*weight01*” e por fim o teste exato de Fisher para determinação das significâncias.

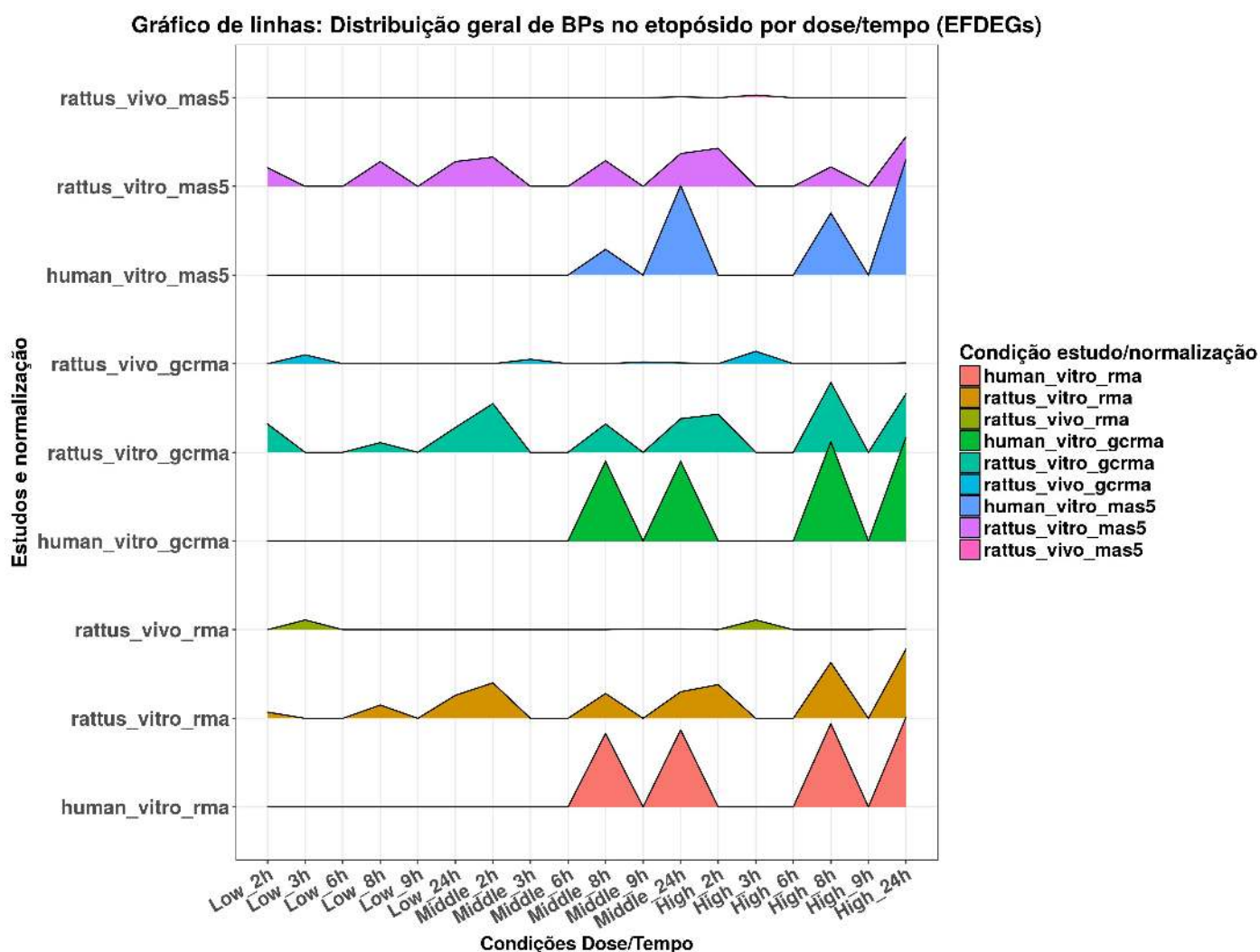


Figura 4.18: Gráfico de linhas para os BPs do etopósido no método EFDEGs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.17, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.17 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

As figuras 4.17 e 4.18 mostram um comportamento parecido com os comportamentos evidenciados em 4.16 e 4.15, portanto assim como nos resultados gerais e na ciclofosfamida os resultados das ontologias significantes estão alinhados com os dos GDEs.

Agora uma breve descrição numérica dos resultados (GDEs/BPs) que foram encontrados nos 3 estudos do etopósido. O estudo *Homo sapiens in vitro* encontrou entre todas as doses e tempos 860 GDEs e 400 ontologias. O estudo do *Rattus norvegicus in vitro* encontrou 1087 GDEs e 359 BPs. Por fim o estudo do *Rattus norvegicus in vivo* encontrou 70 GDEs e 28 BPs. Os resultados dos estudos *in vitro* são especialmente numerosos como esperado,

segundo as figuras 4.16, 4.15, 4.17 e 4.18. O *in vivo* tem pouquíssimos resultados em comparação com os *in vitro*, segundo as mesmas figuras, refletindo talvez uma característica intrínseca do estudo *in vivo*.

Para o etopósido não foram encontrados GDEs em comum aos 3 estudos, nem mesmo quando a comparação foi feita entre *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo*. Entretanto foram encontradas 3 ontologias BP em comum aos 3 estudos (desconsiderando a separação por dose/tempo), algo que não aconteceu para a ciclofosfamida, essas ontologias são: i) GO:0046680–resposta a DDT; ii) GO:2000775–fosforilação da histona H3-S10 envolvida na condensação cromossômica; iii) GO:0071283–resposta celular ao ion de ferro (III).

As 3 ontologias acima tem anotadas em si um gene em comum: o CCNB1, existente no *Homo sapiens* e no *Rattus norvegicus*. Esse gene codifica a proteína ciclina B1, importante na regulação do ciclo celular, especialmente na transição da fase G2 para a fase mitótica (WEIZMANN, 2018). Em nossos estudos do etopósido esse gene sempre apareceu como subexpresso, indicando uma regulação negativa da fase mitótica. O etopósido é um antineoplásico que age como um inibidor da enzima topoisomerase II, que auxilia nos processos de replicação e enrolamento do DNA (MAANEN et al., 1988). Células tumorais, por se dividirem mais, dependem mais do correto funcionamento dessa enzima. Ao inibir essa enzima o etopósido faz com que as fitas de DNA não sejam religadas corretamente durante a replicação do DNA, induzindo a apoptose da célula. Dessa maneira a observação do CCNB1 como um GDE nos 3 estudos do etopósido é algo que faz sentido.

As 3 ontologias que foram encontradas em comum para os 3 estudos, também foram as únicas ontologias BP encontradas em comum em 2 comparações 2x2:

- *Homo sapiens in vitro* vs *Rattus norvegicus in vivo* (hvt-rvv)
- *Rattus norvegicus in vitro* vs *Rattus norvegicus in vivo* (rvt-rvv)

Entretanto a combinação hvt-rvt, ambos estudos *in vitro*, retorna 116 BPs em comum (53 únicas, desconsiderando as repetições) e 132 GDEs, um número bem elevado em comparação com as outras, refletindo talvez uma similaridade entre os estudos *in vitro*. A maioria das ontologias encontradas possuem anotadas em si GDEs associadas com a regulação da replicação do DNA, da evolução do ciclo celular, do citoesqueleto, fuso mitótico, dano ao

DNA e reparo de DNA. Alguns exemplos:

Replicação do DNA e seu reparo (WEIZMANN, 2018): UHRF1, PCLAF, CDT1, EXO1, GMNN, MCM3 MCM5, MCM6, MCM7, PCNA, PTTG1, RFC3, RFC4, RFC5, TOP2A, TMPO, UBE2T, CHEK1, GINS1, POLE2, PRIM1, TK1, TRIP13, MSH2, DSCC1, DUT, FANCD2, FOXM1, GINS4, RRM1, TIMELESS, MGME1, POLA1.

Com as ciclinas e seus reguladores positivos (WEIZMANN, 2018): CCNE2, CCNB1, CCNA2, CDK1, CKS2, CDKN3, SKP2.

Reguladores do ciclo celular e do fuso mitótico (WEIZMANN, 2018): CDC20, PRC1, NUSAP1, AURKA, DLGAP5, FAM83D, MELK, ATAD2, BIRC5, TTK, ZWILCH, ASF1B, ASPM, BUB1B, MAD2L1, PLK4, TCF19, TRIP13, STIL, AURKB, BUB1, CDCA2, CDCA3, CDCA8, CENPF, CENPK, CENPN, CENPW, CKAP2, DSN1, ESCO2, KIF11, KIF18B, KIF23, KIF2C, KIFC1, KNTC1, MIS18A, MKI67, MND1, NCAPD2, NDC1, NUF2, RAC-GAP1, SHCBP1, SKA1, SKA3, SMC2, SPAG5, SPC24, SPC25, STMN1, TACC3, CENPQ e UBE2C.

Todos os genes mencionados acima estão subexpressos nos estudos do *Homo sapiens in vitro* e do *Rattus norvegicus in vitro*, indicando uma regulação negativa dos processos a que estão associados. Uma vez que o etopósido age como um inibidor da topoisomerase II é natural esperar que mecanismos de replicação do DNA, reparo do DNA, progresso da fase mitótico e regulação do fuso mitótico estejam negativamente regulados (MAANEN et al., 1988). É interessante notar que dentro desse universo de 132 GDEs encontrados entre os estudos *in vitro*, poucos estavam superexpressos, entretanto estavam associados com resposta a danos a DNA GDF15, regulação negativa do ciclo celular CDKN1A e apoptose TP53INP1 (WEIZMANN, 2018). Curiosamente o gene MDM2 também está superexpresso nos estudos *in vitro*, um estudo relata que muitos tumores possuem esse gene positivamente regulado entretanto o mesmo estudo afirma que o etopósido eleva os níveis do transcrito desse gene e que o mesmo pode estar associado a resposta de danos ao DNA (MANFREDI, 2010). Um outro estudo mostra que inibidores do gene MDM2 potencializam o efeito de inibidores da topoisomerase II (tal como o etopósido), induzindo mais quebras na fita dupla de DNA e por consequência, mais apoptoses celulares (CONRADT et al., 2013).

O estudo *Homo sapiens in vitro* encontrou 174 ontologias em comum (132 únicas, desconsiderando repetições) a todos os seus casos (dose, desconsiderando o tempo) não nulos no etopósido, no caso os resultados das doses “*High*” e “*Middle*”, visto que a dose “*Low*” não encontrou nenhuma ontologia e também nenhum GDE. Quanto aos GDEs, foram encontrados um total de 210 em comum entre as doses “*High*” e “*Middle*”.

A maioria das ontologias estão relacionadas com a regulação da replicação do DNA, com regulação da mitose, e reparo de danos dano ao DNA. Esse universo de ontologias continha anotado em si, 161 GDEs, que foram utilizados para construir as redes abaixo. Os parametros para a montagem da rede no STRING permanecem inalterados.

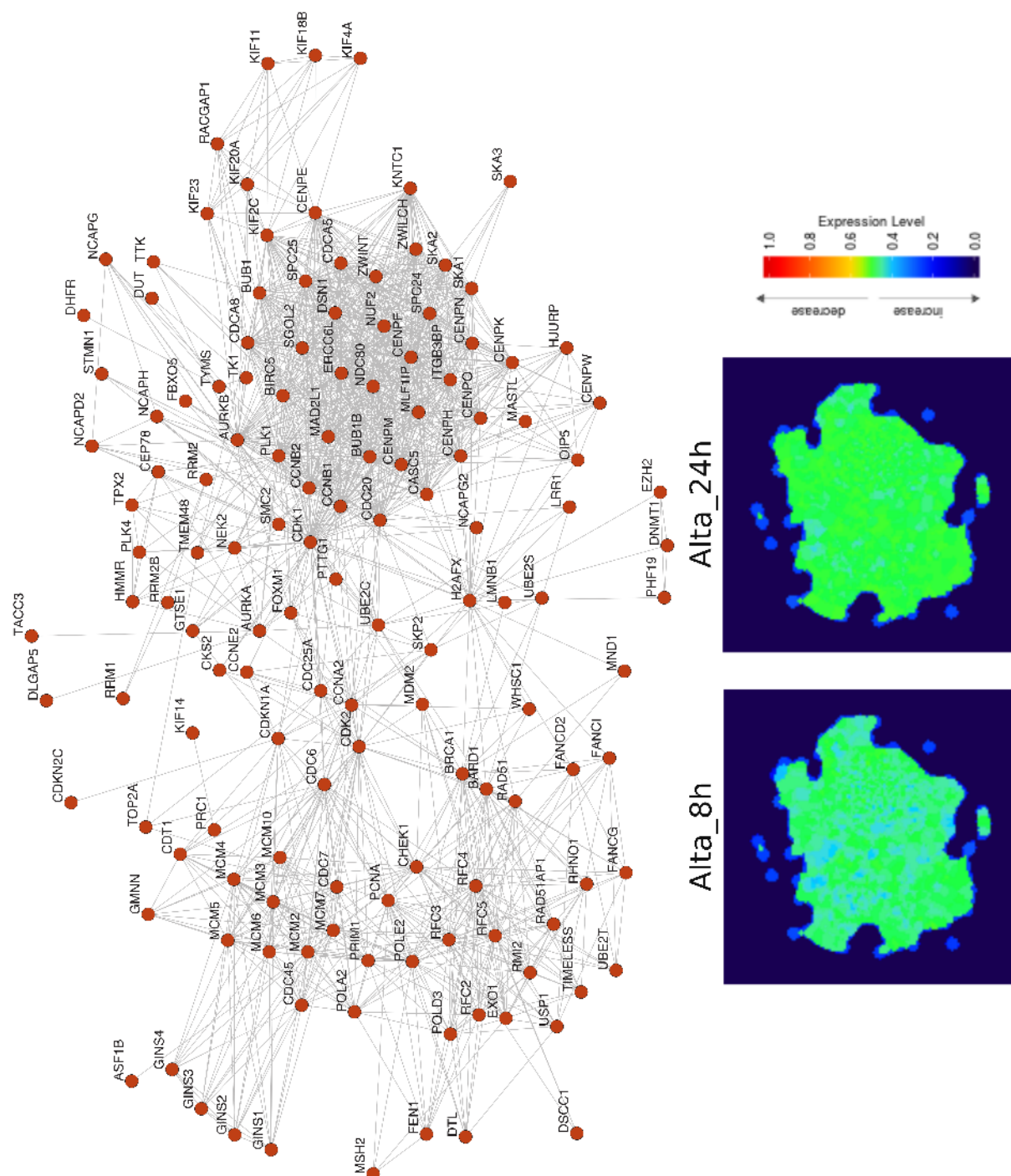


Figura 4.19: Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo *Homo sapiens in vitro* no etopósido com uso do STRING. Ao lado 2 projeções de dados de expressão gênica sobre a mesma rede, cada projeção foi feita com dados de expressão referentes a tempos diferentes: 8 h e 24 h. Nesse estudo em específico o TG-GATES optou por omitir o experimento com 2 h.

Pode-se observar na rede acima que muitos de seus genes estão presentes no universo de GDEs encontrados em comum entre os dois estudos *in vitro*. De fato, genes associados a ciclinas e seus reguladores (CCNA2, CCNE2, CCNB2, CDKN1A, CDK1, CDK2, genes RFC e genes MCM), replicação do DNA e reparo de danos ao DNA (TOP2A, GINS1 GINS4, genes POL, PCNA) e regulação do ciclo celular e do fuso mitótico (genes CENP, genes KIF e genes CDC) (WEIZMANN, 2018). A grande maioria desses genes está negativamente regulado como mostra as regiões com tom azulado na rede (projeção 8 h). Conforme a tempo passa muitos desses genes tem sua expressão aumentada como mostra a projeção de 24 h, entretanto parte considerável continua com regulação negativa.

O estudo do *Rattus norvegicus in vitro* tem um número bem elevado de BPs e GDEs, entretanto parte considerável deles estão associadas com os mesmos processos encontrados para o *Homo sapiens in vitro*: regulação da replicação do DNA, da evolução do ciclo celular, do citoesqueleto, fuso mitótico ,dano ao DNA e reparo de DNA.

O estudo do *Rattus norvegicus in vitro* encontrou 13 ontologias que estavam presentes em todas as 3 doses (desconsiderando o tempo).as ontologias são: i) GO:0000915–montagem do anel contrátil de actomiosina; ii) GO:0070098–via de sinalização mediada por quimiocinas; iii) GO:0008354–migração de células germinativas; iv) GO:0035701–migração de células-tronco hematopoiéticas; v) GO:2000775–fosforilação de histona H3-S10 envolvida na condensação cromossômica; vi) GO:0000281–citocinese mitótica; vi)I) GO:0051256–montagem da zona intermediária do fuso mitótico; viii) GO:0070374–regulação positiva da cascata ERK1 e ERK2; ix) GO:1905322–regulação positiva da migração de células-tronco mesenquimais; x) GO:0035470–regulação positiva da cicatrização vascular; xi) GO:0008360–regulação da forma celular; xii) GO:0046680–resposta ao DDT; xiii) GO:1990478–resposta ao ultra-som.

É interessante notar como algumas dessas ontologias estão relacionadas com processos em que o etopósido atua direta ou indiretamente como a regulação da mitose a condensação cromossômica, entretanto também percebe-se uma elevada atividade de quimiocinas. Esse universo de ontologias possui anotado em si 49 GDEs, que foram utilizados para montar a rede abaixo. Grande parte destes genes estão associados a atividade de quimiocinas.

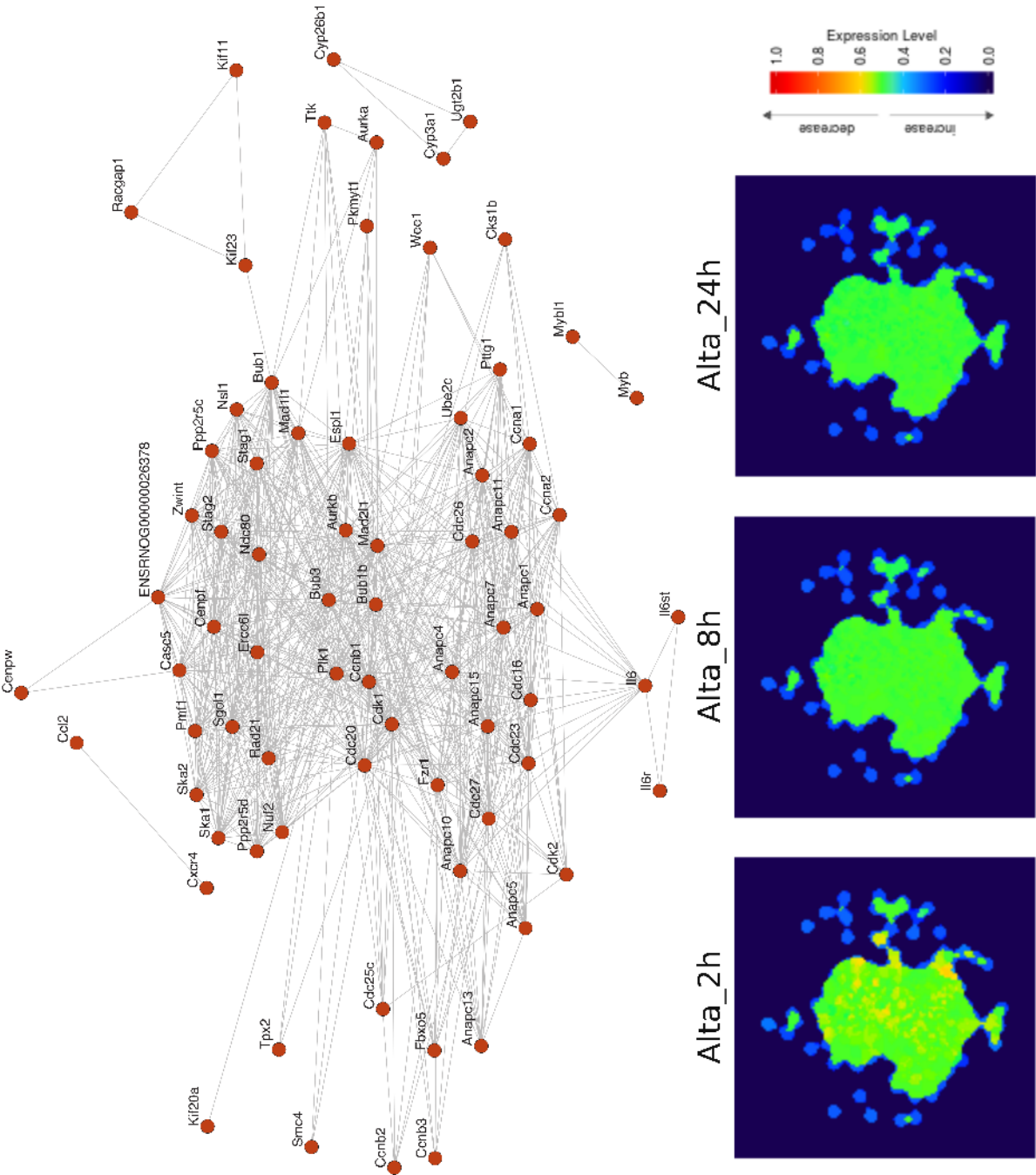


Figura 4.20: Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo *Homo sapiens in vitro* no etopósido com uso do STRING. Ao lado 3 projeções de dados de expressão gênica sobre a mesma rede, cada projeção foi feita com dados de expressão referentes a tempos diferentes: 2 h, 8 h e 24 h.

Nas redes ilustradas em 4.21 fica aparente um comportamento em que conforme o tempo aumenta (de 2 h à 24 h), a expressão diferencial do caso em diversos genes vai desaparecendo, ficando quase que igual ao do controle na dose “Alta_24h”. A maioria desses genes estão relacionados com os ciclinas (genes CCN), ponto de controle da mitose (genes BUB), aurora quinase (AURKA) e fuso mitótico (genes SKA) (WEIZMANN, 2018). Todos esses processos são diretamente afetados pelo mecanismo de ação do etopósido (inibição da topoisomeraseii, portanto observar a alteração de genes ligados a esses processos é algo que faz sentido (MAANEN et al., 1988).

Também foram encontradas, para o estudo *Rattus norvegicus in vitro*, ontologias em comum em outras combinações de doses. Foram encontradas 50 entre as doses “High” e “Middle”, 15 entre “High” e “Low” e por fim, 22 entre “Middle” e “Low”. Todas as ontologias e GDEs encontrados nesses universos, estão relacionados com regulação da mitose a condensação cromossômica, danos ao DNA e atividade quimiocinas. Muitos desses processos também foram observados no universo de ontologias em comum das 3 doses, dando robustez a tal observação.

No estudo do *Rattus norvegicus in vivo*, a comparação entre as doses “High” e “Low” foi a única que encontrou ontologias em comum. Foram aqui encontradas 6 ontologias: i) GO:0071243–resposta celular à substância contendo arsênico; ii) GO:1901143–processo catabólico de insulina; iii) GO:0010801–regulação negativa da fosforilação da peptidil-treonina; iv) GO:0055091–homeostase de fosfolípidios; v) GO:0045821–regulação positiva do processo glicolítico; vi) GO:1901409–regulação positiva da fosforilação do domínio C-terminal da RNA polimerase II.

Este universo de ontologias contém anotado em si 4 GDEs, que foram utilizados para montar as redes abaixo. Curiosamente estes 4 GDEs estão subexpressos. Os genes encontrados foram: CCNT2, DGKQ, GPAM e P2RX7. Esse 4 GDEs foram utilizados para construir uma rede PPI no STRING, utilizando os mesmos parâmetros das outras redes.

Essa rede praticamente não apresentou variação da expressão entre os tempos, o que já era esperado pois apenas 4 GDEs foram utilizados, a maioria dos genes anotados nessa rede, portanto, não tem expressão diferencial segundo os experimentos realizados no TG-GATEs. Os 4 GDEs possuem funções variadas, em especial o CCNT2 é um gene que codifica a proteína “ciclina T2” reguladora de ciclina-quinases e portanto do ciclo celular, processo muito afetado pelo etopósido (MAANEN et al., 1988). Dois outros (DGKQ e GPAM) genes estão associados o metabolismo e síntese de lipídios e o último codifica um receptor purinérgico para ATP responsável pela lise de macrófagos (WEIZMANN, 2018).

4.2.4 Lomustina

Abaixo segue as análises dos resultados para a lomustina no método EFDEGs, seguindo a mesma abordagem utilizada na ciclofosfamida e no etopósido. As figuras 4.22 e 4.23 demonstram a distribuição de GDEs encontrados para a lomustina.

Para facilitar a visualização das tendências/padrões nos resultados da Figura 4.22 foi traçado um gráfico de linhas na Figura 4.23.

Lomustina: Distribuição de GDEs para o método EFDEGs

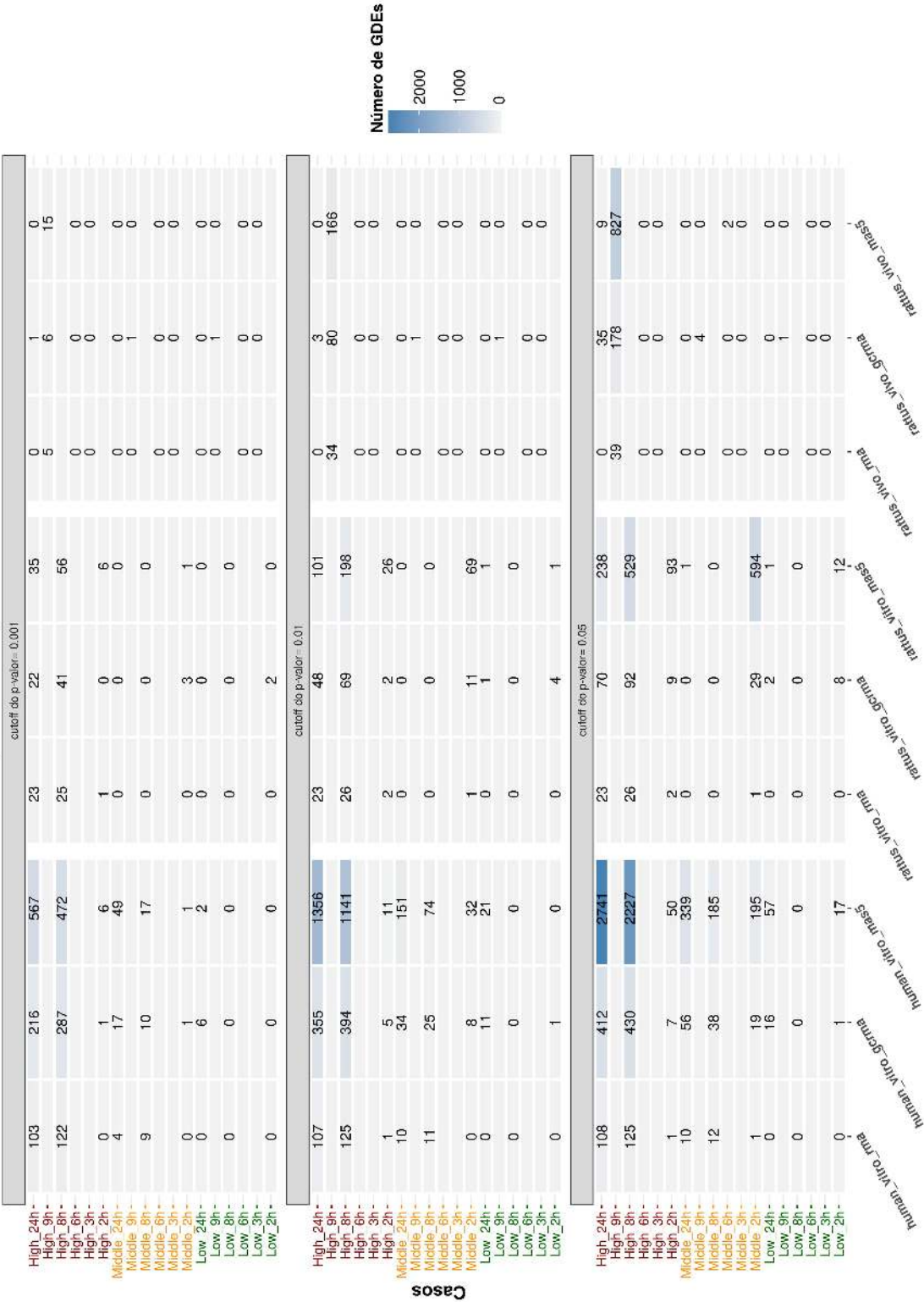


Figura 4.22: GDEs encontrados para a lomustina no método EFDEGs. Distribuição geral de GDEs encontrados no método EFDEGs para a lomustina considerando 3 diferentes valores de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normatização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que as células vazias indicam a inexistência do experimento em tais condições. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

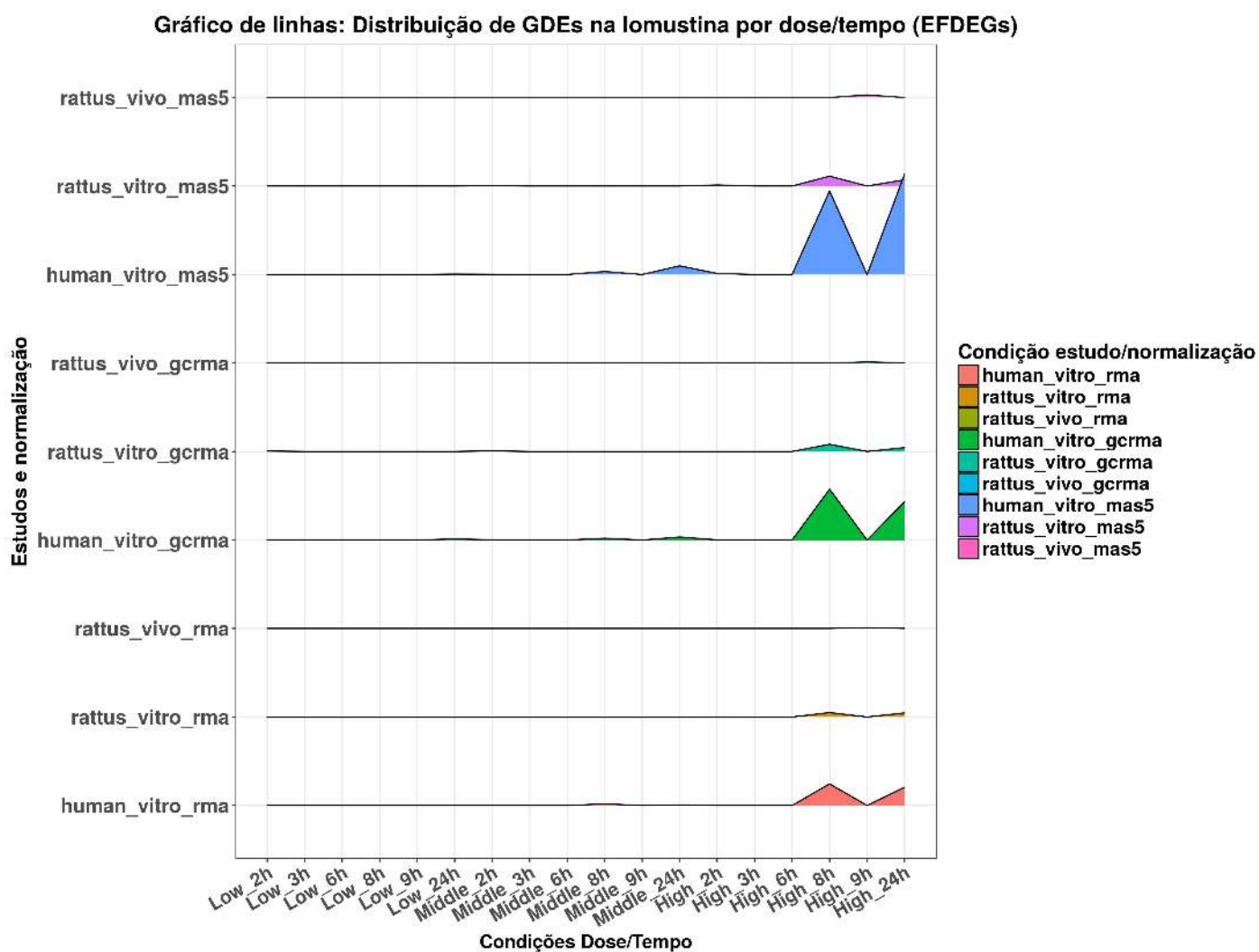


Figura 4.23: Gráfico de linhas para os GDEs da ciclofosfamida no método EFDEGs com cutoff de 0.001.

Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.22, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de GDEs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.22 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

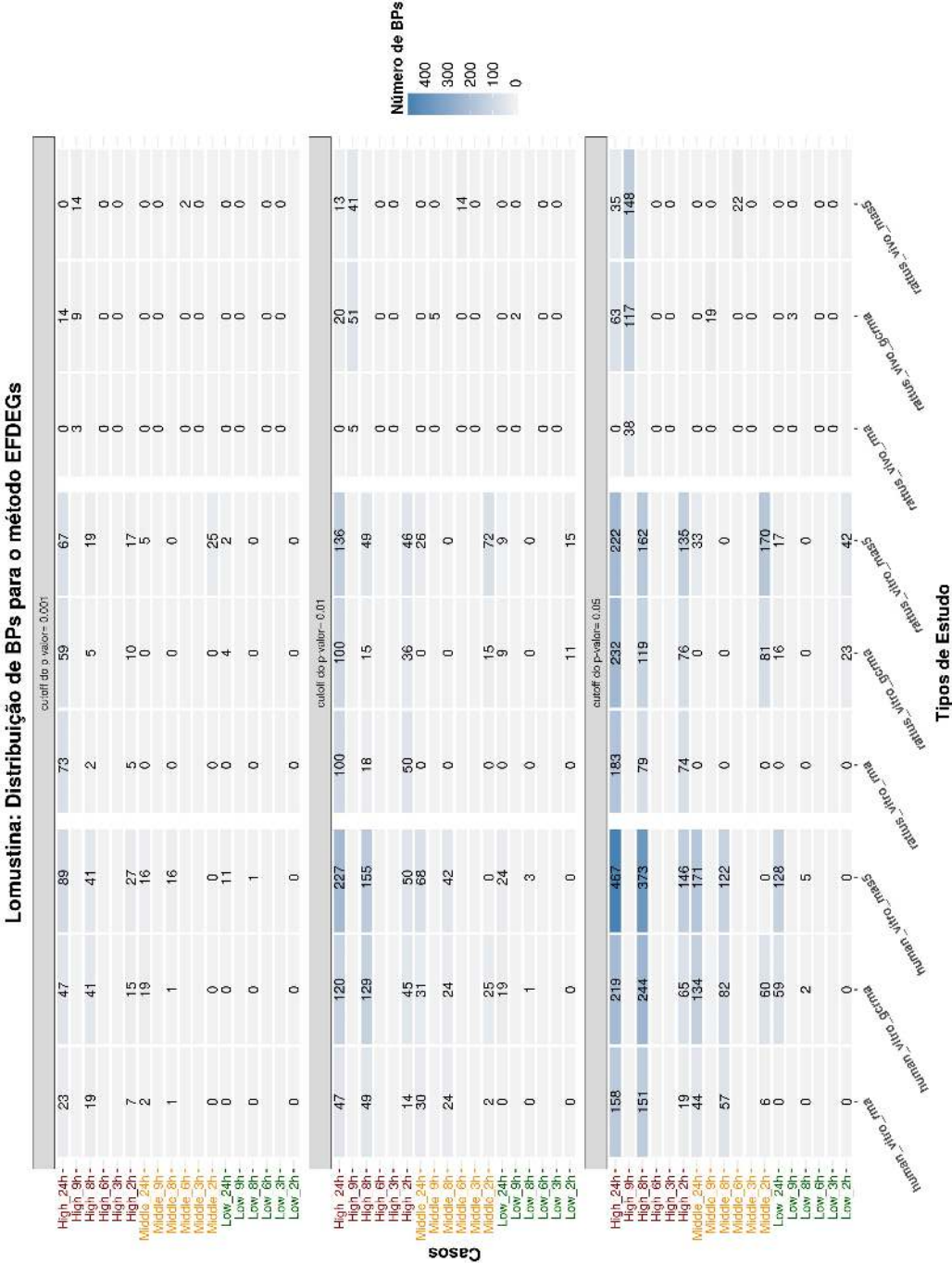


Figura 4.24: BPs encontrados para a lomustina no método EFDEGs. Distribuição geral de BPs encontrados no método EFDEGs para a lomustina considerando 3 valores diferentes de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que as células vazias indicam a inexistência do experimento em tais condições. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de BPs encontrados nas diferentes situações. As significâncias das ontologias são determinadas com uso do pacote “topGO” da plataforma *bioconductor*, utilizando uma lista de GDEs como entrada, um algoritmo de correção gráfica para determinação de pesos para as ontologias chamado “*weight01*” e por fim o teste exato de Fisher para determinação das significâncias.

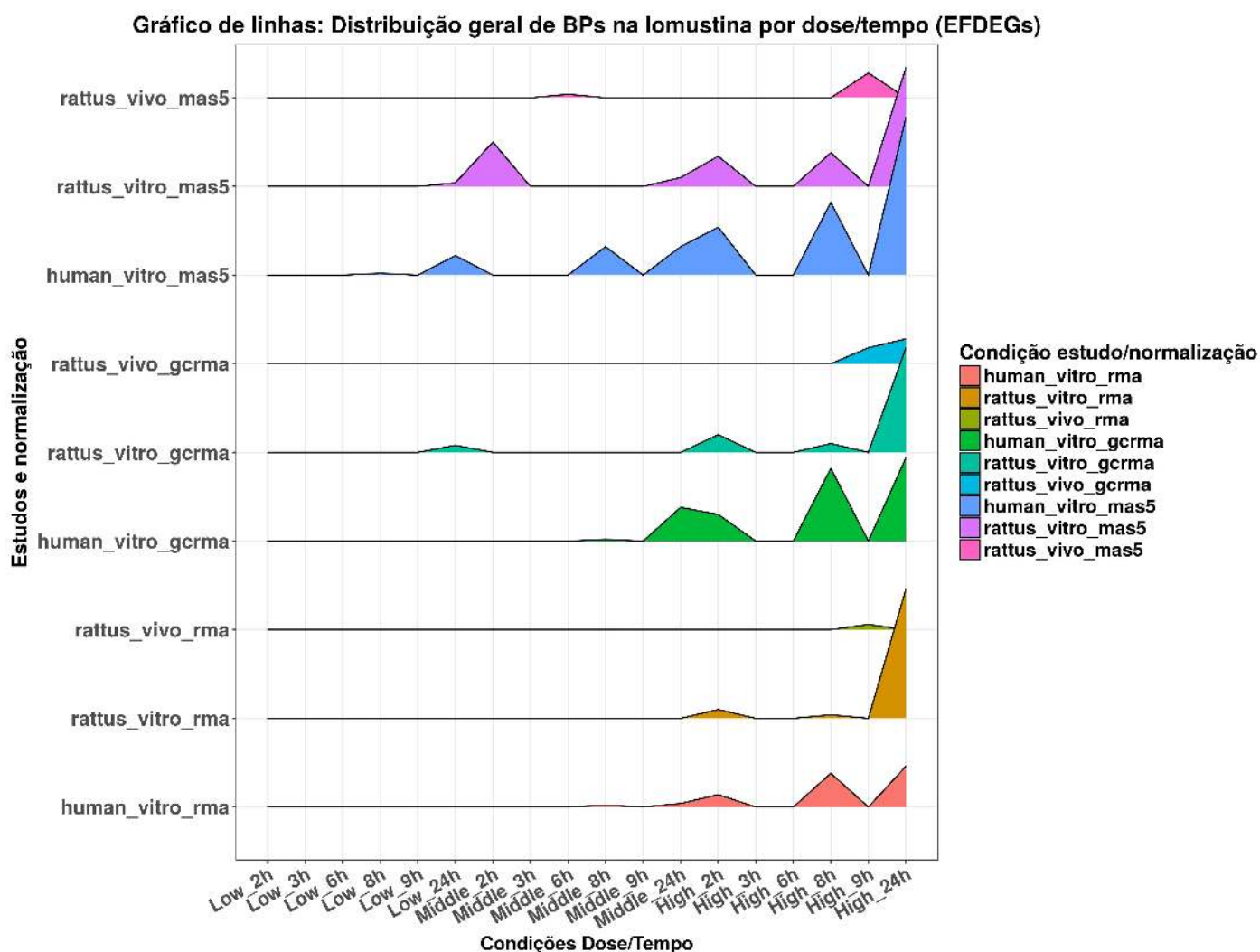


Figura 4.25: Gráfico de linhas para os BPs da lomustina no método EFDEGs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.24, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.25 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

É visível que o comportamento dos resultados das ontologias BP acompanha os resultados dos GDEs, inclusive fica ainda mais fácil observar, nas figuras 4.24 e 4.25, que:

- Os picos de 9 h nos estudos *in vivo* estão mais ativos que os de 24 h ao contrário do que acontece nos estudos *in vitro* onde os números de BPs significativas a 24 h na dose “High” tende a ficar próxima aos do caso High-8h.
- Os estudos *in vitro* apresentam alguns picos nas doses “Low” e “Middle” que quase não aparecem nos estudos *in vivo*.

A lomustina teve, em geral, o menor número de resultados entre as 3 drogas no método EFDEGs. O estudo do *Homo sapiens in vitro* encontrou entre todas as doses e tempos 238 GDEs e 53 ontologias. O estudo do *Rattus norvegicus in vitro* encontrou 49 GDEs e 33 ontologias. Por fim, o estudo do *Rattus norvegicus in vivo* encontrou apenas 3 GDEs em todos os tempos e doses e 3 BPs.

Para a lomustina não foram encontrados ontologia ou GDEs que estivessem simultaneamente presentes nos 3 estudos. No estudo do *Homo sapiens in vitro* foram encontrados ontologias e GDEs em comum apenas na comparação entre doses “High” e “Middle”, 3 e 8 respectivamente. As ontologias encontradas foram: i) GO:0031662—regulação positiva da transição G2 / M do ciclo celular mitótico; ii) GO:0046680—resposta ao DDT; iii) GO:2000775—fosforilação de histona H3-S10 envolvida na condensação cromossômica.

Os GDEs encontrados foram: CCNB1, DEPDC1, CDC20, WASL, MALAT1, ZBED6, NIPBL e BOD1L1.

As ontologias estão associadas com regulação do ciclo celular mitótico e da condensação cromossômica, a lomustina é um agente alquilante (especificamente uma nitrosuréia), dessa maneira espera-se que a droga afete o ciclo celular mitótico. Os GDEs encontrados também evidenciam isso: CCNB1, DEPDC1 e CDC20 são parte fundamental da regulação do ciclo celular mitótico e se encontram subexpressos no estudo *Homo sapiens in vitro* para a lomustina (WEIZMANN, 2018). Os outros 5 se encontram superexpressos, NIPBL, BOD1L1 e WASL estão associados com o citoesqueleto ou cromossomo. O MALAT1 é transcrito em um lncRNA que, segundo estudos recentes, pode atuar como regulador transcricional de numerosos genes, incluindo alguns genes envolvidos na metástase do câncer e migração celular, e está envolvido na regulação do ciclo celular. Sua regulação positiva em múltiplos tecidos cancerosos tem sido associada à proliferação e metástase de células tumorais. No caso do estudo do *Homo sapiens in vitro* do TG-GATEs é importante salientar que as células utilizadas eram hepatócitos, células com notável capacidade de se proliferar mediante danos de diversas origens. A observação do MALAT1 como um GDE superexpresso pode ser uma resposta da cultura celular a danos causados pela nitrosureia lomustina (WEIZMANN, 2018).

Foi construída uma rede para o STRING, utilizando os mesmos parâmetros, a partir dos 8 GDEs (descritos no parágrafo acima) encontrados no universo de BPs em comum as doses “High” e “Middle” para o estudo do *Homo sapiens in vitro*.

Muitas regiões que apresentam tom amarelado podem ser observadas nas 3 projeções em em 4.26, indicando uma superexpressão dos elementos em questão. Podemos observar na rede, entre os pontos mais amarelados, diversos genes associados a ciclinas e seus reguladores (CCNB1, CCNB2, CCNA2, genes CDC, genes CDK e genes CKS), controle da mitose (MAD1L1, BUB1), resposta a danos no DNA (PTTG1), complexo promotor da anafase (genes ANAPC), sinais de de superfície celular e regulação do citoesqueleto (FNBP1L, ARPC2) e regulação negativa do crescimento e divisão celular (genes PPP2R) (WEIZMANN, 2018). O quadro geral da rede implica em um tecido cujas células tentam se dividir mas não o conseguem, provavelmente devido aos danos causados pela lomustina.

O estudo do *Rattus norvegicus in vitro* e *in vivo* não retornaram nenhuma ontologia ou GDE em comum em nenhuma comparação. Estes estudos terão seus GDEs/BPs avaliados em cada caso (considerando dose e tempo) separadamente.

No estudo do *Rattus norvegicus in vitro*, os 49 GDEs encontrados possuem funções bem mistas. Os GDEs superexpressos incluem funções como: catabolismo de xenobióticos mediados por GST (GSTA3) e metabolismo de aldeídos e cetonas (AKR1B1). Entre os subexpressos encontram-se funções como: regulação do crescimento, diferenciação e apoptose celular (IL1B, RARB) e organização do citoesqueleto (ARHGAP26, ACTA1, RTN4RL1, TAGLN) (WEIZMANN, 2018).

Para o estudo do *Rattus norvegicus in vitro* foi contruída uma rede no STRING utilizando os 49 encontrados em todos os tempos e doses, uma vez que não fora encontrado GDEs em comum.

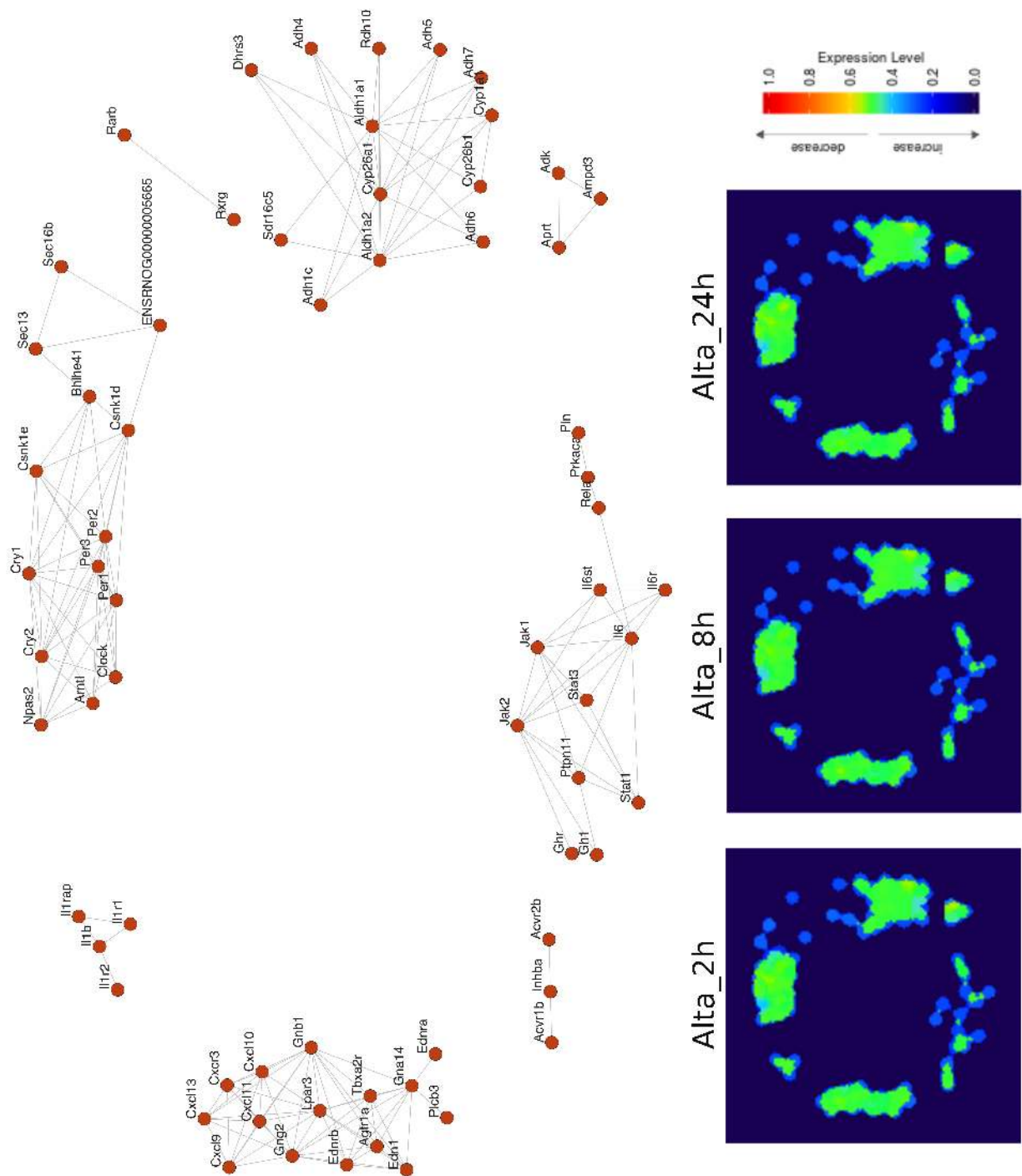


Figura 4.27: Rede PPI derivada dos GDEs em comum aos casos não nulos do estudo *Rattus norvegicus in vitro* na lomustina com uso do STRING. Ao lado 3 projeções de dados de expressão gênica sobre a mesma rede, cada projeção foi feita com dados de expressão referentes a tempos diferentes: 2 h, 8 h e 24 h.

Como pode-se observar na figura 4.27, existem diversas redes isoladas, refletindo a natureza mista dos 49 GDEs utilizados para montá-la. Pode-se notar a presença de algumas áreas levemente amareladas ou azuladas, significando superexpressão e subexpressão, respectivamente. Algumas redes tem funções bem aparentes, por exemplo a rede a esquerda tem relação com metabolismo, genes associados com o metabolismo de aldeídos e cetonas se encontram superexpressos enquanto os que estão relacionados com o citocromo P450 se encontram subexpressos. A rede superior tem genes responsáveis pela regulação de diversas funções: diferenciação celular e ritmo circadiano. Reguladores negativos dos genes CLOCK estão superexpressos enquanto o próprio gene CLOCK e os que são positivamente regulados por esse se mostram subexpressos (CRY2) (WEIZMANN, 2018).

O estudo do *Rattus norvegicus in vivo* encontrou apenas 3 GDEs entre todos os tempos e doses, são eles: Rbm39, Pabpn1 e Fam126b (todos superexpressos). As ontologias encontradas foram: i) GO:0032728—positive regulation of interferon-beta production; ii) GO:0032755—positive regulation of interleukin-6 production; iii) GO:0032760—positive regulation of tumor necrosis factor production.

Todos os 3 genes estão anotados nas 3 ontologias mencionadas e se encontram subexpressos no estudo *Rattus norvegicus in vivo* para a lomustina. Estes foram utilizados para a construção de uma rede no STRING.

A rede praticamente não apresenta mudanças entre as projeções apresentadas, provavelmente este comportamento é fruto do baixíssimo número de GDEs encontrados. Se existem poucos GDEs espera-se naturalmente, que existam poucas diferenças entre as projeções já que a maioria dos genes que as compõe não são GDEs.

4.3 Método AGFAGs

O método AGFAGs (*Analysis of Groups of Functionally Associated Genes*) é bem diferente do EFDEGs, uma vez que este método não utiliza GDEs como dados de entrada. O AGFAGs agrupa sondas/genes de um experimento em grupos de genes funcionalmente associados, sendo essa associação determinada de diversas maneiras como por exemplo ontologias, KEGG e Reactome. Em seguida são definidas para o grupo sua atividade e diversidade, sendo que a atividade é a soma das expressões dos genes contidos nos grupos e a diversidade é calculada segundo a fórmula da entropia de Shannon, isso é feito para os valores de expressão do caso e do controle. Em seguida as atividades e diversidades dos controles e experimentos são utilizadas para determinar a atividade e diversidade relativa (entre caso e controle), posteriormente um método estatístico de *bootstrap* com reposição é utilizado para determinação da significância da atividade relativa e da diversidade relativa. Assim o método AGFAGs permite a avaliação de grupos funcionais sem a pré-determinação de genes diferencialmente expressos.

Como se trata de um método com uma abordagem bem diferente do EFDEGs, espera-se que os seus resultados sejam diferentes dos resultados do EFDEGs. Por essa razão uma análise numérica descritiva dos resultados gerais será realizada para determinar como os resultados do método AGFAGs se comportam.

4.3.1 Aspectos Gerais: Análise Numérica e Descritiva

A Figura 4.29 abaixo ilustra o número de ontologias BP consideradas significantes segundo a variação da atividade entre caso e controle, a Figura 4.30 em seguida é um gráfico de linhas das “colunas” da tabela com *cutoff* de 0.05 feita com a intenção de facilitar a visualização e caracterização de comportamentos dos resultados.

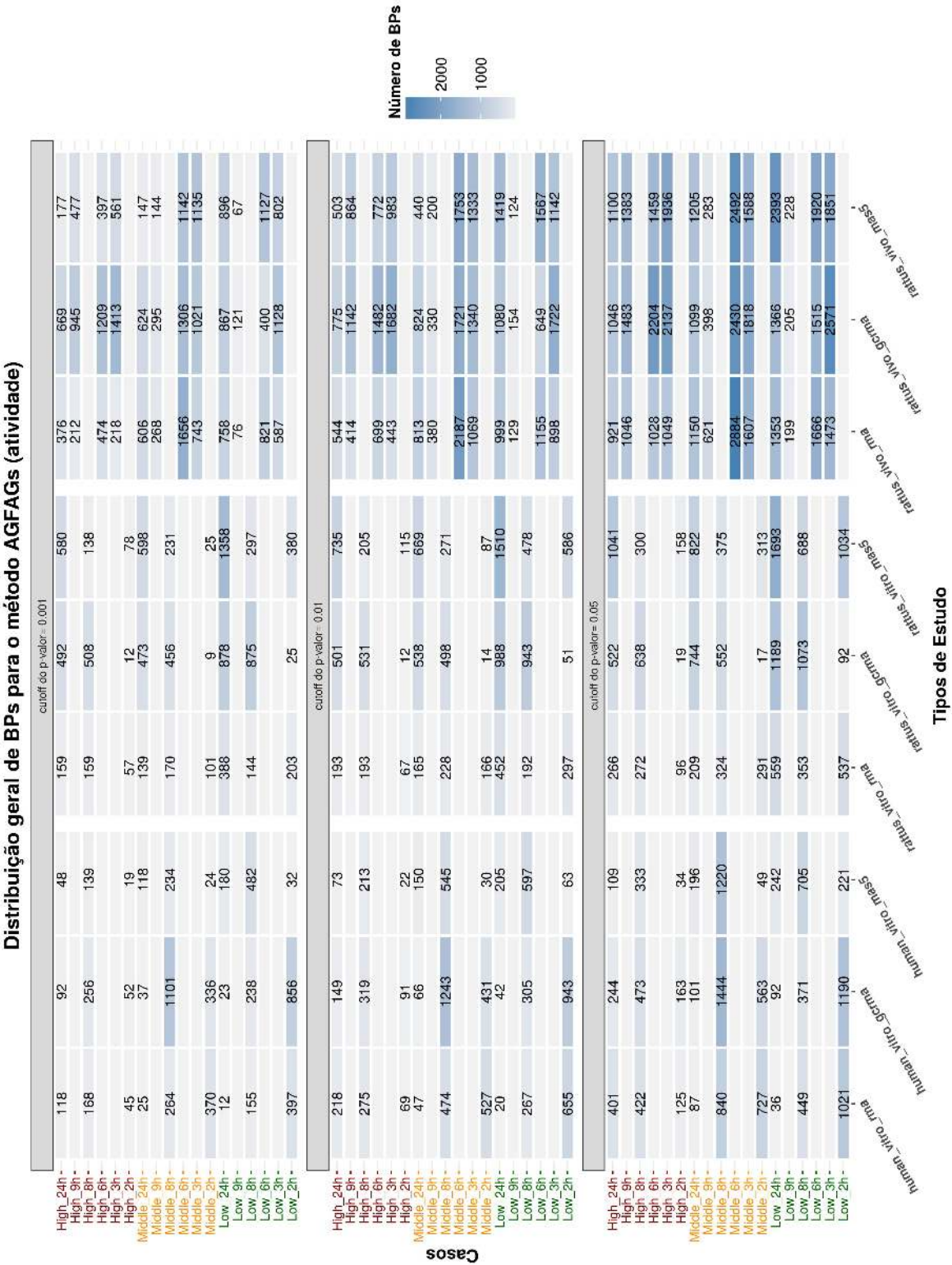


Figura 4.29: BPs encontrados para as 3 drogas antineoplásicas. Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff do p-valor (0,05, 0,01 e 0,001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

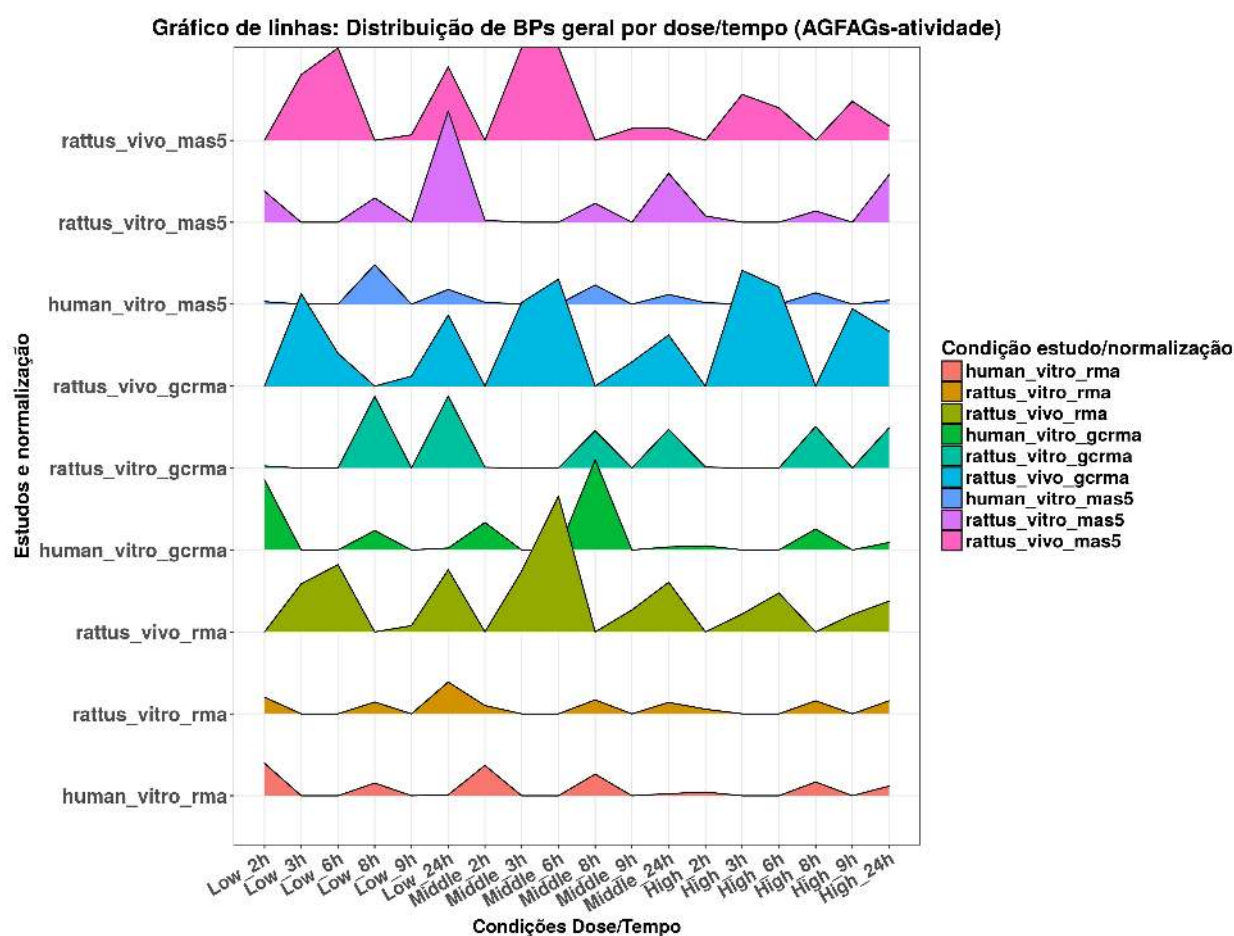


Figura 4.30: Gráfico de linhas para os BPs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.29, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significativos segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.29 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Pode-se notar que o método AGFAGs tem um comportamento muito diferente do EFDEGs. A começar pelo fato de que na figura 4.30 fica evidente que o estudo que mais trouxe resultados são o do *Rattus norvegicus in vivo*, no método EFDEGs o estudo *in vivo* geralmente apresentava poucos resultados. Nota-se, segundo a figura 4.29, que independentemente do método de normalização os estudos *Rattus norvegicus in vitro* foram os que retornaram o menor número de resultados, em seguida temos os estudos do *Homo sapiens in vitro* e por fim, como foi observado, temos os estudos *in vivo* com o maior número de resultados. O comportamento dos resultados segundo as normalizações está bem irregular, ao contrário do que acontece nas figuras 4.2 e 4.3 para o método EFDEGs. Não se encontra no método AGFAGs, os “ciclos de ativação” que encontrávamos no método EF-

DEGs, o único comportamento aparente aparece quando compara-se a figura 4.30 com sua respectiva figura sobre os resultados no método EFDEGs, a figura 4.3. Primeiramente, o comportamento mais notável entre essas duas figuras é que estudos que apresentam poucos ou nenhum GDEs encontrados no método EFDEGs (4.3, lembrando que o método AGFAGs não trabalha com GDEs apenas com ontologias, apresentam muitas ontologias consideradas como significantes pelo método AGFAGs. Observe, por exemplo, os estudos do *Rattus norvegicus in vivo* quase não apresentam resultados no método EFDEGs entretanto apresentam os maiores resultados no método EFDEGS. Isso pode, talvez, ser explicado pela maneira como o método AGFAGs opera, utilizando testes estatísticos para determinar variações significativas nos valores de atividade e/ou diversidade de cada cada ontologia entre seu caso e controle. Com a informação fornecida pela método EFDEGs, que os estudos *in vivo* praticamente não retornam GDEs, pode-se assumir que existem poucas diferenças extremas ou aparentes entre o caso e o controle. Se este é o caso, é natural esperar que a estatística do *bootstrap*, se tratando de uma reamostragem com reposição, considere que quase qualquer diferença entre caso e controle, por mínima que seja, fosse considerada como significativa. Portanto observarmos esse comportamento no estudo *in vivo* na figura 4.30, de maneira análoga podemos esperar que estudos que apresentaram uma grande quantidade de GDEs no método de EFDEGs, como por exemplo os estudos *in vitro* (especialmente na dose “High” e tempo 8 h e 24 h), tenham um menor número de BPs. De fato pode-se observar como os picos vão ficando menores conforme a dose e o tempo aumentam nos estudo *in vitro*, chegando aos picos mínimos na dose “High”. Portanto no método AGFAGs, uma maior quantidade de GDEs, limita o número de ontologias considerados significativas segundo sua atividade.

Distribuição geral de BPs para o método AGFAGs (diversidade)

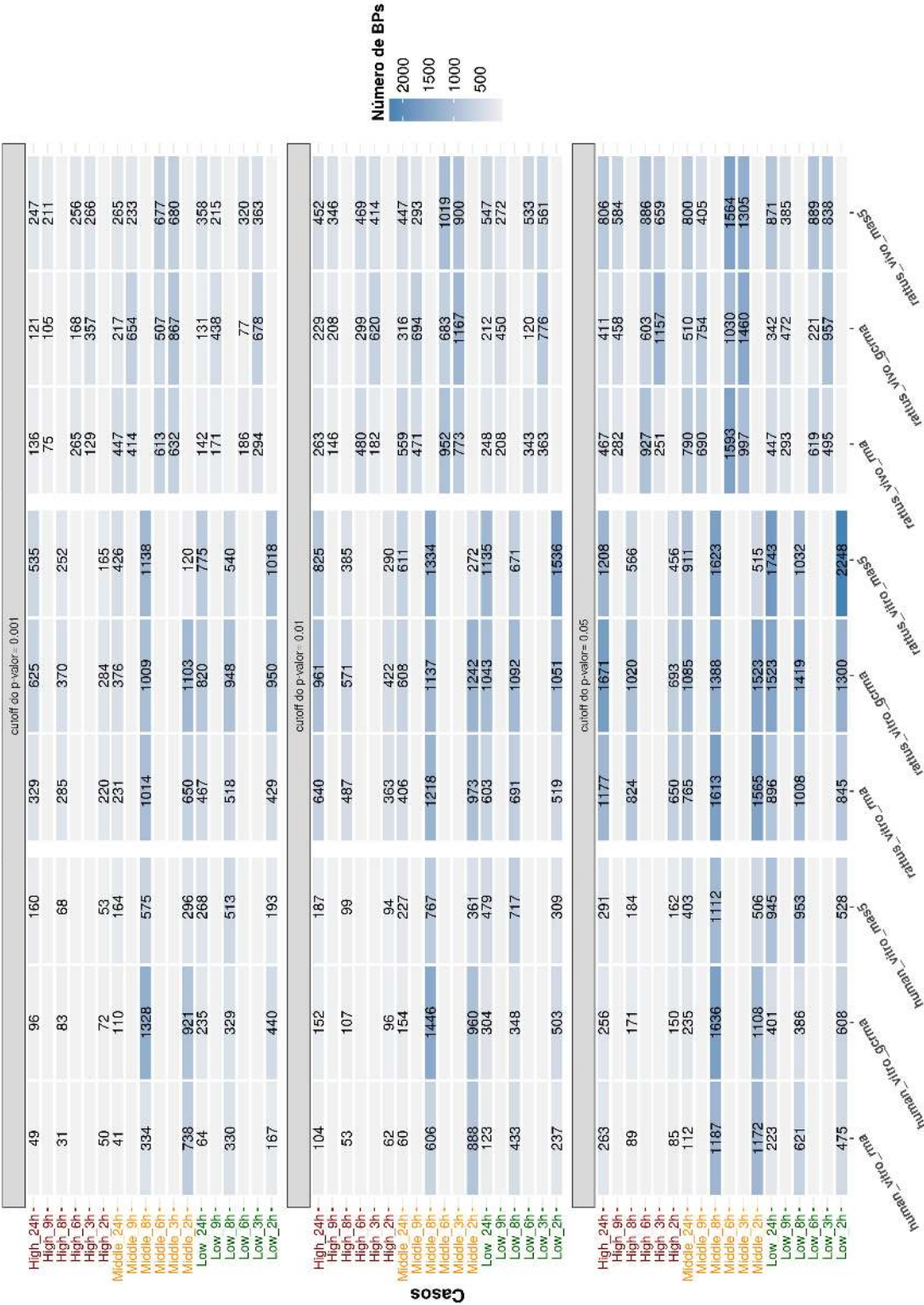


Figura 4.31: BPs encontrados para as 3 drogas antineoplásicas. Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff do p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

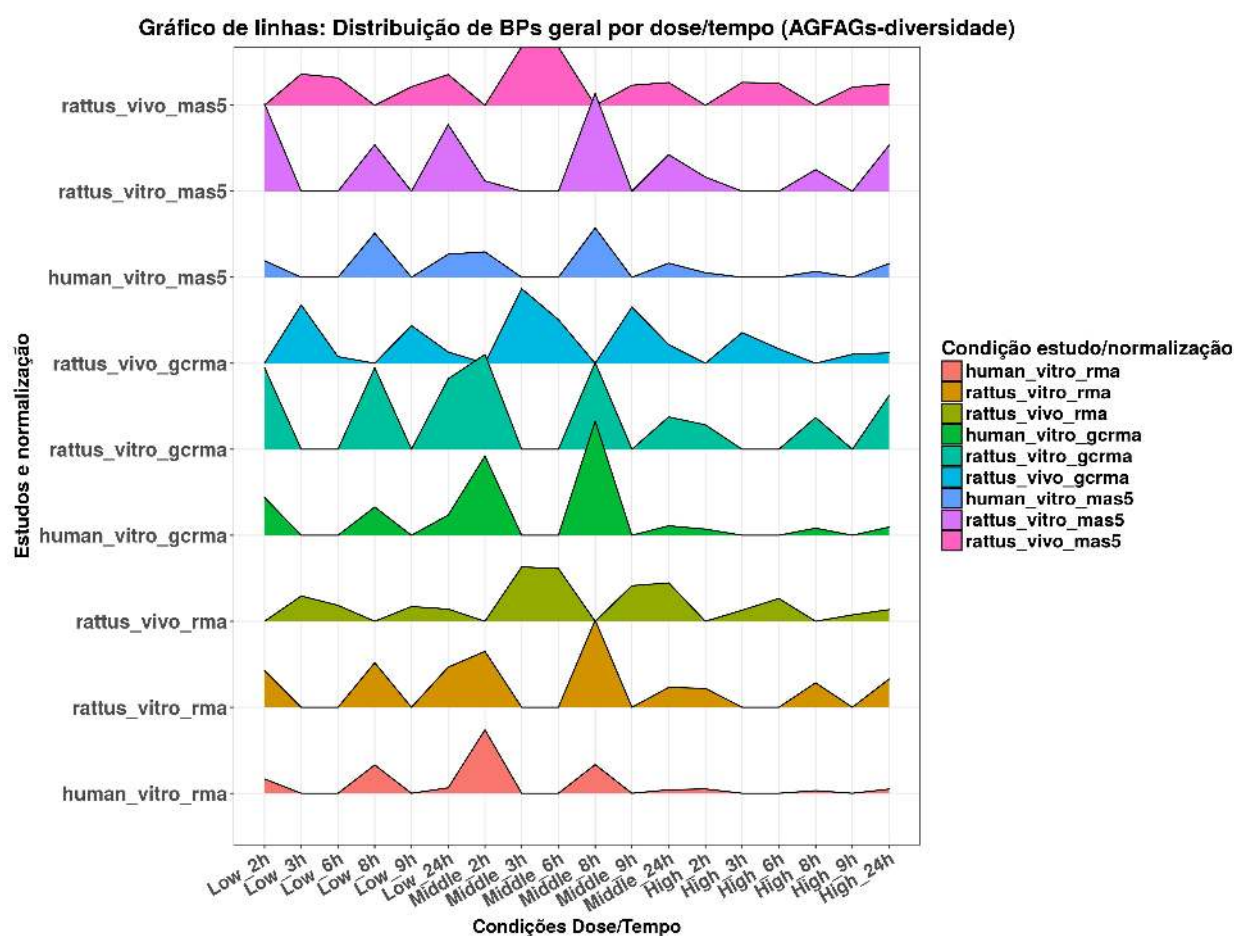


Figura 4.32: Gráfico de linhas para os BPs com cutoff de 0.001. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.31, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.31 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

A primeira vista o comportamento mais notável nos picos da figura 4.32, vão diminuindo conforme a dose e o tempo aumentam, refletindo um comportamento parecido com o da figura 4.30, ou seja, a presença de uma maior quantidade de GDEs limita a quantidade de grupos considerados significantes quanto sua diversidade.

A partir de agora será realizada a análise das drogas separadamente, esta parte é similar à realizada no método EFEDGs, com uma descrição numérica e gráfica dos resultados seguida da análise biológica.

4.3.2 Ciclofosfamida

Na tabela ilustrada na figura 4.33, os estudos do *Homo sapiens* demonstram um comportamento que foi discutido anteriormente nos aspectos gerais: em geral quanto menor o tempo e dose (o que geralmente implica em um menor número de GDEs encontrados no método EFDEGs), maior o número de BPs encontrado no ponto de vista da atividade. Esse comportamento também se repete para os estudos do *Rattus norvegicus in vitro*, no método EFDEGs os estudos do *Rattus norvegicus in vitro* apresentaram o maior número de GDEs, a distribuição de GDEs desses estudos no método EFDEGs é aproximadamente a mesma dos estudos do *Homo sapiens*, isso explica o porque a distribuição de BPs no método AGFAGs também é parecida nesses dois estudos. O estudo do *Rattus norvegicus in vivo* trouxe o maior número de BPs entre os 3 estudos no método AGFAGs, e em contrapartida retornou o menor número de GDEs no método EFDEGs, mantendo a relação que também foi encontrada nos outros dois estudos na ciclofosfamida e na análise geral: quanto menor o numero de GDEs encontrados em estudo no EFDEGs, maior é o número de BPs encontrados para este mesmo estudo no método AGFAGs.

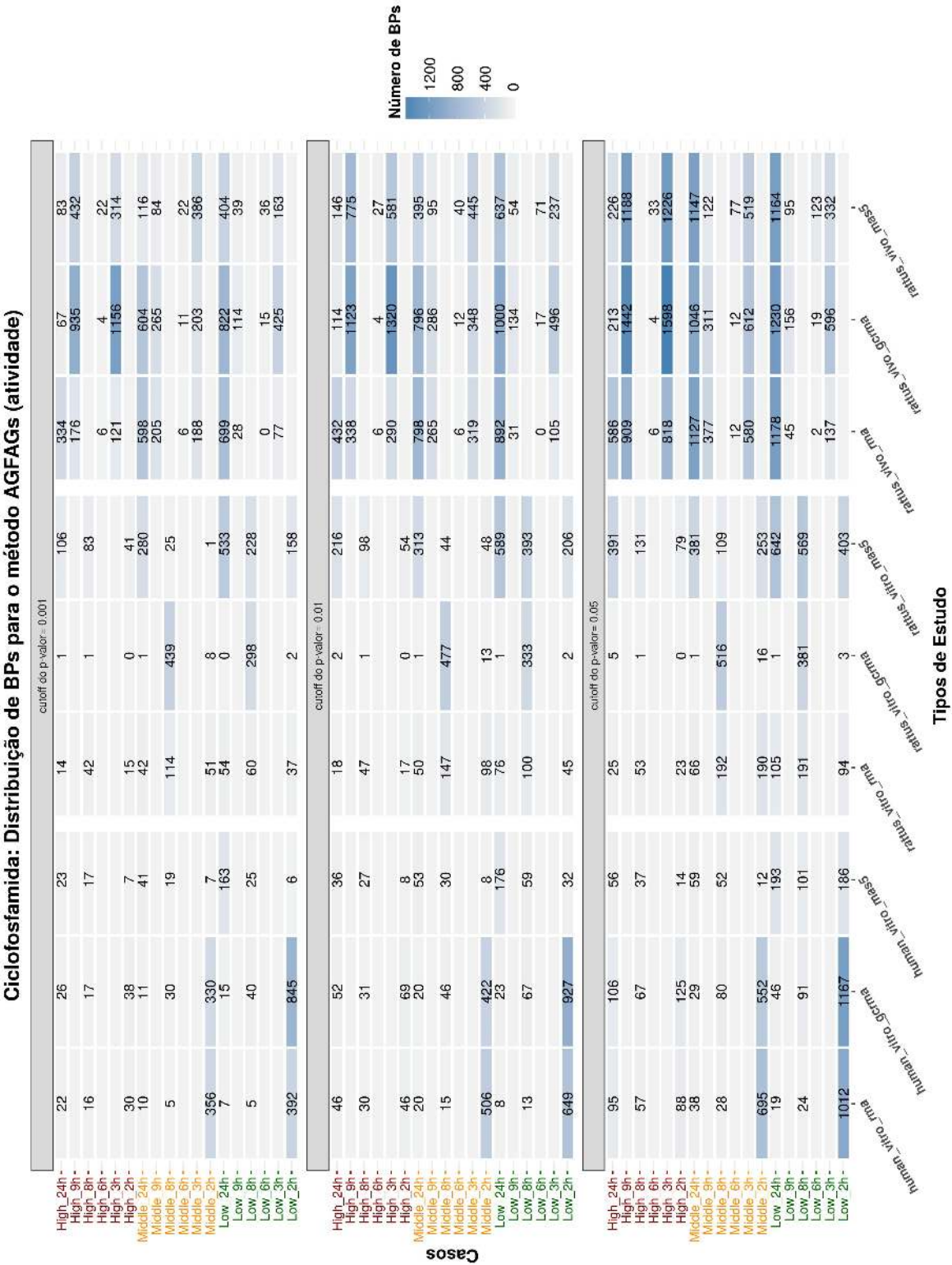


Figura 4.33: BPs encontrados para a ciclofosfamida no AGFAGs (atividade). Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0,05, 0,01 e 0,001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

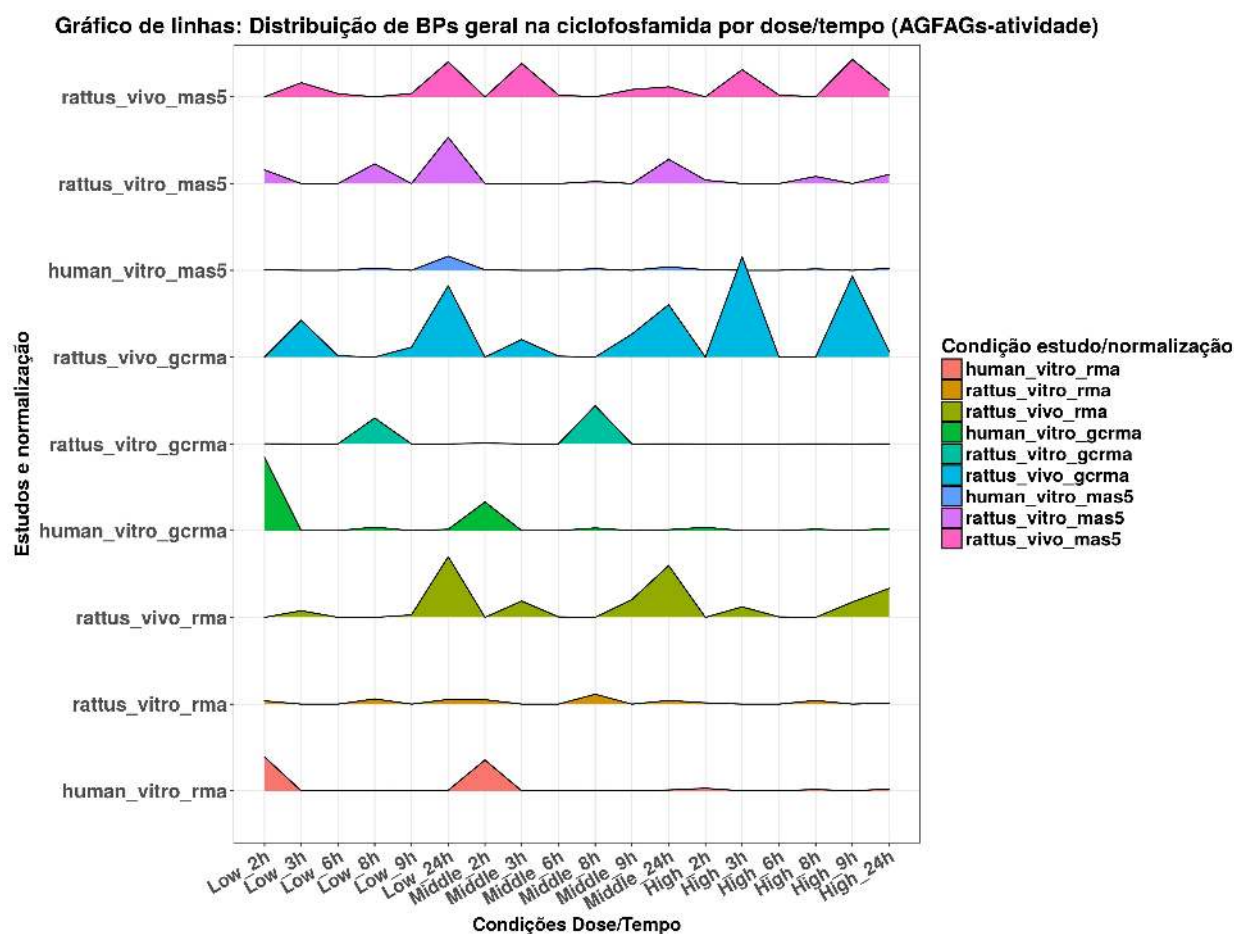


Figura 4.34: Gráfico de linhas para os BPs com cutoff de 0.001 para ciclofosfamida no AGFAGs (atividade). Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.33, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.33 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

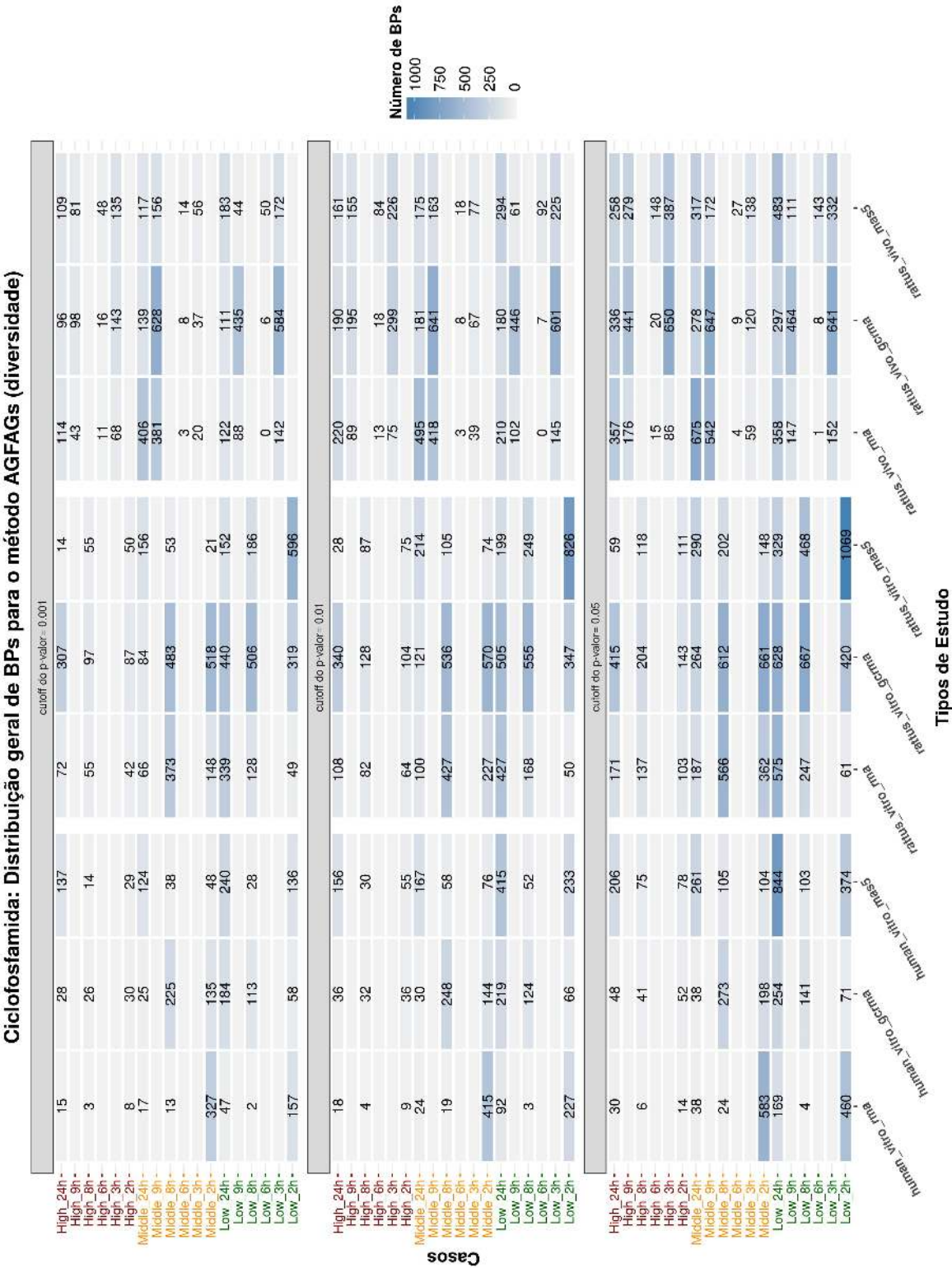


Figura 4.35: BPs encontrados para a ciclofosfamida no AGFAGs (diversidade). Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas , foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

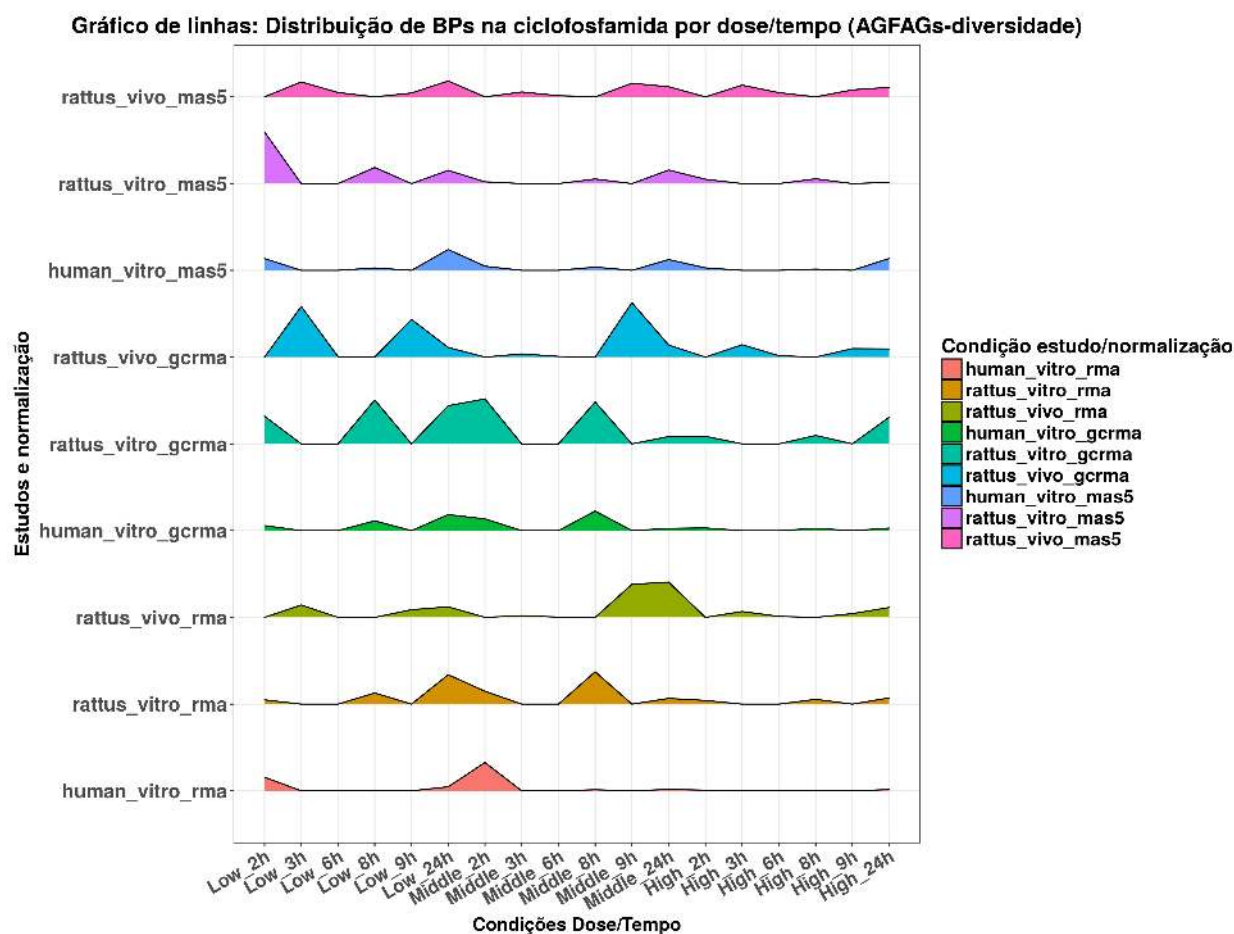


Figura 4.36: Gráfico de linhas para os BPs com cutoff de 0.001 para ciclofosfamida no AGFAGs (diversidade). Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.35, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.35 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

A distribuição de resultados encontrados sob a perspectiva da diversidade na ciclofosfamida, para cada um dos 3 estudos, não é muito diferente da distribuição de resultados encontrados sob a perspectiva da atividade. Pode se observar, por exemplo, que o estudo *Homo sapiens in vitro* em qualquer valor de *cutoff*, geralmente apresenta maior número de BPs em doses e tempo menores, o mesmo comportamento é encontrado nos outros dois estudos.

Portanto a distribuição de resultados do método AGFAGs é parecida entre os resultados de ambas as perspectivas: atividade e diversidade. Entretanto resta saber se os resultados de cada perspectiva tem, em geral, o mesmo significado biológico, processos biológicos conver-

gentes entre essas duas perspectivas podem ter grande importância para os efeitos causados pela droga.

Segundo a atividade, entre todas as doses e tempos, foram encontrados para o estudo do *Homo sapiens in vitro* 2835 BPs. Para o estudo do *Rattus norvegicus in vitro* foram encontrados 3469 BPs e por fim, 9575 BPs para o *Rattus norvegicus in vivo*.

Devido a ao número excessivo de resultados, foram necessárias filtrações adicionais para realizar uma análise biológica. Para elencar ontologias de maior interesse para posterior análise, serão consideradas apenas as ontologias que estavam simultaneamente presentes em todas as doses (desconsiderando o tempo) dentro de um estudo (“low”, “Middle” e “High”). A atividade de uma ontologia é calculada como a soma das expressões dos genes anotadas dentro da ontologia, naturalmente isso cria duas atividades: a do controle e do caso. A atividade relativa é nada mais que o valor da atividade do caso dividida pela soma dos valores do caso e do controle. Assim sendo uma atividade relativa maior que 0.5 indica uma possível superexpressão dos genes anotados no caso em relação ao controle, e uma atividade relativa menor que 0.5 indica uma possível subexpressão desses genes. Portanto ao filtrarmos e analisarmos ontologias de interesse da perspectiva da atividade, estamos considerando ontologias que apresentem uma alteração direta na expressão dos genes constituintes.

As ontologias que apresentaram atividade relativa maior que 0.5, indicando portanto uma superexpressão de genes anotados em si, são: i) GO:0006196–Processo catabólico de AMP; ii) GO:0014740–regulação negativa da hiperplasia muscular; iii) GO:0031937–regulação positiva do silenciamento da cromatina; iv) GO:0032764–regulação negativa da produção de citocinas de mastócitos; v) GO:0034760–regulação negativa do transporte transmembranar de íons de ferro; vi) GO:0036304–morfogênese do cordão umbilical; vii) GO:0046985–regulação positiva do processo biossintético da hemoglobina; viii) GO:0090038–regulação negativa da sinalização da proteína quinase C; ix) GO:0090669–estabilização do RNA da telomerase; x) GO:1901203–regulação positiva da montagem da matriz extracelular; xi) GO:1902310–regulação positiva da desfosforilação da peptidil-serina; xii) GO:1903413–resposta celular ao ácido biliar; xiii) GO:1904816–regulação positiva da localização da proteína no cromossomo, região telomérica; xiv) GO:1990641–resposta à inanição de íons de ferro; xv) GO:2000342–regulação negativa da produção do ligante 2 de quimiocina (motif C-X-C); xvi) GO:0000738–Processo catabólico de DNA, exonucleolítico.

No estudo do *Homo sapiens in vitro* pode-se observar poucas ontologias reguladas ne-

gativamente, todas elas na dose baixa, algumas dessa ontologias inclusive aparecem com reguladas positivamente nas doses maiores. Alguns BPs observados remete-nos diretamente ao que se espera dos efeitos que uma mostarda nitrogenada como ciclofosfamida pode ter nas células, como o silenciamento da cromatina e processo catabólica de DNA, ambos diretamente associados com crescimento e divisão celular, um dos mecanismos mais afetados negativamente pela droga (HUMANS et al., 2012). Outros BPs interessantes incluem os relacionados ao ion de ferro e disfunção muscular, pois a fadiga e anemia são sintomas conhecidos da droga (HUMANS et al., 2012). Observa-se também a regulação negativa componentes importantes da sinalização celular como quimiocinas e quinases. Em geral, as ontologias elencadas como interessantes no método AGFAGs, segundo o protocolo descrito, são diferentes das que são encontradas no método EFDEGs. No método EFDEGs, o estudo *Homo sapiens in vitro* para a ciclofosfamida encontrou principalmene ontologias e GDEs e ontologias que estavam principalmente relacionados ao metabolismo de xenobióticos pelo sistema citocromo P450. No método AGFAGs as ontologias elencadas como interessantes segundo a filtragem realizada estão relacionadas com diversos aspectos de um organismo, embora nenhum deles esteja diretamente relacionado com o metabolismo de xenobióticos. É interessante observar que, as conclusões finais de cada análise biológica tem respaldo na literatura, apesar de que as ontologias elencadas como interessantes a partir dos resultados de cada método tenham sido diferentes.

Para o estudo *Rattus norvegicus in vitro*, foi realizado o mesmo processo de filtragem de BPs utilizado para o estudo *Homo sapiens in vitro*, no fim foram elencadas 49 ontologias de interesse. Todas essas ontologias possuem atividade relativa superior a 0.5, indicando superexpressão dos genes anotados em sí. A maioria deles estão associados com processos de biosíntese e catabolismo, embora outras processos também possam ser encontrados. A seguir uma lista com 25 ontologias selecionadas das 49 observadas: i) GO:0019483–processo biossintético de beta-alanina; ii) GO:0042412–processo biossintético de taurina; iii) GO:0019442–processo catabólico do triptofano a acetil-CoA; iv) GO:0034653–processo catabólico do ácido retinóico; v) GO:0055090–homeostase de acilglicerol; vi) GO:0006212–processo catabólico da uracila; vii) GO:0097089–processo metabólico de ácidos graxos metil ramificados; viii) GO:0070859–regulação positiva do processo biossintético dos ácidos biliares; ix) GO:0045796–regulação negativa da absorção intestinal do colesterol; x) GO:0010949–regulação negativa da absorção intestinal de fitosteróis; xi) GO:0005986–processo biossintético de sacarose; xii) GO:2000482–regulação da secreção de interleucina-8; xiii) GO:0006210–

processo catabólico da timina; xiv) GO:0006642–mobilização de triglicerídeos; xv) GO:0061418–regulação da transcrição do promotor da RNA polimerase II em resposta à hipóxia; xvi) GO:0010727–regulação negativa do processo metabólico de peróxido de hidrogênio; xvii) GO:0019265–processo biossintético da glicina, por transaminação de glioxilato; xviii) GO:0034760–regulação negativa do transporte transmembranar de íons de ferro; xix) GO:0051344–regulação negativa da atividade da fosfodiesterase de nucleotídeo cíclico; xx) GO:0010818–Quimiotaxia de células T; xxi) GO:1901896–regulação positiva da atividade de ATPase transportadora de cálcio; xxii) GO:1903984–regulação positiva da via de sinalização apoptótica ativada por TRAIL; xxviii) GO:1990641–resposta à inanição de íons de ferro; xxiv) GO:0001970–regulação positiva da ativação do complexo de ataque à membrana; xxv) GO:0042977–ativação da atividade da quinase JAK2.

As ontologias selecionadas indicam fortemente a síntese de lipídios e ácidos graxos. também podemos observar ontologias associadas ao transporte de íons de ferro (regulação negativa) e a resposta do organismo a falta de tal ion, ambas também observadas no estudo do *Homo sapiens in vitro*. Outra observação interessante é como as ontologias indicam um aumento nos níveis de peróxido de hidrogênio e reação das células à hipóxia, sabe-se que a hipóxia induz ao aumento de níveis de peróxido de hidrogênio em células eucarióticas e que o aumento deste pode provocar dano por estresse oxidativo em células (NETZER et al., 2015). Isso explicaria a observação de ontologia GO:0001970 no conjunto, que indica a presença de danos a menbrana celular, um dos mais comuns causados por estresse oxidativo. Estas observações continuam tendo respaldo na literatura, visto que os metabólitos da ciclofosfamida são conhecidos pela sua capacidade de produzir espécies reativas de oxigênio (ROS) (SELVAKUMAR et al., 2005). Outra observação interessante é a ontologia GO:190398, indicando a regulação positiva de via de sinalização apoptótica, a ciclofosfamida induz a apoptose celular mediante danos ao DNA (HUMANS et al., 2012).

Os resultados do estudo *Rattus norvegicus in vitro* realizado para ciclofosfamida no método EFDEGs são bem diferentes dos obtidos para o método AGFAGs, a maioria das ontologias encontradas no método EFDEGs fazem referência ao metabolismo de xenobióticos, especialmente envolvendo aldeído-desidrogenases (ALDH) e glucoronidação (UGT), também foram encontrados genes envolvidos no controle do ciclo celular.

O estudo do *Rattus norvegicus in vivo* encontrou, seguindo o mesmo protocolo de filtragem, 46 BPs de interesse. Todas elas também se encontram superexpressas, assim

como no *Rattus norvegicus in vitro*. Alguns deles: i) GO:0090427–ativação da meiose; ii) GO:2001055–regulação positiva do processo apoptótico de células mesenquimais; iii) GO:0032764–regulação negativa da produção de citocinas de mastócitos; iv) GO:0010949–regulação negativa da absorção intestinal de fitosteróis; v) GO:0035407–fosforilação histona H3-T11; vi) GO:0045796–regulação negativa da absorção intestinal do colesterol; vii) GO:2000234–regulação positiva do processamento de rRNA; viii) GO:2000570–regulação positiva da ativação das células T-helper 2; ix) GO:0071812–regulação positiva da geração de febre por regulação positiva da secreção de prostaglandina; x) GO:0071848–regulação positiva da cascata de ERK1 e ERK2 via sinalização mediada por TNFSF11; xi) GO:0061418–regulação da transcrição do promotor da RNA polimerase II em resposta à hipóxia; xii) GO:0036112–processo metabólico de cadeia de ácido graxo-acil-CoA de cadeia média; xiii) GO:0001189–Montagem do complexo de pré-iniciação transcricional de RNA polimerase I no promotor para o transcrito nuclear de rRNA grande; xiv) GO:0036112–processo metabólico de cadeia de gordura-acil-CoA de cadeia média; xv) GO:0014740–regulação negativa da hiperplasia muscular; xvi) GO:0031283–regulação negativa da atividade da guanilato ciclase; xvii) GO:0070562–regulação da via de sinalização do receptor de vitamina D; xviii) GO:0006045–Processo biossintético da N-acetilglucosamina; xix) GO:0010751–regulação negativa de transdução de sinal mediada por óxido nítrico; xx) GO:0045645–regulação positiva da diferenciação de eosinófilos.

As ontologias encontradas para a ciclofosfamida no estudo *Rattus norvegicus in vivo* para o método AGFAGS são mistas, apresentando associação com uma variedade de processos: diferenciação celular, regulação negativa da absorção de lipídios, resposta a hipóxia, metabolismo de ácidos graxos, regulação positiva da cascata de sinalização ERK1 e ERK2, e regulação positiva de processos apoptóticos. A observação da resposta a hipóxia e da regulação positiva da apoptose estão provavelmente ligados com a atuação da droga no organismo, e também são observadas no estudo *Rattus norvegicus in vitro*. Esse mesmo estudo realizado no método EFDEGs encontrou pouquíssimos resultados, a maior parte das ontologias encontradas nesse estudo estavam relacionadas com o metabolismo e regulação do ciclo celular.

Abaixo uma descrição numérica dos resultados da ciclofosfamida da perspectiva da diversidade. Foram encontrados 1311 BPs para o estudo do *Homo sapiens in vitro*, 995 para o estudo do *Rattus norvegicus in vitro* e 3664 no estudo *Rattus norvegicus in vivo*. Devido

ao número elevado de resultados, uma filtragem foi realizada para elencar as ontologias de maior interesse em cada estudo. Para o caso da diversidade foram consideradas as ontologias que estavam simultaneamente nas 3 doses do estudo em questão, ou seja, é idêntica a que foi utilizada para o ponto de vista da atividade. A diversidade é calculada para cada ontologia segundo a fórmula da entropia normalizada de Shannon, também criando os valores de diversidade do caso e do controle, assim como na atividade. A diversidade relativa é calculada de maneira similar ao da atividade relativa, isto é, valor do caso dividido sobre a soma dos valores do caso e do controle. Uma diversidade relativa maior que 0.5 indica uma distribuição de valores de expressão mais heterogênea no caso do que no controle e o oposto caso contrário. Dessa maneira ao analisarmos ontologias de interesse do ponto de vista da diversidade, estamos considerando ontologia que apresentam genes com grande heterogeneidade na expressão de seus genes constituintes. Ou seja, se existe um valor médio de expressão dos genes anotados em uma determinada ontologia também existem nela genes que apresentam uma considerável distância desse valor (subexpressão ou superexpressão).

O estudo do *Homo sapiens in vitro* encontrou 8 ontologias de interesse do ponto de vista da diversidade seguindo a filtragem descrita acima: i) GO:0000493–montagem do box H / ACA snoRNP; ii) GO:0031393–regulação negativa do processo biossintético da prostaglandina; iii) GO:0034059–resposta à anoxia; iv) GO:0090669–estabilização do RNA da telomerase; v) GO:1904816–regulação positiva da localização da proteína no cromossomo, região telomérica; vi) GO:1990619–desacetilação da histona H3-K9; vii) GO:0060532–desenvolvimento de cartilagem de brônquios; viii) GO:0061145–desenvolvimento do músculo liso do pulmão.

Todas as ontologias listadas acima também são estatisticamente relevantes segundo sua atividade e está sempre maior que 0.5, indicando superexpressão de genes anotados nas ontologias. A diversidade relativa também sempre está maior que 0.5, indicando portanto uma maior heterogeneidade de valores de expressão no caso em relação ao controle. Esses processos são bem mistos, envolvendo desenvolvimento de músculos e cartilagem, biossíntese de prostaglandinas, resposta a anoxia, estabilização do RNA da telomerase e localização de proteínas nessa região da cromatina.

No estudo do *Rattus norvegicus in vitro* foram encontrados 34 ontologias de interesse do ponto de vista da diversidade. Algumas delas: i) GO:0001970–regulação positiva da ativação do complexo de ataque à membrana; ii) GO:1902172–regulação do processo apoptótico de queratinócitos; iii) GO:1990636–senescência reprodutiva; iv) GO:0048866–especificação

do destino das células estaminais; v) GO:0007493–determinação do destino da célula endodérmica; vi) GO:0006642–mobilização de triglicerídeos; vii) GO:0034370–Remodelamento de partículas lipoproteicas ricas em triglicéridos; viii) GO:0034445–regulação negativa da oxidação da lipoproteína plasmática; ix) GO:0072703–resposta celular ao metanossulfonato de metila; x) GO:0055090–homeostase de acilglicerol; xi) GO:0097089–processo metabólico de ácidos graxos metil ramificados; xii) GO:0006210–processo catabólico da timina; xiii) GO:0006212–processo catabólico da uracila; xiv) GO:1902033–regulação da proliferação de células-tronco hematopoiéticas; xv) GO:0072703–resposta celular ao metanossulfonato de metila; xvi) GO:0070857–regulação do processo biossintético dos ácidos biliares; xvii) GO:0046209–processo metabólico de óxido nítrico; xviii) GO:0043420–processo metabólico antranilato; xix) GO:0070857–regulação do processo biossintético dos ácidos biliares; xx) GO:0070427–domínio de oligomerização de ligação ao nucleotídeo contendo 1 via de sinalização; xxi) GO:0003308–regulação negativa da via de sinalização Wnt envolvida no desenvolvimento do coração;

Assim como no caso do *Homo sapiens in vitro* as ontologias descritas possuem sua diversidade sempre acima de 0.5 e também apresenta o mesmo comportamento em relação à atividade relativa (sempre estatisticamente significativa e maior que 0.5). Muitas das ontologias encontradas se remetem a descrição dos processos de crescimento, divisão e destino celular celular vias muito afetadas pela ciclofosfamida (HUMANS et al., 2012). Outras ontologias interessantes são as que lidam com vias de apoptose, biosíntese de triglicerídeos e lipídios e evidências de ataque a membrana celular, processos que podem estar associados aos metabólitos da ciclofosfamida. O quadro geral se assemelha um pouco ao dos resultados do ponto de vista da atividade, onde observa-se metabolismo de triglicerídeos e ácidos graxos, alteração de vias apoptóticas e evidencia de danos a membrana celular também são observadas como ontologias alteradas.

O estudo do *Rattus norvegicus in vivo* encontrou 32 ontologias de interesse, algumas listadas abaixo: i) GO:0030174–regulação da iniciação da replicação do DNA dependente do DNA; II)GO:0014908–diferenciação de miotubos envolvida na regeneração do músculo esquelético; iii) GO:0044778–check point de integridade do DNA meiótico; iv) GO:1905342–regulação positiva da localização de proteínas para kinetochore; v) GO:1904582–regulação positiva da localização do mRNA intracelular; vi) GO:1903919–regulação negativa da separação de filamentos de actina; vii) GO:1903673–formação de sulco de clivagem mitótica;

viii) GO:0097252–processo apoptótico oligodendrócito; ix) GO:0046642–regulação negativa da proliferação de células T alfa-beta; x) GO:0030264–fragmentação nuclear envolvida na mudança nuclear apoptótica; xi) GO:0000706–processamento de quebra de cadeia dupla de DNA meiótico; xii) GO:0090063–regulação positiva da nucleação de microtúbulos; xiii) GO:0032425–regulação positiva do reparo de incompatibilidade; xiv) GO:0031247–conjunto de haste de actina; xv) GO:0090063–regulação positiva da nucleação de microtúbulos; xvi) GO:0002933–hidroxilação lipídica; xvii) GO:0032764–regulação negativa da produção de citocinas de mastócitos; xviii) GO:2000296–regulação negativa do processo catabólico de peróxido de hidrogênio; xix) GO:0090218–regulação positiva da atividade da lipase quinase.

Assim como nos últimos 2 estudos envolvendo o ponto de vista da diversidade, as ontologias descritas possuem sua diversidade relativa sempre acima de 0.5 e também apresentam o mesmo comportamento em relação à atividade relativa (sempre estatisticamente significativa e maior que 0.5). É interessante notar como no estudo *Rattus norvegicus in vitro* sob o ponto de vista da diversidade existem várias ontologias referentes a regulação do crescimento e divisão celular, a regulação do citoesqueleto, ao reparo de danos ao DNA e a alteração de vias apoptóticas. Estudos mostram que esses são processos muito afetados pela ciclofosfamida (HUMANS et al., 2012). O mesmo estudo no método AGFAGs, mas sob o ponto de vista da atividade, também fornece essas observações, os resultados portanto convergem quanto a esse tipo de observação. Outra observação convergente entre o a atividade relativa e a diversidade relativa no método AGFAGs no estudo do *Rattus norvegicus* é o metabolismo de lipídios. O mesmo estudo no método EFDEGs encontrou genes e ontologias referentes a regulação do ciclo celular e ao metabolismo de xenobióticos (oxidação dos aldeídos e glucoronidação).

4.3.3 Etopósido

Em geral as figuras 4.37 e 4.38 mostram um quadro parecido com os das figuras 4.33 e 4.34 na ciclofosfamida (AGFAGs), isto é, quanto menor a dose e o tempo mais BPs são encontrados. Esse comportamento é evidente nos 3 estudos realizados no etopósido (atividade).

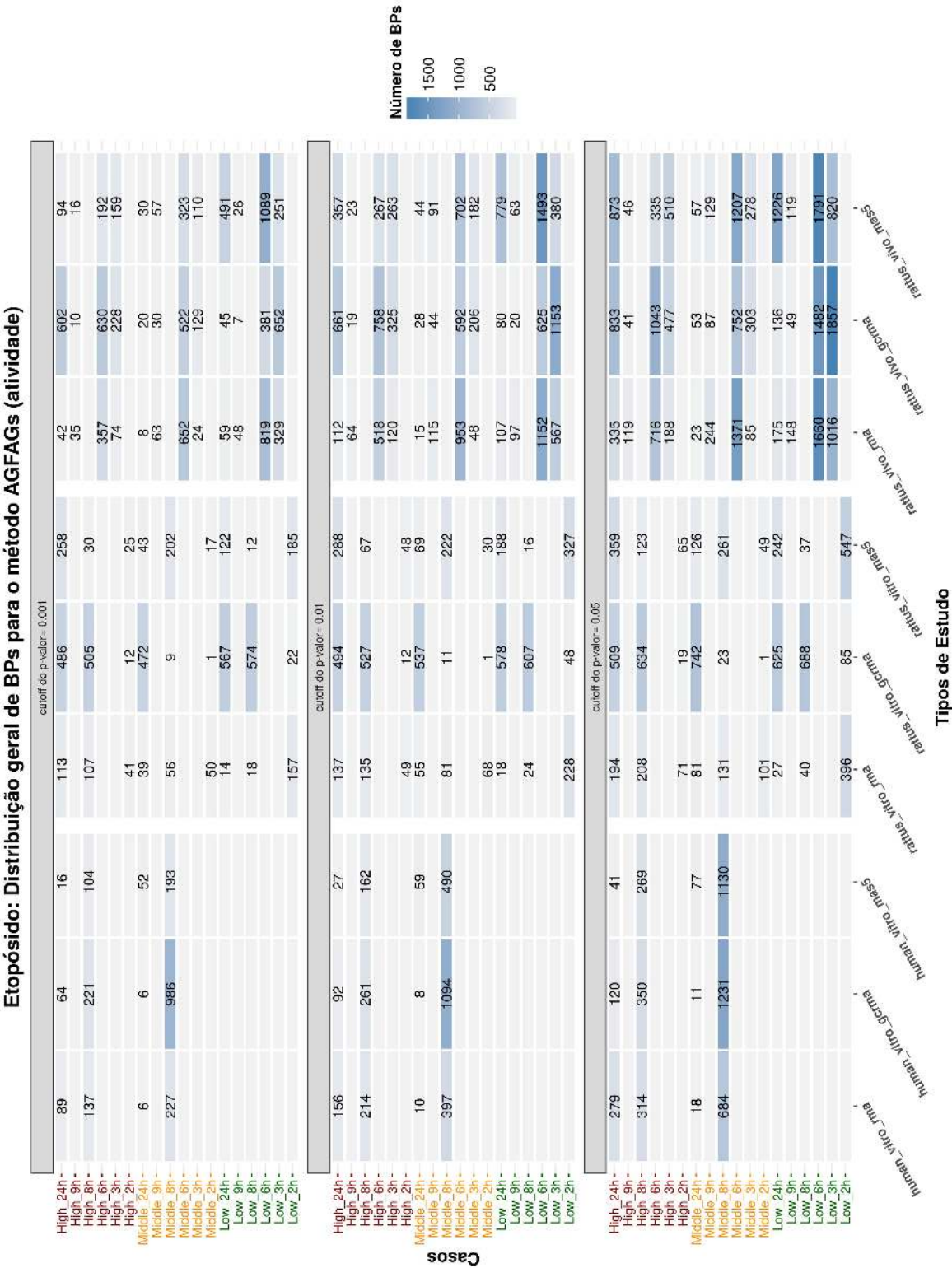


Figura 4.37: BPs encontrados para o etopósido no AGFAGs (atividade). Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0,05, 0,01 e 0,001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

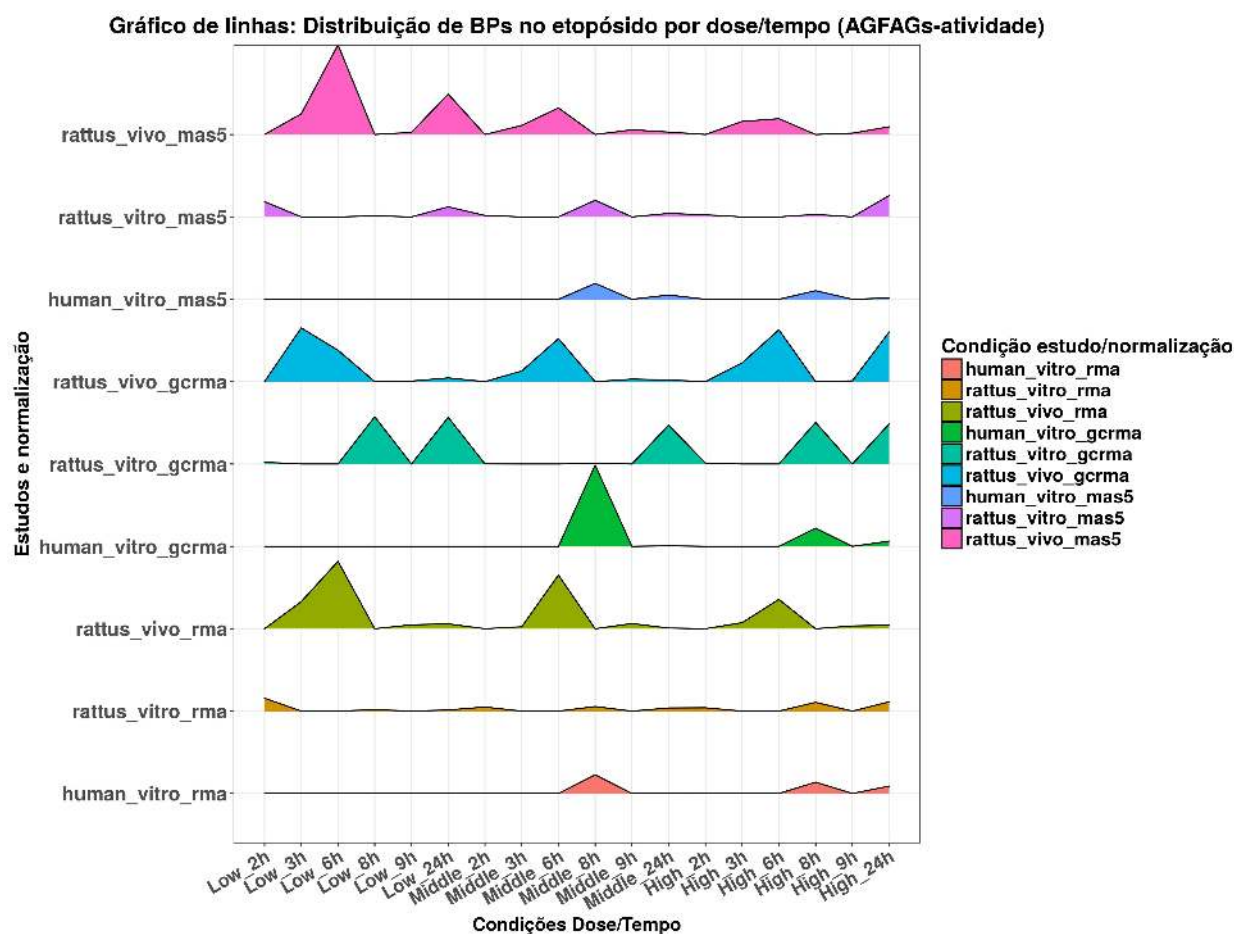


Figura 4.38: Gráfico de linhas para os BPs com cutoff de 0.001 para o etopósido no AGFAGs (atividade). Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.37, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.37 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo*/*in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

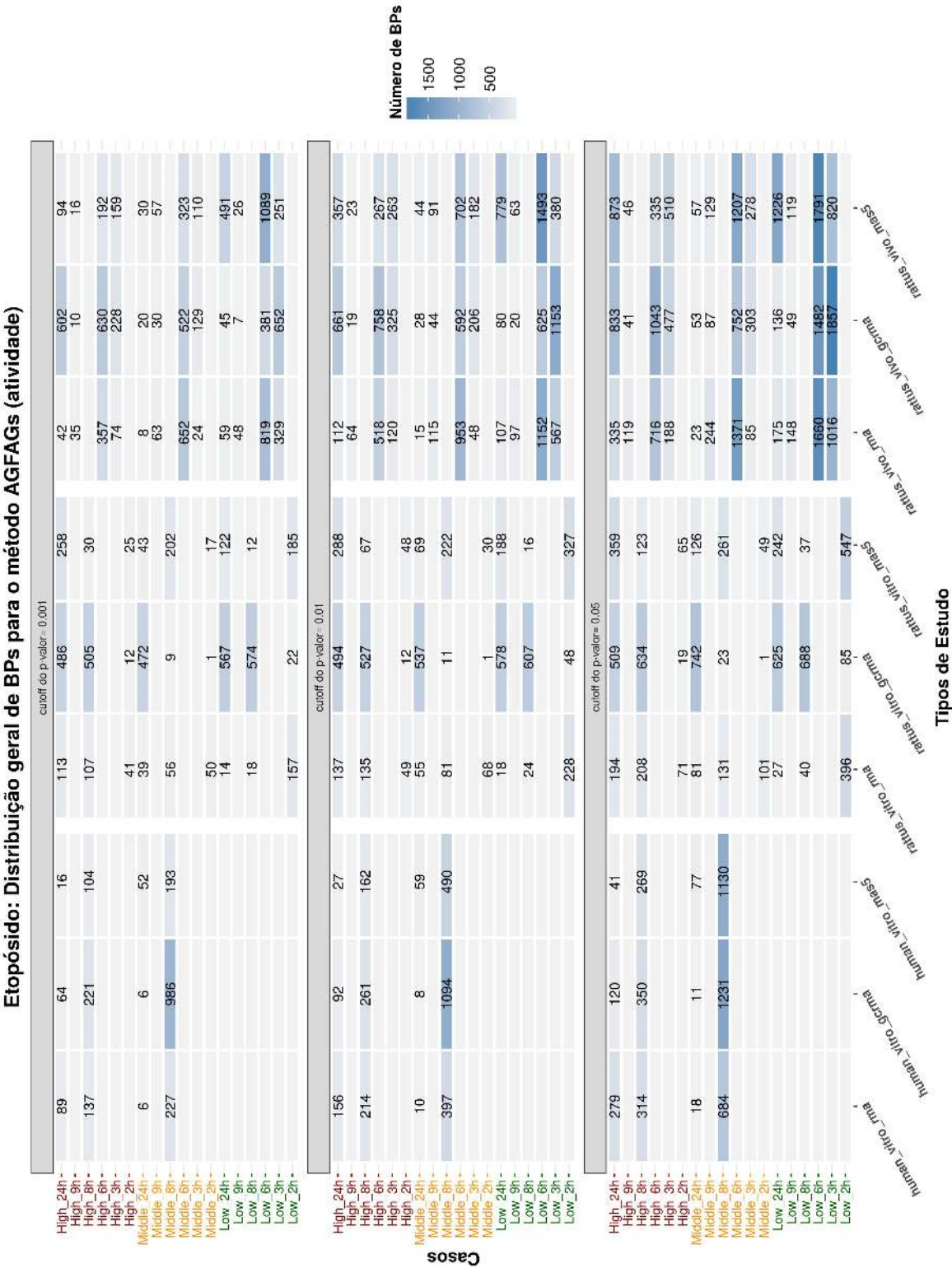


Figura 4.39: BPs encontrados para o etopósido no AGFAGs (diversidade). Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0,05, 0,01 e 0,001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

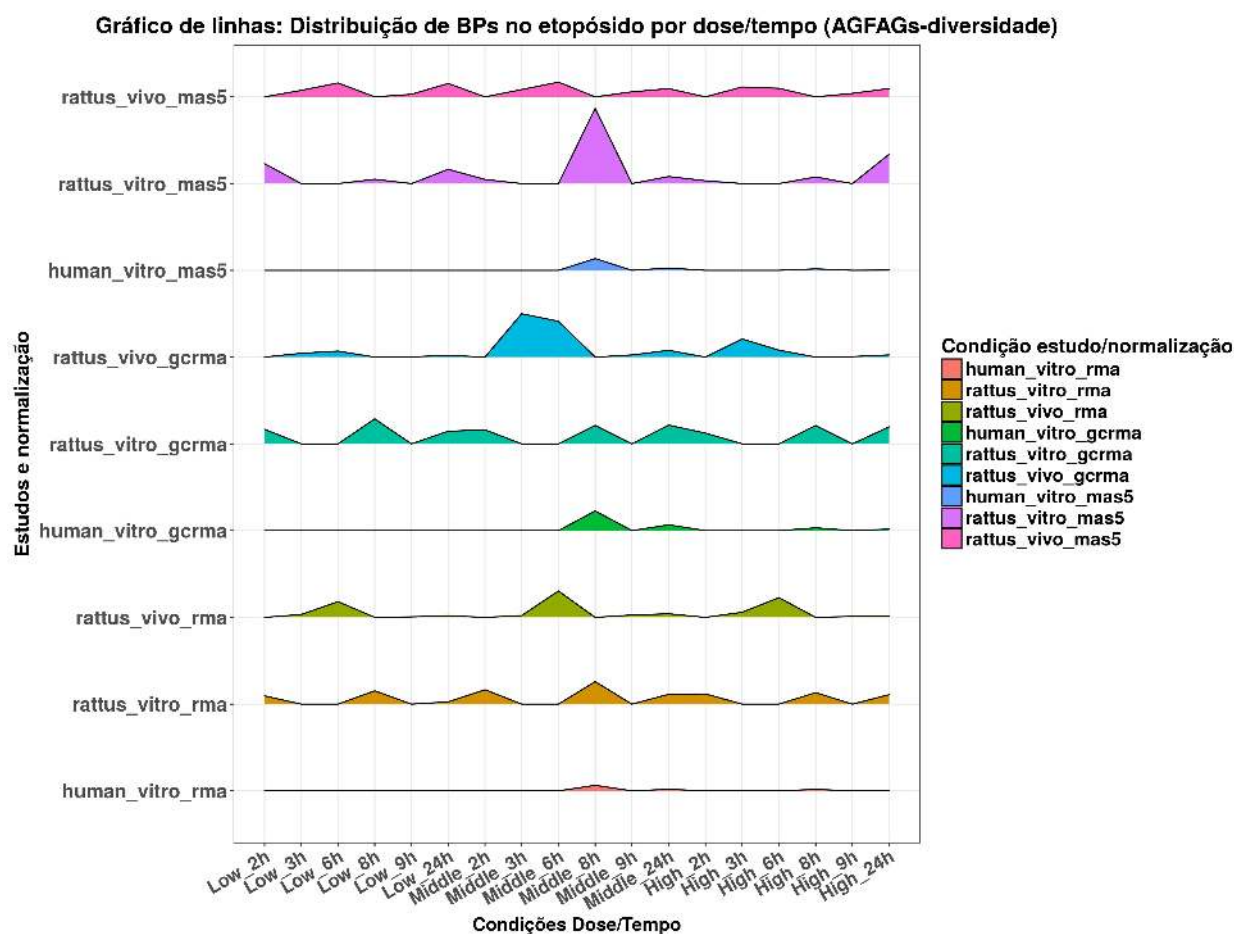


Figura 4.40: Gráfico de linhas para os BPs com cutoff de 0.001 para o etopósido no AGFAGs (diversidade). Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normatizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.39, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.39 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

A perspectiva da diversidade no etopósido (AGFAGs) possui uma distribuição de resultados que exhibe um comportamento similar ao observado para na perspectiva da atividade para o etopósido.

Do ponto de vista da atividade foram encontradas 2078 BPs para o estudo do *Homo sapiens in vitro*, 3621 para o estudo do *Rattus norvegicus in vitro* e 12242 para o estudo do *Rattus norvegicus in vivo*. A filtragem dos resultados, para elencar as ontologias de interesse para posterior análise, foi idêntica à que foi utilizada na ciclofosfamida, sendo consideradas de interesse, as ontologias que estivessem presente em todas as doses do estudo. No caso do etopósido, o estudo do *Homo sapiens in vitro* possui apenas as doses “Middle” e “High”.

No estudo do *Homo sapiens in vitro* foram encontradas 49 ontologias de interesse, segue 25 delas: i) GO:0000710–reparo de incompatibilidade meiótica; ii) GO:0000727–reparo de quebra de fita dupla via replicação induzida por ruptura; iii) GO:0000732–deslocamento de fita de DNA; iv) GO:0001833–proliferação de massa celular interna; v) GO:0006267–formação de complexo pré-replicativo envolvido na replicação do DNA do ciclo celular nuclear; vi) GO:0006268–Desdobramento do DNA envolvido na replicação do DNA; vii) GO:0006269–Replicação de DNA, síntese de primer RNA; viii) GO:0007089–atravessando o ponto de controle inicial do ciclo celular mitótico; ix) GO:0010216–manutenção da metilação do DNA; x) GO:0010767–regulação da transcrição do promotor RNA polimerase II em resposta a danos no DNA induzidos por UV; xi) GO:0031627–formação de loop telomérico; xii) GO:0034395–regulação da transcrição do promotor RNA polimerase II em resposta ao ferro; xiii) GO:0032764–regulação negativa da produção de citocinas de mastócitos; xiv) GO:0043137–Replicação de DNA, remoção de primer RNA; xv) GO:0090267–regulação positiva do posto de verificação do fuso do ciclo celular mitótico; xvi) GO:1900264–regulação positiva da atividade da DNA polimerase dirigida por DNA; xvii) GO:1902595–regulação da ligação da origem da replicação do DNA; xviii) GO:1902990–Manutenção de telômero mitótico via replicação semi-conservadora; xix) GO:1990918–reparo de quebra de fita dupla envolvido na recombinação meiótica; xx) GO:0034760–regulação negativa do transporte transmembranar de íons de ferro; xxi) GO:1990641–resposta à inanição de íons de ferro; xxii) GO:0034395–regulação da transcrição do promotor RNA polimerase II em resposta ao ferro; xxviii) GO:0032764–regulação negativa da produção de citocinas de mastócitos; xxiv) GO:0048478–proteção de garfo de replicação; xxv) GO:0034445–regulação negativa da oxidação da lipoproteína plasmática.

Todas as ontologias listadas acima possuem atividade relativa maior que 0.5 indicando a superexpressão de gens anotados. A maioria das ontologias tem associação com o processos de replicação do DNA, o mais afetado pelo etopósido, que age como um inibidor da enzima topoisomerase II que auxilia no desenrolamento do DNA, sua inibição resulta em diversos danos no DNA o que posteriormente induz à apoptose celular (MAANEN et al., 1988). De fato, pode-se observar entre as ontologias listadas acima algumas que estão associadas à danos ao DNA e reparo dos mesmos. Ainda pode-se notar íons associados a deficiência do ion de ferro, também observados em alguns estudos da ciclofosfamida. O método EFDEGs, para o estudo do *Homo sapiens in vitro* também encontrou diversos genes e ontologias associados com a replicação do DNA, divisão celular e regulação do ciclo celular, tornando

estas observações convergentes entre os dois métodos.

O estudo do *Rattus norvegicus in vitro* encontrou 158 BPS, 20 deles listados a seguir: i) GO:0007493–Determinação do destino celular endodérmico da produção de citocinas mastocitárias; ii) GO:0010623–morte celular programada envolvida no desenvolvimento celular; iii) GO:0042245–Reparo de RNA; iv) GO:0042662–regulação negativa da especificação do destino da célula mesodérmica; v) GO:0048866–especificação do destino das células estaminais; vi) GO:0061364–processo apoptótico envolvido na luteólise; vii) GO:0061402–regulação positiva da transcrição do promotor da RNA polimerase II em resposta ao pH ácido; viii) GO:1990636–senescência reprodutiva; ix) GO:0061418–regulação da transcrição do promotor da RNA polimerase II em resposta à hipóxia; x) GO:0070483–detecção de hipoxia; xi) GO:1990641–resposta à inanição de íons de ferro; xii) GO:0034760–regulação negativa do transporte transmembranar de íons de ferro; xiii) GO:0010727–regulação negativa do processo metabólico de peróxido de hidrogênio; xiv) GO:0045079–regulação negativa do processo biossintético da quimiocina; xv) GO:0050925–regulação negativa da quimiotaxia negativa; xvi) GO:0060302–regulação negativa da atividade da citocina; xvii) GO:0032696–regulação negativa da produção de interleucina-13; xviii) GO:0032714–regulação negativa da produção de interleucina-5; xix) GO:0045401–regulação positiva do processo biossintético da interleucina-3; xx) GO:2000660–regulação negativa da via de sinalização mediada por interleucina-1.

A maior parte das ontologias que não foram listadas acima estavam relacionadas com: regulação negativa da atividade interleucinas e de suas vias de sinalização, regulação negativa da biossíntese de quimiocinas e de suas vias de sinalização, regulação negativa de oxidoredutases e diferenciação celular.

As 13 ontologias anotadas estão principalmente envolvidas com divisão celular, reparo do RNA, controle do ciclo e do destino celular, alteração de vias apoptóticas, senescência reprodutiva, ausência de ferro e resposta a hipóxia. A maioria destes processos são diretamente afetados pelo etopósido, outros como a ausência de ferro e a indução de hipóxia podem estar ligados a outros mecanismos do etopósido (ou de seus metabólitos) que não sejam o seu principal: a inibição da topoisomerase II. No estudo do *Homo sapiens in vitro* no AGFAGs também encontramos esses mesmos processos como alterados, principalmente os envolvidos na replicação do DNA, no reparo de DNA e no controle do ciclo celular, entretanto pode-se também notar que processos envolvidos com a ausência de ferro também

são encontrados. O método EFDEGs para o etopósido no mesmo estudo, encontrou muitas ontologias associadas com processos de replicação do DNA, dano e reparo do DNA e regulação do ciclo celular. Entretanto essas ontologias não foram encontrados em todas as doses no método EFDEGs, a maioria das ontologias encontradas tinham associação com sinalização por quimiocinas e regulação do fuso mitótico e do citoesqueleto. Os processos de sinalização de quimiocinas são observados em ambos os métodos para esta droga, tornando a observação convergente.

O estudo do *Rattus norvegicus in vivo* encontrou, após a filtragem, 174 BPs de interesse, segue listado 15 deles: i) GO:0035986–montagem de foco de heterocromatina associada à senescência; ii) GO:1990636–senescência reprodutiva; iii) GO:0090402–senescência celular induzida por oncogene; iv) GO:0010260–senescência de órgão animal; v) GO:0071140–resolução de intermediários de recombinação mitótica; vi) GO:1904031–regulação positiva da atividade da proteína quinase dependente de ciclina; VII) GO:1900215–regulação negativa do processo apoptótico envolvido no desenvolvimento do ducto coletor metanéfrico; viii) GO:1902806–regulação da transição de fase do ciclo celular G1/S; ix) GO:0034760–regulação negativa do transporte transmembranar de íons de ferro; x) GO:2000340–regulação positiva da produção de ligandos de quimiocina (motif C-X-C) 1; xi) GO:2000538–regulação positiva da quimiotaxia de células B; xii) GO:0050925–regulação negativa da quimiotaxia negativa; xiii) GO:0060302–regulação negativa da atividade da citocina; xiv) GO:0032764–regulação negativa da produção de citocinas de mastócitos; xv) GO:1901843–regulação positiva da atividade do canal de cálcio de alta voltagem.

Todas as ontologias acima possuem atividade relativa superior a 0.5, indicando possível superexpressão dos genes anotados no caso em relação ao controle. Interessante notar a grande quantidade de ontologias associadas com a senescência reprodutiva, algo que se espera acontecer em amostras contendo uma droga antineoplásica. A senescência reprodutiva também foi observada no estudo *Rattus norvegicus in vitro* do AGFAGs e do EFDEGs, tornando essa observação convergente. Também pode se observar a presença de ontologias associadas com a deficiência de ferro, regulação positiva da quimiotaxia e regulação negativa das citocinas. Algumas desses últimos processos citados também aparecem nos outros dois estudos do etopósido (*Homo sapiens in vitro* e *Rattus norvegicus in vitro*) como a deficiência de ferro e a regulação negativa das citocinas.

Agora uma descrição numérica dos resultados da ciclofosfamida da perspectiva da di-

versidade. Foram encontrados 747 BPs para o estudo do *Homo sapiens in vitro*, 1117 para o estudo do *Rattus norvegicus in vitro* e 5253 no estudo *Rattus norvegicus in vivo*. A filtragem desses dados permanece inalterada, sendo consideradas como interessantes as ontologias que aparecerem simultaneamente nas 3 doses do estudo (2 no caso do *Homo sapiens in vitro*).

Para o estudo do *Homo sapiens in vitro* foram encontradas 33 ontologias de interesse, segue uma lista com algumas delas: i) GO:0071163–Montagem do complexo de pré-iniciação de replicação de DNA; ii) GO:0000710–reparo de incompatibilidade meiótica; iii) GO:0051754–Coesão da cromátide irmã meiótica, centromérico; iv) GO:0071038–processo catabólico de tRNA nuclear dependente de poliadenilação; v) GO:0035511 desmetilação oxidativa de DNA; vi) GO:0046601–regulação positiva da replicação do centríolo; vii) GO:0032877–regulação positiva da endoreduplicação do DNA; viii) GO:0006267–formação de complexo pré-replicativo envolvido na replicação do DNA do ciclo celular nuclear; ix) GO:0060623–regulação da condensação cromossômica; x) GO:1902595–regulação da origem da replicação do DNA; xi) GO:0061511–alongamento do centríolo; xii) GO:0034080–CENP-A contendo montagem de nucleossomas; xiii) GO:0006272–alongamento da fita líder; xiv) GO:1990641–resposta à inanição de íons de ferro.

Todas essas ontologias apresentam diversidade relativa e atividade relativa maior que 0.5, indicando que: os casos tem uma distribuição mais heterogênea de valores de expressão do que o controle e que o caso possui em si genes possivelmente superexpressos. A maioria das ontologias elencadas estão associados a processos de replicação do DNA, regulação do fuso mitótico e cromátides e por fim deficiência de ferro. Esse processos também foram encontrados mesmo estudo (AGFAGs-etopósido-*Homo sapiens in vitro*) do ponto de vista da atividade. O método EFDEGs também encontrou diversos GDEs e ontologias associadas a replicação e seu reparo e regulação do fuso mitótico. Podemos concluir que essas observações a respeito da replicação do DNA e regulação do fuso mitótico são convergentes entre os métodos AGFAGs e EFDEGs.

O estudo do *Rattus norvegicus in vitro* encontrou 27 ontologias de interesse segundo a filtragem utilizada, a seguir 12 delas: i) GO:0010260–senescência de órgão animal; ii) GO:0019371–via ciclooxigenase; iii) GO:0045079–regulação negativa do processo biossintético da quimiocina; iv) GO:0032696–regulação negativa da produção de interleucina-13; v) GO:2000660–regulação negativa da via de sinalização mediada por interleucina-1; vi) GO:2000424–regulação positiva da quimiotaxia de eosinófilos; vii) GO:2000404–regulação

da migração de células T; viii) GO:0071436–exportação de íons de sódio; xix) GO:0048635–regulação negativa do desenvolvimento de órgãos musculares; x) GO:0030046–conjunto paralelo de feixes de filamentos de actina; xi) GO:0042662–regulação negativa da especificação do destino da célula mesodérmica; xii) GO:0072137–proliferação de células mesenquimais condensadas.

Assim como no caso do estudo *Homo sapiens in vitro* todas essas ontologias apresentam diversidade relativa e atividade relativa maior que 0.5, indicando que: os casos tem uma distribuição mais heterogênea de valores de expressão do que o controle e que o caso possui em si genes possivelmente superexpressos. A senescência de órgãos também é observada nesse estudo assim como foi observada nos estudos do *Rattus norvegicus in vitro* do ponto de vista da atividade. Também observa-se ontologias a diversos outros processos, principalmente regulação da quimiotaxia, de quimiocinas, de interleucinas, desenvolvimento e proliferação celular. As ontologias referentes ao desenvolvimento e proliferação celular, e regulação da quimiotaxia, das interleucinas e das citocinas também são observadas no mesmo estudo do ponto de vista da atividade. O método EFDEGs encontrou para esse mesmo estudo GDEs e ontologias que estavam associadas regulação da condensação cromossômica e do fuso mitótico, migração de células, controle da quimiotaxia e da sinalização mediada por quimiocinas. O processo de regulação da quimiotaxia é observado tanto no EFDEGs quanto no AGFAGs, portanto é um resultado convergente entre os métodos para o estudo do *Rattus norvegicus in vitro*.

O estudo do *Rattus norvegicus in vivo* encontrou, após a filtragem, 45 ontologias de interesse, segue abaixo 11 delas: i) GO:0060302–regulação negativa da atividade da citocina; ii) GO:2000342–regulação negativa da produção do ligante 2 de quimiocina (motif C-X-C); iii) 300288 GO:0050925–regulação negativa da quimiotaxia negativa; iv) GO:0046851 regulação negativa da remodelação óssea; v) GO:2001287–regulação negativa da endocitose mediada por caveolina; vi) GO:0050712–regulação negativa da secreção de interleucina-1 alfa; vii) GO:1903262–regulação negativa da serina fosforilação da proteína STAT3; viii) GO:0032730–regulação positiva da produção de interleucina-1 alfa; ix) GO:2000366–regulação positiva da importação de proteínas STAT no núcleo; x) GO:0010260–senescência de órgão animal; xi) GO:0071049–retenção nuclear de pré-mRNA com extremidades 3' aberrantes no local da transcrição.

Assim como no caso do estudo *Rattus norvegicus in vitro* todas essas ontologias apre-

sentam diversidade relativa e atividade relativa maior que 0.5, indicando que: os casos tem uma distribuição mais heterogênea de valores de expressão do que o controle e que o caso possui em si genes possivelmente superexpressos. A presença do processo de senescência do órgão animal é algo que faz sentido, já que o etopósido induz à apoptose celular mediante inibição da topoisomerase II, cabe lembrar que processos de senescência foram encontrados para o etopósido no estudo do *Rattus norvegicus in vitro* (em ambos pontos de vista; atividade e diversidade). Outro processo observado é a retenção de mRNA aberrante, algo que se espera acontecer em células onde a topoisomerase II não esteja funcionando corretamente (MAANEN et al., 1988). Observa-se também processos como a regulação de citocinas, regulação de quimiocinas, regulação da quimiotaxia e regulação da produção e secreção de interleucinas. Processos associados com a regulação de quimiocinas, citocinas, quimiotaxis e de interleucinas também são observados no estudo do *Rattus norvegicus in vitro* para ambos os pontos de vista no método AGFAGs e para o estudo *Homo sapiens in vitro* do ponto de vista da atividade, No método EFDEGs os genes e ontologias considerados de interesse estavam principalmente com a regulação do ciclo celular e diversas fosforilação. Portanto os resultados de interesse entre os métodos AGFAGs e EFDEGs para o estudo do etopósido (*Rattus norvegicus in vivo*) são divergentes.

4.3.4 Lomustina

A perspectiva da atividade na lomustina tem resultados cuja distribuição é parecida com as do etopósido e da ciclofosfamida, cujo número de BPs aumenta conforme o tempo e a dose diminuem. Entretanto podemos notar que os estudos do *Rattus norvegicus in vivo* são excessões, para esses estudos as doses médias apresentaram um número de BPs maior do que nas outras doses.

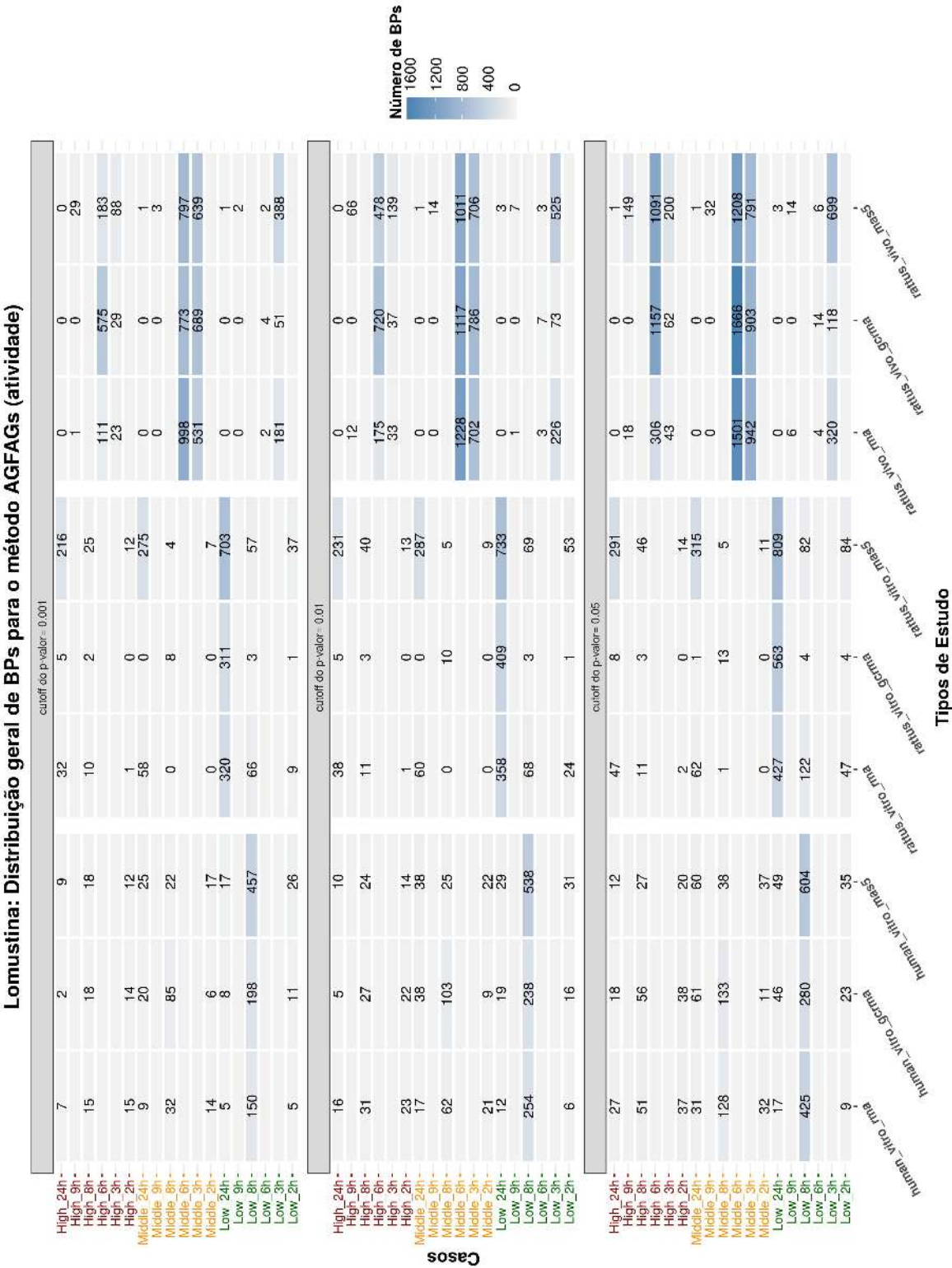


Figura 4.41: BPs encontrados para a lomustina no AGFAGs (atividade). Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0,05, 0,01 e 0,001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normalização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas , foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

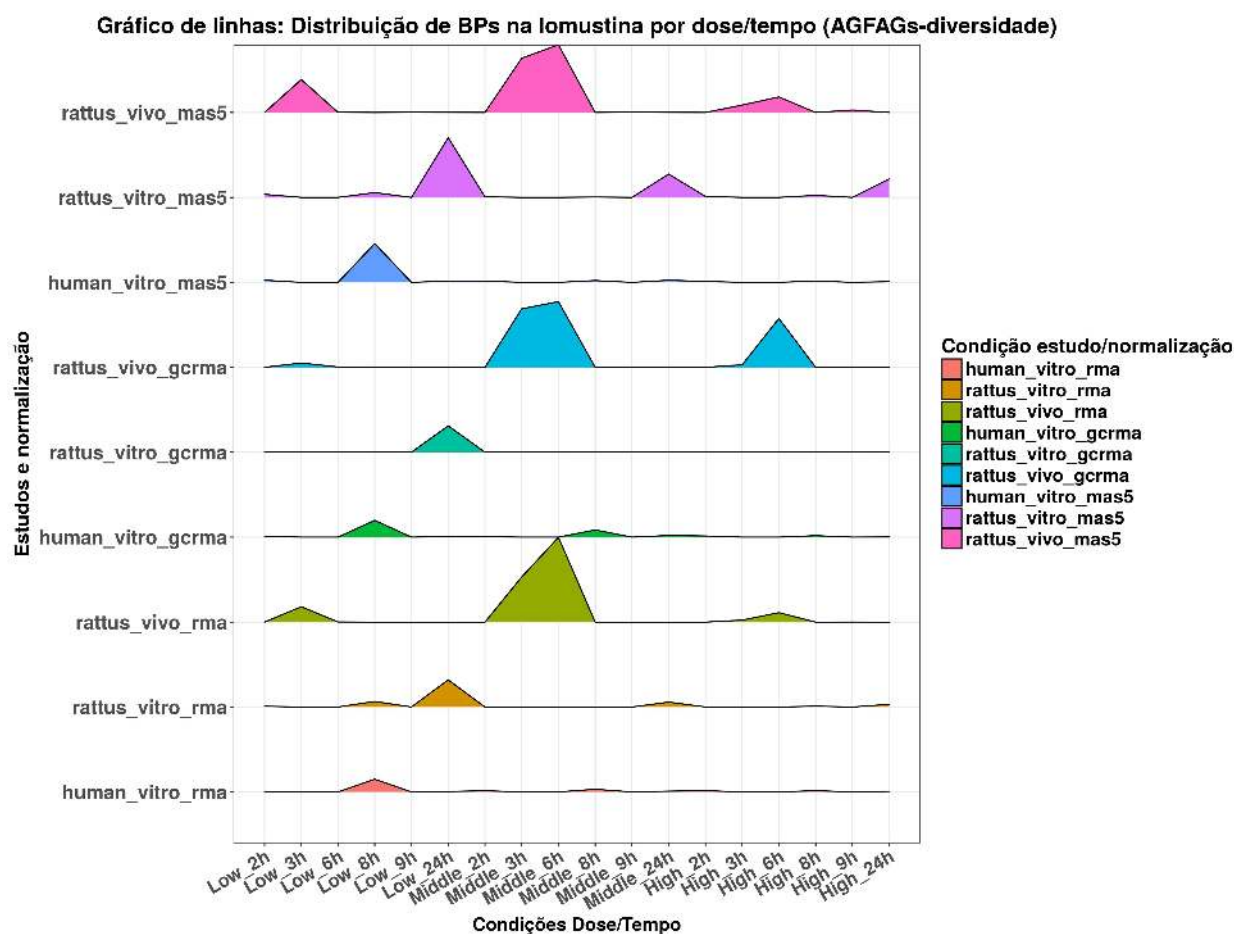


Figura 4.42: Gráfico de linhas para os BPs com cutoff de 0.001 para lomustina no AGFAGs (atividade). Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.41, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.41 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

A perspectiva da diversidade apresenta um comportamento encontrado também na perspectiva da atividade, como pode se observar nas figuras 4.43 e 4.44, até mesmo a concentração de BPs nas doses médias observada na outra perspectiva, também é observado aqui.

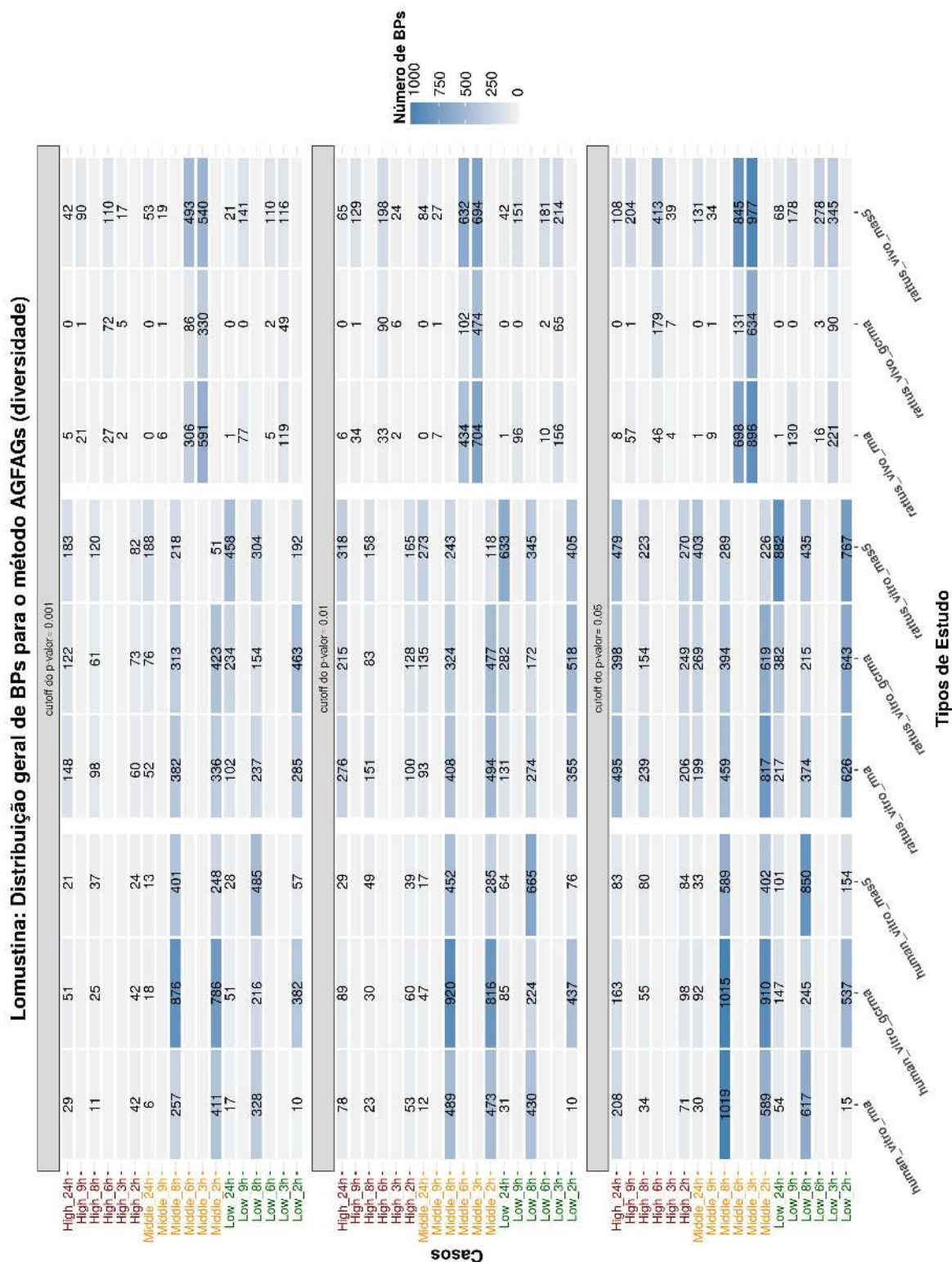


Figura 4.43: BPs encontrados para a lomustina no AGFAGs (diversidade). Distribuição geral de BPs significativos segundo a atividade relativa encontrados no método AGFAGs para as 3 drogas antineoplásicas considerando 3 diferentes valores de cutoff de p-valor (0.05, 0.01 e 0.001). As quantidades estão separadas por condições compostas pelo tipo de estudo/normatização utilizada (nas colunas) e dose/tempo de amostragem (nas linhas), note que as doses estão marcadas com cores diferentes para melhor visualização e que células vazias indicam a inexistência do experimento em tal condição. Um *heatmap*, com uma escala comum a todas tabelas, foi projetado sobre a tabela para mostrar a densidade de GDEs nas diferentes situações.

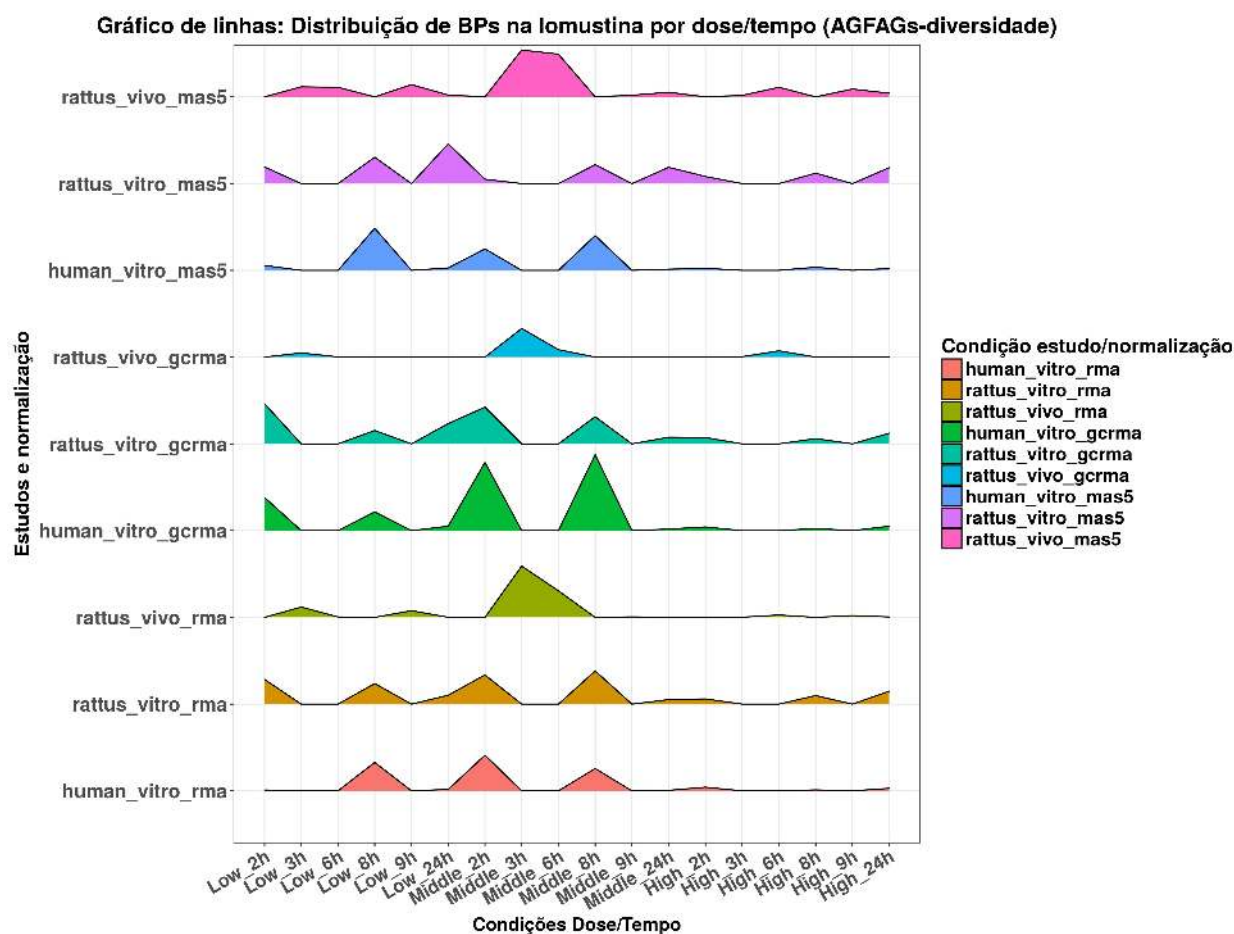


Figura 4.44: Gráfico de linhas para os BPs com cutoff de 0.001 para lomustina no AGFAGs. Na figura o eixo horizontal contém as condições a respeito da dose/tempo de cada caso enquanto o eixo vertical contém os estudos/normalizações. Assim pode-se traçar um gráfico de linhas para cada coluna da figura Figura 4.43, cada cor representa uma coluna diferente. Existe um terceiro eixo “invisível” para as alturas de cada linha, refletindo o número de BPs significantes segundo sua atividade relativa encontrados em cada caso. Note que os estudos estão organizados de forma a colocar normalizações semelhantes perto de si, na Figura 4.43 os estudos estão agrupados segundo as espécies e tipo de estudo (*in vivo/in vitro*). Nesta configuração a imagem permite visualizar o comportamento dos resultados em função da dose/tempo nas normalizações de cada estudo.

Do ponto de vista da atividade foram encontradas 936 BPs para o estudo do *Homo sapiens in vitro*, 2121 para o estudo do *Rattus norvegicus in vitro* e 6092 para o estudo do *Rattus norvegicus in vivo*. A filtragem dos resultados, para elencar as ontologias de interesse para posterior análise, foi idêntica à que foi utilizada na ciclofosfamida e no etopósido, sendo consideradas de interesse, as ontologias que estivessem presente em todas as doses do estudo.

O estudo do *Homo sapiens in vitro* foram elencadas 21 ontologias de interesse, 9 delas listadas a seguir: i) GO:0006788—oxidação heme; ii) GO:1990641—resposta à inanição de íons de ferro; iii) GO:0034395—regulação da transcrição do promotor RNA polimerase II em resposta ao ferro; IV) GO:0034760—regulação negativa do transporte transmembranar de íons de ferro; v) GO:2000482—regulação da secreção de interleucina-8; vi) GO:0090038—regulação

negativa da sinalização da proteína quinase C; vii) GO:0071805–transporte transmembrana de íons de potássio; viii) GO:0032764–regulação negativa da produção de citocinas de mastócitos; ix) GO:0014740–regulação negativa da hiperplasia muscular.

Todas as ontologias listadas acima apresentam atividade relativa maior que 0.5, indicando possível superexpressão dos genes anotados. A regulação da secreção de interleucinas e a regulação produção de citocinas aparecem no conjunto acima, assim como apareceram diversas vezes nos estudos do etopósido (principalmente os estudos do *Rattus norvegicus*), evidenciando uma alteração em comum para duas drogas. Também observa-se muitos processos associados a deficiência do íon de ferro, esses processo também foram observados nos estudos do etopósido e da ciclofosfamida (AGFAGs), tornando essa uma observação robusta.

O estudo do *Rattus norvegicus in vitro* encontrou 123 ontologias de interesse após a filtragem, 20 delas são: i) GO:0035986–montagem de foco de heterocromatina associada à senescência; ii) GO:0061418–regulação da transcrição do promotor da RNA polimerase II em resposta à hipóxia; iii) GO:0010866–regulação do processo biossintético de triglicerídeos; iv) GO:1903984–regulação positiva da via de sinalização apoptótica ativada por TRAIL; v) GO:0070368–regulação positiva da diferenciação de hepatócitos; vi) GO:0090402–senescência celular induzida por oncogene; vii) GO:0010871–regulação negativa do processo biossintético do receptor; viii) GO:0032929–regulação negativa da geração de ânion superóxido; ix) GO:2000660–regulação negativa da via de sinalização mediada por interleucina-1; x) GO:0045796–regulação negativa da absorção intestinal do colesterol; xi) GO:0010949–regulação negativa da absorção intestinal de fitosteróis; xii) GO:0042737–processo catabólico de drogas; xiii) GO:0070483–detecção de hipoxia; xiv) GO:0034760–regulação negativa do transporte transmembranar de íons de ferro; xv) GO:0090212–regulação negativa de estabelecimento de barreira hematoencefálica; xvi) GO:0046320–regulação da oxidação de ácidos graxos; xvii) GO:2000482–regulação da secreção de interleucina-8; xviii) GO:1990641–resposta à inanição de íons de ferro; xix) GO:0070561–via de sinalização do receptor de vitamina D; xx) GO:0097089–processo metabólico de ácidos graxos metil ramificados.

Assim como no caso do estudo *Homo sapiens in vitro* todas essas ontologias apresentam atividade relativa maior que 0.5 (e curiosamente a diversidade relativa também) , indicando que: os casos tem uma distribuição mais heterogênea de valores de expressão do que o controle e que o caso possui em si genes possivelmente superexpressos. Este conjunto de

ontologias de interesse cobre processos bem mistos, a senescência reprodutiva e a deficiência de íons de ferro continuam sendo observadas, assim como foram observadas em diversos estudos do etopósido e da ciclofosfamida no método AGFAGs. Também nota-se processos associados a detecção de hipóxia, metabolismo de ácidos graxos, absorção intestinal de lipídios, metabolismo de drogas e regulação de interleucinas. Outro processo interessante é a regulação negativa da barreira hematoencefálica, pois a lomustina é amplamente utilizada no tratamento de tumores cerebrais por causa de sua capacidade de atravessar essa barreira (LEE et al., 1985). No EFDEGs o estudo *Rattus norvegicus in vitro* para a lomustina encontrou genes e ontologias associados ao metabolismo de xenobióticos, metabolismo de aldeídos e cetonas e alteração de vias apoptóticas. Processos relacionados ao metabolismo de drogas e a alteração de vias apoptóticas são comuns aos dois métodos (EFDEGs e AGFAGs), tornando essas observações convergentes entre os dois métodos.

O estudo do *Rattus norvegicus in vivo* encontrou 39 ontologias de interesse, algumas delas delas alistadas a seguir: i) GO:0070488–agregação de neutrófilos; ii) GO:0047484–regulação da resposta ao estresse osmótico; iii) GO:0032764–regulação negativa da produção de citocinas de mastócitos; iv) GO:0048294–regulação negativa da mudança de isotipo para isotipos IgE; v) GO:2000182–regulação do processo biossintético de progesterona; vi) GO:0035350–Transporte transmembrana FAD; vii) GO:0034653–processo catabólico do ácido retinóico; viii) GO:0002357–resposta de defesa a células tumorais; ix) GO:0070427–domínio de oligomerização de ligação ao nucleotídeo contendo 1 via de sinalização; x) GO:0080121–Transporte AMP, xi) GO:0043132–Transporte NAD; xii) GO:0044339–via de sinalização Wnt canônica envolvida na diferenciação osteoblástica; xiii) GO:2000276–regulação negativa da atividade de desacoplador da fosforilação oxidativa; xiv) GO:0044375–regulação do tamanho do peroxissoma.

Todas as ontologias anotadas acima possuem atividade relativa maior que 0.5. A alteração do transporte transmembranar é algo evidente pelas ontologia elencadas, outro processo observado é a regulação de processos envolvendo oxidação e resposta ao estresse osmótico. São ontologias descrevendo processos bem mistos e variados. Entretanto uma ontologia que merece atenção aqui é a “GO:0002357–resposta de defesa a células tumorais”, que indica a presença de células tumorais ou de efeitos carcinogênicos. A lomustina possui potencial carcinogênico como evidenciado em (MARSELOS; VAINIO, 1991), mostrando que uma droga antineoplásica é também fonte de potenciais efeitos carcinogênicos.

Do ponto de vista da diversidade foram encontradas 284 BPs para o estudo do *Homo sapiens in vitro*, 523 para o estudo do *Rattus norvegicus in vitro* e 1589 para o estudo do *Rattus norvegicus in vivo*. A filtragem dos resultados, para elencar as ontologias de interesse para posterior análise, foi idêntica à que foi utilizada na ciclofosfamida e no etopósido, sendo consideradas de interesse, as ontologias que estivessem presente em todas as doses do estudo.

Para o estudo *Homo sapiens in vitro* foi encontrada apenas uma única ontologia: i) GO:1904816–regulação positiva da localização da proteína no cromossomo, região telomérica. Essa ontologia possui diversidade relativa e atividade relativa maiores que 0.5. Isso indica um cenário em que o caso tem distribuição de valores de expressão mais homogêneos e estes aparentemente estão superexpressos em relação ao controle. Essa ontologia se refere a manutenção de proteínas em um determinado local da célula, no caso a região telomérica. Isso pode ser reflexo da atuação da lomustina como uma nitrosureia, causando danos ao DNA e induzindo à apoptose celular. A perspectiva da diversidade na lomustina para o estudo *Homo sapiens in vitro* foi bem diferente da perspectiva da atividade, encontrando apenas 1 ontologia de interesse segundo o protocolo de filtragem enquanto a outra encontra 21 delas. Isso evidencia algo que já é discutido desde o método EFDEGs: os estudos da lomustina não são muito informativos. Praticamente qualquer outro estudo retorna mais resultados do que os *in vitro* no método EFDEGs, para o AGFAGs isso pode implicar em uma situação em que as diferenças entre caso e controle não são tão evidentes, assim a filtragem realizada para elencar as ontologias de interesse (que se baseia na convergência de observações entre as 3 doses) elimina a maioria das ontologias com diversidade relativa significativa do *pool* de ontologias de interesse. De qualquer o estudo *in vitro* na lomustina se mostra pouco informativo, de forma que fica difícil tirar mais conclusões.

Para o estudo do *Rattus norvegicus in vitro* foram encontradas 16 ontologias de interesse, todas elas anotadas abaixo: i) GO:0055090–homeostase de acilglicerol; ii) GO:0043420–processo metabólico antranilato; iii) GO:0072703–resposta celular ao metanossulfonato de metila; iv) GO:0071379–resposta celular ao estímulo da prostaglandina; v) GO:0019265–processo biossintético da glicina, por transaminação de glioxilato; vi) GO:0071267–Resgate de L-metionina; vii) GO:0097089–processo metabólico de ácidos graxos metil ramificados; viii) GO:0051919–regulação positiva da fibrinólise; ix) GO:1905322–regulação positiva da migração de células-tronco mesenquimais; x) GO:1990743–sialilação proteica; xi) GO:0070857–regulação do processo biossintético dos ácidos biliares; xii) GO:1902033–regulação

da proliferação de células-tronco hematopoiéticas; xiii) GO:0048385–regulação da via de sinalização do receptor de ácido retinóico; xiv) GO:0010034–resposta ao acetato; xv) GO:0090648–resposta ao enriquecimento ambiental; xvi) GO:0006684–processo metabólico da esfingomielina.

As ontologias encontradas no estudo *Rattus norvegicus in vitro* na lomustina do ponto de vista da diversidade, possuem tanto a diversidade relativa quanto a atividade relativa maior que 0.5. Essas ontologias são bem mistas entre si, envolvendo processos de síntese e/ou metabolismo de ácidos graxos e lipídios, regulação de proliferação celular e da migração celular. O ponto de vista da diversidade elencou ontologias bem diferentes das que foram encontradas pela outra perspectiva (diversidade), assim como ocorreu entre os estudos do *Homo sapiens in vitro* das duas perspectivas. Devido a variedade grande dessas ontologias elencadas é difícil explicar, em vias gerais, o que está a ocorrer nesse estudo.

Para o estudo do *Rattus norvegicus in vivo* foram encontradas 18 ontologias de interesse, todas elas anotadas abaixo: i) GO:0031247–formação de haste de actina; ii) GO:1902742–processo apoptótico envolvido no desenvolvimento; iii) GO:0072709–resposta celular ao sorbitol; iv) GO:0035350–Transporte transmembrana FAD; v) GO:0046909–transporte intermembranas; vi) GO:0048294–regulação negativa da mudança de isotipo para isotipos IgE; vii) GO:0090285–regulação negativa da glicosilação de proteínas em Golgi; viii) GO:0008612–modificação de peptidil-lisina para peptidil-hipusina; ix) GO:2000627–regulação positiva do processo catabólico de miRNA; x) GO:1903003–regulação positiva da desubiquitinação proteica; xi) GO:0001798–regulação positiva da hipersensibilidade do tipo IIa; xii) GO:0042701–secreção de progesterona; xiii) GO:1990743–sialilação de proteínas; xiv) GO:2000182–regulação do processo biossintético de progesterona; xv) GO:0002457–Processamento de antígeno de células T; xvi) GO:0070634–transporte de amônio transepitelial; xvi) GO:0007632–comportamento visual; xvii) GO:0042373–processo metabólico de vitamina K.

As ontologias acima também são bem mistas, mas possibilitam o foco em alguns processos, ao contrário das ontologias encontradas no estudo *Rattus norvegicus in vitro* (AGFAGs-diversidade). Aqui observa-se a alteração de vias apoptóticas e alteração do transporte transmembranas de diversas substâncias. A apoptose é algo esperado como consequência da ação da lomustina, que age realizando intercalações no DNA, induzindo a célula à apoptose. Praticamente todas as ontologias elencadas aqui, curiosamente, a atividade relativa maior que 0.5, com exceção da ontologia “GO:0031247–formação de haste de actina”, indicando

uma regulação negativa da formação do citoesqueleto e possivelmente do fuso mitótico no controle em relação ao caso (algo esperado da lomustina).

4.4 Convergência dos Resultados

A análise numérico-descritiva feita para cada droga em cada método deixa alguns comportamentos bem claros. Um exemplo é a tendência que as drogas tem de encontrar mais GDEs conforme aumentan-se a dose e o tempo no método EFDEGs. Ou mesmo a maneira peculiar com que o método AGFAGs funciona em função da estatística de *bootstrap*. Casos que apresentavam menos GDEs no método EFDEGs apresentaram os maiores números de BPs no AGFAGs, a recíproca também é verdadeira. Como consequência desse comportamento, observa-se que a quantidade de BPs aumentava conforme a dose e o tempo diminuiam no método AGFAGs. Por essa razão o comportamento dos gráficos de linhas era bem irregular no método AGFAGs, ao ponto de não se encontrar esse ciclos de ativação. Analogamente era fácil observar a existência dos ciclos de ativação em cada estudo realizado nos gráficos de linhas do método EFDEGs

Não há muito o que discutir sobre a avaliação quantitativa da questão *in vivo* X *in vitro* pois existe uma limitação fundamental no *design* do OPEN TG-GATEs: os estudos *in vitro* tinha amostragem em 2, 8 e 24 horas enquanto os estudos *in vivo* em 3, 6, 9 e 24 horas. Assim a análise numérico-descritiva só pode ser realizada de maneira eficiente somente na comparação dos estudos *in vitro*. Em geral estes estudos (*in vitro*) mostraram comportamentos parecidos para todas as drogas dentro de método EFDEGs. Isso pode ser observado nos gráficos de linhas, a principal diferença entre os gráficos de linhas entre estudos *in vitro* era a magnitude (a quantidade de resultados encontrados), sendo que os estudos do *Rattus norvegicus in vitro* geralmente encontravam mais resultados. No método AGFAGs os gráficos de linhas dos estudos *in vitro* são bem diferentes entre si em função da maneira com que o método AGFAGs calcula algumas estatísticas. Assim concluímos que de um ponto de vista meramente quantitativo o método EFDEGs apresenta uma certa similaridade entre os estudos *in vitro*. Isso não acontece no método AGFAGs por razões que foram discutidas. Quanto ao estudo do *Rattus norvegicus in vivo*, é ineficiente compará-lo com os outros dois estudos utilizando as figuras por causa dos tempos de amostragem. Entretanto é nitido que estes estudos frequentemente são os menos numerosos no método

EFDEGs, o oposto ocorre no método AGFAGs.

Para melhor discutir a questão *in vivo* X *in vitro*, as análises biológicas são mais eficientes. Foram obtidos resultados numerosos no que se refere aos GDEs e BPs em ambos os métodos. Por essa razão neste trabalho optamos por realizar a análise biológica com GDEs/BPs de “interesse”, ou seja, GDEs/BPs filtradas. O filtro foi escolhido de maneira a escolhermos os resultados mais convergentes entre as doses de um estudo. Por essa razão no método EFDEGs, sempre que possível era-se selecionados GDEs/BPs que estavam presentes simultaneamente em todas as 3 doses. Em alguns estudos de algumas droga isso não era possível pois um das doses tinha resultados nulos em todos os seus tempos (frequentemente a dose “Low”). No AGFAGs tendo resultados ainda mais numerosos que o EFDEGs devido a estatística *bootstrap* foi aplicado a mesma filtragem descrita acima.

Em geral os estudos (*in vitro/in vivo*) são bem diferentes entre si, possuindo poucas ontologias em comum dentro de uma droga. Para facilitar essa avaliação foram construídas 3 tabelas, uma para os resultados de cada método: EFDEGs, AGFAGs (atividade) e AGFAGs (diversidade). Cada uma dessas tabelas contém em si as drogas nas linhas e os estudos nas colunas, e em cada célula estarão elencadas “palavras chave” que representam parte dos resultados mais interessantes em cada caso (droga/estudo). Essas palavras estarão seguidas de dos termos “(sup)” e “(sub)” indicando a regulação positiva e negativa, respectivamente, da função representada por essas palavras. Se não foram encontrados resultados o suficiente a frase “poucos resultados” estará escrita.

A tabela 4.1 abaixo mostra alguns processos chave aos quais os GDEs/BPs estão relacionados em cada droga/estudo no método EFDEGs.

Tabela 4.1: Processos chave encontrados nos estudos/drogas no EFDEGs

	HVT	RVT	RVV
Ciclofosfamida	metabolismo CYP (sup) danos ao DNA (sup) ciclo celular (sub) quimiocinas (sub)	metabolismo UGT ALDH (sup) apoptose (sup) transmembrana (sub) citoesqueleto(sub)	apoptose (sup) danos ao DNA (sup) lipídios (sub)
Etopósido	dano ao DNA (sup) reparo do DNA (sub) fuso mitótico (sub)	dano ao DNA (sup) reparo do DNA (sub) fuso mitótico (sub)	ciclo celular (sub) lipídios (sub)
Loustina	danos ao DNA (sup) crescimento - (sup) fuso mitótico (sub) ciclo celular (sub)	crescimento (sub) fuso mitótico (sub) ciclo celular (sub)	apoptose (sup) interleucinas (sup)

Tabela 4.2: Processos chave encontrados nos estudos/drogas no AGFAGs (atividade)

	HVT	RVT	RVV
Ciclofosfamida	cromatina (sub) catabolismo DNA (sup) ferro (sub) quimiocinas (sub)	ácidos graxos (sup) apoptose (sup) absorção (sub) ferro (sub)	apoptose (sup) absorção (sub) citocinas (sub)
Etopósido	dano ao DNA (sup) reparo do DNA (sub) fuso mitótico (sub)	apoptose (sup) senescência (sup) ferro (sub) interleucinas (sub)	apoptose (sup) senescência (sup) ferro (sub) citocinas (sub)
Loustina	danos ao DNA (sup) citocinas (sub) ferro (sub)	senescência (sup) hipóxia (sup) interleucinas (sub) ferro (sub)	apoptose (sup) citocinas (sup)

Tabela 4.3: Processos chave encontrados nos estudos/drogas no AGFAGs (diversidade)

	HVT	RVT	RVV
Ciclofosfamida	hipóxia (sup)	dano membrana (sup) apoptose (sup) ciclo celular (sub)	apoptose (sup) reparo do DNA (sup) ciclo celular (sub) citocinas (sub)
Etopósido	replicação de DNA (sup) reparo do DNA (sup) ciclo celular (sub) ferro (sub)	quimiotaxia(sup) senescência (sup) ciclo celular (sub) interleucinas (sub)	senescência (sup) interleucinas (sub) citocinas (sub)
Lomustina	poucos resultados	ácidos graxos (sup)	apoptose (sup) citoesqueleto (sub) transmembrana (sub)

Ao analisar as 3 tabelas fica nítido que no geral pode-se concluir que os 3 estudos são definitivamente diferentes entre si pois frequentemente exibem resultados bem variados. A pouca convergência que existe entre os 3 tipos de estudo se limita aos termos relacionados com processos que normalmente são alterados pelas drogas antineoplásicas estudadas como controle do ciclo celular, apoptose, dano e reparo do DNA.

Existe também pouca convergência entre os métodos e assim como a pouca convergência que foi observada entre os 3 tipos de estudos, esta se limita a processos que normalmente são alterados pelas drogas antineoplásicas estudadas. A diferença intrínseca entre as estatísticas de cada método é o que ultimamente irá classificar um gene/função como significativamente alterado ou não, obviamente isso irá variar conforme a estatística utilizada. É o que observamos entre essas 3 tabelas, por exemplo a palavra chave “dano a membrana” é observada na tabela 4.3 (ciclofosfamida; RVT), o que é consequência do estresse oxidativo provindo dos metabólitos da ciclofosfamida. Esse processo não é observado nas tabelas 4.1 e 4.2, apesar de que sua observação é algo que tem respaldo na literatura. A integração dos resultados das duas metodologias possibilita realizar um estudo explorativo das drogas tem maior alcance do que com o uso dos resultados de apenas uma metodologia.

5 Conclusão

Com base na análise numérico-descritiva concluímos que o método EFDEGs possui resultados (GDEs e BPs) que variam conforme a dose e o tempo. Em geral o número de resultados é maior conforme a dose da droga é maior, e o mesmo se aplica ao tempo. O método AGFAGs possui resultados (BPs apenas) cuja distribuição varia muito entre as drogas e estudos, não havendo um comportamento claro entre as distribuições como mostram os gráficos de linhas. O único comportamento observado é que quanto menor o número de GDEs em um estudo do EFDEGs menor, maior é o número de BPs encontrados nesse mesmo estudo no método AGFAGs. Os estudos *in vitro* no método EFDEGs tem, em geral, um comportamento semelhante entre si através do aumento da dose e do tempo. Sendo que a magnitude (quantidade) de resultados é o que os difere entre si. No método AGFAGs os estudos *in vitro* não possuem, entre si a mesma similaridade que é mostrada no método EFDEGs. Isso é consequência da maneira peculiar que o AGFAGs funciona.

Quanto a questão *in vivo vs in vitro*, o uso de ambas as metodologias deixou evidente que existe pouca similaridade entre os 3 estudos. As tabelas 4.1, 4.2 e 4.3 na sessão anterior deixam isso claro, a pouca convergência existente entre os 3 tipos de estudo se limita aos processos que normalmente são alterados pelas drogas antineoplásicas como apoptose, dano e reparo do DNA e controle do ciclo celular. A convergência entre os métodos também é escassa e se limita feita aos termos: ciclo celular, apoptose, reparo e dano ao DNA. Todos processos que normalmente são alterados pelas drogas antineoplásicas estudadas. Isso é interessante pois observamos processos cuja alteração era esperada em ambos os métodos, dando robustez a observação. É interessante notar que as diferenças entre os métodos muitas vezes são complementares, como no caso *Homo sapiens in vitro* da ciclofosfamida, no método EFDEGs encontramos GDEs/BPs associados aos seus principais mecanismos (ciclo celular), ao metabolismo e ainda potencial regulação pós-transcricional pelo lncRNA EGOT. O método AGFAGs para esse mesmo estudo encontrou também BPs associados também aos

principais mecanismos da droga (ciclo celular) e a potenciais efeitos adversos como fadiga e anemia (os quais não são observados no método EFDEGs). Isso não apenas deixa evidente a importância da prática de se utilizar dados biológicos pra análises exploratórias na busca de hipóteses. E a meta-análise deste trabalho deixa claro como o uso de diferentes metodologias pode ampliar a gama de pontos interessantes em um mesmo conjunto de dados.

Referências Bibliográficas

AEBERSOLD, R.; MANN, M. Mass spectrometry-based proteomics. *Nature*, Nature Publishing Group, v. 422, n. 6928, p. 198–207, 2003.

AFFYMETRIX, I. Statistical algorithms description document. *Technical paper*, 2002.

AKUTSU, T.; MIYANO, S.; KUHARA, S. Algorithms for identifying boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of computational biology*, Mary Ann Liebert, Inc., v. 7, n. 3-4, p. 331–343, 2000.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Error and attack tolerance of complex networks. *nature*, Nature Publishing Group, v. 406, n. 6794, p. 378–382, 2000.

ALEXA, A.; RAHNENFUHRER, J. topgo: enrichment analysis for gene ontology. *R package version*, v. 2, n. 0, 2010.

ALEXANDER, J. et al. Biomarkers of exposure to heterocyclic amines: approaches to improve the exposure assessment. *Food and chemical toxicology*, Elsevier, v. 40, n. 8, p. 1131–1137, 2002.

ALMEIDA, V. D. et al. Câncer e agentes antineoplásicos ciclo-celular específicos e ciclo-celular não específicos que interagem com o dna: uma introdução. *Quim. Nova*, SciELO Brasil, v. 28, n. 1, p. 118–129, 2005.

ALWINE, J. C.; KEMP, D. J.; STARK, G. R. Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 12, p. 5350–5354, 1977.

AMIN, R. P. et al. Identification of putative gene based markers of renal toxicity. *Environmental health perspectives*, National Institute of Environmental Health Science, v. 112, n. 4, p. 465, 2004.

ANTONIADIS, A.; LAMBERT-LACROIX, S.; LEBLANC, F. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, Oxford University Press, v. 19, n. 5, p. 563–570, 2003.

ARMSTRONG, R. A. When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, Wiley Online Library, v. 34, n. 5, p. 502–508, 2014.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. *Nature genetics*, Nature Publishing Group, v. 25, n. 1, p. 25–29, 2000.

BARABÁSI, A.-L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, Nature Publishing Group, v. 12, n. 1, p. 56–68, 2011.

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, v. 5, n. 2, p. 101–113, Feb 2004. Disponível em: <<http://dx.doi.org/10.1038/nrg1272>>.

BEGLEY, T. J. et al. Damage recovery pathways in *saccharomyces cerevisiae* revealed by genomic phenotyping and interactome mapping. *Molecular cancer research*, American Association for Cancer Research, v. 1, p. 103–112, 2002.

BEGLEY, T. J.; SAMSON, L. D. Network responses to dna damaging agents. *DNA repair*, Elsevier, v. 3, n. 8, p. 1123–1132, 2004.

BERRA, E.; GINOUVÈS, A.; POUYSSÉGUR, J. The hypoxia-inducible-factor hydroxylases bring fresh air into hypoxia signalling. *EMBO reports*, EMBO Press, v. 7, n. 1, p. 41–45, 2006.

BIAGI-JUNIOR, C. A. O. de. Meta-análise do projeto toxicogenômico japonês: diferenças entre modelos in vivo e in vitro.

BLOOM, G. et al. Multi-platform, multi-site, microarray-based human tumor classification. *The American journal of pathology*, Elsevier, v. 164, n. 1, p. 9–16, 2004.

BODDY, A. V.; YULE, S. M. Metabolism and pharmacokinetics of oxazaphosphorines. *Clinical pharmacokinetics*, Springer, v. 38, n. 4, p. 291–304, 2000.

BOLSTAD, B. M. et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, Oxford University Press, v. 19, n. 2, p. 185–193, 2003.

BOULESTEIX, A.-L.; TUTZ, G.; STRIMMER, K. A cart-based approach to discover emerging patterns in microarray data. *Bioinformatics*, Oxford University Press, v. 19, n. 18, p. 2465–2472, 2003.

BROWN, M. P. et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 97, n. 1, p. 262–267, 2000.

BUESEN, R. et al. Applying'omics technologies in chemicals risk assessment: Report of an ecetoc workshop. *Regulatory Toxicology and Pharmacology*, Elsevier, 2017.

BUTTE, A. J.; KOHANE, I. S. Unsupervised knowledge discovery in medical databases using relevance networks. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA Symposium*. [S.l.], 1999. p. 711.

BUTTE, A. J. et al. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 97, n. 22, p. 12182–12186, 2000.

- CASTRO, M. A. A. et al. Viacomplex: software for landscape analysis of gene expression networks in genomic context. *Bioinformatics*, v. 25, n. 11, p. 1468–1469, Jun 2009. Disponível em: <<http://dx.doi.org/10.1093/bioinformatics/btp246>>.
- CHARITOU, T.; BRYAN, K.; LYNN, D. J. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, BioMed Central, v. 48, n. 1, p. 27, 2016.
- CHELLAPPAN, S. P. Hog on the promoter: regulation of the osmotic stress response. *Science Signaling*, Science Signaling, v. 2001, n. 93, p. pe1–pe1, 2001.
- CONRADT, L. et al. Mdm2 inhibitors synergize with topoisomerase ii inhibitors to induce p53-independent pancreatic cancer cell death. *International journal of cancer*, Wiley Online Library, v. 132, n. 10, p. 2248–2257, 2013.
- COUNCIL, N. R. et al. *Health risks from dioxin and related compounds: evaluation of the EPA reassessment*. [S.l.]: National Academies Press, 2006.
- DALMA-WEISZHAUSZ, D. D. et al. [1] the affymetrix genechip® platform: An overview. *Methods in enzymology*, Elsevier, v. 410, p. 3–28, 2006.
- DATTA, A. et al. External control in markovian genetic regulatory networks: the imperfect information case. *Bioinformatics*, Oxford University Press, v. 20, n. 6, p. 924–930, 2004.
- DICKINSON, D. A. et al. Differentiation of dna reactive and non-reactive genotoxic mechanisms using gene expression profile analysis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, Elsevier, v. 549, n. 1, p. 29–41, 2004.
- DOBBIN, K. K. et al. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical Cancer Research*, AACR, v. 11, n. 2, p. 565–572, 2005.
- DONG, Z. et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer*, BioMed Central, v. 15, n. 1, p. 489, 2015.
- DONIGER, S. W. et al. Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data. *Genome biology*, BioMed Central, v. 4, n. 1, p. R7, 2003.
- DRĂGHICI, S. *Statistics and data analysis for microarrays using R and bioconductor*. [S.l.]: Chapman and Hall/CRC, 2016.
- DUDOIT, S. et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, JSTOR, p. 111–139, 2002.
- DYER, M. D.; MURALI, T.; SOBRAL, B. W. The landscape of human proteins interacting with viruses and other pathogens. *PLoS pathogens*, Public Library of Science, v. 4, n. 2, p. e32, 2008.

- EISEN, M. B. et al. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 95, n. 25, p. 14863–14868, 1998.
- EKHART, C. et al. Influence of polymorphisms of drug metabolizing enzymes (cyp2b6, cyp2c9, cyp2c19, cyp3a4, cyp3a5, gsta1, gstp1, aldh1a1 and aldh3a1) on the pharmacokinetics of cyclophosphamide and 4-hydroxycyclophosphamide. *Pharmacogenetics and genomics*, LWW, v. 18, n. 6, p. 515–523, 2008.
- ELLIS, M. et al. Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. *Clinical Cancer Research*, AACR, v. 8, n. 5, p. 1155–1166, 2002.
- ERNST, R. R. et al. Principles of nuclear magnetic resonance in one and two dimensions. Clarendon Press Oxford, 1987.
- FAMILI, I.; MAHADEVAN, R.; PALSSON, B. O. k-cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophysical journal*, Elsevier, v. 88, n. 3, p. 1616–1625, 2005.
- FINKEL, T.; HOLBROOK, N. J. Oxidants, oxidative stress and the biology of ageing. *Nature*, Nature Publishing Group, v. 408, n. 6809, p. 239–247, 2000.
- FREEMAN, T. C. et al. A gene expression atlas of the domestic pig. *BMC biology*, Springer, v. 10, n. 1, p. 90, 2012.
- FRIEDMAN, N. et al. Using bayesian networks to analyze expression data. *Journal of computational biology*, Mary Ann Liebert, Inc., v. 7, n. 3-4, p. 601–620, 2000.
- GALAS, D. J.; MCCORMACK, S. J. An historical perspective on genomic technologies. 2003.
- GAUTIER, L. et al. affy-analysis of affymetrix genechip data at the probe level. *Bioinformatics*, Oxford Univ Press, v. 20, n. 3, p. 307–315, 2004.
- GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, BioMed Central, v. 5, n. 10, p. 1, 2004.
- GIAEVER, G. et al. Functional profiling of the saccharomyces cerevisiae genome. *nature*, Nature Publishing Group, v. 418, n. 6896, p. 387, 2002.
- GOLUB, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, American Association for the Advancement of Science, v. 286, n. 5439, p. 531–537, 1999.
- GRIFFITHS, A. J. *An introduction to genetic analysis*. [S.l.]: Macmillan, 2005.
- HAMADEH, H. K. et al. Prediction of compound signature using high density gene expression profiling. *Toxicological Sciences*, Oxford University Press, v. 67, n. 2, p. 232–240, 2002.

- HANDE, M. P. Book review: Toxicogenomics in predictive carcinogenicity. *Frontiers in Genetics*, Frontiers, v. 8, p. 11, 2017.
- HARTWELL, L. H. et al. From molecular to modular cell biology. *Nature*, Nature Publishing Group, v. 402, p. C47–C52, 1999.
- HASHIMOTO, R. F. et al. Growing genetic regulatory networks from seed genes. *Bioinformatics*, Oxford University Press, v. 20, n. 8, p. 1241–1247, 2004.
- HERRERO, J.; VALENCIA, A.; DOPAZO, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, Oxford University Press, v. 17, n. 2, p. 126–136, 2001.
- HOEN, P. A. t et al. Fluorescent labelling of crna for microarray applications. *Nucleic acids research*, Oxford University Press, v. 31, n. 5, p. e20–e20, 2003.
- HOSACK, D. A. et al. Identifying biological themes within lists of genes with ease. *Genome biology*, BioMed Central, v. 4, n. 10, p. R70, 2003.
- HUANG, Z.; ROY, P.; WAXMAN, D. J. Role of human liver microsomal cyp3a4 and cyp2b6 in catalyzing n-dechloroethylation of cyclophosphamide and ifosfamide. *Biochemical pharmacology*, Elsevier, v. 59, n. 8, p. 961–972, 2000.
- HUMANS, I. W. G. on the Evaluation of Carcinogenic Risks to et al. Pharmaceuticals. volume 100 a. a review of human carcinogens. *IARC monographs on the evaluation of carcinogenic risks to humans*, Various, v. 100, n. PT A, p. 1, 2012.
- IARC et al. *Overall evaluations of carcinogenicity: an updating of IARC monographs volumes 1 to 42*. [S.l.]: IARC Lyon, 1987.
- IMOTO, S. et al. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of bioinformatics and computational biology*, World Scientific, v. 1, n. 02, p. 231–252, 2003.
- IRIZARRY, R. A. et al. Multiple-laboratory comparison of microarray platforms. *Nature methods*, Nature Publishing Group, v. 2, n. 5, p. 345, 2005.
- JANSSEN, A. W. et al. The impact of ppar α activation on whole genome gene expression in human precision cut liver slices. *BMC genomics*, BioMed Central, v. 16, n. 1, p. 760, 2015.
- JELINSKY, S. A. et al. Regulatory networks revealed by transcriptional profiling of damaged saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. *Molecular and cellular biology*, Am Soc Microbiol, v. 20, n. 21, p. 8157–8167, 2000.
- JELINSKY, S. A.; SAMSON, L. D. Global response of saccharomyces cerevisiae to an alkylating agent. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 96, n. 4, p. 1486–1491, 1999.

JENSEN, L. J. et al. String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, Oxford Univ Press, v. 37, n. suppl 1, p. D412–D416, 2009.

JIANG, N. et al. Methods for evaluating gene expression from affymetrix microarray datasets. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 284, 2008.

JONSSON, P. F.; BATES, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, Oxford University Press, v. 22, n. 18, p. 2291–2297, 2006.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 27–30, 2000.

KILLICK, K. E. et al. Key hub and bottleneck genes differentiate the macrophage response to virulent and attenuated mycobacterium bovis. *Frontiers in immunology*, Frontiers Media SA, v. 5, 2014.

KIM, S. K. et al. A gene expression map for caenorhabditis elegans. *Science*, American Association for the Advancement of Science, v. 293, n. 5537, p. 2087–2092, 2001.

KING, C. et al. Udp-glucuronosyltransferases. *Current drug metabolism*, Bentham Science Publishers, v. 1, n. 2, p. 143–161, 2000.

LAGE, K. et al. Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 109, n. 35, p. 14035–14040, 2012.

LARKIN, J. E. et al. Independence and reproducibility across microarray platforms. *Nature methods*, Nature Publishing Group, v. 2, n. 5, p. 337, 2005.

LAWLESS, N. et al. Microrna regulation of bovine monocyte inflammatory and metabolic networks in an in vivo infection model. *G3: Genes, Genomes, Genetics*, G3: Genes, Genomes, Genetics, v. 4, n. 6, p. 957–971, 2014.

LE, Q.-T. et al. Identification of osteopontin as a prognostic plasma marker for head and neck squamous cell carcinomas. *Clinical Cancer Research*, AACR, v. 9, n. 1, p. 59–67, 2003.

LEAR, L.; NATION, R.; STUPANS, I. Effects of cyclophosphamide and adriamycin on rat hepatic microsomal glucuronidation and lipid peroxidation. *Biochemical pharmacology*, Elsevier, v. 44, n. 4, p. 747–753, 1992.

LEE, F. Y. et al. Clinical pharmacokinetics of oral ccnu (lomustine). *Cancer chemotherapy and pharmacology*, Springer, v. 14, n. 2, p. 125–131, 1985.

LEHMANN, J. M. et al. The human orphan nuclear receptor pxx is activated by compounds that regulate cyp3a4 gene expression and cause drug interactions. *The Journal of clinical investigation*, Am Soc Clin Investig, v. 102, n. 5, p. 1016–1023, 1998.

- MAANEN, J. V. et al. Mechanism of action of antitumor drug etoposide: a review. *JNCI: Journal of the National Cancer Institute*, Oxford University Press, v. 80, n. 19, p. 1526–1533, 1988.
- MALIK, A. et al. Network analysis for the identification of differentially expressed hub genes using myogenin knock-down muscle satellite cells. *PloS one*, Public Library of Science, v. 10, n. 7, p. e0133597, 2015.
- MANFREDI, J. J. The mdm2–p53 relationship evolves: Mdm2 swings both ways as an oncogene and a tumor suppressor. *Genes & development*, Cold Spring Harbor Lab, v. 24, n. 15, p. 1580–1589, 2010.
- MARSELOS, M.; VAINIO, H. Carcinogenic properties of pharmaceutical agents evaluated in the iarc monographs programme. *Carcinogenesis*, Oxford University Press, v. 12, n. 10, p. 1751–1766, 1991.
- MEDEIROS, A.; BIRD, M.; WITZ, G. Potential biomarkers of benzene exposure. *Journal of toxicology and environmental health*, Taylor & Francis, v. 51, n. 6, p. 519–539, 1997.
- MERING, C. V. et al. String 7—recent developments in the integration and prediction of protein interactions. *Nucleic acids research*, Oxford Univ Press, v. 35, n. suppl 1, p. D358–D362, 2007.
- MERING, C. V. et al. String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research*, Oxford Univ Press, v. 33, n. suppl 1, p. D433–D437, 2005.
- MONTESANO, R.; PEGG, A. E.; MARGISON, G. P. Alkylation of dna and carcinogenicity of n-nitroso compounds. *Journal of Toxicology and Environmental Health, Part A Current Issues*, Taylor & Francis, v. 6, n. 5-6, p. 1001–1008, 1980.
- NACU, Ş. et al. Gene expression network analysis and applications to immunology. *Bioinformatics*, Oxford University Press, v. 23, n. 7, p. 850–858, 2007.
- NETZER, N. et al. Hypoxia, oxidative stress and fat. *Biomolecules*, Multidisciplinary Digital Publishing Institute, v. 5, n. 2, p. 1143–1150, 2015.
- NGUYEN, D. V.; ROCKE, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, Oxford University Press, v. 18, n. 1, p. 39–50, 2002.
- NIELSEN, H. B.; GAUTIER, L.; GAUTIER, M. L. Package affypdnn.
- NRC et al. *Applications of toxicogenomic technologies to predictive toxicology and risk assessment*. [S.l.]: National Academies Press (US), 2007.
- ORR, M.; SCHERF, U. Large-scale gene expression analysis in molecular target discovery. *Leukemia*, Nature Publishing Group, v. 16, n. 4, p. 473, 2002.

- PEGG, A. E. Repair of o 6-alkylguanine by alkyltransferases. *Mutation Research/Reviews in Mutation Research*, Elsevier, v. 462, n. 2, p. 83–100, 2000.
- PERERA, F.; SANTELLA, R.; POIRIER, M. Biomonitoring of workers exposed to carcinogens: immunoassays to benzo (a) pyrene-dna adducts as a prototype. *Journal of Occupational and Environmental Medicine*, LWW, v. 28, n. 10, p. 1117–1123, 1986.
- PETRICOIN, E. F. et al. Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, Oxford University Press, v. 94, n. 20, p. 1576–1578, 2002.
- PETROS, W. P. et al. Associations between drug metabolism genotype, chemotherapy pharmacokinetics, and overall survival in patients with breast cancer. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 23, n. 25, p. 6117–6125, 2005.
- RAMASWAMY, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 26, p. 15149–15154, 2001.
- RAYCHAUDHURI, S.; STUART, J. M.; ALTMAN, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. In: NIH PUBLIC ACCESS. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. [S.l.], 2000. p. 455.
- RUDEN, D. M.; GURDZIEL, K.; ASCHNER, M. Frontiers in toxicogenomics in the twenty-first century, the grand challenge: To understand how the genome and epigenome interact with the toxic environment at the single-cell, whole-organism, and multi-generational level. *Frontiers in Genetics*, Frontiers, v. 8, p. 173, 2017.
- SAID, M. R. et al. Global network analysis of phenotypic effects: protein networks and toxicity modulation in *saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 101, n. 52, p. 18006–18011, 2004.
- SAVOIE, C. J. et al. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. *Dna Research*, Oxford University Press, v. 10, n. 1, p. 19–25, 2003.
- SCHAEFER, M. H.; SERRANO, L.; ANDRADE-NAVARRO, M. A. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Frontiers in genetics*, Frontiers Media SA, v. 6, 2015.
- SCHLEKER, S.; TRILLING, M. Data-warehousing of protein-protein interactions indicates that pathogens preferentially target hub and bottleneck proteins. *Frontiers in microbiology*, Frontiers Media SA, v. 4, 2013.
- SCHMÄHL, D.; HABS, M. Carcinogenic action of low-dose cyclophosphamide given orally to sprague-dawley rats in a lifetime experiment. *International journal of cancer*, Wiley Online Library, v. 23, n. 5, p. 706–712, 1979.

- SCHULZE, A.; DOWNWARD, J. Navigating gene expression using microarrays—a technology review. *Nature cell biology*, Nature Publishing Group, v. 3, n. 8, p. E190, 2001.
- SELVAKUMAR, E. et al. Mitigation of oxidative stress in cyclophosphamide-challenged hepatic tissue by dl- α -lipoic acid. *Molecular and cellular biochemistry*, Springer, v. 272, n. 1-2, p. 179–185, 2005.
- SHMULEVICH, I. et al. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, Oxford University Press, v. 18, n. 2, p. 261–274, 2002.
- SHUGART, L.; HOLLAND, J.; RANN, R. Dosimetry of pah skin carcinogenesis: covalent binding of benzo [a] pyrene to mouse epidermal dna. *Carcinogenesis*, Oxford University Press, v. 4, n. 2, p. 195–198, 1983.
- SKIPPER, P. L. et al. Protein adducts as biomarkers of human carcinogen exposure. *Drug metabolism reviews*, Taylor & Francis, v. 26, n. 1-2, p. 111–124, 1994.
- 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.
- SOINOV, L. *Supervised classification for gene network reconstruction*. [S.l.]: Portland Press Limited, 2003.
- SOUKAS, A. et al. Leptin-specific patterns of gene expression in white adipose tissue. *Genes & development*, Cold Spring Harbor Lab, v. 14, n. 8, p. 963–980, 2000.
- SOUZA, T. M.; KLEINJANS, J.; JENNEN, D. G. Dose and time dependencies in stress pathway responses during chemical exposure: Novel insights from gene regulatory networks. *Frontiers in genetics*, Frontiers, v. 8, p. 142, 2017.
- SUMIDA, K. et al. Effects of dmso on gene expression in human and rat hepatocytes. *Human & experimental toxicology*, Sage Publications Sage UK: London, England, v. 30, n. 10, p. 1701–1709, 2011.
- SZKLARCZYK, D. et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, Oxford Univ Press, v. 39, n. suppl 1, p. D561–D568, 2011.
- SZKLARCZYK, D. et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, Oxford University Press, v. 43, n. D1, p. D447–D452, 2014.
- TAKASHIMA, K. et al. Effect of the difference in vehicles on gene expression in the rat liver—analysis of the control data in the toxicogenomics project database. *Life sciences*, Elsevier, v. 78, n. 24, p. 2787–2796, 2006.

TAMADA, Y. et al. Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection. *Bioinformatics*, Oxford University Press, v. 19, n. suppl_2, p. ii227–ii236, 2003.

TAMAYO, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 96, n. 6, p. 2907–2912, 1999.

THEILHABER, J. et al. Finding genes in the c2c12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 1, p. 165–176, 2002.

TÖRÖNEN, P. et al. Analysis of gene expression data using self-organizing maps. *FEBS letters*, Wiley Online Library, v. 451, n. 2, p. 142–146, 1999.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 9, p. 5116–5121, 2001.

UEHARA, T. et al. Species-specific differences in coumarin-induced hepatotoxicity as an example toxicogenomics-based approach to assessing risk of toxicity to humans. *Human & experimental toxicology*, Sage Publications Sage UK: London, England, v. 27, n. 1, p. 23–35, 2008.

UEHARA, T. et al. Toxicogenomic biomarkers for renal papillary injury in rats. *Toxicology*, Elsevier, v. 303, p. 1–8, 2013.

UEHARA, T. et al. Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database. *Toxicology and applied pharmacology*, Elsevier, v. 255, n. 3, p. 297–306, 2011.

UEHARA, T. et al. The japanese toxicogenomics project: application of toxicogenomics. *Molecular nutrition & food research*, Wiley Online Library, v. 54, n. 2, p. 218–227, 2010.

VEAL, G. J. et al. Cyclophosphamide pharmacokinetics and pharmacogenetics in children with b-cell non-hodgkin's lymphoma. *European Journal of Cancer*, Elsevier, v. 55, p. 56–64, 2016.

VEZINA, C. M.; WALKER, N. J.; OLSON, J. R. Subchronic exposure to tcdd, pecdf, pcb126, and pcb153: effect on hepatic gene expression. *Environmental health perspectives*, National Institute of Environmental Health Science, v. 112, n. 16, p. 1636, 2004.

WANG, J. et al. Clustering of the som easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC bioinformatics*, BioMed Central, v. 3, n. 1, p. 36, 2002.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, Nature Publishing Group, v. 10, n. 1, p. 57–63, 2009.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

WEINSTEIN, J. N. et al. An information-intensive approach to the molecular pharmacology of cancer. *Science*, American Association for the Advancement of Science, v. 275, n. 5298, p. 343–349, 1997.

WEIS, B. Standardizing global gene expression analysis between laboratories and across platforms. *Nature methods*, Nature Publishing Group, v. 2, n. 5, p. 351, 2005.

WEIZMANN. *GeneCards: Human Gene Database*. jun. 2018. Disponível em: <[www-genecards.org](http://www.genecards.org)>.

WEN, X. et al. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 95, n. 1, p. 334–339, 1998.

WIBACK, S. J.; MAHADEVAN, R.; PALSSON, B. Ø. Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the escherichia coli spectrum. *Biotechnology and bioengineering*, Wiley Online Library, v. 86, n. 3, p. 317–331, 2004.

WU, H. et al. High-throughput tissue extraction protocol for nmr-and ms-based metabolomics. *Analytical biochemistry*, Elsevier, v. 372, n. 2, p. 204–212, 2008.

WU, Y. et al. Long noncoding rna eosinophil granule ontogeny transcript inhibits cell proliferation and migration and promotes cell apoptosis in human glioma. *Experimental and therapeutic medicine*, Spandidos Publications, v. 14, n. 4, p. 3817–3823, 2017.

WU, Z. et al. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American statistical Association*, Taylor & Francis, v. 99, n. 468, p. 909–917, 2004.

XENARIOS, I. et al. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, Oxford University Press, v. 30, n. 1, p. 303–305, 2002.

XU, S.-p. et al. Downregulation of the long noncoding rna egot correlates with malignant status and poor prognosis in breast cancer. *Tumor Biology*, Springer, v. 36, n. 12, p. 9807–9812, 2015.

YU, H. et al. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, Public Library of Science, v. 3, n. 4, p. e59, 2007.

ZEEBERG, B. R. et al. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, BioMed Central, v. 4, n. 4, p. R28, 2003.

ZEMPLICKIS, D. et al. Teratogenicity and carcinogenicity in a twin exposed in utero to cyclophosphamide. *Teratogenesis, carcinogenesis, and mutagenesis*, Wiley Online Library, v. 13, n. 3, p. 139–143, 1993.

ZOU, M.; CONZEN, S. D. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, Oxford University Press, v. 21, n. 1, p. 71–79, 2004.