UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Carlos Alberto Oliveira de Biagi Júnior

Meta-análise do Projeto Toxicogenômico Japonês: diferenças entre modelos *in vivo* e *in vitro*

Botucatu, Julho de 2017

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Carlos Alberto Oliveira de Biagi Júnior

Meta-análise do Projeto Toxicogenômico Japonês: diferenças entre modelos *in vivo* e *in vitro*

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 Biotecnologia.

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Julho de 2017.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Biagi Júnior, Carlos Alberto Oliveira de.
Meta-análise do Projeto Toxicogenômico Japonês : diferenças entre modelos *in vivo* e *in vitro* / Carlos Alberto Oliveira de Biagi Júnior. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Ribarczyk Filho
Capes: 90400003

1. Toxicogenética. 2. Bioinformática. 3. Análise de microarranjo. 4. Drogas - Testes. 5. Técnicas *in vitro*. 6. Meta-análises.

Palavras-chave: Bioinformática; Microarranjo; Toxicogenômica.

Agradecimentos

À CNPq, processo 473789/2013-2, pela disponibilização dos computadores utilizados neste trabalho.

Ao Programa de Pós-Graduação da Biotecnologia, Universidade Estadual Paulista "Júlio de Mesquita Filho"(UNESP), Instituto de Biociências de Botucatu (IBB) e Instituto de Biotecnologia de Botucatu (IBTEC) pelo acolhimento nestes anos e por proporcionar a oportunidade de estudar e desenvolver pesquisa em um dos melhores locais do Brasil.

À meu orientador, Prof. Dr. José Luiz Rybarczyk Filho, a quem posso chamar sem dúvidas como amigo, por ter me mostrado que um orientador não é um ser a ser temido, e sim uma pessoa que, além de te guiar academicamente, está aberto a escutar suas ânsias, desejos e medos.

À Agnes Alessandra Sekijima Takeda, por toda a colaboração e ensinamentos que levarei pelo restante da minha vida.

À banca de qualificação, Prof. Dr. João Pessoa Araújo Júnior e Prof[a]. Dr[a]. Valéria Cristina Sandrim, pelas excelentes sugestões e críticas ao meu trabalho.

Aos meus mais que amigos e companheiros de laboratório: Alex, André, Giordano, Jéssica e José Rafael. Em especial aos amigos André e Giordano pelo companheirismo e amizade verdadeira demonstrada nesses anos. À vocês meu muito obrigado por me inspirarem, aconselharem e sempre estarem presentes nos momentos bons ou ruins. Levarei para sempre essa amizade.

À minha preciosa família, Carlos e Selma, minha irmã Natália, minha namorada Karen e meu cunhado Filipe. Aqueles pelos quais acordo toda manhã e sigo em frente sem medo, pois sei que, mais que quaisquer outras pessoas, me fizeram acreditar que eu posso ter tudo aquilo que quiser, basta não desistir. As pessoas pelas quais eu posso dizer que tenho verdadeiro amor.

À Deus por me conceder forças, disposição e tudo que sempre precisei até aqui.

Aos servidores(as) das instituições citadas acima. Muito obrigado pelas boas conversas, risadas e conselhos.

Por fim, aos professores e amigos que me encorajaram e me inspiraram a trilhar essa caminhada.

Resumo

A toxicogenômica é um campo emergente que possibilita o estudo dos efeitos de uma determinada droga a nível molecular em sistemas modelos. Uma das principais questões é se podemos substituir os estudos *in vivo* pelos estudos *in vitro*. As ciências ômicas possibilitam a resposta para esse tipo de questionamento pois fornecem técnicas como, por exemplo, o *microarray*, que permite o conhecimento dos transcritos (RNAs) de um dado organismo. O Projeto Toxicogenômico Japonês fornece dados para *Homo sapiens*, com somente experimentos *in vitro*, e para *Rattus norvegicus*, com experimentos *in vitro* e *in vivo*, tratados com 131 drogas (aprovadas pelo *FDA*) em diferentes concentrações de dose e tempos de amostragem, totalizando, aproximadamente, 20000 *chips* de *microarray*. A partir desses dados foi possível responder a questão inicial e observar as diferenças existentes entre cada modelo. Por meio da linguagem de programação R normalizamos os dados, obtemos os genes diferencialmente expressos e o respectivo enriquecimento funcional, dessa forma observamos as diferenças entre cada modelo. Em seguida realizamos uma análise comparativa dos modelos *in vivo* e *in vitro* adaptando a metodologia do mapa modular proposta por *Segal* e colaboradores. Essa metodologia tem como objetivo principal obter módulos, que são *sets* de genes (dados do *Gene Ontology*, *KEGG* e *Reactome*) que agem em conjunto para realizar uma função específica. Além de extrair módulos caracterizaremos os valores de expressão em relação as condições clínicas fornecidas pelo Projeto Toxicogenômico Japonês. Com base nessas informações do mapa modular foi possível identificar quais condições estão enriquecidas para um determinado conjunto de *sets* de genes, ou seja, quais processos biológico ou rotas metabólicas estão alteradas em condições específicas. Neste trabalho foi possível identificar diferenças entre os modelos *in vitro* e *in vivo* para *Homo sapiens* e *Rattus norvegicus* por meio da metodologia do mapa modular, avaliamos a quantidade de genes diferencialmente expressos e o enriquecimento funcional para diferentes concentrações de dose e diferentes tempos de amostragem. Concluímos que não é possível substituir os estudos *in vivo* pelo *in vitro* a partir dos dados analisados.

Abstract

Toxicogenomics is an emerging field that allows the study of the effects of a given drug at the molecular level in model systems. One of the key issues is whether we can replace *in vivo* studies with *in vitro* studies. The omics sciences enable researchers to address this problem because it provides techniques, for example the microarray, that allow the knowledge of the transcripts (RNAs) of a given organism. The Japanese Toxicogenomic Project provides data for *Homo sapiens*, with only *in vitro* experiments, and for *Rattus norvegicus*, with *in vitro* and *in vivo* experiments, treated with 131 drugs (FDA approved) at different dose concentrations and sampling times, totaling approximately 20,000 microarray chips. From these data it was possible to answer the initial question and to observe the differences between each model. Using the programming language R we normalized the data, obtained the differentially expressed genes and their respective functional enrichment, so that we could observe the differences between each model. We then performed a comparative analysis of the *in vivo* and *in vitro* models by adapting the modular map methodology proposed by *Segal et al*. The main objective of this methodology is to obtain modules, which are gene sets (Gene Ontology, KEGG and Reactome data) that act together to perform a specific function. In addition to extracting modules we will characterize the expression values in relation to the clinical conditions provided by the Japanese Toxicogenomic Project. Based on the information provided by the modular map it was possible to identify which conditions are enriched for a given set of genes, or in other words, what biological processes or metabolic pathways are altered under specific conditions. In this work it was possible to identify differences between *in vitro* and *in vivo* models for *Homo sapiens* and *Rattus norvegicus* using the modular map methodology, we evaluated the number of differentially expressed genes and the functional enrichment for different dose concentrations and different sampling times. We conclude that it is not possible to replace the *in vivo* studies by *in vitro* from the analyzed data.

Lista de Abreviações

Anvisa	Agência Nacional de Vigilância Sanitária
CAMDA	*Critical Assessment of Massive Data Analysis*
FDA	*Food and Drug Administration*
GO	*Gene Ontology*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MAS5	*Microarray Analysis Suite 5*
MM	*Mismatch*
NIH	*National Institutes of Health*
PM	*Perfect match*
PTGJ	Projeto Toxicogenômico Japonês
Reactome	*Reactome Pathway*
RMA	*Robust MultiArray*
TGPJ	*Toxicogenomics Project in Japan*

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Experimentos *in vivo* e *in vitro*

Atualmente há uma grande demanda de estudos que envolvem a comparação de duas ou mais explicações de um certo fenômeno biológico. Para esse tipo de estudo faz-se necessário utilizar alguns métodos. Entre os métodos mais utilizados estão os estudos *in vivo* e *in vitro*, a fim de comprovar se a hipótese em questão é válida ou não (POLLI, 2008).

O estudo *in vivo* é o experimento ou observações realizadas sobre o tecido em um organismo vivo em um ambiente controlado. Um exemplo é o teste ou ensaio clínico, que pode ser um teste controlado de uma nova droga ou dispositivo em seres humanos. As drogas são administradas a indivíduos que permanecem em observação durante um período. Outro exemplo é a experimentação animal. Os experimentos *in vivo* apresentam custos mais elevados, além de estarem sujeitos a várias restrições em função de se tratar do uso de seres vivos (POLLI, 2008).

Por outro lado, o estudo *in vitro* é o experimento ou observações realizadas no tecido vivo, num ambiente controlado, geralmente usando placas de Petri e tubos de ensaio. A maioria dos experimentos em biologia celular são feitos através de estudos *in vitro* e não são realizados no ambiente natural do organismo. Os resultados desses experimentos são limitados, pois trata-se de uma simulação das condições reais de um organismo e ,em comparação com os experimentos *in vivo*, são mais baratos e fornecem resultados mais rápidos (LODISH et al., 1995).

Os estudos *in vitro* e *in vivo* são muito importantes quando se trata de desenvolvimento de drogas. Cada país possui legislações específicas que guiam as indústrias farmacêuticas e pesquisadores nesse processo.

O desenvolvimento de uma droga no Brasil é regulamentado pela Anvisa, enquanto nos Estados Unidos o órgão responsável é o *FDA*. A diferença entre eles está na rigidez da legislação de cada país. No Brasil, as leis são menos flexíveis quanto ao desenvolvimento de uma droga, ou seja, demanda-se mais tempo para a sua produção em relação aos Estados Unidos. Para desenvolver uma droga são necessários cinco passos, executados, obrigatoriamente, na seguinte

ordem:

1. **Descoberta e Desenvolvimento:** nessa fase do processo, milhares de compostos podem ser potenciais candidatos para o desenvolvimento de um tratamento médico. Após os primeiros testes, no entanto, apenas um pequeno número de compostos parecem promissores e exigem um estudo mais aprofundado;

2. **Avaliação Ética e Pesquisa Pré-Clínica:** antes de testar uma droga em sujeitos de pesquisa, os pesquisadores devem descobrir se ela possui potencial de causar danos graves - também chamado de toxicidade. Dessa forma, as drogas são submetidas a testes laboratoriais e ministradas em animais com o intuito de responder à perguntas básicas sobre segurança (testes *in vitro* e *in vivo*);

3. **Pesquisa Clínica:** embora a investigação pré-clínica responda perguntas básicas sobre segurança de uma droga, ela não substitui estudos que mostram as formas que a droga irá interagir com o corpo humano. A pesquisa clínica refere-se a estudos ou ensaios, que são feitos em pessoas;

4. **Revisão da *FDA*:** se o pesquisador possui provas que seus primeiros testes e pesquisa pré-clínica e clínica de que um medicamento é seguro e eficaz para o uso, a empresa pode apresentar um pedido para comercializar a droga. A equipe de revisão da FDA examina minuciosamente todos os dados apresentados sobre a droga e toma a decisão de aprovar ou não;

5. **Monitoramento de Segurança:** embora os ensaios clínicos forneçam informações importantes sobre a eficácia e segurança de uma droga, é impossível ter informações completas sobre a segurança de um medicamento no momento da aprovação. Portanto, há um monitoramento do medicamento uma vez que o produto está disponível para utilização pelo público.

Ambos os modelos experimentais, *in vitro* e *in vivo*, são primordiais no processo de desenvolvimento de uma droga. A partir da Figura 1.1 podemos destacar algumas fases, como, por exemplo, a fase dos *testes pré clínicos*, que envolve testes laboratoriais em animais para responder perguntas básicas sobre toxicidade e segurança de determinada droga; durante essa fase, são utilizados experimentos *in vitro*. Também destacamos a fase da *pesquisa clínica*, na qual as drogas são testadas *in vivo* para se certificar que são seguras e eficazes.

Os modelos animais *in vitro* e *in vivo* são essenciais na transição da fase pré clínica para a clínica. Caso haja predominância nos estudos *in vitro*, espera-se que as conclusões sobre uma

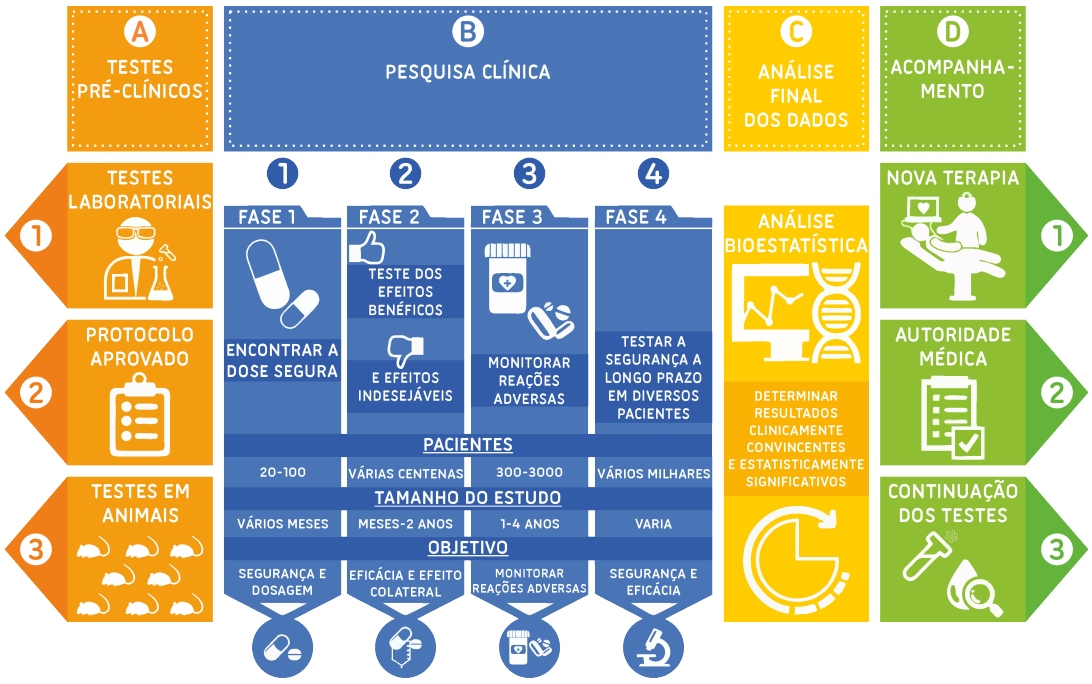

Figura 1.1: O processo de desenvolvimento de uma droga, como mostrado nesta figura, passa por alguns passos que são: testes pré-clínicos (**A**), pesquisa clínica (**B**), análise final dos dados (**C**) e por fim através do acompanhamento (**D**). Cada passo possui suas particularidades e são de suma importância para, no final, uma droga ser autorizada para produção em larga escala. Figura adaptada de `https://vigyanix.com/blog/how-do-clinical-trials-work-from-trial-to-treatment/`

droga específica seja baseada neste mesmo modelo, diminuindo, assim, a utilização de estudos *in vivo* (DENAYER; STÖHR; ROY, 2014).

1.2 As Ciências Ômicas

Nas últimas décadas houve um aumento do número de projetos de sequenciamento, como, por exemplo o Projeto Genoma Humano; esses projetos levaram à otimização e desenvolvimento de novas técnicas, as quais possibilitaram o estudo de processos celulares e moleculares e permitiram maior compreensão dos sistemas biológicos. Entretanto, os organismos atuam como compartimentos moleculares isolados e a única maneira de estudá-los é colocando-os em forma de sistemas. Com isso, é possível ter uma visão global dos processos biológicos. Essas técnicas são denominadas por "ômicas", que são compostas pela genômica, transcriptômica, proteômica e metabolômica (Tabela 1.1), e têm como base a análise de um grande volume de dados (TOXICOLOGY et al., 2007) sendo, para isso, necessário o uso da bioinformática, que permite integrar os dados de forma rápida e com alto rendimento (ESPINDOLA et al., 2010).

Tabela 1.1: As ciências ômicas e suas definições

Ômicas	Definição
Genômica	Estuda o genoma completo de um organismo. Essa ciência pode se dedicar a determinar a seqüência completa do DNA de organismos ou apenas o mapeamento de uma escala genética menor;
Transcriptômica	Permite a análise de mudanças no transcriptoma completo através de uma variedade de condições biológicas;
Proteômica	Envolve o estudo em larga escala das proteínas expressas em uma célula, tecido ou organismo, incluindo a análise quantitativa da expressão ao longo do tempo, em diversas localizações celulares e/ou sob a influência de diferentes estímulos. É complementar ao genoma, pois os genes podem ser transcritos em RNA;
Metabolômica	É o estudo científico que visa identificar e quantificar o conjunto de metabólitos - o metaboloma - produzidos e/ou modificados por um organismo.

1.3 Toxicogenômica

Através da toxicologia clássica, os potenciais efeitos adversos resultantes da exposição à drogas são avaliados por meio de parâmetros como alterações corporais, peso dos órgãos e observações histopatológicas e bioquímicas. Essas observações não fornecem informações sobre o modo de ação da droga. Para melhor avaliar os efeitos adversos associados à sua exposição, precisamos entender o modo de ação específico de cada delas. Com o surgimento de novas tecnologias, foi criada a Toxicogenômica, que através da aplicação das ciências ômicas, é capaz de gerar um melhor entendimento de mecanismos farmacológicos e toxicológicos comparados com a toxicologia clássica (WATERS; FOSTEL, 2004).

A Toxicogenômica é um campo emergente, no qual a elucidação de mecanismos de toxicidade e predição de toxicidade são baseados na compreensão dos dados de expressão gênica, a partir de animais ou células de cultura expostos à drogas ou químicos. A toxicogenômica trabalha com duas estratégias (KANNO, 2003):

- **Toxicologia avançada:** elucida o mecanismo de toxicidade com base nas alterações de expressão gênica resultantes da toxicidade;
- **Toxicologia reversa:** prediz a toxicidade baseado na comparação da alteração da expressão gênica causado por químicos ou drogas tóxicas conhecidas.

Cada ciência ômica tem uma tecnologia que a auxilia em sua pesquisa e desenvolvimento. Por exemplo, a Transcriptômica possibilita o uso do *microarray* para experimentos de análise de expressão gênica em larga escala. Outros exemplos de tecnologias usadas nas ciências ômicas são apresentadas na Figura 1.2.

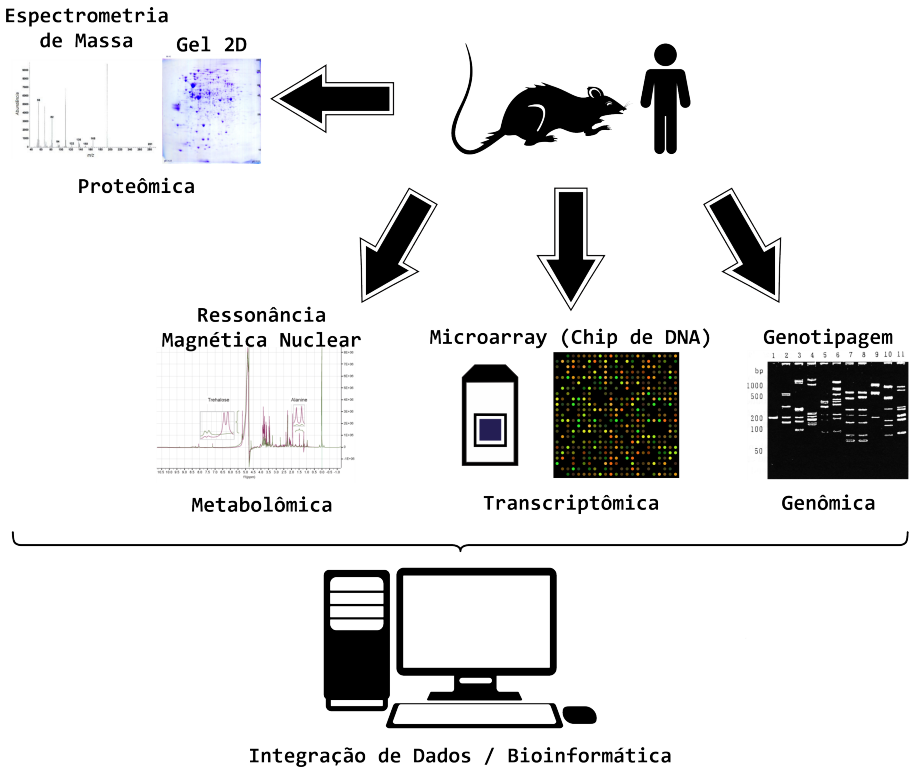

Figura 1.2: Integração das Ciências Ômicas e suas respectivas principais tecnologias.

1.4 Microarranjo

Como comentado nas seções anteriores, atualmente há uma grande quantidade de dados genômicos, ocasionando enorme demanda por tecnologias e métodos que viabilizam o processamento e a análise dos dados de forma eficiente e com elevado grau de confiabilidade. Uma das técnicas utilizadas é a de *microarray* (SCHENA et al., 1995), que proporciona o estudo da expressão gênica perante diversas condições a um baixo custo e tempo. Um experimento de *microarray* produz como resultados imagens de expressão gênica a partir das quais é possível identificar e quantificar os dados biológicos (BRAZMA et al., 2001).

A expressão gênica corresponde ao processo em que a informação codificada em um determinado gene é decodificada. Esse processo pode tanto dar origem a uma proteína como simplesmente controlar a expressão de outros genes (regulação). A síntese proteica é realizada em dois passos. O primeiro refere-se ao processo de transcrição, que corresponde a formação

de uma molécula de RNA mensageiro (RNAm) a partir de uma molécula molde de DNA. O segundo compreende o processo de tradução, que transformará o RNAm em proteína ou em parte dela (aminoácido) (OLSON, 2006) como pode ser observado na Figura 1.3.

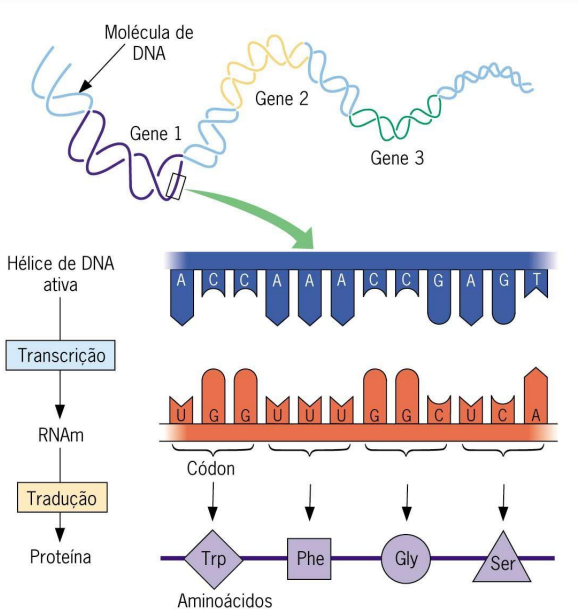

Figura 1.3: Correspondência entre as unidades do DNA e do RNA e os aminiácidos da proteína a ser sintetizada (JÚNIOR; SASSON, 2005).

O nível de expressão gênica é baseado na quantidade de RNAm associado a um gene. As técnicas mais utilizadas atualmente para análise de expressão envolvem a etapa de transcrição.

Os *microarrays* são utilizados como técnica para o estudo da expressão gênica. Desde 1995, quando Schena e colaboradores (SCHENA et al., 1995) a usaram pela primeira vez, a fim de proporcionar a análise do genoma de um organismo eucariótico (*Saccharomyces cerevisiae*), a tecnologia passou a ser amplamente utilizada em experimentos de análise de expressão gênica em larga escala.

Para a realização de um experimento de *microarray*, primeiramente é necessário duas amostras de células cultivadas em soluções distintas: a primeira correspondendo à situação a ser estudada e a segunda à situação controle (normal). Em seguida, faz-se o isolamento do RNA e extrai-se o RNAm das duas amostras. A partir da transcriptase reversa do RNAm é possível obter uma molécula de DNA mais estável, chamada de cDNA. Marca-se, então, o cDNA obtido, com uma substância fluorescente que normalmente são os corantes *cy3* (verde) e *cy5*(vermelho). Os cDNA marcados são chamados de *spots* (sondas) e vão representar as amostras microscópicas depositadas na superfície para atuar como detectores dos genes expressos. Os cDNA são misturados e aplicados nos *microarray*. A partir desse processo ocorrerá a hibridização dos *microarray* com a mistura de cDNA, ou seja, duas sequências complemen-

tares de DNA vão combinar (KNUDSEN, 2005). Todo esse processo citado acima pode ser observado na Figura 1.4.

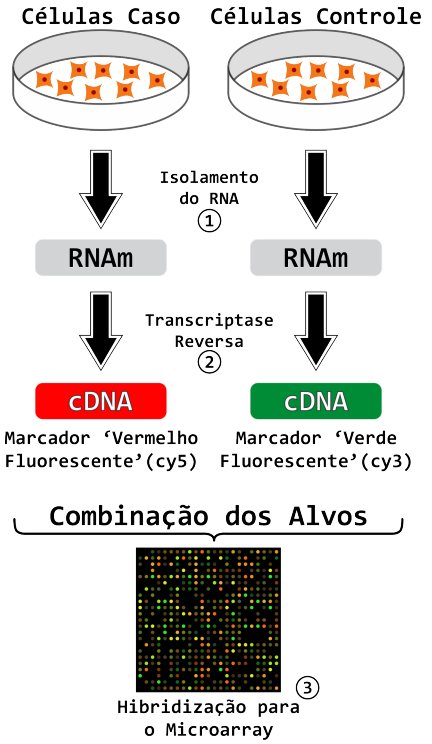

Figura 1.4: Realização de um experimento de microarranjo para amostras de células caso e células controle. Inicialmente são coletadas céluas do caso e do controle. Em seguida é feito o isolamento do RNA, sendo obtido o RNA mensageiro (RNAm). A partir do RNAm e com a utilização da transcriptase reversa é obtido o DNA complementar (cDNA). Por fim, ocorre a combinação dos alvos e a hibridização para o microarranjo.

O processo de hibridização (KOLTAI; WEINGARTEN-BAROR, 2008) é a base do experimento de *microarray*. Somente os fragmentos em que ocorreram hibridização, ou seja, fragmentos que tiverem sequências complementares de DNA, apresentam níveis de expressão. Utilizando-se um comprimento de luz adequado, é possível visualizar o material fluorescente contido no *microarray* hibridizado. As imagens são geradas a partir de um *scanner* especial que utiliza lasers microscópicos e apresentam a reação de fluorescência de todas as sondas contidas na lâmina e varridas pelo laser. Como as sondas foram marcadas pelas cores vermelha e verde, teremos na imagem gerada, representada por círculos verdes mais intensos, as amostras marcadas com "*cy3*" (no caso da Figura 1.4 seria as amostras de células normais). Representadas por círculos vermelhos mais intensos, as amostras marcadas com "*cy5*" (no caso da Figura 1.4 corresponderia as amostras de células cancerosas). Por fim, no caso de quantidades iguais de "*cy3*" e "*cy5*", os círculos aparecerão em amarelo (BOWTELL, 1999).

Após a geração das imagens de *microarray*, é preciso interpretar os dados obtidos (JAIN et al., 2002). Para essa interpretação, seguimos os passos da Figura 1.5. Os dois últimos pas-

sos, quantificação e normalização e identificação dos genes diferencialmente expressos, serão detalhados nos Materiais e Métodos.

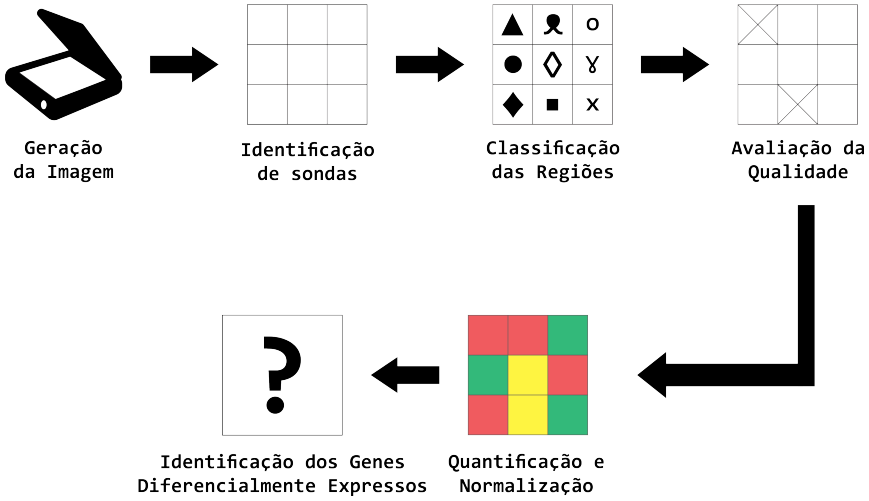

Figura 1.5: Processamento de dados de microarranjo

Desde o processo de obtenção das duas amostras até a obtenção dos genes diferencialmente expressos, que completa o ciclo da geração de um microarranjo, são gerados alguns arquivos (Figura 1.6). Os arquivos gerados são utilizados nas diferentes etapas da análise de microarranjo e estão detalhadas, com cada função, respectivamente, na Tabela 1.2.

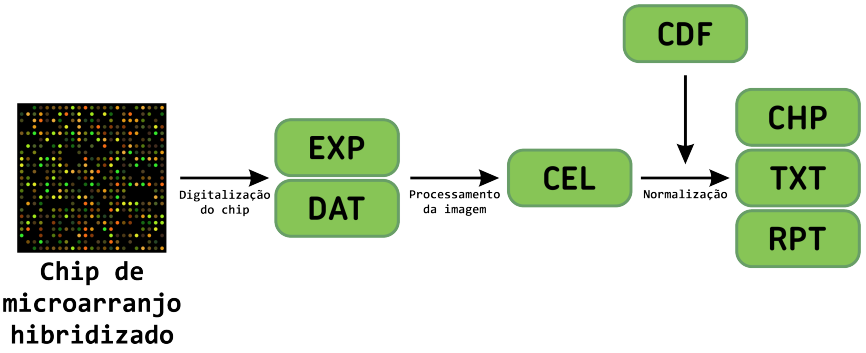

Figura 1.6: *Workflow* dos formatos de arquivos gerados no processamento de um *chip* da *Affymetrix*. Cada formato está especificado na Tabela 1.2.

Tabela 1.2: Descrição de cada arquivo gerado pelo processamento de um *chip* da *Affymetrix*

Arquivo	Descrição
DAT	Imagem óptica bruta do chip hibridizado (TIFF)
CDF	Fornecido pelo *Affy* e descreve o *layout* do chip
CEL	É um arquivo DAT processado (intensidade / valores das posições)
CHP	Resultado experimental criado a partir dos arquivos CEL e CDF
TXT	Valores de expressão das sondas com anotação (arquivo CHP no formato de texto)
EXP	Arquivo texto com detalhes do experimento (tempo, nome, etc)
RPT	Gerado pelo *software Affy* contendo relatório de informações sobre o controle de qualidade

Os métodos de pré-processamento advindos de um microarranjo são de suma importância. Em um microarranjo *Affymetrix* cada gene é representado não por uma sonda, mas sim por um conjunto de sondas, sendo cada conjunto composto por dezenas de pares de sonda. Cada par de sonda, por sua vez, consiste em uma sonda chamada PM (*Perfect Match*) e uma sonda MM (*Mismatch*). Uma sonda PM contém uma sequência de 25 bases que corresponde exatamente a uma sequência passível de hibridização com a amostra. Uma sonda MM, por sua vez, é idêntica à sonda PM com a qual faz par, mas a base do meio, a 13^{a} , é diferente. Assim, uma sonda MM não deveria hibridizar, idealmente, sequência alguma, visto que sua sequência é planejada para não ser complementar a nenhum RNA da amostra. Antes de iniciarmos a descrição dos métodos de pré-processamento, vamos definir inicialmente os índices i, que denota a amostra ou microarranjo, j, que denota o conjunto de sondas destinadas a hibridizar determinada sonda, e k, que denota um par de sonda específico contido em um conjunto de sondas, para identificar cada sonda PM e MM. Ainda, estes métodos envolvem três etapas distintas: a correção de fundo, para retirar um sinal de fundo da medida como um todo, a normalização dos dados, e o tratamento do sinal da sonda PM em relação à sonda MM. Daremos maior atenção a esta última por ser a mais relevante para a compreensão da técnica.

1.4.1 MAS5

Na correção de fundo, para MAS5 (HUBBELL; LIU; MEI, 2002), primeira etapa do método, o microarranjo é dividido em 16 regiões retangulares, e define-se como o sinal de fundo de cada região a média das sondas que estão entre as 2% menos expressas. A correção de fundo se dá então tendo como base sua posição física no microarranjo e o sinal de fundo calculado para cada região do microarranjo. Aqui, ainda se mantém a hipótese de que o erro da medida é lido na sonda MM. Entretanto, nas sondas onde a intensidade lida na MM é maior que a PM, essa hipótese levaria à conclusão de que aquele valor de expressão é negativo, fisicamente impossível. Para contornar esse problema, considera-se que nestes pares a sonda MM falha ao identificar o erro da medida, e este é então estimado a partir das demais sondas do conjunto de sondas a que este pertence. Define-se, portanto, o erro ideal, *IM*,

$$IM_{ij} = \begin{cases} MM_{ij} \mid & MM_{ij} < PM_{ij}, \\ 2^{\frac{PM_{ij}}{SB^{+}_{ij}}} \mid & MM_{ij} \geq PM_{ij} \end{cases} \tag{1.1}$$

que será igual a sonda *MM* caso seu valor seja inferior a *PM*, ou uma fração do sinal da *PM* em função de um ruído específico positivo daquele conjunto de sondas j, SB^{+}_{ij}. Este ruído, por sua

vez, é dado por

$$SB_{ij}^{+} = \begin{cases} SB_{ij} \mid & SB_{ij} > \tau, \\ \frac{\tau}{1+0,1(\tau - SB_{ij})} \mid & SB_{ij} \leq \tau \end{cases} \tag{1.2}$$

$$SB_{ik} = TB_j(\log_2(PM_{ijk}) - \log_2(MM_{ijk})), \tag{1.3}$$

onde TB_j significa o cálculo da média sobre o índice j usando *Tukey's Biweight* (Apêndice A). E o parâmetro $\tau = 0,03$. A Equação 1.2 diz que o ruído é calculado a partir do erro lido naquele conjunto de sondas, mas caso isso também falhe em produzir um erro maior que PM, SB_k^+ segue (Equação 1.2, segundo caso) fracamente baseado nos dados daquele conjunto de sonda. Finalmente, o valor de expressão é corrigido com

$$PM'_{ijk} = max(PM_{ijk} - IM_{ijk}, 2^{-20}) \tag{1.4}$$

onde $max(a,b)$ indica o maior valor entre a e b.

Finalmente, o valor de expressão é calculado por

$$PV_{ij} = TB_k(log_2(PM'_{ijk})) \tag{1.5}$$

Por fim, na normalização, a última etapa do método, realiza-se uma normalização constante, onde todos os valores de expressão de são transladados por um determinado valor de modo que a expressão média de uma amostra seja igual ao de um valor alvo, por padrão, 500.

1.4.2 RMA

No método RMA (IRIZARRY et al., 2003), novamente a primeira etapa é o de retirar o sinal de fundo da medida. Aqui, supõe-se que a distribuição do sinal em relação a sua intensidade é a soma de um sinal verdadeiro, que decai exponencialmente, e um ruído com distribuição normal. Para a normalização, que aqui não é a última etapa, usa-se o método conhecido como normalização quantile (BOLSTAD et al., 2003), onde o objetivo é tornar as distribuições idênticas por meio de métodos estatísticos.

Para a terceira etapa, o RMA assume que as sondas MM não trazem informações confiáveis quanto ao erro medido em PM. Se em um terço dos casos a leitura da sonda MM é maior do

que o da PM, argumenta-se que uma sonda MM também é capaz de identificar sinal verdadeiro, não apenas o erro de medida. Deste modo, calcula-se o valor de expressão para o conjunto de sondas de cada gene baseado apenas nas sondas PM. Assume-se ainda que o erro de medida é multiplicativo e que o sinal identificado é dependente de um termo de afinidade. De fato, neste caso, observa-se que o sinal da sonda MM é tão maior quanto maior for o sinal da sonda PM (IRIZARRY et al., 2003), forte indício de que a sonda MM identifica sinal verdadeiro. Deste modo, seja Y_{ijk} o sinal identificado em PM após correção de fundo e normalização em escala logarítmica, este será dado por

$$Y_{ijk} = \mu_{ij} + \alpha_{jk} + \varepsilon_{ijk} \tag{1.6}$$

onde μ_{ij} é o sinal do gene j na amostra i e α_{jk} é a afinidade da sonda k do gene j. O termo ε_{ijk} representa o ruído da medida. Note, o termo de afinidade da sonda α é o mesmo para toda amostra i, e o pré-processamento conjunto de todas as amostras de um mesmo experimento, para descobrir a afinidade de cada sonda, é um conceito chave que diferencia o RMA de outros métodos de pré-processamento. Após ajuste da Equação 1.6 às expressões observadas nas sondas PM, μ_{ij} é o valor de expressão obtido pelo método RMA.

1.4.3 GCRMA

No método GCRMA (WU et al., 2004), a correção de fundo é igual aos métodos *MAS5* e *RMA*. O que irá diferenciar esse métodos dos outros é a forma que a afinidade da sonda é calculada. Ela é calculada utilizando efeitos de base dependentes da posição, que são mostrados na equação abaixo,

$$ln < B|M >= \sum_{k=1}^{25} \sum_{l \in (A,T,C,G)} S_{l,k} A_{l,k} \tag{1.7}$$

onde B é a intensidade bruta da sonda, M é a intensidade média da matriz, l é o índice do nucleotídeo (A, C, G ou T), k é a posição de l ao longo da sonda (nota-se que k tem uma extensão de 1 até ao comprimento da sequência, que é 25 para as sondas da *GeneChip*), S é uma variável *boolena* igual a 1 se a sequência da sonda tem tamanho de l até k, caso contrário é zero, e A é a afinidade por sítio por nucleotídeo. Outra diferença desse método para os outros é que o ajuste dos dados da sonda MM é baseado na afinidade da mesma, em seguida são subtraídos da sonda PM.

1.4.4 Comparação entre os métodos

Podemos considerar uma questão em aberto sobre qual é o melhor método de pré-processamento possível, havendo trabalhos que se dedicam especificamente a analisar qual produz melhor resultados (LIM et al., 2007) (GYORFFY et al., 2009) (PEPPER et al., 2007) (GHARAIBEH; FODOR; GIBAS, 2008), mas é consenso que o RMA/GCRMA supera outros métodos para genes com baixa expressão, onde o MAS5 produz muitos falsos positivos (IRIZARRY et al., 2003) (PEPPER et al., 2007). Há vantagens e desvantagens em utilizar os métodos citados acima, onde cada um tem as suas peculiaridades e o critério de escolha depende do experimento do pesquisador. Seguem algumas vantagens de utilizar os métodos RMA/GCRMA:

i) Retorna menos falsos positivos que MAS5;

ii) Fornece estimativas de *fold change* mais consistentes;

iii) A exclusão dos dados das sondas MM no RMA reduz o ruído, mas perde informações;

iv) A inclusão do ajuste para a sonda MM no método GCRMA reduz o ruído e mantém os dados dessa sonda.

Em contrapartida, algumas desvantagens em utilizar RMA/GCRMA:

i) Pode ocultar mudanças reais, especialmente em baixos níveis de expressão (falsos negativos);

ii) Realiza controle de qualidade após a normalização;

iii) A normalização assume uma distribuição igual que pode esconder as mudanças biológicas.

1.5 Projeto Toxicogenômico Japonês (PTGJ)

O PTGJ (UEHARA et al., 2010) foi realizado entre 2002 e 2007 em conjunto com o Instituto Nacional de Ciências da Saúde do Japão, Instituto Nacional de Inovação Biomédica e 17 empresas farmacêuticas, com o objetivo de criar um banco de dados toxicológico que permite o uso tanto da toxicologia avançada e como da reversa (KANNO, 2003). No Projeto, foram selecionados como órgãos alvo o rim e o fígado, uma vez que a maioria das toxicidades clínicas surgem nesses órgãos. Os produtos químicos ou drogas em testes foram administrados em ratos ou expostos à células de cultura, de forma a obter os dados de expressão gênica nos órgãos alvos

das células ou animais. As alterações nos marcadores toxicológicos tradicionais também foram recolhidos a partir dos animais. O objetivo é estabelecer um sistema de previsão de toxicidade na fase inicial de desenvolvimento de medicamentos. Foram utilizados apenas dados para o fígado, pois os dados de rim são restritos para acesso.

A Tabela 1.3 mostra algumas informações a respeito do Projeto. Essas informações dizem a respeito da forma de coleta das amostras, célula escolhida para estudo, dose, tempo de sacrifício, amostragem, itens examinados e tratamento. Essas informações são de extrema importância, pois a partir delas podemos identificar o delineamento experimental utilizado.

Tabela 1.3: Informações do Projeto Toxicogenômico Japonês (UEHARA et al., 2010).

	Rat in vivo	*Rat in vitro*	*Human in vitro*
Animal	*Sprague-Dawley*	*Sprague-Dawley*	-
Instrumento de Coleta	- 0.5% de metilcelulose ou óleo de minho (via oral) - Salina ou 5% de solução de glicose (via intravenosa)	- Meio de cultura - Dimetilsulfóxido (DMSO)	- Meio de cultura - Dimetilsulfóxido (DMSO)
Célula	-	Hepatócitos isolados por digestão com colagenase	Hepatócitos congelados
Dose	Baixa, média e alta	-	-
Sacrifice	- 3, 6, 9 e 24h após administração única - 24h após a última dose repetida	-	-
Amostragem	Fígado e rim	Fígado e rim	Fígado e rim
Análise de Microarranjo	GeneChip da Affymetrix	Duplicatas	Duplicatas
Itens examinados	- Peso corporal - Peso dos órgãos - Consumo de comida - Hematologia - Bioquímica do sangue	Viabilidade celular (LDH e conteúdo de DNA)	Viabilidade celular (LDH e conteúdo de DNA)
Tratamento	3, 6, 9 e 24h	2, 8 e 24h	2, 8 e 24h

Os dados fornecidos pelo PTGJ fornece são apresentados na Tabela 1.4. Foram utilizados diferentes drogas, tempos de amostragens, repetição dos experimentos e concentrações de dose (Apêndice D).

A motivação para a criação do Projeto Toxicogenômico Japonês vem com a intenção de contribuir para os progressos em tratamentos médicos através da oferta de novos medicamentos inovadores com alta eficácia e segurança. As empresas farmacêuticas realizam periodicamente programas de investigação para o desenvolvimento de drogas, no entanto, é praticamente impossível evitar efeitos colaterais inesperados. Se os possíveis efeitos colaterais que ocorrem no uso clínico são capazes de ser previstos na fase inicial do desenvolvimento de drogas, as companhias farmacêuticas podem avaliar a segurança de novos produtos químicos ou drogas antes do estudo em larga escala não-clínica ou clínica, e, posteriormente, reduzir os custos, fornecendo medicamentos mais seguros aos pacientes. O projeto tem como objetivo contribuir para o desenvolvimento de drogas com menos efeitos adversos por elucidação da inter-relação entre

Tabela 1.4: Resumo do PTGJ para dados de fígado (UEHARA et al., 2010).

	Homo sapiens in vitro	***Rattus norvegicus in vitro***	***Rattus norvegicus in vivo***	***Rattus norvegicus in vivo***
Dosagem	única	única	única	repetida diariamente
Concentração de Dose	baixa, média e alta	baixa, média e alta	baixa, média e alta	baixa, média e alta
Tempo de Amostragem	2h, 8h e 24h	2h, 8h e 24h	3h, 6h, 9h e 24h	3d, 7d, 14h e 28d
Repetição do Experimento	duplicatas	duplicatas	triplicatas	triplicatas
Arrays	2004	3120	5568	6192
Sondas por *Array*	54675	31099	31099	31099
Medicamentos	119	131	131	131
Quantidade de Dados	54,3 GB	21,9 GB	43,6 GB	43,5 GB

substâncias tóxicas e expressão gênica (CHEN et al., 2011).

Existem outros projetos que também geraram dados toxicogenômicos em grande escala. Um exemplo é o DrugMatrix (GANTER et al., 2005), que foi produzido pela empresa Iconix Pharmaceuticals e depois comprada e disposta como domínio público pelo Instituto Nacional de Saúde (NIH) dos Estados Unidos, e é constituído de experimentos toxicológicos nos quais ratos ou hepatócitos do rato primário foram sistematicamente tratados com produtos químicos terapêuticos, industriais e ambientais em doses não tóxicas e tóxicas. Após a administração destes compostos *in vivo*, foi realizado coleta de dados de expressão gênica para posterior análise dos efeitos destes compostos em diferentes tempos de amostragem e diferentes órgãos alvo (rim, fígado e coração). A principal diferença encontrada entre o Projeto Toxicogenômico Japonês (PTGJ) e o DrugMatrix está na organização dos dados (CHEN et al., 2012). Os dados do PTGJ estão relativamente mais padronizados e organizados no que diz respeito a tempos de amostragem, dosagens, forma de obtenção das amostras, etc. Enquanto o DrugMatrix possui aparentemente um *design* experimental relativamente padronizado, não possui uma organização tão estrita quanto ao do PTGJ. Desta forma, este foi o critério de escolha do PTGJ para ser utilizado neste presente trabalho.

1.6 Avaliação Crítica de Análise de Dados em Massa (CAMDA)

O Projeto Toxicogenômico Japonês é muito utilizado como base para diversos trabalhos e propostas. Uma das utilizações do projeto foi no CAMDA (JOHNSON; LIN, 2001), que é uma conferência internacional anual que teve início em 2000 e ocorre um ano nos Estados Unidos

e outro na Europa. Tem como principal enfoque a análise maciça de dados, introduzindo e avaliando novas abordagens e soluções para o problema de análise de grande quantidade de dados. A conferência apresenta novas técnicas no campo da bioinformática, análise de dados e estatísticas para a manipulação e processamento de grandes conjuntos de dados.

Uma das principais atividades do CAMDA é o desafio proposto, que têm como objetivo analisar grandes quantidades de dados. Pesquisadores de universidades, institutos e de empresas de todo o mundo são convidados a participar dos desafios (TILSTONE, 2003).

O enfoque deste trabalho está nos desafios propostos nos anos de 2012, 2013 e 2014. Nesses anos, os desafios propostos foram baseados no banco de dados criado pelo Projeto Toxicogenômico Japonês (PTGJ) com o propósito de avaliar se há a possibilidade de substituir o estudo *in vivo* pelo *in vitro* e também se é possível predizer doenças relacionadas ao fígado em humanos usando dados toxicogenômicos de animais. Desde então, muitos pesquisadores tentaram responder esses questionamentos propostos neste desafio. Houveram dezenas de publicações tomando diversas frentes de abordagens, por exemplo, selecionando especificamente um pequeno conjunto de drogas a fim de tirar conclusões a partir disto, análises com metodologias diferentes selecionando, mais uma vez, um pequeno conjunto de drogas, entre outras. A partir de uma revisão foi constatado que nenhum pesquisador realizou uma análise completa com todas as 131 drogas do PTGJ a fim de responder os questionamentos. Sendo o PTGJ muito rico em informações, será realizado neste trabalho uma metodologia de análise que englobe todas as drogas.

2 Objetivos

O presente trabalho tem como objetivo a meta-análise de Projeto Toxicogenômico Japonês a fim de analisar as diferenças entre os modelos *in vivo* e *in vitro* verificando, assim, se é possível a substituição do modelo *in vivo* pelo *in vitro* através da metodologia do mapa modular.

2.1 Objetivos Específicos

i) Identificação dos genes diferencialmente expressos;

ii) Realização do enriquecimento funcional;

iii) Aplicação e criação de um pacote em R com a metodologia do mapa modular.

3 *Material e Métodos*

3.1 *Workflow*

O *workflow* representado pela Figura 3.1 mostra todos os passos utilizados nesse trabalho. Abaixo iremos detalhar cada parte.

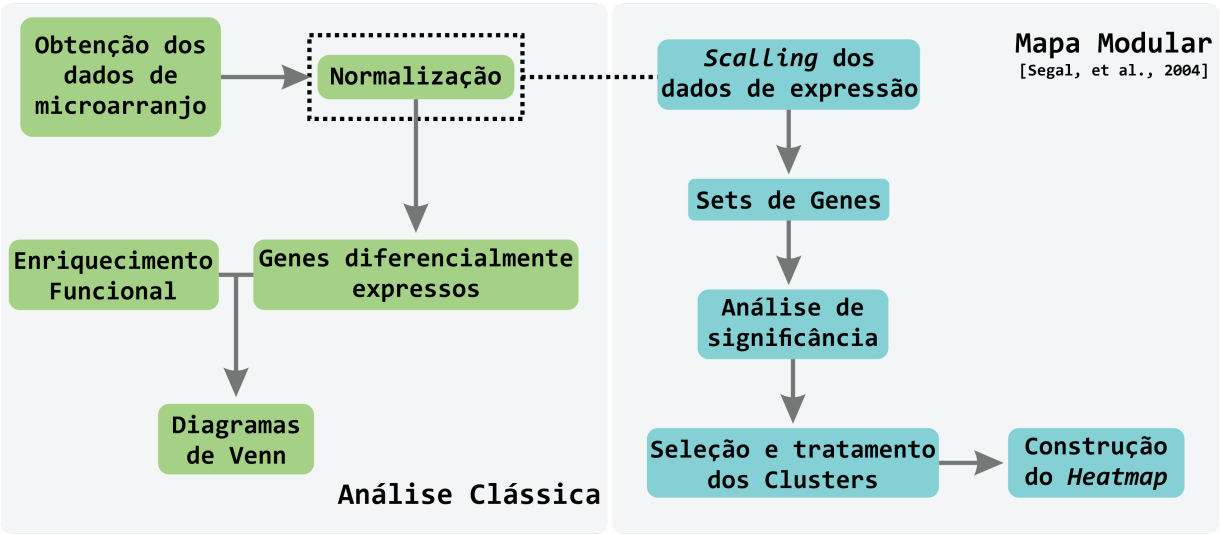

Figura 3.1: Visão geral do material e métodos utilizados.

3.2 *Hardware*

Para as análises foram utilizados computadores do Laboratório de Estudos em Biocomplexidade assim como computadores alocados no Instituto de Biotecnologia de Botucatu (IBTEC). Foi utilizado o sistema operacional *ubuntu* 14.04 LTS, memória RAM de 16 GB, processador Intel Core i7-4770 CPU com taxa de frequência de 3.40 GHz e 8 núcleos disponíveis, além de placa de vídeo modelo GeForce GTX 645/PCIe/SSE2.

3.3 Obtenção dos Dados de Microarranjo

Os dados de microarranjo estão disponíveis no link `http://bioinf.boku.ac.at/camda12/toxdata/`. Os *chips* utilizados no Projeto Toxicogenômico Japonês foram da empresa *Affymetrix*, sendo para *Homo sapiens* o modelo *Affymetrix GeneChip Human Genome U133 Plus 2.0 Array* e para *Rattus norvegicus* o modelo *Affymetrix GeneChip Rat Genome 230 2.0 Array*. A prospecção destes dados foi realizada no dia (11/11/2015). Eles podem ser divididos em quatro partes: *Homo sapiens in vitro*, *Rattus norvegicus in vitro*, *Rattus norvegivus in vivo repeat* e *Rattus norvegicus in vivo single*. Para cada experimento há dois arquivos: arquivo com extensão *.CEL* (plataforma *Affymetrix*) e um arquivo com extensão *.tsv* que indica como cada arquivo *.CEL* está relacionado quanto a espécie, tempo de amostragem, concentração de dose, entre outros. Somente algumas drogas possuem tempo de amostragem e concentração de dose diferentes das apresentadas na Tabela 1.4.

3.4 Pré-processamento dos dados

A fase de pré-processamento dos dados incluem a normalização e obtenção dos genes diferencialmente expressos. Os dados obtidos foram normalizados, ou seja, ajustados para os efeitos que surgem devido à variação na tecnologia e de diferenças biológicas entre as amostras de RNA ou entre sondas. É nesta etapa em que há a transformação de intensidade em expressão de acordo com o que foi explicado na Figura 1.6.

A partir da obtenção dos dados, utilizou-se a linguagem de programação R juntamente com o *software Rstudio* (RStudio Team, 2015), o qual apresenta uma interface mais adequada para a elaboração do código. Além da linguagem de programação e do *software*, foi utilizado alguns pacotes que auxiliaram nos passos posteriores: **affy** (GAUTIER et al., 2004), **gplots** (WARNES et al., 2009), **hgu133plus2.db** (CARLSON et al., 2012), **limma** (RITCHIE et al., 2015), **rat2302.db** (CARLSON, 2002) e **simpleaffy** (MILLER, 2007).

O primeiro passo a ser feito foi a leitura do arquivo com extensão *.tsv* de uma droga para que obtenha-se as informações mais detalhadas da mesma. Um novo arquivo é gerado onde possui somente a relação de cada arquivo *.CEL* da droga em questão, com a concentração de dose e o tempo de amostragem. A Tabela 3.1 mostra um exemplo do arquivo criado.

Após a criação do arquivo anterior, é utilizada a função *readTargets* para a leitura do experimento de microarranjo. Em seguida realiza-se a leitura dos arquivos *.CEL* através da função *ReadAffy* que retorna, além da leitura dos arquivos com extensão *.CEL*, informações relati-

Tabela 3.1: Exemplo do arquivo criado que relaciona cada *.CEL* com a concentração de dose e tempo de amostragem.

Nome do Arquivo	**Dosagem**	**Tempo de Amostragem (h)**
Arquivo_1.CEL	Controle	8
Arquivo_2.CEL	Controle	8
Arquivo_3.CEL	Controle	24
Arquivo_4.CEL	Controle	24
Arquivo_5.CEL	Alta	8
Arquivo_6.CEL	Alta	8
Arquivo_7.CEL	Alta	24
Arquivo_8.CEL	Alta	24
Arquivo_9.CEL	Média	8
Arquivo_10.CEL	Média	8
Arquivo_11.CEL	Média	24
Arquivo_12.CEL	Média	24

vas quanto ao tamanho dos *arrays*, número de amostras, número de genes e tipo de anotação utilizada. Por fim, utiliza-se três funções, descritas a seguir, que representam os tipos de normalização disponíveis para dados de microarranjo: *mas5*, *rma* e *gcrma*. Escolhido o tipo de normalização, obtém-se uma tabela de valores que relaciona o valor de normalização de cada sonda (ou gene) com cada arquivo *.CEL*.

A obtenção dos genes diferencialmente expressos é um passo fundamental para o entendimento das alterações biológicas do tecido/organismo, (por exemplo, ciclo celular, dobramento de proteínas, processo metabólico de drogas, regulação do reparo de DNA etc).

Realizada a normalização dos dados, é gerada uma matriz modelo que contém o *design* do experimento em questão. A partir da obtenção dessa matriz faz-se necessário a escolha de um modelo para o tratamento dos dados, que nesse caso foi um modelo linear. Feito isso, será realizada a comparação entre as concentrações de doses com o tempo de amostragem. Sempre comparamos uma concentração de dose com a concentração de dose controle e, obrigatoriamente, os tempos de amostragem devem coincidir. Por exemplo, cruzamos *High2* (concentração de dose alta com tempo de amostragem de 2 horas) com *Control2* (concentração de dose controle com tempo de amostragem de 2 horas), *Low8* (concentração de dose baixa com tempo de amostragem de 8 horas) com *Control8* (concentração de dose controle com tempo de amostragem de 8 horas), e assim por diante. Dessa forma, podemos obter a matriz de contrastes com todas as combinações respeitando as restrições citadas anteriormente. Após a obtenção da matriz de contrastes o próximo passo é obter tabelas com genes diferencialmente expressos e os respectivos valores estatísticos. Com a seleção dos genes diferencialmente expressos e suas respectivas estatísticas, são realizados alguns filtros a fim de selecionar os genes que possuem mais signi-

ficância. Para isso selecionamos os genes que possuem valores de p-valor menores que 0.05 e valores de *logFC* menores que -1 ou maiores que 1. A Tabela 3.2 mostra um exemplo de tabela de genes diferencialmente expressos para a droga etanol.

Sonda	***Gene Symbol***	***logFC***	***AveExpr***	**t**	**p-valor**	**p-valor ajustado**	**B**
225424_at	GPAM	1.0771	7.229	12.541	3.02E-008	0.002	5.942
213524_s_at	G0S2	0.492	10.281	7.356	8.89E-006	0.069	2.891
224303_x_at	NIN	-0.500	3.817	-5.824	8.24E-005	0.163	1.336
205776_at	FMO5	0.571	5.431	4.852	0.001	0.283	0.132
208990_s_at	HNRNPH3	0.310	7.940	4.343	0.001	0.308	-0.566
216965_x_at	SPG20	-0.389	3.855	-3.438	0.005	0.427	-1.905
1557118_a_at	INTS6-AS1	0.255	3.102	3.212	0.007	0.447	-2.252
1555048_a_at	TSPEAR	-0.260	3.246	-2.924	0.013	0.481	-2.698
210612_s_at	SYNJ2	-0.241	3.786	-2.333	0.038	0.589	-3.600
219317_at	POLI	0.1883	4.413	1.893	0.083	0.677	-4.2346

Tabela 3.2: Genes diferencialmente expressos encontrados para o Etanol com combinação de dosagem alta (*high*) e tempo de amostragem de 8 horas (H8C8) para a normalização do tipo RMA.

A Tabela 3.2 possui alguns parâmetros interessantes que valem a pena ser ressaltados:

- *Sondas*: Nomes das sondas que são específicas e únicas para cada chip de microarranjo;
- *Gene Symbol*: Conversão dos nomes das sondas para a nomenclatura *Gene Symbol*;
- *logFC*: Medida que descreve o quanto uma quantidade muda indo de um valor inicial para um valor final. Fornece o valor do contraste. Geralmente, representa uma mudança de *log2* entre duas ou mais condições experimentais, embora, às vezes, represente um nível de expressão em *log2*;
- *AveExpr*: Fornece o nível médio de expressão em *log2* para determinado gene em todos os *arrays* e canais no experimento;
- *t*: O teste *t* é um teste de hipótese em que se usa conceitos estatísticos para rejeitar ou não uma hipótese nula quando a estatística de teste (*t*) segue uma distribuição t de *student*;
- *P-valor*: Avalia se os dados da amostra suportam o argumento de que a hipótese nula é verdadeira. Mede o quão compatível os dados são com a hipótese nula. Altos valores p-valor significam que os dados são suscetíveis à hipótese nula. Em contrapartida, baixos valores de p-valor significam que os dados são insuscetíveis à hipótese nula;
- *P-valor ajustado*: É o p-valor ajustado para múltiplos testes;
- *B*: A estatística B mostra a probabilidade (em porcentagem) de que determinado gene seja diferencialmente expresso.

3.5 Enriquecimento Funcional

Após a obtenção dos genes diferencialmente expressos, realizamos um enriquecimento funcional destes. O enriquecimento retornará quais rotas, vias metabolicas, funções moleculares, processos biológicos e componentes celulares estão alteradas num determinado grupo de genes, dessa forma identificando um perfil funcional desse grupo. O enriquecimento foi realizado para o *GO*, *KEGG* e *reactome*. Para isso, foram utilizados os seguintes pacotes no R: **DOSE** (YU et al., 2015), **clusterProfiler** (YU et al., 2012), **ReactomePA** (YU; HE, 2016) e **hgu133plus2.db** (CARLSON et al., 2012).

Foram utilizados os seguintes parâmetros estatísticos para a realização do enriquecimento funcional: p-valor < 0.05, FDR (*False Discovery Rate*) como método de ajuste do p-valor e q-valor = 0,1.

3.6 Mapa Modular

A metodologia do mapa modular (SEGAL et al., 2004), originalmente, foi criada para o tratamento de dados de expressão de microarranjo de câncer, mas a adaptamos a fim de utilizar os dados de expressão de microarranjo do Projeto Toxicogenômico Japonês. O principal objetivo dessa metodologia é correlacionar os dados de expressão com *sets* de genes para extrair *clusters* com significado biológico. Um *workflow* da metodologia está representada na Figura 3.2.

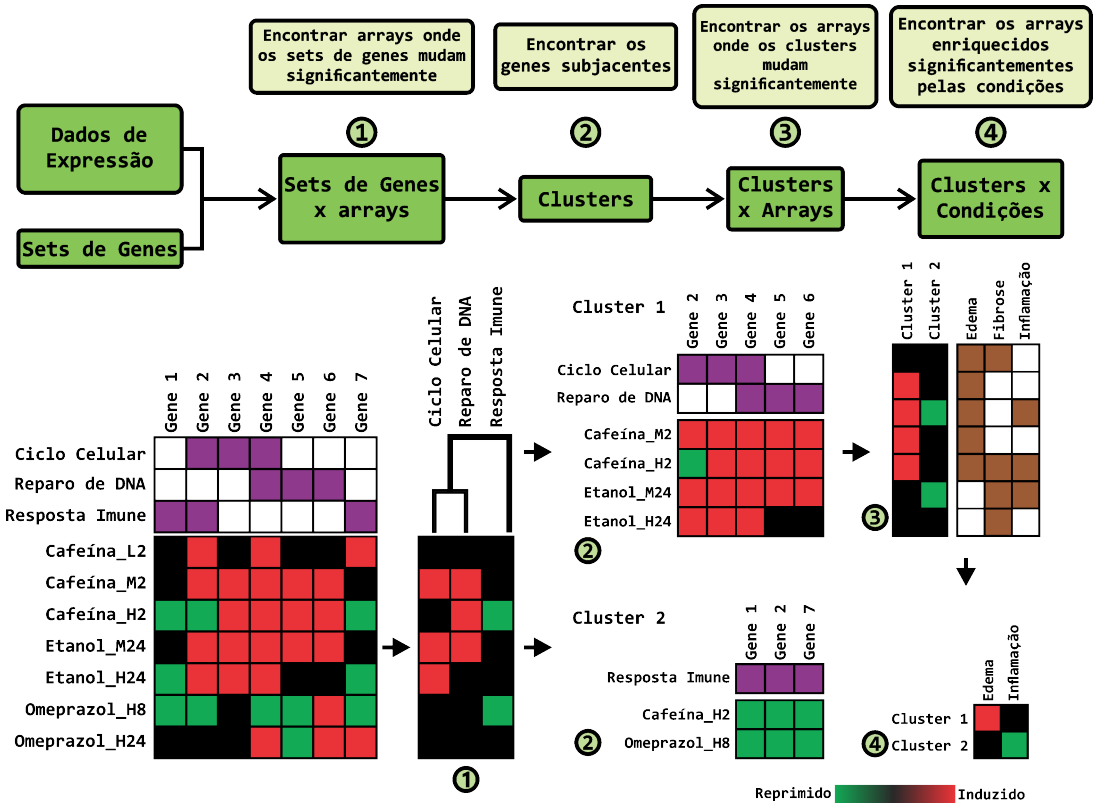

Figura 3.2: Exemplo de análise com um *input* de dados de expressão de sete *arrays* (cafeína_L2, cafeína_M2, cafeína_H2, etanol_M24, etanol_H24, omeprazol_H8 e omeprazol_H24), sete genes (gene 1–7) e três conjuntos de genes (ciclo celular, reparo de DNA e resposta imune). Os números circulados correspondem aos passos no fluxograma. Neste exemplo, os conjuntos de genes ciclo celular e reparo de DNA são significativamente induzidos nos *arrays* cafeína_M2, cafeína_H2, etanol_M24, etanol_H24 e, portanto, constituem um *cluster* de conjuntos de genes, enquanto que o conjunto de genes reposta imune é significativamente reprimido nos *arrays* cafeína_H2 e omeprazol_H8, portanto, constitui seu próprio *cluster* de conjuntos de genes. O módulo resultante do primeiro *cluster* de conjuntos de genes inclui os genes 2, 3, 4, 5 e 6, uma vez que estes genes contribuem para a expressão significativa deste *cluster*. No passo final da análise, os *arrays* são anotados com condições clínicas (edema, fibrose e inflamação); por exemplo, o array cafeína_L2 é anotado com as condições edema e fibrose. O conjunto de *arrays* onde o módulo 1 é significantemente induzido (*arrays* cafeína_M2, cafeína_H2, etanol_M24, etanol_H24) é enriquecido para a condição edema e o conjunto onde o módulo 2 é significativamente reprimido é enriquecido para a condição inflamação. Figura adaptada de (SEGAL et al., 2004).

3.6.1 Obtenção dos Dados de Expressão

A partir dos dados de microarranjo normalizados, foi realizada a média de todos os valores do gene A, por exemplo, e subtraído individualmente da média o valor para cada *array*, de forma que o valor médio do gene A seja 0 no final. Podemos observar esse passo na Figura 3.3.

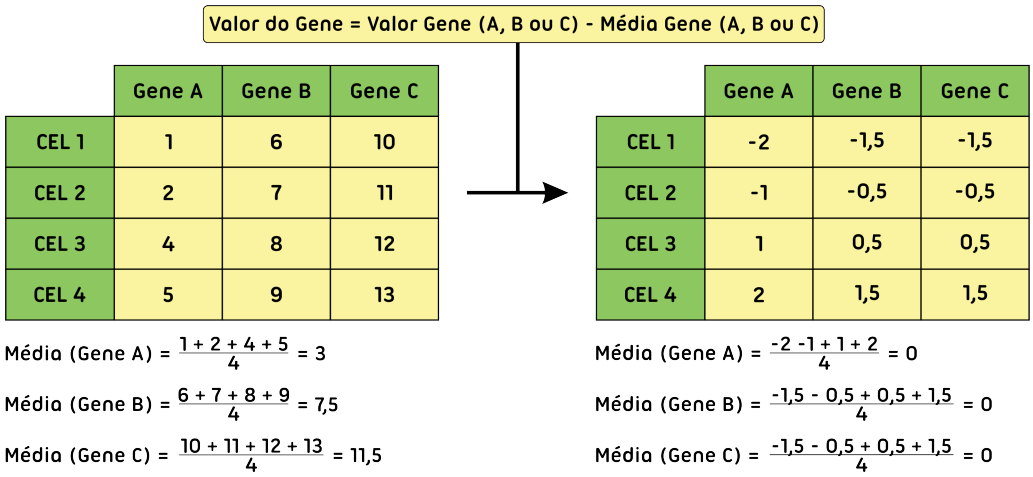

Figura 3.3: Exemplo do método aplicado para obter-se a média de todos os genes, em todos os *arrays*, igual a 0.

3.6.2 Obtenção dos *Sets* de Genes

Para a aplicação da metodologia do mapa modular, é necessário, além dos dados de expressão, obter dados que correspondem às informações do ***GO*** (*Gene Ontology*) (ASHBURNER et al., 2000) (CONSORTIUM et al., 2015), ***KEGG*** (*Kyoto Encyclopedia of Genes and Genomes*) (KANEHISA; GOTO, 2000) (KANEHISA et al., 2015) e ***Reactome Pathway*** (CROFT et al., 2014) (FABREGAT et al., 2016) que são denominados *sets* de genes.

Prospectamos informações referentes aos três bancos de dados citados no parágrafo anterior. O tratamento desses dados é realizado dentro do ambiente do *Rstudio*. Para tanto, utilizou-se três pacotes dentro do *R*: **biomaRt** (DURINCK et al., 2005) (DURINCK et al., 2009), **GO.db** (CARLSON, 2013) e **KEGGREST** (TENENBAUM, 2013).

Com o uso desses três pacotes, é possível obter os genes que compõe cada *set* de genes do *GO*, *KEGG* e *Reactome*. Esses dados foram obtidos tanto para *Homo sapiens* quanto para *Rattus norvegicus*. Um exemplo de arquivos obtidos está na Tabela 3.3.

O banco de dados do *Gene Ontology* (*GO*) foi obtido no dia 20/04/2016, o *KEGG* em 17/04/2016 - mais especificamente a versão 78.0 - e o *Reactome* em 18/04/2016 - na versão 57. A Figura 3.4 mostra a distribuição dos *sets* de genes obtidos para *Homo sapiens* e *Rattus*

GO:0000002	hsa:10000	r-hsa-1059683
Manutenção do Genoma mitocondrial	**AKT serina/ treonina Cinase 3**	**Sinalização da Interleucina 6**
MGME1	AKT3	STAT3
SLC25A4	MPPH	SOCS3
AKT3	MPPH2	CBL
SLC25A33	PKB-GAMMA	JAK2
MPV17	PKBG	IL6
MEF2A	PRKBG	IL6ST
TYMP	RAC-PK-GAMMA	JAK1
SLC25A36	RAC-GAMMA	TYK2
MRPL17	STK-2	PTPN11
LONP1		IL6R
OPA1		
PIF1		
SESN2		
PARP1		
POLG2		
LONP1		

Tabela 3.3: Exemplo de *sets* de genes obtidos com os respectivos genes presentes para *Gene Ontology* (GO:0000002), *KEGG* (hsa:10000) e *Reactome* (r-hsa-1059683).

norvegicus. As ontologias, vias e rotas não possuem quantidades iguais para *Homo sapiens* e *Rattus norvegivus*, portanto, foi realizado uma intersecção de forma que possuam os mesmos *sets* de genes para as duas espécies.

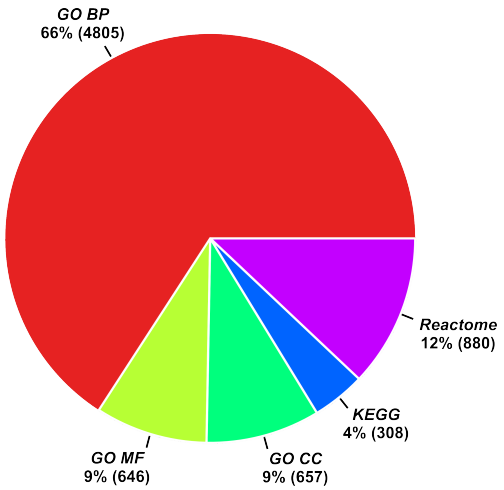

Figura 3.4: Distribuição dos *sets* de genes para *Homo sapiens* e *Rattus norvegicus*. Lembrando que foi realizada uma intersecção de cada *set* de gene para as, duas espécies, de forma que possuam os *sets* iguais.

3.6.3 Identificação dos *Arrays* em que a expressão dos *sets* de genes estão alterados

Para definir os *arrays* em que a expressão dos *sets* de genes alteram significantemente, é necessário encontrar os genes induzidos e reprimidos. Estes genes são aqueles cuja mudança no valor de expressão é maior ou menor a dois em relação a uma referência.

Para isso, é necessário criar aquivos no formato *".txt"* que contenham os genes com os respectivos valores normalizados para cada variação de concentração de dose, tempo de amostragem e tipo de normalização. O primeiro passo para a criação desse novo arquivo é carregar o arquivo gerado a partir da normalização. Em seguida, foram excluídas todas as linhas e colunas cujos genes não possuem valor de normalização. Caso existam genes com nomes iguais, mas com valores de normalização diferentes (isso ocorre pelo fato de várias sondas poderem corresponder ao mesmo gene), fazemos a média dos valores normalizados obtendo somente valores normalizados para únicos genes, sem repetição. Como citado na Seção 3.3, há um arquivo *.tsv* para cada droga que identifica como cada arquivo *.CEL* está relacionado com variações de concentração de dose e tempo de amostragem. A partir dessa identificação, é possível reunir, por exemplo, todos os arquivos *.CEL* que correspondem à concentração de dose média e tempo de amostragem de 8 horas, concentração de dose alta e tempo de amostragem de 24 horas e assim por diante. Um exemplo da metodologia está na Figura 3.5, onde *X8008.CEL* e *X8009.CEL* correspondem à concentração de dose controle e tempo de amostragem de 8 horas, *X8010.CEL* e *X8011.CEL* correspondem a concentração de dose controle e tempo de amostragem de 24 horas, *X8013.CEL* e *X8014.CEL* correspondem a concentração de dose média e tempo de amostragem de 8 horas e, por fim, *X8016.CEL* e *X8017.CEL* correspondem a concentração de dose alta e tempo de amostragem de 24 horas.

Após a geração da tabela presente na Figura 3.5 e sabendo-se o que representa cada arquivo *.CEL*, como citado anteriormente, são feitos os cálculos representados na Figura 3.6 para a geração da nova tabela. Esses cálculos são realizados sempre relacionando a concentração de dose controle com um determinado tempo de amostragem com as variações de dose (alta e/ou baixa e/ou média) e suas respectivas variações de tempos de amostragens. Por exemplo, na Figura 3.6, são obtidas as relações das concentrações de dose controle para os tempos de amostragem de 8 horas e 24 horas, e, em seguida, a relação da concentração de dose média com tempo de amostragem de 8 horas, que é feita inicialmente a partir da média dos dois valores, e na sequência, esse valor é dividido pelo valor da concentração de dose controle para tempo de amostragem 8 horas. Esse raciocínio é o mesmo para todas as linhas que correspondem a cada gene individualmente.

SYMBOL	X8008.CEL	X8009.CEL	X8010.CEL	X8011.CEL	X8013.CEL	X8014.CEL	X8016.CEL	X8017.CEL
A1BG	8,714946	8,741572	8,651843	8,644537	8,703162	8,722072	8,012564	8,064162
A1BG-AS1	4,255945	4,358846	4,081405	4,250012	4,097526	4,172282	4,095089	4,086247
A1CF	7,25908	7,495458	7,481247	7,506941	7,604011	7,603242	8,222587	8,269445
A1CF	5,173816	5,299631	5,514949	5,48873	5,634366	5,628671	6,250759	6,550066
A2M	3,380336	3,525618	3,707762	3,193432	3,53157	3,605133	3,582781	3,522159
A2M	8,670354	8,697425	8,645549	8,605031	8,596569	8,703311	8,168698	8,094056
AA06	7,075738	7,085276	6,975947	6,91455	7,05011	7,096179	7,075028	7,051665
AAK1	4,165193	4,282236	4,350026	4,372497	4,359844	4,142707	4,203867	4,307291
AAK1	4,014377	3,82049	3,787177	3,613966	3,812629	3,911589	3,697893	3,719539
AAK1	3,025211	3,04547	3,02907	2,846776	2,792203	2,724532	2,950688	2,922161
AAK1	3,174059	3,247248	3,037821	3,100755	3,044181	3,198132	3,147607	3,171303
AAK1	7,021389	6,841627	6,805985	6,808023	6,999311	6,990415	7,001227	7,108732
AAK1	3,881909	3,613511	3,611652	3,404445	3,613511	3,60956	3,572903	3,68927
AAK1	3,146618	3,267903	3,334066	3,253275	2,997968	3,221463	3,129834	3,373261

↓ média

SYMBOL	X8008.CEL	X8009.CEL	X8010.CEL	X8011.CEL	X8013.CEL	X8014.CEL	X8016.CEL	X8017.CEL
A1BG	8,714946	8,741572	8,651843	8,644537	8,703162	8,722072	8,012564	8,064162
A1BG-AS1	4,255945	4,358846	4,081405	4,250012	4,097526	4,172282	4,095089	4,086247
A1CF	6,216448	6,397545	6,498098	6,497836	6,619189	6,615956	7,236673	7,409756
A2M	6,025345	6,111521	6,176655	5,899231	6,06407	6,154222	5,875739	5,808108
AA06	4,622554	4,562266	4,513954	4,437288	4,524541	4,52473	4,521316	4,567137
AAK1	3,146618	3,267903	3,334066	3,253275	2,997968	3,221463	3,129834	3,373261

Figura 3.5: Exemplificação da construção da primeira tabela para a identificação dos genes que alteram significantemente para a normalização do tipo RMA.

Linha 1 = gene A1BG

dados_C8 = (8.714946 + 8.741572) / 2 = 8.728267
dados_C24 = (8.651843 + 8.644537) / 2 = 8.64819
M8 = [(8.703162 + 8.722072) / 2] - dados_C8 = -0.015642
H24 = [(8.012564 + 8.064162) / 2] - dados_C24 = -0.689896

Linha 2 = gene A1BG-AS1

dados_C8 = 4.255945 + 4.358846) / 2 = 4.307396
dados_C24 = (4,081405 + 4,250012) / 2 = 4.165709
M8 = [(4,097526 + 4,172282) / 2] - dados_C8 = -0.1724915
H24 = [(4,095089 + 4,086247) / 2] - dados_C24 = -0.2167275

	Middle_8h	High_24h
A1BG	-0.015642	-0.689896
A1BG-AS1	-0.1724915	-0.2167275
A1CF	0.310576	1.016.218
A2M	0.040713	-0.2265095
AA06	-0.0677745	-0.0481835
AAK1	-0.097545	0.044287

Figura 3.6: Exemplificação da construção da nova tabela para a identificação dos genes que alteram significantemente. Esses cálculos são realizados linha a linha da segunda tabela (6 linhas totais) da Figura 3.5. Cálculo baseados na normalização RMA.

Como citado na Seção 1.4, os diferentes tipos de normalização produzem diferentes resultados. A normalização RMA e GCRMA produzem resultados da operação entre logaritmos na base 2 (log_2), enquanto que a normalização MAS5 produz resultados não logarítmicos. Essa diferença nos resultados irá provocar uma alteração nos cálculos. Para MAS5 será realizada a divisão, enquanto que para RMA e GCRMA subtração, pois de acordo com as propriedades de logaritmos, quando temos uma operação de divisão logarítmica podemos fazer a subtração de ambos.

Após a realização dos cálculos da Figura 3.6 para todos os experimentos, drogas e tipos de normalização, será obtido três arquivos para cada droga. Cada um desses arquivos irá mostrar o quanto o valor de expressão variou. Utilizaremos somente os valores que são ≥ 2 ou $\leq$ -2, que representam, respectivamente, valores para os genes induzidos e reprimidos para cada *array*. Esses resultados podem ser expressos em vários arquivos *.txt* ou armazenados em uma lista dentro do R para futuras comparações. Além desses arquivos ou listas geradas, obtemos, também, uma matriz que possui em suas linhas os *arrays* (drogas com variações de concentrações de doses e tempos amostragens) e colunas com os nomes dos genes. Dessa forma, teremos uma matriz que corresponde à relação entre *Arrays* e genes.

3.6.4 Matriz de relação *Sets* de Genes x *Arrays*

O próximo passo é construir uma matriz que mostre a relação entre os *sets* de genes e os *arrays*. Essa relação será construída a partir da informação dos *sets* de genes obtidos anteriormente e dos genes induzidos (ou reprimidos) encontrados. Para relacionar os genes dos *arrays* com os *sets* de genes, utilizamos o Teste Exato de Fisher (que será equivalente a Distribuição Hipergeométrica) que retornará um p-valor que será importante para essa relação que está sendo construída. A partir da verificação da presença dos genes em um determinado *array* em relação aos genes presentes em um certo *set* de gene, é possível montar uma matriz igual como é mostrado na Tabela 3.4, preenchendo tal matriz para cada caso. A montagem da matriz é bem simples. Consiste em preencher qual a fração de genes presentes e ausentes dentro de um dado *set* de genes e a fração de genes que está induzido (ou reprimido) no *set* de gene e *array*.

	induzidos (ou reprimidos) no set de gene	**induzidos (ou reprimidos) no array**	**total**
presença no set de gene	a	b	a + b
ausência no set de gene	c	d	c + d
total	a + c	b + d	N

Tabela 3.4: Matriz montada para a obtenção do p-valor a partir do teste exato de Fisher.

Com a matriz montada, é utilizada a Equação 3.1 para efetuar o cálculo do p-valor. Com os valores de p-valor, é montada uma matriz que relaciona os *sets* de genes com os *arrays*, efetuando-se um arredondamento dos valores de p-valor.

$$p = 1 - \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}} = 1 - \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,N!} \tag{3.1}$$

Também foram gerados arquivos que contêm filtros capazes de selecionar apenas quais *arrays* com seus respectivos *sets* de genes possuem o p-valor $< 0{,}05$. O fator limitante para o p-valor são para valores menores que 0,05 pelo fato de serem esperados por acaso, ou seja, os selecionados da relação *set* de gene com *array* são muito importantes para os dados de microarranjo.

3.6.5 Obtenção e tratamento dos *Clusters*

Com a obtenção da matriz que relaciona os *sets* de genes com os *arrays*, é possível obter *clusters* (agrupamentos) a partir de uma função de correlação entre eles. Os *clusters* foram obtidos através do pacote da linguagem R chamado **pvclust** (SUZUKI; SHIMODAIRA, 2013). Os parâmetros utilizados foram: método aglomerativo (*method.hclust*) como sendo a média, distância euclidiana (*method.dist*) como sendo a correlação, e número de replicações de *bootstrap* igual a 1000. Para cada *cluster* em agrupamento hierárquico são calculados p-valores através do método de *multiscale bootstrap resampling*. O p-valor de um *cluster* é um valor entre 0 e 1 que indica se é correlacionado pelos dados. Há dois tipos de p-valores: o AU (*Approximately Unbiased*) e o BP (*Bootstrap Probability*). A diferença entre os dois p-valores está no modo que são calculados. Enquanto o p-valor AU é calculado através do método *multiscale bootstrap resampling*, o p-valor BP é calculado através do método normal de *bootstrap resampling*. Tomando por base amostras do bootstrap, esse método calcula o p-valor para cada hipótese. Se o p-valor de uma hipótese é muito pequeno (menor que 5%) podemos rejeitar a hipótese. A probabilidade de bootstrap é uma aproximação desse valor, e o método *multiscale bootstrap resampling* corrige o viés da probabilidade de *bootstrap*.

O algoritmo do método *multiscale bootstrap resampling* está esquematizado na Figura 3.7. Primeiro, geramos amostras de *bootstrap* para cada tamanho de amostra. Em seguida, á aplicada a clusterização hierárquica para cada amostra de *bootstrap* para obter os conjuntos das replicações de *bootstrap* de dendogramas. Calcula-se a probabilidade de bootstrap para cada tamanho de amostra. Por último, usando os valores das probabilidades de *bootstrap*, pode-se estimar o p-valor ajustando-os a uma equação teórica. O p-valor estimado é chamado de AU.

Além da obtenção dos *clusters*, também é construida uma "árvore"que contém a relação de cada *cluster*, onde as folhas correspondem a algum *set* de gene G que estão associados com um vetor (indexado por *arrays*) que tem valor zero em qualquer lugar, exceto para entradas que correspondem aos *arrays* nos quais o *set* de gene G é significantemente induzido (ou reprimido), ou seja, é feita a verificação da fração (positiva ou negativa) de genes do *set* de gene G que são induzidos (ou reprimidos) no *array* A, por exemplo. Cada nodo interno é associado com um

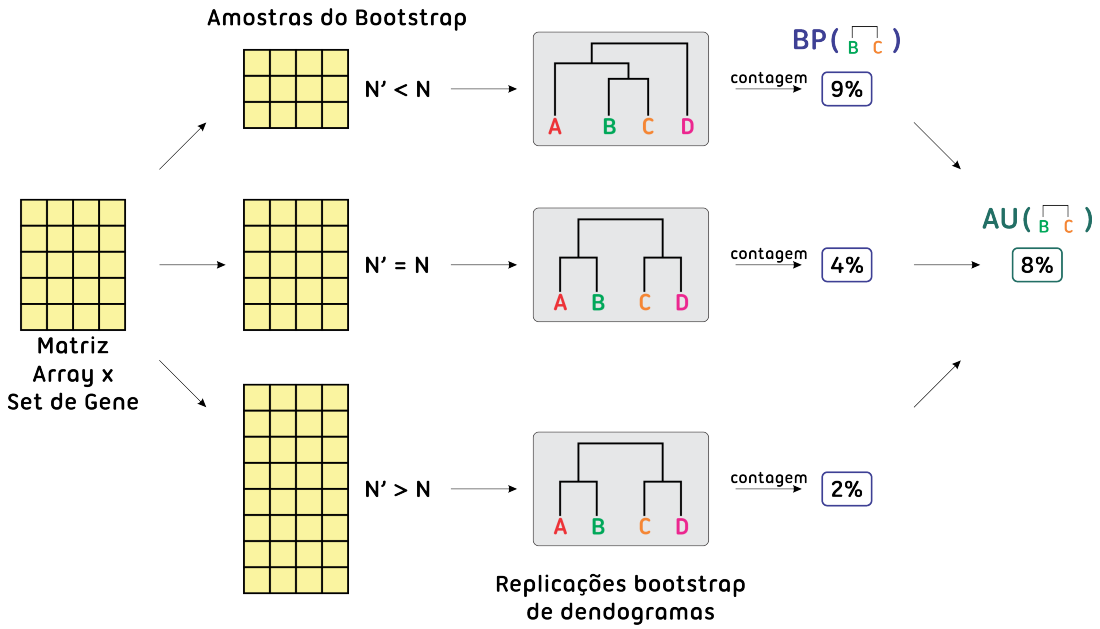

Figura 3.7: Exemplo da metodologia do *multiscale bootstrap resampling*. Nesse exemplo o valo de AU é 8%, portanto não é possível rejeitar a possibilidade de que os dados sejam obtidos sob a hipótese de que B e C são mais próximos.

vetor representando a média de todos os *sets* de genes descendentes na árvore.

A construção da árvore gerada anteriormente se dá pela utilização de dois pacotes em R, o **ade4** (DRAY; DUFOUR et al., 2007) e o **ape** (PARADIS; CLAUDE; STRIMMER, 2004). A principal utilização de ambos possuem é voltada para filogenia. A partir da representação gráfica da árvore (Figura 3.8) é possível definir o valor de cada nodo interior. Dessa forma, obtém-se quais *clusters* são mais consistentes.

Testa-se a expressão do gene g de forma a verificar se esta é consistente com as mudanças significativas na expressão do *set* de gene G. Para isso, é calculado o valor de um *score* que é dado de acordo com a Equação 3.2, onde p_a é a fração dos genes no *array* a que são induzidos (ou reprimidos) para os *arrays* em I (ou em R). Além do valor do *score* são calculados os valores de média e variância de acordo com Equação 3.3 e Equação 3.4, respectivamente. Tais cálculos são realizados para que sejam selecionados conjuntos de *clusters* com características semelhantes.

$$Score(g) = \sum_{\text{a}\in\text{I}|g \text{ está induzido em } a} -log(p_a) + \sum_{\text{a}\in\text{R}|g \text{ está reprimido em } a} -log(p_a) \tag{3.2}$$

$$\mu = \sum_{\text{a}\in\text{I}\cup\text{R}} -p_a \cdot log(p_a) \tag{3.3}$$

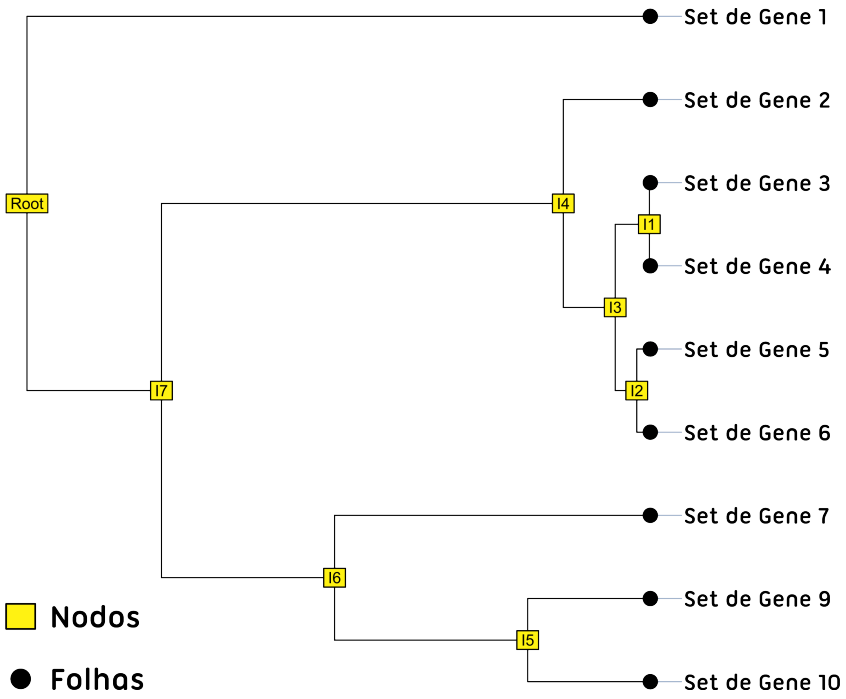

Figura 3.8: Exemplo de uma "árvore"montada com os dados de *sets* de genes além da identificação dos nodos e folhas.

$$\sigma^2 = \sum_{a \in I \cup R} -p_a(1-p_a) \cdot log^2(p_a) \tag{3.4}$$

3.6.6 Construção do *Heatmap*

Para a obtenção dos *heatmaps* foi utilizado o pacote em R chamado **superheat**. Mas para a geração do *heatmap* é necessário obter uma matriz que relacione os *clusters* com as condições clínicas fornecidas pelo Projeto Toxicogenômico Japonês. Essa relação é criada de acordo com os passos do fluxograma da Figura 3.9. Lembrando que os *sets* de genes correspondem aos dados de *GO*, *KEGG* e *Reactome*, enquanto os *arrays* correspondem às drogas relacionadas com diferentes concentrações de dose e tempo de amostragem.

O resultado final da análise do mapa modular está representado na Figura 3.10. As informações presentes são referentes, inicialmente, aos *clusters* e como eles se relacionam com as condições clínicas, dessa forma gerando o *heatmap*. Complementar a essas informações há também a relação entre os *arrays* por condições clínicas e também a relação entre genes por *clusters*. Esse conjunto todo forma, no final, o mapa modular.

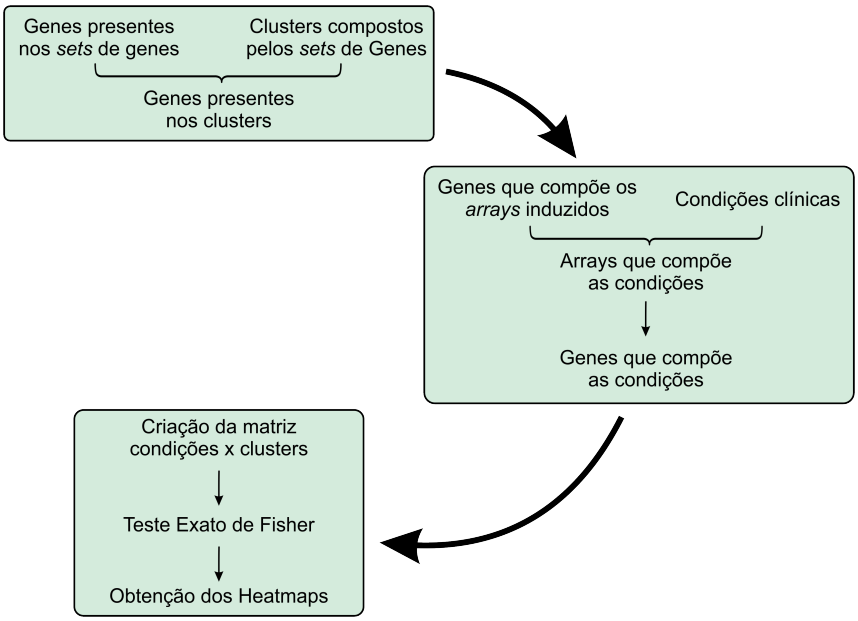

Figura 3.9: Fluxograma para a obtenção do *Heatmap*.

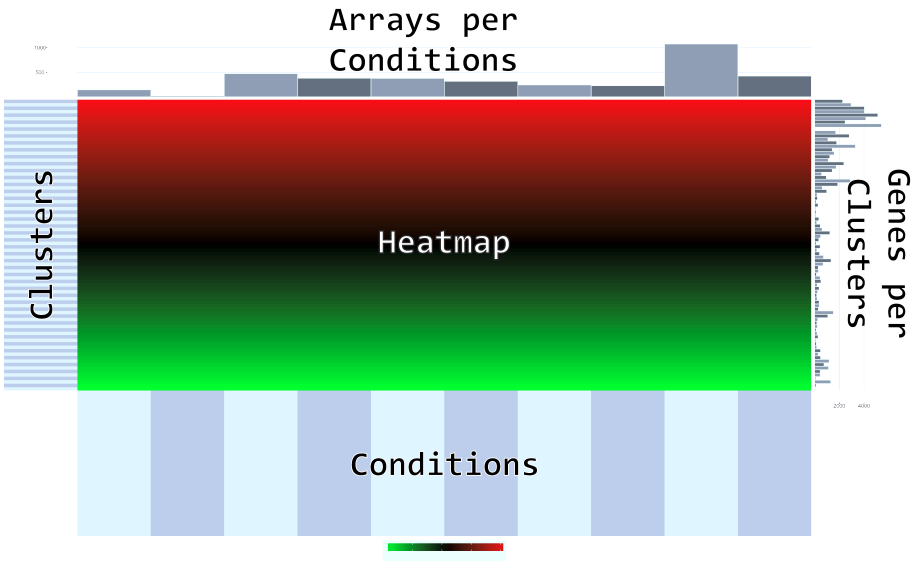

Figura 3.10: Informações contidas no mapa modular. O mapa modular é dividido em 5 partes: condições clínicas, *clusters*, genes por *clusters*, *arrays* por condições clínicas e o heatmap.

4 Resultados e Discussão

Nesse trabalho utilizamos os modelos de experimento para *Homo sapiens in vitro*, *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo single*. Não utilizamos os dados referentes a *Rattus norvegicus in vivo repeat* pelo fato do tempo de amostragem utilizados são variados por dias, dessa forma não há como comparar os dados desse experimento com os outros 3 que estão variando em dias. Então, toda vez que referir a *Rattus norvegicus in vivo* entende-se que são os dados para *Rattus norvegicus in vivo single*. De acordo com a Tabela 1.4 há uma variação entre as concentrações de dose e também entre os tempos de amostragem para cada tipo de experimento. Esses parâmetros coincidem para os três tipos de experimentos quando a concentração de dose é **alta** e tempo de amostragem de **24 horas**, dessa forma podemos comparar os resultados obtidos entre os experimentos para a concentração de dose e tempo de amostragem coincidente.

4.1 Análise Clássica Global

A análise global envolve todas as drogas, concentrações de doses e tempos de amostragem. Foram utilizados dados somente da normalização do tipo RMA.

4.1.1 Genes Diferencialmente Expressos

A partir da Figura 4.1 é possível observar a média da quantidade de genes diferencialmente expressos presentes nas 131 drogas para os três experimentos em relação aos diferentes tempos de amostragens e concentrações de doses. Observa-se que há um padrão em relação a média da quantidade de genes diferencialmente. Quanto mais baixa for a concentração de dose e quanto menor for o tempo de amostragem, menor será a quantidade de genes diferencialmente expressos presentes. Enquanto que quanto mais alta for a concentração de dose e maior o tempo de amostragem, maior a quantidade de genes diferencialmente expressos presentes.

A Figura 4.2 mostra um panorama geral das quantidades de genes diferencialmente ex-

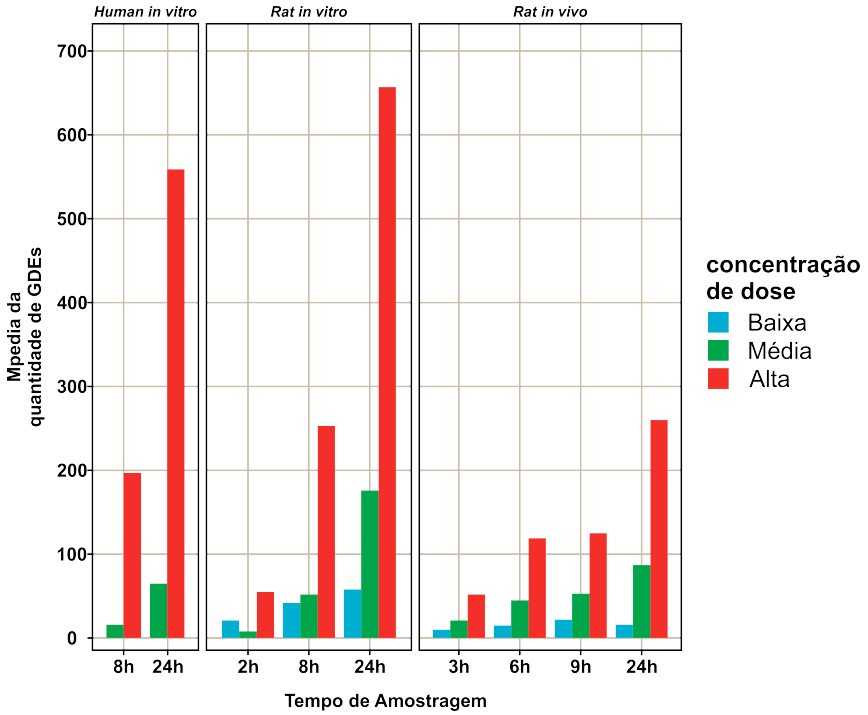

Figura 4.1: Média da quantidade de genes diferencialmente expressos presentes em todas as drogas para os 3 tipos de experimento (*Homo sapiens in vitro*, *Rattus norvegicus in vitro* e *Rattus noevegicus in vivo*) em relação aos diferentes tempos de amostragens e concentrações de doses.

pressos presentes para os experimentos *Homo sapiens in vitro*, *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo* para cada uma das 131 drogas, considerando todas as variações de concentrações de dose e tempos de amostragem. Observa-se que para algumas drogas específicas há uma grande quantidade de genes diferencialmente expressos para um determinado modelo, ou mais, enquanto que os outros modelos possuem uma pequena quantidade de genes diferencialmente expressos. Mais diferenças entre os modelos serão constatadas durante essa seção. A tabela completa com todas as quantidades está no Apêndice B. É possível também analisar as quantidades de genes diferencialmente expressos em concentrações de doses e tempos de amostragem específicos.

Através da análise da tabela presente no Apêndice B é possível identificar algumas drogas que possuem variações muito evidentes para os três experimentos. A partir de uma busca no banco de dados *PubMed* com as palavras chave sendo o nome de determinada droga e seu respectivo tipo de experimento, obteve-se como resultado uma gama de estudos que faziam referência às palavras chave pesquisadas. A busca realizada a partir de drogas com maiores quantidades de genes diferencialmente expressos retornaram mais estudos relacionados, enquanto que a pesquisa realizada com drogas com menores quantidades de genes diferencialmente expressos retornaram uma menor quantidade de estudos relacionados. É necessário uma análise mais aprofundada para obter conclusões mais sólidas.

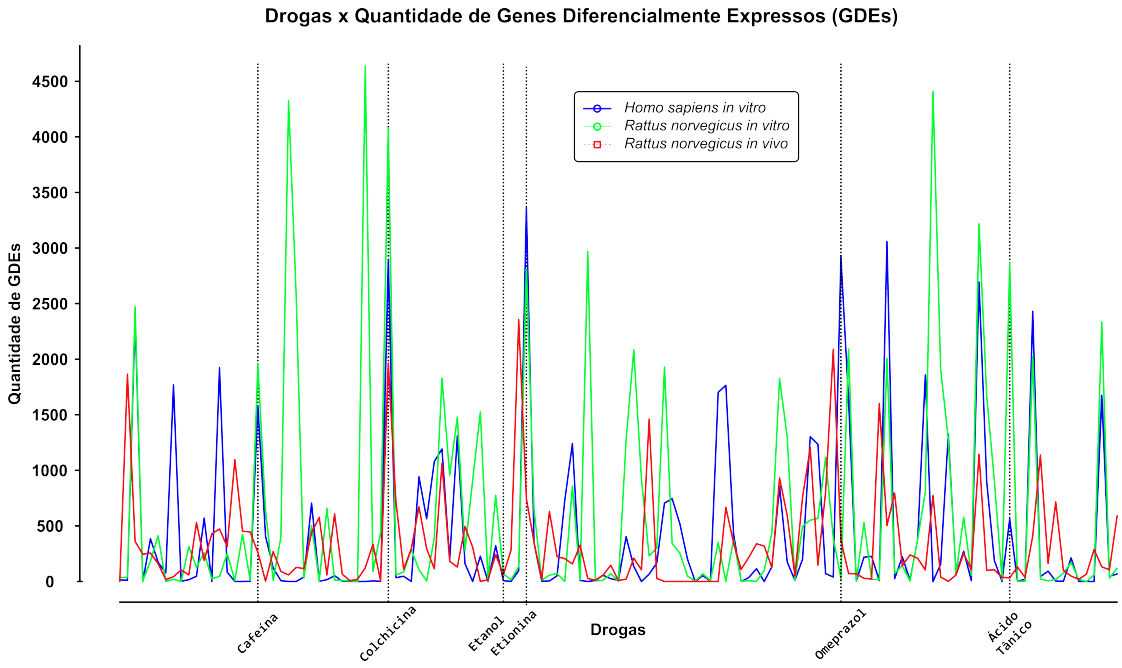

Figura 4.2: Quantidade de genes diferencialmente expressos presentes para cada uma das 131 drogas, considerando todas as variações de concentrações de dose e tempo de amostragem. Em destaque, as 6 drogas selecionadas.

4.1.2 Enriquecimento Funcional

A Figura 4.3, Figura 4.4 e Figura 4.5 mostram a quantidade de GOs, KEGGs e REACTOMES enriquecidos para as 131 drogas e para os três experimentos, respectivamente. Destaca-se nessas figuras as seis drogas selecionadas anteriormete.

Nota-se uma grande variação na quantidade de rotas e vias metabólicas enriquecidas entre GOs, KEGGs e REACTOMEs. Tal diferença se dá pela montagem distinta dos respectivos bancos de dados. Enquanto que o GO possui informações a respeito dos componentes celulares, funções moleculares e processos biológicos obtidos a partir de análises experimentais, o KEGG é montado a partir de informações *online* e o REACTOME a partir de informações da literatura. Todo este conjunto confere esta diferença de enriquecimento entre os bancos de dados.

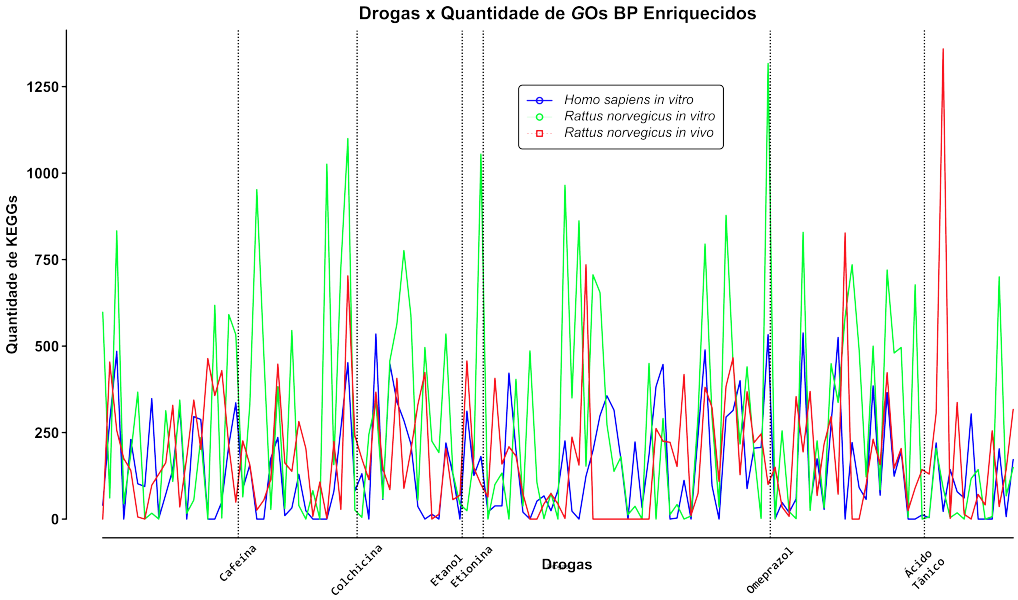

Figura 4.3: Quantidade de GOs enriquecidas para as 131 drogas nos três experimentos: em azul representando *Homo sapiens in vitro*, em verde *Rattus norvegicus in vitro* e em vermelho *Rattus norvegicus in vivo*.

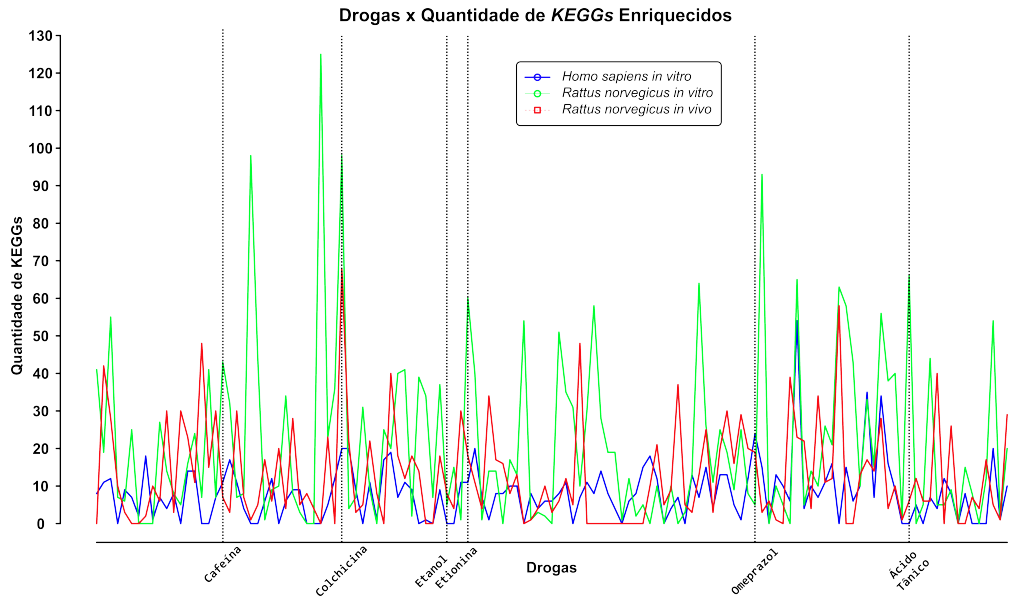

Figura 4.4: Quantidade de KEGGs enriquecidos para as 131 drogas nos três experimentos: em azul representando *Homo sapiens in vitro*, em verde *Rattus norvegicus in vitro* e em vermelho *Rattus norvegicus in vivo*.

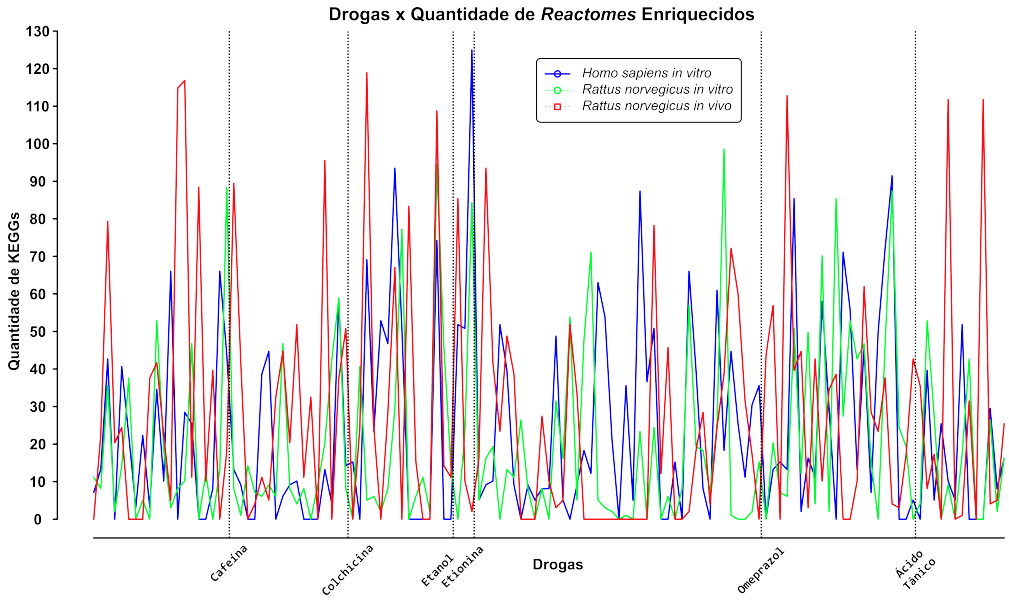

Figura 4.5: Quantidade de REACTOMEs enriquecidos para as 131 drogas nos três experimentos: em azul representando *Homo sapiens in vitro*, em verde *Rattus norvegicus in vitro* e em vermelho *Rattus norvegicus in vivo*.

4.2 Análise Clássica Local

A análise local envolve as 6 drogas selecionadas (ácido tânico, cafeína, omeprazol, etanol, colchicina e etionina) com concentração de dose alta e tempo de amostragem igual a 24 horas. A seleção destas 6 drogas foram baseadas na alta toxicidade. Foram utilizados dados somente da normalização do tipo RMA.

4.2.1 Genes Diferencialmente Expressos

O número de genes diferencialmente expressos para cada experimento teve uma alta variação como pode ser observada na Figura 4.6 para a concentração de dose alta e tempo de amostragem de 24h. Tal fato mostra que há uma diferença considerativa entre os três modelos apresentados. A partir da análise da Figura 4.6 pode-se observar que há um elevado número de genes diferencialmente expressos (maior que 1700) para *Homo sapiens in vitro* nas drogas cafeína, etionina e omeprazol. Em contrapartida, *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo* possuem menor quantidade de genes diferencialmente expressos em relação ao experimento para *Homo sapiens*. Enquanto que nas drogas ácido tânico, colchicina e etanol observa-se uma maior quantidade de genes diferencialmente expressos para *Rattus norvegicus in vitro* quando comparados com *Homo sapiens in vitro* e *Rattus norvegicus in vivo*.

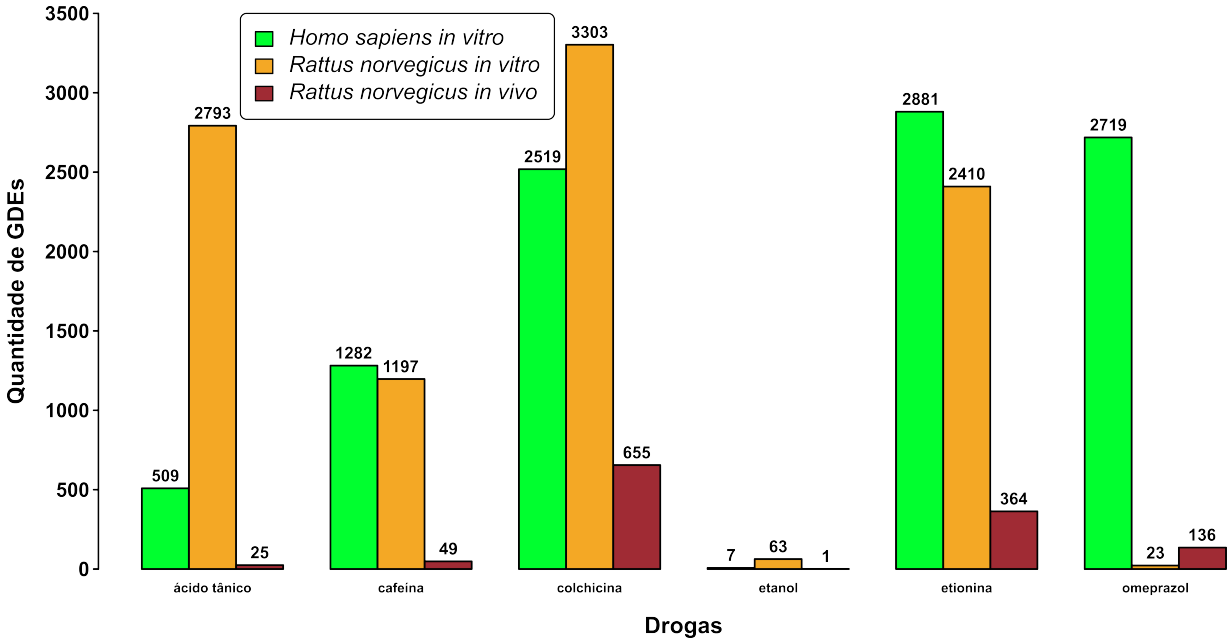

Figura 4.6: Gráfico de barras mostrando a quantidade de genes diferencialmente expressos (GDEs), para concentração de dose alta e tempo de amostragem de 24h para as 6 drogas selecionadas que compõe cada tipo de experimento (*Homo sapiens in vitro*, *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo*).

Quando as 6 drogas são analisadas de maneira individual observa-se a quantidade de genes

Droga	*Human in vitro*	*Rat in vitro*	*Rat in vivo*
Ácido Tânico	508	2792	25
Cafeína	1281	1196	48
Colchicina	2518	3302	654
Etanol	7	62	1
Etionina	2880	2409	363
Omeprazol	2718	22	135

Tabela 4.1: Quantidade de genes diferencialmente expressos presentes para as seis drogas selecionadas e sua respectiva distribuição para os três experimentos. Dados com concentração de dose alta e tempo de amostragem de 24 horas.

diferencialmente expressos presentes em cada uma. A Tabela 4.1 possui a relação de genes diferencialmente expressos presentes nos 3 experimentos para concentração de dose alta e tempo de amostragem de 24 horas, enquanto que a Figura 4.7 mostra a distribuição desses genes diferencialmente expressos nos três experimentos (*Homo sapiens in vitro*, *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo*).

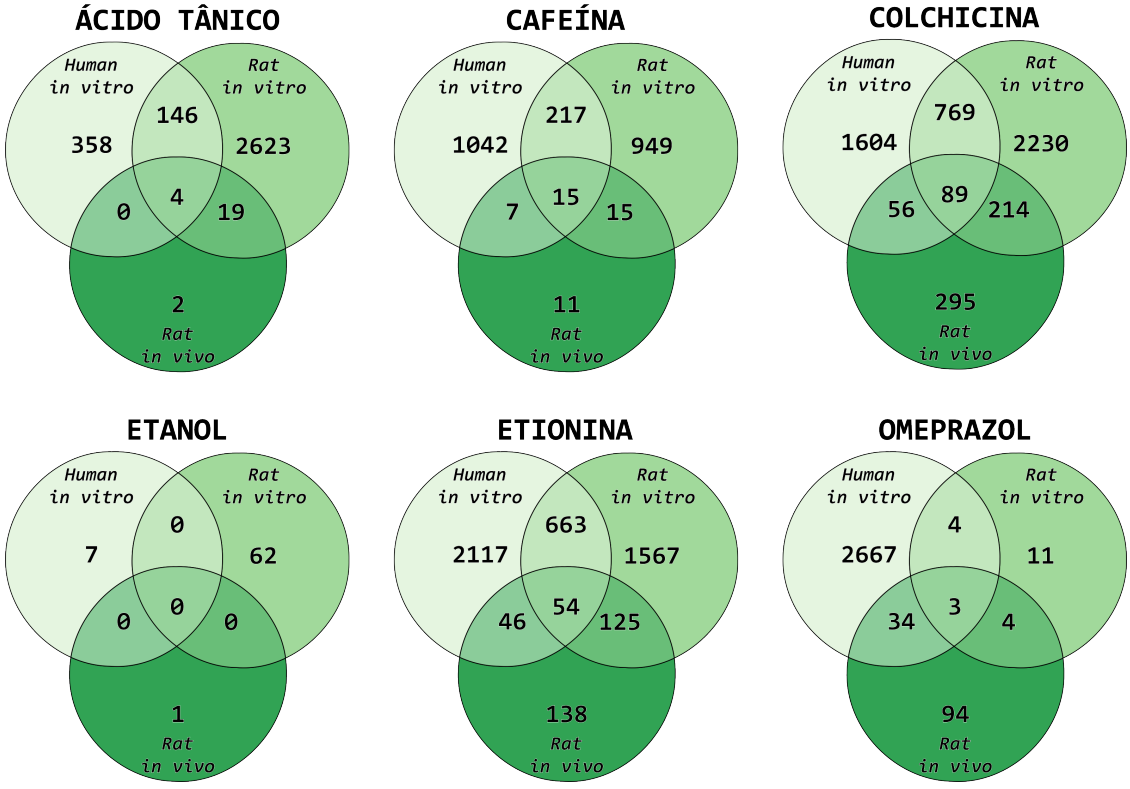

Figura 4.7: Distribuição da quantidade de genes diferencialmente expressos da Tabela 4.1 presentes nos três experimentos, incluindo as intersecções. Dados com concentração de dose alta e tempo de amostragem de 24 horas.

Nota-se na Tabela 4.1 e Figura 4.7 que há uma grande variação na quantidade de genes diferencialmente expressos para cada um dos três experimentos. Por exemplo, observando a droga Ácido Tânico notamos que há 3285 genes diferencialmente expressos para o experimento

Rattus norvegicus in vitro, enquanto que há 628 para *Homo sapiens in vitro* e 31 para *Rattus norvegicus in vivo*. Tal diferença na quantidade se dá pela complexidade e diferenças entre cada modelo. Outra observação importante é que na Figura 4.7 há somente intersecção dos genes diferencialmente expressos para os experimentos *Rattus norvegicus in vitro* e *Rattus norvegivus in vivo*, mostrando, assim, as diferenças de distribuição entre os experimentos.

Foram geradas tabelas e diagramas de *Venn* que contém informações a respeito das quantidades de genes diferencialmente expressos para as 6 drogas selecionadas com todas as variações de concentrações de doses e tempos de amostragens. Essas tabelas e diagramas podem ser encontrados no Apêndice C. Da mesma forma que foi abordado acima, há uma grande variação na quantidade de genes diferencialmente expressos, dessa forma mostrando, mais uma vez, a diferença entre os modelos.

Analisando com mais detalhes 3 (cafeína, colchicina e omeprazol) das 6 drogas selecionadas é possível comparar as doses utilizadas no PTGJ com as doses mais utilizadas no dia a dia.

i) **Cafeína:** no PTGJ foi utilizado a dose média de 15 mg e a dose alta de 75 mg. Se compararmos essas doses com produtos que são consumidos no dia a dia observamos que se assemelham à doses de cafeína presentes, por exemplo, no café americano (60-120 mg), lata de coca-cola (34 mg) e energético (80 mg);

ii) **Colchicina**: no PTGJ foi utilizado a dose média de 8 mg e a dose alta de 40 mg. Quando comparado com a dose utilizada no tratamento de gota aguda, verificamos que é utilizado 1.2 mg. Dose essa que é considerada muito menor que a média para o PTGJ;

iii) **Omeprazol**: no PTGJ foi utilizado a dose baixa de 2.4 mg, dose média de 12 mg e a dose alta de 60 mg. Quando comparamos com as doses utilizadas no tratamento da síndrome de *Zollinger-Elisson* (60 mg ou 80 mg) e no tratamento da úlcera gástrica (40 mg) verificamos que são doses compatíveis com as utilizados no PTGJ.

4.2.2 Enriquecimento Funcional

A partir da obtenção dos genes diferencialmente expressos foi realizado o enriquecimento funcional de *GO*, *KEGG* e *Reactome* para as 6 drogas e constatou-se uma grande variação de quantidades de ontologias e vias para cada experimento. A Figura 4.8 mostra a quantidade de genes diferencialmente expressos quando comparados com a quantidade de *KEGGs*, *REACTOMEs* e ontologias dos tipos processos biológicos, funções moleculares e componentes celulares

presentes que foram enriquecidos para os três experimentos.

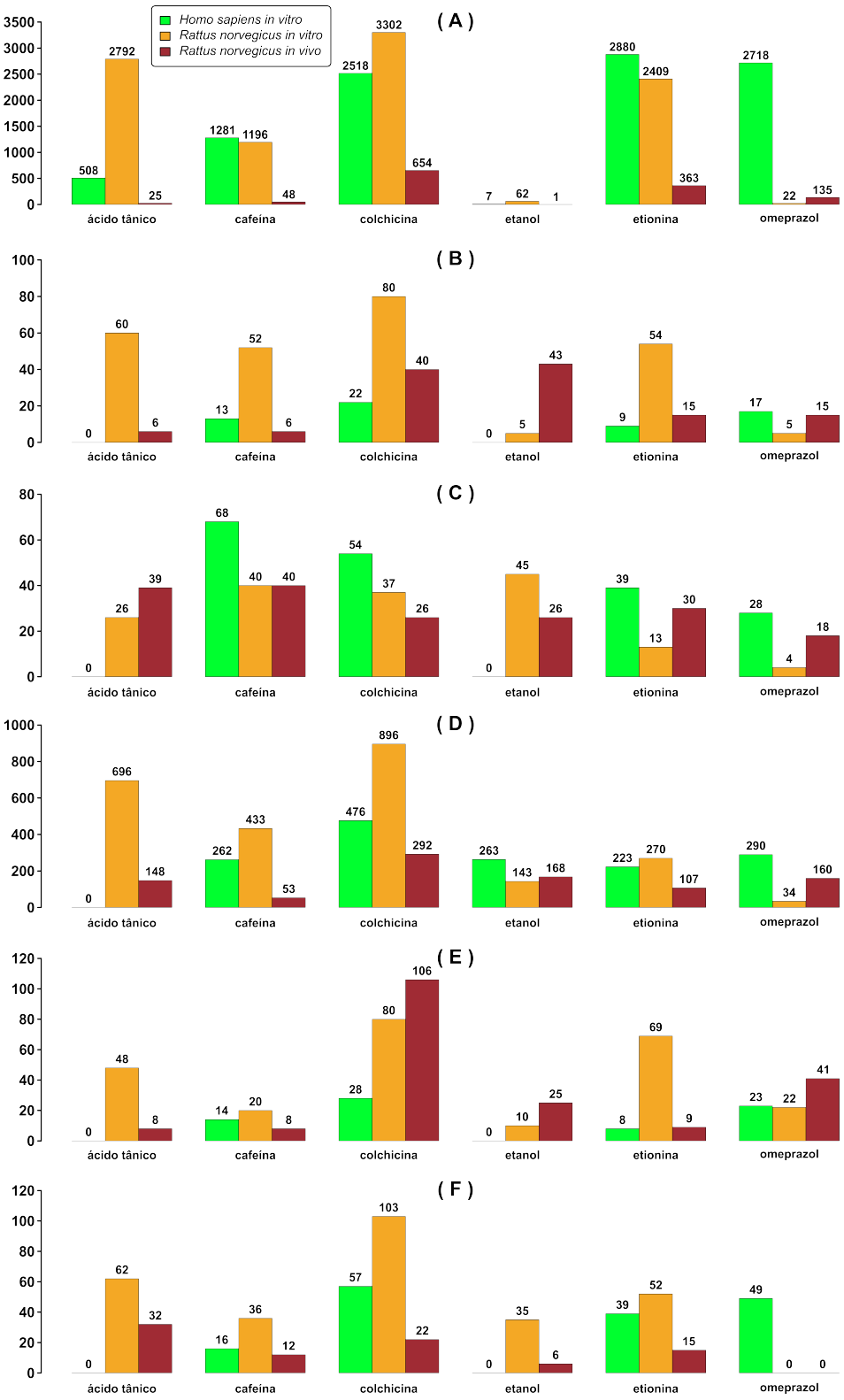

Figura 4.8: Gráficos de barra comparando a quantidade de genes diferencialmente expressos com a quantidade de vias e rotas metabólicas enriquecidas para as 6 drogas e com concentração de dose alta e tempo de amostragem de 24 horas. **(A)** Quantidade de genes diferencialmente expressos. **(B)** Quantidade de *KEGGs* enriquecidos. **(C)** Quantidade de *REACTOMEs* enriquecidos. **(D)** Quantidad de *GOs* do tipo processos biológicos enriquecidos. **(E)** Quantidade de *GOs* do tipo funções moleculares enriquecidos. **(F)** Quantidade de *GOs* do tipo componentes celulares enriquecidos.

Após a quantificação das ontologias e vias presentes em cada experimento para as 6 drogas, iremos analisar com mais detalhes como elas estão distribuídas além de identificar as principais funções e rotas envolvidas nessas drogas. A partir da Figura 4.9 podemos observar essa distribuição para as ontologias do tipo processo biológico e constatamos que das 6 drogas apenas uma (ácido tânico) não possui ontologias do tipo processos biológicos em comum para os 3 experimentos. A partir de uma análise mais detalhada é possível identificar esses processos envolvidos, tanto os exclusivos e a intersecção, para os experimentos. Por exemplo, para a Etionina há 10 processos biológicos em comum para os 3 experimentos. Os principais processos são responsáveis pelo processo catabólico (GO:0046395, GO:0016054, GO:0044282 e GO:0072329). Estudos como (SHARMA; SINGH; KANWAR, 2014) mostram que a enzima L-metionase desempenha um papel importante nas células tumorais. Dessa forma, nesse estudo são abordados meios que se utilizam da L-metionase para o tratamento de doenças.

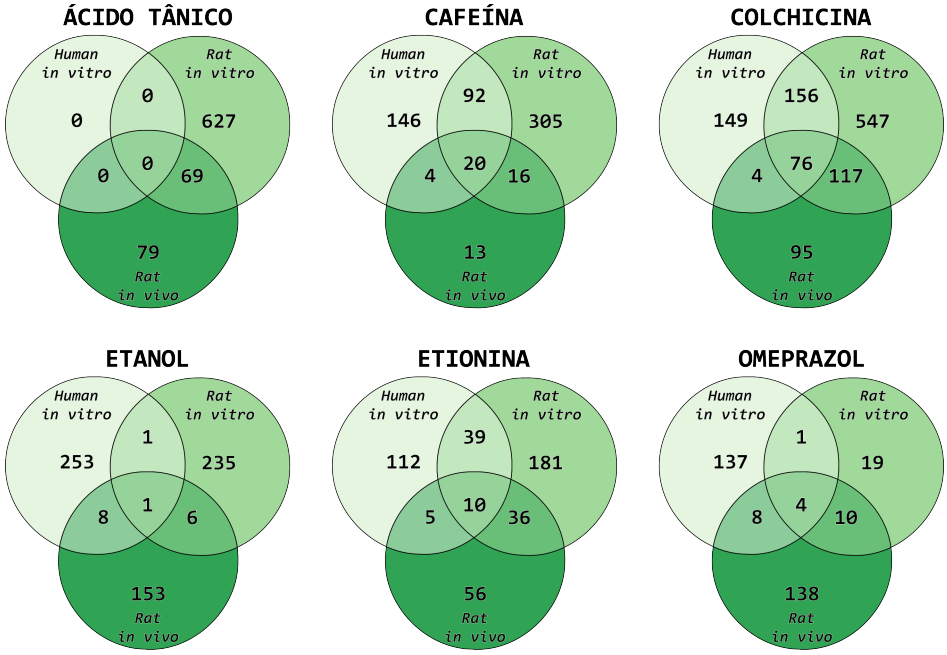

Figura 4.9: Relação da quantidade de *GOs* do tipo processo biológico para as 6 drogas selecionadas em relação a cada um dos três experimentos.

A Figura 4.10 também mostra a relação comentada acima, mas dessa vez para *KEGG*. Olhando para a cafeína podemos observar 3 vias em comum. Essas vias são responsáveis sinalização Fox0, ciclo celular e sinalização do genes p53. Todas essas 3 vias possuem ligação direta com tumores. A sinalização Fox0 é uma família de de fatores de transcrição que regulam a expressão de genes em eventos fisiológicos celulares incluindo apoptose, controle do ciclo celular e etc. Estudos como (BODE; DONG, 2007) desvendam o efeito da cafeína no ciclo celular e câncer, vias muito enriquecidas para os três experimentos. O estudo sugere que o gene supressor de tumor (p53) é o mediador primário do controle do ciclo celular e responsável pela

indução da apoptose na maioria das células responsáveis pelo dano ao DNA.

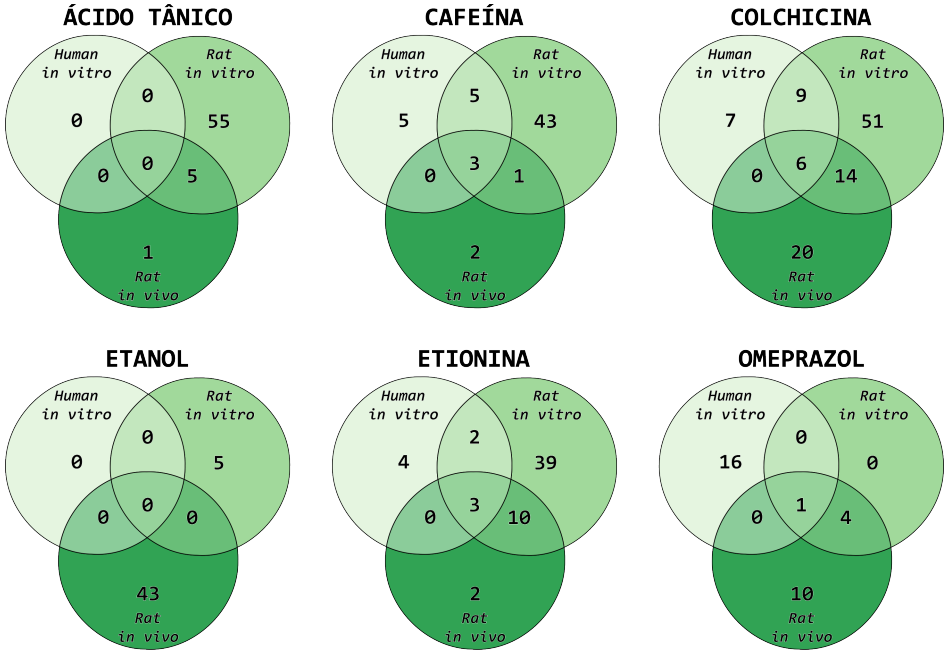

Figura 4.10: Relação da quantidade de *KEGGs* para as 6 drogas selecionadas em relação a cada um dos três experimentos.

A análise de enriquecimento funcional é muito abrangente, pois temos 3 diferentes tipos de experimentos em que há variações tanto individuais e de intersecções, ou seja, há muitas análises que podem ser feitas em torno dessas informações. Acima foram apresentados resultados apenas para ontologias do tipo processos biológicos e *KEGG*. Não será detalhado para *Reactome*, pois a partir do enriquecimento não houveram intersecções entre os experimentos, obteve-se somente vias que eram exclusivas.

4.3 Mapa Modular

4.3.1 Limitação Computacional

Os dados do *Gene Ontology* são divididos em 3 tipos, os processos biológicos, funções moleculares e componentes celulares. Além disso, existe uma hierarquia. A primeira camada são as *GOs* pais, em seguida as ancestrais, filhas e descendentes. A relação entre os termos baseia-se no fato dos pais terem uma especificidade menor que as filhas. A partir dos termos das *GOs* e suas correspondentes filhas, encontramos a relação de genes que compõe cada *GO* com as respectivas filhas. Existe uma limitação computacional para realizar a clusterização hierárquica devido ao grande número de genes e *GOs* totais. Dessa forma, foi necessário realizar um filtro para os três tipos de ontologias (processos biológicos, funções moleculares e componentes celulares) para diminuir esse número total. O primeiro filtro foi descartar todas as *GOs* que possuíam menos de 10 genes e mais de 3000 genes. Esse filtro foi de extrema importância pois a partir dele eliminamos as *GOs* mais genéricas (por exemplo, processo metabólico). A Tabela 4.2 mostra qual foi a diferença das quantidades de *GOs* obtidas antes e após o filtro aplicado tanto para *Homo sapiens in vitro* e *Rattus norvegicus*. Nota-se que houve uma diminuição significativa da quantidade das *GOs*. Assim sendo, atrelada a aplicação de técnicas de computação paralela (várias operações matemáticas são distribuídas para os *cores* do computador para serem realizados ao mesmo tempo e serem solucionados no menor tempo possível), foi possível obter os resultados do mapa modular com uma maior eficiência. Antes da aplicação do filtro o tempo estimado para a obtenção do mapa modular era de aproximadamente 30 dias para alguns casos, mas com a aplicação dos filtros e da paralelização foi possível otimizar esse tempo para alguns minutos. Para as análises a seguir é necessário ressaltar que os valores representados nos *heatmaps* foram normalizados para cada caso específico a fim de facilitar a visualização dos dados.

	Total	**Corte**
Processos Biológicos	28761	9828
Funções Moleculares	8397	2087
Componentes Celulares	3191	1340

Tabela 4.2: Diferença entre a quantidade total de *GOs* disponíveis inicialmente em contraste com a quantidade de *GOs* após a aplicação do filtro.

4.3.2 Análises

Para *Rattus norvegicus in vitro* relacionado com *GOs* do tipo funções moleculares foi gerada a Figura 4.11. A partir do método de *multiscale bootstrap resampling* foram obtidos 40 *clusters*, que tiveram sua consistência testadas e comparadas, dessa forma possibilitando a construção do *heatmap*. Analisando a Figura 4.11 , podemos identificar alguns blocos que estão mais significantemente induzidos (em tons de vermelho) e outros que estão mais reprimidos (em tons de verde). Esses blocos formam perfis de *clusters* que possuem *GOs* diferentes atuando da mesma forma, induzindo ou reprimindo, em condições diferentes. Os blocos amarelos representados na Figura 4.11 destacam um perfil induzido e outro reprimido. A Tabela 4.3 mostra as *GOs* presentes no perfil que está induzido para as condições edema, proliferação, vacuolização nuclear, fibrose, nódulo hepatodiafragmático, morte celular e degeneração acidófila e basófila. A relação dos *clusters* como estão apresentadas no *heatmap* é obtida a partir de qual ou quais drogas com diferentes concentrações de dose e tempos de amostragem estão influenciando diretamente a ocorrência de uma determinada condição. Dessa forma, obtém-se uma lista de drogas que estão diretamente ligadas a uma condição. Por exemplo, tomando por base o perfil escolhido na Figura 4.11, as condições fibrose e nódulo hepatodiafragmático possuem como relação principalmente as drogas: acarbose (STANDL et al., 2014) (dose alta, tempo de amostragem 24 horas), clorofibrato (dose alta, tempo de amostragem 24 horas), amiodarona (SILVA et al., 2006) (dose alta, tempo de amostragem 6 horas), e muitas outras. No caso do outro perfil destacado na figura que está reprimido, temos dois *clusters* que são responsáveis por reprimir as seguintes condições: vacuolização citoplasmática, mudança basofílica, necrose de célula única, microgranuloma, alteração eosinófila, tumor, infiltração celular, necrose, aumento da mitose hipertrofia e alteração acidófila. As funções moleculares responsáveis por reprimir essas condições estão mostradas na Tabela 4.4. Quando esse modelo é comparado com *Homo sapiens in vitro* e *Rattus norvegicus in vivo* nota-se que há diferença na identificação dos *clusters* e dos perfis, ou seja, há muita diferença entre esses modelos para *GO* do tipo função molecular.

Ontologias	Descrição
GO:0051059	NF-kappaB binding
GO:0051536	iron-sulfurcluster binding
GO:0051539	4 iron, 4 sulfur cluster binding
GO:0098641	cadherin binding involved in cell-cell adhesion
GO:0019003	GDP binding
GO:0051087	chaperone binding
GO:0003995	acyl-CoA dehydrogenase activity
GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors
GO:0004177	aminopeptidase activity
GO:0051020	GTPase binding
GO:0004298	threonine-type endopeptidase activity
GO:0016209	antioxidant activity
GO:0046933	proton-transporting ATP synthase activity, rotational mechanism

Tabela 4.3: Tabela com as respectivas *GOs* presentes no perfil muito induzido para as condições: edema, proliferação, vacuolização nuclear, fibrose, nódulo hepatodiafragmático, morte celular e degeneração acidófila e basófila.

Ontologias	DescriÃ§Ã£o
GO:0004176	ATP-dependentpeptidase activity
GO:0005549	odorant binding
GO:0001054	RNA polymerase I activity
GO:0001055	RNA polymerase II activity
GO:0001671	ATPase activator activity
GO:0003954	NADH dehydrogenase activity
GO:0004935	adrenergic receptor activity
GO:0004984	adrenergic receptor activity
GO:0005344	oxygen transporter activity
GO:0008527	taste receptor activity
GO:0016651	oxidoreductase activity, acting on NAD(P)H
GO:0034450	ubiquitin-ubiquitin ligase activity
GO:0043024	ribosomal small subunit binding

Tabela 4.4: Tabela com as respectivas *GOs* presentes no perfil reprimido para as condições: vacuolização citoplasmática, mudança basofílica, necrose de célula única, microgranuloma, alteração eosinófila, tumor, infiltração celular, necrose, aumento da mitose hipertrofia e alteração acidófila.

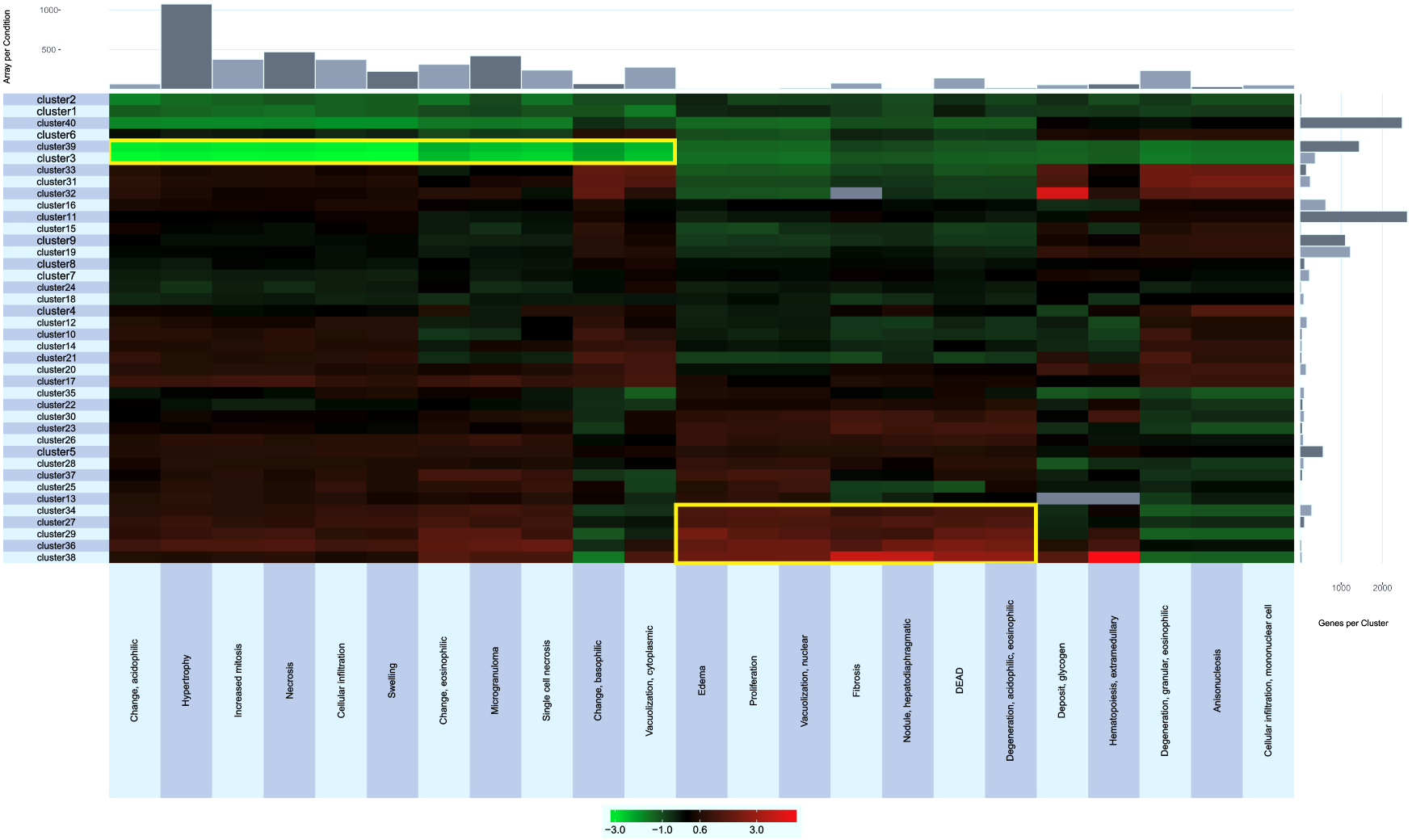

Figura 4.11: *Heatmap* gerado para o experimento *Rattus norvegicus in vitro* relacionado com GO do tipo função molecular. As caixas amarelas estão evidenciando 2 tipos de perfil, um induzido (predominância de vermelho) e outro reprimido (predominância de verde).

No caso de *Homo sapiens in vitro* relacionado com os dados de *Reactome*, dessa relação foi gerado o *heatmap* representado na Figura 4.12. A Tabela 4.5, construída a partir dos genes induzidos/reprimidos, mostram valores do *score* (Equação 3.2), média (Equação 3.3) e variância (Equação 3.4) para cada *set* de gene relacionado com *array*. Essa tabela é importante, pois a partir dela é possível testar se a expressão de um determinado gene é consistente com as mudanças significativas na expressão de um determinado *set* de gene. A partir do método de *multiscale bootstrap resampling* foram obtidos 12 *clusters*. O perfil que representa indução marcado na caixa amarela presente na Figura 4.12 possui 99 vias que fazem parte de 3 *clusters*. Essas vias estão ligadas, principalmente, a funções tais como: metabolismo de aminoácidos e derivados, oxidações biológicas, doenças do metabolismo, complexo promotor da anáfase, etc. Enquanto que o perfil que indica repressão, também marcado em amarelo na Figura 4.12, possui 16 vias que fazem parte de 4 *clusters*. As principais funções dessas vias são as controle da fase G1/S, relacionadas ao DNA (reparo, dano, evitar dano), sinalização do ácido retinoico, etc. Quando esse modelo é comparado com *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo* nota-se que também há diferença na identificação dos *clusters* e dos perfis, ou seja, mais uma vez, há muita diferença entre esses modelos para *Reactome*.

SetGene_Array	***Score***	**Média**	**Variância**
R-HSA-162582_omeprazole_High_24h	37.27287	33.05957	-0.11748
R-HSA-1640170_methimazole_High_24h	35.84561	33.16533	-0.31144
R-HSA-1430728_omeprazole_High_24h	34.99137	31.07374	-0.14475
R-HSA-162582_colchicine_Middle_24h	33.40208	29.05393	-0.10655
R-HSA-162582_phenobarbital_High_24h	33.40208	29.05393	-0.10655
R-HSA-69278_methimazole_High_24h	32.76001	30.17468	-0.32726
R-HSA-1640170_metformin_High_24h	31.97483	29.15188	-0.28826
R-HSA-1430728_acetaminophen_High_24h	31.12952	27.06645	-0.13075
R-HSA-1640170_acetaminophen_High_24h	31.00992	28.14838	-0.28212
R-HSA-1640170_omeprazole_High_24h	31.00992	28.14838	-0.28212
R-HSA-162582_nitrofurantoin_Low_24h	30.51128	26.04955	-0.09802
R-HSA-1640170_colchicine_Middle_24h	30.04628	27.14481	-0.27584
R-HSA-1640170_tetracycline_High_2h	30.04628	27.14481	-0.27584
R-HSA-69278_acetaminophen_High_24h	28.90311	26.15924	-0.30102
R-HSA-69278_metformin_High_24h	28.90311	26.15924	-0.30102

Tabela 4.5: Tabela com os 15 maiores valores de *scores*, média e variância para *Homo sapiens in vitro* com *Reactome*.

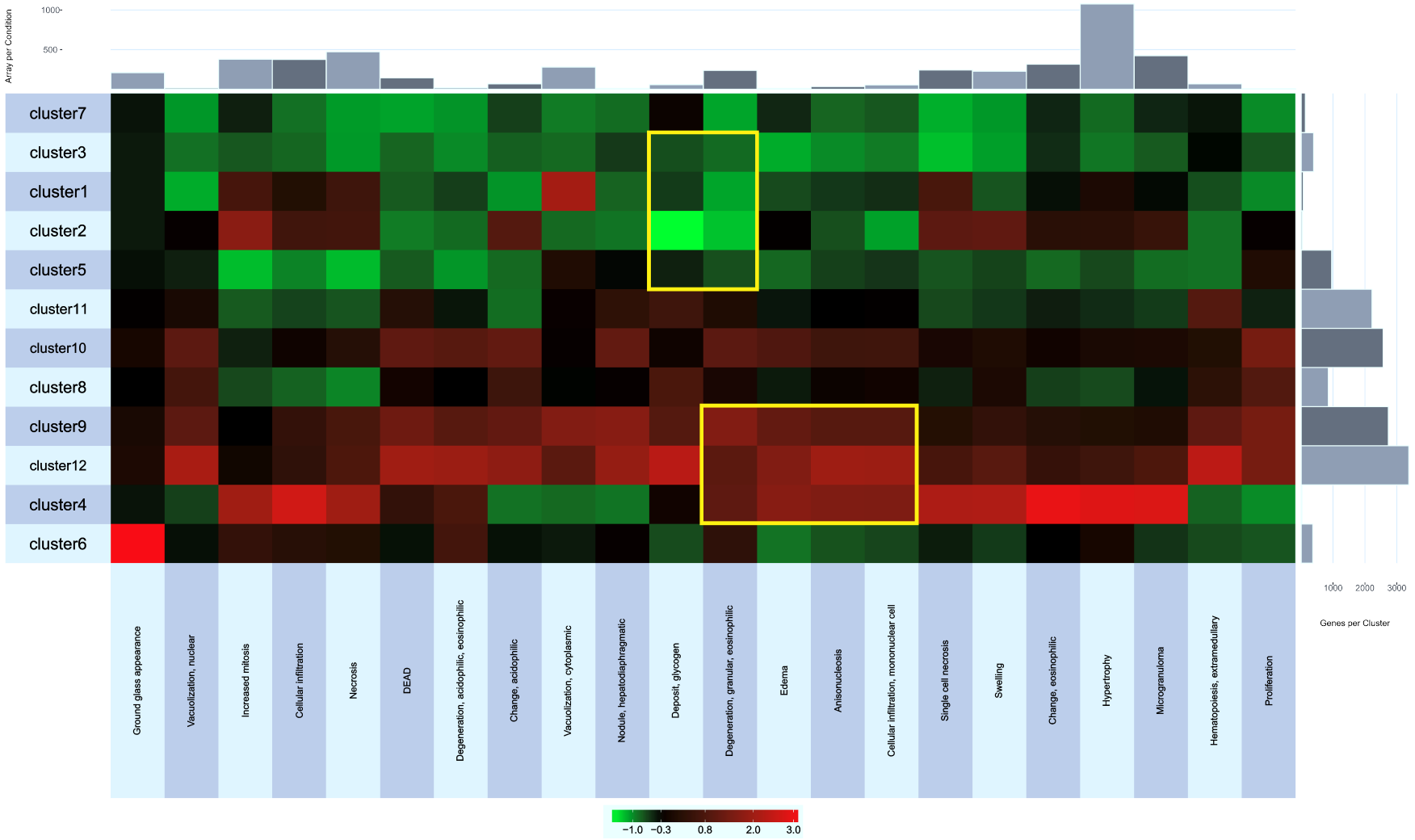

Figura 4.12: *Heatmap* gerado para o experimento *Homo sapiens in vitro* relacionado com *Reactome* As caixas amarelas estão evidenciando 2 tipos de perfil presentes, um induzido (predominância de vermelho) e outro reprimido (predominância de verde).

Para *Homo sapiens in vitro* relacionado com *GO* do tipo processo biológico foi gerado o *heatmap* da Figura 4.13. Também foi obtida a Tabela 4.6, construída a partir dos genes induzidos/reprimidos, que mostra os valores do *score*, média e variância para cada *set* de gene relacionado com *array*. A partir do método de *multiscale bootstrap resampling* foram obtidos 84 *clusters*. Vale ressaltar um dos *clusters* que possui o maior valor de significância para determinada condição: o *cluster* 27 é formado pelas ontologias, com suas respectivas descrições, representadas na Tabela 4.7. As ontologias mais representativas nesse *cluster* são as que possuem relação com dobramento, desdobramento ou reenrolamento de proteínas. De acordo com a Figura 4.13 nota-se que o cluster 27 é significativo para a condição aparição de opacidade em vidro fosco (*ground glass appearace*). Tal condição está associada ao aumento da massa do fígado, que é acompanhada por alguma alteração histopatológica característica tipificada por uma opacidade em vidro fosco (KAORI et al., 2009). Tais ontologias encontradas anteriormente são elucidadas como responsáveis pela indução do *stress* do retículo endoplasmático em diferentes tipos de hepatócitos de opacidade em vidro fosco na infecção crônica do vírus da hepatite B (WANG et al., 2003). Quando comparamos os dados obtidos para esse experimento com os dados obtidos para *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo* também utilizando *GO* do tipo processo biológico, é possível identificar 3 *GOs* que estão presentes no *cluster* 27, nos outros dois experimentos. Os *clusters* gerados para os outros experimentos foram totalmente distintos quando comparados com *Homo sapiens in vitro*. Mesmo assim, as 3 *GOs* em comum estão em *clusters* com mais *GOs*. Essas *GOs* em comum são responsáveis pela resposta celular ao calor (GO:0034605), resposta celular ao *stress* (GO:0033554) e responsável pelo calor (GO:0009408). Apesar de ter elementos em comuns para os 3 experimentos há um predomínio das discrepâncias, sendo assim mais uma vez confirmando a diferença que há entre os modelos.

SetGene_Array	***Score***	**Média (μ)**	**Variância (σ)**
GO_0000278_chlormezanone_High_24h	111.24597	109.23768	-0.42505
GO_0000278_omeprazole_High_24h	108.27387	106.23401	-0.41986
GO_0000278_metformin_High_24h	106.29292	104.23152	-0.41628
GO_0000278_tetracycline_Middle_2h	106.29292	104.23152	-0.41628
GO_0000278_pemoline_High_24h	105.30259	103.23026	-0.41447
GO_0000278_methimazole_High_24h	104.31234	102.22899	-0.41263
GO_0044763_omeprazole_High_24h	104.10445	101.13923	-0.26597
GO_0000278_ethinylestradiol_High_24h	102.33214	100.22642	-0.40888
GO_0000278_ethinylestradiol_Middle_24h	99.3626	97.2225	-0.40309
GO_0000278_tetracycline_High_2h	98.37297	96.22117	-0.40111
GO_0000278_benzbromarone_High_24h	97.38344	95.21983	-0.39911
GO_0000278_danazol_High_24h	97.38344	95.21983	-0.39911
GO_0000278_nitrofurantoin_Low_24h	97.38344	95.21983	-0.39911
GO_0000278_acetaminophen_High_24h	96.39402	94.21848	-0.39709
GO_0000278_monocrotaline_High_24h	96.39402	94.21848	-0.39709

Tabela 4.6: Tabela contendo as *top* 15 informações relativas ao valor de *score*, média e variância para *Homo sapiens in vitro* com GO do tipo processo biológico.

Ontologia	Descrição
GO:0006457	protein folding
GO:0006986	response to unfolded protein
GO:0009408	response to heat
GO:0033554	cellular response to stress
GO:0034605	cellular response to heat
GO:0035966	response to topologically incorrect protein
GO:0042026	protein refolding

Tabela 4.7: GOs do tipo processos biológicos presentes no *cluster* 27 para *Homo sapiens in vitro*.

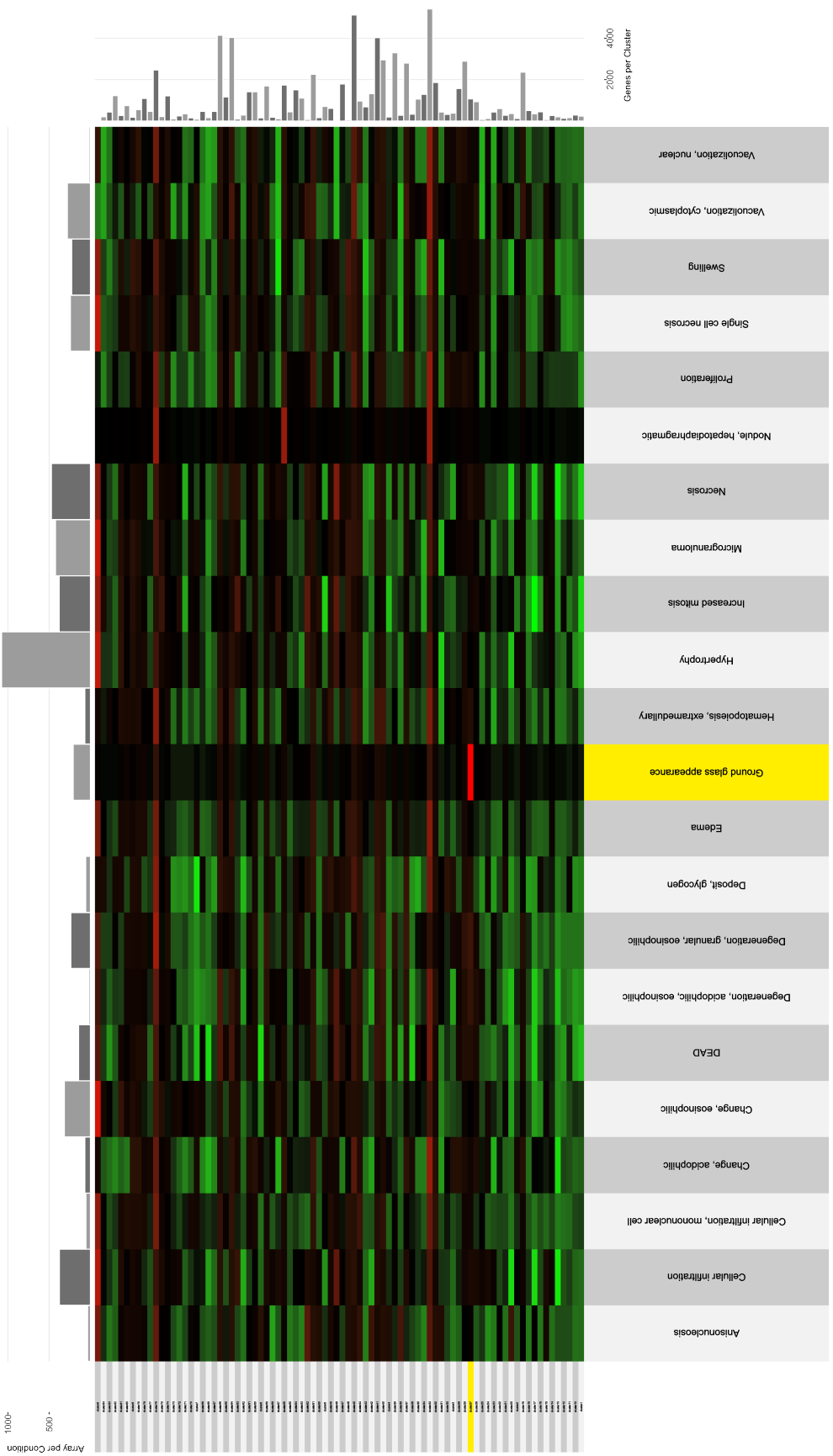

Figura 4.13: *Heatmap* gerado para o experimento *Homo sapiens in vitro* relacionado com GO do tipo processo biológico. Está destacado em amarelo o *cluster* 27 além da condição estudada que foi a aparição de opacidade em vidro fosco.

A próxima análise corresponde ao experimento *Rattus norvegicus in vitro* com concentração de dose alta e tempo de amostragem igual a 24 horas e GO do tipo processo biológico e foi gerado o *heatmap* da Figura 4.14. A Tabela 4.8, construída a partir dos genes induzidos/reprimidos, mostram valores do *score* , média e variância para cada *set* de gene relacionado com *array*. A partir do método de *multiscale bootstrap resampling* foram obtidos 155 *clusters*. O principal *cluster* obtido para esse experimento foi o *cluster* 3 que é formado pelas GOs mostradas na Tabela 4.9. As principais funções presentes nesse *cluster* estão relacionadas a célula, por exemplo o ciclo celular, proliferação, etc. Observa-se na Figura 4.14 que uma das condições mais significativas para o *cluster* 3 é a fibrose. A fibrose é a formação ou desenvolvimento de tecido conjuntivo em determinado órgão ou tecido como parte de um processo de cicatrização ou de degenerescência fibroide. A partir de um estudo que avalia as rotas e processos envolvidos na fibrose pulmonar e hepática (MAKAREV et al., 2016), podemos afirmar que a os principais processos que compõe o *cluster* 3 estão presentes nos processos e rotas evidenciados nesse estudo, tais como processos envolvidos com hipoxia, ciclo celular, fases da mitose, etc. Comparando as *GOs* presentes no *cluster* 3 com experimentos de *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo*, com mesma concentração de dose e tempo de amostragem tratados anteriormente, podemos concluir que as duas *GOs* que compõe o *cluster* 3 estão presentes na grande maioria dos *clusters*. Isso mostra, que além dessas convergências há muitas discrepâncias, mostrando as diferenças entre os modelos.

SetGene_Array	***Score***	**Média**	**Variância**
GO_0044763_colchicine_High_24h	91.49608	89.2057	-0.37749
GO_0044710_colchicine_High_24h	88.60097	86.193	-0.35735
GO_0006468_colchicine_High_24h	75.93138	73.15631	-0.29595
GO_0006915_colchicine_High_24h	74.87247	72.16246	-0.30655
GO_0006357_colchicine_High_24h	72.35624	69.11702	-0.22588
GO_0042221_colchicine_High_24h	71.46293	69.2098	-0.38387
GO_0016310_colchicine_High_24h	67.96527	65.15285	-0.28994
GO_0030154_colchicine_High_24h	65.14596	62.13536	-0.25907
GO_0000278_colchicine_High_24h	65.13913	63.2519	-0.44447
GO_0006950_colchicine_High_24h	64.51274	62.20365	-0.37428
GO_0055114_colchicine_High_24h	62.55205	60.19886	-0.36673
GO_0000278_tacrine_High_24h	62.18792	60.24537	-0.43571
GO_0006357_carboplatin_High_24h	61.5299	58.10345	-0.20084
GO_0000278_nimesulide_High_24h	59.23922	57.23857	-0.42631
GO_0000278_monocrotaline_High_24h	96.39402	94.21848	-0.39709

Tabela 4.8: Tabela contendo as *top* 15 informações relativas ao valor de *score*, média e variância para *Rattus norvegicus in vitro* com GO do tipo processo biológico e concentração de dose alta com tempo de amostragem igual a 24 horas.

Ontologia	Descrição
GO:0000082	G1/S transition of mitotic cell cycle
GO:0001666	response to hypoxia
GO:0001889	liver development
GO:0006950	response to stress
GO:0008283	cell proliferation
GO:0009404	toxin metabolic process
GO:0009636	response to toxic substance
GO:0010033	response to organic substance
GO:0010039	response to iron ion
GO:0017144	drug metabolic process
GO:0030258	lipid modification
GO:0030855	epithelial cell differentiation
GO:0032502	developmental process
GO:0033993	response to lipid
GO:0035902	response to immobilization stress
GO:0036293	response to decreased oxygen levels
GO:0036296	response to increased oxygen levels
GO:0042221	response to chemical
GO:0042493	response to drug
GO:0044237	cellular metabolic process
GO:0044710	single-organism metabolic process
GO:0044763	single-organism cellular process
GO:0046483	heterocycle metabolic process
GO:0046677	response to antibiotic
GO:0048565	digestive tract development
GO:0048609	multicellular organismal reproductive process
GO:0055093	response to hyperoxia
GO:0055114	oxidation-reduction process
GO:0070365	hepatocyte differentiation
GO:0071248	cellular response to metal ion
GO:0071310	cellular response to organic substance
GO:1900087	positive regulation of G1/S transition of mitotic cell cycle
GO:1901992	positive regulation of mitotic cell cycle phase transition
GO:1902808	positive regulation of cell cycle G1/S phase transition
GO:2000045	regulation of G1/S transition of mitotic cell cycle

Tabela 4.9: Principais GOs do tipo processo biológicos presentes no *cluster* 3 para *Rattus norvegicus in vitro* na concentração de dose alta com tempo de amostragem igual a 24 horas.

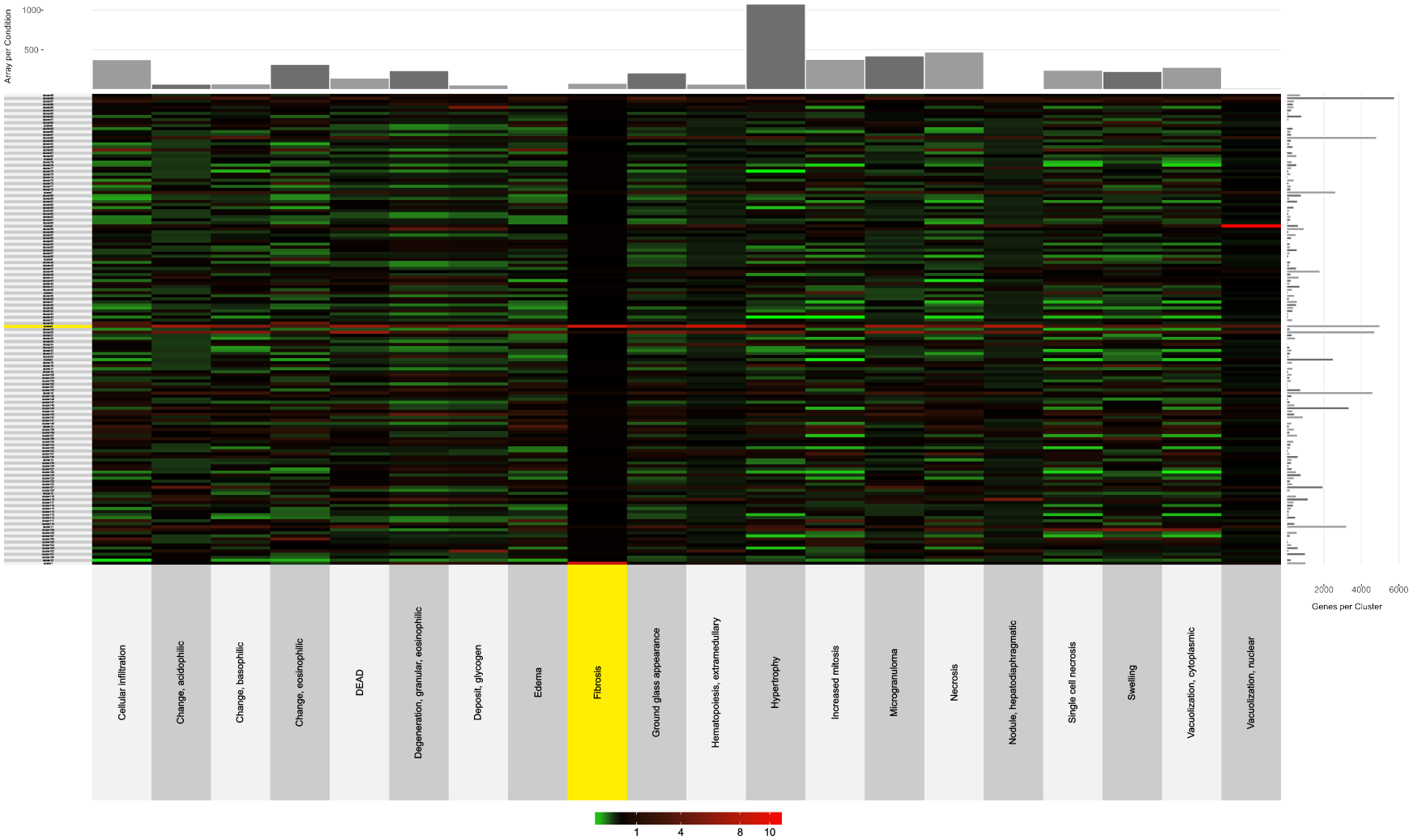

Figura 4.14: *Heatmap* gerado para o experimento *Rattus norvegicus in vitro* relacionado com GO do tipo processo biológico com concentração de dose alta e tempo de amostragem igual a 24h. Está destacado em amarelo o *cluster* 3 além da condição estudada que foi a fibrose.

Para o experimento *Rattus norvegicus in vitro* com *Reactome*, concentração de dose alta e tempo de amostragem igual a 24 horas geramos o *heatmap* representado na Figura 4.15. Além disso, obtivemos a Tabela 4.10, construída a partir dos genes induzidos/reprimidos, mostram valores do *score*, média, e variância para cada *set* de gene relacionado com *array*. A partir do método de *multiscale bootstrap resampling* foram obtidos 33 *clusters*, que tiveram sua consistência testadas e comparadas. De acordo com a Figura 4.15 temos o *cluster* 19 como sendo o mais significante. Através da Figura 4.15 é também possível identificar que a condição fibrose, explicado em um dos itens anteriores, é uma das principais condições significantes para esse experimento. Para o tratamento da fibrose, geralmente, a estratégia recomendada é atuar terapeuticamente no início da síntese do colágeno, pois impedindo essa síntese é possível evitar a proliferação do colágeno, que é o principal componente da fibrose. As rotas mais significantes encontradas no *cluster* 19 são relacionadas com a degradação de colágeno e a formação de colágeno, além da degradação e organização da matriz extracelular Tabela 4.11. Estudos como (WYNN, 2008), (LAURENT, 2009) e (BONNANS; CHOU; WERB, 2014) confirmam que as rotas compostas pelo *cluster* evidenciado anteriormente está muito relacionado com a condição que estava significante. Os *Reactomes* encontrados para esse experimento para o *cluster* 19 não possui relação alguma com os experimentos para *Homo sapiens in vitro* e *Rattus norvegicus in vivo* com concentração de dose alta e tempo de amostragem de 24h. Através dessa informação podemos dizer que há diferenças entre os modelos.

SetGene_Array	Score	Média	Variância
R-RNO-1430728_colchicine_High_24h	104.71001	102.18031	-0.33663
R-RNO-162582_colchicine_High_24h	97.22988	94.12778	-0.24545
R-RNO-1640170_colchicine_High_24h	59.24666	57.23759	-0.42493
R-RNO-1430728_ethambutol_High_24h	56.36469	53.11633	-0.22461
R-RNO-162582_carboplatin_High_24h	54.84135	51.08245	-0.16137
R-RNO-69278_colchicine_High_24h	54.22462	52.2405	-0.429
R-RNO-1430728_cisplatin_High_24h	50.48483	47.10684	-0.20713
R-RNO-1430728_ethionine_High_24h	47.55079	44.10191	-0.19798
R-RNO-1430728_hydroxyzine_High_24h	47.55079	44.10191	-0.19798
R-RNO-1430728_quinidine_High_24h	46.57378	43.10024	-0.19486
R-RNO-162582_papaverine_High_24h	46.0355	42.07133	-0.14015
R-RNO-1430728_cephalothin_High_24h	44.62141	41.09685	-0.18853
R-RNO-1430728_naphthyl_isothiocyanate_High_24h	44.62141	41.09685	-0.18853
R-RNO-194315_colchicine_High_24h	44.2736	42.23405	-0.41991
R-RNO-162582_isoniazid_High_24h	44.08429	40.06876	-0.1352

Tabela 4.10: Tabela contendo as *top* 15 informações relativas ao valor de *score*, média e variância para *Rattus norvegicus in vitro* com *Reactome* e concentração de dose alta com tempo de amostragem igual a 24 horas.

Reactome	Descrição
R.RNO.1442490	Collagen degradation
R.RNO.1474228	Degradation of the extracellular matrix
R.RNO.1474244	Extracellular matrix organization
R.RNO.1474290	Collagen formation
R.RNO.2022090	Assembly of collagen fibrils and other multimeric structures

Tabela 4.11: *Reactomes* presentes no *cluster* 19 para *Homo sapiens in vitro*.

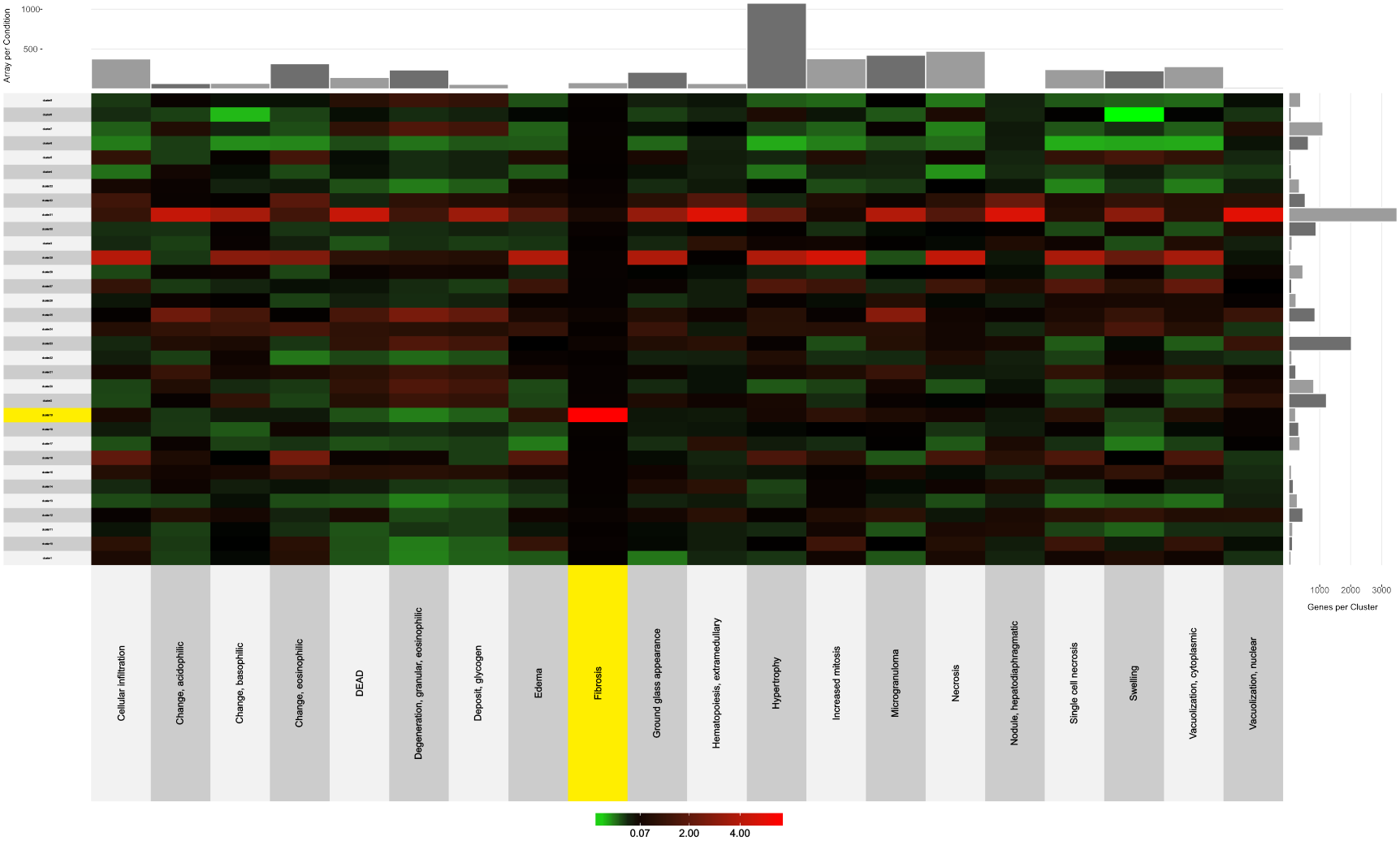

Figura 4.15: *Heatmap* gerado para o experimento *Homo sapiens in vitro* relacionado com *Reactome* com concentração de dose alta e tempo de amostragem igual a 24 horas. Está destacado em amarelo o *cluster* 19 além da condição estudada que foi a fibrose.

Os resultados encontrados para *KEGG* não geraram muitos *clusters* para cada experimento. A principal diferença dos resultados de *KEGG* para *GO* e *Reactome* é que condições que não tinham sido observadas anteriormente passaram a ser observadas a partir do *KEGG*. O experimento mais evidente é para *Rattus norvegicus in vivo* e foi gerado o *heatmap* da Figura 4.16. Além disso obtemos a Tabela 4.12, construída a partir dos genes induzidos/reprimidos, que mostra os valores do *score*, média, e variância para cada *set* de gene relacionado com *array*. A partir do método de *multiscale bootstrap resampling* foram obtidos apenas 2 *clusters*, que tiveram sua consistência testadas e comparadas. O *cluster* 2 é composto pelas *KEGGs* da Tabela 4.13. Vamos analisar o *cluster* 2 e como ele está relacionado com a condição de degeneração gordurosa (esteatose) de acordo com a Figura 4.16. A degeneração gordurosa (esteatose) é a degeneração gordurosa de um tecido, que tem como variação mais conhecida a esteatose hepática que é o acúmulo de gordura nas células do fígado. A partir da Tabela 4.12 é possível observar as vias presentes no *cluster* em questão. A via mais enriquecida para esse *cluster* é a via de sinalização AMPK. De acordo com (WANG; YANG, 2015), o medicamento liraglutide reduz a degeneração gordurosa em células hepáticas a partir da via de sinalização AMPK/SREBP1, ou seja, a via que encontramos é muito significante.

SetGene_Array	***Score***	**Média**	**Variância**
rno01100_indomethacin_Middle_24h	53.28541	50.12296	-0.23671
rno01100_acetamidofluorene_High_24h	29.93934	26.07666	-0.15034
rno01100_bendazac_High_24h	28.97856	25.07445	-0.14611
rno01100_indomethacin_High_24h	28.01938	24.0722	-0.14181
rno01100_isoniazid_High_6h	27.06194	23.06993	-0.13745
rno01100_acetamidofluorene_Low_24h	24.2017	20.0629	-0.12392
rno01100_naphthyl_isothiocyanate_High_24h	24.2017	20.0629	-0.12392
rno01100_phenacetin_High_3h	24.2017	20.0629	-0.12392
rno01100_acetamidofluorene_Middle_24h	23.253	19.06048	-0.11925
rno01100_phenobarbital_High_24h	19.48939	15.0504	-0.09968
rno01100_ticlopidine_High_24h	19.48939	15.0504	-0.09968
rno01100_phenytoin_High_24h	18.55838	14.04777	-0.09453
rno01100_WY-14643_High_6h	18.55838	14.04777	-0.09453
rno01100_bromobenzene_High_24h	17.63249	13.04508	-0.08928
rno01100_naphthyl_isothiocyanate_Middle_24h	17.63249	13.04508	-0.08928

Tabela 4.12: Tabela contendo as *top* 15 informações relativas ao valor de *score*, média e variância para *Rattus norvegicus in vivo* com KEGG.

KEGG	Descrição
rno00051	Fructose and mannose metabolism
rno00053	Ascorbate and aldarate metabolism
rno00071	Fatty acid degradation
rno00120	Primary bile acid biosynthesis
rno00140	Steroid hormone biosynthesis
rno00230	Purine metabolism
rno00240	Pyrimidine metabolism
rno00250	Alanine, aspartate and glutamate metabolism
rno00350	Tyrosine metabolism
rno00380	Tryptophan metabolism
rno00450	Selenocompound metabolism
rno00590	Arachidonic acid metabolism
rno00591	Linoleic acid metabolism
rno00980	Metabolism of xenobiotics by cytochrome P450
rno00982	Drug metabolism - cytochrome P450
rno00983	Drug metabolism - other enzymes
rno01040	Biosynthesis of unsaturated fatty acids
rno01100	Metabolic pathways
rno03320	PPAR signaling pathway
rno04010	MAPK signaling pathway
rno04014	Ras signaling pathway
rno04015	Rap1 signaling pathway
rno04060	Cytokine-cytokine receptor interaction
rno04062	Chemokine signaling pathway
rno04068	FoxO signaling pathway
rno04110	Cell cycle
rno04115	p53 signaling pathway
rno04151	PI3K-Akt signaling pathway
rno04152	AMPK signaling pathway
rno04210	Apoptosis
rno04550	Signaling pathways regulating pluripotency of stem cells
rno04621	NOD-like receptor signaling pathway
rno04668	TNF signaling pathway
rno04910	Insulin signaling pathway
rno04917	Prolactin signaling pathway

Tabela 4.13: *KEGGs* presentes no *cluster* 2 para *Rattus norvegicus in vivo*.

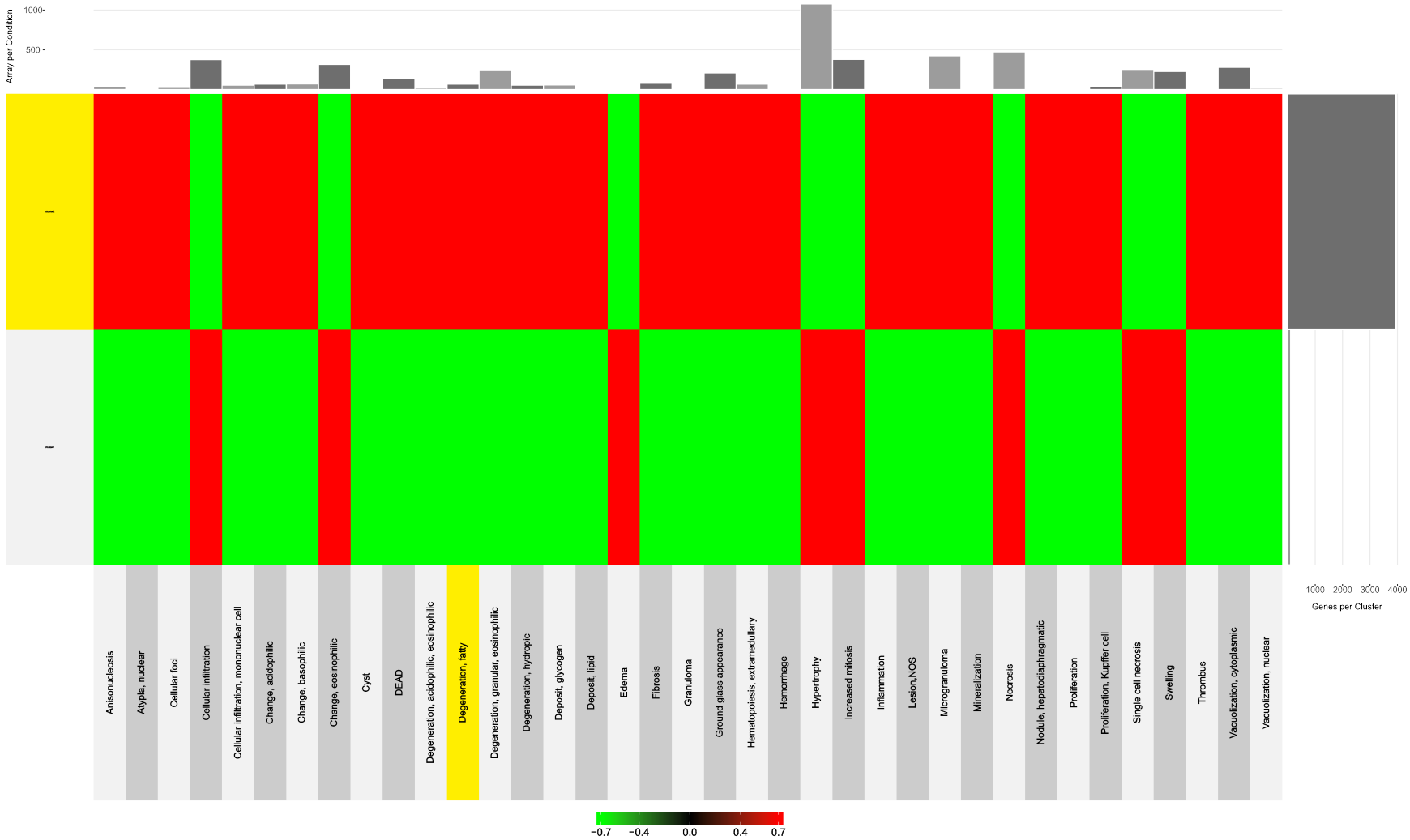

Figura 4.16: *Heatmap* gerado para o experimento *Homo sapiens in vitro* relacionado com *Reactome* com concentração de dose alta e tempo de amostragem igual a 24 horas. Está destacado em amarelo o *cluster* 2 além da condição estudada que foi a degeneração gordurosa.

5 *Conclusões*

A partir da obtenção dos genes diferencialmente expressos é possível afirmar que há diferenças significativas para os modelos *Homo sapiens in vitro*, *Rattus norvegicus in vitro* e *Rattus norvegivus in vivo* quando analisados em pequenos conjuntos de drogas e também quando analisados todas as 131 drogas, para concentrações de doses e tempos de amostragem variados.

Os dados obtidos para as 131 drogas foram enriquecidos para todas as concentrações de dose e tempos de amostragem. A partir do enriquecimento foram encontradas vias que estão alteradas para determinadas drogas. Esse fato é de extrema importância, pois a partir da identificação de quais vias que estão alteradas é possível saber como uma droga específica está agindo no organismo do modelo estudado. Com isso, é pssível analisar as ontologias, vias ou rotas metabólicas significativas para *Homo sapiens in vitro*, por exemplo, e verificar como elas estão relacionadas (se estão alteradas ou não) com os modelos *Rattus norvegicus in vitro* e *Rattus norvegicus in vivo*. Essas comparações variam de droga para droga. Enquanto algumas drogas possuem vias enriquecidas em comum para os 3 modelos, há drogas que não possuem vias em comum, ou seja, possuem apenas vias exclusivas.

E por fim, através da utilização do pacote em R para a geração do mapa modular foi possível identificar perfis correspondentes a indução e repressão. Com isso detecta-se quais são as ontologias, vias ou rotas metabólicas que estão alteradas para determinadas condições analisadas. Tal fato é muito importante, pois além de saber quais ontologias, vias e rotas metabólicas induzidas eu reprimidas, é possível identificar os genes e drogas que estão influenciando diretamente a ocorrência de determinada condição. Quando são comparados os *clusters* e perfis induzidos e reprimidos de um determinado modelo com os outros, nota-se que os *clusters* e perfis formados são totalmente diferentes. Dessa maneira, fica evidente as discrepâncias entre os modelos e também o fato de que há mecanismos específicos que regulam os diferentes experimentos, assim invibializando, por ora, a substituição dos estudos *in vivo* pelos *in vitro*.

Há alguns fatores que dificultam a comparação e substituição de modelos experimentais. Tais fatores implicam nas diferentes comparações realizadas e, principalmente, na ausência de

dados *in vivo* para *Homo sapiens*. Se os dados fossem comparados entre *Homo sapiens in vitro* com *Rattus norvegicus in vitro* e *Homo sapiens in vivo* com *Rattus norvegivus in vivo*, haveria mais comparações e provavelmente mais conclusões a respeito da substituição de modelos.

Referências Bibliográficas

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.

BODE, A. M.; DONG, Z. The enigmatic effects of caffeine in cell cycle and cancer. *Cancer letters*, Elsevier, v. 247, n. 1, p. 26–39, 2007.

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.

BONNANS, C.; CHOU, J.; WERB, Z. Remodelling the extracellular matrix in development and disease. *Nature reviews Molecular cell biology*, Nature Research, v. 15, n. 12, p. 786–801, 2014.

BOWTELL, D. D. Options available – from start to finish – for obtaining expression data by microarray. *Nature genetics*, Nature Publishing Group, v. 21, p. 25–32, 1999.

BRAZMA, A. et al. Minimum information about a microarray experiment (miame) – toward standards for microarray data. *Nature genetics*, Nature Publishing Group, v. 29, n. 4, p. 365–371, 2001.

CARLSON, M. rat2302. db: Affymetrix rat genome 230 2.0 array annotation data (chip rat2302), r package version 2.8. 1. *Santa Clara (California): Affymetrix*, 2002.

CARLSON, M. Go. db: A set of annotation maps describing the entire. gene ontology. 2013. *R package version*, v. 3, n. 2, 2013.

CARLSON, M. et al. hgu133plus2. db: Affymetrix human genome u133 plus 2.0 array annotation data (chip hgu133plus2). *URL http://www. bioconductor. org/packages/2.12/data/annotation/html/hgu133plus2. db. html. R package version*, v. 2, n. 0, 2012.

CHEN, M. et al. Fda-approved drug labeling for the study of drug-induced liver injury. *Drug discovery today*, Elsevier, v. 16, n. 15, p. 697–703, 2011.

CHEN, M. et al. A decade of toxicogenomic research and its contribution to toxicological science. *Toxicological Sciences*, Soc Toxicology, p. kfs223, 2012.

CONSORTIUM, G. O. et al. Gene ontology consortium: going forward. *Nucleic acids research*, Oxford Univ Press, v. 43, n. D1, p. D1049–D1056, 2015.

CROFT, D. et al. The reactome pathway knowledgebase. *Nucleic acids research*, Oxford Univ Press, v. 42, n. D1, p. D472–D477, 2014.

DENAYER, T.; STÖHR, T.; ROY, M. V. Animal models in translational medicine: Validation and prediction. *New Horizons in Translational Medicine*, Elsevier, v. 2, n. 1, p. 5–11, 2014.

DRAY, S.; DUFOUR, A.-B. et al. The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, v. 22, n. 4, p. 1–20, 2007.

DURINCK, S. et al. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, Oxford Univ Press, v. 21, n. 16, p. 3439–3440, 2005.

DURINCK, S. et al. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, Nature Publishing Group, v. 4, n. 8, p. 1184–1191, 2009.

ESPINDOLA, F. S. et al. Recursos de bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica= bioinformatic resources applied on the omic sciences as genomic, transcriptomic, proteomic, interatomic and metabolomic. *Bioscience Journal*, v. 26, n. 3, 2010.

FABREGAT, A. et al. The reactome pathway knowledgebase. *Nucleic acids research*, Oxford Univ Press, v. 44, n. D1, p. D481–D487, 2016.

GANTER, B. et al. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of biotechnology*, Elsevier, v. 119, n. 3, p. 219–244, 2005.

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.

GHARAIBEH, R. Z.; FODOR, A. A.; GIBAS, C. J. Background correction using dinucleotide affinities improves the performance of gcrma. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 452, 2008.

GYORFFY, B. et al. Evaluation of microarray preprocessing algorithms based on concordance with rt-pcr in clinical samples. *PloS one*, Public Library of Science, v. 4, n. 5, p. e5645, 2009.

HUBBELL, E.; LIU, W.-M.; MEI, R. Robust estimators for expression analysis. *Bioinformatics*, Oxford Univ Press, v. 18, n. 12, p. 1585–1592, 2002.

IRIZARRY, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, Biometrika Trust, v. 4, n. 2, p. 249–264, 2003.

JAIN, A. N. et al. Fully automatic quantification of microarray image data. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 2, p. 325–332, 2002.

JOHNSON, K.; LIN, S. Call to work together on microarray data analysis. *Nature*, Nature Publishing Group, v. 411, n. 6840, p. 885–885, 2001.

JÚNIOR, C.; SASSON, S. *Biologia, vol. seriado, 8ª edição*. [S.l.: s.n.], 2005.

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

KANEHISA, M. et al. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, Oxford Univ Press, p. gkv1070, 2015.

KANNO, J. Reverse toxicology as a future predictive toxicology. In: *Toxicogenomics*. [S.l.]: Springer, 2003. p. 213–218.

KAORI, A.-T. et al. Use of toxicogenomics for discrimination between the types of liver weight increase. In: JAPAN TOXICOLOGY SOCIETY. *Academic Year of Japan Toxicology Society 36 th Annual Meeting of Japanese Toxicology Society*. [S.l.], 2009. p. 4121–4121.

KNUDSEN, S. *Guide to analysis of DNA microarray data*. [S.l.]: John Wiley & Sons, 2005.

KOLTAI, H.; WEINGARTEN-BAROR, C. Specificity of dna microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic acids research*, Oxford Univ Press, v. 36, n. 7, p. 2395–2405, 2008.

LAURENT, G. J. Biochemical pathways leading to collagen deposition in pulmonary fibrosis. *Fibrosis*, John Wiley & Sons, v. 832, p. 222, 2009.

LIM, W. K. et al. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, Oxford Univ Press, v. 23, n. 13, p. i282–i288, 2007.

LODISH, H. et al. *Molecular cell biology*. [S.l.]: Scientific American Books New York, 1995.

MAKAREV, E. et al. Common pathway signature in lung and liver fibrosis. *Cell Cycle*, Taylor & Francis, v. 15, n. 13, p. 1667–1673, 2016.

MILLER, C. simpleaffy: Very simple high level analysis of affymetrix data. *R package version 2.28*, 2007.

OLSON, N. E. The microarray data analysis process: from raw data to biological significance. *NeuroRx*, Elsevier, v. 3, n. 3, p. 373–383, 2006.

PARADIS, E.; CLAUDE, J.; STRIMMER, K. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, Oxford Univ Press, v. 20, n. 2, p. 289–290, 2004.

PEPPER, S. D. et al. The utility of mas5 expression summary and detection call algorithms. *BMC bioinformatics*, BioMed Central, v. 8, n. 1, p. 273, 2007.

POLLI, J. E. In vitro studies are sometimes better than conventional human pharmacokinetic in vivo studies in assessing bioequivalence of immediate-release solid oral dosage forms. *The AAPS journal*, Springer, v. 10, n. 2, p. 289–299, 2008.

RITCHIE, M. E. et al. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, Oxford Univ Press, p. gkv007, 2015.

RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2015. Disponível em: <http://www.rstudio.com/>.

SCHENA, M. et al. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, The American Association for the Advancement of Science, v. 270, n. 5235, p. 467, 1995.

SEGAL, E. et al. A module map showing conditional activity of expression modules in cancer. *Nature genetics*, Nature Publishing Group, v. 36, n. 10, p. 1090–1098, 2004.

SHARMA, B.; SINGH, S.; KANWAR, S. S. L-methionase: a therapeutic enzyme to treat malignancies. *BioMed research international*, Hindawi Publishing Corporation, v. 2014, 2014.

SILVA, C. P. et al. Importância da toxicidade pulmonar pela amiodarona no diagnóstico diferencial de paciente com dispnéia em fila para transplante cardíaco. *Arq Bras Cardiol*, v. 87, n. 3, p. 4–7, 2006.

STANDL, E. et al. On the potential of acarbose to reduce cardiovascular disease. *Cardiovascular diabetology*, BioMed Central, v. 13, n. 1, p. 81, 2014.

SUZUKI, R.; SHIMODAIRA, H. Hierarchical clustering with p-values via multiscale bootstrap resampling. *R package*, 2013.

TENENBAUM, D. Keggrest: Client-side rest access to kegg. *R package version*, v. 1, n. 1, 2013.

TILSTONE, C. Dna microarrays: vital statistics. *Nature*, Nature Publishing Group, v. 424, n. 6949, p. 610–612, 2003.

TOXICOLOGY, N. R. C. U. C. on Applications of Toxicogenomic Technologies to P. et al. *Applications of toxicogenomic technologies to predictive toxicology and risk assessment*. [S.l.]: National Academies Press (US), 2007.

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.

WANG, H.-C. et al. Different types of ground glass hepatocytes in chronic hepatitis b virus infection contain specific pre-s mutants that may induce endoplasmic reticulum stress. *The American journal of pathology*, Elsevier, v. 163, n. 6, p. 2441–2449, 2003.

WANG, Y.-G.; YANG, T.-L. Liraglutide reduces fatty degeneration in hepatic cells via the ampk/srebp1 pathway. *Experimental and therapeutic medicine*, Spandidos Publications, v. 10, n. 5, p. 1777–1783, 2015.

WARNES, G. R. et al. gplots: Various r programming tools for plotting data. *R package version*, v. 2, n. 4, 2009.

WATERS, M. D.; FOSTEL, J. M. Toxicogenomics and systems toxicology: aims and prospects. *Nature Reviews Genetics*, Nature Publishing Group, v. 5, n. 12, p. 936–948, 2004.

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.

WYNN, T. Cellular and molecular mechanisms of fibrosis. *The Journal of pathology*, Wiley Online Library, v. 214, n. 2, p. 199–210, 2008.

YU, G.; HE, Q.-Y. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, Royal Society of Chemistry, v. 12, n. 2, p. 477–479, 2016.

YU, G. et al. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 16, n. 5, p. 284–287, 2012.

YU, G. et al. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, Oxford Univ Press, v. 31, n. 4, p. 608–609, 2015.

APÊNDICE A – Estimativa da média por Tukey's Biweigth

No método *Tukey Biweigth*, o objetivo é calcular a média de um conjunto de dados impedindo que pontos que fujam da distribuição alterem muito o resultado. Inicialmente, calcula-se a mediana da distribuição, M, e então a mediana das distâncias absolutas até M, S. Para cada valor x_i da distribuição, calcula-se

$$u_i = \frac{x_i - M}{cS + \varepsilon} \tag{A.1}$$

onde c é uma constante de ajuste e ε impede divisão por zero. Por padrão do método MAS5 descrito neste trabalho, usa-se $c = 5$ and $\varepsilon = 0.0001$.

Finalmente, a estimativa da média da distribuição é dada por

$$w_i = \begin{cases} (1-u^2)^2, & |u| < 0 \\ 0, & |u| > 0 \end{cases} \tag{A.2}$$

$$TB_{i(x_i)} = \frac{\sum_i w_i x_i}{\sum_i w_i} \tag{A.3}$$

APÊNDICE B – Tabela de drogas e as quantidades de genes diferencialmente expressos para os experimentos

Droga	***Homo sapiens* *in vitro***	***Rattus norvegicus* *in vitro***	***Rattus norvegicus* *in vivo***
acarbose	14	36	0
acetamidofluorene	11	37	1865
acetaminophen	2364	2470	361
acetazolamide	0	3	245
adapin	384	186	256
ajmaline	173	410	159
allopurinol	77	0	17
allyl_alcohol	1768	20	43
amiodarone	2	0	104
amitriptyline	16	314	59
aspirin	47	128	530
azathioprine	570	257	190
bendazac	0	24	429
benzbromarone	1922	46	472
benziodarone	84	250	313
bromobenzene	0	8	1096
bromoethylamine	0	425	452
bucetin	2	12	446
caffeine	1581	1957	262
captopril	406	642	3
carbamazepine	121	9	269
carbon_tetrachloride	7	401	91
carboplatin	0	4326	59
cephalothin	0	2562	127
chloramphenicol	40	32	116
chlormadinone	705	501	438
chlormezanone	2	4	577
chlorpheniramine	16	658	59
chlorpromazine	52	13	609
chlorpropamide	3	9	65
cimetidine	1	6	8
ciprofloxacin	0	1	20
cisplatin	0	4635	123

Tabela B.1: Tabela das 131 drogas e como elas se relacionam com as quantidades de genes expressos para cada um dos 3 experimentos disponíveis. Esses genes diferencialmente expressos foram obtidos para todas as concentrações de doses e tempos de amostragem.

Droga	***Homo sapiens* *in vitro***	***Rattus norvegicus* *in vitro***	***Rattus norvegicus* *in vivo***
clofibrate	5	89	335
clomipramine	2	444	4
colchicine	2891	4080	1966
coumarin	34	61	717
cyclophosphamide	47	84	111
cyclosporine_A	0	288	295
danazol	941	106	671
dantrolene	564	4	296
diazepam	1079	393	114
diclofenac	1192	1830	1065
diltiazem	190	953	182
disopyramide	1310	1479	130
disulfiram	160	278	494
doxorubicin	0	912	307
enalapril	227	1525	1
erythromycin_ethylsuccinate	0	7	15
ethambutol	322	775	240
ethanol	7	68	61
ethinylestradiol	1	16	282
ethionamide	96	135	2356
ethionine	3359	2811	747
etoposide	345	647	352
famotidine	2	13	18
fenofibrate	5	56	631
fluphenazine	51	72	224
flutamide	726	0	206
furosemide	1241	859	158
gemfibrozil	9	16	327
gentamicin	0	2967	27
glibenclamide	9	5	12
griseofulvin	56	11	58
haloperidol	25	70	146
hexachlorobenzene	9	1	9
hydroxyzine	403	1272	21
ibuprofen	135	2085	210
imipramine	0	921	105
indomethacin	66	230	1460
iproniazid	167	302	27
isoniazid	705	1929	0
ketoconazole	745	341	0
labetalol	521	256	0
lomustine	201	49	0
lornoxicam	0	4	0
mefenamic_acid	53	69	0
meloxicam	3	8	0
metformin	1701	349	0
methapyrilene	1765	1	667
methimazole	435	374	348

Tabela B.2: Tabela das 131 drogas e como elas se relacionam com as quantidades de genes expressos para cada um dos 3 experimentos disponíveis. Esses genes diferencialmente expressos foram obtidos para todas as concentrações de doses e tempos de amostragem (continuação).

Droga	*Homo sapiens in vitro*	*Rattus norvegicus in vitro*	*Rattus norvegicus in vivo*
methyldopa	0	3	110
methyltestosterone	36	7	218
mexiletine	115	1	339
monocrotaline	0	100	319
moxisylyte	133	437	126
naphthyl_isothiocyanate	853	1828	932
naproxen	176	1299	593
nicotinic_acid	10	11	51
nifedipine	201	498	760
nimesulide	1302	552	1204
nitrofurantoin	1235	567	147
nitrofurazone	71	1117	803
nitrosodiethylamine	40	395	2087
omeprazole	2917	26	371
papaverine	1625	2091	71
pemoline	0	0	69
penicillamine	218	531	29
perhexiline	226	19	22
phenacetin	7	16	1601
phenobarbital	3057	2007	503
phenylanthranilic_acid	25	73	795
phenylbutazone	222	143	142
phenytoin	12	4	239
promethazine	376	380	209
propylthiouracil	1858	803	104
puromycin_aminonucleoside	0	4408	773
quinidine	155	1908	40
ranitidine	1327	1278	1
rifampicin	66	66	58
simvastatin	272	579	252
sulfasalazine	7	56	107
sulindac	2694	3216	1144
sulpiride	899	1667	101
tacrine	188	906	105
tamoxifen	1	16	37
tannic_acid	563	2868	32
terbinafine	2	1	131
tetracycline	20	3	36
theophylline	2430	2019	409
thioacetamide	44	23	1138
thioridazine	93	6	162
ticlopidine	4	19	716
tiopronin	2	75	99
tolbutamide	212	170	49
triamterene	0	18	20
triazolam	0	0	66
trimethadione	0	59	288
valproic_acid	1675	2335	132
vitamin_A	45	31	105
WY-14643	68	119	590

Tabela B.3: Tabela das 131 drogas e como elas se relacionam com as quantidades de genes expressos para cada um dos 3 experimentos disponíveis. Esses genes diferencialmente expressos foram obtidos para todas as concentrações de doses e tempos de amostragem (continuação).

APÊNDICE C – Tabela de drogas e as quantidades de genes diferencialmente expressos para os experimentos

C.1 *Homo sapiens in vitro*

Droga	Cód.	Tempo de Amostragem (h)	Baixa	Média	Alta	Baixa ∩ Média	Baixa ∩ Alta	Média ∩ Alta	Baixa ∩ Média ∩ Alta
Ácido Tânico	A1	8	-	0	110	-	-	6	-
	A2	24	-	0	469	-	-	39	-
Cafeína	B1	8	-	1	518	-	-	17	-
	B2	24	-	20	1137	-	-	144	-
Colchicina	C1	8	-	42	578	-	-	248	-
	C2	24	-	75	1780	-	-	738	-
Etanol	D1	8	-	0	1	-	-	0	-
	D2	24	-	0	7	-	-	0	-
Etionina	E1	2	0	0	21	0	0	0	0
	E2	8	0	41	772	1	0	134	11
	E3	24	6	80	2059	11	1	719	101
Omeprazol	F1	2	0	1	36	0	0	2	3
	F2	8	0	11	699	1	0	8	3
	F3	24	0	44	2521	10	1	174	22

Tabela C.1: Quantidade de genes diferencialmente expressos encontrados para *Homo sapiens in vitro* com a normalização RMA.

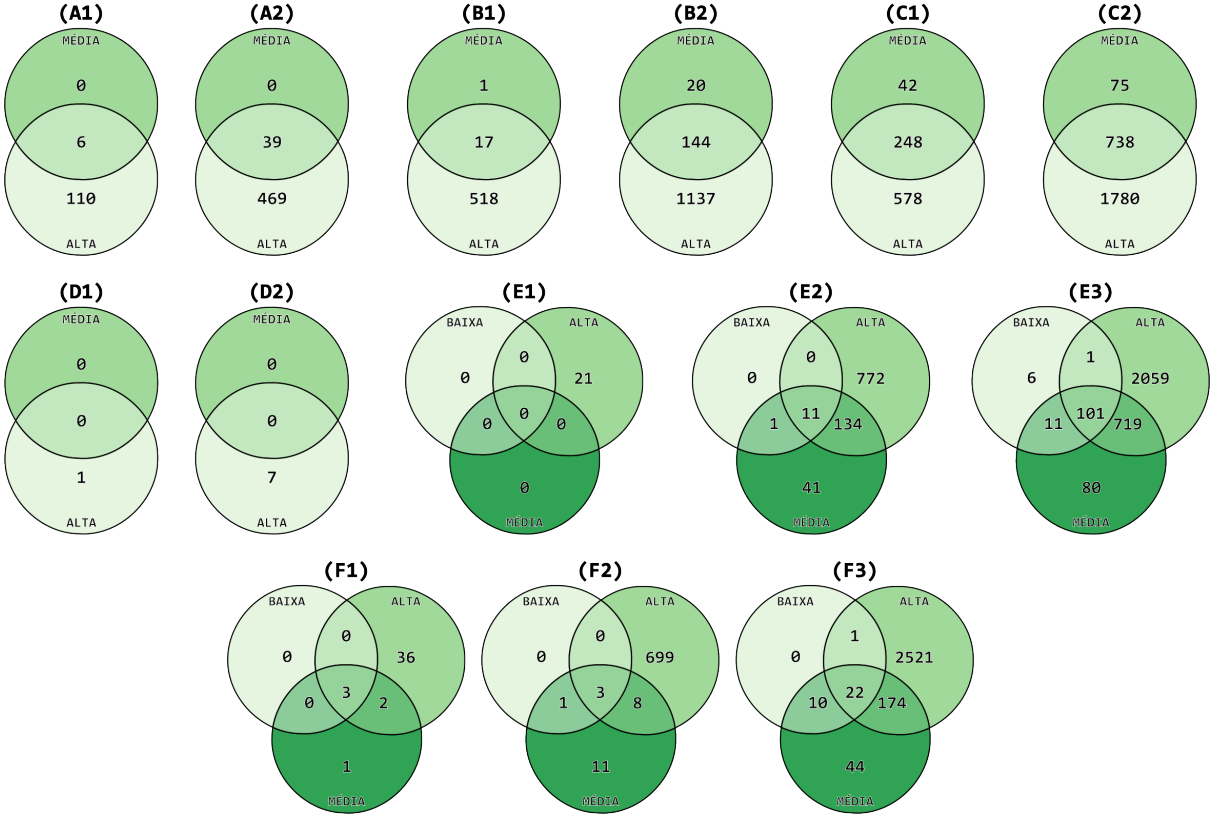

Figura C.1: Diagramas de *Venn* com a respectiva correspondência da Tabela C.1.

C.2 *Rattus norvegicus in vitro*

Droga	Cód.	Tempo de Amostragem (h)	Baixa	Média	Alta	Baixa ∩ Média	Baixa ∩ Alta	Média ∩ Alta	Baixa ∩ Média ∩ Alta
Ácido Tânico	A1	8	0	0	703	0	0	2	0
	A2	24	0	2	2778	0	1	13	0
Cafeína	B1	2	0	4	51	0	0	12	0
	B2	8	0	27	900	0	0	128	1
	B3	24	0	15	1012	0	0	179	5
Colchicina	C1	2	6	24	128	9	2	34	29
	C2	8	17	83	981	32	2	390	206
	C3	24	50	64	2163	85	31	505	603
Etanol	D1	2	0	0	1	0	0	1	0
	D2	8	0	0	7	0	0	0	0
	D3	24	0	0	61	0	0	1	0
Etionina	E1	2	0	0	35	0	0	7	1
	E2	8	4	91	390	11	0	280	39
	E3	24	18	138	1035	34	5	1004	365
Omeprazol	F1	2	0	0	0	0	0	0	0
	F2	8	0	0	4	0	0	2	0
	F3	24	0	0	21	0	0	1	0

Tabela C.2: Quantidade de genes diferencialmente expressos encontrados para *Rattus norvegicus in vitro* com a normalização RMA.

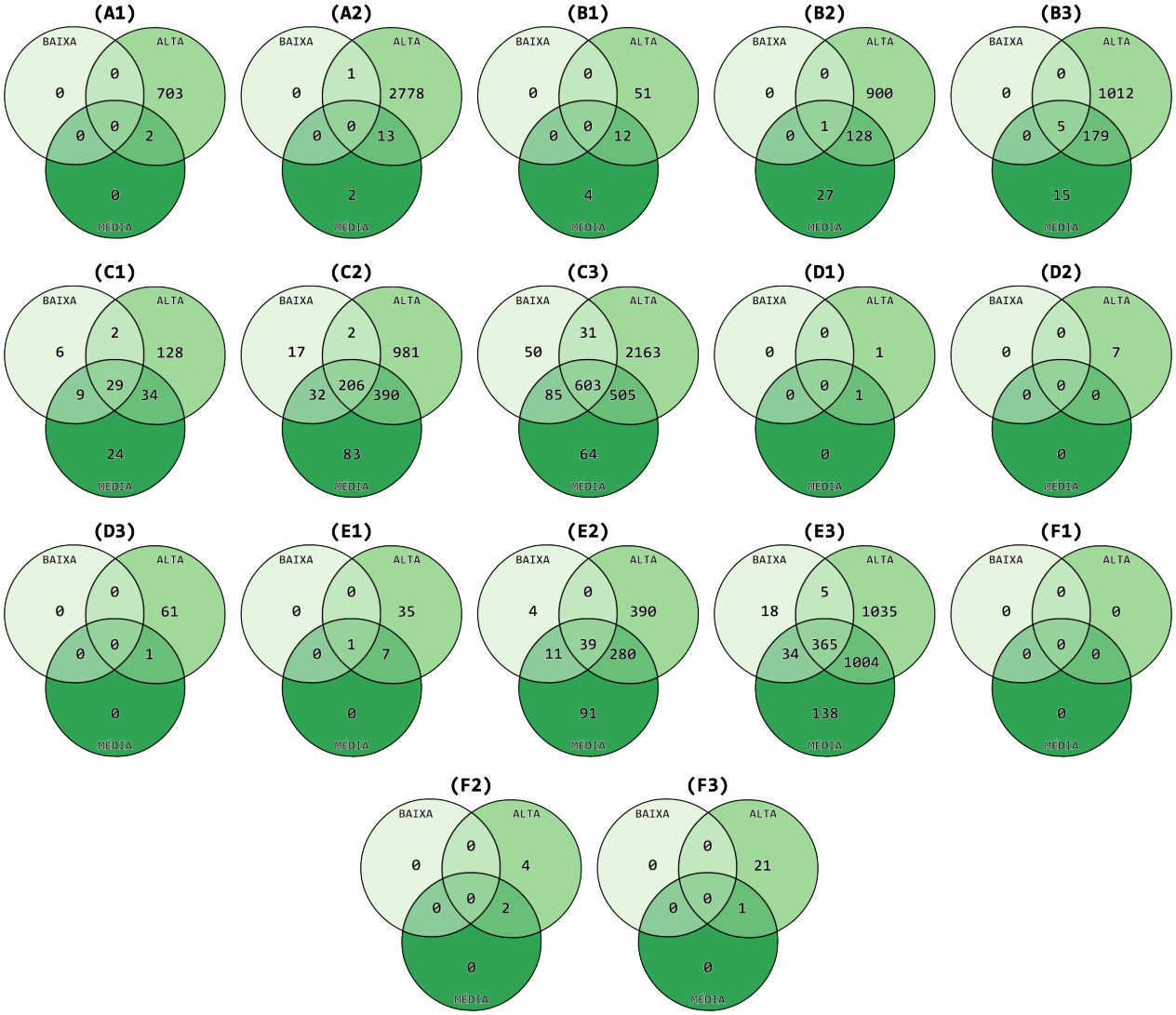

Figura C.2: Diagramas de *Venn* com a respectiva correspondência da Tabela C.2.

C.3 *Rattus norvegicus in vivo*

Droga	Cód.	Tempo de Amostragem (h)	Baixa	Média	Alta	Baixa ∩ Média	Baixa ∩ Alta	Média ∩ Alta	Baixa ∩ Média ∩ Alta
Ácido Tânico	A1	6	0	0	7	0	0	0	0
	A2	24	0	0	25	0	0	0	0
Cafeína	B1	3	0	4	94	0	0	7	0
	B2	6	0	7	87	0	0	1	0
	B3	9	0	0	62	0	0	0	0
	B4	24	0	4	46	0	0	2	0
Colchicina	C1	3	0	0	37	0	0	0	1
	C2	6	0	6	719	0	0	138	0
	C3	9	2	11	1205	0	0	124	5
	C4	24	0	44	592	0	0	62	0
Etanol	D1	3	3	2	12	0	0	4	0
	D2	6	0	0	37	0	0	0	0
	D3	9	0	0	11	0	0	0	0
	D4	24	0	0	1	0	0	1	0
Etionina	E1	3	1	3	88	0	0	11	4
	E2	6	13	18	94	2	10	58	35
	E3	9	2	23	183	1	3	107	27
	E4	24	0	19	263	0	0	100	0
Omeprazol	F1	3	1	13	5	1	0	6	8
	F2	6	3	8	92	0	0	68	29
	F3	9	3	18	80	4	1	60	14
	F4	24	0	0	122	0	0	12	1

Tabela C.3: Quantidade de genes diferencialmente expressos encontrados para *Rattus norvegicus in vivo* com a normalização RMA.

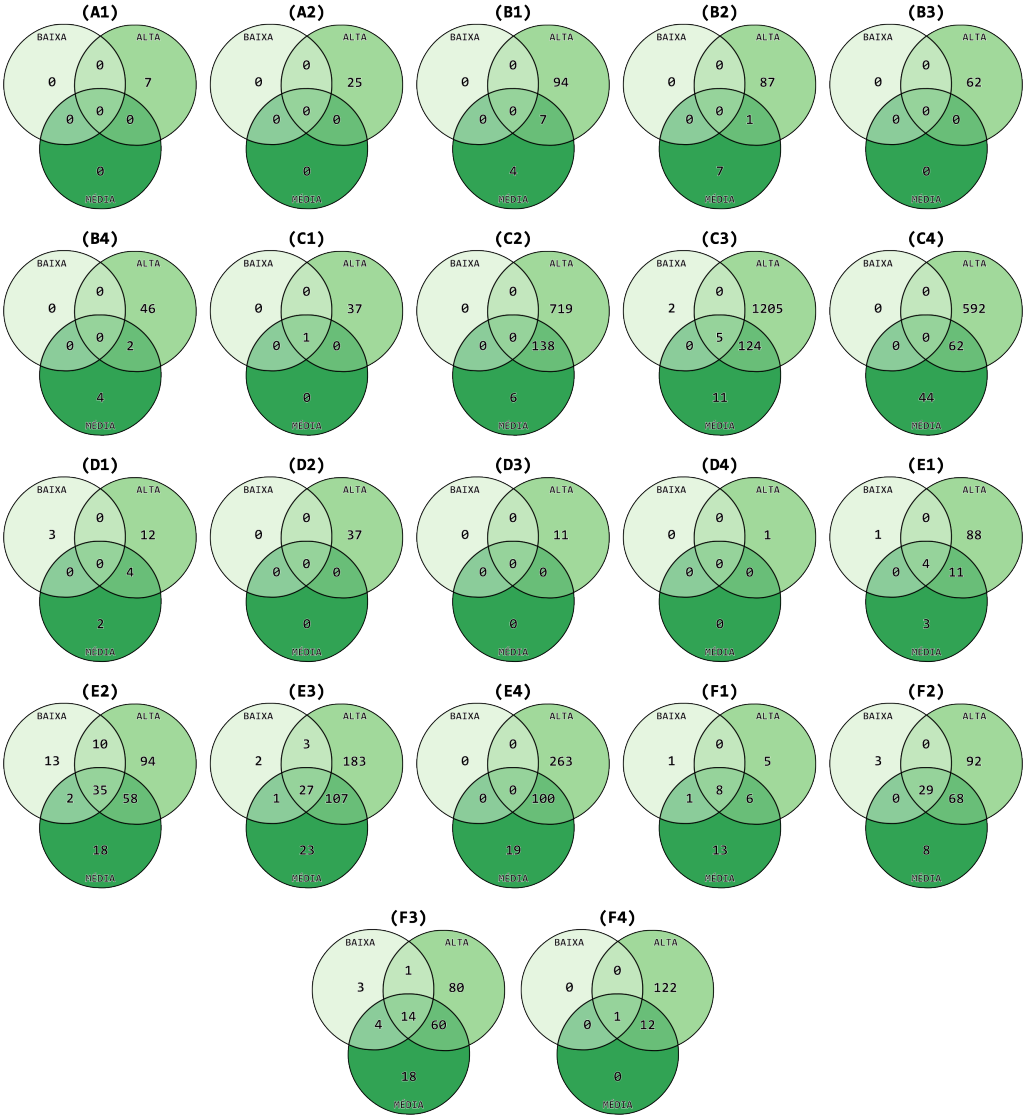

Figura C.3: Diagramas de *Venn* com a respectiva correspondência da Tabela C.3.

APÊNDICE D – Tabela de drogas e doses

D.1 *Homo sapiens in vivo*

Droga	Controle	Baixa	Média	Alta
acarbose	0	-	2000	10000
acetamidofluorene	0	-	10	50
acetaminophen	0	200	1000	5000
acetazolamide	0	-	120	600
adapin	0	3	15	75
ajmaline	0	-	60	300
allopurinol	0	5	28	140
allyl alcohol	0	2	14	70
amiodarone	0	0	1	7
amitriptyline	0	-	3	15
aspirin	0	120	600	3000
azathioprine	0	2	14	72
benzbromarone	0	4	20	100
benziodarone	0	-	8	40
bromobenzene	0	8	40	200
bucetin	0	-	60	300
caffeine	0	-	1500	7500
captopril	0	-	1600	8000
carbamazepine	0	12	60	300
carbon tetrachloride	0	300	1500	7500
chloramphenicol	0	-	90	450
chlormadinone	0	-	8	40
chlormezanone	0	-	50	250
chlorpheniramine	0	-	18	90
chlorpromazine	0	0	4	20
chlorpropamide	0	-	150	750
cimetidine	0	12	60	300
ciprofloxacin	0	-	5	25
clofibrate	0	12	60	300
clomipramine	0	-	2	10
colchicine	0	-	800	4000
coumarin	0	12	60	300
cyclophosphamide	0	80	400	2000
danazol	0	-	7	35
dantrolene	0	-	2	10
diazepam	0	10	50	250
diclofenac	0	16	80	400
diltiazem	0	-	30	150
disopyramide	0	-	700	3500
disulfiram	0	-	12	60
enalapril	0	-	400	2000
erythromycin ethylsuccinate	0	-	1	5
ethambutol	0	-	800	4000

Tabela D.1: Tabela de drogas e suas respectivas doses (em μM) para *Homo sapiens in vitro*

Droga	Controle	Baixa	Média	Alta
ethanol	0	-	2000	10000
ethinylestradiol	0	-	3	15
ethionamide	0	-	120	600
ethionine	0	400	2000	10000
etoposide	0	-	66	330
famotidine	0	-	140	700
fenofibrate	0	-	6	30
fluphenazine	0	0	4	20
flutamide	0	2	10	50
furosemide	0	-	500	2500
gemfibrozil	0	4	20	100
glibenclamide	0	0	4	20
griseofulvin	0	0	4	20
haloperidol	0	0	4	20
hexachlorobenzene	0	1	6	30
hydroxyzine	0	-	8	40
ibuprofen	0	-	30	150
imipramine	0	-	3	15
indomethacin	0	8	40	200
iproniazid	0	-	240	1200
isoniazid	0	400	2000	10000
ketoconazole	0	0	3	15
labetalol	0	5	28	140
lomustine	0	4	24	120
lornoxicam	0	-	3	15
mefenamic acid	0	-	30	150
meloxicam	0	-	10	50
metformin	0	-	200	1000
methapyrilene	0	24	120	600
methimazole	0	-	2000	10000
methyldopa	0	-	10	50
methyltestosterone	0	0	4	20
mexiletine	0	-	60	300
monocrotaline	0	-	18	90
moxisylyte	0	-	80	400
naphthyl isothiocyanate	0	8	40	200
naproxen	0	-	120	600
nicotinic acid	0	-	2000	10000
nifedipine	0	-	30	150
nimesulide	0	-	66	330
nitrofurantoin	0	5	25	125
nitrofurazone	0	-	10	50
nitrosodiethylamine	0	-	2000	10000
omeprazole	0	24	120	600
papaverine	0	-	12	60
pemoline	0	-	15	75
penicillamine	0	-	2000	10000
perhexiline	0	0	3	15
phenacetin	0	-	120	600
phenobarbital	0	400	2000	10000
phenylanthranilic acid	0	-	40	200
phenylbutazone	0	16	80	400
phenytoin	0	2	12	60
promethazine	0	-	7	35
propylthiouracil	0	160	800	4000
quinidine	0	-	10	50
ranitidine	0	-	800	4000
rifampicin	0	2	14	70
simvastatin	0	-	6	30

Tabela D.2: Tabela de drogas e suas respectivas doses (em μM) para *Homo sapiens in vitro*

Droga	Controle	Baixa	Média	Alta
sulfasalazine	0	6	30	150
sulindac	0	-	600	3000
sulpiride	0	-	1000	5000
tacrine	0	-	16	80
tamoxifen	0	-	5	25
tannic acid	0	-	1	5
terbinafine	0	-	3	15
tetracycline	0	1	5	25
theophylline	0	-	2000	10000
thioacetamide	0	400	2000	10000
thioridazine	0	0	3	15
ticlopidine	0	-	4	20
tiopronin	0	-	400	2000
tolbutamide	0	-	400	2000
valproic acid	0	200	1000	5000
vitamin A	0	-	1	7
WY-14643	0	6	30	150

Tabela D.3: Tabela de drogas e suas respectivas doses (em μM) para *Homo sapiens in vitro*

D.2 *Rattus norvegicus in vitro*

Droga	Controle	Baixa	Média	Alta
acarbose	0	400	2000	10000
acetamidofluorene	0	2	10	50
acetaminophen	0	1000	3000	10000
acetazolamide	0	24	120	600
adapin	0	3	15	75
ajmaline	0	12	60	300
allopurinol	0	5	28	140
allyl alcohol	0	0	4	20
amiodarone	0	0	1	7
amitriptyline	0	2	12	60
aspirin	0	120	600	3000
azathioprine	0	0	0	3
bendazac	0	8	40	200
benzbromarone	0	0	3	15
benziodarone	0	1	5	25
bromobenzene	0	8	40	200
bromoethylamine	0	20	100	500
bucetin	0	12	60	300
caffeine	0	400	2000	10000
captopril	0	400	2000	10000
carbamazepine	0	12	60	300
carbon tetrachloride	0	1000	3000	10000
carboplatin	0	120	600	3000
cephalothin	0	120	600	3000
chloramphenicol	0	18	90	450
chlormadinone	0	1	8	40
chlormezanone	0	10	50	250
chlorpheniramine	0	8	40	200
chlorpromazine	0	0	4	20
chlorpropamide	0	30	150	750
cimetidine	0	12	60	300
ciprofloxacin	0	1	5	25
cisplatin	0	8	40	200
clofibrate	0	12	60	300
clomipramine	0	1	8	40
colchicine	0	200	1000	5000
coumarin	0	12	60	300
cyclophosphamide	0	8	40	200
cyclosporine A	0	0	1	6
danazol	0	1	7	35
dantrolene	0	0	2	10
diazepam	0	5	25	125
diclofenac	0	16	80	400
diltiazem	0	10	50	250
disopyramide	0	100	500	2500
disulfiram	0	2	12	60
doxorubicin	0	0	0	2
enalapril	0	80	400	2000
erythromycin ethylsuccinate	0	3	15	75
ethambutol	0	160	800	4000
ethanol	0	400	2000	10000
ethinylestradiol	0	0	3	15
ethionamide	0	24	120	600
ethionine	0	400	2000	10000
etoposide	0	14	70	350
famotidine	0	28	140	700
fenofibrate	0	1	6	30
fluphenazine	0	1	6	30
flutamide	0	3	15	75
furosemide	0	100	500	2500
gemfibrozil	0	4	20	100
gentamicin	0	1	6	30
glibenclamide	0	2	12	60

Tabela D.4: Tabela de drogas e suas respectivas doses (em μM) para *Rattus norvegicus in vitro*

Droga	Controle	Baixa	Média	Alta
griseofulvin	0	1	6	30
haloperidol	0	2	10	50
hexachlorobenzene	0	0	3	15
hydroxyzine	0	6	30	150
ibuprofen	0	40	200	1000
imipramine	0	4	20	100
indomethacin	0	12	60	300
iproniazid	0	80	400	2000
isoniazid	0	400	2000	10000
ketoconazole	0	0	3	15
labetalol	0	5	28	140
lomustine	0	4	24	120
lornoxicam	0	0	3	15
mefenamic acid	0	6	30	150
meloxicam	0	2	10	50
metformin	0	40	200	1000
methapyrilene	0	0	3	15
methimazole	0	400	2000	10000
methyldopa	0	2	10	50
methyltestosterone	0	1	8	40
mexiletine	0	0	3	15
monocrotaline	0	3	18	90
moxisylyte	0	24	120	600
naphthyl isothiocyanate	0	8	40	200
naproxen	0	80	400	2000
nicotinic acid	0	400	2000	10000
nifedipine	0	10	50	250
nimesulide	0	3	15	75
nitrofurantoin	0	5	25	125
nitrofurazone	0	12	60	300
nitrosodiethylamine	0	400	2000	10000
omeprazole	0	4	24	120
papaverine	0	4	20	100
pemoline	0	3	15	75
penicillamine	0	400	2000	10000
perhexiline	0	0	2	10
phenacetin	0	24	120	600
phenobarbital	0	1000	3000	10000
phenylanthranilic acid	0	8	40	200
phenylbutazone	0	16	80	400
phenytoin	0	2	12	60
promethazine	0	3	16	80
propylthiouracil	0	160	800	4000
puromycin aminonucleoside	0	100	500	2500
quinidine	0	8	40	200
ranitidine	0	160	800	4000
rifampicin	0	2	14	70
simvastatin	0	2	12	60
sulfasalazine	0	4	20	100
sulindac	0	80	400	2000
sulpiride	0	200	1000	5000
tacrine	0	8	40	200
tamoxifen	0	0	0	3
tannic acid	0	0	2	10
terbinafine	0	0	3	15
tetracycline	0	1	5	25
theophylline	0	400	2000	10000
thioacetamide	0	400	2000	10000
thioridazine	0	0	2	10
ticlopidine	0	2	12	60
tiopronin	0	1	5	25
tolbutamide	0	80	400	2000
triamterene	0	1	6	30
triazolam	0	0	2	10
trimethadione	0	400	2000	10000
valproic acid	0	400	2000	10000
vitamin A	0	0	1	7
WY-14643	0	8	40	200

Tabela D.5: Tabela de drogas e suas respectivas doses (em μM) para *Rattus norvegicus in vitro*

D.3 *Rattus norvegicus in vivo*

Droga	Controle	Baixa	Média	Alta
acarbose	0	100	300	1000
acetamidofluorene	0	100	300	1000
acetaminophen	0	300	600	1000
acetazolamide	0	60	200	600
adapin	0	100	300	-
ajmaline	0	30	100	300
allopurinol	0	15	50	150
allyl alcohol	0	3	10	30
amiodarone	0	200	600	2000
amitriptyline	0	15	50	150
aspirin	0	450	1000	2000
azathioprine	0	3	10	30
bendazac	0	100	300	1000
benzbromarone	0	20	60	200
benziodarone	0	30	100	300
bromobenzene	0	30	100	300
bromoethylamine	0	6	20	60
bucetin	0	300	1000	2000
caffeine	0	10	30	100
captopril	0	100	300	1000
carbamazepine	0	30	100	300
carbon tetrachloride	0	30	100	300
carboplatin	0	10	30	100
cephalothin	0	300	1000	2000
chloramphenicol	0	100	300	1000
chlormadinone	0	300	1000	2000
chlormezanone	0	50	150	500
chlorpheniramine	0	3	10	30
chlorpromazine	0	45	150	-
chlorpropamide	0	30	100	300
cimetidine	0	100	300	1000
ciprofloxacin	0	100	300	1000
cisplatin	0	0	1	3
clofibrate	0	30	100	300
clomipramine	0	10	30	100
colchicine	0	1	5	15
coumarin	0	15	50	150
cyclophosphamide	0	15	50	150
cyclosporine A	0	30	100	300
danazol	0	300	1000	2000
dantrolene	0	25	75	250
diazepam	0	25	75	250
diclofenac	0	10	30	100
diltiazem	0	80	240	800
disopyramide	0	40	120	400
disulfiram	0	60	200	600
doxorubicin	0	1	3	10
enalapril	0	60	200	600
erythromycin ethylsuccinate	0	100	300	1000
ethambutol	0	100	300	1000
ethanol	0	400	1200	4000
ethinylestradiol	0	1	3	10
ethionamide	0	100	300	1000
ethionine	0	25	80	250
etoposide	0	10	100	1000
famotidine	0	100	300	1000
fenofibrate	0	10	100	1000
fluphenazine	0	2	6	20

Tabela D.6: Tabela de drogas e suas respectivas doses (em *mg/kg*) para *Rattus norvegicus in vivo*

Droga	Controle	Baixa	Média	Alta
flutamide	0	15	50	150
furosemide	0	30	100	300
gemfibrozil	0	30	100	300
gentamicin	0	10	30	100
glibenclamide	0	100	300	1000
griseofulvin	0	100	300	1000
haloperidol	0	3	10	30
hexachlorobenzene	0	300	1000	2000
hydroxyzine	0	10	30	100
ibuprofen	0	60	200	400
imipramine	0	10	30	100
indomethacin	0	5	15	50
iproniazid	0	6	20	60
methapyrilene	0	10	30	100
methimazole	0	10	30	100
methyldopa	0	60	200	600
methyltestosterone	0	30	100	300
mexiletine	0	40	120	400
monocrotaline	0	3	10	30
moxisylyte	0	50	150	500
naphthyl isothiocyanate	0	15	50	150
naproxen	0	20	60	200
nicotinic acid	0	100	300	1000
nifedipine	0	100	300	1000
nimesulide	0	30	100	300
nitrofurantoin	0	100	300	600
nitrofurazone	0	30	100	300
nitrosodiethylamine	0	10	30	100
omeprazole	0	100	300	1000
papaverine	0	40	120	400
pemoline	0	7	25	75
penicillamine	0	100	300	1000
perhexiline	0	15	50	150
phenacetin	0	300	1000	2000
phenobarbital	0	100	150	300
phenylanthranilic acid	0	300	1000	2000
phenylbutazone	0	20	60	200
phenytoin	0	600	1200	2000
promethazine	0	20	60	200
propylthiouracil	0	10	30	100
puromycin aminonucleoside	0	12	40	120
quinidine	0	20	60	200
ranitidine	0	100	300	1000
rifampicin	0	20	60	200
simvastatin	0	40	120	400
sulfasalazine	0	100	300	1000
sulindac	0	15	50	150
sulpiride	0	300	1000	2000
tacrine	0	3	10	30
tamoxifen	0	6	20	60
tannic acid	0	100	300	1000
terbinafine	0	75	250	750
tetracycline	0	100	300	1000
theophylline	0	20	60	200
thioacetamide	0	4	15	45
thioridazine	0	10	30	100
ticlopidine	0	100	300	1000
tiopronin	0	100	300	1000
tolbutamide	0	100	300	1000
triamterene	0	15	50	150
triazolam	0	100	300	1000
trimethadione	0	50	150	500
valproic acid	0	45	150	450
vitamin A	0	10	30	100
WY-14643	0	10	30	100

Tabela D.7: Tabela de drogas e suas respectivas doses (em *mg/kg*) para *Rattus norvegicus in vivo*