

RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo deste trabalho será disponibilizado somente a partir de 14/07/2019.

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
INSTITUTO DE BIOCÊNCIAS
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 BIOCÊNCIAS
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	<i>Critical Assessment of Massive Data Analysis</i>
FDA	<i>Food and Drug Administration</i>
GO	<i>Gene Ontology</i>
KEGG	<i>Kyoto Encyclopedia of Genes and Genomes</i>
MAS5	<i>Microarray Analysis Suite 5</i>
MM	<i>Mismatch</i>
NIH	<i>National Institutes of Health</i>
PM	<i>Perfect match</i>
PTGJ	Projeto Toxicogenômico Japonês
Reactome	<i>Reactome Pathway</i>
RMA	<i>Robust MultiArray</i>
TGPI	<i>Toxicogenomics Project in Japan</i>

Lista de Figuras

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/	p. 3
1.2	Integração das Ciências Ômicas e suas respectivas principais tecnologias.	p. 5
1.3	Correspondência entre as unidades do DNA e do RNA e os aminoácidos da proteína a ser sintetizada (JÚNIOR; SASSON, 2005).	p. 6
1.4	Realização de um experimento de microarranjo para amostras de células caso e células controle. Inicialmente são coletadas células 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.	p. 7
1.5	Processamento de dados de microarranjo	p. 8
1.6	<i>Workflow</i> dos formatos de arquivos gerados no processamento de um <i>chip</i> da <i>Affymetrix</i> . Cada formato está especificado na Tabela 1.2.	p. 8
3.1	Visão geral do material e métodos utilizados.	p. 17

3.2	Exemplo de análise com um <i>input</i> de dados de expressão de sete <i>arrays</i> (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 circulos correspondem aos passos no fluxograma. Neste exemplo, os conjuntos de genes ciclo celular e reparo de DNA são significativamente induzidos nos <i>arrays</i> cafeína_M2, cafeína_H2, etanol_M24, etanol_H24 e, portanto, constituem um <i>cluster</i> de conjuntos de genes, enquanto que o conjunto de genes resposta imune é significativamente reprimido nos <i>arrays</i> cafeína_H2 e omeprazol_H8, portanto, constitui seu próprio <i>cluster</i> de conjuntos de genes. O módulo resultante do primeiro <i>cluster</i> de conjuntos de genes inclui os genes 2, 3, 4, 5 e 6, uma vez que estes genes contribuem para a expressão significativa deste <i>cluster</i> . No passo final da análise, os <i>arrays</i> 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 <i>arrays</i> onde o módulo 1 é significativamente induzido (<i>arrays</i> 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).	p. 22
3.3	Exemplo do método aplicado para obter-se a média de todos os genes, em todos os <i>arrays</i> , igual a 0.	p. 23
3.4	Distribuição dos <i>sets</i> de genes para <i>Homo sapiens</i> e <i>Rattus norvegicus</i> . Lembrando que foi realizada uma intersecção de cada <i>set</i> de gene para as, duas espécies, de forma que possuam os <i>sets</i> iguais.	p. 24
3.5	Exemplificação da construção da primeira tabela para a identificação dos genes que alteram significativamente para a normalização do tipo RMA.	p. 26
3.6	Exemplificação da construção da nova tabela para a identificação dos genes que alteram significativamente. Esses cálculos são realizados linha a linha da segunda tabela (6 linhas totais) da Figura 3.5. Cálculo baseados na normalização RMA.	p. 26
3.7	Exemplo da metodologia do <i>multiscale bootstrap resampling</i> . Nesse exemplo o valor 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.	p. 29
3.8	Exemplo de uma “árvore” montada com os dados de <i>sets</i> de genes além da identificação dos nodos e folhas.	p. 30
3.9	Fluxograma para a obtenção do <i>Heatmap</i> .	p. 31

3.10	Informações contidas no mapa modular. O mapa modular é dividido em 5 partes: condições clínicas, <i>clusters</i> , genes por <i>clusters</i> , <i>arrays</i> por condições clínicas e o heatmap.	p. 31
4.1	Média da quantidade de genes diferencialmente expressos presentes em todas as drogas para os 3 tipos de experimento (<i>Homo sapiens in vitro</i> , <i>Rattus norvegicus in vitro</i> e <i>Rattus noevigicus in vivo</i>) em relação aos diferentes tempos de amostragens e concentrações de doses.	p. 33
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.	p. 34
4.3	Quantidade de GOs enriquecidas para as 131 drogas nos três experimentos: em azul representando <i>Homo sapiens in vitro</i> , em verde <i>Rattus norvegicus in vitro</i> e em vermelho <i>Rattus norvegicus in vivo</i>	p. 35
4.4	Quantidade de KEGGs enriquecidos para as 131 drogas nos três experimentos: em azul representando <i>Homo sapiens in vitro</i> , em verde <i>Rattus norvegicus in vitro</i> e em vermelho <i>Rattus norvegicus in vivo</i>	p. 35
4.5	Quantidade de REACTOMEs enriquecidos para as 131 drogas nos três experimentos: em azul representando <i>Homo sapiens in vitro</i> , em verde <i>Rattus norvegicus in vitro</i> e em vermelho <i>Rattus norvegicus in vivo</i>	p. 36
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 (<i>Homo sapiens in vitro</i> , <i>Rattus norvegicus in vitro</i> e <i>Rattus norvegicus in vivo</i>).	p. 37
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.	p. 38
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 <i>KEGGs</i> enriquecidos. (C) Quantidade de <i>REACTOMEs</i> enriquecidos. (D) Quantidade de <i>GOs</i> do tipo processos biológicos enriquecidos. (E) Quantidade de <i>GOs</i> do tipo funções moleculares enriquecidos. (F) Quantidade de <i>GOs</i> do tipo componentes celulares enriquecidos.	p. 41

4.9	Relação da quantidade de <i>GOs</i> do tipo processo biológico para as 6 drogas selecionadas em relação a cada um dos três experimentos.	p. 42
4.10	Relação da quantidade de <i>KEGGs</i> para as 6 drogas selecionadas em relação a cada um dos três experimentos.	p. 43
4.11	<i>Heatmap</i> gerado para o experimento <i>Rattus norvegicus in vitro</i> relacionado com <i>GO</i> do tipo função molecular. As caixas amarelas estão evidenciando 2 tipos de perfil, um induzido (predominância de vermelho) e outro reprimido (predominância de verde).	p. 47
4.12	<i>Heatmap</i> gerado para o experimento <i>Homo sapiens in vitro</i> relacionado com <i>Reactome</i> As caixas amarelas estão evidenciando 2 tipos de perfil presentes, um induzido (predominância de vermelho) e outro reprimido (predominância de verde).	p. 49
4.13	<i>Heatmap</i> gerado para o experimento <i>Homo sapiens in vitro</i> relacionado com <i>GO</i> do tipo processo biológico. Está destacado em amarelo o <i>cluster 27</i> além da condição estudada que foi a aparição de opacidade em vidro fosco.	p. 52
4.14	<i>Heatmap</i> gerado para o experimento <i>Rattus norvegicus in vitro</i> relacionado com <i>GO</i> do tipo processo biológico com concentração de dose alta e tempo de amostragem igual a 24h. Está destacado em amarelo o <i>cluster 3</i> além da condição estudada que foi a fibrose.	p. 55
4.15	<i>Heatmap</i> gerado para o experimento <i>Homo sapiens in vitro</i> relacionado com <i>Reactome</i> com concentração de dose alta e tempo de amostragem igual a 24 horas. Está destacado em amarelo o <i>cluster 19</i> além da condição estudada que foi a fibrose.	p. 57
4.16	<i>Heatmap</i> gerado para o experimento <i>Homo sapiens in vitro</i> relacionado com <i>Reactome</i> com concentração de dose alta e tempo de amostragem igual a 24 horas. Está destacado em amarelo o <i>cluster 2</i> além da condição estudada que foi a degeneração gordurosa.	p. 60
C.1	Diagramas de <i>Venn</i> com a respectiva correspondência da Tabela C.1.	p. 73
C.2	Diagramas de <i>Venn</i> com a respectiva correspondência da Tabela C.2.	p. 75
C.3	Diagramas de <i>Venn</i> com a respectiva correspondência da Tabela C.3.	p. 77

Lista de Tabelas

1.1	As ciências ômicas e suas definições	p. 4
1.2	Descrição de cada arquivo gerado pelo processamento de um <i>chip</i> da <i>Affymetrix</i> . . .	p. 8
1.3	Informações do Projeto Toxicogenômico Japonês (UEHARA et al., 2010).	p. 13
1.4	Resumo do PTGJ para dados de fígado (UEHARA et al., 2010).	p. 14
3.1	Exemplo do arquivo criado que relaciona cada <i>.CEL</i> com a concentração de dose e tempo de amostragem.	p. 19
3.2	Genes diferencialmente expressos encontrados para o Etanol com combinação de dosagem alta (<i>high</i>) e tempo de amostragem de 8 horas (H8C8) para a normalização do tipo RMA.	p. 20
3.3	Exemplo de <i>sets</i> de genes obtidos com os respectivos genes presentes para <i>Gene Ontology</i> (GO:0000002), <i>KEGG</i> (hsa:10000) e <i>Reactome</i> (r-hsa-1059683).	p. 24
3.4	Matriz montada para a obtenção do p-valor a partir do teste exato de Fisher.	p. 27
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.	p. 38
4.2	Diferença entre a quantidade total de <i>GOs</i> disponíveis inicialmente em contraste com a quantidade de <i>GOs</i> após a aplicação do filtro.	p. 44
4.3	Tabela com as respectivas <i>GOs</i> presentes no perfil muito induzido para as condições: edema, proliferação, vacuolização nuclear, fibrose, nódulo hepatodiafragmático, morte celular e degeneração acidófila e basófila.	p. 46
4.4	Tabela com as respectivas <i>GOs</i> 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.	p. 46

4.5	Tabela com os 15 maiores valores de <i>scores</i> , média e variância para <i>Homo sapiens in vitro</i> com <i>Reactome</i>	p. 48
4.6	Tabela contendo as <i>top 15</i> informações relativas ao valor de <i>score</i> , média e variância para <i>Homo sapiens in vitro</i> com GO do tipo processo biológico.	p. 50
4.7	GOs do tipo processos biológicos presentes no <i>cluster 27</i> para <i>Homo sapiens in vitro</i>	p. 51
4.8	Tabela contendo as <i>top 15</i> informações relativas ao valor de <i>score</i> , média e variância para <i>Rattus norvegicus in vitro</i> com GO do tipo processo biológico e concentração de dose alta com tempo de amostragem igual a 24 horas.	p. 53
4.9	Principais GOs do tipo processo biológicos presentes no <i>cluster 3</i> para <i>Rattus norvegicus in vitro</i> na concentração de dose alta com tempo de amostragem igual a 24 horas.	p. 54
4.10	Tabela contendo as <i>top 15</i> informações relativas ao valor de <i>score</i> , média e variância para <i>Rattus norvegicus in vitro</i> com <i>Reactome</i> e concentração de dose alta com tempo de amostragem igual a 24 horas.	p. 56
4.11	<i>Reactomes</i> presentes no <i>cluster 19</i> para <i>Homo sapiens in vitro</i>	p. 56
4.12	Tabela contendo as <i>top 15</i> informações relativas ao valor de <i>score</i> , média e variância para <i>Rattus norvegicus in vivo</i> com KEGG.	p. 58
4.13	KEGGs presentes no <i>cluster 2</i> para <i>Rattus norvegicus in vivo</i>	p. 59
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.	p. 69
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).	p. 70
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).	p. 71
C.1	Quantidade de genes diferencialmente expressos encontrados para <i>Homo sapiens in vitro</i> com a normalização RMA.	p. 72

C.2	Quantidade de genes diferencialmente expressos encontrados para <i>Rattus norvegicus</i> <i>in vitro</i> com a normalização RMA.	p. 74
C.3	Quantidade de genes diferencialmente expressos encontrados para <i>Rattus norvegicus</i> <i>in vivo</i> com a normalização RMA.	p. 76
D.1	Tabela de drogas e suas respectivas doses (em μM) para <i>Homo sapiens in vitro</i> . . .	p. 78
D.2	Tabela de drogas e suas respectivas doses (em μM) para <i>Homo sapiens in vitro</i> . . .	p. 79
D.3	Tabela de drogas e suas respectivas doses (em μM) para <i>Homo sapiens in vitro</i> . . .	p. 80
D.4	Tabela de drogas e suas respectivas doses (em μM) para <i>Rattus norvegicus in vitro</i> .	p. 81
D.5	Tabela de drogas e suas respectivas doses (em μM) para <i>Rattus norvegicus in vitro</i> .	p. 82
D.6	Tabela de drogas e suas respectivas doses (em mg/kg) para <i>Rattus norvegicus in vivo</i>	p. 83
D.7	Tabela de drogas e suas respectivas doses (em mg/kg) para <i>Rattus norvegicus in vivo</i>	p. 84

Sumário

Resumo	p. iv
Abstract	p. v
1 Introdução	p. 1
1.1 Experimentos <i>in vivo</i> e <i>in vitro</i>	p. 1
1.2 As Ciências Ômicas	p. 3
1.3 Toxicogenômica	p. 4
1.4 Microarranjo	p. 5
1.4.1 MAS5	p. 9
1.4.2 RMA	p. 10
1.4.3 GCRMA	p. 11
1.4.4 Comparação entre os métodos	p. 12
1.5 Projeto Toxicogenômico Japonês (PTGJ)	p. 12
1.6 Avaliação Crítica de Análise de Dados em Massa (CAMDA)	p. 14
2 Objetivos	p. 16
2.1 Objetivos Específicos	p. 16
3 Material e Métodos	p. 17
3.1 <i>Workflow</i>	p. 17
3.2 <i>Hardware</i>	p. 17
3.3 Obtenção dos Dados de Microarranjo	p. 18
3.4 Pré-processamento dos dados	p. 18

3.5	Enriquecimento Funcional	p. 21
3.6	Mapa Modular	p. 21
3.6.1	Obtenção dos Dados de Expressão	p. 23
3.6.2	Obtenção dos <i>Sets</i> de Genes	p. 23
3.6.3	Identificação dos <i>Arrays</i> em que a expressão dos <i>sets</i> de genes estão alterados	p. 25
3.6.4	Matriz de relação <i>Sets</i> de Genes x <i>Arrays</i>	p. 27
3.6.5	Obtenção e tratamento dos <i>Clusters</i>	p. 28
3.6.6	Construção do <i>Heatmap</i>	p. 30
4	Resultados e Discussão	p. 32
4.1	Análise Clássica Global	p. 32
4.1.1	Genes Diferencialmente Expressos	p. 32
4.1.2	Enriquecimento Funcional	p. 34
4.2	Análise Clássica Local	p. 37
4.2.1	Genes Diferencialmente Expressos	p. 37
4.2.2	Enriquecimento Funcional	p. 39
4.3	Mapa Modular	p. 44
4.3.1	Limitação Computacional	p. 44
4.3.2	Análises	p. 45
5	Conclusões	p. 61
	Referências Bibliográficas	p. 63
	Apêndice A	p. 68
	Apêndice B	p. 69
	Apêndice C	p. 72

C.1	<i>Homo sapiens in vitro</i>	p. 72
C.2	<i>Rattus norvegicus in vitro</i>	p. 74
C.3	<i>Rattus norvegicus in vivo</i>	p. 76

Apêndice D p. 78

D.1	<i>Homo sapiens in vivo</i>	p. 78
D.2	<i>Rattus norvegicus in vitro</i>	p. 81
D.3	<i>Rattus norvegicus in vivo</i>	p. 83

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

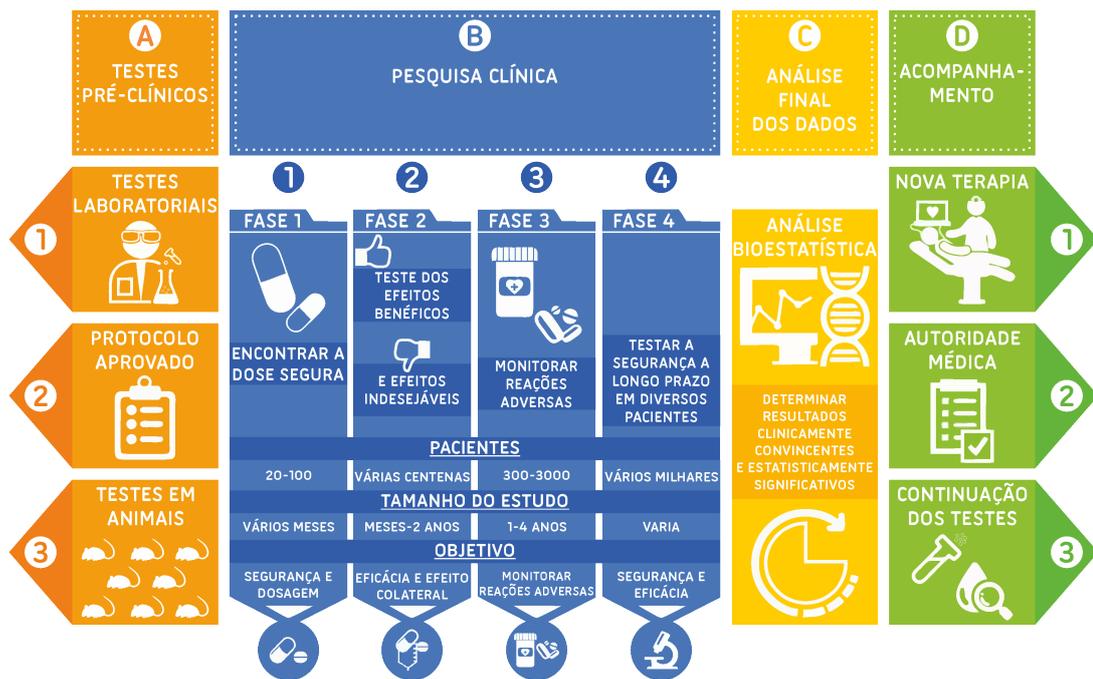


Figura 1.1: O processo de desenvolvimento de uma droga, como mostrado nesta figura, passa por alguns passos que são: testes pré-clínicos (A), pesquisa clínica (B), análise final dos dados (C) e por fim através do acompanhamento (D). Cada passo possui suas particularidades e são de suma importância para, 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.

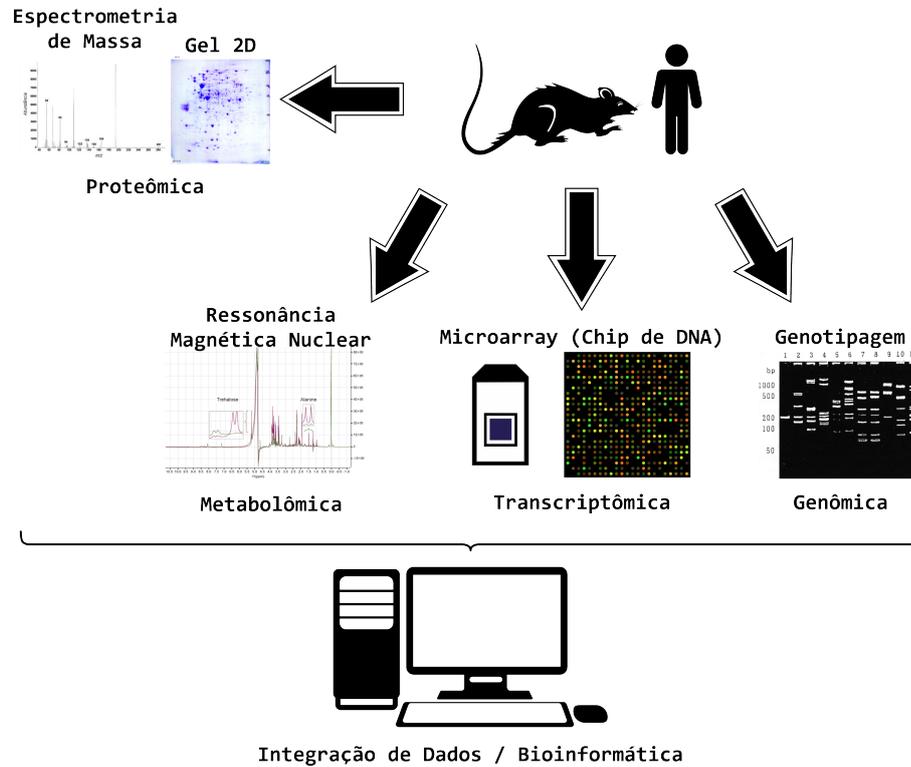


Figura 1.2: Integração das Ciências Ômicas e suas respectivas principais tecnologias.

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.

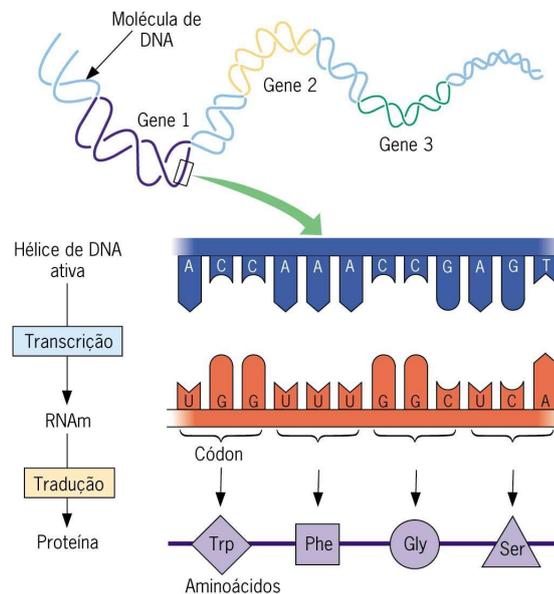


Figura 1.3: Correspondência entre as unidades do DNA e do RNA e os aminoácidos da proteína a ser sintetizada (JÚNIOR; SASSON, 2005).

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.

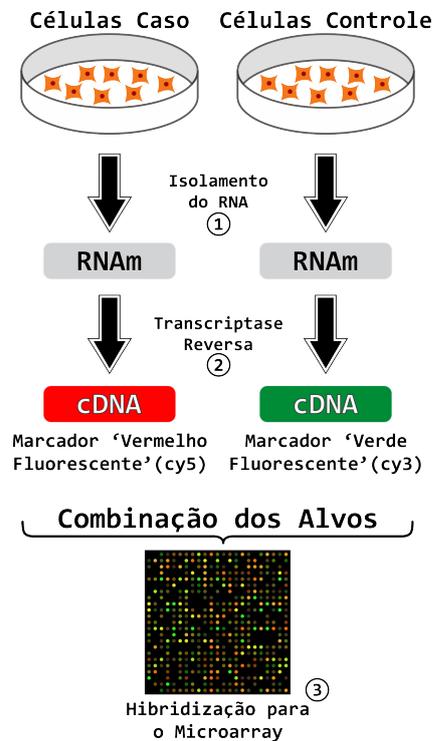


Figura 1.4: Realização de um experimento de microarray para amostras de células caso e células controle. Inicialmente são coletadas células do caso e do controle. Em seguida é feito o isolamento do RNA, sendo obtido o RNA mensageiro (RNAm). A partir do RNAm e com a utilização da transcriptase reversa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

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-

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

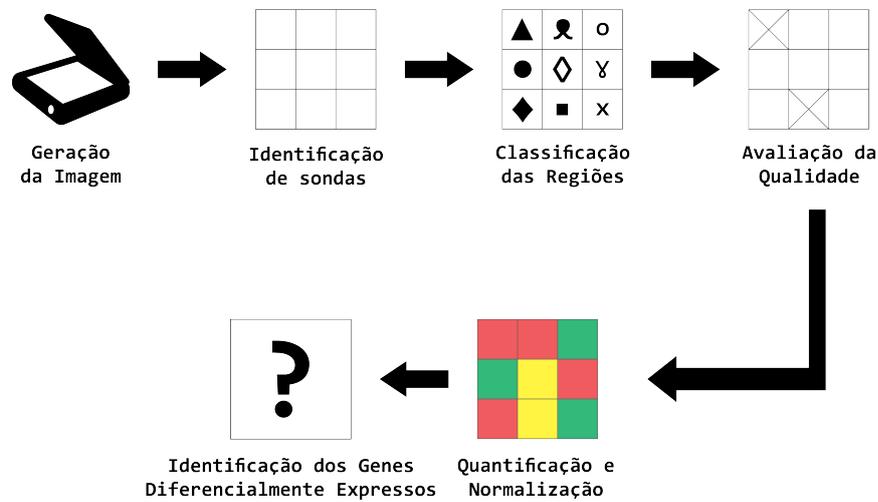


Figura 1.5: Processamento de dados de microarranjo

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.

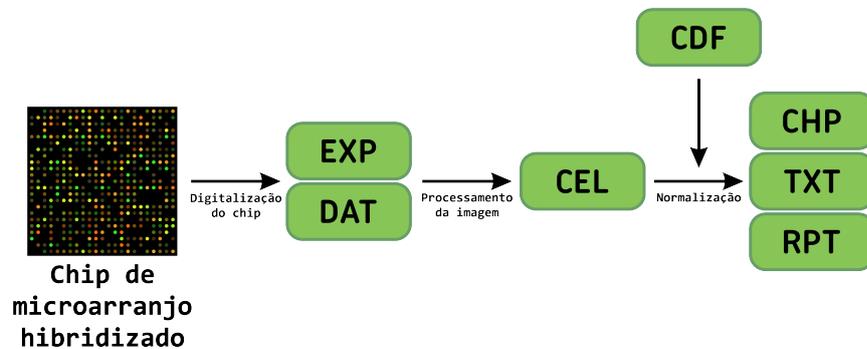


Figura 1.6: *Workflow* dos formatos de arquivos gerados no processamento de um *chip* da *Affymetrix*. Cada formato está especificado na Tabela 1.2.

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 <i>Affy</i> e descreve o <i>layout</i> 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 <i>software Affy</i> 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 ,

$$IM_{ij} = \begin{cases} MM_{ij} & | \quad MM_{ij} < PM_{ij}, \\ \frac{PM_{ij}}{2^{SB_{ij}^+}} & | \quad MM_{ij} \geq PM_{ij} \end{cases} \quad (1.1)$$

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

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

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

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

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

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

Finalmente, o valor de expressão é calculado por

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

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

$$Y_{ijk} = \mu_{ij} + \alpha_{jk} + \varepsilon_{ijk} \quad (1.6)$$

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,

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

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 *booleana* 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) (GHARAIIBEH; 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).

	<i>Rat in vivo</i>	<i>Rat in vitro</i>	<i>Human in vitro</i>
Animal	<i>Sprague-Dawley</i>	<i>Sprague-Dawley</i>	-
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).

	<i>Homo sapiens</i> <i>in vitro</i>	<i>Rattus norvegicus</i> <i>in vitro</i>	<i>Rattus norvegicus</i> <i>in vivo</i>	<i>Rattus norvegicus</i> <i>in vivo</i>
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 prever 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.

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 norvegicus 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 possível 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 ou 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 inviabilizando, 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 norvegicus 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.