UNIVERSIDADE FEDERAL DO PAMPA
CAMPUS SÃO GABRIEL

LEONARDO OTAKE

# ANÁLISE E AVALIAÇÃO QUALITATIVA DA EXPRESSÃO DE TRANSCRITOS QUE CODIFICAM PROTEÍNAS DE MEMBRANA A PARTIR DA DETERMINAÇÃO DE LIMIAR DE PRESENÇA/AUSÊNCIA

SÃO GABRIEL
2015

LEONARDO OTAKE

# ANÁLISE E AVALIAÇÃO QUALITATIVA DA EXPRESSÃO DE TRANSCRITOS QUE CODIFICAM PROTEÍNAS DE MEMBRANA A PARTIR DA DETERMINAÇÃO DE LIMIAR DE PRESENÇA/AUSÊNCIA

Trabalho de conclusão de curso III, como pré-requisito parcial para obtenção do grau de Bacharel em Biotecnologia na Universidade Federal do Pampa.

Orientador: Andrés Delgado Cañedo

SÃO GABRIEL

2015

Dedico este trabalho ao caos por sua aleatoriedade intrínseca, concedendo sempre as oportunidades para viver minhas experiências.

**LEONARDO OTAKE**

# ANÁLISE QUALITATIVA DA EXPRESSÃO DE TRANSCRITOS QUE CODIFICAM PROTEÍNAS DE MEMBRANA A PARTIR DA DETERMINAÇÃO DE LIMIAR DE PRESENÇA/AUSÊNCIA

Trabalho de conclusão de curso III, como pré-requisito parcial para obtenção do grau de Bacharel em Biotecnologia na Universidade Federal do Pampa.

Orientador: Andrés Delgado Cañedo

Trabalho de conclusão de curso defendido em: 4 de Dezembro de 2015.
Banca examinadora:

_____

Prof. Doutor. Andrés Delgado Cañedo
Orientador
Biotecnologia - UNIPAMPA

_____

Prof. Doutor. Cristhian Augusto Bugs
Engenharia Florestal -UNIPAMPA

_____

Prof. Doutor. José Ricardo Inácio Ribeiro
Biologia -UNIPAMPA

# RESUMO

Com o desenvolvimento de técnicas *hightroughput* na biologia molecular, houve o nascimento da genômica. Desde então a geração de dados biológico cresceu de forma exponencial e grande parte destes dados estão disponíveis gratuitamente em banco de dados biológicos como o do NCBI. A técnica do microarranjo analisa os transcritos de mRNA através de uma abordagem global e foi por muito tempo a técnica mais popular na área da genômica funcional. Os dados brutos de microarranjo são produzidos através do escaneamento das intensidades produzidas pela hibridização dos transcritos com as sondas alvo. É necessário uma etapa de pré-processamento para converter o sinal de intensidade de luz para valores de expressão. Diversos algorítimos de pré-processamento foram desenvolvidos e nenhum foi considerado "absoluto", pois dependendo do método de pré-processamento utilizados nas análise, resultados distintos podem ser encontrado. Poucos métodos de detecção de presença/ausência de genes foram desenvolvidos para microarranjos, e os que foram desenvolvidos possuem limitações de uso que limitam a exploração de dados entre diferentes plataformas ou o tipo de método de pré-processamento. No presente trabalho, um novo método de detecção de genes (*detection call)* que permite a utilização de qualquer método de pré-processamento e plataformas, foi desenvolvido baseados em dados empíricos de citometria de fluxo de células do sistema imunológico.



Palavras-chave: Microrranjo - Detecção de Genes - Validação Interna - Expressão Gênica - Genômica Funcional

ABSTRACT

The development of high-throughput techniques in the field of molecular biology brought the born? of genomics. Since then the generation of biological data rose in a exponential way and a good share of these data are freely available in biological databases like NCBI. The microarray technique analyses the mRNA transcripts in a global approach and for a long time the most popular technique in the field of functional genomics. The microarray raw data are created trough the scanning of light signals intensities produced by the hybridization of transcripts with its probes targets. It is necessary a preprocessing step to convert the light signal to expression values. Several preprocessing algorithms was developed but none is absolute, due to different results produced depending on which preprocessing method was choose. Few methods to determine the presence/absent of genes was created for microarray. Those whom was developed have some limitation concerning data exploration on different microarray platforms or the method of preprocessing. In the present work, a new method of detection call which permits the use of any preprocessing method and microarray platform was developed based on empirical data of flow cytometry using human cells of the immunological system.

Keywords: Microarray - Detection Call - Internal Validation - Gene Expression - Functional Genomics

SUMÁRIO

# 1. Introdução

## 1.1. O gene e a genômica funcional

Conceitualmente o gene é definido como uma seqüência de nucleotídeos necessária para a síntese de uma molécula funcional, a qual pode ser uma cadeia de polipeptídio ou uma molécula de RNA. É estimado que o genoma humano codifica aproximadamente 30.000 genes, entretanto cerca de 10.000 destes genes são transcritos ativamente. A expressão gênica é o processo que converte a informação do gene em produtos funcionais para célula e é regido pelo "Dogma Central da Biologia Molecular". O nível de expressão gênica teoricamente se refere ao número de cópias de transcrito em um dado momento, muitos desses genes são ubiquamente expressos em diferentes tecidos e tipos celulares; entretanto, existem genes que possuem expressão diferencial, restrita a determinados tecidos (ALBERTS, 2007). Enquanto o primeiro conjunto é responsável por processos comuns em todos os tecidos como, por exemplo, os processos metabólicos e fisiológicos, o segundo desempenha funções celulares e moleculares específicas ao tecido onde são preferencialmente expressos (SU; C. et al., 2002).

Mesmo contendo a mesma informação genética em todas as células do organismo, as células de tecidos distintos são diferentes. Essas diferenças são originadas por mudanças nos níveis de expressão e no padrão de ativação dos genes, o que é conhecido como perfil da expressão gênica. O processo da biologia molecular que realiza a análise sistemática destes perfis é conhecido como genomica funcional ou "transcriptômica" (HIETER; BOGUSKI, 1997). Diferentemente das ferramentas mais tradicionais da biologia molecular que avaliam um único ou poucos genes, a tecnologia do microarranjo permitiu mensurar a abundância de milhares de transcritos de mRNA ao mesmo tempo e devido a esta característica, a tecnologia do microarranjo foi popularizado no campo da biotecnologia como ferramenta para diagnóstico, identificação de biomarcadores e no estudo da função de genes. Nos últimos dez anos, com a popularização das tecnologias de abrangencia genômica, houve uma grande deposição de dados microarranjos e também de dados de RNA-seq em banco de dados público como o Gene Expression Omnibus (GEO) (BARRETT; T. et al., 2013)

## 1.2. Microarranjo Affymetrix

Os microarranjos do sistema *GeneChip Microarray* comercializados pela empresa Affymetrix são compostos por um conjunto de milhares de sequencias de sondas de DNA sintetizada em superfícies de quartzo organizadas em "spots". Uma plataforma da Affymetrix pode representar o genoma inteiro de um organismo. Cada spot possui um par de sonda, a sonda "*Perfect Match"* (PM) e a sonda "*Mismatch*" (MM), estas sondas possuem sequência complementar única de 25 bases que representa uma porção da sequência alvo, entretanto a sonda MM possui uma modificação na 13º base. Teoricamente a sonda MM ajudaria a eliminar o sinal de intensidade de hibridização inespecífica devido a diferença de temperatura de hibridização entre as sondas PM e MM com o gene alvo. Quando o alvo hibridiza no seu respectivo conjunto de sondas, um sinal de luz correspondente à média da intensidade de todo o conjunto é capturado por escaneamento no sistema de detecção. Contudo, outros fabricantes de microarranjos somente utilizam a sonda PM. (GOHLMANN; TALLOEN, 2009)

A intensidade de luz é relativa à expressão do gene do material analisado. Quando o chip contendo o microarranjo é escaneado, cada imagem do chip é depositada em um arquivo DAT. O valor numérico da intensidade das sondas obtido através da imagem escaneada, é convertido em arquivo CEL. Para obter um único valor que representa à abundancia do transcrito alvo é preciso derivar a intensidade das sondas específicas para esse transcrito. Nas plataformas Affymetrix que possuem as sondas *Mismatch*, a expressão do gene precisa ser calculada através das intensidades do conjunto de sondas (11-20 pares PM e MM) designadas a um único alvo. O nível de expressão resultante pode ser classificado qualitativamente com o método de "*Detection Call*" (P - Presente, A - Ausente, M - Marginal) (BESSANT; CONRAD, 2011; DZIUDA, 2010)

A geração de dados confiáveis é fundamental para qualquer análise. Devido às fontes de variações técnicas introduzidas pela marcação, hibridização e escaneamento dos microarranjos, os dados brutos contidos no arquivo CEL necessitam de uma etapa inicial de pré-processamento para obter o valor de expressão absoluto de cada gene (RAMASAMY et al., 2008). Esta etapa consiste em três partes: correção do "*background"*, normalização e sumariação. A correção do *background* corrige o sinal de intensidade do ruído do *background*

e ajusta os valores de hibridização não específica, e tem o maior efeito sobre os resultados pré-processados (IRIZARRY; WU; JAFFEE, 2006). A normalização tem como objetivo diminuir a variação não biológica interna e entre microarranjos. Assim que o sinal de intensidade das sonda é determinado, a sumariação combina todos os sinais em um único valor de expressão para o conjunto de sondas (IRIZARRY, 2003; IRIZARRY; WU; JAFFEE, 2006).

Diferentes algorítimos de pré-processamento para microarranjos foram desenvolvidos ao longo dos anos, e cada um deles possuem formas diferentes para calcular o valor de expressão. Consequentemente, cada algorítimo pode apresentar valores de expressão divergentes para um mesmo gene (LUO et al., 2010). Usualmente, trabalhos com experimento de microarranjo adotam apenas um método de pré-processamento para obtenção de valores de expressão. Alguns algorítimos tornaram-se populares, porém até o momento nenhum método tornou-se absoluto. Em um trabalho feito por Millenaar et al, seis algorítimos de pré-processamento foram analisados (MAS5, dChip PMMM, dChip PM, RMA, GC-RMA e PDNN) utilizando o mesmo conjunto de dados, e surpreendentemente sobreporam-se apenas 27-36% dos resultados entre os algorítimos comparados. (MILLENAAR et al., 2006).

## 1.2. Projeto Bioconductor

O projeto Bioconductor é uma iniciativa de colaboração com a finalidade de criação de softwares *open source*, extensível para biologia computacional e bioinformática. A linguagem de programação R foi escolhida para o desenvolvimento dos seus softwares, em virtude de ter implementado em seu *core* um amplo repertório de algorítmos matemáticos e estatístico com grande capacidade numérica, flexibilidade de visualização e diversas outras características que tornam esta linguagem um ambiente atrativo para desenvolvimento de *softwares* de bioinformática (R CORE TEAM; 2015). Um dos pilares deste projeto é a reprodutibilidade dos resultados, assim como os protocolos experimentais em biologia molecular os algorítimos desenvolvidos nas análises devem ser reproduzíveis e disponíveis para publicação. (GENTLEMAN et al., 2004).

Os pacotes disponíbilizados e mantidos pelo projeto Bioconductor promovem formas eficientes para métodos de processamento e mineração de dados biológicos. Alguns dos principais pacotes com algorítimos de análise e pré-processamento de microarranjos são: *affy* (GAUTIER; 2015), *gcrma* (JEAN; 2015) *e plier* (AFFYMETRIX; 2015). O pacote *affy* tem como objetivo ser um ambiente de exploração e de análise dos dados de microarranjos da Affymetrix. Este pacote inclui funções de controle de qualidade, avaliação de degradação de RNA, procedimentos de normalização, correção de brackground e de sumariação (GAUTIER et al., 2004). Contém os algorítimos implementados MAS5.0 (Affymetrix, 2001), DChip's MBEI e RMA (GAUTIER et al., 2004; LI; WONG, 2001). Já os pacotes gcrma e plier possuem os algoritmos GCRMA e PLIER respectivamente.

O algorítimo MAS5 realiza análises de forma independente para cada microarranjo, ajustando o *background* e o processo de sumariação e, então, todos os arranjos são dimensionados para ter o mesmo valor de expressão da média aparada (*"trimmed mean")*, mas devido à característica de analizar cada arranjo independentemente, o algoritmo MAS5 não leva em consideração afinidades sondas-específicas, reduzindo a capacidade de detectar pequenas mudanças entre arranjos se comparado com abordagens *"multichip"*. Na descrição do algorítimo MAS5, o manual do usário do software Affymetrix de expressão recomenda: "O uso primário do algorítmo MAS5 é obter um relatório rápido em relação a performance dos arranjos e para identificar problemas obvios, antes de submeter o conjunto final de arranjos para algum algorítimo de análise *multichip* (RMA, PLIER)" (TALLOEN, W. & GÖHLMANN, H., 2009)

O algorítimo RMA ("*Robust Multichip Analysis")* realiza a correção de *background* usando apenas as intensidades das sondas PM, normalizacão quantile e a sumarição robusta *multichip*. O ajuste de *background* leva em consideração as seguintes premissas: i) Cada arranjo possui a média de *background* em comum; ii) As sondas MM são ignoradas; iii) A sonda PM observada é modelada por uma convolução de um sinal S distribuido exponencialmente e um sinal de background B com distribuição normal ($Y = S + B$). O algorítimo GCRMA é uma versão alternativa do RMA que tem os mesmo processos de normalização e sumariação. Contudo, o ajuste da intensidade do *background* leva em consideração as informações das sequências de sondas MM, estimando as características de

hibridização específicas pelo conteúdo GC. O algorítimo PLIER (*"Probe Logarithmic Intensity Error")* desenvolvido pela Affymetrix, é similar ao GCRMA por contar com as informações de sequencia e por normalizar seus dados através do quantile. A principal diferença entre esses algorítimos é o seu modelo de erro. (TALLOEN; GÖHLMANN, 2009)

Por ser possível avaliar as consequências globais nos resultados de um único estímulo, sem o conhecimento prévio de quais genes devem ser afetados por esta variação, o experimento com a técnica do microarranjo torna-se uma ótima ferramenta para a geração de hipótese. Um dos aspectos limitantes dos microarranjos é a baixa sensibilidade para a alteração de genes com baixa expressão(ref). E a detecção de alterações em até 50% não são significantes, sendo assim o microarranjo possui menor sensibilidade em relação a técnicas de quantificação de RNA que testam alguns genes por vez. Sendo necessário a validação dos resultados utilizando técnicas como o qRT-PCR. (CANALES et al., 2006)

## 1.3 Sistema de proteinas CD (*Cluster definition*) do sistema hematopoiético

A interação das células com o seu universo é realizada através de proteínas na superfície de membrana. Conhecidas como "Cluster Definition" ou "Cluster Differentiation", os CD são um código criado usado para diferenciar e caracterizar diferentes proteínas da superficie de membrana das células hematopoiéticas. Atualmente mais de 500 CDs foram caracterizados e há estimativas de que exista cerca de 2.500 na superficie de leucócitos. Alguns CDs tem destaque no uso em citômica e no diagnóstico clínico via citometria de fluxo, como exemplo podemos citar o CD20, alvo para terapia por anticorpo, com rendimento de mercado anual de 2 bilhões de dólares e o CD4 alvo de entrada do HIV usado para diagnosticar a doença de forma rápida e confiável. Todos esses CDs são usados como alvos de anticorpos em técnicas como a citometria de fluxo (BREM . et al., 2007).

# 2. Justificativa

Embora tenha se avançado muito no processamento de dados para avaliar quantitativamente a expressão genica por microarranjos, poucos recursos para detecção de presença ou ausência de genes foram estabelecidos para dados de microarranjo. Por muito tempo o método de "*Detection Call*" era exclusividade do software MAS5 com a ferramenta de detecção MAS-P/A, até ser implementado no pacote *affy*. Este algorítimo de detecção suporta somente o método de pré-processamento MAS5, nenhum outro algorítimo de pre-processamento possui embutido algorítimo de análise de detecção de genes. O pacote PANP ("Presence-Absence calls with Negative Probesets") do projeto bioconductor foi desenvolvido para detecção da presença/ausência dos genes. Diferente da fator limitante do algoritmo MAS-P\A, esta ferramenta permite a utlização de diversos algorítimos de pré-processamento, entretanto processa apenas as séries da plataforma AffyMetrix HG-U133. Essa técnica utiliza como *cutoff* a média das intensidades do conjunto de sondas conhecidas por não haver alvo de hidrização, chamadas "*Negative Strand Matching Probesets"* (NSMPs). Contudo, a importância da determinação desse limite entre os valores de fluorescência de genes presente e não presente demanda o estabelecimento de um método confiável e que possa ser utilizado em qualquer plataforma.

Devido ao fato que a citômica do sistema hematológico provê dados consistentes e curados sobre a expressão de proteínas de membrana nos diferentes componentes celulares deste sistema e existem dados do perfis de expressão gênica dessas células obtidos através do uso de microarranjos, no presente trabalho nós comparamos diversos modelos de análise de dados de microarranjos de células hematopoiéticas e comparamos os dados obtidos com os dados já conhecidos de expressão de proteínas de superfície nessas células.

# 3. Objetivo Geral

Descrever um método para determinar a presença e ausência de transcritos em dados de microarranjos.

## 3.1 Objetivos específicos

I. Realizar análise transcritômica dos perfis de expressão gênica de células humanas do sistema imunológico.

II. Analisar os níveis de expressão correspondente aos *clusters definition* de dados de citometria de fluxo.

III. Escolher o melhor algorítimo de pré-processamento para estabelecer um nível de corte que possa determinar a presença/ausência de genes.

IV. Inferir possíveis marcadores de superfície de membrana, baseado no nível de corte determinado.

# 4. Metodologia

## 4.1 Prospecção dos dados de expressão global e dos *cluster definition*

Os dados brutos de microarranjo em formato CEL foram obtidos através do banco de dado do NCBI Gene Expression Omnibus. A série selecionada (GSE22886) da plataforma HG-U133A Affymetrix Human Genome U133A Array (GPL96), contém dados de expressão global de doze tipos célulares do sistema hematológico humano: linfócitos T CD8 e CD4, linfócitos B Naive e de memória, células Natural Killer, Monocitos, Macrófagos, células Dendríticas, Neutrófilos, entre outros. Extraidos de sangue periférico ou a partir da medula óssea.

Os dados de presença/ausência das proteínas de membrana que definem os subgrupos celulares dos sistema hematológico humano conhecidas como "*Cluster Definition*" ou "Clusters of Differentiation" (CD), foram obtidos da tabela *human CD marker chart* (https://www.bdbiosciences.com/documents/cd_marker_handbook.pdf), disponibilizada pela empresa BD. Os dados do chart são atualizados a cada 4 anos pelo consórcio agora conhecido como "*Human Cell Differentiation Molecules"* (HCDM) (ZOLA, H. et al., 2005). As respectivas sondas que correspondem as proteínas de marcador CD foram obtidas manualmente através do arquivo de anotação da plataforma de microarranjo Affymetrix.

## 4.2 Mineração de dados

A análise *in silico* foi realizada através da linguagem de programação R para garantir reprodutibilidade. Todos os códigos com comentários, pacotes utilizados e as funções criadas, estão disponíveis no material suplementar.

## 4.3 Pré-Processamento e controle de qualidade

Antes de seguir para o processo de obtenção dos dados de expressão, os métodos do pacote *affyPLM* (Brettschneider et al., 2007) *"Relative Log Expression"* (RLE) e o "*Normalized unscaled standard error"* (NUSE) foram aplicados para testar a qualidade dos microarranjos analisados. O RLE compara a expressão de um conjunto de sondas de um arranjo, a mediana

do nível de expressão da mesma sonda sobre todos os arranjos. O NUSE normaliza o erro padrão das intensidades através do arranjos para a mediana igual a um. Em seguida, quatro métodos foram utilizados para determinar os valores de expressão ou *"absolut call*": MAS5, PLIER, RMA e GCRMA. Cada um deles possuem diferentes suposições para o cálculo do valor de expressão e de processamento para as etapas de correção de *background*, normalização e da sumariação do conjunto de sondas. Os pacotes do Bioconductor gcrma, plier, e affy contém a implementação dos algoritmos supracitados e foram utilizados para o processamento de nível inferior das amostras.

### 4.3 *Detection Call*

O algorítmo "*Presence-Absence calls with Negative Probesets"* (PANP) implementado no pacote "PANP" foi utilizado para determinar a presença ou ausencia dos genes. Esse método utiliza o valor de expressão de conjuntos de sondas denominadas "*Negative Strand Matching Probests*" (NSMP) que não hibridizam com seus transcritos alvos por não serem complementares com seu gene alvo (um erro de desenho no chip que foi mantido pelos fabricantes) . Um método alternativo de *detection call* baseado em um valor de *cutoff,* gerado através da média da soma dos valores de expressão do segundo quartil de intensidade dos dados de presença, e o terceiro quartil da intensidade dos ausentes em citometria.

### 4.4. Diagramas de Venn

Os resultados de comparação dos resultados de detecção de gene entre o diferentes métodos de pré-processamento, foi realizado através da ferramenta on-line Venny, que possibilita extrair os valores de intersecção. (http://bioinfogp.cnb.csic.es/tools/venny/index.html)

# 5. Resultados e discussão

## 5.1 Controle de qualidade

O boxplot da distribuição dos valores RLE gerado pela função RLE (fig.1a), foi usado para identificar arranjos com baixa qualidade. Os valores RLE que possuem maior distribuição interquartil, ou os que não estão centrado ao redor de zero, indicam que o arranjo possui baixa qualidade. Já o boxplot com os valores NUSE (fig.1b) que possua um alto valor da mediana e maior abrangencia interquartil tambem indicam baixa qualidade de arranjos.

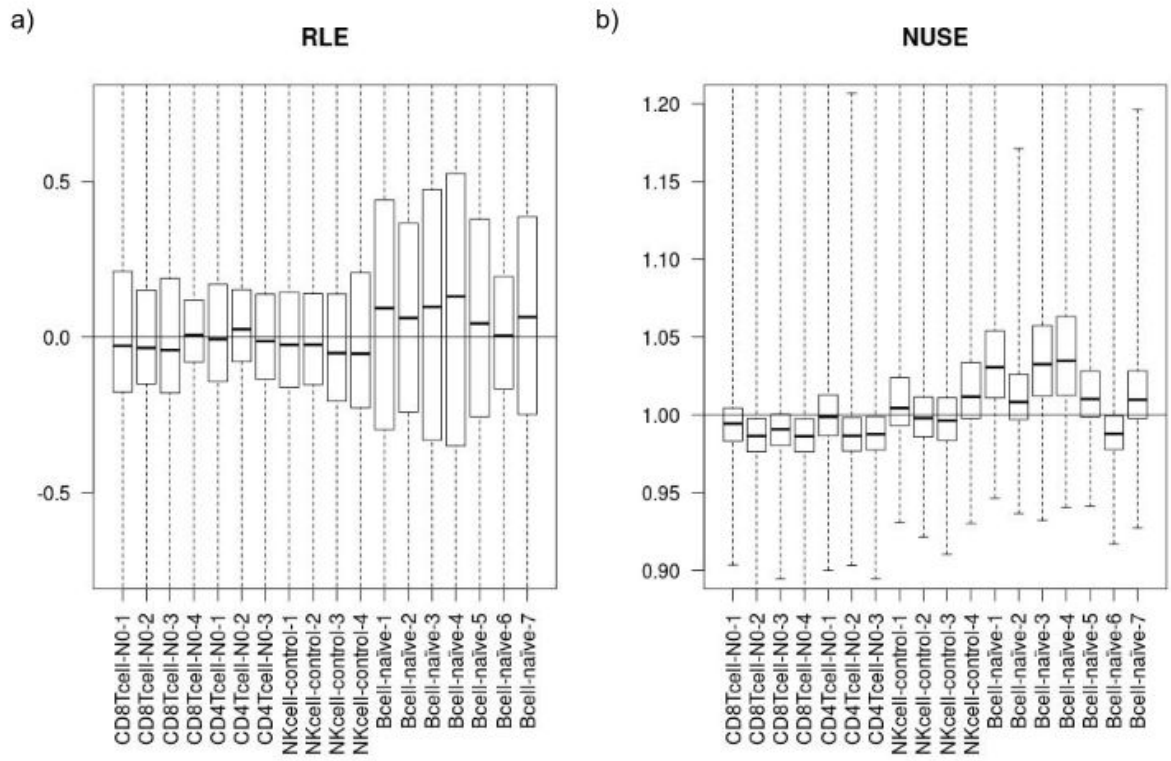


Figura 1. Boxplot dos valores de RLE e NUSE para análise de qualidade dos dados de cada experimento. Como pode ser observado os dados de células B possuem baixa qualidade por apresentar e afastados do valor 1.

Os microarranjo com amostras das células B possuem grande variação interquartil nos valores RLE e valor mediano de NUSE acima do padrão. Os resultados gerados por seus níveis de expressão não são confiáveis e portanto não serão considerados nas análises de detecção de genes. Todos os dados, independentemente do resultado obtido pelas análises com os metodos

RLE e NUSE, foram então pre-processados pelos algorítmos: MAS5, PLIER, RMA e GCRMA para normalização dos dados. Nos gráficos obtidos é possível observar que as amostras com dados de baixa qualidade (Células B) apresentam-se com comportamento diferente ao apresentado pelos outros tipos celulares (fig. 2).

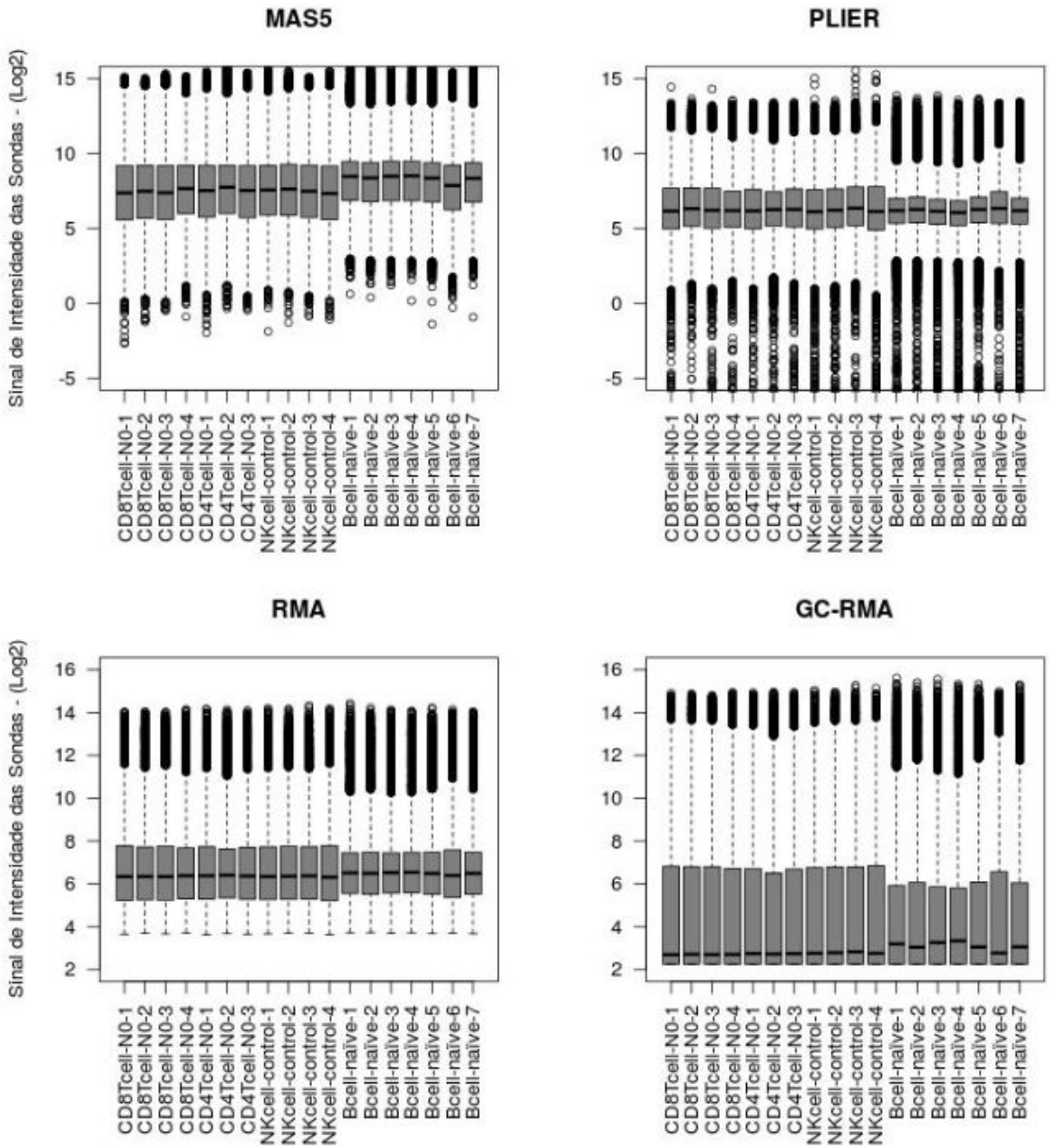


Figura 2. Diferentes algorítimos de pré-processamento: MAS5, PLIER, RMA e GCRMA.

Após obter as matrizes de expressão através dos métodos de pré-processamento, os valores de expressão de proteínas de membrana (*cluster definitions)* foram filtrados e a distribuição de intensidade dos respectivos sinais foi observada (anexo 1). Uma amostra de microarranjo de

célula T foi utilizada para representar a distribuição do sinal de intensidade das sondas CDs filtradas de citometria (fig. 3). Esses dados mostraram que os métodos PLIER e CGRMA seriam os mais recomendados para o tipo de análise aqui proposto.

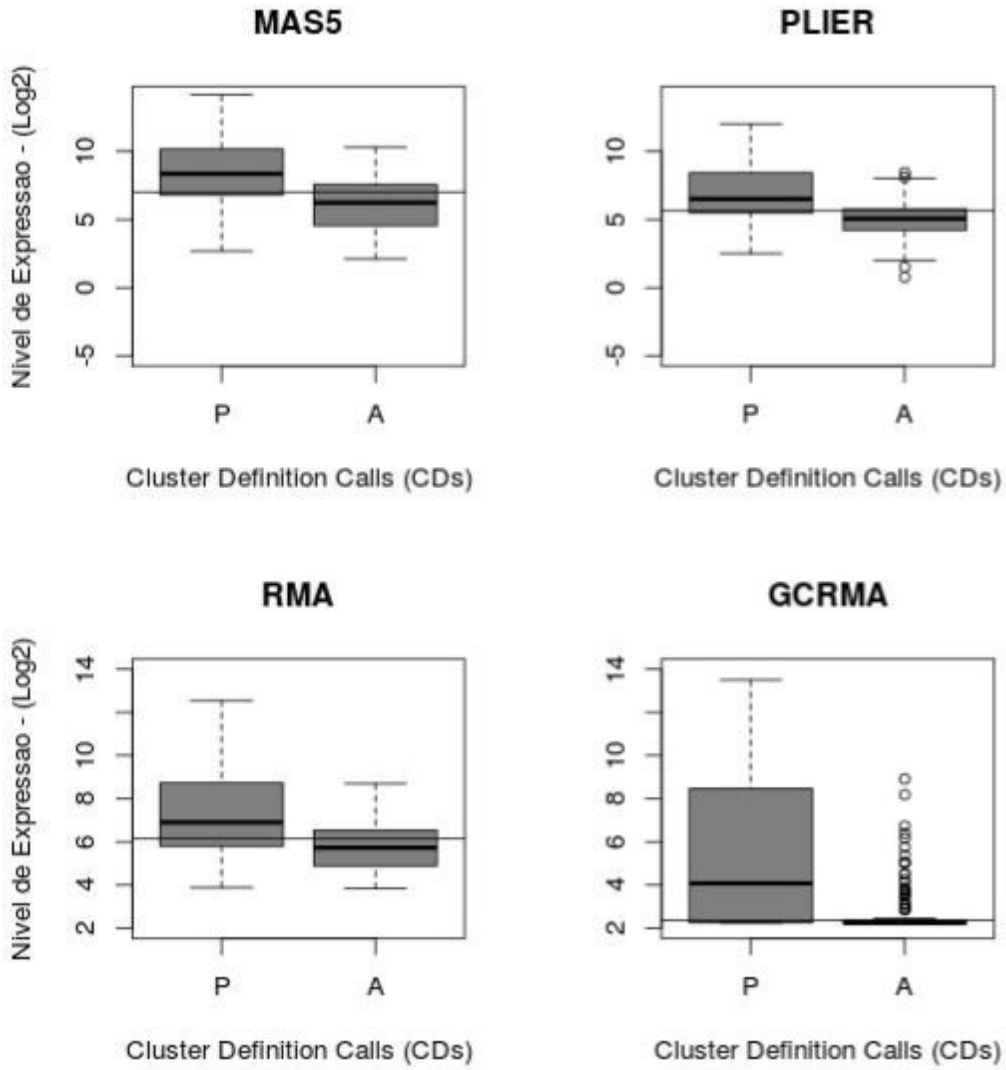


Figura 3. Distribuição da intensidade de sinal de expressão dos transcritos quando comparados com valores qualitativos de presença/ausência da proteina obtidos por citometria de fluxo. P: Sondas de proteínas presentes em citometria A: Sondas de proteínas ausentes em citometria.

## 5.2 Detection Call: PANP e o valor de *cutoff*

Para determinar a presença ou ausência dos genes baseados no valores de expressão, foi necessário observar uma distinção entre os sinais de intensidade das sondas filtradas das proteínas presentes e ausentes em dados de citometria. Um valor de *cut-off* foi determinado

para os resultados de cada algorítmo de pré-processamento analisado. Baseado na média da soma dos valores de expressão do primeiro quartil das sondas presentes com o terceiro quartil das sondas ausentes em citometria. O valor do *cut-off* está representado como um limite nos boxplots da figura 3.

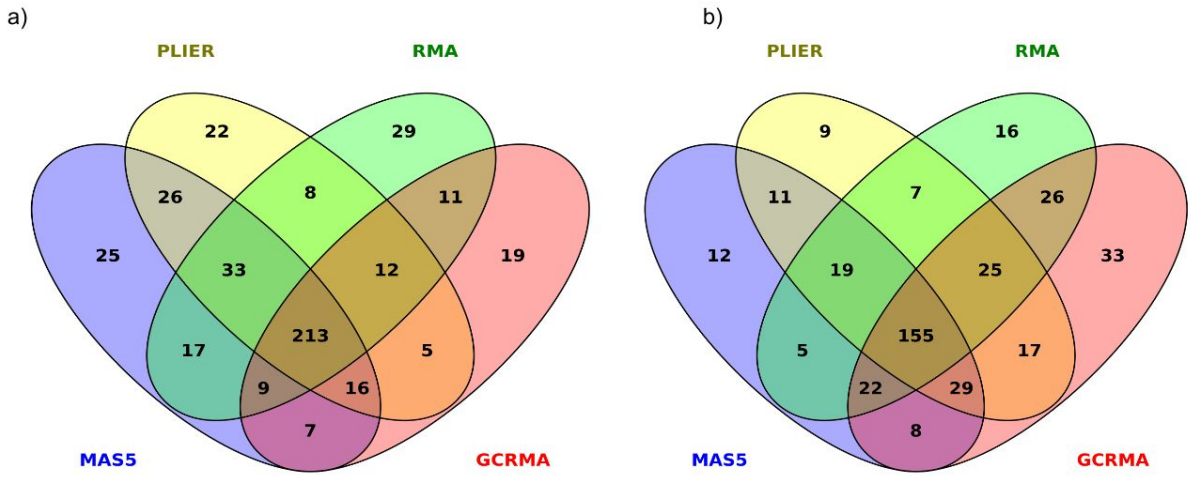


Figura 4. Detection Call baseado em *cutoff* entre os diferente métodos de pré-processamento. a) Presentes. b) Ausentes.

Após identificar os valores de *cutoff* para cada método, foi determinado que valores de intensidade maior ou igual ao *cutoff* foram considerados como presentes e abaixo deste valor como ausentes (fig. 4). Simultaneamente nos analisamos os valores de expressão derivado de cada método de pré-processamento usando o algorítmo de *detection call* PANP (fig. 5).

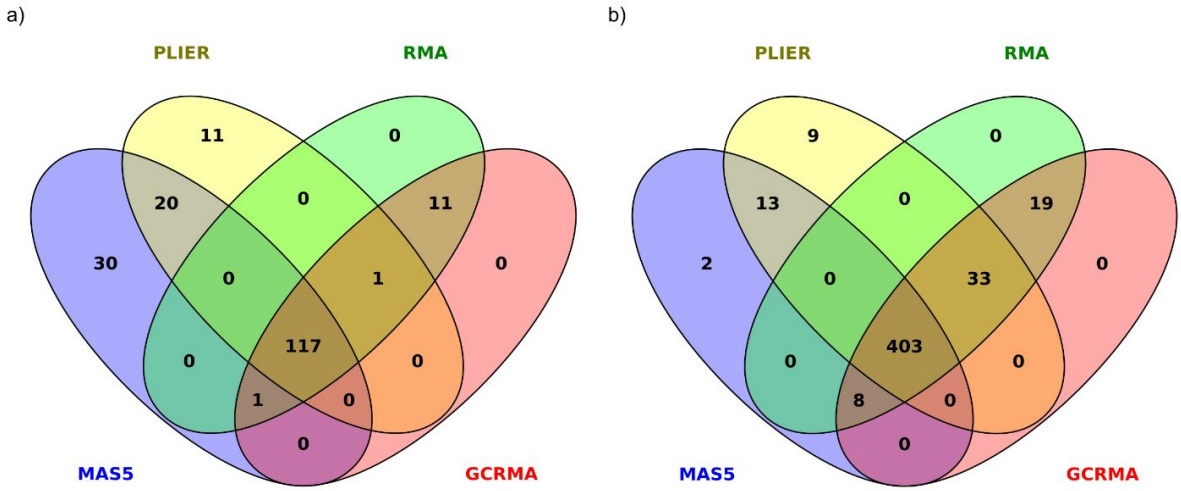


Figura 5. Detection Call baseado em PANP entre os diferente métodos de pré-processamento. a) Presentes. b) Ausentes.

A figura 6 apresenta os resultados dos valores desconhecidos de citometria, comparando os resultados obtidos entre os métodos PANP e de *cutoff*. Houve uma boa concordancia entre os valores desconhecidos de ausência (102), o método PANP apresentou em média 16 classificações como presente e 13 deles foram compartilhados com o *cutoff*. Enquanto que a detecção por *cutoff* classificou 34 valores como presente, as mesma sondas foram definidas como ausente pelo método PANP.

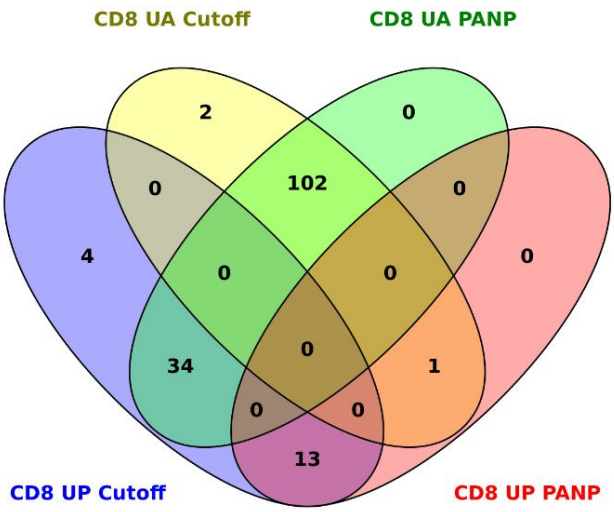


Figura 6. Comparação dos métodos de detecção PANP e *cutoff* obtidos por GCRMA. UA: Ausente desconhecidos; UP: Presentes desconhecidos.

## 5.3 Comparação entre os resultados de *detection call* e os dados de citometria de fluxo em células T

Os resultados de detecção de genes de ambos os métodos foram comparados com os valores qualitativos dos dados de citometria de fluxo de células T CD8 e CD4. O método PANP apresentou as maiores taxas de acerto dos dados ausentes e também uma maior taxa de classificação de dados presentes em citometria de fluxo e ausentes em microarranjo (que nós denominamos PA) (Tabela 1). Biologicamente, a categorização "PA" significa que há a produção de proteína de membrana sem a transcrição do seu gene. Portanto, viola o dogma central da biologia molecular sendo considerado como falso-negativo.

Os resultados baseados no *cutoff* (dados sumarizados na Tabela 2) apresentaram maior taxa de acerto dos dados de presença e uma menor taxa de dados na classificação PA quando comparados com os resultados de PANP, mas houve uma maior taxa de classificação de ausência em citometria e presença em microrranjo (denominado AP), isso pode ser considerado uma maior taxa de falso-positivo para o microarranjo; entretanto, essa classificação não viola o dogma central da biologia molecular, por existir mecanismos

Tabela 1 - Comparação dos dados de citometria com as *detection calls* baseado em PANP

| | CD8 | | | | CD4 |
|---|---|---|---|---|---|
| | MAS | PLIER | RMA | GCRMA | GCRMA |
| PP (320) | 145 | 125 | 112 | 112 | 125 |
| AA (131) | 121 | 127 | 127 | 127 | 121 |
| AP | 7 | 4 | 4 | 4 | 6 |
| PA | 168 | 195 | 200 | 200 | 183 |
| UP | 16 | 20 | 14 | 14 | 15 |
| UA | 137 | 136 | 136 | 136 | 137 |
| NA (156) | 13 | 0 | 14 | 14 | 20 |
| % | | | | | |
| PP% | 45,31% | 39,06% | 35,00% | 35,00% | 39,06% |
| AA% | 92,37% | 96,95% | 96,95% | 96,95% | 92,37% |

Tabela 2 - Comparação dos dados de citometria com as *detection calls* baseado no cutoff

| | CD8 | | | | CD4 |
|---|---|---|---|---|---|
| | MAS | PLIER | RMA | GCRMA | GCRMA |
| PP (320) | 231 | 270 | 257 | 320 | 317 |
| AA (131) | 80 | 96 | 78 | 96 | 90 |
| AP | 51 | 35 | 53 | 35 | 41 |
| PA | 89 | 38 | 45 | 0 | 3 |
| UP | 64 | 73 | 60 | 51 | 53 |
| UA | 92 | 83 | 96 | 105 | 103 |
| NA (156) | 0 | 12 | 18 | 0 | 0 |
| % | | | | | |
| PP% | 72,19% | 84,38% | 80,31% | 100,00% | 99,06% |
| AA% | 61,07% | 73,28% | 59,54% | 73,28% | 68,70% |

conhecidos de regulação gênica pós-transcricional que permite a presença do transcrito sem que esse seja traduzido em proteína ou, ainda, que a proteína seja traduzida, mas mantida em vesiculas sem serem mostradas na membrana até a sua necessidade biológica. O valor de *cut-off* aliado ao método de pré-processamento GCRMA apresentou os melhores resultados, com 100% de acerto dos dados de presença e com uma taxa de classificação PA menor (0% para CD4) que o método PANP e menor AP em comparação aos outros métodos de pré-processamento; entretanto, ainda assim a categoria AP apresentou-se aproximadamente 9 vezes maior que o resultado de PANP.

Por apresentar os melhores resultados, o método gcrma foi escolhido para realizar uma comparação entre os dados de células T CD8 e CD4 (figura 7).

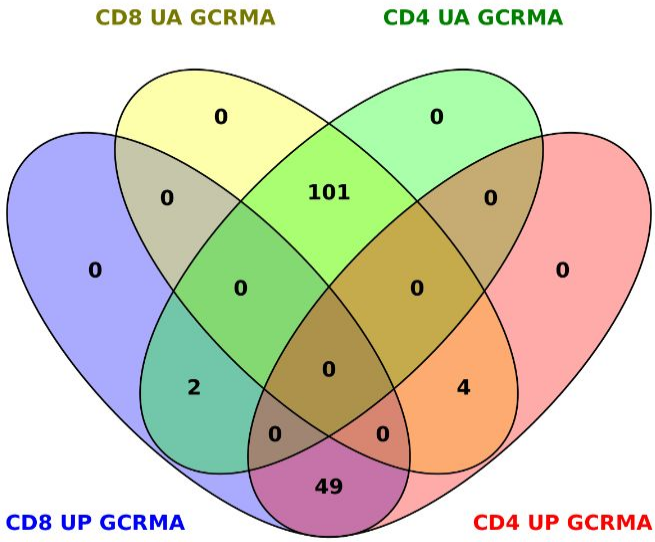


Figura 7. Comparação entre as células CD8 e CD4. UA: Ausente desconhecidos; UP: Presentes desconhecidos.

Cento e um valores classificados como ausentes e quarenta e nove como presentes, previamente desconhecidos por dados de citometria, foram compartilhados entre CD8 e CD4. Duas sondas foram classificadas como ausentes em CD4 e presentes em CD8, e outras quatro sondas apresentaram como presente em CD4 e ausentes em CD8. As últimas seis sondas presentes na tabela 3 são potenciais novos marcadores de membrana.

Tabela 3 - Possíveis CDs inferidos por cutoff-GCRMA

| CDs | CD8 | CD4 |
|---|---|---|
| CD201 | + | - |
| CD248 | + | - |
| CD85e | - | + |
| CD85f | - | + |
| CD158d | - | + |
| CD312 | - | + |

Desta forma nossos dados demonstram que o metodo do *cut-off* baseado no modelo CGRMA aqui proposto comporta-se como o mais confiável entre todos os propostos até o presente momento para determinação do *absolute call*.

# 6. Conclusão

O algorítmo de pré-processamento GCRMA e o valor de *cut-off* baseado na distribuição da intensidade de sinal das respectivas sondas de microarranjo de proteínas presentes e ausentes em superfície de membrana de células do sistema imune humanos determinadas através de citometria de fluxo, apresenta ser um método alternativo para avaliação de presença e ausência de genes de proteína de membrana. O presente método de detecção de gene é superior em relação ao MAS5/P-A por permitir o uso de outros algorítimos de pré-processamento, e ao PANP por apresentar resultados com maior precisão na comparação com dados empíricos de citometria.

# 7. Perspectivas

Aplicar o método proposto para preencher as lacunas nas tabelas de CDs de células hematopoiéticas visando ampliar os perfis de proteínas de membrana para diagnóstico de doenças hematológicas.

# 8. Referencias


SU, A. I. et al. Large-scale analysis of the human and mouse transcriptomes. **Proceedings of the National Academy of Sciences USA**, v. 99, n. 7, p. 4465–4470, 2002.

BARRETT, T. et al. NCBI GEO: archive for functional genomics data sets--update. **Nucleic Acids Research**, v. 41, n. D1, p. D991–D995, 2013.

BESSANT, C.; CONRAD, B. Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data Darius M. Dziuda John Wiley and Sons, 2010, p. 327 ISBN: 978-0-470-16373-3. **Proteomics**, v. 11, n. 18, p. 3768–3768, 2011.

DZIUDA, D. M. **Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data**. [s.l.] John Wiley & Sons, 2010.

GAUTIER, L. et al. affy--analysis of Affymetrix GeneChip data at the probe level. **Bioinformatics** , v. 20, n. 3, p. 307–315, 12 fev. 2004.

TALLOEN, W.; GÖHLMANN, H. Gene Expression Studies Using Affymetrix Microarrays. **Biomedical Engineering**, v. 23, n. 2009, p. 2010–2010, 2009.

AFFYMETRIX, I. Affymetrix Expression Console Software, v1.4 — User Guide. **Flying**, 2001.

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

IRIZARRY, R. A.; WU, Z.; JAFFEE, H. A. Comparison of Affymetrix GeneChip expression measures. **Bioinformatics** , v. 22, n. 7, p. 789–794, 1 abr. 2006.

LI, C.; WONG, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 1, p. 31–36, 2 jan. 2001.

LUO, J. et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. **The pharmacogenomics**

**journal**, v. 10, n. 4, p. 278–291, ago. 2010.

MILLENAAR, F. F. et al. How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. **BMC bioinformatics**, v. 7, p. 137, 15 mar. 2006.

RAMASAMY, A. et al. Key issues in conducting a meta-analysis of gene expression microarray datasets. **PLoS medicine**, v. 5, n. 9, p. e184, 30 set. 2008.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome biology**, v. 5, n. 10, p. R80, jan. 2004.

R CORE TEAM, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2015.

GAUTIER, L., COPE, L., BOLSTAD, B. M., and IRIZARRY, R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. **Bioinformatics** 20, 3 (Feb. 2004), 307-315.

JEAN, W., IRIZARRY, R., with contributions from James MacDonald Jeff Gentry (). gcrma: Background Adjustment Using Sequence Information. **R package** version 2.42.0.

AFFYMETRIX Inc., MILLER, C.J., and PICR (). plier: Implements the Affymetrix PLIER algorithm. **R package** version 1.40.0.

CANALES, R. D. et al. Evaluation of DNA microarray results with quantitative gene expression platforms. **Nature Biotechnology**, v. 24, n. 9, p. 1115–1122, 2006.

TALLOEN, W.; GÖHLMANN, H. Gene Expression Studies Using Affymetrix Microarrays. **Biomedical Engineering**, v. 23, n. 2009, p. 2010–2010, 2009.

BRETTSCHENEIDER, J.; COLIN, F., BOLSTAD, B.M., SPEED, T.P,. "Quality assessment for short oligonucleotide arrays". **Technometrics**. (2007)

WARREN, P. et al. PANP - A new method of gene detection on oligonucleotide expression arrays Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE. Anais...2007

OLIVEROS, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html

ZOLA, H.; SWART, B. The human leucocyte differentiation antigens (HLDA) workshops: the evolving role of antibodies in research, diagnosis and therapy. Cell research, v. 15, n. 9, p. 691–4, 2005.

BREM, H. et al. Molecular Markers in Patients with Chronic Wounds Guide to Surgical Debridement. Molecular medicine Cambridge Mass, v. 13, n. 9, p. 30–39, 2007.

# ANEXOS

## Anexo 1. Script utilizado e funções desenvolvidas

https://www.dropbox.com/s/f9eef0sdacotip8/tcc.R?dl=0

```
typecell = function(j){ #extrai os nomes das amostras por categorias celular
  as.vector(rownames(control)[control$type == colnames(cytoflow)[j]])
}
extractvexp = function(a,j, method){ #extrai os valores de expressão por categoria celular; a = 'P' ou 'A'
  if(method == 1){log2(exp.cdsmas5[which(cytoflow[j] == a), typecell(j)])}
  else {if(method == 2){exp.cdsplier[(cytoflow[j] == a), typecell(j)]}
    else{if(method == 3){exp.cdsrma[(cytoflow[j] == a), typecell(j)]}
      else{if(method == 4){exp.cdsgcrma[(cytoflow[j] == a), typecell(j)]}}}}
}
extractpanpcall = function(i,j, method){ #extrai as absoluts calls por categoria celular, baseado em PANP
  if(method == 1){cds.mas5[i, typecell(j)]}
  else {if(method == 2){cds.plier[i, typecell(j)]}
    else{if(method == 3){cds.rma[i, typecell(j)]}
      else{if(method == 4){cds.gcrma[i, typecell(j)]}}}}
}
makecall = function(b,s,method){ #funcao detection call baseado em cutoff, b = valor escolhido
  if(method == 1){ifelse(log2(exp.cdsmas5[,s]) >= b, 'P', 'A')}
  else {if(method == 2){ifelse(exp.cdsplier[,s] >= b, 'P', 'A')}
    else{if(method == 3){ifelse(exp.cdsrma[,s] >= b, 'P', 'A')}
      else{if(method == 4){ifelse(exp.cdsgcrma[,s] >= b, 'P', 'A')}}}}
}

tcc1 <- function(j,s,method){
  as.data.frame(ifelse(cytoflow[,j] == 'NA' & extractpanpcall(,j,method)[,s] == 'P', 'UP',
        ifelse(cytoflow[,j] == 'NA' & extractpanpcall(,j,method)[,s] == 'A', 'UA',
              ifelse(cytoflow[,j] == 'P' & extractpanpcall(,j,method)[,s] == 'P', 'PP',
                    ifelse(cytoflow[,j] == 'A' & extractpanpcall(,j,method)[,s] == 'A', 'AA',
                          ifelse(cytoflow[,j] == 'A' & extractpanpcall(,j,method)[,s] == 'P', 'AP',
                                ifelse(cytoflow[,j] == 'P' & extractpanpcall(,j,method)[,s] == 'A', 'PA', 'NA')))))))
}
tcc2 <- function(j,b,s,method){
```

```
as.data.frame(ifelse(cytoflow[,j] == 'NA' & makecall(b,s,method) == 'P', 'UP',
      ifelse(cytoflow[,j] == 'NA' & makecall(b,s,method) == 'A', 'UA',
           ifelse(cytoflow[,j] == 'P' & makecall(b,s,j) == 'P', 'PP',
                ifelse(cytoflow[,j] == 'A' & makecall(b,s,method) == 'A', 'AA',
                     ifelse(cytoflow[,j] == 'A' & makecall(b,s,method) == 'P', 'AP',
                          ifelse(cytoflow[,j] == 'P' & makecall(b,s,method) == 'A', 'PA', 'NA')))))))
```

## Anexo 2. Sondas de CDS inferidas através do método de detecção *cutoff*-GCRMA

CD8 Presentes

| | | | | | |
|---|---|---|---|---|---|
| 204007_at | 205745_x_at | 217240_at | 204832_s_at | 211398_at | 216836_s_at |
| 201646_at | 213532_at | 214088_s_at | 213578_at | 211399_at | |
| 215754_at | 205179_s_at | 219669_at | 203799_at | 222006_at | |
| 202351_at | 205180_s_at | 203650_at | 218451_at | 204579_at | |
| 204628_s_at | 202603_at | 207278_s_at | 218529_at | 217045_x_at | |
| 216261_at | 202604_x_at | 202947_s_at | 207822_at | 217493_x_at | |
| 206049_at | 214895_s_at | 219025_at | 210973_s_at | 210763_x_at | |
| 201998_at | 216676_x_at | 37408_at | 211535_s_at | 211010_s_at | |
| 214971_s_at | 1007_s_at | 209280_at | 215404_x_at | 211583_x_at | |
| 214444_s_at | 210749_x_at | 210176_at | 222164_at | 209097_s_at | |

CD8 Ausentes

| | | | | | |
|---|---|---|---|---|---|
| 208592_s_at | 205455_at | 208422_at | 207426_s_at | 203440_at | 209098_s_at |
| 215784_at | 32699_s_at | 208423_s_at | 207037_at | 203441_s_at | 209099_x_at |
| 201819_at | 212662_at | 211887_x_at | 218368_s_at | 210796_x_at | 216268_s_at |
| 215834_x_at | 214443_at | 214770_at | 206641_at | 207937_x_at | 210930_s_at |

| | | | | | |
|---|---|---|---|---|---|
| 215835_at | 216283_s_at | 220428_at | 205858_at | 203638_s_at | 219764_at |
| 201647_s_at | 205746_s_at | 205569_at | 204924_at | 203639_s_at | |
| 204625_s_at | 205715_at | 207277_at | 221060_s_at | 208225_at | |
| 204626_s_at | 208426_x_at | 206172_at | 207446_at | 208228_s_at | |
| 204627_s_at | 211242_x_at | 220043_s_at | 210523_at | 208229_at | |
| 211579_at | 211245_x_at | 202242_at | 207995_s_at | 208234_x_at | |
| 215240_at | 207169_x_at | 207459_x_at | 210481_s_at | 211400_at | |
| 206211_at | 208779_x_at | 214407_x_at | 206682_at | 211401_s_at | |
| 215925_s_at | 206934_at | 216398_at | 210510_s_at | 204379_s_at | |
| 214970_s_at | 216010_x_at | 206077_at | 210615_at | 204380_s_at | |
| 206881_s_at | 221349_at | 216317_x_at | 212298_at | 211237_s_at | |
| 215838_at | 206660_at | 210429_at | 203934_at | 207860_at | |
| 215839_at | 206206_at | 210430_x_at | 207610_s_at | 217088_s_at | |
| 213325_at | 206702_at | 210586_x_at | 219213_at | 217095_x_at | |
| 222167_at | 217711_at | 204844_at | 201130_s_at | 221074_at | |
| 205876_at | 219912_s_at | 204845_s_at | 201131_s_at | 221075_s_at | |

CD4 Presentes

| | | | | | |
|---|---|---|---|---|---|
| 204007_at | 215838_at | 216676_x_at | 210176_at | 215404_x_at | 211583_x_at |
| 201646_at | 214444_s_at | 1007_s_at | 204832_s_at | 222164_at | 209097_s_at |
| 215754_at | 205745_x_at | 210749_x_at | 213578_at | 211398_at | 216836_s_at |
| 202351_at | 213532_at | 217240_at | 203799_at | 211399_at | |

| | | | | |
|---|---|---|---|---|
| 204628_s_at | 205179_s_at | 214088_s_at | 207610_s_at | 222006_at |
| 216261_at | 205180_s_at | 219669_at | 218451_at | 204579_at |
| 206049_at | 202603_at | 207278_s_at | 218529_at | 217045_x_at |
| 201998_at | 202604_x_at | 202947_s_at | 207822_at | 217493_x_at |
| 214971_s_at | 214895_s_at | 37408_at | 210973_s_at | 210763_x_at |
| 206881_s_at | 208426_x_at | 209280_at | 211535_s_at | 211010_s_at |

CD4 Ausentes

| | | | | | |
|---|---|---|---|---|---|
| 208592_s_at | 212662_at | 211887_x_at | 207037_at | 210796_x_at | 216268_s_at |
| 215784_at | 214443_at | 214770_at | 218368_s_at | 207937_x_at | 210930_s_at |
| 201819_at | 216283_s_at | 220428_at | 206641_at | 203638_s_at | 219764_at |
| 215834_x_at | 205746_s_at | 205569_at | 205858_at | 203639_s_at | |
| 215835_at | 205715_at | 207277_at | 204924_at | 208225_at | |
| 201647_s_at | 211242_x_at | 206172_at | 221060_s_at | 208228_s_at | |
| 204625_s_at | 211245_x_at | 220043_s_at | 207446_at | 208229_at | |
| 204626_s_at | 207169_x_at | 202242_at | 210523_at | 208234_x_at | |
| 204627_s_at | 208779_x_at | 207459_x_at | 207995_s_at | 211400_at | |
| 211579_at | 206934_at | 214407_x_at | 210481_s_at | 211401_s_at | |
| 215240_at | 216010_x_at | 216398_at | 206682_at | 204379_s_at | |
| 206211_at | 221349_at | 206077_at | 210510_s_at | 204380_s_at | |
| 215925_s_at | 206660_at | 216317_x_at | 210615_at | 211237_s_at | |
| 214970_s_at | 206206_at | 210429_at | 212298_at | 207860_at | |

| | | | | |
|---|---|---|---|---|
| 215839_at | 203650_at | 210430_x_at | 203934_at | 217088_s_at |
| 213325_at | 206702_at | 210586_x_at | 219213_at | 217095_x_at |
| 222167_at | 217711_at | 219025_at | 201130_s_at | 221074_at |
| 205876_at | 219912_s_at | 204844_at | 201131_s_at | 221075_s_at |
| 205455_at | 208422_at | 204845_s_at | 203440_at | 209098_s_at |
| 32699_s_at | 208423_s_at | 207426_s_at | 203441_s_at | 209099_x_at |