Daniele Yumi Sunaga de Oliveira

**Biologia computacional aplicada para a
análise de dados em larga escala**

Computational biology for
high-throughput data analysis

São Paulo
2013

Daniele Yumi Sunaga de Oliveira

**Biologia computacional aplicada para a
análise de dados em larga escala**

Computational biology for
high-throughput data analysis

Tese apresentada ao Instituto de
Biociências da Universidade de São
Paulo, para a obtenção de Título de
Doutor em Ciências, na Área de
Biologia/Genética.

Orientadora: Profa. Dra. Maria Rita
Passos-Bueno
Co-orientador: Prof. Dr. Ronaldo
Hashimoto

São Paulo
2013

Ficha catalográfica

Sunaga de Oliveira, Daniele Yumi
Biologia computacional
aplicada para a análise de dados em larga escala
138 páginas

Tese (Doutorado) - Instituto de Biociências da
Universidade de São Paulo. Departamento de Genética e
Biologia Evolutiva.

1. dados em larga escala 2. expressão gênica 3.
sequências genômicas I. Universidade de São Paulo.
Instituto de Biociências. Departamento de Genética e
Biologia Evolutiva.

Comissão Julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Profa. Dra. Maria Rita Passos-Bueno
Orientadora

Agradecimentos

Meu agradecimento especial é para a professora Maria Rita, que é um exemplo de profissionalismo e competência, com quem sempre pude contar para tirar dúvidas, discutir ideias e sempre me deu oportunidades para enfrentar novos desafios, muito obrigada!

Também agradeço a todos os colegas de laboratório que sempre me ajudaram com as minhas questões de biologia e sempre souberam manter um clima de descontração e competência, fazendo do nosso lab um ótimo ambiente de trabalho. À Daniela Bueno e ao Gerson Kobayashi, em especial, com quem compartilhei a primeira publicação na área e ótimos momentos. À Karina Griesi por sempre me ajudar com as questões gerais da pós. À Cibele Masotti que teve participação especial na definição da minha qualificação e tese, sempre muito querida e prestativa. À Erika Takamoto, Shirlene Barros e Deisy Santos de Moraes, da Secretaria de Pós-graduação e do Departamento, que sempre responderam com muita atenção às minhas dúvidas com relação às normas e regimentos do programa.

Ao professor Ronaldo Hashimoto pelas sugestões, correções da tese e pelas suas ótimas aulas. A todos os meus colegas do IME, em especial à Paola Gaviria, que se tornou uma grande amiga.

À minha grande amiga Ana Paula, Sandro e Branquinha que sempre me acolheram tão carinhosamente em sua casa todas as vezes que fui a São Paulo. À Renata Pellegrino que foi uma das grandes motivadoras para eu iniciar os trabalhos com *microarrays*.

Agradeço muito ao meu pai, mãe, Sô e Guizinho que sempre apoiaram e incentivaram as minhas decisões. Dedico um agradecimento super especial ao Léo, meu marido querido, que sempre está do meu lado e foi o meu grande suporte e guru da parte computacional.

Ao amigo Misha Sukchev, que foi muito importante durante o tempo em que passei no Max Planck e à Shini, que foi quem me ensinou todo o projeto do departamento e sempre estava disposta a me ajudar dentro e fora do Instituto.

À FAPESP pelo apoio financeiro que tornou possível a realização deste trabalho.

Índice

Agradecimentos	i
Introdução geral e objetivos	1

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o método Rank Products

Resumo	5
1.1 Introdução	6
1.1.1 Transcriptoma de doenças complexas	6
1.1.2 Rank Products	9
1.1.3 SAM (<i>Significance Analysis of Microarrays</i>)	10
1.2 Motivação	11
1.3 Materiais e métodos.....	13
1.3.1 Enfoque computacional	13
1.3.1.1 Simulação dos dados.....	13
1.3.1.2 Teste de heterogeneidade.....	17
1.3.1.3 Teste de sensibilidade	18
1.3.1.4 Controle de qualidade	18
1.3.1.5 Pré-processamento	19
1.3.2 Enfoque biológico.....	19
1.3.2.1 Fissura Lábio-Palatina Não-Sindrômica (FLP/NS).....	19
1.3.2.2 Autismo.....	19
1.3.2.3 Sono	20
1.4 Resultados e discussões	20
1.4.1 Enfoque computacional	20
1.4.1.1 Teste de heterogeneidade.....	20
1.4.1.2 Teste de sensibilidade	22
1.4.2 Enfoque biológico.....	29
1.4.2.1 Fissura Lábio-Palatina Não-Sindrômica (FLP/NS).....	29
1.4.2.2 Autismo.....	30
1.4.2.3 Sono	30
1.5 Conclusões.....	31
1.6 Referências bibliográficas.....	31
1.7 Anexos	34

Capítulo II

Busca por genes alvos do T com o programa $hunT$

Resumo	79
2.1 Introdução	80

Índice

2.1.1	Sítios de ligação na regulação da expressão gênica	80
2.2	Motivação	84
2.3	Materiais e métodos.....	88
2.3.1	Funcionamento do <i>hunT</i>	88
2.3.2	Comparação com programa similar	96
2.3.3	<i>hunT</i> para a busca de alvos de diferentes FTs	98
2.3.4	<i>hunT</i> para a busca de genes alvos do <i>T</i>	98
2.3.5	Obtenção da sequência genômica	100
2.4	Resultados e discussões	103
2.4.1	Comparação com programa similar	103
2.4.2	Busca por alvos de diferentes FTs	104
2.4.3	Busca por alvos do <i>T</i>	105
2.4.4	Expressão gênica dos alvos do <i>T</i>	106
2.4.5	Considerações e perspectivas futuras	110
2.5	Conclusões.....	110
2.6	Referências bibliográficas.....	111
2.7	Anexo	114
Conclusão geral		131
Resumo geral		132
Abstract		133
Referências Bibliográficas		134
Produção científica durante o doutorado		135

Introdução geral e objetivos

Pela primeira vez, em centenas de anos de pesquisas nas áreas biológica e médica, o principal gargalo para o progresso científico está na coleção de dados. Tecnologias como a de sequenciamento e de *microarray*, por exemplo, são as responsáveis por um novo cenário onde dados têm sido produzidos em larga escala, de forma rápida e custo cada vez mais acessível. O resultado são milhares de genes com níveis de expressão medidos em diferentes condições experimentais, o sequenciamento de genomas completos e a classificação de milhares de SNPs¹ em amostras individuais, o que tem inspirado a criação de neologismos como "ORFeoma", "transcriptoma", "proteoma", "metaboloma", "exoma" e o mais recente "incidentaloma", entre muitos outros que descrevem um conjunto completo de dados biológicos. A preocupação por trás desta nova realidade está na capacidade de armazenamento, manipulação e extração de conhecimento dessa enorme quantidade de dados.

Felizmente para a computação, a capacidade de armazenar e processar estes dados não é um fator limitante. A biologia pode aproveitar os avanços de áreas como climatologia, telecomunicações e internet, que processam rapidamente grandes quantidades de dados diariamente. A ferramenta de busca da Google é um bom exemplo disso. Ela varre uma extensa quantidade de dados em busca de uma palavra chave e retorna o resultado em segundos. Em 2008, ela processou mais de 20 petabytes² de dados por dia (Dean & Ghemawat, 2008).

Processamentos desta proporção, contudo, requerem uma infra-estrutura de computadores compatível, que tipicamente não é encontrada nos pequenos centros de pesquisa. O investimento em um aglomerado de computadores (do inglês, *cluster computing*) envolve custos para a aquisição e hospedagem das máquinas, assim como para o gerenciamento e manutenção. Uma alternativa viável que tem se tornado cada vez mais acessível é chamada computação em nuvem (do inglês, *cloud computing*), que consiste na disponibilização de serviços virtuais pela internet. Processamentos complexos podem ser feitos em servidores localizados em algum lugar do mundo (por isso o termo nuvem), que podem ser acessados a partir de um computador comum com bom acesso à internet. O Dropbox³, que disponibiliza serviços de sincronização de arquivos, é um bom exemplo de uso desta tecnologia. Na biologia estes assuntos estão também cada

¹ Polimorfismos de nucleotídeo único, do inglês *Single Nucleotide Polymorphism*.

² 1 petabyte corresponde à 1.000 terabytes e 1 terabyte corresponde à 1.000 gigabytes.

³ <https://www.dropbox.com/>

Introdução geral e objetivos

vez mais presentes. Técnicas avançadas da computação têm sido empregadas para o alinhamento de sequências (Matsunaga *et al.*, 2008), para a análise de dados de expressão gênica de microarray (Zhang *et al.*, 2012), de RNA-seq (Langmead *et al.*, 2010), para a análise de dados de genotipagem (Gurtowski *et al.*, 2012) e de sequenciamento de nova geração (McKenna *et al.*, 2010).

O principal desafio da produção de dados em larga escala fica por conta da nossa habilidade de minerar informação biologicamente relevante destes dados, seja através do desenvolvimento de novos métodos de análise ou do uso de métodos existentes.

Neste trabalho, nós mostramos 2 aplicações de bioinformática para a análise de dados de duas das principais fontes em larga escala: *microarrays* e sequenciamento. Em uma delas, usamos um método existente e, na outra, desenvolvemos uma nova ferramenta de análise. Como cada uma destas aplicações possui metodologia e resultados próprios, elas foram organizadas em capítulos distintos para facilitar a leitura. Uma introdução mais específica de cada uma delas também é apresentada nos respectivos capítulos.

No capítulo I, avaliamos o desempenho do método de seleção de genes diferencialmente expressos Rank Products (Breitling *et al.*, 2004) para o estudo de doenças complexas, que é um dos focos de pesquisa do grupo da Professora Maria Rita Passos Bueno. Os objetivos deste capítulo são apresentados a seguir:

- (1) Avaliar se as características do Rank Products fazem dele uma opção adequada para lidar com a heterogeneidade genética de indivíduos com uma mesma doença complexa;
- (2) Mostrar a importância de conhecer as características e limitações dos métodos de análise existentes para escolher o mais adequado deles e poder aplicá-lo corretamente para conseguir extrair resultados de maior confiabilidade.

No capítulo II, apresentamos um novo programa chamado *hunT*, que foi desenvolvido para buscar genes alvos do fator de transcrição *T* - que é um marcador de mesoderma com papel fundamental no desenvolvimento de vertebrados -, através da identificação de sítios de ligação para este fator em suas sequências reguladoras. Os objetivos deste capítulo são:

- (1) Desenvolver uma ferramenta adaptada ao conhecimento existente sobre o modo como o *T* atua para regular seus alvos;

Introdução geral e objetivos

(2) Identificar genes alvos do T e ajudar na compreensão das redes regulatórias por trás da formação do mesoderma e do importante papel do T no desenvolvimento de vertebrados.

Uma página *web* foi criada para disponibilizar todo o material suplementar:

<http://danieleyumi.sunaga.de/>.

Capítulo I

Busca por genes diferencialmente expressos em
doenças complexas com o método Rank Products

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Resumo

A análise de dados de transcriptoma envolve uma série de etapas, incluindo a de seleção de genes com expressão diferenciada entre as condições avaliadas. Existem vários métodos para esta seleção. O emprego de um método inadequado pode ter um impacto dramático no resultados, no entanto, não é raro que pesquisadores escolham o mais popular deles ou aquele de um trabalho bem sucedido. Neste capítulo, nós avaliamos o método Rank Products (RP), que é baseado em *ranking* e permutações. Nossa hipótese é que estas características fazem dele uma solução interessante para o estudo de doenças complexas, onde não é esperado que todos os indivíduos apresentem uma etiologia comum. O método trata cada amostra de forma independente, o que permite a identificação de genes desregulados em somente um subgrupo de indivíduos. Para testar esta hipótese, avaliamos o desempenho do RP sob um enfoque computacional e biológico. No primeiro, testamos a habilidade do método para lidar com perfil heterogêneo de expressão entre as amostras. Também avaliamos sua sensibilidade a diferentes tamanhos amostrais e número de genes analisados. Os testes foram feitos usando dados parcialmente simulados de um estudo de asma (E-GEOD-8052) e outro de leucemia (E-GEOD-11877). Todos os testes foram também rodados com o popular método SAM. Para avaliar o RP sob um enfoque biológico, o empregamos em um estudo de Fissura Lábio-Palatina Não-Sindrômica (FLP/NS), de autismo e também num estudo que avalia o efeito da privação do sono em humanos. Os testes com dados simulados revelaram notável habilidade do RP para detectar genes consistentemente alterados em apenas um subgrupo de amostras. Também mostraram que esta habilidade é mantida com poucas amostras, mas que seu desempenho é prejudicado quando são analisados poucos genes. Nos estudos de FLP/NS, autismo e sono, obtivemos fortes evidências biológicas da eficiência do método através da identificação de genes e vias previamente associados à estas doenças e da validação de novos genes candidatos através da técnica de PCR quantitativo em tempo real (qRT-PCR). Juntos, estes resultados confirmam a nossa hipótese de que o RP é uma opção interessante para o estudo de doenças complexas.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

1.1 Introdução

Neste capítulo avaliamos o desempenho do método RankProd de seleção de genes diferencialmente expressos para o estudo de doenças complexas. Inicialmente apresentamos uma breve introdução de conceitos importantes associados ao tema e, em seguida, apresentamos a motivação deste trabalho, a metodologia, resultados e discussões e, por último, as conclusões.

1.1.1 Transcriptoma de doenças complexas

A busca por genes de predisposição à doenças complexas é um grande desafio. A variedade de fatores (genéticos e não-genéticos) envolvidos nestas etiologias e a complexidade das interações que pode haver entre estes fatores não podem ser explicadas por técnicas reducionistas que estudam gene a gene. Os *microarrays* representam um grande avanço metodológico para as pesquisas dessas doenças pois permitem o monitoramento da variação de transcriptomas inteiros. A comparação do padrão de expressão global de indivíduos sadios e afetados pode levar à identificação de múltiplos genes, o que representa um passo importante para a compreensão de vias biológicas críticas para a manifestação dessas doenças.

Existem atualmente várias plataformas de *microarrays*. Os dados utilizados neste capítulo foram todos produzidos com o GeneChip da Affymetrix⁴. A principal característica que o distingue das outras tecnologias é que ele é confeccionado pela síntese direta de oligonucleotídeos⁵ *in situ* através da combinação de um semicondutor de fotolitografia e técnicas de síntese química. Em geral, o termo "sonda" é usado para descrever a sequência de oligonucleotídeo fixada no *chip* e "alvo" refere-se à sequência que se liga às sondas. O resultado, é que mais de um milhão de diferentes sondas podem ser sintetizadas em um *array*, permitindo que múltiplas delas sejam usadas para interrogar a mesma sequência alvo, oferecendo maior robustez para a análise estatística dos dados (Dalma-Weiszhausz *et al.*, 2006).

A análise de dados de *microarrays* envolve uma série de etapas que vão desde o controle da qualidade dos *arrays* até a seleção de genes diferencialmente expressos e a posterior interpretação da relevância biológica dos genes selecionados. Elas podem ser vistas em mais detalhes no trabalho de Slonin &

⁴ www.affymetrix.com/

⁵ Moléculas formadas por uma cadeia de 25 nucleotídeos que são chamadas de sondas.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Yanai, 2009.

Este trabalho refere-se à etapa de seleção de genes diferencialmente expressos. A entrada para os métodos de seleção segue o formato da Figura 1.1.

<i>n arrays, m genes</i>						
	array 1	array 2	array 3	array 4	...	array <i>n</i>
gene 1	1.23	2.61	3.57	4.22	...	5.12
gene 2	3.98	2.94	1.73	2.97	...	2.43
⋮			⋮			⋮
gene <i>m</i>	2.46	3.72	1.83	1.10	...	2.94

Figura 1.1: Formato de entrada para os métodos de seleção de genes diferencialmente expressos. Os dados de todos os *arrays* são concentrados numa única matriz. As linhas correspondem aos genes (*m*) e as colunas correspondem às amostras (*n*). Os valores correspondem à intensidade da expressão de cada gene em cada amostra. Os valores de expressão de um gene em todas as amostras é chamado de perfil de expressão.

O que estes métodos fazem é buscar por genes, dentre milhares, que apresentam alteração de expressão estatisticamente significativa entre as diferentes condições avaliadas (indivíduos sadios e afetados, por exemplo). Isto é feito, tipicamente, gene-por-gene separadamente. Um gene é considerado significativo quando seu nível de significância (o conhecido valor de *p*, do inglês *p-value*) é menor que um valor de corte definido pelo usuário.

O problema é que estes métodos ignoram a dependência onipresente entre os genes. Genes trabalham em conjunto (como nas vias de sinalização, por exemplo). Seus níveis de expressão claramente não são independentes. Um assunto importante, neste sentido, é o problema dos testes múltiplos⁶. Por causa da abordagem gene-por-gene, o número de testes realizados é o mesmo do número de genes analisados, o que geralmente resulta numa grande quantidade

⁶ Refere-se ao problema de ter um resultado com grande número de falsos positivos porque a mesma hipótese foi testada múltiplas vezes.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

de falsos positivos⁷. Por exemplo, numa análise de 10.000 genes com $p < 0,05$, a expectativa é que cerca de 500 deles (5%) sejam selecionados ao acaso.

Para lidar com este problema, os métodos de seleção de genes geralmente incorporam a opção de correção para testes múltiplos, cujo objetivo é ajustar os valores de p para quantificar e corrigir esta ocorrência de falsos positivos. Bonferroni (Lin, 2005) e FDR (*False Discovery Rate*, Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) são algumas das opções mais utilizadas. Uma alternativa que pode ser útil nesta etapa é reduzir a quantidade de genes que serão analisados pelos métodos de seleção, o que é comumente feito através da aplicação de filtros de dados. Numa análise de 100 genes com $p < 0,05$, por exemplo, seriam esperados somente 5 falsos positivos.

Nos estudos de doenças complexas, a identificação de genes diferencialmente expressos é ainda dificultada pelo fato de que seus níveis de expressão podem apresentar alta variabilidade entre os pacientes estudados. Um gene pode, por exemplo, estar aumentado em apenas 3 de 8 dos afetados (Figura 1.2). Isto acontece devido a riscos individuais, que dependem de funções desconhecidas da genética, heterogeneidade genética, fatores ambientais e estocásticos (Pritchard & Cox, 2002).

Figura 1.2: Variabilidade da expressão gênica entre pacientes com uma mesma doença complexa. O eixo x corresponde às amostras analisadas (pacientes) e o y ao nível de expressão.

⁷ Genes que foram selecionados como estatisticamente significantes quando na realidade eles não são.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Existem vários métodos desenvolvidos para a seleção de genes diferencialmente expressos (Tusher *et al.*, 2001; Efron *et al.*, 2001; Newton *et al.*, 2004; Smyth *et al.*, 2004; Breitling *et al.*, 2004). Conhecer as características, limites e deficiências de vários deles é crucial para escolher o mais adequado e poder aplicá-lo corretamente.

A seguir, descrevemos o método Rank Products, que avaliamos com o propósito de verificar se as suas características o tornam adequado para lidar com o perfil de expressão gênica encontrado nas doenças complexas. Também descrevemos o SAM, que serviu de referência para os nossos testes por ser um método de seleção mais conhecido.

1.1.2 Rank Products

O Rank Products (RP) é um método não paramétrico, desenvolvido por Breitling *et al.* (2004) para buscar por genes diferencialmente expressos em experimentos de *microarrays*. Ele é baseado em *ranking*, que consiste na classificação de todos os genes em todas as amostras por ordem decrescente de valor de expressão. O produto dos *rankings* de cada gene, é então calculado por:

$$RP_g^{up} = \prod_{i=1}^k (r_{i,g}^{up} / n_i)$$

onde, $r_{i,g}^{up}$ é a posição do gene g na lista de genes classificada em ordem decrescente de valores de expressão na amostra i th e n_i é o número total de genes.

Desta forma, os menores valores de RP indicam menor probabilidade de observar um gene no topo da lista de diferencialmente expressos (genes com expressão aumentada) ao acaso. O mesmo procedimento é feito para detectar os genes com expressão diminuída, mas classificando estes genes em ordem crescente de valor de expressão. No final, é gerado uma lista de genes com expressão aumentada e diminuída baseado no percentual de falsos positivos (pfp) estimados, também conhecido como FDR (*False Discovery Rate*).

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

O cálculo do nível de significância dos genes diferencialmente expressos é descrito a seguir. Ele é baseado em permutação, que gera um número de experimentos aleatórios que mudam os *rankings* dos genes do conjunto de dados original.

Passo 1) Gera p permutações de k lista de *rankings* de tamanho n .

Passo 2) Calcula o produto dos *rankings* de n genes em p permutações.

Passo 3) Conta quantas vezes o produto dos *rankings* dos genes nas permutações são menores ou iguais ao produto dos *rankings* observados. Atribui este valor a uma variável c .

Passo 4) Calcula o valor médio esperado para o produto dos *rankings*:

$$E(RP_g) = c / p$$

Passo 5) Calcula a porcentagem de falsos positivos:

$$pfp_g = E(RP_g) / \text{ranking}(g)$$

onde, $\text{ranking}(g)$ corresponde à posição do gene g na lista de todos os n genes classificados em ordem crescente de valor de RP.

1.1.3 SAM (*Significance Analysis of Microarrays*)

Este método foi introduzido por Tusher *et al.* (2001) também com o objetivo de identificar genes diferencialmente expressos. O método atribui um escore estatístico para cada gene baseado na razão da mudança da expressão em relação ao desvio padrão nos dados para este gene:

$$d(i) = \frac{\bar{x}_a(i) - \bar{x}_b(i)}{s(i) + s_0}$$

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

onde $\bar{x}_a(i)$ e $\bar{x}_b(i)$ correspondem à média dos níveis de expressão do gene i nas condições a e b , respectivamente, $s(i)$ é o desvio padrão e s_0 é uma pequena constante que serve para minimizar o coeficiente de variação, reduzindo a chance de um gene com baixo desvio padrão ser selecionado ao acaso.

$$s(i) = \sqrt{\left(\frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2} \right) \left\{ \sum_m \left[x_m(i) - \bar{x}_a(i) \right]^2 + \sum_n \left[x_n(i) - \bar{x}_b(i) \right]^2 \right\}}$$

onde n_1 e n_2 são os valores de expressão nas condições a e b . $\sum_m$ e $\sum_n$ são as somas dos níveis de expressão nas condições a e b , respectivamente.

Genes com escore maior que um limiar são considerados significantes. O percentual de falsos positivos é estimado através de permutação. Primeiro conta-se o número de genes que excedem um valor de corte nas permutações, que é definido pelo usuário. Em seguida, divide este número pelo número de genes considerados significantes.

1.2 Motivação

O SAM é um dos métodos mais populares para a seleção de genes diferencialmente expressos. Junto com outros métodos igualmente conhecidos, como o ANOVA (Kerr *et al.*, 2000) e o Limma (Smyth *et al.*, 2004), eles formam uma classe de métodos baseados no Teste-t, que é a abordagem mais comum para esta tarefa (Hong & Breitling, 2008). Isto significa que, dado dois grupos de amostras, um deles formado por indivíduos saudáveis e outro por indivíduos afetados, estes métodos basicamente decidem se um gene é ou não um diferencialmente

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

expresso comparando a média da expressão dos dois grupos. De uma forma simplificada, se a média for igual, significa que não há alteração de expressão. Se for diferente, maior ou menor, significa que o gene é um diferencialmente expresso com expressão aumentada ou diminuída, respectivamente (Dunkler *et al.*, 2011).

O RP é um método ainda pouco conhecido, citado 598 vezes até o momento, enquanto o SAM já foi citado 7.994 vezes⁸. Ao invés de média, ele é baseado em *ranking*, conforme descrito anteriormente. Isto significa que, dado dois grupos de amostras, ele decide se um gene é ou não diferencialmente expresso se este gene ficou bem "ranqueado" na maioria das amostras, sem levar em conta a variabilidade entre elas.

Nossa hipótese é que esta característica torna o RP uma solução adequada para dados de doenças complexas. Um gene com expressão alterada somente em um subgrupo de, por exemplo, 5 de 10 indivíduos afetados, ficaria bem "ranqueado" 5 vezes, deixando-o numa posição significativa no *ranking* final, fazendo dele um candidato. A motivação deste trabalho é mostrar se isso funciona, ou seja, se o RP é capaz de selecionar genes diferencialmente expressos relevantes para a patologia de doenças complexas.

Desde que foi criado, o RP vem sendo aplicado em problemas que envolvem heterogeneidade. Em 2008, uma versão do programa foi adaptada para analisar dados de diferentes experimentos de *microarrays* (gerados em diferentes estudos e/ou diferentes plataformas), que é a chamada meta-análise (Hong & Breitling, 2008). Em 2009, uma adaptação do método chamada MASTA, foi criada para buscar por genes presentes em diferentes listas de diferencialmente expressos, uma espécie de meta-análise de resultados (Reina-Pinto *et al.*, 2009). Recentemente, uma ferramenta chamada modRP, foi criada para analisar dados de diferentes GWAS⁹, com o objetivo de identificar efeitos pleiotrópicos na comorbidade de doenças. O *ranking* dos valores de expressão foi substituído por valores de *p*, resultantes dos GWAS (McEachin *et al.*, 2012). Embora estes trabalhos tenham mostrado a eficiência do RP para analisar dados com alta

⁸ Fonte: Google Scholar

⁹ do inglês: *genome-wide association studies*, a técnica consiste no estudo da associação entre variantes individuais (ex. SNPs) e o fenótipo de uma doença.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

variabilidade, ninguém mostrou ainda se o método pode ser útil para lidar com a heterogeneidade genética das doenças complexas.

1.3 Materiais e Métodos

A estratégia para testar a nossa hipótese foi avaliar o RP através de um enfoque computacional e biológico. No enfoque computacional, realizamos uma série de testes utilizando dados simulados. No enfoque biológico, aplicamos o método nos estudos de Fissura Lábio Palatina Não-Sindrômica (FLP/NS), autismo e num estudo que avalia os efeitos da privação do sono em humanos. A Figura 1.3 ilustra o cenário da nossa estratégia de testes, que são descritos nos subitens a seguir.

Figura 1.3: Estratégia de testes do método RP. A) Enfoque computacional: dados foram simulados para avaliar o método sob dois aspectos: heterogeneidade e sensibilidade. Nos testes de heterogeneidade verificamos como o RP lida com perfil de expressão heterogêneo entre as amostras. Nos testes de sensibilidade, avaliamos como ele lida com diferentes quantidades de genes e amostras. B) Enfoque biológico: o método foi aplicado em dados reais de estudos de FLP/NS, autismo e sono.

1.3.1 Enfoque computacional

1.3.1.1 Simulação dos dados

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Uma questão comum da simulação de dados é se o mesmo desempenho obtido com testes usando dados simulados poderia ser obtido com dados reais. Isto porque dados simulados não representam o verdadeiro comportamento biológico e por isso não podem ser utilizados como padrão ouro (Nykter *et al.*, 2006; Jeffery *et al.*, 2006; Breitling *et al.*, 2004). Neste trabalho, nós optamos por gerar os nossos dados mantendo a sua realidade biológica e as variáveis inerentes à produção dos *microarrays*. Desta forma, podemos garantir que não fomos tendenciosos na criação de valores que beneficiassem o algoritmo do RP e garantir também que qualquer evidência dos nossos testes serão o mais próximo possível da realidade.

Os nossos dados foram parcialmente simulados a partir dos dados de um estudo de asma (E-GEOD-8052) e de leucemia (E-GEOD-11877), que estão disponíveis publicamente no repositório *ArrayExpress*¹⁰. O estudo de asma conta com 404 *arrays* de expressão, sendo 136 de indivíduos controles (58 meninos e 78 meninas) e 268 de afetados (163 meninos e 105 meninas). O estudo de leucemia conta com um total de 270 *arrays* de indivíduos afetados, sendo 137 de meninos e 70 de meninas. Ambos utilizaram os chips Affymetrix Human Genome U133 Plus 2.0.

Criamos dois conjuntos de dados, um de tamanho relativamente grande (80 amostras) e outro pequeno (20 amostras). Abaixo descrevemos a simulação de dados nas 80 amostras. O mesmo procedimento foi usado para o conjunto menor.

Selecionamos 80 *arrays* de meninas com asma, seguindo os critérios de controle de qualidade descritos no item 1.3.1.4. Não usamos meninas controle pois tínhamos somente 78 *arrays*, muitos dos quais não passaram no nosso controle de qualidade. Os 80 *arrays* de meninas com asma foram divididos aleatoriamente em dois grupos: controle e teste. A obtenção dos valores de expressão destes *arrays* está descrita no item 1.3.1.5. Selecionamos 80 genes para terem seus valores de expressão simulados e, em seguida, analisados pelo RP. São todos genes do cromossomo Y e, portanto, ligados à determinação do sexo masculino. Utilizamos a expressão destes genes em meninas por não apresentarem variação entre elas,

¹⁰ <http://www.ebi.ac.uk/arrayexpress/>

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

representando uma boa fonte para testes que necessitam de uma base homogênea, com a vantagem de serem dados reais e não inventados.

Os passos para simular a heterogeneidade destes dados são descritos a seguir:

1) Calculamos a média da expressão de cada um dos 80 genes dos 80 indivíduos (Tabela 1.1):

Tabela 1.1: Simulação-Passo 1: Calculado a média dos valores de expressão de cada gene.

Genes do Y	CONTROLE							TESTE							Média
	C1	C2	C3	C4	C5	...	C40	T1	T2	T3	T4	T5	...	T40	
Y1	2.1	2.4	1.9	2.1	2.0	...	2.2	2.2	2.0	1.8	2.4	2.1	...	2.1	2.1
Y2	3.5	3.1	3.2	3.7	2.9	...	3.2	3.4	3.4	3.2	3.0	2.9	...	3.2	3.2
Y3	1.5	1.1	1.2	1.7	1.9	...	1.2	1.4	1.4	1.2	1.0	1.9	...	1.8	1.4
Y4	5.2	4.8	4.7	4.8	5.1	...	5.0	5.9	5.0	5.7	5.3	5.0	...	4.2	5.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Y80	2.9	3.2	3.4	3.0	3.0	...	3.2	3.1	3.0	2.9	3.2	3.1	...	3.0	3.1

2) As médias foram ordenadas em ordem crescente. Os genes foram então separados em dois grupos: um deles com $média < 3.0$ e o outro com $média \geq 3.0$. Na Tabela 1.2 eles são mostrados nas cores verde e vermelho, respectivamente:

Tabela 1.2: Simulação-Passo 2: Ordenada as médias. Em verde estão os genes com $média < 3.0$ e em vermelho os genes com $média \geq 3.0$.

Genes do Y	CONTROLE							TESTE							Média
	C1	C2	C3	C4	C5	...	C40	T1	T2	T3	T4	T5	...	T40	
Y3	1.5	1.1	1.2	1.7	1.9	...	1.2	1.4	1.4	1.2	1.0	1.9	...	1.8	1.4
Y1	2.1	2.4	1.9	2.1	2.0	...	2.2	2.2	2.0	1.8	2.4	2.1	...	2.1	2.1
Y80	2.9	3.2	3.4	3.0	3.0	...	3.2	3.1	3.0	2.9	3.2	3.1	...	3.0	3.1
Y2	3.5	3.1	3.2	3.7	2.9	...	3.2	3.4	3.4	3.2	3.0	2.9	...	3.2	3.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Y4	5.2	4.8	4.7	4.8	5.1	...	5.0	5.9	5.0	5.7	5.3	5.0	...	4.2	5.0

Este passo serviu apenas de marcação, nenhuma alteração foi feita nos dados. Ele foi feito para que, em seguida, pudéssemos inverter a direção da expressão desses genes nos indivíduos do grupo teste com relação ao controle. Como o perfil da expressão destes genes em todas as amostras é homogêneo, esta é uma forma de simular as duas situações encontradas em dados reais, a de

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

aumento e diminuição de expressão, e de testar como o RP lida com estas duas situações.

3) O primeiro passo da simulação foi a adição do valor 3.0 em cada um dos valores de expressão dos genes com média < 3.0 e a subtração deste mesmo valor dos genes com média ≥ 3.0. Iniciamos fazendo isso em somente 10% das 40 meninas do grupo teste. O resultado desta etapa é ilustrado na Tabela 1.3:

Tabela 1.3: Simulação-Passo 3: Simulação de heterogeneidade em 80 genes de 4 meninas do grupo teste, indicadas por flechas. Foi adicionado o valor 3.0 na expressão dos genes com média < 3.0 e subtraído o mesmo valor dos genes com média ≥ 3.0.

TESTE							
T1	T2	T3	T4	T5	...	T40	Média
1.4 + 3.0	1.4 + 3.0	1.2	1.0 + 3.0	1.9	...	1.8 + 3.0	1.4
2.2 + 3.0	2.0 + 3.0	1.8	2.4 + 3.0	2.1	...	2.1 + 3.0	2.1
3.1 - 3.0	3.0 - 3.0	2.9	3.2 - 3.0	3.1	...	3.0 - 3.0	3.1
3.4 - 3.0	3.4 - 3.0	3.2	3.0 - 3.0	2.9	...	3.2 - 3.0	3.2
...	...	...	...	...	...	...	...
5.9 - 3.0	5.0 - 3.0	5.7	5.3 - 3.0	5.0	...	4.2 - 3.0	5.0

O valor 3.0 foi definido a partir da observação da diferença real existente entre a expressão de meninos e meninas. Esta observação foi feita a partir de um teste onde analisamos 40 *arrays* de meninas junto com 10 *arrays* de meninos, todos afetados. Observamos que o nível da expressão de vários dos genes do Y nos meninos é, no mínimo, 2 vezes maior que nas meninas. O gene *DDX3Y*, por exemplo, apresentou média de expressão de 8.4 nos meninos e 2.1 nas meninas; o *EIF1AY*¹¹, apresentou média de expressão de 8.9 nos meninos e 2.3 nas meninas. Estes dados estão disponíveis no endereço http://danieleyumi.sunaga.de/wp-uploads/80genes_10meninos+40meninas.xls do material suplementar. Estas diferenças são, inclusive a razão de muitos trabalhos utilizarem estes dados para caracterizarem o sexo das amostras (Göhlmann & Talloen, 2009).

¹¹ Cada um destes genes possuem mais de um transcrito no *microarray*. O *EIF1AY*, por exemplo, possui os transcritos 204409_s_at e 204410_at. Nós usamos o termo "gene" como sinônimo de transcrito por questão de simplicidade.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

A simulação da variação de menor efeito (<3.0) também seria um bom critério de teste, mas neste trabalho nos limitamos a testar variações maiores e concentrar o nosso foco na variação do número de amostras simuladas.

No final desta etapa, as 4 meninas tinham os genes, antes aumentados, para diminuídos e os genes diminuídos para aumentados, como se tivéssemos as transformado em meninos. A Tabela 1.4 ilustra o resultado desta simulação.

Tabela 1.4: Resultado da simulação.

Genes do Y	CONTROLE							TESTE						
	C1	C2	C3	C4	C5	...	C40	T1	T2	T3	T4	T5	...	T40
Y3	1.5	1.1	1.2	1.7	1.9	...	1.2	4.4	4.4	1.2	4.0	1.9	...	4.8
Y1	2.1	2.4	1.9	2.1	2.0	...	2.2	5.2	5.0	1.8	5.4	2.1	...	5.1
Y80	2.9	3.2	3.4	3.0	3.0	...	3.2	0.1	0.0	2.9	0.2	3.1	...	0.0
Y2	3.5	3.1	3.2	3.7	2.9	...	3.2	0.4	0.4	3.2	0.0	2.9	...	0.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Y4	5.2	4.8	4.7	4.8	5.1	...	5.0	2.9	2.0	5.7	2.3	5.0	...	1.2

Criamos uma planilha Excel com cada um dos passos da simulação organizados em abas. Ela está disponível no endereço http://danieleyumi.sunaga.de/wp-uploads/passos_simulacao.xls do material suplementar. Neste material é possível ver os valores de expressão antes e após a simulação. A primeira aba contém os dados antes da simulação (originais) junto com o valor da média e desvio padrão (σ) dos níveis de expressão de cada gene. Todos eles apresentaram $\sigma < 0,5$, confirmando que a expressão de genes do Y em meninas trata-se de uma base de dados bastante homogênea.

1.3.1.2 Teste de heterogeneidade

Para testar se o RP é capaz de selecionar genes com perfil de expressão heterogêneo entre indivíduos com uma mesma doença, comparamos os grupos controle e teste, simulando a alteração da expressão dos 80 genes em, inicialmente, 10% dos indivíduos do grupo teste, conforme descrito no item anterior. O RP, disponível no pacote RankProd (Hong *et al.*, 2006) do programa Bioconductor/R¹², foi rodado seguindo o desenho de análise de duas classes não-

¹² <http://www.bioconductor.org/> e <http://cran.r-project.org/>.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

pareadas, com valor de $p < 0,05$ corrigidos por FDR e 3.000 permutações. Este procedimento foi repetido com 20, 40, 60, 80 e 100% dos indivíduos do grupo teste.

Todas as análises foram também rodadas com o método SAM, também seguindo o desenho de análise de duas classes não-pareadas, com valor de $p < 0,05$ corrigidos por FDR e 3.000 permutações.

1.3.1.3 Teste de sensibilidade

A sensibilidade do RP foi testada com relação à diferentes quantidades de genes e amostras. Para testar como o RP lida com diferentes quantidades de genes, repetimos o mesmo procedimento dos testes de heterogeneidade, mas usando um conjunto de 1.495 genes (80 do cromossomo Y e 1.415 do X), ao invés de somente 80 genes do Y. A adição de genes do cromossomo X foi definida arbitrariamente, sem qualquer intenção de comparação entre gêneros masculino e feminino. Eles foram usados somente para aumentar o número de genes de entrada para os métodos RP e SAM. Seus valores de expressão não foram alterados. A simulação permaneceu sendo feita somente nos 80 genes do cromossomo Y e em diferentes quantidades de indivíduos.

Para avaliar o RP sob o impacto da redução do número de amostras, comparamos 10 indivíduos do grupo controle contra 10 do grupo teste e repetimos todo o procedimento do teste anterior. Dez-20 é o número mínimo de indivíduos que tem sido recomendado para estudos em humano para ter acesso às mudanças de expressão gênica com precisão (Gosse *et al.*, 2008). Os testes com 20 amostras foram feitos com dados do estudo de asma e leucemia.

Todas as análises foram também rodadas com o método SAM. Os parâmetros, tanto do RP como do SAM foram os mesmos que usados no teste de heterogeneidade.

1.3.1.4 Controle de qualidade

Os *arrays* dos estudos de asma (E-GEOD-8052) e leucemia (E-GEOD-11877) que utilizamos em nossos testes foram selecionados seguindo os critérios de controle de qualidade recomendados pela fabricante Affymetrix. Os valores de BG_avg, Noise_avg e %P foram obtidos com o método de pré-processamento MAS5, disponível no programa Expression Console™ da Affymetrix. BG_avg

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

corresponde à média da intensidade de background de cada transcrito no *array*. Noise_avg corresponde à média da intensidade de ruído de cada transcrito no *array*. O %P refere-se ao percentual de presença de cada gene no *array*. Um *array* de boa qualidade deve basicamente apresentar baixo valor de BG_avg (<100) e Noise_avg (<5) e alto valor de P% (>30) (Affymetrix, 2007). Selecionamos os *arrays* com a melhor combinação destes 3 critérios.

1.3.1.5 Pré-processamento

Os valores de expressão foram obtidos a partir dos três passos do método de pré-processamento RMA (*Robust Multi-array Average*): correção de background, normalização e sumarização (Irizarry *et al.*, 2003), disponível no pacote affy do Bioconductor/R.

1.3.2 Enfoque biológico

Para verificar o desempenho do RP com dados reais, aplicamos o método num estudo de Fissura Lábio-Palatina Não-Sindrômica (FLP/NS), de autismo e num estudo de sono. Os estudos de FLP/NS e sono foram publicados e o manuscrito de autismo está em preparação. A seguir, nós descrevemos como o RP foi aplicado em cada um deles. Mais detalhes podem ser vistos nos respectivos artigos anexados no final deste capítulo.

1.3.2.1 Fissura Lábio-Palatina Não-Sindrômica (FLP/NS)

Neste estudo nós comparamos o transcriptoma de 6 indivíduos controles e 6 indivíduos com FLP/NS. O RP foi aplicado seguindo o desenho de duas classes não pareadas, com valor de $p < 0,05$ ajustado por FDR e 3.000 permutações.

1.3.2.2 Autismo

Neste estudo foram avaliados 7 indivíduos autistas e dois diferentes grupos de 6 controles cada. A primeira comparação de 6 controles e 7 afetados foi usada como referência e a segunda como validação dos resultados da primeira. Utilizamos a versão gráfica do RP disponível no programa MeV¹³ (Saeed *et al.*, 2003). Em ambas as comparações, ele foi aplicado seguindo o desenho de duas

¹³ <http://www.tm4.org/mev/>

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

classes não pareadas, com valor de $p < 0,05$ ajustado por FDR e 4.000 permutações. O SAM foi aplicado com os mesmos parâmetros que o RP.

1.3.2.3 Sono

O estudo do sono foi feito em colaboração com o grupo do Departamento de Psicobiologia da UNIFESP. Avaliamos o transcriptoma de 9 indivíduos sadios antes da privação do sono (controle), após 60h de privação (privado) e após 12h de sono (rebote). O RP foi aplicado seguindo o desenho de duas classes pareadas, com valor de $p < 0,01$ ajustado por FDR e 3.000 permutações. Comparamos os grupos controle x privado e privado x rebote.

1.4 Resultados e discussões

Todos os resultados deste capítulo estão disponíveis no endereço http://danieleyumi.sunaga.de/?page_id=30 do material suplementar.

1.4.1 Enfoque computacional

1.4.1.1 Teste de heterogeneidade

Iniciamos comparando os grupos controle e teste usando os valores de expressão originais, sem nenhuma simulação, para certificar que os dados eram homogêneos e que tanto o RP como o SAM não achariam nenhum DEG (do inglês *Differentially Expressed Gene*). O RP detectou 2 DEGs e o SAM nenhum.

Em seguida, calculamos a média da expressão dos 80 genes para dar início às simulações. Cinquenta e três deles apresentaram média < 3.0 e 27 apresentaram média ≥ 3.0 . Estes genes tiveram então seus valores de expressão simulados em 10% das amostras do grupo teste ($=4$). Os 53 genes que antes possuíam valores de expressão inferiores a 3.0 em todas as amostras passaram a ter expressão aumentada nas 4 amostras do grupo teste. Os 27 genes que antes possuíam expressão superior a 3.0 em todas as amostras passaram a ter expressão diminuída nestes mesmos indivíduos.

O RP detectou apenas 2 dos 27 genes que tiveram a expressão diminuída e nenhum com a expressão aumentada. O SAM detectou 26 dos 27 genes com a expressão diminuída e 35 dos 53 genes com expressão aumentada.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Na simulação da expressão de 20% das amostras (=8), o RP detectou 8 dos 27 genes que tiveram a expressão diminuída. O SAM detectou todos os genes com expressão diminuída e 34 dos 53 que tiveram a expressão aumentada.

Na simulação de 40% das amostras (=16), o RP detectou 19 dos 27 genes com expressão diminuída. O SAM detectou todos os genes com expressão diminuída e 31 dos 53 que tiveram a expressão aumentada.

Na simulação de 60% das amostras (=24), o RP detectou todos os 27 genes que tiveram a expressão diminuída e 9 dos 53 com expressão aumentada. O SAM detectou todos os genes com expressão diminuída e 29 dos genes com expressão aumentada.

Na simulação de 80% das amostras (=32), o RP detectou todos os 27 genes com expressão diminuída e 26 dos 53 com expressão aumentada. O SAM detectou todos os genes com expressão diminuída e 28 com expressão aumentada.

Por último, a simulação da expressão de todas as amostras, situação onde era esperado encontrar todos os genes pelos métodos RP e SAM, o RP detectou todos os 27 com expressão diminuída e 39 dos 53 genes com expressão aumentada. O SAM detectou 26 dos genes com expressão diminuída e 27 com expressão aumentada.

Os resultados destes testes estão reunidos na Tabela 1.5 e foram chamados de 40Cx40T_80g (40 controles x 40 testes usando 80 genes).

Tabela 1.5: Resultados do teste de heterogeneidade. Cada linha da tabela corresponde ao percentual de amostras que tiveram a expressão dos 80 genes simulada. A porção destes genes que teve a expressão diminuída e aumentada está entre parênteses na coluna "Dim. / Aum.", respectivamente. Por exemplo, 2(27) / 0(53), significa que o RP detectou 2 dos 27 genes que tiveram a expressão diminuída e nenhum dos 53 genes com expressão aumentada.

		RP	SAM
		#DEGs detectados (#DEGs simulados)	#DEGs detectados (#DEGs simulados)
		Dim. / Aum.	Dim. / Aum.
40Cx40T_80g	original	2(27) / 0(53)	0
	10%	2(27) / 0(53)	26(27) / 35(53)
	20%	8(27) / 0(53)	27(27) / 34(53)
	40%	19(27) / 0(53)	27(27) / 31(53)
	60%	27(27) / 9(53)	27(27) / 29(53)
	80%	27(27) / 26(53)	27(27) / 28(53)

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

100%	27(27) / 39(53)	26(27) / 27(53)
------	-----------------	-----------------

Estes resultados estão disponíveis no endereço http://danieleyumi.sunaga.de/wp-uploads/resultados_40Cx40T_80g.xls do material suplementar. Eles revelaram o melhor desempenho do SAM para detectar genes que não estão igualmente variáveis entre as amostras. A simulação da expressão de apenas 10% das amostras foi suficiente para o método encontrar quase todos os nossos genes simulados. No entanto, interessante, a performance do método para detectar genes com expressão diminuída piorou conforme foi aumentado o número de amostras simuladas. Ele detectou 35 dos 53 genes aumentados em 10% das amostras mas esse número caiu para 27 quando a simulação foi feita em todas as amostras. O desempenho do RP, por outro lado, melhorou conforme aumentado o número de simulações. O método, contudo, apresentou pior desempenho que o SAM. Ele só foi capaz de encontrar o número de genes que o SAM encontrou com a simulação de 10% das amostras quando todas elas foram simuladas.

Resumindo, estes resultados mostram que o SAM é mais eficiente que o RP quando o estudo envolve um número grande de amostras e um pequeno número de genes de entrada. É possível que a performance do RP tenha sido influenciada pelo pequeno número de genes. Num esquema de *ranking*, quanto maior o número de elementos (genes) maior a chance de distinguir suas posições no *ranking* (Gosse *et al.*, 2008).

1.4.1.2 Teste de sensibilidade

Os resultados do teste de sensibilidade do RP à diferentes quantidades de genes estão reunidos na Tabela 1.6 e foram chamados de 40Cx40T_1495g. Neste teste analisamos um total de 1.495 genes (80 do cromossomo Y e 1.415 do X), embora somente os 80 do Y tiveram a expressão simulada. Por esta razão, a Tabela 1.6 também mostra a quantidade total de DEGs detectados, que pode incluir os genes do X.

Tabela 1.6: Resultados do teste de sensibilidade à diferentes quantidades de genes. A coluna "#Total DEGs detectados" corresponde à quantidade total de genes diferencialmente expressos detectados, que pode incluir os genes que tiveram e não tiveram a expressão simulada. Por exemplo, 53 / 108, significa que, de um total de 1.495 genes, o RP detectou

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

53 com expressão diminuída e 108 com expressão aumentada e, conforme mostrado na coluna "#DEGs detectados", apenas 1 deles é um dos simulados.

		RP		SAM	
		#DEGs detectados (#DEGs simulados)	#Total DEGs detectados	#DEGs detectados (#DEGs simulados)	# Total DEGs detectados
		Dim. / Aum.	Dim. / Aum.	Dim. / Aum.	Dim. / Aum.
40Cx40T _1495g	original	1(27) / 0(53)	53 / 108	0	0
	10%	6(27) / 0(53)	53 / 97	0	0
	20%	27(27) / 42(53)	76 / 138	0(27) / 50(53)	0 / 51
	40%	27(27) / 53(53)	68 / 129	27(27) / 53(53)	27 / 54

Estes resultados foram organizados em dois arquivos e estão disponíveis nos endereços http://danieleyumi.sunaga.de/wp-uploads/resultados_40Cx40T_1495g_1.zip e http://danieleyumi.sunaga.de/wp-uploads/resultados_40Cx40T_1495g_2.zip do material suplementar. O primeiro deles contém os resultados das simulações e o segundo os resultados do RP e SAM. Eles mostraram que o desempenho do RP superou o do SAM quando usado um número maior de genes de entrada, corroborando nossa observação prévia de que quanto maior o número de genes disponível para *ranking*, menos provável a chance de um gene ser bem "ranqueado" ao acaso. A variação da expressão em somente 8 (20%) dos 40 indivíduos foi suficiente para o RP conseguir detectar quase todos os nossos genes alvos, enquanto o SAM só detectou parte deles. A Figura 1.4 ilustra o desempenho do RP com diferentes quantidades de genes.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Figura 1.4: Sensibilidade do RP à diferentes quantidades de genes. A linha rosa corresponde ao teste onde foram analisados 80 genes (40Cx40T_80g). A linha laranja corresponde ao teste onde foram analisados 1.495 genes (40Cx40T_1495g). O eixo x corresponde ao percentual de amostras que tiveram a expressão simulada. O eixo y corresponde à quantidade de DEGs detectados pelo RP. A linha pontilhada vermelha marca o número total de DEGs que tiveram a expressão simulada.

Estes resultados também mostraram que, além dos genes do cromossomo Y que tiveram a expressão simulada, o RP também selecionou genes do cromossomo X, cuja expressão foi mantida inalterada. Como usamos dados reais em nossos testes, é possível que estes genes tenham sido selecionados por apresentarem variações biológicas reais entre as amostras. Variações técnicas decorrentes da confecção dos *arrays* também podem explicar a seleção excedente. Nossa hipótese, é de que o mesmo recurso que confere ao RP a capacidade de identificar genes de relevância biológica em um subgrupo de amostras também leva à identificação de genes ao acaso. Há caminhos, contudo, que podem ajudar a remover estes genes. Corte por *fold change* e por funções biológicas distantes do foco do estudo são alguns deles (Reina-Pinto *et al.*, 2009). O que não há, é a chance de descobrir genes ligados à doença sem que eles tenham sido selecionados. O SAM, na situação de 20% de simulação, por exemplo, detectou um total de 51 DEGs sendo 50 deles simulados. Estes números podem demonstrar precisão, no entanto, 30 genes simulados não foram detectados. Estes resultados destacam a importância de conhecer as limitações do método estatístico empregado. É melhor lidar com uma seleção de genes que inclui a maioria dos candidatos e alguns falsos positivos, como o resultado do RP, ou lidar com resultados de uma seleção mais rigorosa, como o resultado do SAM?

No teste de sensibilidade do RP à diferentes quantidades de amostras foram usadas 20 amostras (10 do grupo controle e 10 do grupo teste) e analisados 1.495 genes.

Inicialmente comparamos os dois grupos sem nenhuma simulação. Tanto o RP como o SAM não detectaram nenhum DEG. Simulando a expressão de 20% das amostras, eles também não detectaram nenhum DEG. Simulando 40%, o RP notavelmente detectou todos os genes que tiveram a expressão diminuída e 30 dos 53 genes que tiveram a expressão aumentada. O SAM não detectou nenhum. Na simulação de 60%, o RP identificou todos os genes simulados e o SAM

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

identificou 26 dos 27 genes com expressão diminuída e todos os que tiveram a expressão aumentada. Estes resultados foram chamados de 10Cx10T_1495g_1.

Eles corroboram a nossa observação prévia de que o RP é eficiente para detectar genes que não variam igualmente entre as amostras e revela que essa eficiência é mantida mesmo quando o tamanho amostral é pequeno. Para confirmar essa evidência e que elas não ocorreram ao acaso ou que são dependentes dos dados de asma, repetimos o mesmo procedimento usando três diferentes conjuntos de 20 amostras. No primeiro, usamos as outras 20 amostras de meninas afetadas do estudo de asma, no segundo, usamos 20 amostras de meninas controles e, no terceiro, usamos 20 amostras de meninas do estudo de leucemia. Os resultados foram chamados de 10Cx10T_1495g_2, 10Cx10T_1495g_3 e 10Cx10T_1495g_4, respectivamente, e estão reunidos na Tabela 1.7.

No teste em que usamos as 20 amostras de meninas afetadas do estudo de asma, tivemos que recalcular as médias dos genes pois as amostras são diferentes das utilizadas no teste anterior. Trinta e um dos 80 genes do cromossomo Y apresentaram média < 3.0 e 49 apresentaram média ≥ 3.0. O mesmo foi feito com os dados de leucemia, onde encontramos 67 genes com média < 3.0 e 13 com média ≥ 3.0.

Tabela 1.7: Resultados do teste de sensibilidade à diferentes quantidade de amostras.

		RP		SAM	
		#DEGs detectados (#DEGs esperados)	# Total DEGs detectados	#DEGs detectados (#DEGs esperados)	#Total DEGs detectados
		Dim. / Aum.	Dim. / Aum.	Dim. / Aum.	Dim. / Aum.
10Cx10T _1495g_ _1	original	0	0	0	0
	20%	0	6 / 15	0	0
	40%	27(27) / 30(53)	37 / 47	0	0
	60%	27(27) / 53(53)	35 / 64	26(27) / 53(53)	26 / 56
10Cx10T _1495g_ _2	original	0	0	0	0
	20%	0	5 / 5	0	0
	40%	25(27) / 28(53)	30 / 33	0	0
	60%	27(27) / 53(53)	32 / 58	27(27) / 53(53)	27 / 53
10Cx10T _1495g_ _3	original	0	0	0	0
	20%	0	1 / 6	0	0
	40%	26(31) / 38(49)	34 / 50	0	0
	60%	31(31) / 49(49)	36 / 56	21(31) / 49(49)	21 / 49

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

10Cx10T _1495g_ 4	original	0	34 / 37	0	3 / 0
	20%	0	32 / 18	0	3 / 0
	40%	13(13) / 14(67)	47 / 31	0	3 / 0
	60%	13(13) / 67(67)	46 / 86	11(13) / 67(67)	14 / 68

Os resultados destas quatro análises estão disponíveis nos endereços:
http://danieleyumi.sunaga.de/wp-uploads/resultados_10Cx10T_1495g_1.xls,
http://danieleyumi.sunaga.de/wp-uploads/resultados_10Cx10T_1495g_2.xls,
http://danieleyumi.sunaga.de/wp-uploads/resultados_10Cx10T_1495g_3.xls,
http://danieleyumi.sunaga.de/wp-uploads/resultados_10Cx10T_1495g_4.xls do material suplementar. Eles confirmam o melhor desempenho do RP com poucas amostras, comparado ao SAM. Conclusão similar tem sido mostrada em vários trabalhos (Breitling & Herzyk, 2005; Jeffery *et al.*, 2006; Hong & Breitling, 2008; Gosse *et al.*, 2008). Eles, contudo, só consideraram o tamanho amostral. Aqui, nós mostramos que, além do RP ser eficaz com poucas amostras ele é eficaz em detectar genes alterados em apenas uma fração delas. A alteração da expressão de um gene em 40% de 10 amostras foi suficiente para ele ser detectado pelo RP, enquanto para o SAM, foram necessárias 60% das amostras para alcançar resultado equivalente. Uma observação similar foi mostrada por Breitling e colaboradores no trabalho em que eles introduzem o RP. Num dos testes onde compararam a performance do RP com o SAM, eles observaram que um subconjunto de 4 amostras analisadas pelo RP teve poder equivalente a 8 amostras analisadas pelo SAM para identificar genes que eles já sabiam serem verdadeiros diferencialmente expressos (Breitling *et al.*, 2004).

A diferença observada entre o desempenho do RP e SAM com um conjunto menor de amostras é principalmente devido à estimativa da variância. Quando o tamanho amostral é pequeno, essa estimativa se torna muito instável. A lógica por trás deste conceito é simples. Se estimássemos o peso médio dos homens no Brasil baseado em um único indivíduo, a estimativa seria bastante instável. Provavelmente seria alterada a cada nova amostra analisada. Por outro lado, se a amostragem fosse de 100 homens, aleatoriamente selecionados no país, isto faria da estimativa não somente mais representativa, mas também mais estável.

O bom desempenho do RP, neste caso, é explicado pelo fato de que ele não depende da estimativa da variância de cada gene, ao contrário dos métodos

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

baseados no teste-t, como o SAM. Os gráficos da Figura 1.5 ilustram o desempenho do RP e SAM com poucas amostras, nas quatro análises descritas anteriormente.

Figura 1.5: Desempenho do RP com poucas amostras. A linha azul corresponde aos resultados do RP e a linha verde aos resultados do SAM. O eixo x corresponde ao percentual de amostras que tiveram a expressão simulada, de um total de 10 amostras. O eixo y corresponde à quantidade de DEGs detectados. A linha pontilhada vermelha indica o número total de genes que tiveram a expressão simulada. As imagens A, B, C e D representam os testes 10Cx10T_1495g_1, 10Cx10T_1495g_2, 10Cx10T_1495g_3 e 10Cx10T_1495g_4, respectivamente.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Considerando que estudos com *microarrays* geralmente envolvem poucas amostras devido à dificuldade de encontrar indivíduos afetados, de obter RNAm em quantidade e qualidade suficientes e do custo dos *arrays*, estes achados são importantes para mostrar que resultados de confiabilidade tão boa quanto a do uso de muitas amostras podem ser obtidos pelo RP com poucas amostras.

Por último, fizemos um teste para verificar a relevância das nossas simulações. Nós repetimos uma das análises (10Cx10T_1495g_1), mas ao invés de simular, nós substituímos 40% das meninas do grupo teste por meninos. O SAM não detectou nenhum DEG do cromossomo Y. O RP, em contrapartida, detectou 17 genes do Y com expressão consistentemente alterada nos 4 meninos, validando a nossa estratégia de simulação.

A Figura 1.6 ilustra a diferença da expressão de um destes genes, o *KDM5D*, em meninas e meninos através de um *dot plot*. Cada ponto (*dot*) representa uma amostra. No eixo x estão os grupos controle e teste. No eixo y estão os valores de expressão.

Figura 1.6: *Dot plot* dos valores de expressão do gene *KDM5D*, do cromossomo Y, que foi detectado como DEG pelo RP mas não pelo SAM na comparação de meninas e meninos. Cada ponto corresponde a um indivíduo. Os pontos do grupo controle correspondem a 10 meninas (afetadas asma). Os pontos do grupo teste ao lado correspondem a 6 meninas (afetadas asma). Os pontos do grupo teste no topo do gráfico correspondem a 4 meninos (afetados asma).

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Este resultado é importante para mostrar que os dados simulados que usamos em nossos testes serviram como um bom modelo da realidade. Mostraram que os valores da expressão dos genes do cromossomo Y em meninas são bastante homogêneos e a constante 3.0 que foi adicionada/subtraída destes valores corresponde à variação real observada nos genes do Y em meninos. A nossa estratégia de simulação tem a vantagem de poder ser aplicada em qualquer conjunto de dados de meninos e meninas (sadios ou afetados), desde que a doença estudada não seja ligada ao sexo. Estes dados são facilmente encontrados nos repositórios de dados públicos.

Os resultados deste último teste foram chamados de 10Cx10T_1495g_reais e estão disponíveis no endereço http://danieleyumi.sunaga.de/wp-uploads/resultados_10Cx10T_1495g_reais.xls do material suplementar.

1.4.2 Enfoque biológico

1.4.2.1 Fissura Lábio-Palatina Não-Sindrômica (FLP/NS)

O RP identificou 87 genes com expressão diferenciada entre os grupos controle e afetado. Três deles, selecionados aleatoriamente, foram confirmados com a técnica de qRT-PCR¹⁴ usando as mesmas amostras dos *microarrays*. Estes genes, junto com outros 12, foram também confirmados em novas amostras (16 controles e 13 afetados, sendo 4 dos controles e 4 dos afetados, os mesmos usados nos *microarrays*). Identificamos uma rede biológica que reúne 13 dos 87 genes, vários deles que codificam proteína extracelular, sugerindo o envolvimento do processo de Transição Eitélíio Mesênquima (EMT) - que desempenha papel fundamental no fechamento do palato e na fusão do lábio - na etiologia da doença. Selecionamos 6 genes que fazem parte desta rede e que foram confirmados com qRT-PCR para uma análise de agrupamento usando os dados de expressão de ambas as técnicas. Interessantemente, observamos um padrão na expressão destes genes compartilhado por 4 dos 6 afetados dos *microarrays* e o mesmo padrão foi também observado em 7 dos 13 afetados do qRT-PCR. Este resultado confirma os nossos achados, de que o RP é uma boa solução para o

¹⁴ PCR quantitativo em tempo real

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

estudo de doenças complexas. Ele foi capaz de identificar genes com expressão consistentemente alterada em somente um subgrupo de amostras, mesmo num pequeno conjunto amostral.

1.4.2.2 Autismo

O RP selecionou 684 genes diferencialmente expressos que levaram ao enriquecimento de funções e vias biológicas ($p < 0,05$) previamente associadas à doença. Dentre estes genes está o *CHD8*, que foi o único identificado por três trabalhos recentes que usaram a tecnologia de sequenciamento de nova geração para encontrar mutações em indivíduos autistas. Como forma de testar a validade dos nossos achados, comparamos as amostras de afetados com um diferente grupo de controles. Identificamos 701 genes diferencialmente expressos, 206 deles comuns ao resultado anterior. As mesmas funções e vias biológicas foram novamente encontradas enriquecidas significativamente. Além disso, 25 destes genes estão no resultado de uma compilação recente de 430 genes associados ao autismo. Também rodamos o método SAM usando as mesmas amostras que levou aos 684 genes iniciais. Foram encontrados 149 genes com expressão diferenciada (dados não mostrados). Nenhuma função ou via biológica interessante foi identificada e nem o gene *CHD8* foi selecionado, mostrando que o RP é mais eficiente para lidar com a heterogeneidade genética da doença.

1.4.2.3 Sono

O RP detectou um total de 500 genes diferencialmente expressos. Funções como a de resposta a estresse, reparo e dano ao DNA, assim como diversas funções de resposta do sistema imune foram enriquecidas, dando suporte à idéia existente de que a perda do sono pode levar à alterações em processos moleculares que resultam na perturbação da imunidade celular, indução de resposta inflamatória e desequilíbrio homeostático. Três destes genes já foram previamente associados à privação do sono. Interessantemente, 76 deles apresentaram alteração de expressão coordenada nas três condições avaliadas. Eles tiveram a expressão diminuída na privação com relação aos controles e aumentada no rebote com relação à privação, servindo como uma assinatura molecular dos efeitos da privação do sono neste estudo. Dentre eles, destacam-se genes associados à homeostase, indicando a tentativa do corpo de restabelecer o equilíbrio com a recuperação do sono. Nestes estudo também observamos um

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

padrão de expressão consistente em um subgrupo de 4 dos 9 indivíduos analisados. Embora desordens do sono não sejam doenças complexas, estes resultados confirmam a habilidade do RP para lidar com a variabilidade inter-indivíduo que é conhecida em resposta à privação do sono.

1.5 Conclusões

Nossos resultados mostram que o método RP, em geral, tem desempenho superior a do popular método SAM para o estudo de doenças complexas em ambos, dados simulados e reais, especialmente em conjuntos de poucas amostras e maior número de genes. A aplicação do RP em dados simulados demonstrou que o *ranking* de genes leva a uma discriminação acurada daqueles que estão consistentemente alterados em apenas um subgrupo de amostras. Nos testes com maior número de amostras e genes, o método foi capaz de detectar alteração significativa de expressão em apenas 20% das amostras, enquanto para o SAM foram necessárias 40%. Em contrapartida, o desempenho do SAM neste mesmo teste foi bastante superior quando usado um conjunto de poucos genes. Estes achados revelam a sensibilidade de ambos os métodos à quantidade de genes analisados, o que na prática, significa que deve-se atentar para a aplicação de filtros de dados rigorosos dependendo do método de seleção a ser empregado em seguida. O resultado mais importante contudo, foi o notável desempenho do RP com poucas amostras. Esta habilidade do método já foi descrita pelo autor do programa e vários outros trabalhos. Aqui, no entanto, nós mostramos que ela é mantida mesmo com dados de alta variabilidade, como é esperado encontrar em doenças complexas. A aplicação do RP em dados reais levou à identificação de genes alterados em subgrupos de indivíduos afetados, vários deles, incluindo funções e vias biológicas, previamente associados às desordens estudadas. Juntos, os resultados dos testes com dados simulados e reais confirmam a nossa hipótese de que o RP é uma solução adequada e poderosa para a seleção de genes diferencialmente expressos em estudos de doenças complexas.

1.6 Referências Bibliográficas

Affymetrix (2007). Affymetrix GeneChip® Gene and Exon Array Whitepaper Collection: Quality Assessment of Exon and Gene Arrays. Obtido de

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

http://media.affymetrix.com/support/technical/whitepapers/exon_gene_arrays_ga_whitepaper.pdf, Último acesso: 19/12/2012.

Benjamini Y & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

Benjamini Y & Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Statist*, 29:1165–1188.

Breitling R, Armengaud P, Amtmann A, Herzyk P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573:83–92.

Breitling R & Herzyk P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol.*, 3(5):1171–89.

Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. (2006). The Affymetrix GeneChip Platform: An Overview. *Methods Enzymol*, 410:3–28.

Dunkler D, Sánchez-Cabo F, Heinze G. (2011). Statistical analysis principles for Omics data. *Methods Mol Biol.*, 719:113–131.

Efron B, Tibshirani R, Storey JD, Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*, 96:1151–1160.

Göhlmann H & Talloen W. (2009). Gene Expression Studies Using Affymetrix Microarrays, Chapman and Hall, CRC mathematical & computational biology series, pg. 54.

Gosse JA, Hampton TH, Davey JC, Hamilton JW. (2008). A New Approach to Analysis and Interpretation of Toxicogenomic Gene Expression Data and its Importance in Examining Biological Responses to Low, Environmentally Relevant Doses of Toxicants. *John Wiley & Sons, Ltd*, Chapter2.

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

Hong F & Breitling R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4), e15.

Jeffery IB, Higgins DG, Culhane AC. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 26:7:359.

Kerr MK, Martin M, Churchill GA. (2000). Analysis of variance for gene expression microarray data. *J Comput Biol.*, 7(6):819–37.

Lin DY. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21:781–787.

McEachin R, Sannareddy K, Cavalcoli J, Karnovsky A, Vink J, Sartor M. (2012). Convergence of genetic influences in comorbidity. *BMC Bioinformatics*, 13(Suppl2):S8.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

Newton MA, Noueiry A, Sarkar D, Ahlquist P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155-176.

Nykter M, Aho T, Ahdesmäki M, Ruusuvaari P, Lehmuusola A, Yli-Harja O. (2006). Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, 7:349.

Pritchard JK, Cox NJ. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet.*, 11(20):2417-23.

Reina-Pinto JJ, Voisin D, Teodor R, Yephremov A. (2009). Probing differentially expressed genes against a microarray database for in silico suppressor/enhancer and inhibitor/activator screens. *Plant J.*,61(1):166-75.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, *et al.* (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374-8.

Slonin D & Yanai I. (2009). Getting Started in Gene Expression Microarray Analysis. *PLOS Computational Biology*, 5(10): e1000543.

Smyth GK. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3.

Tan YD, Fornage M, Fub YX. (2006). *Ranking* Analysis of Microarray Data: A Powerful Method for Identifying Differentially Expressed Genes. *Genomics*, 88(6): 846–854.

Tusher VG, Tibshirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121.

Capítulo I

Busca por genes diferencialmente expressos em doenças complexas com o Rank Products

1.7 Anexos

Human Stem Cell Cultures from Cleft Lip/Palate Patients Show Enrichment of Transcripts Involved in Extracellular Matrix Modeling By Comparison to Controls

Daniela Franco Bueno · Daniele Yumi Sunaga · Gerson Shigeru Kobayashi ·
Meire Agüena · Cassio Eduardo Raposo-Amaral · Cibele Masotti ·
Lucas Alvizi Cruz · Peter Lees Pearson · Maria Rita Passos-Bueno

Published online: 30 October 2010

© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Nonsyndromic cleft lip and palate (NSCL/P) is a complex disease resulting from failure of fusion of facial primordia, a complex developmental process that includes the epithelial-mesenchymal transition (EMT). Detection of differential gene transcription between NSCL/P patients and control individuals offers an interesting alternative for investigating pathways involved in disease manifestation. Here we compared the transcriptome of 6 dental pulp stem cell (DPSC) cultures from NSCL/P patients and 6 controls. Eighty-seven differentially expressed genes (DEGs) were identified. The most significant putative gene network comprised 13 out of 87 DEGs of which 8 encode extracellular proteins: *ACAN*, *COL4A1*, *COL4A2*, *GDF15*, *IGF2*, *MMP1*, *MMP3* and *PDGFA*. Through clustering analyses we also observed that *MMP3*, *ACAN*, *COL4A1*

and *COL4A2* exhibit co-regulated expression. Interestingly, it is known that *MMP3* cleavages a wide range of extracellular proteins, including the collagens IV, V, IX, X, proteoglycans, fibronectin and laminin. It is also capable of activating other MMPs. Moreover, *MMP3* had previously been associated with NSCL/P. The same general pattern was observed in a further sample, confirming involvement of synchronized gene expression patterns which differed between NSCL/P patients and controls. These results show the robustness of our methodology for the detection of differentially expressed genes using the RankProd method. In conclusion, DPSCs from NSCL/P patients exhibit gene expression signatures involving genes associated with mechanisms of extracellular matrix modeling and palate EMT processes which differ from those observed in controls. This comparative approach should lead to a more rapid identification of gene networks predisposing to this complex malformation syndrome than conventional gene mapping technologies.

Daniela Franco Bueno and Daniele Yumi Sunaga contributed equally to this work

Electronic supplementary material The online version of this article (doi:10.1007/s12015-010-9197-3) contains supplementary material, which is available to authorized users.

D. F. Bueno · D. Y. Sunaga · G. S. Kobayashi · M. Agüena ·
C. Masotti · L. A. Cruz · P. L. Pearson · M. R. Passos-Bueno
Human Genome Research Center,
Biosciences Institute of University of São Paulo (USP),
São Paulo, São Paulo, Brazil

C. E. Raposo-Amaral
Sobrapar Hospital,
Campinas, São Paulo, Brazil

M. R. Passos-Bueno (✉)
Depto. Genética e Biologia Evolutiva, Instituto de Biociências,
Universidade de São Paulo,
Rua do Matão, 277,
São Paulo, SP 05508-900, Brazil
e-mail: passos@ib.usp.br

Keywords Nonsyndromic cleft lip and palate ·
Gene expression profile · Dental pulp · Stem cell ·
Epithelial-mesenchymal transition · Extracellular matrix

Introduction

Nonsyndromic cleft lip and palate (NSCL/P [MIM 119530]), a complex multifactorial disorder, is one of the most common congenital malformations, with a prevalence of 0.69 to 2.35 per 1,000 births in the Caucasian population [1]. Taking account of the complexities of this orofacial malformation and the long rehabilitation period following surgery, cleft lip and palate is considered to be a major psychosocial and economic burden for families and society. Gaining insight

into the genetic causes of NSCL/P should lead to future improvement of genetic counseling, preventive and curative measures.

The development of the human face begins with migration of neural crest cells that combine with the core mesoderm and pharyngeal ectoderm, establishing the facial primordia, which in turn give rise to structures associated with upper lip and palate formation [2, 3]. The growth, differentiation and fusion of these structures are genetically determined, and it is likely that disturbances in genetic pathways orchestrating these processes result in facial abnormalities, such as cleft lip and palate [2–5]. In this context, the epithelial-mesenchymal transition (EMT) plays a central role in generating the cranial neural crest cells as well as ensuring palate and lip fusion [6–8].

The large number of genes known or suspected to be involved in clefting probably reflects the diversity of embryological events that contribute to the formation of these facial structures [4, 9–16]. Although gene mapping approaches, such as Genome-Wide Association Studies (GWAS), appeared to offer an option to identify at-risk alleles associated with NSCL/P, with better reproducibility among different studies [10, 11, 17, 18] than candidate genes, lack of progress over the last decade suggests that GWAS are still unlikely to provide sufficient information on the genetic etiology underlying the disease. However, genome-wide expression analyses based on differential gene expression associated with NSCL/P, as proposed here; present a viable and challenging alternative, since patterns of co-expression can be used to identify biological pathways or gene networks associated with disease predisposition. The current data supporting this supposition can be summarized as follows: transcriptome analysis in tissues of cleft palate (CP) patients showed a distinct gene expression signature when compared to CL/P [19]. It has been reported that a few genes coding for extracellular matrix proteins, such as *TGFB3* and *MMP3*, are differentially expressed in fibroblasts of NSCL/P patients as compared to controls [20, 21], suggesting that the transcriptome of NSCL/P cells might exhibit a specific expression signature irrespective of origin of the cells concerned.

The use of mesenchymal stem cells (MSCs) or induced pluripotent stem cells (iPSs) has been shown to be a promising new approach to study gene function and signaling pathways in genetic disorders [22–24]. MSCs constitute a long-lived population of cells possessing self-renewal and differentiation properties [25–29]. Accordingly, these cells are a good model to study the *in vitro* characteristics of NSCL/P, since in addition to gene expression, they can be tested for proliferation, migration and differentiation properties, including EMT, functions that are presumed to be altered in cells of NSCL/P patients during embryonic development. MSCs were originally isolated from bone

marrow, and subsequently, similar cell populations were isolated from other tissues, such as adipose tissue [27], dental pulp [28, 30], *orbicularis oris* muscle [26], umbilical cord blood, and umbilical cord tissue [31]. Moreover, MSCs can be easily obtained from non-invasive sources, such as exfoliated teeth, both from NSCL/P patients and control individuals [28].

The main aim of this study was to verify if there are consistent gene expression profile differences between mesenchymal stem cells from NSCL/P patients and controls. We chose to study stem cells from dental pulp as they can be obtained relatively easy from both NSCL/P patients and controls. In addition, these stem cells, as for any other cells obtained from craniofacial tissues, are derived from cranial neural crest cells which play an important role in craniofacial development, including lip and palate [32, 33]. Our results will provide a base line for further investigation and insights into genetic pathway irregularities associated with craniofacial clefts.

Materials and Methods

Sample: NSCL/P Patients and Controls

Ethical approval to obtain stem cells from dental pulp of deciduous teeth was obtained from the Biosciences Institute Research Ethics Committee (Protocol 037/2005). Samples were included only after signed informed consent by the parents or legal guardians.

Deciduous teeth from controls were obtained from odontopediatric clinics in Sao Paulo, while teeth from NSCL/P patients were excised during the exfoliation period by Dr. Bueno D.F. at Sobrapar, Campinas, Sao Paulo. An individual was classified as NSCL/P if no other malformations than clefting of both lip and palate were present.

We analyzed mRNA of dental pulp stem cell cultures (DPSC) obtained from 6 controls and 6 NSCL/P patients (Table 1, supplementary material) for microarray expression analysis and for validation of 4 genes by quantitative Real-Time PCR (qRT-PCR). Validation of the microarray expression analysis by qRT-PCR for a larger number of genes was also done in mRNA obtained from DPSC cultures of 16 additional controls and 13 NSCL/P patients.

Stem Cell Culture

Stem cell cultures obtained from DPSC of deciduous teeth were established according to previously published protocols [28]. Cells were cultured at 37°C with 5% CO₂ in DMEM-F12 (Invitrogen, UK) supplemented with 15% Fetal Bovine Serum (FBS, HyClone, USA) and frozen in

Table 1 List of 87 differentially expressed genes sorted out by comparing controls and NSCL/P patients with the RankProd analysis

AFFY ID	Gene symbol	FC ^a	Pfp ^b	Cytoband
8176375	RPS4Y1	5.1099	0.0007	Yp11.3
8176624	DDX3Y	4.9358	0.0000	Yq11
8041206	LBH	4.6707	0.0000	2p23.1
8118509	PPT2 ^c	4.2212	0.0008	6p21.3
7972750	COL4A1 ^{d e}	3.6284	0.0006	13q34
7932254	ITGA8 ^{d e}	3.2852	0.0037	10p13
8125537	HLA-DMA ^e	3.2541	0.0005	6p21.3
7998927	—	3.0562	0.0053	—
8108370	EGR1	3.0266	0.0124	5q31.1
8176719	EIF1AY	2.9797	0.0055	Yq11.222
7952205	MCAM ^{d e}	2.9129	0.0123	11q23.3
7953200	CCND2	2.7894	0.0089	12p13
8056491	SCN9A ^e	2.7724	0.0223	2q24
7964388	NDUFA4L2	2.7670	0.0112	12q13.3
8113504	C5orf13	2.7473	0.0077	5q22.1
8176578	USP9Y	2.7360	0.0219	Yq11.2
8104663	CDH6 ^{d e}	2.7285	0.0263	5p15.1-p14
7985786	ACAN ^{d e}	2.6889	0.0042	15q26.1
8121838	TPD52L1	2.6688	0.0106	6q22-q23
8177137	UTY	2.6455	0.0280	Yq11
7970033	COL4A2 ^{d e}	2.6123	0.0210	13q34
7965573	NTN4 ^e	2.5853	0.0129	12q22-q23
8003298	SLC7A5 ^e	2.5767	0.0174	16q24.3
8156783	COL15A1 ^e	2.5452	0.0250	9q21-q22
8090565	SNORA7B	2.5214	0.0225	3q21.3
7954293	PDE3A	2.5063	0.0121	12p12
8137670	PDGFA ^{d e}	2.4802	0.0179	7p22
8049888	ATG4B	2.4474	0.0115	2q37.3
7974895	FLJ43390	2.4260	0.0427	14q23.2
7958262	TCP11L2	2.4172	0.0253	12q23.3
7989985	ITGA11 ^e	2.4038	0.0447	15q23
8104035	SORBS2	2.3912	0.0291	4q35.1
8102800	SLC7A11 ^e	2.3697	0.0390	4q28-q32
7912157	ERRFI1	2.3596	0.0308	1p36
7981728	—	2.3590	0.0407	—
7985493	TM6SF1	2.3535	0.0442	15q24-q26
7981962	SNORD116-5	2.3518	0.0419	15q11.2
7931977	ITIH5	2.3354	0.0486	10p14
7928308	DDIT4	2.3277	0.0467	10pter-q26.12
8034940	NOTCH3 ^e	2.3234	0.0396	19p13.2-p13.1
7966089	CMKLR1 ^e	2.2946	0.0455	12q24.1
8068024	JAM2 ^{d e}	2.2538	0.0397	21q21.2
8023152	TCEB3CL	2.2139	0.0407	18q21.1
8109159	LOC728264	2.2134	0.0459	5q33.1
8117018	JARID2	2.2051	0.0408	6p24-p23
8086752	RNU13P3	2.2046	0.0452	3p21.31
7982868	CHAC1	2.2041	0.0273	15q15.1
8139207	INHBA ^e	2.1839	0.0405	7p15-p13
8027352	—	2.1608	0.0399	—

Table 1 (continued)

AFFY ID	Gene symbol	FC ^a	Pfp ^b	Cytoband
8027002	GDF15 ^c	2.1594	0.0285	19p13.11
8048749	KCNE4 ^c	2.1436	0.0184	2q36.3
8007850	LRRC37A	2.0206	0.0244	17q21.31
8160138	NFIB	1.9790	0.0182	9p24.1
8131867	–	1.8907	0.0477	–
8064978	JAG1 ^{d c}	1.8529	0.0178	20p12.1-p11.23
8042788	ACTG2	1.7905	0.0414	2p13.1
8045804	–	1.7730	0.0492	–
8092726	CLDN1 ^d	1.4872	0.0445	3q28-q29
7937772	IGF2	–1.8634	0.0165	11p15.5
7915592	–	–1.8745	0.0195	–
8124391	HIST1H2AB	–2.0165	0.0426	6p21.3
8163202	SVEP1	–2.0970	0.0214	9q32
8037240	PSG1 ^c	–2.1467	0.0273	19q13.2
8044605	LOC654433	–2.2491	0.0468	–
8117594	HIST1H2BM	–2.2734	0.0128	6p22-p21.3
7951271	MMP1 ^{d c}	–2.3465	0.0026	11q22.3
8138888	PDE1C	–2.4348	0.0263	7p15.1-p14.3
8140668	SEMA3A ^c	–2.5040	0.0042	7p12.1
7951284	MMP3 ^{d c}	–2.5261	0.0433	11q22.3
8107044	ERAP2	–2.5407	0.0291	5q15
8003667	SERPINF1 ^c	–2.5688	0.0109	17p13.1
8110916	LOC442132	–2.5879	0.0258	5p15.31
7904293	PTGFRN ^c	–2.6619	0.0164	1p13.1
8113800	FBN2 ^c	–2.7034	0.0000	5q23-q31
8116418	GFPT2	–2.8180	0.0212	5q34-q35
8152617	HAS2 ^d	–2.8501	0.0038	8q24.12
8129573	MOXD1 ^c	–2.8809	0.0181	6q23.1-q23.3
7925929	AKR1C3	–3.1055	0.0165	10p15-p14
7917850	ARHGAP29	–3.1793	0.0031	1p22.1
8037251	PSG7 ^c	–3.2618	0.0006	19q13.2
8180291	–	–3.2801	0.0006	–
8165808	XG ^c	–3.6591	0.0065	Xp22.33
8037272	PSG5 // PSG5 ^c	–3.7775	0.0000	19q13.2 // 19q13.2
8083887	CLDN11 ^d	–4.2221	0.0000	3q26.2-q26.3
7909730	KCNK2 ^c	–4.7183	0.0000	1q41
8037283	PSG4 ^c	–4.9049	0.0000	19q13.2
8152522	ENPP2 ^{d c}	–5.3472	0.0000	8q24.1

^a FC = Fold change^b Pfp = estimated percentage of false positive predictions.^(c) Genes that were functionally categorized as glycoproteins by DAVID ($p=8,0E-6$).^(d) Genes involved in EMT.

40% FBS for storage. Frozen cells were thawed and grown until 80% confluency in a 75 cm² flask and submitted to serum starvation (12 h) prior to RNA extraction. All experiments were conducted with cells between the 4th and 8th subculture.

Characterization of Mesenchymal Stem Cell Populations

Cell populations used in the microarray assays were analyzed in a flow cytometer for specific cell surface markers to evaluate homogeneity. Cells were harvested

with TrypLe (Invitrogen, UK), washed with PBS, and incubated at 4°C for 30 min with the following antibodies: CD29-PE CY5, CD90 (Thy-1), CD45-FITC, CD31-PE (Becton Dickinson, USA), SH2, SH3, and SH4 (Case Western Reserve University, USA). After a second wash, samples incubated with unconjugated primary antibodies were then incubated with anti-mouse-PE secondary antibody (Guava Technologies, Hayward, CA) for an additional 15 min at 4°C. Finally, the cell suspension was washed with PBS, and signals from 10^5 cells were acquired with an EasyCyte Flow cytometer (Guava Technologies). Control samples for determining background noise were incubated with PBS instead of primary antibody followed by incubation with anti-mouse-PE secondary antibody. All plots generated were analyzed with Guava ExpressPlus software (Guava Technologies).

The *in vitro* differentiation into bone, cartilage, muscle and fat had previously been demonstrated in 2 of the NSCL/P patients (F4280, F4285) and 3 of the control (F3363, F4271, F4272) samples included in this study [28, 29]. Because of the high reproducibility of our protocols, stem cell cultures are currently characterized only with respect to their immunophenotype.

RNA Processing

Total RNA was isolated using TRIzol (Invitrogen, UK) according to the manufacturer's protocol and purified with RNeasy Mini-kit (QIAGEN). RNA quality and concentration were assessed using Nanodrop 1000 and gel electrophoresis. Only samples with a preserved rRNA ratio (28S/18S) and no evidence of RNA degradation were used in the microarray hybridization and qRT-PCR.

Microarray Processing

Expression measurements were performed using the Affymetrix Human Gene 1.0 ST array, which interrogates 28,869 genes, following RNA labeling and hybridization protocols as recommended by the manufacturer. After array scanning, quality control was performed with GCOS software (Affymetrix, USA) according to the manufacturer's recommendations.

Transcriptome Analysis

Gene expression values were obtained using the three-step Robust Multi-array Average (RMA) pre-processing method, implemented in the Affy package from R/Bioconductor [34].

We employed the RankProd method for the selection of differentially expressed genes (DEGs), considering a p-value cut-off of 0.05 adjusted for FDR (False Discovery Rate) [35]. RankProd is a rank-based nonparametric proce-

dure [36]. The method has the advantage of being able to deal with few samples for identifying biologically relevant expression changes [37]. Genes selected by RankProd do not necessarily need to present homogenous expression levels across all the samples of the test and control groups. Accordingly, RankProd seems to be a good choice for identifying differential gene expression in complex diseases, in which altered expression of a given candidate gene is expected in just a subgroup of patients due to both multi-locus genetic heterogeneity and the stochastic nature of gene expression in complex systems [38].

Functional Annotation

We performed functional annotation analysis of the differentially expressed genes with the DAVID (Database for Annotation, Visualization and Integrated Discovery, <http://david.abcc.ncifcrf.gov>) and IPA (Ingenuity Pathway Analysis, <http://www.ingenuity.com>) tools. In IPA, we considered the default parameter Molecules per Network=35, Networks per Analysis=25, only direct relationships between genes and the “Ingenuity Expert Information” Data Source, including the new option of “Ingenuity ExpertAssist Findings”, in which the information has been manually reviewed and curated from full-text scientific publications.

Validation of Microarray Expression Using Quantitative Real-Time PCR (qRT-PCR)

Initially we performed qRT-PCR for 4 genes (*COL15A1*, *ERAP2*, *PPT2* and *EGFL8*) on the same samples used in the microarray assay. These genes were randomly selected, but they were amongst those with the highest fold change. *PPT2* and *EGFL8* are represented by a common probe set on the Affymetrix Human Gene 1.0 ST array and therefore both genes were tested for qRT-PCR with gene-specific primers. Next, we performed qRT-PCR for further 12 genes (*ACAN*, *CDH6*, *CLDN1*, *CLDN11*, *COL4A1*, *COL4A2*, *ENPP2*, *HAS2*, *ITGA8*, *JAG1*, *MCAM* and *MMP3*) plus the genes *COL15A1*, *ERAP2* and *PPT2* in 16 controls and 13 NSCL/P patients. Only 4 of the control and 4 of the NSCL/P patients samples were the same as those used in the microarray assay due to unavailability of RNA from the remaining 4 individuals (Table 1 in supplementary material).

One microgram of total RNA from each cell culture was reverse transcribed with Superscript II (Invitrogen, UK), according to manufacturer's recommendations. Quantitative Real-Time PCR reactions were performed in duplicates with a final volume of 20 µl, using 20 ng of cDNA, 1X SYBR Green PCR Master Mix (*Applied Biosystems*) and 100–400 nM of each primer. We used

ABI Prism 7500 Sequence Detection System (*Applied Biosystems*) with standard temperature protocol. Primers were designed with Primer Express software V.2.0 (*Applied Biosystems*; primers sequence in supplementary Table 2) and the amplification efficiency (E) of each primer was calculated according to the equation $E=10^{(-1/\text{slope})}$. The expression data of the target transcripts were determined by relative quantification in comparison to a pool of RNAs (4 controls and 4 patients). GeNorm v3.4 was used to determine the most stable endogenous controls (*SDHA*, *HPRT1* and *GAPDH*) and to calculate the normalization factors for each sample [39]. Expression values were calculated according to reference [40].

To compare the expression of *COL15A1*, *ERAP2*, *PPT2* and *EGFL8*, obtained by qRT-PCR and microarray assay in the same samples, we used an unpaired t-test with Welch's correction.

To compare the results obtained by microarray with those obtained by qRT-PCR in the novel samples, for which we do not have microarray data, we performed the following strategy: for each of the 15 genes, we calculated the average (avg) expression for controls and for NSCL/P patients obtained by both methods (Table 3 in supplementary material). The correlation between the ratios “avg_patients/avg_controls” from microarray and qRT-PCR assays was calculated using Spearman's correlation test.

Clustering Analysis

The cluster analysis of the differentially expressed genes was performed using GEDI (Gene Expression Dynamics Inspector) software. This tool creates a 2 dimensional gene expression image for each sample in which each gene retains exactly the same position in the image of each sample and in which the gene positions are computed to give the most parsimonious gene arrangement for depicting expression level differences between the patient and control groups for all differentially expressed genes [41]. In the analysis, a 10×11 grid configuration of SOM (Self-Organizing Map) was used. Inspection of GEDI images allows a straightforward classification of the samples into subgroups without the aid of a clustering algorithm, but simply based on the visual differences in the patterns [42].

We also used two other conventional clustering methods: K-means and Hierarchical, both available in the MeV (MultiExperiment Viewer) software [43], with Spearman's correlation as the distance metric. The clustering analysis of qRT-PCR data followed two criteria: a) only DEGs from network 1 (Fig. 4) and b) only DEGs that showed the same tendency of expression in the qRT-PCR and microarray assays (Fig. 2 in supplementary material).

Results

Characterization of Mesenchymal Stem Cell Populations

The cell cultures presented positive labeling for cell adhesion (CD29, CD90) and mesenchymal stem cell markers (SH2, SH3, SH4) in most of the cells (>90%) and were negative for endothelial and hematopoietic cell markers (Fig. 2 in supplementary material). Moreover, 2 NSCL/P patients and 3 control cell cultures used in the microarray analyses had been previously shown to differentiate into bone, muscle, cartilage and fat upon *in vitro* induction [28, 29]. Therefore the cell populations used in this study had the main properties of stem cells.

Controls versus NSCL/P Patients

We identified a total of 87 differentially expressed genes (DEGs; 58 upregulated and 29 downregulated) in the comparison between controls and NSCL/P patients (adjusted $p \leq 0.05$; Table 1).

Next, in order to visualize the expression behavior of these DEGs, we performed clustering analysis with the GEDI software. An image was created reflecting the 87 DEGs' transcriptional behavior for each individual and where the gene position was fixed to give the most parsimonious arrangement to show differential gene expression between controls and NSCL/P patients. Upon visual inspection of the GEDI images, we observed that 4 of the NSCL/P patients showed a similar expression pattern (F4280, F4281, F4282, F4283; Fig. 1).

The clustering analysis with the k-means method resulted in 9 gene clusters. Four of them exhibited a similar expression profile among NSCL/P patients, most particularly in the afore-mentioned group (F4280, F4281, F4282, F4283; Fig. 2). The similarities and dissimilarities in expression levels observed between NSCL/P patients were similar for both clustering methods. The expression pattern found in 4 out of 6 NSCL/P patients illustrates the characteristic of RankProd of being capable of selecting genes with differential expression in just a subgroup of samples.

Functional Annotation

Differentially expressed genes were functionally annotated and analyzed with two different tools. First, the DAVID tool led to identification of three main canonical pathways from the KEGG Database: Focal adhesion, Cell adhesion molecules and ECM-receptor interaction (Fig. 3, 4 and 5 in supplementary material). Moreover, the most relevant functional category identified through DAVID was that of Glycoproteins ($p=8.0E-6$), which included 36 of the 87

Fig. 1 Clustering of 87 DEGs resulted from the comparison between 6 controls and 6 NSCL/P patients. Each GEDI map (or mosaic) represents a gene expression profile of a single individual. The blue color

represents the lowest expression level and red color represents the highest expression level on a scale of -4.70 to 7.98 , respectively. The black frame highlights four patients with similar gene expression profile

DEGs (Table 1). Subsequently, the IPA tool was used to characterize the 87 DEGs regarding possible biological functions. We observed 3 relevant functions enriched with a significant number of our genes: Cellular movement (20 genes, $p=6.4E-06$ – $2.76E-02$), Cellular growth and proliferation (27 genes, $p=3.11E-05$ – $2.68E-02$) and Cellular development (27 genes, $p=3.3E-05$ – $2.47E-02$) (Table 4 in supplementary material). A putative network with the largest number of DEGs built by IPA (13 DEGs; Fig. 4) suggests functional relationship among several extracellular proteins: *ACAN*, *COL4A1*, *COL4A2*, *GDF15*, *IGF2*, *MMP1*, *MMP3* and *PDGFA*. All of these 8 genes are DEGs.

Validation of the Microarray Analysis

The reproducibility of gene expression assayed by Affymetrix microarrays is high [44] and 4 genes (*COL15A1*, *ERAP2*, *PPT2* and *EGFL8*) were initially selected for validation through qRT-PCR. Except for *ERAP2* ($p=0.0397$), that showed lower expression levels, the other genes showed higher transcript levels in NSCL/P patients than in controls: *PPT2* ($p=0.0087$) and *COL15A1* ($p=0.0355$) (Fig. 3), which confirms the expression patterns observed in the microarray assays. *EGFL8* is represented by the same probe set of *PPT2*. Considering that we did not find significant differential expression levels through qRT-PCR with specific primers

Fig. 2 Gene clusters 1, 4, 6 and 9 resulted from k-means method ($k=9$). In both clusters it is possible to observe a similar gene expression profile among 4 out of 6 patients (F4280, F4281, F4282

and F4283), indicating that many of the 87 selected DEGs are co-regulated in these 4 NSCL/P patients

Fig. 3 Quantitative Real-Time PCR (qRT-PCR) initial analysis of NSCL/P patients and control samples for *ERAP2*, *PPT2*, *COL15A1*. E = primer amplification efficiency; CT = cycle threshold; delta_CT = sample's average_CT normalized by pool's average_CT; NF = normalization factor

for *EGFL8* ($p=0.1781$), we kept only *PPT2* among our candidate genes.

For further validation of our results, in addition to *COL15A1*, *PPT2* and *ERAP2*, we analyzed a further 12 other genes known to be involved with matrix remodeling

in a novel sample of individuals (16 controls and 13 NSCL/P patients). By comparing the ratios of the average expression from patients/controls, we observed that *ENPP2* exhibited the highest discordance between the microarray and qRT-PCR ratios (0.645 and 4.324, respectively), therefore we considered this gene expression as not validated. However, we observed a significant positive correlation between microarray and qRT-PCR expression data ($p=0.0114$, $r=0.653$) for the 14 genes chosen for validation (Table 3 in supplementary material). Accordingly, the analyses thus show that the data obtained from microarrays and qRT-PCR are consistent with each other, even in an enlarged novel sample, attesting to the reliability of the microarray analysis.

Using NSCL/P patients expression data from qRT-PCR (13 NSCL/P patients) and microarray assays (6 NSCL/P patients), we also performed a clustering analysis for the following DEGs: *ACAN*, *COL4A1*, *COL4A2*, *ERAP2*, *HAS2*, and *MMP3*. These genes belong to network 1 (Fig. 4) and exhibited the same expression tendency in both experiments (Fig. 2 in supplementary material). The hierarchical clustering performed with qRT-PCR data

Fig. 4 The most significant network built by IPA tool with the highest number of differentially expressed genes. The upregulated genes in NSCL/P patients are indicated in red and the downregulated genes in green. The blank symbols pertain to genes that were either not present in our array or not differentially expressed. Solid lines indicate a direct linkage among two genes. Lines with arrows indicate that one gene acts on the other, and lines without arrows indicate that the corresponding proteins interact with each other. The 6 genes circled in orange were used in clustering analysis of qRT-PCR and microarray

NETWORK 1

Fig. 5 Hierarchical clustering of all the patients analyzed by qRT-PCR and microarray, considering only the genes from IPA network 1 (Fig. 4). **a** Clustering of expression values obtained by qRT-PCR. **b** Clustering of expression values obtained by microarray. **c** Correlations (Spearman's correlation test, r and p -values) between each co-regulated gene from qRT-PCR clustering. It is possible to observe that the gene *MMP3* is significantly and inversely correlated to *ACAN*, *COL4A1*, *COL4A2* genes in a subgroup of patients from qRT-PCR. This same pattern of co-regulation is also present in another group of patients analyzed by microarray, which includes two individuals (F4280 and F4281) from the mentioned qRT-PCR subgroup. In controls expression data of both assays, *MMP3* is upregulated and *ACAN*, *COL4A1* and *COL4A2* downregulated

revealed a homogeneous cluster with 7 out of 13 NSCL/P patients (F4312, F4311, F4281, F4243, F4388, F5398 and F4280), in which *MMP3* is downregulated and *ACAN*, *COL4A1* and *COL4A2* upregulated. *ERAP2* and *HAS2* did not exhibit a consistent expression pattern in these patients (Fig. 5a). Interestingly, the hierarchical clustering using microarray data showed the same transcriptional behavior among these genes in 4 out of 6 NSCL/P patients (F4280, F4281, F4282 and F4283), although in this case, *ERAP2* and *HAS2* seem to be co-regulated with *MMP3* and collagens (Fig. 5b). Spearman's correlation test (Fig. 5c) calculated for each relationship among the genes from Fig. 5a confirmed that *MMP3* and *ACAN* are negatively correlated ($r=-0.889$; $p<0.05$) while *ACAN* exhibited positive correlation with *COL4A1* and *COL4A2* ($r=0.921$ and $r=0.872$, respectively; $p<0.05$), corroborating our findings thus far. On the other hand, we did not achieve significance for *ERAP2* and *HAS2*, confirming the lack of a well-defined expression pattern for these genes, as observed in Fig. 5a.

Discussion

In this study, we conducted the first comparative transcriptome analysis between dental pulp stem cells from NSCL/P patients and controls to obtain more information on the genetic etiology of this malformation and explore new possibilities to identify pathways associated with disease pathology.

Genome expression microarray analysis is a powerful tool for assessing changes in the transcription patterns of related genes and identification of signaling pathways associated with specific cell types, culture conditions or disease states [45, 46]. Considering that the cell populations from NSCL/P patients and controls were established and maintained under similar identical conditions, it is very likely that a large proportion of the DEGs identified is related to the genetic constitutional differences between cells from NSCL/P patients and controls. The observations that NSCL/P disease-specific expression profiles have also been previously reported in tissue biopsies and fibroblast cultures [19, 21] suggests that such profiles may be ubiquitous; in this respect our findings suggest that the disease-specific transcription profile is also present in mesenchymal stem cells. Accordingly, transcriptome analysis of stem cells represents an additional option to the study of the transcriptome and genetics of NSCL/P.

Of the 87 DEGs obtained in our microarray analysis, we noted that *MMP3* had previously been proposed as a candidate gene for NSCL/P [47]. A further 2 genes, *PDGFA* and *ERRFI1*, had previously been indirectly associated with NSCL/P, since their receptors, respectively *PDGFR* and *EGFR*, were identified as predisposing loci for this form of clefting [48–50]. These observations provide further confirmation that the transcriptome of DPSCs from NSCL/P patients can also be used to identify genetic variations associated with the disease.

The functional annotation through DAVID showed that nearly half of the 87 DEGs correspond to glycoprotein molecules, including collagens, metalloproteinases, integrins, and adhesion proteins, which are important to orchestrate the signaling between the extracellular and intracellular compartment. Indeed, the three canonical pathways we identified through DAVID are mainly related to extracellular matrix components and signaling molecules located on the cell membrane (Fig. 3, 4 and 5 in supplementary material). Functional relationships among several extracellular proteins that are deregulated in our studies are also suggested by the putative network built by IPA (Fig. 4). These analyses suggest that a large proportion of the DEGs in the DPSC from NSCL/P patients are extracellular matrix components (ECM) involved in several cellular processes in facial development, such as extracellular remodeling during the epithelial-mesenchymal transition (EMT).

EMT is a fundamental mechanism behind palatal fusion. This process occurs through a regulated sequence of events determined both by the extracellular environment and the gene expression program of the cell, leading to loss of cell-cell adhesion, breakdown of basal laminae, and increased cell invasion and mobility [51, 52].

Fifteen genes enrolled in EMT are within the 87 DEGs (Table 1). Further, we observed the enrichment of biological functions involving cell proliferation, movement and invasion (Table 4 in supplementary material), all of which are expected phenotypic outcomes for genes involved with EMT.

In the most relevant IPA network we observed that 8 out of 13 DEGs (*ACAN*, *COL4A1*, *COL4A2*, *GDF15*, *IGF2*, *MMP1*, *MMP3* and *PDGFA*, Fig. 4) are extracellular matrix components, suggesting that an abnormal expression behavior of these genes may affect EMT during palate development.

Clustering analyses showed that *MMP3*, *ACAN*, *COL4A1* and *COL4A2* transcripts are co-regulated in 4 out of 6 NSCL/P patients analyzed by microarray as well as in 7 out of 13 NSCL/P patients from qRT-PCR analysis. Therefore, it seems that in NSCL/P mesenchymal cells, the down-regulation of *MMP3* is associated with up-regulation of *ACAN*, *COL4A1* and *COL4A2*. Such a deregulated pathway probably has serious consequences in embryonic development, since it is known that the *MMP3* protein cleaves a wide range of ECM proteins, including the collagens IV, V, IX, X, proteoglycans, fibronectin and laminin. It is also capable of activating other MMPs, as shown in network 1 (Fig. 4), and playing a key role in ECM degradation and remodeling [47]. In this context it has already been experimentally shown that lower levels of MMPs can block palatal fusion [7]. Therefore, it is possible that failure of lip and palate fusion in these groups of patients is at least partly associated with down-regulation of MMPs and up-regulation of collagens. It will be important in the near future to identify the leading causative mechanism of altered expression of MMPs in these individuals. These results also show the robustness of RankProd to detect DEGs on a limited and heterogeneous group of samples, in contradistinction to a popular method SAM [53] which appears to require larger sample sizes. Moreover, RankProd is able to identify expression patterns in just a subgroup of affected samples, which is the ideal, considering that this is the expected to occur in a complex disease such as NSCL/P. Notwithstanding our favorable impression of RankProd, we are acutely aware of the small sample sizes employed in detecting DEGs in the initial microarray assay. Future studies must be directed towards confirming and/or expanding the pattern of DEGs using novel and much larger sample sizes.

In summary our results suggest that NSCL/P patients exhibit deregulation of genes participating in either EMT or embryonic processes that depend on extracellular modeling. Abnormal expression of some genes encoding for extracellular matrix proteins in NSCL/P fibroblasts has also been reported by others, reinforcing the concept that disease expression signatures for NSCL/P are present in various tissues and not necessarily confined to a specific cell type. Moreover, the penetrance of the phenotype can depend on exposure to environmental factors and it is of interest that a recent report claimed a positive association between nicotine exposure and the CL/P phenotype in conjunction with deregulation of gene expression involved in ECM synthesis and degradation [21].

The observation that the DEGs in NSCL/P patients are associated with ECM component signaling suggests that further analysis of these cells presents unique opportunities to study the complexity of molecular pathways and alleles involved in NSCL/P etiology. However, arguably the major advantage of DPSCs above other cell types, such as fibroblasts, will be their potential to study *in vitro* differentiation into muscle, bone and cartilage that are affected tissues in NSCL/P. In this context, it will be possible to integrate genomic and transcriptome analysis under different experimentally induced environmental insults on identical cell cultures. Our results open new avenues for the study of novel candidate genes for NSCL/P, since most of the 87 DEGs have not been previously associated with NSCL/P. In particular, the potential offered by our approach can be best visualized in the gene network 1 (Fig. 4), in which several of our DEGs are ECM components, suggesting that these genes might be enrolled in EMT during the development of NSCL/P phenotype.

If wide-spread differences in co-regulated gene transcription, as observed in our experiments, are indeed a primary cause of NSCL/P, then the research emphasis must move away from naively cataloging which genes are being differentially expressed to defining the central transcription factors and regulatory elements that are driving the disease specific transcription patterns. In this respect, identifying the putative gene networks involved, as in our observations, will be an initial crucial step towards defining which regulatory elements are the most important.

Acknowledgements The authors would like to express their gratitude to: Mrs. Constancia Urbani for secretarial assistance, and the following colleagues: Rodrigo Atique, BSc, for helping us with figures; Eder Zucconi, PhD, and Natassia Vieira, BSc, for helping us with flow cytometry; Andrea Sertié, PhD, and Keith Okamoto, PhD for their helpful reviews and suggestions regarding the manuscript. FAPESP/CEPID, CNPq, MCT, FINEP sponsored this study.

Disclosures The authors indicate no potential conflicts of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Gundlach, K. K. H., & Maus, C. (2006). Epidemiological studies on the frequency of clefts in Europe and world-wide. *Journal of Cranio-Maxillo-Facial Surgery*, 34(2), 1–2.
- Kerrigan, J. J., Mansell, J. P., Sengupta, A., Brown, N., & Sandy, J. R. (2000). Palatogenesis and potential mechanisms for clefting. *Journal of the Royal College of Surgeons of Edinburgh*, 45(6), 351–358.
- Jiang, R., Bush, J. O., & Lidral, A. C. (2006). Development of the upper lip: morphogenetic and molecular mechanisms. *Developmental Dynamics*, 235(5), 1152–1166.
- Jugessur, A., Farlie, P. G., & Kilpatrick, N. (2009). The genetics of isolated orofacial clefts: from genotypes to subphenotypes. *Oral Diseases*, 15(7), 437–453.
- Suzuki, K., Hu, D., Bustos, T., et al. (2000). Mutations of PVRL1, encoding a cell-cell adhesion molecule/herpesvirus receptor, in cleft lip/palate-ectodermal dysplasia. *Nature Genetics*, 25(4), 427–430.
- Acloque, H., Adams, M. S., Fishwick, K., Bronner-Fraser, M., & Nieto, M. A. (2009). Epithelial-mesenchymal transitions: the importance of changing cell state in development and disease. *The Journal of Clinical Investigation*, 119(6), 1438–1449.
- Yu, W., Ruest, L. B., & Svoboda, K. K. (2009). Regulation of epithelial-mesenchymal transition in palatal fusion. *Experimental Biology and Medicine*, 234(5), 483–491.
- Kang, P., & Svoboda, K. K. (2005). Epithelial-mesenchymal transformation during craniofacial development. *Journal of Dental Research*, 84(8), 678–690.
- Juriloff, D. M., & Harris, M. J. (2008). Mouse genetic models of cleft lip with or without cleft palate. *Birth Defects Research. Part A: Clinical and Molecular Teratology*, 82(2), 63–77.
- Mangold, E., Ludwig, K. U., Birnbaum, S., et al. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, 42(1), 24–26.
- Mangold, E., Reutter, H., Birnbaum, S., et al. (2009). Genome-wide linkage scan of nonsyndromic orofacial clefting in 91 families of central European origin. *American Journal of Medical Genetics. Part A*, 149A(12), 2680–2694.
- Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Steffens, M., Rubini, M., et al. (2009). Key susceptibility locus for non-syndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature Genetics*, 41, 473–477.
- Marazita, M. L., Murray, J. C., Lidral, A. C., Arcos-Burgos, M., Cooper, M. E., Goldstein, T., et al. (2004). Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32–35. *American Journal of Human Genetics*, 75, 161–173.
- Cai, J., Chen, J., Liu, Y., et al. (2006). Assessing self-renewal and differentiation in human embryonic stem cell lines. *Stem Cells*, 24(3), 516–530.
- Gong, S. G., Gong, T. W., & Shum, L. (2005). Identification of markers of the midface. *Journal of Dental Research*, 84(1), 69–72.
- Brown, N. L., Knott, L., Halligan, E., Yarram, S. J., Mansell, J. P., & Sandy, J. R. (2003). Microarray analysis of murine palatogenesis: temporal expression of genes during normal palate development. *Development, Growth & Differentiation*, 45(2), 153–165.
- Nikopensius, T., Birnbaum, S., Ludwig, K. U., et al. (2010). Susceptibility locus for non-syndromic cleft lip with or without cleft palate on chromosome 10q25 confers risk in Estonian patients. *European Journal of Oral Sciences*, 118(3), 317–319.
- Beatty, T. H., Murray, J. C., Marazita, M. L., et al. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genetics*, 42(6), 525–529. Erratum in: *Nature Genetics*, 42(8), 727, 2010.
- Jakobsen, L. P., Borup, R., Vestergaard, J., et al. (2009). Expression analyses of human cleft palate tissue suggest a role for osteopontin and immune related factors in palatal development. *Experimental & Molecular Medicine*, 41, 77–85.
- Brown, N. L., Yarram, S. J., Mansell, J. P., & Sandy, J. R. (2002). Matrix metalloproteinases have a role in palatogenesis. *Journal of Dental Research*, 81, 826–830.
- Baroni, T., Bellucci, C., Lilli, C., et al. (2010). Human cleft lip and palate fibroblasts and normal nicotine-treated fibroblasts show altered in vitro expressions of genes related to molecular signaling pathways and extracellular matrix metabolism. *Journal of Cellular Physiology*, 222(3), 748–756.
- Gunaseeli, I., Doss, M. X., Antzelevitch, C., Hescheler, J., & Sachinidis, A. (2010). Induced pluripotent stem cells as a model for accelerated patient- and disease-specific drug discovery. *Current Medicinal Chemistry*, 17(8), 759–766.
- Masotti, C., Ormelas, C., Splendore-Gordonos, A., Moura, R., Felix, T., Nivaldo, A., et al. (2009). Reduced transcription of TCOF1 in adult cells of Treacher Collins syndrome patients. *BMC Medical Genetics*, 10(1), 136.
- Tsai, M.-S., Hwang, S.-M., Chen, K.-D., Lee, Y.-S., et al. (2007). Functional network analysis of the transcriptomes of mesenchymal stem cells derived from amniotic fluid, amniotic membrane, cord blood, and bone marrow. *Stem Cells*, 25(10), 2511–2523.
- Zucconi, E., Vieira, N. M., Bueno, D. F., et al. (2010). Mesenchymal stem cells derived from canine umbilical cord vein—a novel source for cell therapy studies. *Stem Cells and Development*, 19(3), 395–402.
- Bueno, D. F., Kerkis, I., Costa, A. M., et al. (2009). New source of muscle-derived stem cells with potential for alveolar bone reconstruction in cleft lip and/or palate patients. *Tissue Engineering*, 15, 427–435.
- Vieira, N., Bueno, C. R., Jr., Brandalise, V., et al. (2008). Dystrophic mice express a significant amount of human muscle proteins following systemic delivery of human adipose-derived stromal cells without immunosuppression. *Stem Cells*, 26(9), 2391–2398.
- Costa, A. M., Bueno, D. F., Martins, M. T., et al. (2008). Reconstruction of large cranial defects in nonimmunosuppressed experimental design with human dental pulp stem cells. *The Journal of Craniofacial Surgery*, 19, 204–210.
- Bueno, D. F. (2007). The use of adult stem cells to study the etiopathogeny of cleft lip and palate and tissue engineering. Phd in human genetics, Institute of Biosciences, University of Sao Paulo (USP), Rua do Matao, trav. 14, n321—Cidade Universitaria—Sao Paulo—Brazil.
- Gronthos, S., Braham, J., Li, W., Fisher, L. W., et al. (2002). Stem cell properties of human dental pulp stem cells. *Journal of Dental Research*, 81(8), 531–535.
- Secco, M., Moreira, Y., Zucconi, E., et al. (2009). Gene expression profile of mesenchymal stem cells from paired umbilical cord units: Cord is different from blood. *Stem Cell Reviews and Reports*. Epub ahead of print.
- Widera, D., Zander, C., Heidbreder, M., et al. (2009). Adult palatum as a novel source of neural crest-related stem cells. *Stem Cells*, 27(8), 1899–1910.

33. Grenier, J., Teillet, M. A., Grifone, R., Kelly, R. G., & Duprez, D. (2009). Relationship between neural crest cells and cranial mesoderm during head muscle development. *PLoS ONE*, 4(2), e4381.
34. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4), e15.
35. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.
36. Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., & Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22), 2825–2827.
37. Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573, 83–92.
38. Raj, A., Rifkin, S. A., Andersen, E., & van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283), 913–918.
39. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, 3(7), research0034.1–0034.11.
40. Pfaffl, M. W. (2001). A new mathematical model for quantification in real-time-RT-PCR. *Nucleic Acids Research*, 9, 2002–2007.
41. Eichler, G. S., Huang, S., & Ingber, D. E. (2003). Gene expression dynamics inspector (GEDI): for integrative analysis of expression profiles. *Bioinformatics*, 19, 2321–2322.
42. Guo, Y., Eichler, G. S., Feng, Y., Ingber, D. E., & Huang, S. (2006). Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers. *Journal of Biomedicine and Biotechnology*, 69141.
43. Saeed, A. I., Sharov, V., White, J., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2), 374–378.
44. Bochukova, E., Soneji, S., Wall, S., & Wilkie, A. O. (2009). Scalp fibroblasts have a shared expression profile in monogenic craniosynostosis. *Journal of Medical Genetics*, [Epub ahead of print].
45. LaGamba, D., Nawshad, A., & Hay, E. D. (2005). Microarray analysis of gene expression during epithelial-mesenchymal transformation. *Developmental Dynamics*, 234(1), 132–142.
46. Fanganiello, R. D., Sertié, A. L., Reis, E. M., et al. (2007). Apert pSer252Trp mutation in FGFR2 alters osteogenic potential and gene expression of cranial periosteal cells. *Molecular Medicine*, 13(7,8), 422–442.
47. Letra, A., Silva, R. A., Menezes, R., et al. (2007). MMP gene polymorphisms as contributors for cleft lip/palate: association with MMP3 but not MMP1. *Archives of Oral Biology*, 52(10), 954–960.
48. Xu, X., Bringas, P., Jr., Soriano, P., & Chai, Y. (2005). PDGFR- α signaling is critical for tooth cusp and palate morphogenesis. *Developmental Dynamics*, 232(1), 75–84.
49. Ferby, I., Reschke, M., Kudlacek, O., et al. (2006). Mig6 is a negative regulator of EGF receptor-mediated skin morphogenesis and tumor formation. *Nature Medicine*, 12, 568–573.
50. Miettinen, P. J., Chin, J. R., Shum, L., Slavkin, H. C., Shuler, C. F., Derynck, R., et al. (1999). Epidermal growth factor receptor function is necessary for normal craniofacial development and palate closure. *Nature Genetics*, 22, 69–73.
51. Yu, W., Ruest, L. B., & Svoboda, K. K. (2009). Regulation of epithelial-mesenchymal transition in palatal fusion. *Experimental Biology and Medicine*, 234(5), 483–491.
52. Khudyakov, J., & Bronner-Fraser, M. (2009). Comprehensive spatiotemporal analysis of early chick neural crest network genes. *Developmental Dynamics*, 238(3), 716–723.
53. Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), 5116–5121.

Capítulo II

Busca por genes alvos do T com o programa hunT

Capítulo II

Busca por genes alvos do *T* com o programa hun*T*

Resumo

T ou *Brachyury* é um fator de transcrição da família T-box. Ele é fundamental durante o desenvolvimento embrionário, com papel chave na formação do mesoderma e da notocorda de embriões de vertebrados. A ausência completa do *T* leva à morte prematura do embrião de camundongo, por volta do 11º dia de gestação. Embora muito tem sido feito para caracterizar o *T* pouco se sabe dos genes que ele regula. Neste capítulo nós apresentamos o programa hun*T*, que foi desenvolvido para buscar por alvos do *T* através da identificação do sítio de ligação em suas regiões *cis*-reguladoras. Diferente dos programas existentes, ele permite a busca simultânea de sítios de ligação para diferentes fatores de transcrição e permite proteger trechos de um sítio da ocorrência de *mismatches*. Nossos testes revelaram que estas funcionalidades tornam os resultados do hun*T* mais precisos se comparados com os resultados do programa similar fuzznuc do pacote EMBOSS. O hun*T* identificou 4.602 genes de camundongo com o sítio de ligação para o domínio T-box. São genes que podem ser alvo de qualquer um dos membros desta família. Dentre eles, estão o *Dll1* e *Msgn1* que foram demonstrados experimentalmente como sendo alvos do *Tbx6* (Hoffman *et al.* 2004 e Wittler *et al.*, 2007) e vários outros genes que recentemente vem sendo encontrados através das novas tecnologias baseadas na imunoprecipitação de cromatina. Para melhorar a confiabilidade das nossas predições, integramos os resultados do hun*T* com os dados de expressão gênica de um estudo onde avaliamos o transcriptoma da diferenciação *in vitro* de células tronco embrionárias de camundongo para mesoderma em uma linhagem de células com a presença normal do *T* e outra onde o *T* foi reprimido por um RNAi. Os dados do transcriptoma ajudaram a apontar quais dos 4.602 genes podem ser alvos específicos do *T*. Encontramos 32 deles com a expressão alterada após 96h de diferenciação somente nas células normais e não nas células onde o *T* foi reprimido ($\geq 1,5$ fold, FDR<0,05). Este período é quando é iniciada a formação do mesoderma lateral, sugerindo a participação destes genes neste processo sendo regulados pelo *T*.

Capítulo II

Busca por genes alvos do T com o programa hunT

2.1 Introdução

Neste capítulo apresentamos o programa hunT, que foi desenvolvido em colaboração com o grupo do Prof. Dr. Bernhard Herrmann do Max Planck *Institute for Molecular Genetics* de Berlim-Alemanha, para buscar por genes alvos de um fator de transcrição chamado T , que tem papel chave no desenvolvimento de vertebrados. Iniciamos o capítulo com uma breve introdução da regulação da expressão gênica e da tarefa de buscar por sítios de ligação em sequências genômicas. Em seguida, descrevemos a motivação para a criação do hunT, o funcionamento do programa, os resultados e discussões e, por último, as conclusões.

2.1.1 Sítios de ligação na regulação da expressão gênica

Quase todas as células de um organismo contêm o mesmo genoma. A informação contida no genoma é a base para a existência de todas as linhagens celulares. A regulação da expressão gênica é o mecanismo responsável pela distinção da ampla variedade de tipos celulares existentes.

Durante o desenvolvimento de mamíferos, a expressão gênica é espacial e temporalmente coordenada para garantir a adequada especificação e maturação celular. Esta coordenação é feita, em parte, pelos fatores de transcrição (FTs), que são reguladores essenciais da expressão gênica. FTs se ligam a elementos reguladores de genes alvos, os chamados sítios de ligação, para modular diretamente suas atividades. O trecho da sequência de DNA do FT que se liga a estes sítios é chamado de domínio de ligação (Figura 2.1).

Figura 2.1: Representação de um sítio de ligação.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Um sítio de ligação para um fator de transcrição consiste numa pequena sequência de tipicamente 8-10 a 16-20 nucleotídeos. Pode ser representado por uma sequência consenso, que reúne os nucleotídeos mais frequentes de cada posição de um alinhamento (Figura 2.2A). Sequências consenso podem ser representadas através de uma expressão regular¹⁵ (Figura 2.2B), símbolos da nomenclatura IUPAC¹⁶ (Figura 2.2C) ou, ainda, através de uma sequência logo¹⁷. Uma sequência logo pode ser criada a partir da contagem dos nucleotídeos de cada posição (Figura 2.2D) ou do cálculo de uma matriz de peso escore-posição (PSSMs, Figura 2.2E), cujas equações podem ser vistas no trabalho de Wasserman & Sandelin, 2004.

¹⁵ Em ciência da computação, uma expressão regular é uma composição de símbolos e caracteres com funções especiais que, combinados com caracteres literais, formam uma sequência, uma expressão. Essa expressão é interpretada como uma regra e serve para identificar cadeias de caracteres de interesse em texto.

¹⁶ do inglês: *International Union of Pure and Applied Chemistry*. É uma organização reconhecida no desenvolvimento de padrões para a denominação de compostos químicos (NC-IUB, 1986).

¹⁷ Em bioinformática, uma sequência logo é uma representação gráfica da conservação de sequências de nucleotídeos ou aminoácidos (Schneider & Stephens, 1990).

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Figura 2.2: Representação de um sítio de ligação. A) Alinhamento da sequência-sítio de uma coleção de genes regulados por um mesmo fator de transcrição. B) Sequência consenso representada através de uma expressão regular, onde [CT] indica que esta posição pode ser ocupada por um C ou um T. C) Sequência consenso representada através de símbolos IUPAC, onde Y=C ou T e H=A, C ou T. Por convenção, um único nucleotídeo é mostrado se ele ocorre em mais da metade dos sítios e, no mínimo, duas vezes a ocorrência do segundo nucleotídeo mais frequente. Um símbolo IUPAC duplo como o Y é usado se os dois nucleotídeos ocorrerem em mais que 75% dos sítios e um símbolo triplo como o H quando um dos nucleotídeos não ocorrem D) Sequência logo criada a partir do cálculo da matriz PSSMs (do inglês *Position-Specific Scoring Matrices*), cuja equação está descrita em Wasserman & Sandelin, 2004. Imagem adaptada do trabalho de D'haeseleer, 2006.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

A identificação de elementos regulatórios em sequências de nucleotídeos, como os sítios de ligação, é um desafio tanto para a biologia como para a bioinformática.

Do ponto de vista biológico, a identificação de FTs e alvos pode oferecer meios para interpretar e modelar a resposta de células a diversos estímulos. É o caminho para a compreensão dos mecanismos moleculares que levam à doenças e problemas no desenvolvimento.

Para a bioinformática, o desafio está no desenvolvimento de algoritmos que sejam capazes de extrair informações biologicamente significantes da enorme quantidade de dados que vem sendo gerada por tecnologias em larga escala como de sequenciamento de DNA e as mais recentes baseadas na imunoprecipitação de cromatina¹⁸, como ChIP-chip¹⁹ e ChIP-seq²⁰.

O problema computacional da identificação de sítios de ligação, também chamados de padrões ou motivos, pode ser dividido em dois subproblemas distintos. No primeiro, dado um conjunto de sequências de genes, geralmente co-regulados, o objetivo é identificar um padrão conservado entre estas sequências, que seja alvo de um FT que supostamente regula todos estes genes (problema da descoberta de padrões). No segundo, e menos comum, a sequência do domínio de ligação do FT é conhecida e o objetivo é buscar os genes que possuem o sítio de ligação para este domínio sendo, portanto, possíveis alvos do FT conhecido (problema da busca por padrões) (Stormo, 2000). O programa que apresentamos neste capítulo refere-se a esse último problema.

Há uma série de programas disponíveis para a busca por padrões em sequências. O *fuzznuc* (Rice *et al.*, 2000), *tacg* (Mangalam, 2002), *PatSearch* (Grillo *et al.*, 2003) e *PatMatch* (Yan *et al.*, 2005) são alguns exemplos. Em todos eles, a entrada é um conjunto de regiões regulatórias de genes. A principal fonte para a criação dessa entrada são os resultados de sequenciamento de genomas completos, mas buscas mais específicas podem ser feitas nos resultados de experimentos de expressão gênica. Os algoritmos devem então percorrer milhares de As, Cs, Gs e Ts atrás de pequenas sequências, o que representa um problema para os métodos de busca. Quanto menor o tamanho do sítio, maior a chance de detectá-los ao acaso. Além disso, estas sequências geralmente variam bastante.

¹⁸ Técnica que permite a identificação das interações DNA-proteína. Também chamada de ChIP.

¹⁹ ChIP combinada com a tecnologia de *microarrays* (ChIP-chip).

²⁰ ChIP combinada com as novas tecnologias de sequenciamento (ChIP-seq).

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Por exemplo, dois dos conhecidos sítios de ligação para o *MEF2*, são conservados em somente 7 de 14 posições (Wasserman & Sandelin, 2004).

A predição de erros pode ser minimizada com a combinação de informações de conservação evolutiva e resultados de experimentos como de expressão gênica, por exemplo.

2.2 Motivação

T ou *Brachyury* é um fator de transcrição, membro fundador da família T-box. Ele é fundamental durante o desenvolvimento embrionário, com papel chave na formação do mesoderma e da notocorda de embriões de vertebrados (Herrmann *et al.*, 1990; Wilkinson *et al.*, 1990). Sua falta resulta na morte prematura do embrião de camundongo aproximadamente no 11º dia de gestação (Yanagisawa *et al.*, 1981). Vários trabalhos têm mostrado que o padrão da expressão do *T* em diferentes organismos é conservado e que estes genes desempenham papel similar no desenvolvimento, indicando que a função do *T* tem sido conservada ao longo da evolução (Herrmann & Kispert, 1994; Conlon *et al.*, 1996; Schulte-Merker *et al.*, 1994).

Na tentativa de entender como o *T* exerce seu efeito, nós desenvolvemos o programa hunT para buscar por genes alvos deste fator de transcrição através da identificação de sítios de ligação em suas sequências *cis*-reguladoras.

T é foco de estudo do grupo do Prof. Dr. Bernhard Herrmann. Em 1990, o gene foi clonado pela primeira vez em camundongo (Herrmann *et al.*, 1990). Em 1993, o grupo mostrou através de um procedimento baseado em PCR, que o *T* se liga a uma sequência parcialmente palindrômica de 20 nucleotídeos: 5'-T[GC]²¹ACACCTAGGTGTGAAATT-3' (Figura 2.3C, Kispert & Herrmann, 1993). Em 1995, mostraram que a presença de múltiplos sítios de ligação para o *T* aumenta o nível de ativação de genes alvos (Kispert *et al.*, 1995). Em 1997, a estrutura do domínio de ligação do *T* foi mostrada através de cristalografia de raio-X em *Xenopus laevis* (Muller & Herrmann, 1997). Este domínio, então chamado de T-box, passou a servir como critério para a admissão de novos membros à família (Smith, 1997). Estudos recentes têm corroborado estes achados através da identificação de dois genes, o *Dll1* e o *Msgn1*, ambos alvos do *Tbx6* (que é outro FT da família T-box), contendo vários sítios de ligação para o domínio T-box distribuídos ao longo de suas regiões *cis*-regulatórias (Hoffman *et al.*, 2004; Wittler

²¹ [GC] = G ou C

Capítulo II

Busca por genes alvos do *T* com o programa hunT

et al., 2007). Além disso, eles mostraram que estes genes também possuem sítio de ligação para o FT Tcf/Lef e são, portanto, regulados pela ação conjunta destes dois fatores. As Figuras 2.3A e B ilustram os sítios de ligação para o *Tbx6* nos genes *Dll1* e *Msgn1*, respectivamente. Os sítios para o Tcf/Lef não são mostrados.

Figura 2.3: Regulação dos genes *Dll1* e *Msgn1* por *Tbx6*. A) Trecho da sequência à montante do gene *Dll1* (4.3kb do ATG) onde foram encontrados 9 sítios de ligação T-box que estão representados em pequenos círculos. A flecha indica o sítio de início da transcrição do *Dll1*. Logo abaixo, está o alinhamento das sequências dos 9 sítios, cada uma seguida por uma flecha que indica sua orientação na fita de DNA. No final, é apresentada a sequência consenso, que reúne os nucleotídeos mais conservados de cada posição das 9 sequências. B) Trecho da sequência à montante do gene *Msgn1* (1.2kb do ATG) onde foram encontrados 7 sítios de ligação T-box. C) Sítio de ligação T-box definido por Kispert & Herrmann, 1993. O quadro vermelho destaca o trecho deste sítio que alinha com aqueles identificados nos genes *Dll1* e *Msgn1*. As Figuras A e B foram adaptadas dos trabalhos de Hoffman *et al.*, 2004 e Wittler *et al.*, 2007, respectivamente.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Embora muito tem sido feito para caracterizar o *T*, ainda pouco se sabe dos genes que ele regula. A tecnologia de ChIP-seq, apesar de estar revolucionando a descoberta de alvos de fatores de transcrição, é muito difícil de ser aplicada quando o material é de tecido embrionário, pois ela requer, no mínimo, dez milhões de células. Fatores de transcrição que controlam uma linhagem embrionária em particular são expressos regionalmente no embrião. Isto significa, por exemplo, que um embrião de camundongo na fase E9.5 tem somente cerca de 1.000-2.000 células do mesodérmicas pré-somíticas. Seria, portanto, necessário cerca de 500-1000 embriões, no mínimo, para um fator tecido-específico, tornando praticamente impossível a determinação de sítios de ligação de reguladores mesodérmicos, como o *T* ou o *Tbx6*, em material embrionário usando esta tecnologia.

Como alternativa, o grupo recentemente investiu num estudo de expressão gênica. Em 2010, eles geraram uma linhagem de ESC (do inglês *Embryonic Stem Cell*) de camundongo com a expressão de um repórter mCherry-*T*, construído geneticamente para conter um sistema de RNA de interferência (RNAi) para *T* (Vidigal *et al.*, 2010). Este trabalho foi a chave para o estudo onde avaliamos o transcriptoma da formação do mesoderma *in vitro*, pois permitiu que a especificação de células tronco embrionária de camundongo para mesoderma fosse monitorada discriminando o desenvolvimento entre uma linhagem de célula normal (WT, do inglês *wildtype*) e outra onde o *T* foi reduzido com um RNAi (*T*-KD, do inglês *knockdown*). Mais detalhes podem ser vistos no anexo 1 deste capítulo.

A expectativa com essa abordagem foi identificar os genes co-regulados com o *T* nas células normais e inalterados nas células *T*-KD. Embora o resultado das análises destes dados tenha revelado evidências interessantes, a metodologia de *microarrays* sozinha é limitante. Ela não permite distinguir quais dos genes que estão co-expressos com um FT são seus alvos diretos ou indiretos. A Figura 2.4 ilustra um exemplo onde o gene X é alvo direto do *T* e o Y é um alvo indireto. Os resultados de expressão gênica incluem X + Y. Nosso desafio é encontrar genes como o X, ou seja, os alvos diretos do *T*.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Figura 2.4: Exemplo de alvo direto e indireto do *T*. O X representa um gene regulado diretamente pelo *T* e o Y representa um gene regulado indiretamente.

A Figura 2.5 ilustra o desenho experimental dos ensaios de *microarray*. A diferenciação foi monitorada nas transições de ESC para aESC (ESCs agregadas que foram induzidas à diferenciação com Bmp4) e nos tempos 48h, 72h e 96h. Em 72h foi observada a formação do mesoderma e em 96h a formação do mesoderma lateral *in vitro*.

Figura 2.5: Desenho experimental dos ensaios de *microarray*. Os dois ramos no desenho indicam os diferentes cenários em que a diferenciação *in vitro* foi conduzida. No ramo superior, chamado de normal ou WT, a diferenciação foi conduzida com a presença do *T*, que levou ao desenvolvimento normal do camundongo. No ramo inferior, chamado de *T*-KD, a ação do *T* foi reprimida pelo RNAi, resultando no desenvolvimento anormal do camundongo.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

O *hunT* foi criado com o propósito de tornar a nossa predição de novos alvos do *T* mais precisa pois com ele poderíamos unir duas importantes fontes de informação: sequência e expressão gênica. Poderíamos detectar genes com sítios de ligação para o *T* em suas sequências reguladoras e com a expressão alterada durante a formação do mesoderma.

Diferente dos programas existentes, o *hunT* foi criado para ser uma ferramenta mais robusta biologicamente. Ele foi adaptado ao conhecimento existente sobre o *T*. Primeiro, ele permite a busca simultânea de n sítios na mesma sequência. Segundo, permite proteger trechos de um sítio da ocorrência de *mismatches*²². Terceiro, permite definir a quantidade mínima de vezes que cada um dos sítios deve ser encontrada.

2.3 Materiais e Métodos

2.3.1 Funcionamento do *hunT*

O programa *hunT* foi desenvolvido na linguagem Perl versão 5.0. Sua lógica está na criação de uma máquina de estado finito para cada padrão buscado. Uma máquina de estado finito (MEF) é um modelo matemático usado para representar programas de computador ou circuitos lógicos. Tem como função, receber uma cadeia de símbolos, construída a partir de um determinado alfabeto, processá-los e gerar um resultado. Uma MEF é composta por um conjunto de estados predefinidos e um conjunto de regras que definem a transição entre os estados. Um estado descreve um comportamento do sistema, que está à espera de uma condição para executar uma transição. Uma transição indica uma mudança de estado e é descrita por uma condição que precisa ser realizada para que a transição ocorra.

Uma máquina pode estar em apenas um estado por vez, que é chamado de estado atual. O conjunto de estados serve como uma espécie de memória da máquina. MEFs podem ser aplicadas em um grande número de problemas, é um modelo útil para representar aspectos dinâmicos que não se expressam em outros

²² *Mismatch* refere-se a troca de um nucleotídeo em uma determinada posição. Pode ser resultado da incorporação incorreta de um nucleotídeo durante a replicação do DNA, uma mutação, ou resultado de problemas na leitura dos equipamentos de sequenciamento.

Capítulo II

Busca por genes alvos do T com o programa hunT

diagramas. Algumas das aplicações mais comuns incluem analisador léxico, casamento de padrões e modelagem de comportamento em jogos. No jogo Pac-man, por exemplo, a MEF é usada para definir o comportamento dos fantasmas (Gallagher & Ryan, 2003).

Uma máquina de estado finito M pode ser descrita como uma 5-tupla $M = \{Q, \Sigma, \rho, f_s, f_o\}$, onde Q é um conjunto finito de estados, Σ é um conjunto finito de símbolos de entrada, ρ é um conjunto finito de símbolos de saída, $f_s: Q \times \Sigma \rightarrow Q$ é a função que mapeia para o próximo estado e $f_o: Q \times \Sigma \rightarrow \rho$ é a função que mapeia para a próxima saída (Cormen *et al.*, 2002).

O quadro a seguir mostra o pseudocódigo de como esta abordagem foi utilizada no hunT.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

```

Entrada: Arquivo com sequências de nucleotídeos de genes;
Padrão(oes), padrões[]; número mínimo de ocorrência de cada
padrão(oes), num_padrao; número de mismatch, num_mismatch

padroes[] = ler_padroes()
gerar_maquina_estados(padroes) # Gera um grafo para cada padrão, inclusive para padrões
buscados em fita de DNA complementar
PARA CADA sequencia_gene FACA # Para cada uma das sequências de genes
    genes_alvos[] # Cria novo array
    PARA CADA maquina_estado FACA # Para cada um dos grafos
        indice = 0
        hit[] # Cria novo array
        FACA ENQUANTO indice < sequencia_gene.tamanho() # Para cada um dos
nucleotídeos da sequência atual
            indice_estado = indice
            mismatch_encontrado = 0
            maquina_estado.reinicia() # Retorna ao estado inicial da máquina
            FACA ENQUANTO indice_estado < sequencia_gene.tamanho() # Loop
MEF
                estado_atual = maquina_estado.proximo_estado()
                SE estado_atual.nao_contem(sequencia_gene[indice_estado])
ENTAO # Verifica todos os possíveis nucleotídeos de um estado, por ex. [AT]
                    SE estado_atual.aceita_mismatch()) ENTAO # Estado não esta
protegido contra mismatch
                        mismatch_encontrado = mismatch_encontrado + 1
                        SE mismatch_encontrado > num_mismatch ENTAO
                            indice_estado = sequencia_gene.tamanho()
                        FIM SE
                    SENAO
                        indice_estado = sequencia_gene.tamanho()
                    FIM SE
                FIM SE
continua...
```

Capítulo II

Busca por genes alvos do *T* com o programa hunT

```

        SE estado_atual.estado_final() E indice_estado <>
sequencia_gene.tamanho() ENTAO
    hit.adiciona(indice, indice_estado, maquina_estado) #
    Armazena coordenadas do padrão encontrado
    indice_estado = sequencia_gene.tamanho()
    SENAO
        indice_estado = indice_estado + 1
    FIM SE
FIM ENQUANTO
indice_nucleotideo = indice_nucleotideo + 1
FIM ENQUANTO
SE maquina_de_estado.num_padrao(hit) ENTAO # Verifica se número de vezes que
padrão foi encontrado é >= que o definido pelo usuário
    genes_alvos.adiciona(hit)
    FIM SE
FIM PARA
SE genes_alvos.tamanho() == padroes.tamanho() ENTAO # Verifica se foram
encontrados todos os padrões
    imprime(sequencia_gene, genes_alvos)
    FIM SE
FIM PARA

Saída: Lista de genes que possuem o(s) padrão(oes) buscado(s) pelo
usuário.

```

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

CONFIGURAÇÃO

O programa *hunT* é multiplataforma. Para rodá-lo, somente é necessário ter o Perl instalado. O arquivo de entrada - que contém as sequências de DNA onde os padrões serão buscados - deve estar no formato *fasta*²³. Somente sequências de nucleotídeos são aceitas. O *hunT* suporta a busca por padrões exato ou aproximado. A busca por padrão aproximado pode ser feita através do uso de um caractere coringa N ou do uso de colchetes, ver Tabela 2.1.

Tabela 2.1. Sintaxe suportada pelo programa *hunT*.

Padrão	Significado	Exemplo	Explicação
A	Adenina	AAA	
C	Citosina	CCC	
G	Guanina	GGG	
T	Timina	TTT	
N	Símbolo coringa	GCCNTTA	N é usado para posição onde qualquer nucleotídeo é aceito
[]	Subconjunto de símbolos aceitos em uma dada posição	AT[TC]ATA	AT seguido por um T ou C, seguido por ATA
()	Subconjunto protegido de símbolos	GC(CTG)AA	GC, seguido por CTG que está protegido da ocorrência de mismatch, seguido por AAA

Uma busca com o *hunT* pode ser customizada de várias formas. Ela pode ser feita em somente em uma ou ambas as fitas do DNA, pode-se buscar por *n* padrões simultaneamente e é possível especificar parâmetros específicos para cada um dos padrões, tornando a busca mais sofisticada. É possível, por exemplo, buscar por genes que possuam, no mínimo, um padrão sem *mismatch* e dois com até 2 *mismatches* nas suas sequências reguladoras ou buscar genes regulados por fatores de transcrição distintos.

A Figura 2.6 ilustra dois exemplos de máquinas de estado criadas pelo *hunT* para os padrões ACCGT e ACNC[GT]ATT.

²³ Formato de texto para representar sequências de nucleotídeos ou de aminoácidos. Cada sequência no arquivo começa com uma linha de descrição iniciada pelo símbolo ">", seguida pelas linhas da sequência.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Figura 2.6: Dois exemplos de máquinas de estado criadas pelo *hunT*. A primeira, para o padrão ACCGT e a segunda para o padrão ACNC[GT]ATT. Neste último, o N corresponde a um caractere coringa. Significa que na terceira posição do padrão será aceito qualquer um dos quatro nucleotídeos: A, C, G ou T. O G e o T entre colchetes indicam que na quinta posição será aceito um destes dois nucleotídeos.

USO

Sintaxe para executar o *hunT* via linha de comando:

```
Usage ./hunT.pl [ARGS]
        --input-file=<file_name>
        --output-file=<file_name>
        --search-type=[0|1]
        --pattern=<search_pattern>
        --label=<label>
        --pattern-min=[0..n]
        --mismatch=[0..n]
```

SAÍDA

Capítulo II

Busca por genes alvos do *T* com o programa hunT

O hunT gera dois arquivos de resultados no formato html: um sumário e um detalhado. No sumário, a primeira coluna ("*Sequence Name*") traz a relação de todos os genes onde foram encontrados os padrões, a segunda coluna ("*Name*") mostra o nome do padrão, a terceira coluna ("*Match*") mostra o trecho da sequência de DNA que casou com o padrão buscado, a quarta coluna ("*Mismatch*") à quantidade de *mismatch* presente no trecho da sequência da coluna anterior, a quinta coluna ("*Strand*") à orientação da fita de DNA e as duas últimas colunas ("*Start*" e "*End*") mostram a posição inicial e final do padrão na sequência de DNA (Figura 2.7A). No final do arquivo é apresentado o total de genes em que foram encontrados os padrões. O arquivo detalhado traz a sequência desses genes. Os padrões encontrados são mostrados em negrito e na cor verde (Figura 2.7B).

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Figura 2.7: Trecho dos dois arquivos de saída gerados pelo hunT. A) Arquivo sumário com a relação dos genes onde foram encontrados os padrões. Ao lado de cada gene estão os padrões encontrados, seguidos da quantidade de *mismatch*, da orientação da fita de DNA e da posição do padrão na sequência de DNA. No gene *Dbt*, por exemplo, foram encontrados 12 padrões, sendo 1 forte (*strong*) e 11 fracos (*weak*). B) Arquivo detalhado, onde estão as sequências de DNA dos genes onde os padrões foram encontrados. Os padrões estão em negrito e na cor verde. Os *mismatches* estão sublinhados. A sequência do gene *Dbt* possui 11 padrões. O sumário mostrou 12 pois o padrão "*strong*" foi considerado também como "*weak*". Ele foi encontrado na fita reversa do DNA, por isso ele aparece como TTTTACACCT.

CÓDIGO

O código fonte do programa hunT está disponível no endereço: http://danieleyumi.sunaga.de/?page_id=23 do material suplementar.

2.3.2 Comparação com programa similar

Para testar as funções do hunT de busca simultânea por diferentes padrões e de proteção contra *mismatches*, utilizamos 3 padrões que foram buscados com o hunT e com um programa similar chamado fuzznuc (Rice *et al.*, 2000)²⁴:

- P1: 5' - [CG]CTGTA[GT] [AT] [AT] [AT] - 3'
- P2: 5' - [CG] (CTGTA) [GT] [AT] [AT] [AT] - 3'
- P3: 5' - AGG[CG]TT - 3'

O fuzznuc inteligentemente seleciona o algoritmo de busca mais adequado, dependendo da complexidade do padrão a ser buscado. Foi usado como referência para os nossos testes por ser um programa conhecido, que faz parte do pacote de programas de alta qualidade EMBOSS²⁵.

Os testes foram realizados usando um arquivo de entrada contendo 20 sequências de até 600 nucleotídeos cada, que foram criadas arbitrariamente. Foram usados os mesmos critérios de busca no hunT e fuzznuc.

Inicialmente testamos a eficiência do hunT para uma busca simples usando o padrão P1 de tamanho médio e permitindo a ocorrência de até 2 *mismatches* (teste-simples-padrão-grande):

²⁴ <http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/fuzznuc.html>

²⁵ <http://emboss.sourceforge.net/>

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Comando *hunT*:

```
./hunT.pl --input-file=20seqs.fas --output-file=P1
--pattern="[CG]CTGTA[GT][AT][AT][AT]" --label="P1" --search-type=0
--mismatch=2 --pattern-min=1
```

Comando *fuzznuc*:

```
fuzznuc -sequence 20seqs.fas -pattern [CG]CTGTA[GT][AT][AT][AT]
-pmismatch=2 -complement 0 -pname P1 -rformat srs
```

Em seguida, testamos a função de proteção do *hunT* contra *mismatches*.

Para isso, usamos o padrão P2, permitindo a ocorrência de até 2 *mismatches* na porção não protegida do padrão (teste-proteção):

Comando *hunT*:

```
./hunT.pl --input-file=20seqs.fas --output-file=P1-protegido
--pattern="[CG](CTGTA)[GT][AT][AT][AT]" --label="P2" --search-type=0
--mismatch=2 --pattern-min=1
```

Comando *fuzznuc*: operação não suportada

O padrão P3 foi buscado sem nenhum *mismatch* (teste-simples-padrão-pequeno):

Comando *hunT*:

```
./hunT.pl --input-file=20seqs.fas --output-file=P3
--pattern="AGG[CG]TT" --label="P3" --search-type=0 --mismatch=0
--pattern-min=1
```

Comando *fuzznuc*:

```
fuzznuc -sequence 20seqs.fas -pattern AGG[CG]TT -pmismatch=0
-complement 0 -pname P3 -rformat srs
```

Por último, os padrões P1 e P3 foram combinados para testar a função de busca simultânea do *hunT* (teste-busca-simultânea):

Comando *hunT*:

```
./hunT.pl --input-file=20seqs.fas --output-file=P1+P3-protegido
--pattern="[CG]CTGTA[GT][AT][AT][AT]" --pattern="(AGG[CG]TT)"
--label="P1" --label="P3-protegido" --search-type=0 --mismatch=2
--pattern-min=1 --pattern-min=1
```

Comando *fuzznuc*: operação não suportada

Neste teste, permitimos a ocorrência de até 2 *mismatches* mas protegemos todo o padrão P3, servindo como um teste completo das funcionalidades do *hunT* pois reúne a busca simultânea por diferentes padrões e a proteção contra *mismatches*.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

2.3.3 *hunT* para a busca de alvos de diferentes FTs

Utilizamos os achados dos trabalhos de Hoffman *et al.*, 2004 e Wittler *et al.*, 2007 para validar o funcionamento do *hunT* na busca por alvos de diferentes FTs. Nós buscamos pela ocorrência simultânea de sítios de ligação para o domínio T-box e Tcf/Lef nas sequências *cis*-reguladoras dos genes *Dll1* e *Msgn1* de camundongo. A metodologia para a obtenção destas sequências está descrita no item 2.3.5.

Definimos os padrões T-box e Tcf/Lef de acordo com os achados destes trabalhos:

T-box: 5'-A[AG]GTGT[GA][AT][AT][AT]-3'
 Tcf-Lef: 5'-[CT]CTTTG[AT][AT]-3'

Buscamos pela ocorrência simultânea destes dois padrões, sendo no mínimo uma de cada, e até dois *mismatches*, conforme sintaxe abaixo:

```
./hunT.pl
--input-file=Dll1_Msgn1.fas
--output-file=Dll1_Msgn1.out
--pattern1="A[AG]GTGT[GA][AT][AT][AT]"
--pattern2="[CT]CTTTG[AT][AT]"
--label1="T-box"
--label2="Tcf-lef"
--pattern-min1=1
--pattern-min2=1
--mismatch=2
```

2.3.4 *hunT* para a busca de genes alvos do *T*

Para a busca por alvos do *T*, definimos duas variantes da sequência do domínio T-box mostrada acima, que foram denominadas de padrão *T-fraco* e *T-forte*:

T-fraco: A[AG] (GTG) T[GA][AT][AT][AT]
T-forte: (A[AG]GTGT[GA][AT][AT][AT])

Capítulo II

Busca por genes alvos do *T* com o programa hunT

O que os difere são somente os parênteses, que servem para proteger a sequência da ocorrência de *mismatch*. No padrão *T-forte*, os parênteses envolvem a sequência completa, o que significa que não é aceito nenhum *mismatch* em toda a sequência, ou seja, ela deve ser exata. Já no padrão *T-fraco*, os parênteses protegem somente o trecho GTG, o que significa que podem haver *mismatches* em todo o restante da sequência menos nesta trinca, que temos assumido como identidade do sítio. Nos trabalhos com os genes *Dll1* e *Msgn1*, a mutação de um nucleotídeo nesta trinca levou à perda da especificidade entre sítio e fator de transcrição. As máquinas de estado dos padrões *T-fraco* e *T-forte* são ilustradas na Figura 2.8.

Figura 2.8: Máquinas de estado dos padrões *T-fraco* e *T-forte*. O que as diferencia são somente os estados protegidos contra *mismatch*. No padrão fraco, somente a sequência de estados G, T e G está protegida, enquanto no padrão forte todos os estados estão

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

protegidos. Se permitida a ocorrência de até 2 *mismatches*, por exemplo, significa que eles podem ocorrer em qualquer um dos estados do *T*-fraco, menos no trecho protegido.

Buscamos estes dois padrões na região *cis*-reguladora de todos os genes de camundongo. A metodologia para a obtenção destas sequências está descrita no item 2.3.5. Abaixo, segue a sintaxe do *hunT* utilizada nesta busca:

```
./hunT.pl
--input-file=mousegenome.fas
--output-file=mousegenome.out
--pattern1="(A[AG]GTGT[GA][AT][AT][AT])"
--pattern2="A[AG](GTG)T[GA][AT][AT][AT]"
--label1="T-strong"
--label2="T-weak"
--pattern-min1=1
--pattern-min2=1
--mismatch=2
```

2.3.5 Obtenção da sequência genômica

Utilizamos a ferramenta *biomaRt*²⁶ do pacote *Bioconductor/R* para extrair os trechos das sequências reguladoras onde buscar pelos padrões *T*-forte e *T*-fraco. A ferramenta permite extrair grandes quantidades de dados de grandes repositórios do *BioMart*²⁷, como por exemplo, o *Ensembl*, *Uniprot* e *HapMap*.

Foram extraídos 3kb da região à montante do TSS (do inglês, *Transcription Start Site*) + a região 5'UTR (do inglês, *Untranslated region*; região transcrita, mas não-traduzida) de todos os genes de camundongo do repositório *Ensembl*, conforme representado no esquema da Figura 2.9.

²⁶ <http://www.bioconductor.org/packages/2.2/bioc/html/biomaRt.html>

²⁷ <http://www.biomart.org>

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Figura 2.9: Representação esquemática da região genômica que foi extraída do Ensembl. Para a busca pelos padrões *T*-forte e fraco, utilizamos o fragmento que está em vermelho na imagem, que corresponde a 3kb à montante do TSS + a região 5'UTR.

Esta ferramenta, contudo, não permite extrair todas as sequências do genoma de uma só vez. Para lidar com essa limitação, delimitamos grupos de 5.000 sequências cada e repetimos 7 vezes os comandos descritos no quadro abaixo até cobrir todas as ~40.000 sequências genoma de camundongo.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

```
$library(biomaRt)
###Carrega repositório mmusculus_gene_ensembl na variável ensembl
$ensembl=useMart("ensembl", dataset="mmusculus_gene_ensembl")
###Obtem lista de nomes de genes usando o atributo external_gene_id
$mousegenomelist=getBM("external_gene_id", mart=ensembl)
###Obtem a região 5UTR+3000upstream dos primeiros 5000 genes da lista
mousegenomelist
$utr5batch1=getSequence(id=mousegenomelist[1:5000,1],type="wik
igene_name", seqType="coding_gene_flank", upstream=3000,
mart=ensembl)
###Exporta resultado para arquivo no formato FASTA
$exportFASTA(sequence=utrbatch1, file="utrbatch1.fas")

---> A obtenção de sequências com o getSequence teve que ser
repetida 7 vezes até cobrir o genoma de camundongo do
repositório mmusculus_gene_ensembl:

$utr5batch2=getSequence(id=mousegenomelist[5001:10000,1],type=
"wikigene_name", seqType="coding_gene_flank", upstream=3000,
mart=ensembl)
$exportFASTA(sequence=utrbatch2, file="utrbatch2.fas")
.
.
$utr5batch7=getSequence(id=mousegenomelist[30001:40000,1],type
="wikigene_name", seqType="coding_gene_flank", upstream=3000,
mart=ensembl)
$exportFASTA(sequence=utrbatch7, file="utrbatch7.fas")

---> No final, os 7 arquivos .fas foram concatenados com o
comando cat do linux
```

Capítulo II

Busca por genes alvos do *T* com o programa hunT

2.4 Resultados e discussões

Neste item, primeiro apresentamos os resultados da comparação do hunT com uma ferramenta similar chamada fuzznuc, em seguida, apresentamos os resultados do teste onde buscamos por alvos de dois FTs distintos. Por último, mostramos os resultados da busca por alvos do *T* e a integração destes resultados com os dados de expressão gênica.

Todos os resultados deste capítulo estão disponíveis no endereço http://danieleyumi.sunaga.de/?page_id=23 do material suplementar.

2.4.1 Comparação com programa similar

No teste-simples-padrão-grande, o hunT identificou 65 padrões P1 distribuídos entre 18 das 20 sequências analisadas. O fuzznuc encontrou o mesmo resultado. Em ambos, a busca foi feita considerando a ocorrência de até 2 *mismatches* em qualquer posição do padrão (Tabela 2.2).

No teste-proteção, repetimos a mesma busca anterior usando o padrão P2, que é igual ao P1, exceto pelo trecho GTGTA que foi protegido da ocorrência de *mismatches*. Esta proteção fez reduzir os achados do hunT para 6 padrões distribuídos entre 5 das 20 sequências. O fuzznuc não suporta este tipo de busca. Estes resultados mostram que a proteção do hunT serve como um filtro eficiente, uma vez que sem ela, o usuário teria que lidar com 65 padrões candidatos ao invés de 6. Numa situação real onde geralmente são analisadas milhares de sequências, esta diferença representa uma diminuição de 91% dos padrões encontrados (Tabela 2.2).

No teste-simples-padrão-pequeno, tanto o hunT como o fuzznuc encontraram 7 padrões P3 distribuídos entre 7 das 20 sequências (Tabela 2.2).

Por último, no teste-busca-simultânea, o hunT encontrou ambos os padrões, P1 e P3, em 5 das 20 sequências, o que numa situação real, poderiam corresponder a genes regulados pela combinação de dois fatores distintos (Tabela 2.2). O fuzznuc não suporta este tipo de busca. Num problema como este, a busca teria que ser feita separadamente para cada padrão. O usuário teria então que lidar com as 18 sequências com o padrão P1 e as 7 com o padrão P3, totalizando 25 sequências ao invés das 5 encontradas pelo hunT. Ou ainda, poderia usar os resultados da busca do primeiro padrão como entrada para a busca do segundo. A dificuldade neste caso, seria que as informações do primeiro, como por exemplo: a

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

posição onde o padrão foi encontrado, quantidade de *mismatches*, e fita do DNA, seriam perdidas.

Também testamos a hipótese de o sítio de um destes fatores ser encontrado várias vezes na sequência reguladora dos genes que eles regulam, assim como o que tem sido observado nos genes regulados pelo *Tbx6*. O *hunT* encontrou 3 sequências contendo no mínimo um padrão P3 e 4 padrões P1. Todos estes resultados estão disponíveis no material suplementar.

Tabela 2.2: Resultados dos testes com o *hunT* e *fuzznuc*. O número de *hits* corresponde ao número de padrões.

Testes	Padrões	hunT		Fuzznuc	
		#hits	#seqs	#hits	#seqs
teste-simples-padrão-grande	P1: [CG]CTGTA[GT] [AT] [AT] [AT]	65	18	65	18
teste-proteção	P2: [CG] (CTGTA) [GT] [AT] [AT] [AT]	6	5	NS*	
teste-simples-padrão-pequeno	P3: AGG[CG]TT	7	7	7	7
teste-busca-simultânea	P1+P3	30	5	NS*	

NS* = Operação não suportada

2.4.2 Busca por alvos de diferentes FTs

Extraímos um total de 21.020 sequências reguladoras de genes de camundongo do repositório Ensembl usando a ferramenta *biomaRt*.

Neste teste, usamos somente as sequências dos genes *Dll1* e *Msgn1* como entrada para o *hunT*. O arquivo *fasta* está disponível no endereço http://danieleyumi.sunaga.de/wp-uploads/Dll1_Msgn1.fas.zip do material suplementar.

O programa encontrou sítios de ligação para o domínio T-box e Tcf/Lef nas sequências de ambos os genes, corroborando os achados de Hoffman *et al.* 2004 e Wittler *et al.*, 2007 e confirmando a funcionalidade do programa de buscar por sítios de FTs distintos na mesma sequência. Os resultados estão disponíveis nos endereços <http://danieleyumi.sunaga.de/wp->

Capítulo II

Busca por genes alvos do *T* com o programa hunT

[uploads/summary_Dll1_Msgn1.out_.html](http://danieleyumi.sunaga.de/wp-uploads/summary_Dll1_Msgn1.out_.html) (resultado resumido) e http://danieleyumi.sunaga.de/wp-uploads/detailed_Dll1_Msgn1.out_.html (resultado detalhado) do material suplementar.

2.4.3 Busca por alvos do *T*

Analisando o total de 21.020 sequências de genes de camundongo, o hunT encontrou a ocorrência simultânea de, no mínimo, um padrão *T*-forte e um *T*-fraco em 4.602 destes genes. Os resultados estão disponíveis nos endereços: http://danieleyumi.sunaga.de/wp-uploads/summary_mousegenome230912.out_.zip (resultado resumido) e http://danieleyumi.sunaga.de/wp-uploads/detailed_mousegenome230912.out_.zip (resultado detalhado) do material suplementar.

Dentre eles, estão novamente os genes *Dll1* e o *Msgn1*, corroborando que eles, de fato, possuem vários sítios de ligação para o domínio T-box distribuídos ao longo de suas sequências reguladoras (Hoffman *et al.* 2004 e Wittler *et al.*, 2007). Vários outros genes interessantes também fazem parte desta lista, mas ainda sem evidência experimental. O *Dkk1* é um exemplo. Assim como o *T*, ele é um conhecido alvo e inibidor da via de sinalização Wnt, que tem papel fundamental na formação do mesoderma (Martin & Kimelman, 2008).

Recentemente, dois trabalhos também buscaram por alvos do *T* através de tecnologias baseada em ChIP (Garnett *et al.*, 2009 e Evans *et al.*, 2012). No primeiro, de 44 genes de zebrafish, 27 dos quais possuem ortólogos em camundongo, 16 deles (59%) estão entre os resultados do hunT. No segundo, de 396 genes encontrados, 78 (20%) estão entre os resultados do hunT (http://danieleyumi.sunaga.de/wp-uploads/hunT_Garnett_Evans.xls do material suplementar).

Há uma série de razões que explicam por que alguns dos genes identificados nestes trabalhos não foram identificados pelo hunT. A primeira delas é de que estes genes certamente não possuem o sítio de ligação que usamos nas nossas buscas, ou estes sítios variam de alguma forma que não tratamos (como gaps, inserções e deleções) ou, ainda, estão localizados numa região diferente da que delimitamos as nossas buscas. Já o maior número de genes encontrados pelo hunT deve-se ao fato de que o sítio que buscamos é comum para todos os membros da família T-box, o que significa que podem ser alvos de 18 FTs

Capítulo II

Busca por genes alvos do *T* com o programa hunT

diferentes²⁸. Experimentos adicionais são necessários para mostrar quais deles podem ser alvos específicos do *T*. Nós fizemos isso através de estudos de expressão gênica, cujos resultados são mostrados a seguir.

2.4.4 Expressão gênica dos alvos do *T*

Encontramos 32 dos 4.602 genes identificados pelo hunT com expressão alterada após 96h de diferenciação das células WT e não das T-KD. Este período foi quando observamos a formação do mesoderma lateral *in vitro* e quando o *T* foi encontrado com a expressão aumentada em nossas análises, indicando a participação destes genes neste processo sendo regulados pelo *T* (Tabela 2.3).

Tabela 2.3: Trinta e dois genes candidatos-alvos do *T*. Além de terem suas sequências reguladoras enriquecidas com sítios de ligação para o *T*, eles estão co-expressos com o *T* em 96h das células WT (*fold change* > 1,5), quando é iniciada a formação do mesoderma lateral.

ILMN_ID	Gene	Fold change	Descrição
ILMN_1259418	<i>Aifm2</i>	-2.1915	Mus musculus apoptosis-inducing factor, mitochondrion-associated 2 (<i>Aifm2</i>), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
ILMN_3009792	<i>Aim2</i>	-2.0732	Mus musculus absent in melanoma 2 (<i>Aim2</i>), mRNA.
ILMN_1238558	<i>Arid3b</i>	2.0802	Mus musculus AT rich interactive domain 3B (BRIGHT-like) (<i>Arid3b</i>), mRNA.
ILMN_1247811	<i>Ass1</i>	-3.5839	Mus musculus argininosuccinate synthetase 1 (<i>Ass1</i>), mRNA.
ILMN_3065164	<i>Bex2</i>	2.1043	Mus musculus brain expressed X-linked 2 (<i>Bex2</i>), mRNA.
ILMN_2792868	<i>Blvrb</i>	-2.202	Mus musculus biliverdin reductase B (flavin reductase (NADPH)) (<i>Blvrb</i>), mRNA.
ILMN_1224754	<i>Ckb</i>	-2.037	Mus musculus creatine kinase, brain (<i>Ckb</i>), mRNA.
ILMN_2658908	<i>Cxcl12</i>	2.4322	
ILMN_2622983	<i>Dusp1</i>	-2.0318	Mus musculus dual specificity phosphatase 1 (<i>Dusp1</i>), mRNA.
ILMN_2846194	<i>Echdc2</i>	-2.5142	Mus musculus enoyl Coenzyme A hydratase domain containing 2 (<i>Echdc2</i>), mRNA.
ILMN_2688653	<i>Epha2</i>	-2.2148	Mus musculus Eph receptor A2 (<i>Epha2</i>), mRNA.
ILMN_1256354	<i>Gclm</i>	-2.3301	Mus musculus glutamate-cysteine ligase, modifier subunit (<i>Gclm</i>), mRNA.
ILMN_3053593	<i>Glipr2</i>	2.0189	Mus musculus GLI pathogenesis-related 2 (<i>Glipr2</i>), mRNA.
ILMN_2433318	<i>Greb1</i>	2.2322	
ILMN_2776909	<i>Irx3</i>	9.0565	
ILMN_2614462	<i>Krt19</i>	2.4516	Mus musculus keratin 19 (<i>Krt19</i>), mRNA.
ILMN_1213273	<i>Lrig3</i>	2.716	Mus musculus leucine-rich repeats and immunoglobulin-like domains 3 (<i>Lrig3</i>), mRNA.
ILMN_1216748	<i>Nt5m</i>	2.0653	Mus musculus 5',3'-nucleotidase, mitochondrial (<i>Nt5m</i>), nuclear gene encoding mitochondrial protein, mRNA.

²⁸ 18 membros da família T-box são conhecidos em mamíferos, 11 em *Drosophila melanogaster*, 14 em *Caenorhabditis elegans*, entre outros (Wilson & Conlon, 2002).

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

ILMN_2842877	<i>Pja1</i>	2.1621	Mus musculus prja1, RING-H2 motif containing (Pja1), mRNA.
ILMN_2732576	<i>Pml</i>	-2.6636	Mus musculus promyelocytic leukemia (Pml), transcript variant 1, mRNA.
ILMN_2720836	<i>Rbm47</i>	-2.7116	Mus musculus RNA binding motif protein 47 (Rbm47), mRNA.
ILMN_2661650	<i>Reep6</i>	-2.2843	Mus musculus receptor accessory protein 6 (Reep6), mRNA.
ILMN_1219289	<i>Rgs17</i>	2.0405	
ILMN_2761046	<i>Slc25a33</i>	-2.0656	Mus musculus solute carrier family 25, member 33 (Slc25a33), mRNA.
ILMN_2938893	<i>Smad3</i>	2.1034	Mus musculus MAD homolog 3 (Drosophila) (Smad3), mRNA.
ILMN_2644632	<i>Stxbp1</i>	2.041	Mus musculus syntaxin binding protein 1 (Stxbp1), mRNA.
ILMN_1222489	<i>Sulf1</i>	2.1797	Mus musculus sulfatase 1 (Sulf1), mRNA.
ILMN_1217144	<i>Tbc1d5</i>	2.0788	Mus musculus TBC1 domain family, member 5 (Tbc1d5), mRNA.
ILMN_3084954	<i>Tes</i>	2.0872	Mus musculus testis derived transcript (Tes), transcript variant 1, mRNA.
ILMN_2484147	<i>Tro</i>	2.5087	Mus musculus trophinin (Tro), transcript variant 1, mRNA.
ILMN_3129213	<i>Tro</i>	2.0477	Mus musculus trophinin (Tro), transcript variant 1, mRNA.
ILMN_2817714	<i>Tspan17</i>	-2.3566	Mus musculus tetraspanin 17 (Tspan17), mRNA.

A expressão temporal dos 32 genes nas células WT é ilustrada na Figura 2.10. No topo da figura é mostrado o perfil da expressão do *T* nas células WT, para servir de referência para a expressão dos genes que podem estar sendo regulados por ele.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Figura 2.10: Expressão temporal dos 32 genes candidatos-alvos do *T*. O perfil da expressão do *T* está representado no topo da imagem. Cada coluna no *heat map* corresponde à média da expressão dos *arrays* de cada um dos tempos da diferenciação. As linhas representam o perfil da expressão dos genes. Vermelho corresponde à expressão aumentada e verde diminuída, numa escala de -2.0 a 2.0.

Cxcl12 é um ligante para o receptor *Cxcr4*, que é expresso no mesoderma pré-somítico (McGrath *et al.*, 1999). *Smad3* é um mediador da via de sinalização TGF-beta, atuando como alvo dos genes *Nodal* e *Activin* que são requeridos na formação do mesoderma (Jia *et al.*, 2008). *Irx3* é um fator de transcrição que atua no mesoderma lateral, induzido por Bmp4 (Mahlpuu *et al.*, 2000). *Dusp1* é induzido por um receptor tirosina quinase da via MAP-quinase, também envolvida na formação do mesoderma (Umbhauer *et al.*, 1995). Está claro que estes candidatos a alvos do *T* identificados correspondem a somente uma peça de um enorme quebra cabeças. Estudos adicionais serão necessários. Eles, no entanto,

Capítulo II

Busca por genes alvos do *T* com o programa hunT

destacam o fato de que os alvos do *T* devem incluir um amplo espectro de funções moleculares, que compreende fatores de transcrição, componentes das vias de sinalização, proteínas estruturais e enzimas de vias metabólicas.

A Figura 2.11 ilustra imagens da hibridização *in situ* de 5 dos 32 genes (*Cxcl12*, *Epha2*, *Greb1*, *Lrig3* e *Sulf1*), que são expressos nas mesmas células onde o *T* está expresso, reforçando a evidência de que eles podem ser seus alvos diretos.

Figura 2.11: Expressão do *T* e de 5 dos 32 genes na cauda do camundongo. Imagens de hibridização *in situ* obtidas do repositório Mamep (<http://mamep.molgen.mpg.de/>) criado pelo grupo do Prof. Bernhard Herrmann.

Estes resultados não incluem os genes com expressão alterada na transição de 72h para 96h, quando é iniciada a formação do mesoderma e a participação do *T* também é conhecida. Cinquenta e sete dos genes do hunT foram

Capítulo II

Busca por genes alvos do *T* com o programa hunT

identificados com expressão diferenciada neste intervalo nas células WT, mas ainda não pudemos verificar o comportamento destes genes nas células *T*-KD devido a problemas com os *arrays* de 72h.

2.4.5 Considerações finais e perspectivas futuras

O hunT também tem sido aplicado para buscar por alvos do *T* em listas específicas de genes. O que difere estas buscas da anterior (quando foi avaliado todo o genoma do camundongo) é a configuração da região reguladora, como o tamanho da sequência, se ela está à montante ou à jusante dos genes e se inclui ou não as regiões 3' e 5'UTRs. Numa destas aplicações, utilizamos uma lista composta pelos dos 4 grupos da família Hox: *hoxA*, *hoxB*, *hoxC* e *hoxD*. Assim como o *T*, estes genes também estão envolvidos na formação do mesoderma lateral. O que não se sabe ainda é qual a relação destes genes neste processo. Foram utilizados 50kb de cada uma das extremidades de cada grupo de genes. Os resultados estão disponíveis nos endereços http://danieleyumi.sunaga.de/wp-uploads/summary_HOX.zip e http://danieleyumi.sunaga.de/wp-uploads/detailed_HOX.zip do material suplementar.

Embora idealizado para buscar alvos do *T*, o hunT é perfeitamente aplicável para buscar alvos de qualquer outro fator de transcrição. Ele é especialmente vantajoso quando se sabe da existência de vários sítios de ligação para um fator, como no caso do *T*, ou para buscar genes que são regulados por fatores de transcrição distintos. A principal limitação do programa é que os padrões encontrados são retornados na ordem em que foram detectados, sem qualquer avaliação de significância. Além disso, hunT somente considera *mismatches* de substituição e não de inserção ou deleção. Nós continuamos a melhorar o programa e esperamos em breve disponibilizar na internet uma nova versão de código livre e futuramente permitir que o usuário possa rodar o programa *online*.

2.5 Conclusões

hunT é uma ferramenta prática e eficiente. Suas funções de busca simultânea por diferentes padrões e de restrição de *mismatches* a certas posições, servem como filtro para a identificação de padrões mais precisos. Mais importante que isso, elas aproximam o resultado de um significado mais biológico, uma vez que a combinação de múltiplos fatores que se ligam a múltiplos sítios são essenciais para a regulação de alguns genes. Comparado com tecnologias como

Capítulo II

Busca por genes alvos do *T* com o programa hunT

ChIP-chip, o hunT tem a vantagem de poder ser estendido à buscas em regiões mais distantes, como os *enhancers*. Já o ChIP-seq, requer uma grande quantidade de células, o que é bastante difícil obter de material embrionário. Combinado com os dados de transcriptoma das condições WT e de embriões mutantes, a predição de sítios de ligação se torna realmente poderosa pois pode indicar quais os genes com expressão diferenciada podem ser alvos diretos do *T*. Os resultados podem ser confirmados com a tecnologia de ChIP-qPCR²⁹, que é factível usando material embrionário e tem a vantagem dos alvos poderem ser analisados em múltiplas amostras, descartando a necessidade do sequenciamento de genomas completos. A metodologia e resultados preliminares apresentados neste capítulo mostram que estamos no caminho para decifrar a rede regulatória por trás da formação do mesoderma. É o primeiro passo para a compreensão de alguns problemas do desenvolvimento de vertebrados, assim como a origem de algumas doenças humanas.

2.6 Referências Bibliográficas

Conlon FL, Sedgwick SG, Weston KM, Smith JC. (1996). Inhibition of Xbra transcription activation causes defects in mesodermal patterning and reveals autoregulation of Xbra in dorsal mesoderm. *Development*, 122:2427–2435.

Cormen TH, Leiserson CE, Rivest RL, Stein C. (2002). *Algoritmos Teoria e Prática 2ªed.* Editora Campus.

D'haeseleer P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, 24:423-425.

Evans AL, Faial T, Gilchrist MJ, Down T, Vallier L, *et al.* (2012). Genomic Targets of Brachyury (T) in Differentiating Mouse Embryonic Stem Cells. *PLoS ONE*, 7(3):e33346. doi:10.1371/journal.pone.0033346.

Farin HF, Mansouri A, Petry M, Kispert A. (2008). T-box Protein Tbx18 Interacts with the Paired Box Protein Pax3 in the Development of the Paraxial Mesoderm. *The Journal of Biological Chemistry*, 283:25372-25380.

Gallagher M & Ryan A. (2003). Learning to play Pac-Man: An evolutionary, rule-based approach. In *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on Evolutionary Computation*- Piscataway, NJ. IEEE, 4:2462–2469.

Garnett AT, Han TM, Gilchrist MJ, Smith JC, Eisen MB, Wardle FC e Amacher SL. (2009). Identification of direct T-box target genes in the developing zebrafish mesoderm. *Development*, 136:749-760.

Grillo G, Licciulli F, Liuni S, Sbisà E, Pesole G. (2003). PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.*, 31(13):3608-3612.

²⁹ Reação em Cadeia da Polimerase em tempo real.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Herrmann BG, Labeit S, Poustka A, King TR, Lehrach H. (1990). Cloning of the T gene required in mesoderm formation in the mouse. *Nature*, 343(6259):617–22.

Herrmann BG & Kispert A. The T genes in embryogenesis. (1994). *Trends in Genetics*, 10:280–286.

Hofmann M, Schuster-Gossler K, Watabe-Rudolph M, Aulehla A, Herrmann BG, Gossler A. (2004). WNT signaling, in synergy with T/TBX6, controls Notch signaling by regulating Dll1 expression in the presomitic mesoderm of mouse embryos. *Genes Dev.*, 15,18(22):2712–7.

Jia S, Ren Z, Li X, Zheng Y, Meng A. (2008). smad2 and smad3 Are Required for Mesendoderm Induction by Transforming Growth Factor- β /Nodal Signals in Zebrafish. *J. Biol. Chem.*, 283: 2418–2426

Kispert A & Herrmann BG. (1993). The Brachyury gene encodes a novel DNA binding protein. *EMBO J.*, 12:3211–3220.

Kispert A, Koschorz B, Herrmann BG. (1995). The T protein encoded by Brachyury is a tissue-specific transcription factor. *The EMBO Journal*, 14(9):4763–4772.

Mahlapuu M, Ormestad M, Enerbäck S, Carlsson P. (2000). The forkhead transcription factor Foxf1 is required for differentiation of extraembryonic and lateral plate mesoderm. *Development*, 128:155–166.

McGrath KE, Koniski AD, Maltby KM, McGann JK, Palis J. (1999). Embryonic expression and function of the chemokine SDF-1 and its receptor, CXCR4. *Dev Biol.*, 213:442–456.

Mangalam HJ. (2002). tacg – a grep for DNA. *BMC Bioinformatics*, 3–8.

Martin BL & Kimelman D. (2008). Regulation of canonical Wnt signaling by Brachury is essential for posterior mesoderm formation. *Dev Cell*, 15(1):121–133.

Muller CW & Herrmann BG. (1997). Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. *Nature*, 389:884–888.

NC-IUB (Nomenclature Committee of the International Union of Biochemistry). (1986). Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. Recommendations 1984. *Proc. Nat. Acad. Sci. (U. S.)*, 83:4–8.

Rice P, Longden I, Bleasby A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277.

Schulte-Merker S, van Eeden FJ, Halpern ME, Kimmel CB, Nusslein-Volhard C. (1994). No tail (ntl) is the zebrafish homologue of the mouse T (Brachyury) gene. *Development*, 120:1009–1015.

Schneider TD & Stephens RM. (1990). Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res*, 18(20):6097–6100.

Smith J. (1997). Brachyury and the T-box genes. *Curr. Opin. Genet. Dev.*, 7:474–480.

Stormo GD. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16:16–23.

Umbhauer M, Marshall CJ, Mason CS, Old RW, Smith JC. (1995). Mesoderm induction in *Xenopus* caused by activation of MAP kinase. *Nature*, 6;376(6535):58–62.

Vidigal JA, Morkel M, Wittler L, Brouwer-Lehmitz A, Grote P, Macura K, Herrmann BG. (2010). An inducible RNA interference system for the functional dissection of mouse embryogenesis. *Nucleic acids research*, 38, e122.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Wasserman WW & Sandelin A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5:276-287.

Wilkinson DG, Bhatt S, Herrmann BG. (1990). Expression pattern of the mouse *T* gene and its role in mesoderm formation. *Nature*, 343:657-659.

Wilson V & Conlon FL. (2002). The T-box family. *Genome Biol.*, 3(6):reviews3008.1-reviews3008.7.

Wittler L, Shin E, Grote P, Kispert A, Beckers A, Gossler A, Werber M, Herrmann BG. (2007). Expression of *Msgn1* in the presomitic mesoderm is controlled by synergism of WNT signalling and *Tbx6*. *EMBO Rep.*, 8(8):784-789.

Yan T, Yoo D, Berardini TZ, *et al.* (2005). PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Res.*, 33(Web Server issue): W262-W266.

Yanagisawa KO, Fujimoto H, Urushihara H. (1981). Effects of the brachyury (T) mutation on morphogenetic movement in the mouse embryo. *Developmental biology*, 87:242-248.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

2.7 Anexo

Este material descreve de forma simplificada o trabalho de "Análise do transcriptoma da formação do mesoderma *in vitro*", feito em colaboração com o grupo do Prof. Dr. Bernhard Herrmann do Max Planck *Institute for Molecular Genetics*, em Berlim-Alemanha. Os resultados das análises de bioinformática foram destacados pois são fruto da minha participação no projeto. Descrevo brevemente a interpretação do significado biológico destes resultados.

Análise do transcriptoma da formação do mesoderma *in vitro*

INTRODUÇÃO

Durante a embriogênese de mamíferos, a gastrulação é um processo fundamental que gera três camadas germinativas primárias: ectoderma, mesoderma e endoderma, que darão origem a um órgão ou tecido específico. O presente estudo investiga a especificação de células para mesoderma em nível transcricional, discriminando entre o desenvolvimento com a presença do *T* e com a redução do *T*.

T ou *Brachyury* é um fator de transcrição, membro fundador da família T-box, que consiste numa grande família de fatores de transcrição, vários deles associados à doenças humanas, como fissura palatina e anquiloglossia ligada ao X (*TBX22*), síndrome de Holt–Oram (*TBX5*), Ulnar-Mammary (*TBX3*) e DiGeorge (*TBX1*, Packham & Brook, 2003). O *T* é um marcador de mesoderma, conhecido por desempenhar um papel fundamental durante o desenvolvimento. A falta deste gene resulta na morte prematura do embrião de camundongo, aproximadamente no 11º dpc (dia pós concepção) (Yanagisawa *et al.*, 1981). Embriões heterozigotos para um alelo do *T* sobrevivem até a idade adulta mas apresentam a cauda encurtada. No trabalho de Stott e colaboradores, eles mostraram que a introdução de uma única cópia de um transgene do *T* foi suficiente para recuperar completamente o fenótipo da cauda encurtada em camundongos heterozigotos, confirmando que o fenótipo observado é de fato originado pela redução da atividade do *T* (Stott *et al.*, 1993).

Neste estudo, uma linhagem de célula tronco embrionária de camundongo, derivada da massa interna do blastocisto, foi diferenciada *in vitro* para mesoderma. Em paralelo, uma linhagem da mesma célula foi também diferenciada mas com o *T* reduzido por um sistema de RNA de interferência. Estes dois cenários foram monitorados em nível transcricional através da tecnologia de *microarrays*. O objetivo foi investigar a função molecular do *T* durante a formação do mesoderma a partir da discriminação destas duas linhagens celulares.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

METODOLOGIA

Sistema de RNA de interferência do *T*

A diferenciação *in vitro* para mesoderma sem a presença do *T* foi possível através de um sistema *knockdown*, que é um dos aspectos chaves deste trabalho. Uma linhagem de célula tronco embrionária pluripotente (ESC, *Embryonic Stem Cell*) de camundongo foi gerada com a expressão de um repórter mCherry-*T*, construído geneticamente para conter um sistema de RNA de interferência (RNAi) para *T*, induzido por doxíciclina (Vidigal *et al.*, 2010). Neste estudo, eles mostraram que embriões que não foram tratados com doxíciclina (DOX) se desenvolveram normalmente (Figura A1a), no entanto, a indução do *knockdown* do *T* resultou em embriões gravemente comprometidos que sobreviveram até o 11º dpc (Figura A1b).

Figura A1: Efeito da indução do sistema de RNAi com doxíciclina. a) Desenvolvimento normal do embrião após 9.5 dpc - sem DOX. b) Embrião gravemente comprometido após 9.5 dpc - com DOX (Vidigal *et al.*, 2010).

Diferenciação *in vitro* de ESCs para mesoderma

A diferenciação *in vitro* de ESC para mesoderma foi feita em dois cenários distintos: WT (*wild-type*) e T-KD (*T-knockdown*). No primeiro, foi utilizada uma linhagem de ESC com a presença normal do *T*. No segundo, foi utilizada uma linhagem de ESC onde *T* foi reduzido com um RNAi.

Cinco diferentes tempos foram selecionados dos ensaios de diferenciação e hibridados nos *microarrays* Illumina Sentrix BeadChip Array - Mouse Ref-8 V2. Primeiro, as ESCs pluripotentes não diferenciadas; segundo, as ESC agregadas (aESC) que foram induzidas à diferenciação com Bmp4; terceiro, o tratamento após 48h de diferenciação, mas

Capítulo II

Busca por genes alvos do *T* com o programa hunT

antes do surgimento das células mCherry-*T* positivas; quarto, o tratamento após 72h de diferenciação, logo após as células começarem a transcrever o *T*; e por último, após 96h, quando as células se diferenciaram para mesoderma e uma grande quantidade de células com a expressão de *T* foi gerada.

Células mesodérmicas expressando mCherry-*T* - que codifica uma proteína vermelha fluorescente - puderam ser distinguidas e selecionadas por FACs (*Fluorescence Activated Cell Sorting*). Uma triplicata biológica foi produzida para cada um dos tempos da diferenciação, totalizando 42 *microarrays*. (Figura A2).

Figura A2: Desenho experimental dos ensaios de *microarray*. Os dois ramos no desenho indicam os diferentes cenários em que foi feita a diferenciação *in vitro*. No ramo superior não houve a adição de DOX e por isso a diferenciação ocorreu normalmente com a presença do *T*. Este cenário foi chamado de WT. No ramo inferior, ao contrário, houve a adição de DOX, que disparou a ação do RNAi e consequentemente interferiu na transcrição do *T*. Este cenário foi chamado de *T*-KD. As ESCs foram diferenciadas com Bmp4, foram então desagregadas nos tempos ilustrados na imagem e selecionadas por FACs. As linhas vermelhas indicam as células fluorescentes vermelhas (mCherry-*T*) que foram selecionadas após 48h.

Análise dos dados de transcriptoma

Os dados dos *microarrays* foram analisados com a ferramenta lumi disponível no pacote Bioconductor/R. Utilizamos os métodos quantil para a normalização dos dados, bgAdjust para a subtração do *background* e transformação logarítmica para a estabilização da variância. Aplicamos um filtro que selecionou 20% dos genes de maior variância. Para a seleção dos genes diferencialmente expressos (DEGs) empregamos o método SAM (*Significance Analysis of Microarrays*, Tusher *et al.*, 2001) seguindo o desenho de análise de duas classes não pareada, com valor de $p < 0,05$ ajustado por FDR (*False Discovery Rate*,

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Benjamini & Hochberg, 1995). Foram comparados os intervalos: ESCxaESC, aESCx48h, 48hx72h e 72hx96h e considerados somente os DEGs com *fold change* > 1.5.

Este modelo do SAM não é adequado para a análise de dados em série temporal pois ele ignora a natureza sequencial da expressão ao longo do tempo. Ele foi empregado nestas análises pois não estávamos interessados em identificar genes alterados nos tempos *t* e *t-n*, mas nos genes alterados em um dos intervalos da diferenciação. Nosso objetivo com esta estratégia foi identificar quais dos genes alterados num determinado intervalo das células com o *T* também se alteraram no mesmo intervalo das células sem o *T*. Para minimizar o efeito da maximização do erro deste procedimento, dedicamos maior esforço na etapa de anotação funcional dos genes, concentrando naqueles com funções conhecidamente associadas ao desenvolvimento. Para visualizar o perfil da expressão destes genes ao longo de toda a diferenciação e identificar grupos de genes com perfil similar, reunimos os genes resultantes das 4 comparações em um único conjunto que foi então analisado com uma ferramenta de agrupamento de dados.

Análise funcional e de agrupamento

A análise de agrupamento foi feita com o método CLICK, disponível na ferramenta Expander (Sharan *et al.*, 2003). O algoritmo se baseia na teoria de grafos para identificar grupos de elementos altamente similares (núcleos). Em seguida, ele aplica um procedimento heurístico para expandir os núcleos em grupos maiores. Empregamos a ferramenta IPA (*Ingenuity Pathway Analysis*, <http://www.ingenuity.com>) para a análise funcional dos DEGs. Este programa categoriza os genes em funções, redes biológicas e vias canônicas baseado em dados da literatura. Consideramos todas as espécies e tecidos de todos os repositórios de dados disponíveis.

RESULTADOS E DISCUSSÕES

Análise do transcriptoma da diferenciação de células WT

Foram detectados 48 DEGs na transição de ESC para aESC, 1.792 na transição de aESC para 48h, 547 DEGs de 48h para 72h e 346 DEGs na transição de 72h para 96h (Figura A3). O maior número de DEGs foi identificado no intervalo entre aESC e 48h, provavelmente devido ao forte impacto da diferenciação causado pelo *Bmp4*. A lista de todos os DEGs está

Capítulo II

Busca por genes alvos do *T* com o programa hunT

disponível no endereço: http://danieleyumi.sunaga.de/wp-uploads/DEGs_diferenciacao_WT.xls do material suplementar.

ESC 48 aESC 1792 48h 547 72h 346 96h WT

Figura A3: Quantidade de DEGs detectados pelo método SAM. Todos com $p > 0,05$ corrigidos por FDR e $fold\ change > 1,5$.

Após a detecção dos DEGs, verificamos o comportamento da expressão de alguns genes marcadores conhecidos, como forma de confirmar a reprodutibilidade dos *microarrays*. Genes de pluripotência conhecidos como o *Pou5f*, *Rex1*, *Esrrb* e *Fbxo15* foram expressos entre os estados pluripotentes ESC e aESC e apresentaram queda da regulação após a indução da diferenciação (Figura A4a). Genes que são conhecidos por se expressarem no epiblasto, como o *Fgf5*, *Otx2*, *Cdh1*, *Rab25* e *Enpp2* apresentaram aumento da expressão após 48h (Figura A4b). Genes ligados ao desenvolvimento do mesoderma lateral, como o *Myl4*, *T*, *Tnni1* e *Actc1*, mostraram-se aumentados após 72h, indicando que a formação do mesoderma foi induzida neste tempo (Figura A4c). Genes conhecidos por seu envolvimento na progressão do ciclo celular como o *p53*, *Ccnd1*, *Ccne1*, *Ccn1* e *Ccnc* apresentaram diminuição dos níveis de expressão durante o curso da diferenciação (Figura A4d). Por fim, verificamos a expressão dos genes que desempenham papel fundamental na aquisição de traços mesenquimais como o *Vim*, *Zeb1*, *Zeb2* e *Twist2*. Todos eles apresentaram aumento de expressão em 96h de diferenciação, quando é iniciada a formação do mesoderma lateral (Figura A4e).

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Figura A4: Avaliação do desempenho dos *microarrays* através da expressão de genes conhecidos envolvidos em alguma das etapas de diferenciação. O eixo *x* indica os tempos da diferenciação. O eixo *y* representa o nível de expressão relativa em percentual. O valor mais alto de expressão para cada gene foi setado como 100%, enquanto os demais foram calculados em relação a este valor. TD: *T-positive derivatives*.

Estes resultados indicam que a expressão de genes envolvidos na formação do mesoderma *in vivo* estão também representados nas células *in vitro*. Além disso, estes dados demonstram que o tratamento de ESCs *in vitro* com Bmp4 reflete os estágios sucessivos do desenvolvimento do mesoderma e por isso pode ser considerado um bom recurso para o estudo das mudanças do transcriptoma durante a formação do mesoderma. A análise funcional com o IPA identificou uma série de funções biológicas enriquecidas com os DEGs de cada intervalo da diferenciação. A Tabela A1 apresenta as funções mais significativas associadas ao desenvolvimento do mesoderma, endoderma e ectoderma.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Tabela A1: Funções biológicas do desenvolvimento que foram enriquecidas com os DEGs das células WT. Para cada transição: de aESC para 48h, 48h para 72h e 72h para 96h são apresentados o número de genes e seu respectivo valor de significância.

	aESC para 48h		48h para 72h		72h para 96h	
Funções do desenvolvimento	valor de p	# de genes	valor de p	# de genes	valor de p	# de genes
Mesodermal derivatives						
Development of blood vessels	1.10E ⁻⁰⁵	71	2.09E ⁻⁰⁹	38	7.47E ⁻⁰⁷	26
Development of muscle	4.59E ⁻⁰⁵	41	-	-	-	-
Development of heart	4.43E ⁻⁰⁴	35	6.80E ⁻⁰⁴	16	-	-
Development of lymphocytes	1.49E ⁻⁰³	75	-	-	-	-
Development of skeletal system	1.83E ⁻⁰³	26	-	-	-	-
Development of vasculature	-	-	-	-	4.96E ⁻⁰⁴	5
Development of artery	-	-	-	-	7.00E ⁻⁰³	4
Endodermal derivatives						
Development of lung	-	-	1.29E ⁻⁰⁶	14	1.03E ⁻⁰⁴	9
Development of bronchus	-	-	6.38E ⁻⁰⁵	3	-	-
Development of gastrointestinal tract	-	-	3.09E ⁻⁰⁴	6	-	-
Ectodermal derivatives						
Development of neurites	1.66E ⁻⁰³	35	-	-	3.13E ⁻⁰³	11

Estes resultados revelam que genes envolvidos no desenvolvimento do ectoderma (neuritos) e mesoderma (vasos sanguíneos, músculo, coração, linfócitos e esqueleto) foram diferencialmente expressos no primeiro intervalo da diferenciação. Entre 48h e 72h, genes ligados à formação do mesoderma (vasos sanguíneos, coração) e endoderma (pulmão, brônquios e trato gastrointestinal) foram detectados. Entre 72h e 96h, foram detectados genes ligados à formação do mesoderma (vasos sanguíneos, sistema vascular e artérias), do endoderma (pulmão) e do ectoderma (neuritos). Para a interpretação destes resultados, no entanto, é necessário considerar a tendência da expressão desses genes, se aumentada ou diminuída. Por exemplo, verificamos que a maioria dos genes diferencialmente expressos na transição de 72h para 96h - que enriqueceram as funções de desenvolvimento do pulmão e de neuritos - estão com a expressão diminuída, sugerindo que após 72h de diferenciação o desenvolvimento do endoderma e ectoderma é reprimido. Além disso, os genes associados ao desenvolvimento de vasos sanguíneos, sistema vascular e artérias, também apresentaram expressão diminuída neste último estágio da diferenciação, indicando também a repressão do desenvolvimento do sistema circulatório. Para verificar esta observação, genes conhecidos por desempenharem um papel no desenvolvimento vascular e do endoderma (*Gata6*, *Sox17*, *Foxq1*, *Spink3* e *Tmprss2*) foram selecionados para uma verificação mais detalhada. Estes genes apresentaram expressão aumentada em 72h e diminuição drástica da expressão em 96h

Capítulo II

Busca por genes alvos do *T* com o programa hunT

(Figura A5a). O mesmo comportamento de expressão foi observado nos genes que contribuem para a formação do mesoderma vascular (*Par1*, *Egfl7*, *Anxa3*, *Vegfa* e *Vegfr1*) (Figura A5b).

Figura A5: Perfil de expressão de genes envolvidos na formação do endoderma e mesoderma vascular. O eixo *x* indica os tempos da diferenciação. O eixo *y* representa o nível de expressão relativa em percentual. O valor mais alto de expressão para cada gene foi considerado como 100%, enquanto os demais foram calculados em relação a este valor. a) Perfil da expressão de genes associados ao desenvolvimento do endoderma. b) Perfil da expressão de genes associados ao desenvolvimento do mesoderma vascular.

Como forma de visualizar a dinâmica expressão temporal dos DEGs identificados, reunimos todos eles em um único conjunto que totalizou 2.733 genes. Os genes comuns a mais de uma comparação foram removidos. Os 2.195 genes restantes foram analisados com o método de agrupamento CLICK que resultou em 7 grupos de genes com perfil de expressão similar (Figura A6). A lista de genes está disponível no endereço: http://danieleyumi.sunaga.de/wp-uploads/DEGs_diferenciacao_WT.xls do material suplementar

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Figura A6: Agrupamento CLIC de todos os DEGs das células WT. a) *Heat map* dos 7 grupos. As linhas correspondem aos genes e as colunas aos tempos da diferenciação. Vermelho indica expressão aumentada e verde diminuída. b) Dinâmica do comportamento da expressão gênica de cada grupo.

Cerca de 74% dos genes foram alocados em um dos dois maiores grupos: I e II. Estes grupos apresentam perfil de expressão oposta, enquanto no grupo I os genes estão aumentados nas condições ESC e aESC e diminuídos após 48h, no grupo II os genes estão diminuídos nas condições iniciais e aumentados após 48h. Esta observação corrobora os resultados das nossas análises iniciais, onde o maior número de DEGs foi detectado na transição de aESC para 48h. Os demais grupos apresentaram alterações de expressão específicas, permitindo a visualização de alterações transientes em um ou outro ponto da diferenciação. O grupo III reúne genes com expressão diminuída nas condições iniciais, aumentada em 48h e novamente diminuída após 72h. Os grupos IV e VI reúnem genes com expressão diminuída até 48h, aumentada em 72h e novamente diminuída em 96h. O grupo V reúne genes com expressão diminuída em 72h e o grupo VII reúne poucos genes com expressão diminuída no período de 48h até 72h e aumentada em 96h.

Em resumo, os grupos III, IV, VI e VII apresentam significativa mudança de expressão em 48h, 72h e 96h, respectivamente. Já os genes dos grupos IV e VI são

Capítulo II

Busca por genes alvos do *T* com o programa hunT

especificamente transcritos em 72h de diferenciação, sugerindo que um importante processo molecular e celular específico de mesoderma ocorre neste tempo.

Os genes de cada grupo foram classificados pelo IPA de acordo com suas funções biológicas (Tabela A2).

Tabela A2: Funções molecular e celular enriquecidas com os genes dos 7 grupos.

Grupo I	valor de p	# de genes
Transcription	2.34E ⁻⁰⁷	107 / 421
Proliferation of cells	2.97E ⁻⁰⁷	134 / 421
Maintenance of stem cells	2.06E ⁻⁰⁵	6 / 421
Cell division process	7.81E ⁻⁰⁵	80 / 421
Grupo II	valor de p	# de genes
Movement of cells	5.75E ⁻¹¹	118 / 657
Migration of cells	1.20E ⁻¹⁰	116 / 657
Organization of cytoskeleton	1.48E ⁻¹⁰	39 / 657
Shape change	2.56E ⁻¹⁰	62 / 657
Grupo III	valor de p	# de genes
Specification of rostrocaudal axis	1.59E ⁻⁰³	2 / 43
Transport of molecule	2.44E ⁻⁰³	5 / 43
Transcription of DNA	6.31E ⁻⁰³	6 / 43
Grupo IV	valor de p	# de genes
Metabolic process of cholesterol	6.35E ⁻⁰⁸	9 / 109
Migration of eukaryotic cells	1.74E ⁻⁰⁷	28 / 109
Proliferation of cells	9.41E ⁻⁰⁶	38 / 109
Grupo V	valor de p	# de genes
Transcription of rRNA	1.40E ⁻⁰⁴	2 / 9
Transcription	2.76E ⁻⁰²	5 / 9
Grupo VI	valor de p	# de genes
Synthesis of lipids	1.05E ⁻⁰³	7 / 52
Growth of cells	3.93E ⁻⁰²	12 / 52
Proliferation of cells	2.71E ⁻⁰²	15 / 52
Grupo VII	valor de p	# de genes
Assembly of filaments	1.16E ⁻⁰³	3 / 17
Differentiation of cells	2.72E ⁻⁰²	6 / 17

O grupo I reúne genes com conhecidas propriedades de pluripotência. Divisão celular, proliferação e genes com funções na manutenção de células tronco foram identificados, como o *Esrrb*, *Fgf4*, *Klf4*, *Pou5f1*, *Sox2*. No grupo II estão os genes envolvidos na mudança do formato celular e também na transcrição, mobilidade e migração celular. Neste grupo estão genes como o *Cspr1*, *Efnb1*, *Epha4*, *Flna*, *Pdlim5* e *Srf*. O grupo III reúne genes que apresentam uma ativação transiente em diversos processos, como os genes *Lefty1* e *Nodal*,

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

característicos do estágio de epiblasto em camundongo, indicando que a população de células do epiblasto são estabelecidas após 48h de cultura *in vitro*. Além disso, genes envolvidos na transcrição do DNA e transporte de moléculas também foram expressados nesse tempo. Genes dos grupos IV e VI estão aumentados após 72h de diferenciação (Figura A6). O IPA associou estes genes ao metabolismo de colesterol, síntese de lipídios, migração, proliferação e crescimento celular. Colesterol e lipídio são componentes da membrana plasmática, sugerindo que a ativação destes processos ocorreu para aumentar a geração de membrana plasmática durante a formação do mesoderma lateral. No grupo V estão os genes associados à transcrição. Os genes do grupo VII estão associados à montagem de filamentos e indução da diferenciação. Neste grupo estão os genes *Vim* e o fator de transcrição *Zeb1*, característicos da aquisição de traços mesenquimais.

In vivo, a gastrulação e a subsequente formação do mesoderma é mediada pela interação das vias Tgf β , Wnt, Bmp e Fgf (Arnold & Robertson, 2009). Para verificar qual dessas vias foram reguladas na cultura *in vitro*, selecionamos os membros de cada uma delas e examinamos o comportamento de cada um durante o processo de diferenciação (Figura A7).

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Figura A7: Regulação das vias de sinalização. Cada gráfico ilustra o perfil de expressão de um membro de uma das vias indicadas no topo da imagem. O eixo *x* indica os tempos da diferenciação. O eixo *y* representa o nível de expressão relativa em percentual. O valor mais alto de expressão foi usado para calcular os níveis de expressão relativa.

Na via *Tgfβ*, *Nodal* é um gene chave. A expressão deste gene nos *microarrays* mostra que ele está presente desde o estado de ESC até 48h, sugerindo que esta via é ativada nas fases iniciais da diferenciação e não parece ser requerida para a formação do mesoderma e mesoderma lateral. Na via Wnt, encontramos os genes *Wnt6* e *Wnt7b* com expressão aumentada em 72h, sugerindo sua participação na formação do mesoderma. Na via BMP, o *Bmp4* apresentou notável aumento de expressão nos estados ESC e aESC, diminuição em 48h e aumentou novamente em 72h, sugerindo que o *Bmp4* é requerido durante estado pluripotente da célula e também exerce uma função na formação do mesoderma e do mesoderma lateral. O

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Bmp8b também apresentou aumento da expressão em 72h, indicando seu envolvimento na formação do mesoderma. A via de sinalização Bmp é conhecida por desempenhar múltiplos papéis durante o desenvolvimento. Nossos dados indicam que dentre esses papéis está a formação do mesoderma e mesoderma lateral. Com relação à via de sinalização Fgf, o *fgf4* apresentou aumento da expressão nos estados ESC e aESC e diminuição nos tempos seguintes, sugerindo um papel durante o estado pluripotente da célula. Os genes *fgf5* e *fgf8*, por outro lado, apresentaram aumento da expressão em 48h e 96h. O *fgf5* é um gene marcador de células do epiblasto (Figura A4). O aumento da expressão do *fgf5* e *fgf8* em 96h, contudo, aponta para o envolvimento destes genes também na formação do mesoderma lateral. O *fgf10* apresentou aumento da expressão em 72h mas diminuiu em 96h. O *fgf13* apresentou aumento da expressão em 48h e também diminuiu em 96h, sugerindo que estes genes não participam da formação do mesoderma lateral. Este perfil de expressão divergente entre os fgfs indicam que, assim como a via Bmp, a via Fgf também está envolvida em diferentes estágios do desenvolvimento.

De um modo geral, estes resultados sugerem que um único gene pode participar de vários processos do desenvolvimento; que as vias Tgf β , Wnt, Bmp e Fgf parecem regular e direcionar o desenvolvimento durante a formação do mesoderma. Especificamente, o gene *Bmp7* da via de sinalização Bmp, assim como os genes *fgf5* e *fgf8* da via Fgf, parecem estar envolvidos no processo de especificação de mesoderma para mesoderma lateral.

Para resumir os processos que ocorrem durante a formação do mesoderma lateral *in vitro*, nós compilamos o modelo que é apresentado na Figura A8.

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Figura A8: Representação esquemática das vias de sinalização envolvidas no desenvolvimento do mesoderma lateral *in vitro*. Nos círculos, estão os genes marcadores de cada etapa da diferenciação. As barras coloridas indicam quando os fatores de crescimento são expressados.

A análise dos dados de expressão revelou sucessivos estágios do desenvolvimento que ocorreram durante a diferenciação *in vitro*, iniciando com as ESCs pluripotentes, passando pelo epiblasto, mesoderma para finalmente das origem às células do mesoderma lateral. Nas ESCs não diferenciadas observamos a desregulação de genes marcadores de pluripotência como o *Klf4*, *Pou5f1* e o *Sox2*. Neste estágio, também observamos o aumento da expressão do gene *Nodal*, da via Tgf β (barra verde da Figura A8). Em 48h, as células já começaram a expressar genes que são transcritos no epiblasto, como o *Otx2*, *Nodal* e *Fgf5*. Neste estágio também observamos alterações na sinalização da via Bmp com os genes *Bmp1*, *Bmp4* e *Bmp7* (barra roxa - Figura A8). Nesse momento a expressão do *Nodal* passou a diminuir (barra verde - Figura A8). Após 72h de diferenciação, foram observadas células expressando os genes do endoderma (*Gata6*) e do mesoderma (*Tnncl*, *Vegfr1*). Além disso, genes de migração, proliferação celular e processamento de colesterol e lipídios foram transcritos neste tempo, conforme identificado na análise funcional dos grupos (Tabela A2). Ainda neste período da diferenciação, observamos a expressão de membros da via Wnt (*Wnt6* e *Wnt7b*, barra rosa - Figura A8) e a redução da sinalização de Fgf (barra amarela - Figura A8). Após 96h, as células WT apresentaram diminuição da transcrição dos genes do endoderma mas passaram a expressar o marcador mesenquimal *Vim* e genes específicos do mesoderma lateral. A sinalização da via Wnt desaparece (barra rosa - Figura A8) enquanto a sinalização de Bmp e Fgf aumentam (barra roxa e amarela - Figura A8), indicando o envolvimento destas vias na formação do mesoderma lateral.

Análise do transcriptoma da diferenciação de células *T*-KD: ausência do *T*

Para avaliar o efeito da redução do *T* na formação do mesoderma, comparamos a expressão de 96h de ambas as células (WT e *T*-KD) com a expressão das ESCs, que foram usadas como controles pois deram origem a ambas as linhagens.

Foram detectados 734 DEGs nas células WT, sendo 397 com expressão aumentada e 337 com expressão diminuída. Nas células *T*-KD foram detectados 897 DEGs, sendo 511 com expressão aumentada e 386 diminuída (Figura A9A). Estes resultados foram comparados para identificar os genes comuns e os específicos de cada condição (Figura A9b).

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Figura A9: DEGs da comparação entre ESC e 96h nas células WT e *T-KD*. a) As barras indicam a quantidade de DEGs. As cores vermelha e laranja indicam expressão aumentada e os dois tons de verde indicam expressão diminuída. b) Sobreposição do número de genes com expressão aumentada e diminuída.

Foram identificados 297 DEGs comuns entre os 397 e os 511 DEGs com expressão aumentada nas células WT e *T-KD*, respectivamente. Cem DEGs tiveram a expressão aumentada somente nas células WT e 214 somente nas células *T-KD*. Com relação aos DEGs com expressão diminuída, foram identificados 259 comuns entre as células WT e *T-KD*, 78 DEGs com a expressão diminuída somente na WT e 127 na *T-KD*. Também comparamos os DEGs aumentados na WT com os diminuídos na *T-KD* e vice-versa para verificar se houve alguma desregulação inversa entre as diferentes células, mas nenhuma sobreposição foi encontrada. Todas as listas de genes estão disponíveis no endereço: http://danieleyumi.sunaga.de/?page_id=23 do material suplementar.

Os dois conjuntos de DEGs (734 das células WT e 897 das células *T-KD*) foram analisados com o IPA. Na Tabela A3 estão as funções biológicas associadas ao desenvolvimento.

Tabela A3: Funções biológicas do desenvolvimento que foram enriquecidas com os DEGs das comparações de ESCx96h (células WT e *T-KD*).

	WT		T-KD	
Funções do desenvolvimento	valor de p	# de genes	valor de p	# de genes
Development of blood vessel	2.92E ⁻⁰⁹	47	7.13E ⁻⁰⁷	47
Development of muscle	5.13E ⁻⁰⁵	23	-	-
Development of heart	2.82E ⁻⁰⁴	20	-	-
Development of notochord	4.89E ⁻⁰⁴	4	-	-
Development of skeletal system	1.23E ⁻⁰³	15	-	-
Development of blood cells	-	-	3.76E ⁻⁰³	47

Capítulo II

Busca por genes alvos do *T* com o programa hunT

Os resultados do IPA revelaram uma diferença considerável entre as duas populações de células. Quando *T* estava presente, genes que participam do desenvolvimento de vasos sanguíneos, músculo, coração, notocorda e esqueleto foram detectados. Já na ausência do *T*, somente genes associados ao desenvolvimento de células e vasos sanguíneos foram detectados.

Estes resultados indicam que funções importantes do desenvolvimento são prejudicadas quando *T* está reprimido. Além disso, a sobreposição do número de DEGs das células WT e T-KD mostrou que cerca de duas vezes mais genes é alterada quando *T* está reprimido (214 aumentados e 127 diminuídos). Sendo o *T* um fator de transcrição, ele pode atuar como indutor e repressor de outros genes envolvidos na formação do mesoderma. Estes dados sugerem, portanto, que a falta do *T* pode ter levado ao maior número de genes aumentados pois é quando ele deveria tê-los reprimido e de genes diminuídos quando ele deveria tê-los induzido. Saber quais destes genes são regulados diretamente pelo *T* ajudaria a entender melhor as funções moleculares e vias de sinalização onde o *T* exerce papel que o torna fundamental para o desenvolvimento. A identificação de genes alvos, contudo, não é possível somente usando dados de transcriptoma.

O capítulo II apresenta o programa hunT, que foi desenvolvido para buscar por sítios de ligação de fatores de transcrição em sequências de DNA. Nossa expectativa é este programa nos ajude a identificar quais dos genes alterados nos *microarrays* são alvos do *T*.

REFERÊNCIAS BIBLIOGRÁFICAS

- Arnold SJ, Robertson EJ. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Biol*, 10:91–103.
- Benjamini Y & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.
- Packham EA & Brook JD. (2003). T-box genes in human disorders. *Human Molecular Genetics*, 12: R37–R44.
- Sharan R, Maron-Katz A, Shamir R. (2003). Click and Expander: a system for clustering and visualizing gene expression data. *Bioinformatics*, 19, 1787–1799.
- Stott D, Kispert A, Herrmann BG. (1993). Rescue of the tail defect of Brachyury mice. *Genes & Development*, 7:197–203.

Capítulo II

Busca por genes alvos do *T* com o programa *hunT*

Tusher VG, Tibshirani R, Chu G. (2001). Significance analysis of *microarrays* applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121.

Vidigal JA, Morkel M, Wittler L, Brouwer-Lehmitz A, Grote P, Macura K, Herrmann BG. (2010). An inducible RNA interference system for the functional dissection of mouse embryogenesis. *Nucleic acids research*, 38, e122.

Yanagisawa KO, Fujimoto H, Urushihara H. (1981). Effects of the brachyury (T) mutation on morphogenetic movement in the mouse embryo. *Developmental biology*, 87:242–248.

Conclusão geral

Embora haja uma série de programas e métodos disponíveis para a análise dos dados que vêm sendo produzidos em larga escala, a escolha de qual deles usar continua sendo uma tarefa desafiadora para os pesquisadores. O emprego de um método inadequado pode levar a resultados espúrios, comprometendo todo o investimento de um estudo. Uma boa prática, neste sentido, é conhecer diferentes métodos, preferencialmente que implementem diferentes abordagens estatísticas, e, então, escolher o que melhor se enquadra ao problema biológico em questão ou concluir que o melhor a fazer é desenvolver um novo.

Neste trabalho, nós mostramos que o popular método SAM de seleção de genes diferencialmente expressos não é uma boa opção para a análise de dados de doenças complexas, principalmente quando o tamanho amostral é pequeno. Ele deixou de selecionar genes com expressão consistentemente alterada em nossos testes usando dados simulados. Em contrapartida, mostramos que a característica de *ranking* do método Rank Products o tornou capaz de lidar com a variabilidade da expressão gênica entre indivíduos com uma mesma doença. Ele conseguiu detectar genes desregulados em apenas uma porção destes indivíduos e esta habilidade foi mantida em estudos com poucas amostras. Esta última observação é especialmente importante, uma vez que a obtenção de material para o estudo dessas doenças não é trivial.

Já para a tarefa de identificar genes alvos de fatores de transcrição, nós desenvolvemos um novo programa pois, dentre várias opções disponíveis, não encontramos um que atendesse a todas as nossas necessidades. O *hunT* foi adequado ao conhecimento existente sobre o modo como o fator de transcrição *T* regula seus alvos. Sua capacidade de buscar por genes com sítios de ligação para diferentes fatores e de proteção contra *mismatches* levou à identificação de conhecidos alvos do *T* assim como de candidatos interessantes. Comparado com uma ferramenta equivalente, os resultados do *hunT* foram mais precisos pois atenderam à especificação de uma combinação de diferentes parâmetros. Este teste inclusive mostrou que se tivéssemos usado esse programa ao invés do *hunT*, seriam necessárias etapas adicionais de análise para apontar os genes com características que já conhecíamos *a priori*. O desenvolvimento do *hunT* permitiu que isso fosse feito automaticamente.

Resumo geral

A enorme quantidade de dados que vem sendo gerada por tecnologias modernas de biologia representam um grande desafio para áreas como a bioinformática. Há uma série de programas disponíveis para a análise destes dados, mas que nem sempre são compreendidos o suficiente para serem corretamente aplicados, ou ainda, há problemas que requerem o desenvolvimento de novas soluções. Neste trabalho, nós apresentamos a análise de dados de duas das principais fontes de dados em larga escala: *microarrays* e sequenciamento. Na primeira, avaliamos se a estatística do método Rank Products (RP) é adequada para a identificação de genes diferencialmente expressos em estudos de doenças complexas, cujo uma das características é a heterogeneidade genética entre indivíduos com o mesmo fenótipo. Na segunda, desenvolvemos uma ferramenta chamada *hunT* para buscar por genes alvos do fator de transcrição *T* - um importante marcador de mesoderma com papel chave no desenvolvimento de vertebrados -, através da identificação de sítios de ligação para o *T* em suas sequências reguladoras. O desempenho do RP foi testado usando dados simulados e dados reais de um estudo de fissura lábio-palatina não-sindrômica, de autismo e também de um estudo que avalia o efeito da privação do sono em humanos. Nossos resultados mostraram que o RP é uma solução eficiente para detectar genes consistentemente desregulados em somente um subgrupo de pacientes, que esta habilidade é mantida com poucas amostras, mas que o seu desempenho é prejudicado quando são analisados poucos genes. Obtivemos fortes evidências biológicas da eficiência do método nos estudos com dados reais através da identificação de genes e vias previamente associados às doenças e da validação de novos genes candidatos através da técnica de PCR quantitativo em tempo real. Já o programa *hunT* identificou 4.602 genes de camundongo com o sítio de ligação para o domínio do *T*, sendo alguns deles já demonstrados experimentalmente. Identificamos 32 destes genes com expressão alterada em um estudo onde avaliamos o transcriptoma da diferenciação *in vitro* de células tronco embrionárias de camundongo para mesoderma, sugerindo a participação destes genes neste processo sendo regulados pelo *T*.

Abstract

The large amount of data generated by modern technologies of biology provides a big challenge for areas such as bioinformatics. In order to analyze these data there are several computer programs available; however these are not always well understood enough to be correctly applied. Moreover, there are problems that require the development of new solutions. In this work, we present the data analysis of two main high-throughput data sources: microarrays and sequencing. Firstly, we evaluated whether the statistic of Rank Products method (RP) is suitable for the identification of differentially expressed genes in studies of complex diseases, which are characterized by the vast genetic heterogeneity among the individuals affected. Secondly, we developed a tool named *hunT* to search for target genes of *T* transcription factor - an important mesodermal marker that plays a key role in the vertebrate development -, by identifying binding sites for *T* in their regulatory sequences. The RP performance was tested using both simulated and real data from three different studies: non-syndromic cleft lip and palate, autism and sleep deprivation effect in Humans. Our results have shown that RP is an effective solution for the identification of consistently deregulated genes in a subgroup of patients, this ability is maintained even with few samples, however its performance is impaired when only few genes are analyzed. We have obtained strong biological effectiveness of the method in the studies with real data by not only identifying genes and pathways previously associated with diseases but also corroborating the behavior of novel candidate genes with the real-time PCR technique. The *hunT* program has identified 4,602 mouse genes containing the binding site for the *T* domain, some of which have already been demonstrated experimentally. We identified 32 of these genes with altered expression in a study which evaluated the transcriptome of *in vitro* differentiation of mouse embryonic stem cells to mesoderm, suggesting the involvement of these genes in this process regulated by *T*.

Referências Bibliográficas

Dean J & Ghemawat S. (2008). MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51(1): 107-113.

Breitling R, Armengaud P, Amtmann A, Herzyk P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573:83-92.

Gurtowski J, Schatz MC, Langmead B. (2012). Genotyping in the cloud with Crossbow. *Curr Protoc Bioinformatics*, 39:15.3.1-15.3.15.

Langmead B, Hansen KD, Leek JT. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology*, 11:R83.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9): 1297-1303.

Matsunaga A, Tsugawa M, Fortes J. (2008). CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications. *Fourth IEEE International Conference on eScience* 2008.

Zhang L, Gu S, Liu Y, Wang B, Azuaje F. (2012). Gene set analysis in the cloud. *Bioinformatics*, 28(2):294-295.

Produção científica durante o doutorado

Lista de artigos

- 1) Bueno DF*, Sunaga DY*, Kobayashi GS, Agüena M, Raposo-Amaral CE, Masotti C, Cruz LA, Pearson PL, Passos-Bueno MR. (2011). **Human Stem Cell Cultures from Cleft Lip/Palate Patients Show Enrichment of Transcripts Involved in Extracellular Matrix Modeling By Comparison to Controls.** *Stem Cell Rev and Rep*, 7(2):446-457. *os autores contribuíram igualmente.
- 2) Pellegrino R, Sunaga DY, Guindalini C, Martins RCS, Mazzotti DR, Wei Z, Daye J, Andersen ML, Tufik S. (2012). **Whole-blood genome-wide gene expression profile in males after prolonged wakefulness and sleep.** *Physiological Genomics*, 44(21):1003-1012.
- 3) Griesi-Oliveira K, Sunaga DY, Cruz LA, Vadasz E, Passos-Bueno MR. **Expression analysis in dental pulp stem cells of idiopathic autistic patients reveals alterations in important processes for neurogenesis**, manuscrito em preparação.
- 4) Karina Griesi-Oliveira, Allan Acab, Daniele Yumi Sunaga, Xavier Nicol, Nicole Davis, Kristin Rose, Yanelli Nunez, Stephan Sanders, Christopher Mason, Estevão Vadasz, Alexander Dietrich, Maria Rita Passos-Bueno, Matthew W. State, Alysson Muotri. **Identification of TRPC6 gene, a Ca²⁺ ion channel, as a novel candidate gene for autism and modeling neuronal development in vitro using patient's cells**, manuscrito em preparação.
- 5) Kobayashi GS, Alvizi L, Sunaga DY, Francis-West P, Kuta A, Almada BVP, Ferreira SG, Andrade-Lima LC, Bueno DF, Raposo-Amaral CR, Menck CF, Passos-Bueno MR. **Susceptibility to DNA Damage as a Molecular Mechanism for Non-Syndromic Cleft Lip and Palate**, manuscrito em preparação.
- 6) Erika Yeh, Roberto Dalto Fanganiello, Daniele Yumi Sunaga, Xueyan Zhou, Gregory Holmes, Katia Maria Rocha, Nivaldo Alonso, Hamilton Matushita, Yingli Wang, Ethylin W. Jabs, Maria Rita Passos-Bueno. **Novel molecular pathways elicited by mutant FGFR2 may account for brain abnormalities in Apert syndrome**, manuscrito em preparação.

Produção científica durante o doutorado

Trabalhos apresentados em congressos

1. **Non-syndromic cleft lip/palate and DNA repair.** Kobayashi GS, Alvizi L, Sunaga DY, Francis-West P, Kuta A, Almada BVP, Ferreira SG, Andrade-Lima LC, Bueno DF, Raposo-Amaral CR, Menck CF, Passos-Bueno MR. The 62nd Annual Meeting of the American Society of Human Genetics. San Francisco, CA, USA, 2012.
2. **Non-syndromic cleft lip/palate: a disease driven by deficiency in genomic repair?** Kobayashi GS, Alvizi L, Sunaga DY, Francis-West P, Kuta A, Almada BVP, Ferreira SG, Andrade-Lima LC, Bueno DF, Raposo-Amaral CR, Menck CF, Passos-Bueno MR. The Annual Meeting of the Society of Craniofacial Genetics and Developmental Biology. San Francisco, CA, USA, 2012.
3. **Modulation of autophagy pathway gene expression after crotonamine treatment of tumor cell line.** Márcia Neiva, Camila M Yonamine, Marcela B Nering, Eduardo B Oliveira, Daniele Y Sunaga, Mirian A F Hayashi. In: 10th International Congress on Cell Biology and the XVI Meeting of the Brazilian Society for Cell Biology, Rio de Janeiro, RJ, Brazil, July 25-28, 2012.
4. **The whole-blood genome-wide gene expression profiles in human volunteers after sleep deprivation.** R. Pellegrino, C.S. Guindalini, D.Y. Sunaga, M.L. Andersen, R.C. Martins, K.S. Barrueco, S. Tufik. In: 20th Congress of the European Sleep Research Society, Lisbon, Portugal, September 14-18, 2010.
5. **Identificação de vias de sinalização associadas à predisposição às fissuras lábio-palatinas não sindrômicas.** Lucas Alvizi Cruz, Daniele Yumi Sunaga, Daniela Franco Bueno, Gerson Shigeru Kobayashi, Meire Agüena, Simone Ferreira, Cássio Eduardo Raposo do Amaral, Luciano Brito, Maria Rita Passos-Bueno. In: 56th Congresso Brasileiro de Genética, Guarujá, SP, Brazil, September 14-17, 2010.
6. **Efeito da sincronização celular na análise do transcriptoma de culturas celulares de pacientes portadores de fissura lábio-palatina.** Kobayashi G.S.,

Produção científica durante o doutorado

Sunaga D.Y., Bueno D.F., Cruz L.A., Ferreira S.G., Passos-Bueno M. In: 56th Congresso Brasileiro de Genética, Guarujá, SP, Brazil, September 14-17, 2010.

7. Use of Stem cells to Identify novel signaling pathways and candidate genes associated with Non Syndromic Cleft Lip and Palate (NSCL/P).

Bueno D.F., Sunaga D.Y., Kobayashi G.S., Amaral C.E.R., Agüena M., Tanikawa D., Alonso N., Passos-Bueno M.R. In: 16th International Society of Developmental Biologists Congress, Edinburgh, Scotland, September 6–10, 2009.

8. A relevant chondrogenesis pathway is identified through transcriptome analysis in stem cells of a patient with condyle hypoplasia, severe micrognathia and auricular defects (Auriculo Condylar Syndrome).

Vanessa L.R. Tavares, Daniele Yumi Sunaga, Daniela F. Bueno, Roberto D. Fanganiello, Maria Rita Passos-Bueno. In: Gordon Research Conferences, 2009.

9. Análise do transcriptoma de células-tronco mesenquimais de pacientes com Síndrome Aurículo Condilar (ACS1) durante a condrogênese.

Tavares V.L.R., Sunaga D.Y., Bueno D.F. Passos-Bueno M.R. In: 55th Congresso Brasileiro de Genética, Águas de Lindóia, SP, Brazil, August 30-02, 2009.

10. Estudo de expressão gênica entre os sexos e suas implicações na busca por genes candidatos para as fissuras lábio-palatinas não-sindrômicas.

Kobayashi G.S., Sunaga D.Y., Bueno D.F., Cruz L.A., Ferreira S.G., Passos-Bueno M. In: 55th Congresso Brasileiro de Genética, Águas de Lindóia, SP, Brazil, August 30-02, 2009.

11. Differential expression signature in Stem Cells of non-syndromic cleft lip/cleft palate patients: A new approach to understand this orofacial malformation?

D.F. Bueno, G. Kobayashi, D.Y. Sunaga, C. Amaral, I. Zambra, M. Agüena, I. Kerkis, M.R. Passos-Bueno. In: 58th The American Society of Human Genetics, Philadelphia, EUA, November 11-15, 2008.

Produção científica durante o doutorado

12. **Expression signature in stem cells of patients with non syndromic cleft lip with or without cleft palate (NSCL/P): A new approach to reveal the signaling pathways associated with this orofacial malformation?** D.F. Bueno, G. Kobayashi, D.Y. Sunaga, C. Amaral, I. Zambra, M. Agüena, I. Kerkis, M.R. Passos-Bueno. In: 58th The American Society of Human Genetics. Philadelphia, EUA, November 11-15, 2008.

13. **Tumor responses to chemotherapy: searching for gene expression signatures that may explain why treatment often fails.** Costa P.L.N; Bigaton F.J.; Sunaga D.Y.; Nonogaki S.; Nantel F.; Battistini, B.; Sirois P.; Chammas R. In: XIV Congresso Brasileiro de Biologia Celular, Sao Paulo, SP, Brazil: Elsevier, p114, 2008.