

IDENTIFICAÇÃO E ANÁLISE DE NOVOS SNPs NÃO SINÔNIMOS UTILIZANDO ORESTES

PEDRO EDSON MOREIRA GUIMARÃES

Dissertação apresentada à Fundação Antônio Prudente
para a obtenção do título de Mestre em Ciências.

Área de concentração: Oncologia

ORIENTADOR: PROFº. DR.
EMMANUEL DIAS NETO

CO-ORIENTADORA: PROFª. DR.
MARIA CRISTINA RAMOS COSTA

São Paulo
2003

EXEMPLAR
ESPECIAL

Fundação Antônio Prudente
Ana Maria Rodrigues Alves Kunze
Coordenadora Pós-Graduação

FUNDAÇÃO ANTONIO
PRUDENTE
BIBLIOTECA

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa
Hospital do Câncer A.C. Camargo

Guimarães, Pedro Edson Moreira

Identificação e análise de novos SNPs não sinônimos utilizando Orestes /
Pedro Edson Moreira Guimarães -- São Paulo, 2003.

154p.

Dissertação(mestrado)-Fundação Antônio Prudente.

Curso de Pós-Graduação em Ciências-Área de concentração: Oncologia.

Orientador: Emmanuel Dias Neto

Descritores: 1. ORESTES. 2. ESTs. 3. SNPs. 4. NEOPLASIAS DE CABEÇA
E PESCOÇO. 5. DOENÇA DE ALZHEIMER. 6. SCHISTOSOMA MANSONI.



Ministério da Ciência e Tecnologia



“Se muito vale o já feito, mais vale o que será.

E o que foi feito é preciso conhecer para melhor prosseguir”

Milton Nascimento e Fernando Brant

*A meus amados pais, Filomeno e Maria Cleusa,
exemplos de luta, dedicação, fé e integridade,
e irmãos, Paulo, Eliana e Frederico,
meus fiéis companheiros de caminhada,
todos presentes divinos, sem os quais a vida não teria o mesmo sabor.*

*A meu Grande Amor, Marília,
quem dá um toque de graça a minha vida
e que, apesar da distância, sempre esteve ao meu lado.*

Ao Emmanuel,

mais que um orientador, foi pai, irmão e amigo.

*Agradeço também por ter me presenteado com sua inteligência e por ter
me mostrado que, apesar das inúmeras dificuldades que sempre encontramos,
dignidade, honestidade e companheirismo são, acima de tudo,
os maiores valores de um grande homem.*

À Cris,

Orientadora mais que competente e dedicada.

*Amiga sempre alegre e paciente, mesmo às 4 da manhã, depois de
muito trabalho, fome, sono e com o Rodrigues para “ajudar”.*

Agradecimentos

Ao Eduardo e ao CH, meus grandes e bons amigos, sempre dispostos a temperar com humor, a nem sempre tranqüila, rotina de cada dia;

Aos integrantes do Carrossel, Breno, Carlos Henrique, Casssy, Eduardo, Elildo, Flávio e Júlio, pelas “discussões científicas” que tanto ajudaram, direta ou indiretamente, meu amadurecimento pessoal e profissional;

A minha segunda família, Irani e Joelma, por todo o carinho e atenção recebidos;

Aos amigos de Ribeirão Preto, Adriano, Caetano, Cris, Francis, Giziane, Israel, Karine, Kiyoko, Marlize, Rochele e Virgínia, uma das maiores conquistas deste Mestrado;

Aos “meninos” Alexandre Hamilton, Cláudia, Dani, Denise, Eduardo Tarazona, Fabrício, Glória, Márcio, Rinaldo Bacamarte, Tadeu e Túlio, que mais uma vez sempre estiveram comigo;

À Carol, ao Gabriel, à Ana Maria, à Vó Tininha e ao Fausto por todo o carinho, atenção e dedicação dados a este não tão grato hóspede;

A todos do GREA, pelo ambiente sempre agradável, acolhedor e descontraído de trabalho;

A meus colegas de Pós-graduação, com muito carinho ao Lúcio, Marcilei e Renata;

Aos filhos de Emmanuel, Alexandre, Anísio, Carol, Cíntia, Élide e Sheila, que possamos cada vez mais caminhar como um só grupo;

À Carolina Dias pelo auxílio com os resumos em inglês;

Ao parecerista anônimo da FAPESP, por seus comentários sempre gratificantes e estimulantes;

Ao Instituto Ludwig de Pesquisa sobre o Câncer, de forma especial a Ana Paula, Anamaria, Tia Ceíça, Seu Eudes, Fabrício, Gaëlle, Lourdes, Ricardo Gabriel e Ricardo Moura que tanto ajudaram no início deste Mestrado;

A todos da Biblioteca do Hospital do Câncer, de modo muito especial à Sueli;

À Pós-graduação do Hospital do Câncer, especialmente à Márcia e à Ana, pela paciência e competência;

À Dra. Maria do Rosário Dias de Oliveira Latorre e à Dra. Karina de Cássia Braga Ribeiro pelo auxílio com as dúvidas em estatística;

Ao Dr. Luiz Paulo Kowalski, Cirurgião Chefe do Depto de Cabeça e Pescoço e Otorrinolaringologia do Hospital do Câncer/Fundação Antônio Prudente, pelas amostras tumorais utilizadas neste projeto e por toda a atenção e receptividade que sempre encontrei;

Ao grupo Jamboree-Orestes, pela oportunidade de discussões relevantes e enriquecedoras;

Ao Dr. Marco Antonio Zago, ao Dr. Wilson Araújo da Silva Jr. e à Dra Maria Cristina Costa, Coordenadores do grupo de sequenciamento RC e Coordenadores da iniciativa de análise de SNPs dentro do Projeto Genoma do Câncer Ludwig/FAPESP, pela colaboração que possibilitou o bom andamento desta tese;

Ao Departamento de Psiquiatria do Hospital das Clínicas da Faculdade de Medicina da USP, em particular ao Dr. Wagner Farid Gattaz;

À toda a Equipe do Laboratório de Neurociências - LIM27, meus companheiros neste último ano de Mestrado;

À Paula Kageyama pela sempre otimista visão da vida;

À equipe do dbSNP pelo auxílio com os dados desse banco;

À *Applied Biosystems*, em particular ao Alessandro e de forma mais que especial ao Rinaldo, que esta colaboração continue prosperando junto a nossas discussões e “viagens” científicas;

À toda a Equipe do Projeto Genoma de *Schistosoma mansoni*, pelo alto padrão de discussões que tanto me ensinaram durante todo o decorrer do projeto;

Ao Dr. Arthur Gruber, pelo auxílio com os experimentos de padronização da PCR em *multiplex*;

Ao Dr. David Rollinson, do Museu de História Natural de Londres, à Dra. Norma Lucena e ao Dr. Mittermayer Reis, ambos da FIOCRUZ, que gentilmente cederam as amostras de *Schistosoma mansoni* utilizadas nas análises dos SNPs;

À Dra. Erika Meirelles Kalil Pessoa de Barros e ao Dr. Tarcísio Eloy Pessoa de Barros Filho do Instituto de Ortopedia e Traumatologia do Hospital das Clínicas da Faculdade de Medicina da USP pelas amostras de pacientes com Paralisia Cerebral, Trauma Raquimedular e seus respectivos controles;

À Associação Beneficente Alzira Denise Hertzog Silva (ABAHDS), pelo contínuo apoio que recebemos;

À FAPESP e ao Ministério de Ciência e Tecnologia pelo auxílio financeiro ao Projeto Genoma de *Schistosoma mansoni*;

À FAPESP pela bolsa de mestrado, sem a qual este trabalho não teria sido possível;

A todos aqueles que direta ou indiretamente contribuíram para o desenvolvimento desta tese;

A Deus e a meu Anjo da Guarda.

Prefácio

Esta dissertação teve seu início no Laboratório de Descoberta Gênica, do Instituto Ludwig de Pesquisa sobre o Câncer, onde nosso grupo permaneceu até dezembro de 2001; a partir desta data nos transferimos para o Laboratório de Neurociências (LIM-27) do Departamento e Instituto de Psiquiatria do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.

A mudança de instituição nos abriu novos horizontes, novos campos de pesquisa e nos proporcionou novas colaborações, os quais se refletiram no decorrer deste projeto de mestrado.

O objetivo inicial desta dissertação estava restrito à avaliação de polimorfismos de nucleotídeo único em pacientes com Tumores de Cabeça e Pescoço, buscando associar os achados aos diferentes aspectos clínicos destes indivíduos.

Devido às características das informações que usamos, dados recentes derivados das ORESTES geradas no contexto do Projeto Genoma do Câncer Humano (HCGP), antes de iniciarmos nossa busca de associação genótipo/fenótipo nos deparamos com a necessidade de trabalhar na avaliação da qualidade/quantidade de SNPs derivados das ORESTES, um objetivo precedente ao anterior, e que naturalmente se inseriu no contexto do Mestrado.

O projeto foi assim crescendo, junto com nossa curiosidade e nossas colaborações, expandindo o painel de sub-projetos onde os SNPs identificados pudessem ser avaliados.

Não somente os dados foram testados, a capacidade de gerar informações sobre polimorfismos a partir de ORESTES também o foi. Para isso, utilizamos o

Projeto Genoma de *Schistosoma mansoni* como um segundo conjunto de dados para a identificação de SNPs.

No decorrer deste Mestrado a dissertação foi então adaptada de modo a incluir as novas possibilidades que surgiam, e permitiam vislumbrar as diversas aplicabilidades resultantes de análises de polimorfismos. Para facilitar a compreensão dos trabalhos realizados e para melhor dispor os dados obtidos, esta dissertação será apresentada em duas partes.

Na primeira temos a etapa de geração dos dados, referentes aos polimorfismos por nós identificados, e na segunda são apresentadas as informações que pudemos extrair destes dados.

Lista de figuras

Figura 1 - Identificação de um SNP.....	10
Figura 2 - Distribuição dos SNPs humanos já identificados por cromossomo	11
Figura 3 - Polimorfismos sinônimos e não sinônimos	14
Figura 4 - Comparação da distribuição posicional relativa das seqüências ORESTES e ESTs em seqüências de cDNA completas	20
Figura 5 - Esquema da rotina de bioinformática desenvolvida para a identificação de SNPs	29
Figura 6 - Falso SNP gerado por erro de seqüenciamento	33
Figura 7 - Reações de padronização da PCR	35
Figura 8 - Avaliação do efeito da retirada do passo de extensão sobre a PCR	37
Figura 9 - Alinhamento de ORESTES sugerindo a existência de um SNP.....	50
Figura 10 - Resultado do Primer3.....	53
Figura 11 - Validação do SNP no gene da proteína similar a Podocalixina por RFLP.....	56
Figura 12 - Validação por seqüenciamento do SNP presente no gene KiSS-1	57
Figura 13 - Exemplo de SNP encontrado em gene candidato a vacina para <i>Schistosoma mansoni</i>	71
Figura 14 - Validação do SNP em GST de <i>Schistosoma mansoni</i>	73
Figura 15 - Primeiros resultados obtidos com o kit <i>SNaPshot</i>	77
Figura 16 - Genotipagem com o Kit <i>SNaPshot</i> em reação de triplex	78
Figura 17 - Redução do volume de reação de genotipagem	80
Figura 18 - Tentativa de genotipagem em sistema de pentaplex	81
Figura 19 - Experimento de padronização da PCR em sistema multiplex.....	84
Figura 20 - Distribuição da freqüência do alelo variante de GST nos diferentes isolados seqüenciados	129
Figura 21 - Estruturas preditas pelo SWISS-MODEL da proteína RAB1 com e sem o SNP	131
Figura 22 - Alteração de aminoácido na seqüência do gene TIM de <i>Schistosoma mansoni</i>	132



Lista de tabelas

Tabela 1 - Sumário dos polimorfismos depositados no dbSNP	15
Tabela 2 - Número total de dados utilizados na identificação de SNPs e porcentagens de <i>clusters</i> onde estas variações foram encontradas.....	45
Tabela 3 - Classificação dos candidatos a SNPs identificados antes do uso do filtro de duas bibliotecas	47
Tabela 4 - Classificação dos candidatos a SNPs identificados após o uso do filtro de duas bibliotecas	48
Tabela 5 - SNPs validados	58
Tabela 6 - Frequência dos SNPs validados em amostras populacionais	59
Tabela 7 - SNPs identificados a partir das ORESTES geradas pelo PGSM nos genes a candidatos vacinais de <i>Schistosoma mansoni</i>	70
Tabela 8 - Distribuição de seqüências de cabeça e pescoço por localização topográfica e histologia.....	110
Tabela 9 - Genótipos dos pacientes com câncer de cabeça para IP30, Tax1 e KiSS-1	125
Tabela 10 - Genótipos dos pacientes com trauma raquimedular e controles para Ninjurina 1 e Alox12.....	vide Anexo 3

Lista de abreviações

- cDNA – (do inglês: *complementary DeoxyriboNucleic Acid*) Ácido desoxirribonucleico complementar
- CDS – (do inglês: *CoDing Sequence*) Região codificadora
- CGAP – (do inglês: *Cancer Genome Anatomy Project*) Projeto de Anatomia do Genoma do Câncer
- CIAP – (do inglês: *Calf Intestinal Alkaline Phosphatase*) Fosfatase alcalina intestinal de bezerro
- DA – Doença de Alzheimer
- dbSNP – (do inglês: *data base SNP*) Banco de dados de SNP
- dNTP – (do inglês: *desoxy riboNucleotides Tri-Phosphate*) Desoxiribonucleotídeos trifosfato
- EST – (do inglês: *Expressed Sequence Tag*) Etiqueta de sequência expressa
- Exo I – Exonuclease I
- FM – Faculdade de Medicina
- HapMap – (do inglês: *Haplotype Map*) Projeto Mapa de Haplótipos Humanos
- HC – Hospital das Clínicas
- HCGP – (do inglês: *Human Cancer Genome Project*) Projeto Genoma do Câncer Humano
- mRNA – (do inglês: *messenger RiboNucleic Acid*) Ácido ribonucleico mensageiro ou RNA mensageiro
- NCBI – (do inglês: *National Center for Biotechnology Information*) Centro Nacional para Informação em Biotecnologia
- OMS – Organização Mundial da Saúde
- OR – (do inglês: *Odds Ratio*)
- ORESTES – (do inglês: *Open Reading Frame ESTs*) ESTs derivadas da região codificadora dos genes
- pb – pares de base
- PCR – (do inglês: *Polymerase Chain Reaction*) Reação em Cadeia da Polimerase
- PGSM – Projeto Genoma de *Schistosoma mansoni*
- RFLP – (do inglês: *Restriction Fragment Length Polymorphism*) Polimorfismo de tamanho do fragmento de restrição
- S. mansoni* – *Schistosoma mansoni*
- SCI – (do inglês: *Spinal Cord Injury*) Injúria na coluna vertebral ou trauma raquimedular

SNC – Sistema Nervoso Central

SNP – (do inglês: *Single Nucleotide Polymorphism*) Polimorfismo de Nucleotídeo Único

T_m – (do inglês: *melting Temperature*) Temperatura média de desnaturação

TSC – (do inglês: *The SNP Consortium*) Consórcio de SNP

U – unidade

USP – Universidade de São Paulo

Índice

PARTE I - Identificação de novos SNPs não sinônimos utilizando ORESTES	1
Resumo.....	2
<i>Summary</i>	5
Introdução	8
I.1.1 Polimorfismos de nucleotídeo único	9
I.1.2 Bancos de dados de SNPs	13
I.1.3 Métodos de genotipagem	16
I.1.4 Aplicações	17
I.1.5 ORESTES: informações derivadas de regiões codificadoras de genes.....	18
I.1.5.1 O uso de ORESTES para investigar polimorfismos genéticos.....	21
I.1.5.2 A Geração de ORESTES para a análise do transcriptoma de tumores humanos.....	22
I.1.5.3 A Geração de ORESTES para a análise do transcriptoma de <i>Schistosoma mansoni</i> ..	23
I.1.6 SNPs em genes de <i>Schistosoma mansoni</i> candidatos a vacina	24
Objetivos	25
I.2.1 Objetivo geral.....	26
I.2.2 Objetivos específicos	26
Material e Métodos	27
I.3.1 Seleção inicial dos SNPs humanos	28
I.3.1.1 Estruturação dos dados a serem analisados	28
I.3.1.2 Alinhamento das ORESTES e das seqüências âncora	30
I.3.1.3 Identificação dos SNPs humanos	30
I.3.1.4 Resultados e estatísticas.....	31
I.3.1.5 Busca em bancos de dados de SNPs	31
I.3.1.6 Montagem do banco de dados.....	31
I.3.2 Seleção final dos SNPs	32
I.3.3 Mapeamento genômico e desenho dos iniciadores	32
I.3.4 Padronização das amplificações	34
I.3.5 Validação dos SNPs	36

I.3.5.1 Amostras humanas	38
I.3.6 SNPs em genes de <i>Schistosoma mansoni</i> candidatos a vacina	38
I.3.7 Mapeamento gênico e desenho dos iniciadores.....	39
I.3.8 Implementação de novas técnicas de genotipagem	40
I.3.8.1 PCR <i>multiplex</i>	41
Resultados e Discussão	43
I.4.1 Análise dos SNPs.....	44
I.4.2 Mapeamento genômico e desenho de iniciadores	49
I.4.3 Padronização das amplificações	52
I.4.4 Validação dos SNPs em amostras populacionais.....	54
I.4.4.1 Tax1 (Vírus tipo 1 da leucemia de célula T)	60
I.4.4.2 Ladinina (LAD).....	60
I.4.4.3 Inibidor específico de elastase (Elafin).....	61
I.4.4.4 Colagenase tipo IV	61
I.4.4.5 Inibidor de cisteino-protease.....	62
I.4.4.6 Proteína induzida por interferon Gama (IP-30)	62
I.4.4.7 Proteína similar a Podocalixina	63
I.4.4.8 Ninjurina 1	64
I.4.4.9 Proteína ligadora de estimulador de adipócitos 1 (AEBP1).....	64
I.4.4.10 Gene da Síndrome de Usher tipo 1C (USH1C)	65
I.4.4.11 Nidogene	65
I.4.4.12 Serino protease (SerProt).....	65
I.4.4.13 Proteína coativadora do receptor de hormônio da tireóide (THRCP).....	66
I.4.4.14 Membro 2 da família de antígenos carcinoembrionários (CGM2).....	66
I.4.4.15 Membro 3 da superfamília 4 de proteínas transmembrana (TM4SF3)	67
I.4.4.16 Supressor de metástase de melanoma maligno (KiSS-1).....	68
I.4.5 SNPs em genes de <i>Schistosoma mansoni</i> candidatos a vacina	69
I.4.6 Falsos SNPs.....	72
I.4.7 Implementação do <i>SNaPshot</i>	75

I.4.7.1 PCR <i>multiplex</i>	82
Conclusões	85
Perspectivas	87
Artigos relacionados a tese.....	90
PARTE II: Análise de novos SNPs não sinônimos utilizando ORESTES.....	98
Resumo.....	99
<i>Summary</i>	102
Introdução	105
II.1.1 Importância dos SNPs em questões médicas relevantes.....	106
II.1.1.1 Câncer como uma doença genética: O impacto do Projeto Genoma Humano	107
II.1.1.1.1 Câncer de Cabeça e Pescoço no mundo.....	107
II.1.1.1.2 Gerando ORESTES de amostras de cabeça e pescoço	108
II.1.1.1.3 SNPs em Tumores de Cabeça e Pescoço	109
II.1.1.2 Distúrbios psiquiátricos como doenças genéticas	111
II.1.1.3 SNPs em genes de <i>Schistosoma mansoni</i> candidatos à vacina	113
Objetivos	114
II.2.1 Objetivo geral	115
II.2.2 Objetivos específicos	115
Material e Métodos	116
II.3.1 Escolha dos SNPs a serem avaliados.....	117
II.3.2 Amostras humanas	117
II.3.2.1 Amostras de pacientes com Câncer de Cabeça e Pescoço	117
II.3.2.2 Amostras de pacientes com a Doença de Alzheimer	118
II.3.3 Amostras de <i>Schistosoma mansoni</i>	118
II.3.4 Predição de alteração da estrutura proteica pelo SNP.....	119
II.3.5 Análises estatísticas	120
Resultados e Discussão	121
II.4.1 Avaliação dos SNPs em pacientes com Câncer de Cabeça e Pescoço.....	122
II.4.2 SNPs em pacientes com a Doença de Alzheimer	124

II.4.3 Outras patologias onde os SNPs validados podem estar associados	126
II.4.4 SNPs em genes de <i>Schistosoma mansoni</i> candidatos a vacina	127
II.4.5 Predição de estrutura das proteínas onde foram validados SNPs.....	128
Conclusões	133
Perspectivas	135
Artigos relacionados a tese	138
Referências bibliográficas	144
Endereços eletrônicos	152
Anexos	
Colaborações	
<i>Curriculum vitae</i> (resumo)	

PARTE I

Identificação de novos SNPs não sinônimos

utilizando ORESTES

Resumo

SNPs ou polimorfismos de nucleotídeo único são as variações mais comuns do genoma humano. Na sua maioria estas variações são inertes, mas em alguns casos podem levar a uma série de alterações moleculares que, em última instância, podem modificar a atividade biológica da proteína codificada.

Aparentemente, a maior parte do genoma humano é não codificante e a pressão seletiva sobre estas áreas deve ser menor, de maneira que alterações nestes segmentos, em geral, são menos danosas ao organismo. Sendo assim, a maioria dos polimorfismos causais, associados a patologias, está presente dentro das regiões codificadoras dos genes ou bem próximas a estas.

Como a maioria das ESTs disponíveis no GenBank é derivada das extremidades dos genes humanos, a técnica ORESTES, desenvolvida por nosso grupo e usada como a base da geração de seqüências do Projeto Genoma do Câncer Humano FAPESP/Ludwig (HCGP) e do Projeto Genoma de *Schistosoma mansoni* FAPESP/MCT (PGSM), tem uma vantagem para o encontro de SNPs relevantes, uma vez que a maioria das seqüências produzidas são derivadas da região central dos genes.

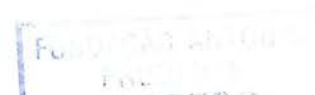
Neste projeto, nos propomos a identificar novos SNPs sugeridos a partir dos dados gerados pelo uso da técnica ORESTES, com foco principal em SNPs não sinônimos, buscando eventuais associações com questões de relevância médica.

Para o encontro de SNPs humanos, colaboramos com o grupo do Dr. Marco Antonio Zago (Fundação Hemocentro de Ribeirão Preto) na criação de uma rotina de bioinformática especificamente trabalhada para a busca de SNPs a partir das

ORESTES derivadas do HCGP. A partir de 167.194 ORESTES similares a genes humanos conhecidos, foram identificados 2.845 SNPs dos quais 275 passaram por nossos filtros. A partir deste ultimo conjunto, 22 polimorfismos foram selecionados para os experimentos de validação em amostras populacionais sendo que 13 deles (60%) foram confirmados.

Já em *S. mansoni* contamos com o auxílio das equipes de bioinformática do Projeto Genoma de *Schistosoma mansoni* na busca de polimorfismos em genes específicos, candidatos ao desenvolvimento de vacinas contra o parasita, tendo sido possível identificar a existência de SNPs não sinônimos em todos os genes candidatos à vacina pela OMS. Para os genes GST28, 14-3-3 e TPI, foram realizados experimentos de validação e dois polimorfismos foram confirmados (em GST28 e em TPI) sendo estes os primeiros SNPs identificados para este organismo.

A partir dos dados gerados podemos concluir que ORESTES é uma valiosa fonte para o encontro de SNPs não sinônimos.



Summary

Fundação Antonio Prudente

Maria Rodrigues Alves Kuninari
Pós-Graduação

Single nucleotide polymorphisms (SNPs) are the most frequent variation found in the human genome. While the majority of the SNPs seem to be inert, a few can promote a series of molecular alterations that, may culminate in alterations of the biological activity of the corresponding protein.

Most of the human genome is non coding and the selective pressure over these regions is lower as, in general, alterations in these segments are less prone to be dangerous to the organism. In this way, most of the causal polymorphisms(those directly associated with pathologies), are located near or inside of coding regions of the genome.

As the largest portion of the ESTs set available in GenBank is derived from the extremities of human genes, the ORESTES technique, developed by our group and used as the basis of the HCGP (Projeto Genoma do Câncer Humano FAPESP/Ludwig) and the PGSM (Projeto Genoma de *Schistosoma mansoni* FAPESP/MCT) have significantly contributed to the large scale sequencing of the human transcriptome. In this project we aimed to evaluate the utility of ORESTES to unravel new SNPs in the human genome, with a main focus directed to new non-synonymous SNPs. After the confirmation of new SNPs, we decided to look for putative associations of these with relevant medical questions.

In order to identify new human SNPs, we began a project with the collaboration of Dr. Marco Antonio Zago's group (Fundação Hemocentro de Ribeirão Preto) to elaborate a bioinformatic pipeline, specially designed to find SNPs using HCGP derived ORESTES. Using an initial set of 167,194 ORESTES similar to know human genes, 2,845 putative SNPs were identified. A subset of 275 was still

considered after the adoption of some filters. About ten percent of these polymorphisms were selected for validation experiments and 60% could be confirmed.

In order to expand our SNP analysis to a different organism, *S. mansoni*, we got the support of the bioinformatic groups of SMGP that enable SNP minning in specific genes, the WHO candidates to vaccine development against this parasite. Our data suggests that all WHO candidate genes are polymorphic. To GST28, 14-3-3 and TPI genes, validation experiments were performed and two variations were confirmed (one to GST28 and one to TPI), these are the first SNPs identified for this organism.

Using the generated data we can conclude that ORESTES is a valuable source of non synonymous SNPs.

Introdução

I.1.1 Polimorfismos de nucleotídeo único

No lançamento do Projeto Genoma Humano, há mais de uma década, houve discussões em torno de qual amostra de DNA deveria ser seqüenciada. É possível que a importância desta discussão não tenha sido percebida naquela época, mas amostras de diversos indivíduos acabaram sendo usadas no seqüenciamento, o que contribuiu decisivamente para a relevante geração de dados referentes à extensão da diversidade genética humana, tendo com base a seqüência do DNA. Sabemos hoje que existe uma variação de cerca de 0,16% entre a seqüência de DNA de quaisquer dois indivíduos (SACHIDANANDAM et al. 2001) considerando a combinação linear dos quatro nucleotídeos (simbolicamente A, C, G e T). Quando duas seqüências de DNA da mesma região são comparadas e dois nucleotídeos diferentes são encontrados em uma mesma posição, dizemos que existe um Polimorfismo de Nucleotídeo Único, ou SNP (Figura 01).

Um fator que reveste os SNPs de grande importância é a sua abundância. Estima-se que exista, no genoma humano, um SNP a cada 600-2000 nucleotídeos (KRUGLYAK e NICKERSON 2001, SACHIDANANDAM et al. 2001, Figura 02); isto significa que no genoma humano, temos algo entre 1,5 a 5 milhões destas variações.

Em abril de 1999, foi criado um consórcio, denominado *The SNP Consortium* (TSC - <http://snp.cshl.org>), envolvendo onze gigantes da indústria farmacêutica e de biotecnologia. O projeto se estendeu até final de 2001 e, utilizando um orçamento



```

GAAAAGTCACTGATGATCATGGGCACGT
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;
gaaagt cact gat gat cat gggcacgt;
gaaagt cact gat aat cat gggcacgt;

```

Figura 1 - Identificação de um SNP. No alinhamento de diversas seqüências correspondentes a um mesmo segmento genômico (ou cDNA) podemos encontrar variações de ponto, estas variações de seqüência são denominados SNPs ou Polimorfismos de Nucleotídeo Único.

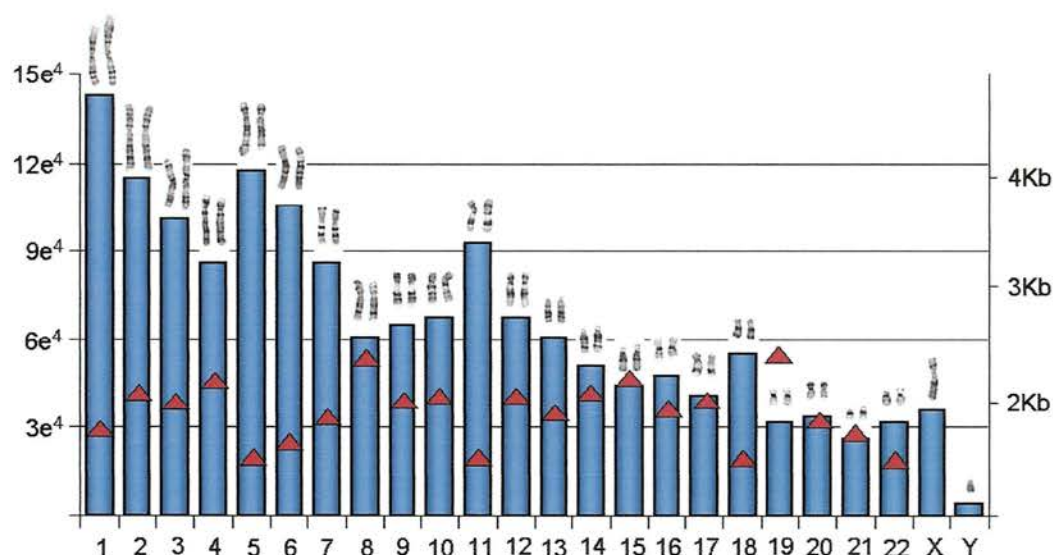


Figura 2 - Distribuição dos SNPs humanos já identificados por cromossomo. O número de polimorfismos (ne^4) por cromossomo é proporcional ao tamanho do mesmo, quanto maior o cromossomo mais variações são identificadas. Os triângulos vermelhos representam a densidade média de SNPs (número de SNPs/Kb) em cada um dos cromossomos. Utilizando somente os dados depositados até setembro de 2001 temos a média da distribuição variando de 1 SNP a cada 0,8Kb (crom22) até 1 SNP a cada 2,3Kb (crom19), estes valores tendem a diminuir à medida em que novos polimorfismos são identificados.

estimado em 50 milhões de dólares, foi capaz de identificar, juntamente com o Projeto Genoma Humano, 1,42 milhão de SNPs (SACHIDANANDAM et al. 2001).

Os SNPs representam uma das mais poderosas ferramentas para a identificação de haplótipos, que são combinações específicas de alelos (GABRIEL et al. 2002). Estes haplótipos são muito úteis na determinação da história evolutiva humana, já que a taxa de alterações recorrentes e remissivas é muito baixa (BROOKES 1999), e também em estudos de associação com patologias (PATIL et al. 2001), pois vários SNPs situados dentro de um mesmo segmento genômico podem estar em desequilíbrio de ligação e, com isso, um pequeno número de polimorfismos pode ser utilizado para analisar grandes regiões genômicas. Atualmente, encontra-se em andamento um projeto, *HapMap*, financiado pelo Instituto Nacional de Saúde dos Estados Unidos (NIH) que tem por objetivo gerar um mapa dos haplótipos humanos, para tanto, este projeto conta com 100 milhões de dólares para três anos de pesquisa (COUZIN 2002).

Cerca de 97% do DNA humano não contém genes e trabalhos recentes indicam que a diversidade nucleotídica (a quantidade de variações dividida pelo tamanho da sequência) é bastante similar entre regiões ricas ou pobres em genes. A diversidade nucleotídica em regiões ricas em genes foi estimada como sendo de 8 diferenças a cada 10 mil nucleotídeos. Níveis similares, 7,51 diferenças a cada 10 mil nucleotídeos, foram encontrados em outras regiões genômicas (SACHIDANANDAM et al. 2001). Estas informações nos permitem estimar a existência de aproximadamente 42 mil SNPs em genes humanos, concordante com dados recentes do TSC que indicam cerca de 60 mil SNPs dentro de genes. Já foram

encontrados SNPs em 93% dos genes humanos conhecidos até agora (CHAKRAVARTI 2001).

A maior parte dos SNPs identificados em genes humanos não resulta em alterações na sequência protéica, graças à sua localização em regiões não traduzidas da molécula de mRNA, ou ainda devido à sua posição dentro dos *codons* (SNPs sinônimos). Assim, apenas uma pequena parcela dos SNPs tem a capacidade de alterar a composição das proteínas, o que por sua vez poderia levar à modificação da sua atividade biológica. São esses polimorfismos, denominados “SNPs não sinônimos”, que têm maior potencial de reflexo fisiológico (Figura 03).

I.1.2 Bancos de dados de SNPs

A fim de organizar as informações acerca dos polimorfismos humanos, vários bancos de dados, públicos e privados, foram criados. O mais popular e completo deles é o dbSNP (SHERRY et al. 1999), que se encontra sediado no Centro Nacional para Informação em Biotecnologia (NCBI) e cujas informações estão disponibilizadas no endereço eletrônico: <http://www.ncbi.nlm.nih.gov/SNP>. Atualmente, já se encontram depositados no dbSNP cerca de 3 milhões de SNPs humanos (4,9 milhões de submissões, montagem 110 de 02/01/2003), além de entradas para dados de outras espécies, incluindo *Gallus*, *Mus*, *Plasmodium*, *Aedes*, etc (Tabela 1). Entretanto, poucos destes polimorfismos apresentam informações referentes à sua frequência ou mesmo à sua validação em amostras individuais (Tabela 1).

Existem também outros bancos de SNPs, tais como o banco público HGBase que contém 62907 entradas (<http://www.hgbase.interactiva.de>), o banco privado da

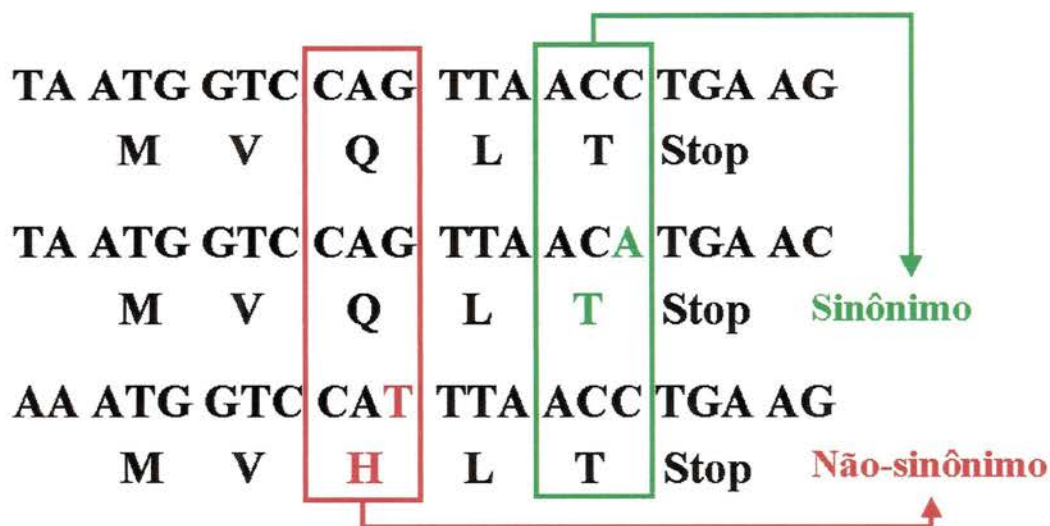


Figura 3 - Polimorfismos presentes dentro da região codificadora dos genes podem alterar ou não a sequência proteica final. No caso deste gene hipotético temos o alinhamento de três diferentes sequências, sendo a primeira a sequência de referência. Na segunda sequência existe uma alteração de C>A a qual não altera o aminoácido a codificado, estes polimorfismos são denominados SNP sinônimos. Entretanto temos uma variação de G>T na terceira sequência que provoca uma mudança de glutamina para histidina na proteína, estes polimorfismos que promovem alterações nas sequências proteicas correspondentes são denominados SNPs não sinônimos.

Tabela 1 - Sumário dos polimorfismos depositados no dbSNP

Organismo	Número de submissões - ss	Número de clusters RefSNP - rs (validados)	Número de ss com frequência	Número de ss com genótipo
<i>Homo sapiens</i>	4894587	3049569 (522072)	199849	35785
<i>Anopheles gambiae</i>	430291	429587 (0)	-	-
<i>Mus musculus</i>	115466	111235 (79747)	-	79752
<i>Danio rerio</i>	2031	2025 (2025)	-	2031
<i>Gallus gallus</i>	1210	1204 (0)	-	-
<i>Caenorhabditis elegans</i>	1065	1065 (0)	-	-
<i>Glycine max</i> (Soja)	281	278 (278)	-	281
<i>Plasmodium falciparum</i>	203	203 (199)	-	199
<i>Arabidopsis thaliana</i>	184	184 (184)	-	184
<i>Ficedula albicollis</i> (Ave)	37	37 (37)	37	-
<i>Ficedula hypoleuca</i> (Ave)	28	20 (20)	28	-
<i>Sus scrofa</i>	4	4 (0)	-	-
<i>Gorilla gorilla</i>	4	4 (0)	-	-
<i>Pan paniscus</i>	2	2 (2)	2	-
<i>Rattus norvegicus</i>	1	1 (0)	-	-

Celera que possui cerca de 200 mil polimorfismos validados em amostras populacionais (<http://www.celeradiscoverysystem.com>), o PicSNP (CHANG e FUJITA 2001) que é um banco de polimorfismos não sinônimos associado ao dbSNP, atualmente composto por 3793 SNPs, e o Go!Poly (ZHANG et al. 2001) que é um banco de dados associado ao dbSNP e ao HGBase mais direcionado a genes específicos.

Estudos isolados envolvendo a identificação de SNPs relevantes para a agropecuária e agricultura estão sendo feitos, dentro destes podemos citar estudos envolvendo milho (CHING et al. 2002), arroz (NASU et al. 2002), ovelha (MCRAE et al. 2002) e cavalo (SHUBITOWSKI et al. 2001).

Além do mais, vários bancos privados de SNPs focados nas mais diversas áreas estão sendo montados, mas as informações pertinentes a estes bancos são restritas.

I.1.3 Métodos de genotipagem

Inúmeros são os métodos hoje disponíveis para a genotipagem de SNPs, não mais havendo restrições ao número de amostras nem mesmo ao número de polimorfismos a serem avaliados. O método mais simples, o RFLP (*Restriction Fragment Length Polymorphism*), pode ser utilizado quando o polimorfismo, cria ou abole um determinado sítio de restrição. A principal limitação do RFLP é a sua dependência da existência de um sítio de restrição que seja alterado pelo polimorfismo em estudo.

O seqüenciamento de DNA e suas variantes também podem ser citados como métodos mais antigos mas um pouco mais sofisticados. Entretanto, dispomos hoje de

uma grande variedade de protocolos, *kits*, aparelhos e sistemas de detecção, envolvendo ensaios bioquímicos e imunológicos. A aplicação de cada procedimento depende do número de amostras a serem trabalhadas, número de SNPs a serem genotipados, estrutura do laboratório, necessidade de automatização, custos e tempo a ser gasto. De qualquer forma, é seguro afirmar que para cada combinação das variáveis acima citadas existe um método adequado.

I.1.4 Aplicações

Os polimorfismos do genoma humano constituem um ponto chave na definição da base etiológica das diferenças fenotípicas entre indivíduos, diferentes níveis de resposta a medicamentos, predisposição a doenças e outros (BROOKES et al. 2000). Alterações funcionais relevantes no DNA são provavelmente localizadas dentro ou próximas de seqüências transcritas, incluindo uma série extensa de SNPs que podem ter como consequência: *splicing* alternativo do mRNA, alteração na seqüência de aminoácidos da proteína, geração ou supressão de códons de terminação de tradução, adenilação alternativa, alteração na estrutura e estabilidade do RNA. Cada uma destas pode levar a alterações do fenótipo normal. Além disto, estas variações podem ser úteis como marcadoras de grandes blocos genômicos que são herdados em conjunto, definindo os haplótipos.

A capacidade de identificar polimorfismos em genes humanos, e associá-los a diferentes graus de susceptibilidade ou evolução das principais doenças humanas, pode representar um dos maiores progressos da pesquisa biológica recente. Trabalhos desta natureza, feitos com amostras de DNA associadas a bancos de dados complexos e completos, analisados com abordagem estatística adequada, podem

revelar marcadores moleculares relevantes em relação à evolução clínica de várias doenças, assim como a predição de resposta a diferentes tratamentos, ou seja, um verdadeiro diagnóstico e tratamento baseado na nossa individualidade genética.

Outra frente na qual os SNPs são de grande importância são os estudos de evolução molecular, divergência nucleotídica e variabilidade protéica. Tais aspectos são de grande importância para estudo de diversas formas de vida, e podem ter diversas aplicações práticas. Uma destas aplicações é o estudo de variação gênica em candidatos vacinais. A fim de se gerar uma resposta imunológica a uma determinada proteína de um parasita, por exemplo, é desejável que estes sejam pouco polimórficos dentro das populações de parasitas, e como sugerido para *Plasmodium falciparum*, um dos prováveis mecanismos de escape do sistema imune do hospedeiro é a seleção positiva favorecendo a variabilidade de proteínas de membrana (MU et al. 2002; VOLKMAN et al. 2002). Esta variabilidade poderia aniquilar a resposta imune para certos indivíduos da população, e de acordo com o caso em questão, a vacina perderia sua efetividade. Assim, para os organismos onde informações genômicas são disponíveis, é possível realizar uma avaliação de potenciais candidatos vacinais, averiguando a presença ou não de polimorfismos nestes genes, o que eventualmente levaria à sua exclusão ou inclusão como reais candidatos vacinais.

I.1.5 ORESTES: informações derivadas de regiões codificadoras de genes

O levantamento dos genes ativos em uma célula ou tecido é fundamental para a compreensão de seu fenótipo. Assim, conhecendo este perfil de RNAs expressos,

podemos tentar entender os mecanismos que disparam uma condição patológica em um tecido normal ou de que forma um tecido ou um organismo reage às alterações que ocorrem no ambiente em que este se encontra (RIGGINS e STRAUSBERG 2001).

Uma das abordagens utilizadas para realizar um levantamento dos genes expressos em determinados tecidos é a geração de ESTs (*Expressed Sequence Tags*) ou etiquetas de seqüências expressas. Para isto, é feita uma extração do RNA total do tecido em questão. A partir destes, os RNAs mensageiros (mRNAs) são purificados e, através de uma reação de transcrição reversa, são geradas moléculas de cDNA. Os cDNAs são clonados e suas extremidades seqüenciadas, gerando as etiquetas de seqüências expressas. Uma desvantagem desta técnica é que a maior parte das seqüências geradas correspondem a regiões não traduzidas das moléculas de mRNA, o que impede o conhecimento total da mensagem, e dificulta de sobremaneira as análises de similaridade dos genes.

Alternativamente, foi desenvolvida a técnica ORESTES, ou *Open Reading frame ESTs* (DIAS-NETO et al. 2000), que permite gerar ESTs derivadas da região central dos genes que estão sendo expressos no modelo biológico estudado (Figura 04). Na geração convencional de ESTs, é utilizado um iniciador composto somente por timinas para a síntese das moléculas de cDNA, já que os mRNAs são poliadelinados os iniciadores irão se ligar neste segmento. Na técnica ORESTES, são utilizados iniciadores arbitrários para a síntese da molécula de cDNA. Já que a transcrição reversa é feita a 37°C, os iniciadores usados irão se ligar de maneira parcialmente complementar e promover a síntese de parte da molécula de mRNA. Após a transcrição reversa, segue-se uma PCR de baixa estringência de anelamento,

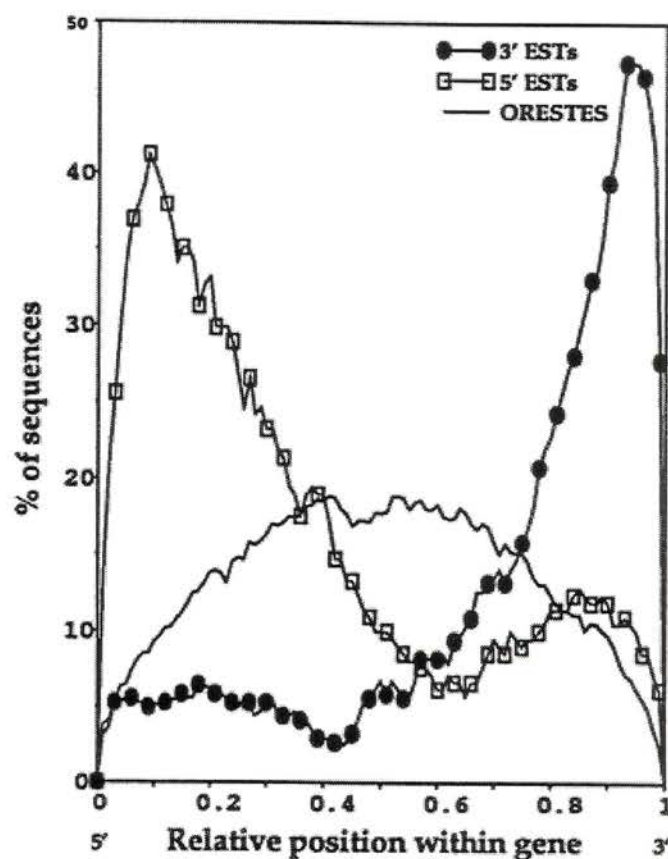


Figura 4 - Comparação da distribuição posicional das seqüências ORESTES e ESTs 5'e 3' em relação à sua posição relativa em seqüências completas de cDNA. Pode-se ver claramente que a maioria das ESTs tradicionais está localizada nas extremidades não transcritas dos genes enquanto ORESTES tem sua distribuição preferencial na região codificadora dos mesmos (figura extraída de Dias-Neto et al. 2000).

utilizando o mesmo iniciador da síntese de cDNA. Esta combinação de fatores, iniciadores arbitrários e condições de baixa estringência, favorece a amplificação da região central das moléculas de mRNA que é a porção codificante (Figura 04), e normaliza parcialmente a população de genes amostrados, permitindo assim que sejam obtidas seqüências de genes raros (detalhes em DIAS-NETO et al. 2000). Após a construção de diversas bibliotecas ORESTES, é de se esperar a obtenção de fragmentos gênicos complementares, que se sobrepõem parcialmente e permitem a montar rascunhos de moléculas de mRNA originais. Nas regiões de sobreposição, é possível identificar variações pontuais nas seqüências alinhadas. Uma fração destas variações é meramente artefactual, normalmente derivada de erros das polimerases, ou erros de sequenciamento. No entanto, de posse de alguns critérios, podemos identificar com segurança alguns polimorfismos reais, que posteriormente podem ser validados em uma maior série de amostras.

1.1.5.1 O uso de ORESTES para investigar polimorfismos genéticos

Apenas em regiões codificadoras de genes é possível identificar SNPs não sinônimos. Desta maneira, pela utilização da técnica de ORESTES, temos uma posição privilegiada para a obtenção de novos polimorfismos desta categoria. Graças à utilização da técnica em projetos de sequenciamento em larga escala, em duas diferentes espécies (humanos e *Schistosoma mansoni*), podemos investigar os resultados sugestivos de SNPs de uma maneira mais ampla. Se por um lado a qualidade dos dados favorece o achado de SNPs não sinônimos, a quantidade dos mesmos nos permite uma cobertura significativa dos transcriptomas estudados.

I.1.5.2 A Geração de ORESTES para a análise do transcriptoma de tumores humanos

Em 1997, foi lançado nos Estados Unidos um projeto denominado CGAP (*Cancer Genome Anatomy Project*), que pretendia identificar e catalogar o padrão de expressão gênica de tecidos normais, pré-cancerosos e cancerosos, dados estes que poderiam levar a um aumento do conhecimento do câncer, com o conseqüente aperfeiçoamento na detecção das lesões, no diagnóstico e tratamento dos pacientes (STRAUSBERG et al. 1997). No contexto do CGAP, foram geradas bibliotecas de cDNAs de diferentes tecidos e condições histológicas. Os insertos clonados nestas bibliotecas tiveram as suas extremidades seqüenciadas gerando as ESTs. O CGAP produziu cerca de 1,2 milhão de seqüências tornando-se um dos maiores projetos mundiais de geração de ESTs humanas (RIGGINS e STRAUSBERG, 2001; <http://cgap.nci.nih.gov/>).

No Brasil foi lançado em 1999 um projeto complementar ao CGAP, denominado Projeto Genoma do Câncer Humano (HCGP – *Human Genome Cancer Project*). A base da geração de seqüências do HCGP, financiado pela FAPESP em conjunto com o Instituto Ludwig de Pesquisas sobre o Câncer, foi a técnica ORESTES (DIAS-NETO et al. 2000). Concluído em 2001, o HCGP produziu cerca de 1,1 milhão de seqüências. Como a maioria das ESTs disponíveis no GenBank é derivada das extremidades dos genes humanos e a técnica ORESTES gera preferencialmente seqüências da região central dos genes, temos hoje um cenário favorável para a identificação de polimorfismos humanos utilizando estes dois grupos de ESTs. O alinhamento destas seqüências permite a obtenção de consenso gênicos e também permite preencher os segmentos incompletos entre as ESTs

(CAMARGO et al. 2001), gerando assim dados que contribuem para a identificação de seqüências completas de cDNA (Figura 04).

I.1.5.3 A Geração de ORESTES para a análise do transcriptoma de *Schistosoma mansoni*

O *Schistosoma mansoni* é um parasita responsável pela esquistossomose, doença que ocorre em regiões tropicais e sub-tropicais e afeta atualmente cerca de 200 milhões de pessoas, sendo uma das principais causas de morbidade humana (CHITSULO et al. 2000; BERGQUIST et al. 2002; ENGELS et al. 2002).

Encontra-se em fase de finalização no Estado de São Paulo, o Projeto Genoma de *Schistosoma mansoni* (PGSM, <http://bioinfo.iq.usp.br/schisto>), financiado pela FAPESP e pelo Ministério de Ciência e Tecnologia, que teve como objetivo gerar 120 mil ESTs usando bibliotecas de cDNA normalizadas do tipo ORESTES. O PGSM tinha como objetivo a geração de um painel dos genes expressos nos diversos estágios de vida do parasita a fim de se entender melhor os processos biológicos envolvidos com o desenvolvimento e a interação deste parasita com seus hospedeiros, entre eles o homem.

Até o momento da redação desta dissertação já haviam sido produzidas pelo PGSM 179.155 seqüências (03/02/2003) de seis estágios de vida do parasita (adultos, ovos, miracídeos, *germ balls*, cercárias e esquistossômulos de sete dias), ultrapassando em 50% o objetivo inicialmente proposto. Dos 163 genes completos já descritos em *S. mansoni*, 92% já estão representados por nossas ESTs, o que permite inferir que um bom percentual dos transcritos deste parasita esteja representado em nosso banco de dados.

I.1.6 SNPs em genes de *Schistosoma mansoni* candidatos a vacina

A esquistossomose é um dos maiores problemas de saúde humana em regiões tropicais e sub-tropicais, sendo que cerca de 600 milhões de pessoas vivem em regiões endêmicas (CHITSULO et al. 2000; BERGQUIST et al. 2002; ENGELS et al. 2002). Um dos enfoques dados na tentativa de reverter este quadro é a produção de vacinas contra o parasita. Várias são as proteínas candidatas à vacina e entre elas podemos destacar: 14-3-3, glutathione-S-transferase (GST), paramiosina, triose fosfatase isomerase (TPI) e o antígeno Sm14. Com exceção de 14-3-3, todos os genes anteriormente citados são preconizados pela Organização Mundial de Saúde (OMS) como candidatos à vacina (YANG et al. 2000; SCHECHTMAN et al. 2001). Embora alguns testes de campo já estejam em andamento, os resultados são pouco promissores, sendo que baixos níveis de proteção têm sido observados em testes realizados pelos laboratórios participantes.

Possuindo em mãos as seqüências geradas pelo PGSM, foi possível vislumbrar a possibilidade de avaliar possíveis polimorfismos nos genes candidatos a vacinas, sugeridos pelas ORESTES geradas em nosso projeto. Este foi então o primeiro estudo de SNPs já realizado para este importante parasita humano. Em *S. mansoni*, avaliamos SNPs que alterassem a seqüência protéica dos genes candidatos a vacinas, resultados estes que podem ajudar a explicar o insucesso dos ensaios imunológicos realizados até o momento.

Objetivos

I.2.1 Objetivo geral

Avaliar as seqüências ORESTES quanto ao seu potencial de identificação de novos SNPs em regiões codificadoras de genes de diferentes organismos, utilizando para isso os dados gerados pelos projetos HCGP e PGSM.

I.2.2 Objetivos específicos

- Validar SNPs não sinônimos sugeridos no Projeto Genoma do Câncer Humano;
- Avaliar a frequência dos SNPs humanos em amostras populacionais;
- Validar SNPs não sinônimos sugeridos no Projeto Genoma de *Schistosoma mansoni* localizados em genes candidatos a vacina;
- Implementar novos protocolos de genotipagem de SNPs.

Material e Métodos

I.3.1 Seleção inicial dos SNPs humanos

Anteriormente a este projeto já estava em andamento uma iniciativa de identificação e validação de SNPs utilizando como base de dados as seqüências ORESTES do Projeto Genoma do Câncer Humano. Esta iniciativa, coordenada e executada pelo grupo do Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto, já havia desenvolvido, em colaboração com nosso grupo, uma rotina computacional para identificação e análise de SNPs (Figura 05), que consistia resumidamente nos passos descritos a seguir.

Os parâmetros dos diversos programas utilizados (BLAST, *PolyBayes*, *RepeatMasker*, etc) seguiram, quando não descritos, os especificados como padrão.

I.3.1.1 Estruturação dos dados a serem analisados

Esta etapa constitui na preparação dos dados, tanto das ORESTES como do conjunto de seqüências a serem usadas como referência no alinhamento (âncoras). Os eletroferogramas das ORESTES foram submetidos a um *base calling* utilizando o programa *Phred* (EWING e GREEN 1998), sendo obtidos como resultados arquivos *.phd*, que continham, além da seqüência correspondente, informações sobre a qualidade de cada base. Para a obtenção do conjunto de seqüências âncora, foi feita uma busca no GenBank por genes humanos que continham a região codificadora (CDS) completa. A estas seqüências, foi atribuído um valor de qualidade alto e homogêneo, uma vez que estão disponíveis apenas em formato texto, não sendo possível submetê-las ao *Phred*, que requer eletroferogramas como arquivos de

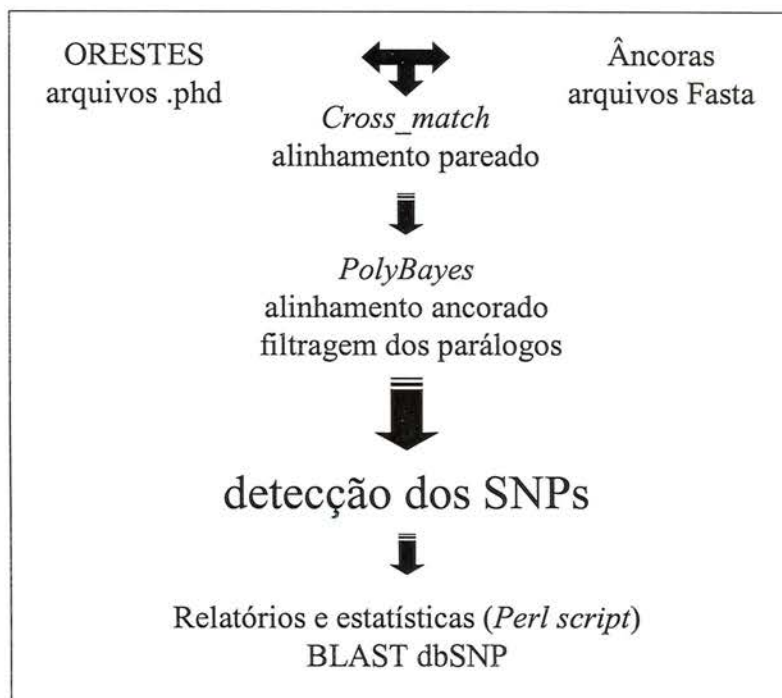


Figura 5 - Esquema da rotina de bioinformática, desenvolvida pelo Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto, para a identificação dos SNPs nas seqüências ORESTES do Projeto Genoma do Câncer Humano.

entrada. Assim, foram selecionadas como seqüências âncora 7.950 CDS completas que foram identificadas no GenBank pela equipe de bioinformática. Atualmente estamos preparando uma nova rotina de bioinformática para que as análises sejam refeitas utilizando como âncoras o *RefSeq* (MAGLOTT et al. 2000).

I.3.1.2 Alinhamento das ORESTES e das seqüências âncora

Os dois conjuntos de seqüências, ORESTES e seqüências âncora, foram submetidos a um alinhamento pareado através do programa *Cross_match* (<http://genome.uc.edu/genome/HelpPages>). Nesta etapa, foram identificadas e retiradas seqüências de vetor e iniciadores presentes nas ORESTES. O resultado desse programa funcionou como o arquivo de entrada para o próximo passo.

I.3.1.3 Identificação dos SNPs humanos

Nesta etapa, ocorreu a identificação dos SNPs utilizando o programa *PolyBayes* (MARTH et al. 1999), por meio dos seguintes passos: alinhamento múltiplo ancorado a uma seqüência referência, detecção e exclusão de parálogos (seqüências muito similares, mas contendo discrepâncias de alta qualidade ao longo de todo o alinhamento) e identificação dos sítios polimórficos. A identificação de um sítio variante e a designação do mesmo como um possível SNP levou em consideração o número de seqüências alinhadas na referida posição e a qualidade das bases (valor de *Phred* de 15).

I.3.1.4 Resultados e estatísticas

Como resultado da análise anterior, o *PolyBayes* gerou um relatório extenso e detalhado, contendo informações sobre todos os *clusters* formados, os SNPs identificados e vários dados probabilísticos usados na análise. Foi desenvolvido um *script* para retirar as informações desse relatório e gerar um arquivo conciso, contendo os dados mais relevantes sobre as variações identificadas. Foram gerados dois arquivos distintos, um para os SNPs sinônimos e outro para os não sinônimos. Nesse relatório, além da alteração de nucleotídeo, era apresentado o efeito da substituição na sequência protéica, e a sequência que flanqueia a variação. Nesse passo, foram também determinadas as estatísticas obtidas nas análises.

I.3.1.5 Busca em bancos de dados de SNPs

Após a identificação dos possíveis SNPs, foi feita uma busca automática no banco de dados dbSNP (<http://www.ncbi.nlm.nih.gov/SNP>). Dessa forma, foi possível verificar quais SNPs identificados já haviam sido descritos na literatura.

I.3.1.6 Montagem do banco de dados

Terminado o processo de análise, os dados obtidos foram armazenados em um banco de dados relacional o HCGP – SNP (SNPs do Projeto Genoma do Câncer Humano) que pode ser acessado via *web page* (<http://bit.fmrp.usp.br/bit/projects.php?ptitle=Projects>).

I.3.2 Seleção final dos SNPs

O banco de dados acima descrito fornece informações a respeito de todos os SNPs encontrados pela rotina de bioinformática. Entretanto, nosso interesse principal está nos SNPs não sinônimos e, por enquanto, os SNPs identificados que não levam a alteração de aminoácido não foram submetidos a validações em bancada.

Como os dados gerados no HCGP são derivados de PCR, falsos SNPs podem surgir por alguns motivos básicos: erro da transcriptase reversa (que é uma enzima de origem viral que promove a variabilidade do vírus através de mutações espontâneas; BOYER et al. 1992; MENENDEZ-ARIAS 2002), erro da *Taq* DNA polimerase no momento da geração dos perfis de ORESTES ou no momento da geração dos moldes de seqüenciamento (SANSON et al. 2002), ou ainda por erros durante a reação de seqüenciamento (Figura 06) e baixa qualidade de algumas seqüências. Diante destas possibilidades, antes de iniciar a confirmação dos SNPs em bancada, optamos por adotar dois critérios de seleção, baseados em uma confirmação prévia, *in silico*, dos possíveis SNPs. Em primeira instância, foram excluídos todos os SNPs que não fossem confirmados por pelo menos duas seqüências (para evitar erros da reação de seqüenciamento) e também todos os polimorfismos que não fossem confirmados por pelo menos duas bibliotecas ORESTES distintas (para evitar erros como o de incorporação pela *Taq* DNA polimerase ou pela transcriptase reversa).

I.3.3 Mapeamento genômico e desenho dos iniciadores

Para a validação experimental dos SNPs que passaram pelo critério de duas bibliotecas, como mencionado anteriormente, o primeiro passo foi determinar a localização da região gênica que continha a variação no genoma humano. Para isto,

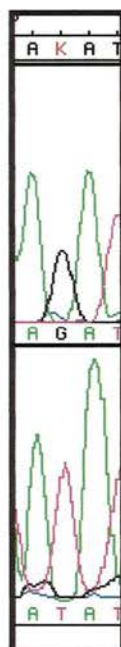


Figura 6 - Falso SNP gerado por erro de seqüenciamento. O mesmo produto de PCR foi seqüenciado duas vezes e na mesma posição pode ser observado uma variação de base G>T que não existe. Em 48 amostras, este tipo de evento pôde ser verificado três vezes.

em cada uma dos SNPs candidatos à validação, submetemos a respectiva sequência de mRNA ao programa BLAST utilizando a versão mais atual do Genoma Humano disponível no momento da análise (<http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html>). Após uma inspeção visual dos resultados obtidos, a sequência genômica, onde o provável SNP estava presente, foi isolada e submetida ao programa *RepeatMasker* (<http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker>), que mascara regiões repetitivas dentro do genoma, assim como regiões de baixa complexidade, retrovírus, minissatélites, Alus, SINEs, LINEs, entre outras. O arquivo proveniente da análise deste programa foi editado e submetido a programas para desenhos de iniciadores. Foi utilizado o programa *Primer3* (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi), que gera uma lista de pares de iniciadores que podem ser utilizados para a amplificação do fragmento de interesse. Os iniciadores, sempre que possível, foram desenhados a uma distância mínima de 50pb e máxima de 120pb do SNP, o que facilitou as reações de PCR e sua confirmação por seqüenciamento. Preferencialmente, um dos iniciadores foi desenhado dentro de um intron a fim de diminuir a possibilidade de amplificar pseudogenes e parálogos. Os iniciadores foram desenhados de forma que a temperatura ótima de amplificação dos fragmentos fosse sempre próxima a 60°C.

I.3.4 Padronização das amplificações

Embora a temperatura ótima predita para amplificação dos fragmentos fosse de 60°C, na prática ela teve que ser otimizada para cada par de novos iniciadores com os quais trabalhamos. Na primeira bateria de amplificações foram utilizadas temperaturas de anelamento que variaram de 53°C a 66°C (Figura 07). A

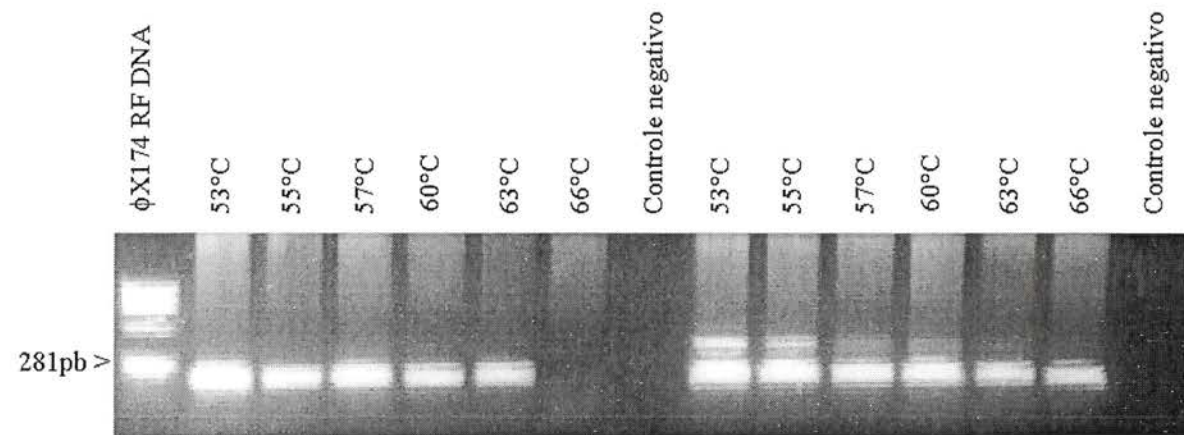


Figura 7 - Reações de padronização da PCR. Podemos ver os resultados das padronizações de dois SNPs. Para o primeiro SNP não há amplificação do fragmento a temperatura de 66°C, para o segundo SNP o aumento de temperatura aumenta a estringência da reação.

temperatura na qual foi obtido um melhor rendimento de amplificação e uma maior especificidade, foi escolhida para os experimentos de bancada.

Após avaliar diferentes protocolos de amplificação, optamos por um programa de *two step* PCR que não possui a etapa de extensão. Já que o tamanho do fragmento amplificado não passou de 200pb para a grande maioria dos SNPs avaliados, o tempo gasto pelo termociclador entre as temperaturas de anelamento e desnaturação é suficiente para a amplificação do fragmento. Além de diminuir o tempo das reações de PCR, esta estratégia mostrou-se mais eficaz para reduzir ou eliminar a maioria das amplificações inespecíficas (Figura 07 e Figura 08).

I.3.5 Validação dos SNPs

Para a validação experimental dos SNPs foram utilizados o seqüenciamento direto e a digestão com enzimas de restrição (RFLP). Havendo a possibilidade da utilização de enzimas de restrição na identificação dos alelos (<http://darwin.bio.geneseo.edu/~yin/WebGene/RE.html>) e havendo a disponibilidade da referida enzima, o RFLP foi o método preferencialmente utilizado na tentativa de validação dos SNPs. Nos casos de em que não foi possível utilizar este método, foi feito o seqüenciamento do fragmento contendo o polimorfismo, utilizando os mesmos iniciadores da PCR.

O seqüenciamento dos produtos de PCR foi feito utilizando os seqüenciadores automáticos MegaBace 1000, ABI 377 e ABI 3100.

Os experimentos de validação dos SNPs selecionados foram feitos utilizando um painel de 150 amostras, ou 300 alelos. Quando as variações não foram validadas nestas amostras voltamos às amostras de pacientes, com as quais foram feitas as

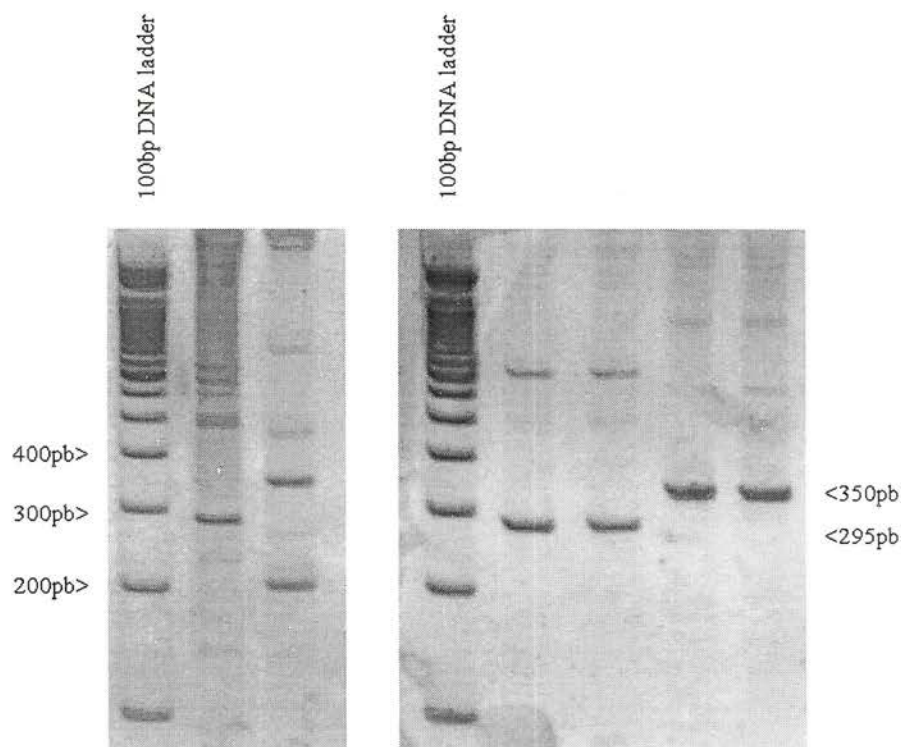


Figura 8: Avaliação do efeito da retirada do passo de extensão sobre o a PCR. No primeiro gel foram aplicadas as PCRs feitas com tempo de extensão e no segundo as mesmas PCRs feitas sem o tempo de extensão. Notar que no primeiro gel há uma quantidade significativa de bandas inespecíficas. No segundo gel cada uma das PCRs foi feita em duplicata.

seqüências que sugeriam a existência do SNP, para confirmar a existência do mesmo. Infelizmente nem todas as amostras de pacientes estavam disponíveis.

I.3.5.1 Amostras humanas

Para a validação dos SNPs humanos foi utilizado um painel de indivíduos de três diferentes grupos populacionais. Entretanto, definir grupos populacionais e étnicos no Brasil é algo extremamente falível e um levantamento mais detalhado sobre este ponto se encontra em Resultados e Discussão.

Foram “definidos” três grupos populacionais: brancos, negros e asiáticos. Para um indivíduo ser considerado asiático, seus quatro avós deveriam possuir origem asiática. Para brancos e negros os quatro avós deveriam possuir a mesma cor de pele. As informações pertinentes à definição dos grupos foi dada pelo próprio entrevistado.

Os grupos consistiam de 50 indivíduos de origem asiática, 50 indivíduos negros e 50 indivíduos brancos. Cada conjunto era composto por 25 homens e 25 mulheres. Estas amostras foram obtidas junto ao Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto. Cada doador assinou um termo de consentimento e a coleta deste material foi aprovada pelo Comitê de Ética dessa instituição.

I.3.6 SNPs em genes de *Schistosoma mansoni* candidatos a vacina

O Projeto Genoma de *Schistosoma mansoni* possui um núcleo de bioinformática distribuído entre os Laboratórios do Instituto de Química da USP, sob

a coordenação do Dr. Sergio Verjovski-Almeida, e do Laboratório de Bioinformática da Unicamp, sob a coordenação do Dr. João Carlos Setubal.

Cada sequência gerada pelo projeto teve sua qualidade avaliada pelo programa *Phred*, os iniciadores e vetores retirados e, utilizando os programas *Phrap* e *Cap3* as sequências foram alinhadas formando os *contigs*.

Estes *contigs* foram disponibilizados para inspeção através de uma interface *.html*, permitindo que as sequências que o compunham, a qualidade de cada uma das bases destas sequências e as discrepâncias pudessem ser visualizadas.

Quando em um mesmo ponto encontramos sequências originadas de bibliotecas diferentes, apresentando a mesma discrepância, classssificamos esta posição como possivelmente polimórfica. É importante ressaltar que para considerarmos a existência de um SNP as bases variantes necessariamente devem possuir alta qualidade (*Phred* acima de 20). As ESTs de *S. mansoni* depositadas no GenBank também foram alinhadas aos *contigs* já formados pelas sequências do PGSM para a identificação de outras variações. Nesta etapa inicial, a busca por SNPs em *S. mansoni* foi direcionada aos genes candidatos ao desenvolvimento de uma vacina contra este parasita.

Para os experimentos de validação foram selecionadas alterações nos genes candidatos a vacina, que fossem de alta qualidade e que fossem confirmadas por duas bibliotecas diferentes.

I.3.7 Mapeamento gênico e desenho dos iniciadores

O sequenciamento do genoma de *S. mansoni* ainda está em sua etapa inicial (<http://www.tigr.org/tdb/e2k1/sma1/intro.shtml>); desta forma foram utilizadas como

molde as seqüências de cDNA dos genes de interesse, para o desenho dos iniciadores a serem utilizados na validação dos polimorfismos deste parasita. As seqüências completas dos genes de *S. mansoni* candidatos à vacina foram obtidas junto ao GenBank (<http://www.ncbi.nlm.nih.gov>). Com o auxílio do grupo da bioinformática do PGSM as seqüências foram mascaradas contra repetições específicas de *S. mansoni* e contra regiões de baixa complexidade. Estas seqüências mascaradas foram submetidas ao programa *Primer3* para a construção dos iniciadores, seguindo, sempre que possível, os mesmos parâmetros descritos anteriormente para os SNPs humanos.

I.3.8 Implementação de novas técnicas de genotipagem

Nosso grupo, em colaboração com a *Applied Biosystems*, iniciou os primeiros testes para a implementação no Brasil de uma nova técnica de genotipagem de SNPs. Este protocolo baseia-se em um novo *kit* da *Applied Biosystems*, o *SNaPshot*, onde é feito, o seqüenciamento de uma única base, sendo a mesma o SNP de interesse. A partir da seqüência genômica, na qual se localiza o SNP, é desenhado um iniciador imediatamente anterior à variação, de modo que durante a reação de seqüenciamento a base do SNP seja a primeira a ser incorporada. São utilizados na reação somente os quatro ddNTPs, de forma que uma única base é incorporada para cada um dos alelos presentes. Cada um dos ddNTPs é marcado com um fluoróforo que emite luz em comprimento de onda diferente, assim podemos identificar qual das bases (A, C, G ou T) foi incorporada. O desenho dos iniciadores seguiu o mesmo protocolo para os desenhos dos iniciadores utilizados na PCR, mas de modo que a extremidade 3' do iniciador estivesse sempre uma base antes da variação.

Além de implementar este novo sistema foram feitos ensaios para otimizá-lo de forma que as genotipagens fossem feitas em sistemas de multiplex e em volumes de reação menores do que os recomendados pelo fabricante, o que representa economia de tempo e redução de custos. Os dados referentes a padronização do *SNaPshot* encontram-se detalhados em Resultados e Discussão.

I.3.8.1 PCR *multiplex*

Para o *kit SNaPshot*, existe a possibilidade de se genotipar até dez polimorfismos em uma mesma reação. Podemos, para tanto, fazer as PCRs, dos segmentos contendo os SNPs de interesse, em separado ou padronizar as amplificações em um sistema multiplex, o que facilita o manuseio das amostras e permite utilizar uma menor quantidade de material (DNAs e reagentes), além de economizar tempo e reduzir custos.

HENEGARIU et al. (1997) descrevem um protocolo simples para a implementação de PCR em multiplex e seguindo algumas de suas sugestões, foram feitas alterações nas concentrações de iniciadores e nos programas de amplificação. Inicialmente, todos os iniciadores foram adicionados ao *mix* da PCR nas concentrações utilizadas em reações individuais. Observamos que alguns fragmentos foram preferencialmente amplificados e, para normalizar os resultados, os iniciadores que mostraram melhores resultados tiveram suas concentrações diminuídas, enquanto aqueles que não apresentaram amplificações satisfatórias tiveram suas concentrações aumentadas na reação de PCR.

Outra alteração trabalhada foi o programa da PCR; segundo a experiência do Dr. Arthur Gruber (comunicação pessoal), o aumento de 1°C por segundo entre as

etapas de anelamento e extensão dos iniciadores aumenta a eficiência da reação. A própria temperatura de anelamento foi mais uma vez trabalhada, variando de 57 a 63°C.

Várias combinações das variantes acima mencionadas, foram testadas em amostras padrão; as combinações com melhores resultados foram aplicadas em grupos de amostras maiores para averiguar a reprodutibilidade dos sistemas escolhidos.

Resultados e Discussão

I.4.1 Análise dos SNPs

Utilizando a rotina de bioinformática descrita anteriormente, foram feitas as análises de todas as ORESTES, classificadas como fragmentos de genes humanos conhecidos, produzidas até janeiro de 2001, quando o banco de dados do HCGP atingiu a meta prevista de 1 milhão de seqüências. Dentre estas ORESTES, selecionamos apenas aquelas que mostraram identidade com genes humanos completos, contidos no nosso banco de âncoras (Tabela 02). Desta maneira, nos concentramos em 167.194 ORESTES geradas a partir de 4.390 bibliotecas (Tabela 02).

Um ponto comum aos diversos bancos de dados de seqüências é o seu dinamismo, e nosso banco não é uma exceção. Os resultados aqui apresentados se referem a uma das diversas análises as quais os dados foram submetidos, já que as consecutivas alterações no pipeline se refletem no conjunto final obtido. Na sessão de *Artigos relacionados a tese* podem ser identificadas discrepâncias entre os resultados apresentados no artigo de COSTA et al. e os apresentados nesta dissertação consequentes de diferentes análises de nosso banco de dados.

Nas estatísticas apresentadas como resultados da rotina de bioinformática acrescentamos filtros para a eliminação das seqüências que apresentavam muitas discrepâncias no alinhamento e um filtro que mantivesse na estatística final apenas variações confirmadas em pelo menos duas bibliotecas diferentes (filtro de duas bibliotecas), além de SNPs já depositados no dbSNP.

Tabela 2 - Número total de dados utilizados na identificação de SNPs e porcentagens de *clusters* onde estas variações foram encontradas.

	Total	Clusters com SNPs	%
ORESTES	167.194	915	0,7
CDS	7.950	516	6,5
Bibliotecas	4.390	1.527	34,8

Na ausência do uso do filtro de duas bibliotecas houve um total insucesso total nas validações. Em um painel de 20 possíveis polimorfismos selecionados, sem o uso do filtro de duas bibliotecas, nenhum foi validado em nossas amostras. Este resultado pode ser explicado pelas etapas existentes na geração de ORESTES onde são utilizadas duas polimerases, a transcriptase reversa e a DNA polimerase. Ambas as enzimas, podem inserir incorretamente nucleotídeos na nova fita de DNA que está sendo polimerizada, assim, mesmo seqüências de alta qualidade podem apresentar erros, não por falhas na reação de seqüenciamento mas por erros de incorporação das polimerases.

Para efeito de comparação, são apresentados os resultados obtidos, antes (Tabela 03) e após (Tabela 04) aplicar-se o filtro de duas bibliotecas. Observa-se uma redução de cerca de 90% no número de SNPs candidatos identificados, de um total de 2.845 para 275 SNPs. Esta drástica redução do número de SNPs pode ser atribuída a grande estringência dos parâmetros utilizados em nossas análises. Vale ressaltar que no banco de dados HCGP-SNP foram depositadas todos os SNPs sugeridos pela rotina de bioinformática, pois esta estringência elevada elimina polimorfismos reais que não foram representados por seqüências de bibliotecas diferentes.

Após a identificação dos 275 SNPs foi feita uma busca no banco de dados dbSNP para verificar quantos já haviam sido descritos. Obtivemos um total de 51 *hits*, ou seja, 18,5% dos SNPs identificados já haviam sido depositados no dbSNP, 27 sinônimos (20,3% de 133) e 24 não sinônimos (16,9% de 142), esta diferença na validação pelo dbSNP não foi significativa ($p \sim 0,95$).

Tabela 3 - Classificação dos candidatos a SNPs identificados antes do uso do filtro de duas bibliotecas.

SNPs	Transições	Transversões	Total
Sinônimos	797	141	938
Não sinônimos	1.349	558	1.907
Total	2.146	699	2.845

Tabela 4 - Classificação dos candidatos a SNPs identificados após o uso do filtro de duas bibliotecas.

SNPs	Transições	Transversões	Total
Sinônimos	110	23	133
Não sinônimos	85	57	142
Total	195	80	275

Apesar de todos os passos descritos acima para a seleção dos SNPs, fizemos ainda uma última verificação manual de cada um dos possíveis polimorfismos relatados e somente aqueles que passaram por esta inspeção prosseguiram nas validações em bancada. Cerca de 60% das variações foram excluídas das validações experimentais por não passarem na inspeção manual, isto se deve ao fato de haver no banco de dados seqüências submetidas várias vezes, com nomes ou datas diferentes, embora sejam as mesmas seqüências (Figura 9), ou ainda, por serem variações decorrentes da presença de parálogos que não foram excluídos pelo *PolyBayes*. Após todas as etapas de filtragem foram selecionados 26 SNPs para os experimentos de validação.

Vale mais uma vez lembrar que a alta estrigência dos parâmetros de bioinformática impossibilita que SNPs pouco freqüentes sejam identificados, já que há a necessidade de duas seqüências de bibliotecas diferentes para que consideremos a existência de uma possível variação.

I.4.2 Mapeamento genômico e desenho de iniciadores

Os passos aqui descritos foram de fundamental importância para o sucesso das validações. Somente após termos em mãos um banco curado de SNPs os experimentos de validação puderam ser iniciados de uma forma mais sólida. Para cada SNP foi selecionada uma seqüência no GenBank que possuía a maior cobertura do gene em questão. Por diversas vezes nos deparamos com seqüências de genes parálogos, sendo necessária uma busca cuidadosa pela seqüência correta, já que uma escolha errônea impossibilitava a validação do SNP. Possuindo a seqüência do gene, foram feitas análises por BLAST contra a versão mais atual do Genoma Humano

ORESTES Alignments

```

HSD00138 - TCGTTTTTCATTAAATATGAACTTTAGCAGTAAAA*GAACAAAATTTAAAAAT
RC3-BT0333-180700-111-d01 - ..a.....G.....g...
QV2-BT0463-150200-048-h05 - ..a.....G.....g...
QV2-BT0619-130400-139-b08 - ..a.....A.....g...
QV2-BT0636-130400-138-b08 - ..a.....A.....g...

```

Figura 9 -Alinhamento de ORESTES sugerindo a existência de um SNP. Notar no quadro em vermelho um possível erro de submissão onde a mesma sequência foi submetida duas vezes com nomes diferentes.

disponível no momento da análise. Para alguns casos as validações foram suspensas nesta fase, pois, ou não foi possível identificar o segmento genômico que continha a região da variação, ou não foi possível diferenciar o gene de interesse de seus parálogos.

Tendo sido identificado o segmento genômico correto, este foi submetido a um programa que identifica e mascara regiões repetitivas ou de pouca complexidade, o *RepeatMasker*. Este passo foi inserido para impedir que os iniciadores fossem desenhados nestas regiões o que levaria a um complexo padrão de amplificação devido ao grande número de regiões idênticas ou similares dispersas no genoma. Após acrescentarmos este passo, nenhum SNP deixou de ter sua validação continuada pela presença de elementos repetitivos.

De posse da região genômica onde o polimorfismo estava presente, conhecendo a distribuição dos exons dentro da mesma e tendo sido mascaradas as regiões repetitivas e de baixa complexidade, demos prosseguimento ao desenho dos iniciadores para a amplificação do segmento genômico que continha o SNP. Foi utilizado o programa *Primer3* que, dentro de nossa experiência, se mostrou altamente eficiente, sendo modificados os seguintes parâmetros:

- i) Tamanho do fragmento amplificado, mínimo 120pb, ótimo 150pb e máximo de 250pb;
- ii) Tamanho dos iniciadores, mínimo 17 e máximo 23 bases;
- iii) Os demais parâmetros utilizados foram os padrões do próprio programa, valendo ressaltar que a temperatura ótima de anelamento dos iniciadores era de 60°C, também padrão do programa.

Podemos ver na Figura 10 um exemplo de resultado obtido. O par de iniciadores mais indicado para ser utilizado aparece no cabeçalho enquanto que os demais iniciadores sugeridos são apresentados ao fim do relatório gerado por este programa.

Para cada polimorfismo foi criada uma pasta onde todas as informações anteriormente descritas foram depositadas. Nesta pasta podemos encontrar o alinhamento das seqüências ORESTES que sugeriam a existência do SNP; a seqüência e o número de acesso do GenBank do gene que foi utilizado nas buscas; o número de acesso do clone genômico onde foi localizado o gene; os alinhamentos: do SNP com o gene, do SNP com o clone genômico e do gene com o clone genômico; o resultado do *RepeatMasker*, e finalmente o resultado do *Primer3* (ver exemplo no Anexo 01).

I.4.3 Padronização das amplificações

Tendo em mãos os iniciadores, partimos para as padronizações das PCRs dos fragmentos genômicos que continham os SNPs. Inicialmente foram utilizadas para as padronizações amostras de DNA de voluntários do laboratório, para poupar amostra mais raras.

A PCR foi realizada nas seguintes condições: 50ng de DNA, 100µM de dNTPs, 2,5pmol de cada um dos iniciadores, 1U de *Taq* DNA polimerase e seu tampão 1× em volume final de reação de 25µL. O programa utilizado foi: 95°C por 3', 38x (95°C por 45'', Tm por 30''), 72°C por 5'.

Na padronização da ciclagem foram utilizadas as seguintes temperaturas de anelamento: 53°C, 55°C, 57°C, 60°C, 63°C e 66°C (Figura 07). Para cada par de

Identificação e análise de novos SNPs não sinônimos utilizando ORESTES

iniciadores foi escolhida a temperatura na qual havia um melhor rendimento de amplificação e uma menor presença de bandas inespecíficas.

Inicialmente, no programa da amplificação, havia um passo de extensão após o passo de anelamento, mas este foi retirado da PCR visto que aumentava o tempo de ciclagem e, principalmente, aumentava a quantidade de amplificação inespecífica da reação (Figura 07 e 08).

I.4.4 Validação dos SNPs em amostras populacionais

Em colaboração com o grupo do Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto, foram conduzidos os experimentos de validação dos SNPs em genes humanos, focando nossa atenção em SNPs que causassem alterações não sinônimas. As primeiras tentativas de validar por seqüenciamento os SNPs identificados foram realizadas em um seqüenciador automático MegaBace 1000 (*Amersham-Pharmacia Biotech*), a partir de protocolos já padronizados no laboratório de Descoberta Gênica da Instituto Ludwig de Pesquisa sobre o Câncer. Após varias tentativas sem resultados satisfatórios, partimos para a validação no ABI 377 (*Applied Biosystems*), do Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto, onde foi possível seguir com os experimentos de validação por seqüenciamento. Atualmente nosso Laboratório no Instituto de Psiquiatria, conta com um ABI 3100, (cedido por empréstimo pela empresa *Applied Biosystems*), e os experimentos de seqüenciamento neste aparelho já foram padronizados.

Vale ressaltar que o seqüenciador automático MegaBace 1000 foi e tem sido utilizado em vários projetos de seqüenciamento em larga escala (Projeto Genoma do

Cancer Humano, Projeto Genoma de *Schistosoma mansoni*, Projeto Genoma Nacional - *Agrobacterium violaceum*, entre outros), sendo obtidos resultados muito satisfatórios com o aparelho. Entretanto, nossa experiência sugere que, utilizando PCR de amostras de DNA genômico e não de DNA plasmidial a eficiência do seqüenciador é bastante reduzida, uma hipótese é que o aparelho é bastante sensível a oscilações em quantidades reduzidas de material o que não ocorre quando trabalhamos com material amplificado de plasmídeos.

Foram validados 13 SNPs, tanto por ensaios com enzima de restrição (Figura 11) quanto por seqüenciamento (Figura 12), e estes estão listados na Tabela 05. Na Tabela 06 são apresentadas as frequências encontradas para cada um dos alelos nas amostras populacionais e a frequência de indivíduos heterozigotos.

É interessante notar que a frequência de determinados polimorfismos varia consideravelmente entre os diferentes grupos, como exemplo extremo podemos citar o SNP no gene Tax1 cujo alelo variante está presente somente em indivíduos brancos (11/50 ou 22%) não sendo encontrado nenhum alelo variante entre os 50 indivíduos negros nem nos 50 indivíduos de origem asiática analisados. Este resultado vem a corroborar os dados de que estes polimorfismos humanos podem ajudar a traçar nossa história evolutiva (GABRIEL et al. 2002), embora a definição dos grupos amostrais e do N populacional avaliado ainda seja questionável. Podemos aqui citar PARRA et al. (2003), em seu artigo sobre a relação de caracteres fenotípicos (cor e textura dos cabelos, cor da pele e dos olhos, formato do nariz e espessura ds lábios) com marcadores genéticos de ancestralidade: “Nosso estudo torna claro como é perigoso associar cor ou raça com a origem geográfica, e o uso de termos como

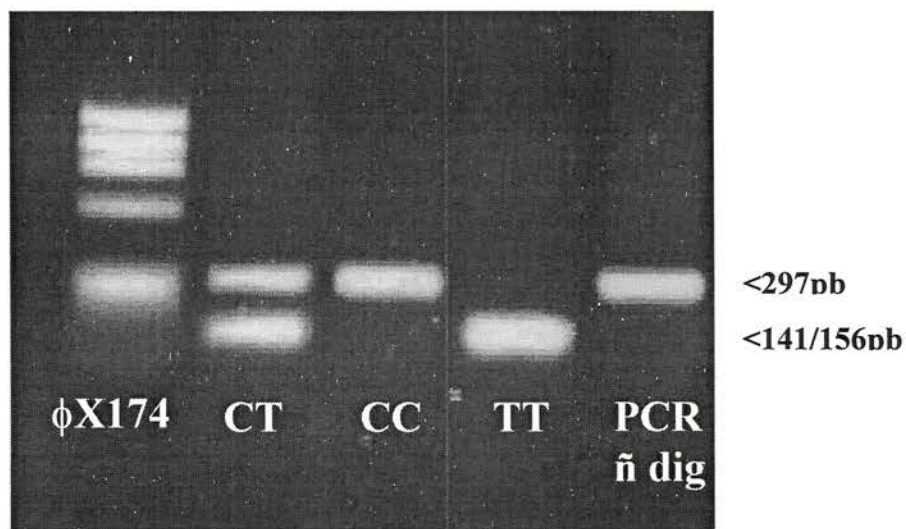


Figura 11 - Validação do SNP no gene da proteína similar a Podocalixina por RFLP. Aqui foi utilizada a enzima *Taq1* que digere o alelo T. Podemos ver, na seqüência, um indivíduo heterozigoto, um homozigoto C (297pb), um homozigoto T (um fragmento de 156 e outro de 141 que se sobrepõe no gel) e na última canaleta o produto de PCR não digerido (297pb). O padrão de peso utilizado foi o ϕ X174 RF DNA.

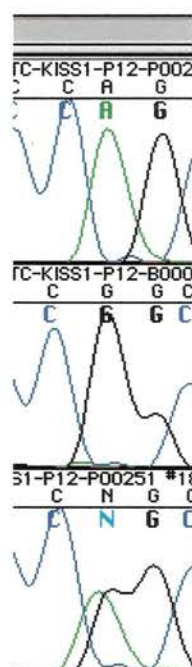


Figura 12 - Validação por seqüenciamento do SNP presente no gene KiSS-1. Podemos ver dois indivíduos homozigotos, cada um para um alelo (A e G), e um indivíduo heterozigoto (N).

Tabela 5 – Novos SNPs não sinônimos validados.

Gene	Variação de sequência	Variação de aminoácido
Human nidogen mRNA, complete cds (NID)	G>A	R>Q
<i>Homo sapiens</i> PDZ-73 protein (PDZ-73/NY-CO-38)	C>T	T>I
<i>Homo sapiens</i> mRNA for AEBP1, complete cds	A>G	K>E
Human ninjurin1 mRNA, complete cds (NINJ1)	A>C	D>A
<i>Homo sapiens</i> podocalyxin-like protein mRNA, complete cds (PODXL)	T>C	L>S
<i>Homo sapiens</i> Tax1 (human T-cell leukemia virus type I) binding protein 1 (TAX1BP1)	T>A	L>I
<i>Homo sapiens</i> thyroid hormone receptor coactivating protein mRNA, complete cds (SMAP)	C>T	T>M
Human serine protease mRNA, complete cds (TMPRSS2)	A>G	M>V
Human ladinin (LAD) mRNA, complete cds (LAD1)	A>G	K>E
Human radiated keratinocyte mRNA for cysteine protease inhibitor (CSTA)	C>T	T>M
Human elastase-specific inhibitor – Elafin (PI-3)	C>T	T>M
Human gamma-interferon-inducible protein (IP-30) mRNA, complete cds (IFI30)	G>A	R>Q
Human type IV collagenase mRNA, complete cds (MMP9)	G>A	R>Q
Human malignant melanoma metastasis-supressor (KiSS-1) gene, complete cds	A>G	Q>R
<i>Homo sapiens</i> transmembrane 4 superfamily member 3 (TM4SF3), mRNA	G>C	G>A
<i>H. sapiens</i> mRNA for carcinoembryonic antigen family member 2 (CGM2)	T>A	F>I

Tabela 6 - Frequência nas amostras populacionais dos novos SNPs não sinônimos validados.

Gene	Branços	Asiáticos	Negros	Heterozigotos
NID	100 alelos (16G-74A)	100 alelos (71G-29A)	100 alelos (30G-70A)	61/150 indivíduos 40,67%
PDZ-73	90 alelos (16T-74C)	64alelos (19T-45C)	80alelos (46T-34C)	13/117indivíduos 11,11%
AEBP1	68 alelos (34A-34G)	40 alelos (23A-17G)	74 alelos (35A-39G)	36/91indivíduos 39,56%
NINJ1	100 alelos(84A-16C)	98 alelos (80A-18C)	90 alelos (63A-27C)	54/144indivíduos 37,50%
PODXL	100 alelos (39T-61C)	100 alelos (55T-45C)	100 alelos (11T-89C)	59/150indivíduos 39,33%
TAX1BP1	100 alelos (87T-13A)	100 alelos (100T)	100 alelos (100T)	13/150 indivíduos 8,67%
SMAP	100 alelos (87C-13T)	100 alelos (96C-4T)	100 alelos (98C-2T)	6,04%
TMPRSS2	-	-	-	-
LAD1	14 alelos (11A-3G)	-	-	1/7 indivíduos 14,28%
CSTA	100 alelos (76C-24T)	100 alelos (62C-38T)	-	19/100 indivíduos 19%
PI-3	100 alelos (86C-14T)	100 alelos (100C)	100 alelos (96C-4C)	4/150 indivíduos 2,66%
IFI30	24 alelos (19G-5A)	14 alelos (14G)	-	3/19 indivíduos 15,79%
MMP9	24 alelos (11G-13A)	-	-	9/12 indivíduos 75%
KiSS-1	38 alelos (35A-3G)	-	Pacientes: 8 alelos (7A - 1G)	2/23 indivíduos 8,67%
TM4SF3	102 alelos(62G-40C)	100alelos (73G-27C)	100 alelos (92G - 8C)	54/151indivíduos 35,76%
CGM2	48 alelos (28T-20A)	-	-	12/24indivíduos 50,00%

branco, Caucasiano ou Europeu e seus correspondentes negro ou Africano, como são freqüentemente empregados na literatura médica e científica”.

Torna-se então arbitrária a distinção de etnias dentro da população Brasileira; entretanto, para efeitos de comparação seguem os resultados das genotipagens feitas utilizando os três diferentes grupos populacionais e um breve resumo sobre cada gene para o qual um SNP foi validado.

I.4.4.1 Tax1 (Vírus tipo 1 da leucemia de célula T)

A proteína HTLV-1 Tax (LocusLink: 8887, Citogenética: 7p15) ativa transcricionalmente o promotor de HTLV-1, ela também se liga e estimula a expressão de genes celulares incluindo fatores de transcrição (OMIM 605326). A interação de Tax1 com a proteína anti-apoptótica A20, inibe apoptose induzida por TNF.

Como já mencionado o polimorfismo em Tax1 foi validado somente em indivíduos brancos, estando ausente no grupo de negros e asiáticos.

I.4.4.2 Ladinina (LAD)

Ladinina é uma proteína de ancoragem da membrana basal na junção dermo-epidermal (LocusLink: 3898, Citogenética: 1q25.1-q32.3). A proteína humana é um autoantígeno associado a IgA (OMIM 602314).

AL MOUSTAFA et al. (2002) descrevem um menor expressão de Ladinina, cerca de 50 vezes, em Carcinomas de Células Escamosas de Cabeça e Pescoço em relação às amostras de tecido normal, retiradas dos mesmos pacientes.

Não há sítio de restrição para a genotipagem deste polimorfismo e o seqüenciamento desta região não foi padronizado, desta forma somente 7 indivíduos foram genotipados com segurança, assim encontramos somente um heterozigoto e um homozigoto para a variação.

I.4.4.3 Inibidor específico de elastase (Elafin)

Descrita como um inibidor de elastase, PI3 ou Elafin foi isolado de lesão de psoríase (LocusLink: 5266, Citogenética: 20q12-q13), embora possa ser encontrada em várias outras patologias inflamatórias da pele, não sendo encontrada em tecido normal. Ela é encontrada na urina de pacientes com psoríase e testes imunohistoquímicos demonstraram que Elafin pode ser encontrada em queratinócitos diferenciados da suprabasal em epiderme psorítica (OMIM: 182257).

Este polimorfismo não foi encontrado no grupo de asiáticos e sua frequência foi de 14% em brancos e 4% em negros.

I.4.4.4 Colagenase tipo IV

Proteínas da família de metaloproteinases de matriz estão envolvidas em processos fisiológicos normais de quebra da matriz extracelular, assim como desenvolvimento embrionário, reprodução, remodelamento de tecidos e com processos patológicos bem como artrite e metástase. A maioria das metaloproteinases de matriz são secretadas como proteínas inativas que somente após clivagem por proteinases extracelulares se ativam. A colagenase tipo IV ou metaloproteinase de matriz 9 (LocusLink: 4318, Citogenética: 20q11.2-q13.1), é uma enzima que degrada colágenos tipo IV e V. Estudos em modelos murinos sugerem que esta

enzima está envolvida em processos associados a remodelamento de tumor (OMIM: 120361).

Na população branca genotipada o alelo variante está presente com uma frequência de 54%, ou seja, em uma maior proporção em relação ao alelo selvagem descrito.

I.4.4.5 Inibidor de cisteino-protease

Este gene codifica uma proteína que possui atividade inibidora de cisteino-protease (LocusLink: 1475, Citogenética: 3q21), formando complexos com papaina e catepsinas B, H e L. Esta proteína é uma das precursoras do envelope celular em queratinócitos e possui um papel no desenvolvimento e manutenção da epiderme, tendo sido proposta como um marcador para diagnóstico e prognóstico de câncer (OMIM: 184600).

I.4.4.6 Proteína induzida por interferon Gama (IP-30)

Proteínas internalizadas pela via de endocitose são normalmente degradadas. Uma proteólise eficiente requer desnaturação induzida por um ambiente ácido dentro dos lisossomos e redução das ligações dissulfeto inter e intra-cadeias. IP-30 ou GILT (LocusLink: 10437, Citogenética: 19p13.1) é uma tiol redutase lisossomal que possui atividade ótima em pHs baixos e é capaz de catalizar a redução de ligações dissulfeto *in vitro* e *in vivo*. IP-30 codifica uma proteína de 261 aminoácidos composta de um peptídeo sinal de 37 aminoácidos e uma porção ácida de 224 resíduos. A enzima é expressa constitutivamente em células apresentadoras de antígenos e é induzida por

interferon Gama em outros tipos celulares sugerindo um importante papel no processamento de antígenos (OMIM 604664).

O alelo variante não foi encontrado em asiáticos e sua frequência no grupo de brancos foi de 20,8%.

I.4.4.7 Proteína similar a Podocalixina

As células renais do epitélio glomerular ou podócitos, são células altamente diferenciadas com processos interdigitantes característicos voltados para a membrana basal glomerular. Estes processos são revestidos em sua superfície apical por um glicocálix polianiónico o qual é essencial para o filtro glomerular, sendo a podocalixina (LocusLink: 5420, Citogenética: 7q32-q33) o maior componente deste glicocálix. Análises de *Northern blot* revelaram que a Podocalixina é expressa em três diferentes transcritos, um de 5,9kb (mais abundante) e outros dois de 4,4kb e 9,6kb em vários tecidos, sendo a maior expressão nos rins, pâncreas e coração. A proteína predita tem 528 aminoácidos, iniciando com um peptídeo sinal de 21 aminoácidos, seguindo um domínio transmembrana e um domínio extracelular fortemente ácido. A sequência de proteína da podocalixina humana é 48% idêntica à de coelho possuindo 96% de identidade nos domínios transmembrana e extracelular (OMIM:602632).

A frequência do alelo variante também oscila nos três grupos indo de 45% em asiáticos a 89% em negros passando por 61% em brancos.

I.4.4.8 Ninjurina 1

Danos em axônios podem alterar a expressão de determinados genes, sendo um destes a Ninjurina 1 (LocusLink: 4814, Citogenética: 9q22). A expressão deste gene é aumentada após o dano sendo que a proteína por ele codificada está relacionada a mecanismos de adesão celular e reparo axonal (ARAKI e MILBRANDT 1996). O domínio crítico de adesão mediada pela Ninjurina 1 está entre os aminoácidos 26 e 37 da proteína estando estes na porção extracelular. O gene encontra-se localizado entre os marcadores D9S196 e D9S197, região candidata à associação com Neuropatia sensorial hereditária tipo 1 e Síndrome de predisposição a epiteloma escamoso auto-remissivo (OMIM 602062).

O alelo variante está presente em 30% dos negros, 18% dos asiáticos e 16% dos brancos. Dos 144 indivíduos genotipados somente 7 são homozigotos para o alelo variante.

I.4.4.9 Proteína ligante de estimulador de adipócitos 1 (AEBP1)

A proteína ligadora de estimulador de adipócitos 1 (LocusLink: 165, Cytogenetic: 7p13) é um repressor transcricional com atividade carboxipeptidase, sua atividade consta da clivagem de fatores envolvidos com a transcrição.

A proteína predita de AEBP1 contém 1158 aminoácidos, sendo composta de um provável peptídeo sinal, um motivo de 11 resíduos rico em Prolina e Lisina repetido quatro vezes, um domínio com 30% de identidade com a proteína discoidal e uma porção carboxi-terminal com 39% de identidade com a carboxipeptidase E.

Camundongos *Knock-out* para o gene AEBP1 apresentaram problemas de desenvolvimento e de reparo de tecidos; os embriões abortados apresentavam múltiplas alterações nos órgãos viscerais abdominais e os indivíduos adultos eram incapazes de curar feridas apresentando pobre proliferação de fibroblastos (OMIM 602981).

I.4.4.10 Gene da Síndrome de Usher tipo 1C (USH1C)

A síndrome de Usher é um defeito sensorial autosômico recessivo envolvendo surdez e cegueira devido a um retinite pigmentosa progressiva. O gene responsável pela síndrome de Usher codifica uma proteína, denominada harmonina (LocusLink: 10083, Citogenética: 11p14.3), que possui um domínio PDZ. Em pacientes com a síndrome de Usher já foram descritas mutações nos sítios de splicing, inserções e deleções na janela de leitura do gene e expansões de microsatélites em segmentos intrônicos (OMIM: 605242).

I.4.4.11 Nidogene

Nidogene ou entactina (LocusLink: 4811, Citogenética: 1q43) é uma glicoproteína sulfatada de 150KDa que liga colágeno tipo IV e laminina sendo o maior componente das membranas basais (OMIM: 131390).

I.4.4.12 Serino protease (SerProt)

Este gene codifica uma proteína que pertence à família de serino proteases (LocusLink: 7113, Citogenética: 21q22.3). As serino proteases são conhecidas por atuar em uma série de processos fisiológicos e patológicos. Este gene foi

demonstrado ser induzido por hormônios andrôgenos em tumores de próstata, sendo verificada uma menor expressão em tecidos tumorais de próstata não dependentes de andrógeno. A função biológica desta proteína é desconhecida (OMIM: 602060).

I.4.4.13 Proteína coativadora do receptor de hormônio da tireóide (THRCP)

THRCP (LocusLink: 10902, Citogenética: 5q31) interage com o receptor de hormônio da tireóide de forma ligante dependente, aumentando assim a ativação dependente de hormônio da tireóide por elementos de resposta da tireóide. A proteína contém um bromodomínio através do qual é um coativador nuclear (OMIM: 602848).

I.4.4.14 Membro 2 da família de antígenos carcinoembriônicos (CGM2)

A família gênica de antígenos carcinoembriônicos (CEA), pertence à superfamília de genes de imunoglobulinas. O membro 2 da família de antígenos carcinoembriônicos ou molécula 7 de adesão celular relacionada a antígeno carcinoembriônico é um potencial supressor tumoral (LocusLink: 1087, Citogenética: 19q13.2).

Este polimorfismo foi avaliado em 42 cromossomos de indivíduos brancos, e encontramos uma frequência de 41,6% para o alelo variante.

I.4.4.15 Membro 3 da superfamília 4 de proteínas transmembrana (TM4SF3)

Também denominado de antígeno CO-029 associado a tumor (LocusLink: 7103, Citogenética: 12q14.1-q21.1) este gene codifica uma proteína membro da superfamília 4 de proteínas transmembrana ou família tetraspanina. A maior parte dos membros desta família são proteínas de superfície celular, caracterizadas pela presença de quatro domínios hidrofóbicos. Estas proteínas medeiam eventos de transdução de sinal que participam da regulação do desenvolvimento, ativação, crescimento e motilidade celular. TM4SF3 codifica uma glicoproteína de superfície celular que sabidamente forma complexos com integrinas. Este gene é expresso em diferentes carcinomas (gástrico, cólon-retal e pancreático) mas sua expressão não é observada na maioria dos tecidos normais.

A proteína é contituida de 237 aminoácidos e possui similaridade com outras proteínas transmembrana como: antígeno associado a melanoma ME491 ou CD63; SAS antígeno de sarcoma e antígeno de superfície celular de leucócito CD37 (OMIM 600769). A alteração encontrada está na posição 73 da proteína (NM_004616), dentro do segundo segmento transmembrana, e consiste em uma troca de glicina para alanina. Esta região da proteína, incluindo a posição em questão, é extremamente conservada em relação às demais proteínas transmembrana acima mencionadas e ao antígeno Sm23 de *S. mansoni* (SZALA et al. 1990).

As oscilações entre as frequências observadas nos três grupos são interessantes, 4% em negros, 27% em asiáticos e 40% em brancos, ou seja, aumentando à medida em que nos distanciamos da África.

I.4.4.16 Supressor de metástase de melanoma maligno (KiSS-1)

O gene KiSS-1 (LocusLink: 3814, Citogenética: 1q32) é um regulador de metástase em melanomas malignos e em câncer de mama, entretanto sua expressão não afeta a tumorigênese (WELCH et al. 1994; LEE e WELCH 1997).

A proteína codificada por este gene consiste de 145 aminoácidos sendo que OHTAKI et al. (2001) demonstraram que um peptídeo de 54 resíduos isolado de placenta humana é um ligante de um receptor acoplado a proteína G (aminoácidos 68 a 121, o polimorfismo validado está na posição 36). Este peptídeo denominado metastina, inibe quimiotaxia e invasão em culturas celulares, também atenuando metástases pulmonares de algumas linhagens de melanoma *in vivo*. Análises de northern blot demonstraram que nenhuma linhagem metastática em camundongos nude atímicos expressavam KiSS-1 (OMIM 603286).

O polimorfismo validado consiste em uma troca de A para G acarretando a alteração de uma glutamina para arginina. Esta posição está entre o peptídeo sinal da proteína e o peptídeo que interage com o receptor acoplado a proteína G.

O alelo variante tem uma frequência de 7,9% dentro de nossa população branca genotipada (19 indivíduos). Também foi possível genotipar os pacientes de onde foi extraído mRNA para a geração das ORESTES, do total de quatro pacientes somente um alelo foi identificado.

Vale ressaltar que acima estão descritos 16 e não 13 genes para os quais foram validados SNPs. No decorrer deste Projeto de Mestrado foi iniciado um projeto para a análise geral das ORESTES geradas pelo HCGP, esta iniciativa liderada pelo Instituto Ludwig de Pesquisa sobre o Câncer contou com a participação de uma

parcela significativa dos envolvidos no HCGP além de pesquisadores estrangeiros. Uma das metas desta iniciativa, denominada Jamboree-ORESTES, foi a identificar e validar SNPs em genes relacionados ao processo de tumorigênese. Nosso grupo aliado ao grupo do Hemocentro de Ribeirão Preto liderou e executou as tarefas referentes a esta etapa que culminou com a validação dos SNPs em KiSS-1, TM4SF3 e CGM2 utilizando para tanto as mesmas amostras populacionais.

I.4.5 SNPs em genes de *Schistosoma mansoni* candidatos a vacina

Ao avaliar os alinhamentos montados pelo *Cap3* e pelo *Phrap* pudemos identificar três polimorfismos presentes em genes de *S. mansoni* candidatos à vacina. Os critérios aqui utilizados foram: i) o polimorfismo deveria ser encontrado em regiões de alta qualidade de sequência e, ii) deveria estar presente em no mínimo duas sequências de bibliotecas diferentes, iii) gerando alteração não sinônima. Foram encontrados SNPs em todos os genes candidatos a vacina preconizados pela Organização Mundial da Saúde (OMS) (Tabela 07; Figura 13, exemplo).

Para a validação experimental dos SNPs nos genes candidatos à vacina, foram selecionados 3 polimorfismos não sinônimos (1 em 14-3-3, 1 em glutathione-S-transferase e 1 em triose fosfato isomerase) para os quais foram desenhados iniciadores utilizando a sequência dos mRNAs. Não foram usados os segmentos genômicos correspondentes, já que o genoma de *S. mansoni* ainda não foi sequenciado. A princípio, o fato de termos utilizado a sequência dos mRNAs não constitui um problema em si, já que os *introns* de *S. mansoni* são geralmente pequenos, em torno de 35pb, e por isso não impossibilitam a amplificação dos fragmentos de interesse a partir de DNA genômico.

Tabela 7 - SNPs identificados a partir das ORESTES gerados pelo PGSM nos genes candidatos a vacinais de *S. mansoni*.

	Sm14	GST	TPI	Paramiosina
Quantos SNPs sugeridos	1	5	2	1
Quantos SNPs em UTRs sugeridos	0	1	0	0
Quantos SNPs sinônimos sugeridos	0	2	1	0
Quantos SNPs não sinônimos sugeridos	1	2	1	1
Quantos SNPs avaliados	0	1	1	0
Quantos SNPs confirmados	0	1	1	0

Para os experimentos de validação o segmento genômico contendo a possível variação foi amplificado a partir de um *pool* de vermes, clonado e sequenciado. É importante ressaltar que a PCR foi feita utilizando-se enzimas que possuíam atividade corretora (*Pfu* e *Platinum*), a fim de evitar possíveis erros inseridos pelas mesmas. Dois polimorfismos foram confirmados, os SNPs dos genes triose fosfato isomerase e glutathione-S-transferase (Figura 14), não sendo possível confirmar o SNP em 14-3-3.

I.4.6 Falsos SNPs

Como já mencionado, no dbSNP já estão submetidos cerca de 4,9 milhões de polimorfismos (28/01/2003) número muito próximo do valor total estimado de 5 milhões. Entretanto, este elevado número de polimorfismos já submetidos é, em parte, fictício já que uma parcela considerável é redundante, (cerca de 38% dos SNPs submetidos, Tabela 01) e outra parcela destas variações de ponto são irreais, refletindo na realidade discrepâncias entre parálogos (BAILEY et al. 2002) ou mesmo diferenças entre seqüências de qualidade duvidosa. BANSAL et al. (2002) relatam uma taxa de validação de ~60% em um grupo de 200 mil SNPs depositados neste banco.

Em nosso projeto, existem alguns pontos que devem ser considerados já que a técnica ORESTES favorece o surgimento de variações virtuais. Devemos lembrar que a técnica contém um passo de transcrição reversa e outro de PCR, ambos podendo inserir alterações, ou erros, nas seqüências geradas. SANSON et al. (2002) utilizam destas variações adquiridas para construir uma filogenia experimental, e um dos dados obtidos é que as polimerases que hoje são usadas em PCRs possuem uma

[illegible]

Figura 14 - Validação do SNP em GST de *Schistosoma mansoni*. Aqui podemos ver o alinhamento que confirma a existência da variação, nas seqüências geradas a partir dos diferentes isolados.

tendência a aumentar o conteúdo de GC das fitas que estão sendo polimerizadas pois são derivadas, em sua maioria, de organismos provenientes de fontes termais onde há a necessidade de um alto conteúdo de GC para a estabilidade do genoma.

Nenhum dos SNPs não validados foi encontrado nas amostras de pacientes utilizadas na geração das ORESTES (quando disponíveis), vale também ressaltar que dos 9 SNPs não validados 6 são confirmados por seqüências do CGAP, aqui vale recordar que as seqüências geradas pelo CGAP também passam por um passo de transcrição reversa. Destes 9 polimorfismos não validados 4 são alterações do tipo H>G (H = A,C ou T), ou seja quase 50% um número significativamente maior do que $\frac{1}{4}$ o qual seria o valor esperado. A partir destes dados podemos tentar entender, um pouco melhor, os eventos que estão por detrás das não validações. É sabido que a transcriptase reversa, uma enzima retroviral, tem por finalidade promover a variabilidade de seu genoma por indução de mutações e vem de encontro com parte de nossos achados o fato de que existe uma tendência de indução de hipermutações, do tipo A>G, por enzimas retrovirais (KIM et al., 1996). Os dados por nós encontrados unidos aos dados de KIM et al. (1996) e SANSON et al. (2002) ajudam então a explicar o porque do número de SNPs não validados .

Apesar dos 60% de taxa de validação, este número é bem confortável já que coincide com os dados descritos por Bansal *et al.* (2002, ~60%) para os dados de validação encontrados em 200 mil SNPs depositados no dbSNP. Outro dado relevante é de que nossos dados favorecem o encontro de polimorfismos não sinônimos, os quais possuem uma maior aplicabilidade pelo seu maior potencial de reflexo fisiológico, podendo por vezes serem polimorfismos causais e não apenas marcadores.

I.4.7 Implementação do *SNaPshot*

A fim de inserir na rotina do laboratório novos métodos de genotipagem, nosso grupo em colaboração com a *Applied Biosystems* tem testado um novo *kit* de genotipagem de SNPs, o *SNaPshot*.

Os objetivos eram: implementar o *kit* e testar um sistema de multiplex, onde vários SNPs são genotipados ao mesmo tempo. Inicialmente foram escolhidos cinco SNPs já validados (LAD, SerProt, IP30, Tax1 e KiSS-1) para os primeiros testes com o *kit*. A escolha destes SNPs se deu pela dificuldade encontrada para genotipagem por sequenciamento ou RFLP. Para cada um deles foi desenhado um iniciador que estivesse imediatamente anterior à base que continha a variação e somente para um deles (LAD) foram construídos iniciadores nos dois sentidos. Os seis iniciadores foram desenhados de maneira que cada iniciador diferia de quatro bases um do outro, isto foi feito de forma a tentar realizar a reação de multiplex. A diferença de quatro bases é sugerida pelo *kit* e serve para que os alelos de cada SNP possam ser identificados por diferença de tamanho quando reunidos na mesma eletroforese.

Foi feita a reação de PCR em 12,5µL para a amplificação de cada um dos segmentos genômicos que continham o SNP. Cada uma das PCRs foi tratada com 3U de Fosfatase alcalina bovina (CIP) e com 1U de Exonuclease I (Exo I), por 1h a 37°C, seguido de 15min a 75°C. A CIP tem como função retirar os grupos fosfatos dos nucleotídeos que não foram incorporados na PCR e que poderiam influenciar na reação de genotipagem, já a Exo I retira todo o restante dos iniciadores que não foram utilizados na PCR e que também poderiam intervir na reação de genotipagem. Após o tratamento existe um passo de 15 min a 75°C para desnaturar as duas enzimas.

A reação de genotipagem consta basicamente de DNA molde (produto da PCR), ddNTPs marcados, uma DNA polimerase e o iniciador. A reação é feita em um volume final de 10 μ L onde temos 3 μ L da PCR, 5 μ L do *kit* do *SNaPshot* (aqui temos a polimerase e os dideoxiribonucleotídeos marcados) e 0,2 μ M do iniciador. Para podermos testar a quantidade de produto de PCR a ser utilizado na reação de genotipagem foram feitas várias diluições: 1:1, 1:3, 1:6, 1:12, 1:24 e 1:48.

No protocolo para várias genotipagens na mesma reação (multiplex) as PCRs foram realizadas separadamente mas os produtos e os iniciadores das mesmas foram diluídos em conjunto. Independentemente do número de iniciadores na reação a concentração final de cada um deveria sempre estar em 0,2 μ M.

Aqui procedemos então com a reação de genotipagem que consta apenas da extensão de uma única base sendo ela o SNPs de interesse. O protocolo seguido foi o seguinte: 25x (96°C por 10", 50°C por 5", 60°C por 30").

Após a reação os produtos foram tratados com 1U de CIP por 1h a 37°C e 15min a 75°C, para a retirada dos dideoxinucleotídeos não incorporados, os quais co-migram com os fragmentos de interesse interferindo nas análises. Sendo assim, 0,5 μ L de cada uma das reações foi adicionado em 9 μ L de Formamida e 0,5 μ L do padrão de peso molecular, o LIZ-120. As amostras foram então desnaturadas a 96°C por 5min e submetidas a eletroforese no ABI3100 utilizando capilares de 36cm.

Os primeiros resultados estão na Figura 15. As diluições mais adequadas foram as 1:24 e 1:48. Podemos verificar a presença do alelo G para o gene IP30 (ou C se fosse na fita complementar), a presença dos dois alelos T e A para o gene Tax1 e a presença do alelo A para o gene KiSS-1. Na Figura 16 podemos ver também que obtivemos bons resultados com a reação em *triplex*.

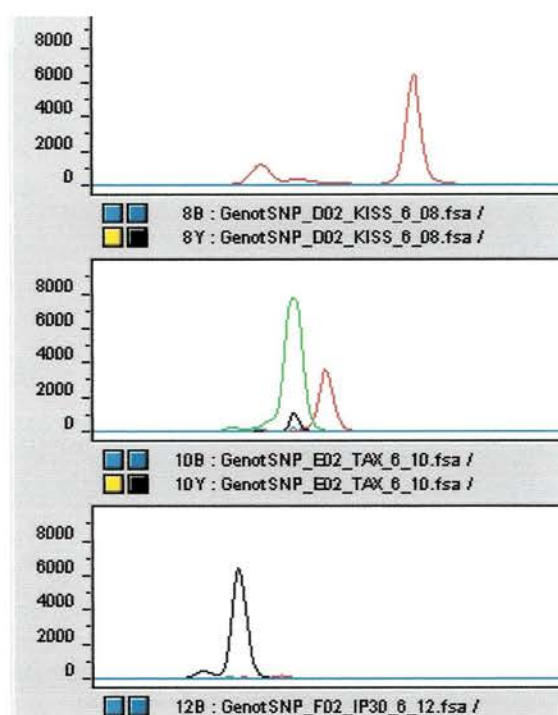


Figura 15 - Primeiros resultados obtidos com o kit *SNaPshot*. Podemos ver a genotipagem para os SNPs dos genes KiSS-1, Tax1 e IP30 respectivamente. Notar que este indivíduo é heterozigoto para o gene Tax1.

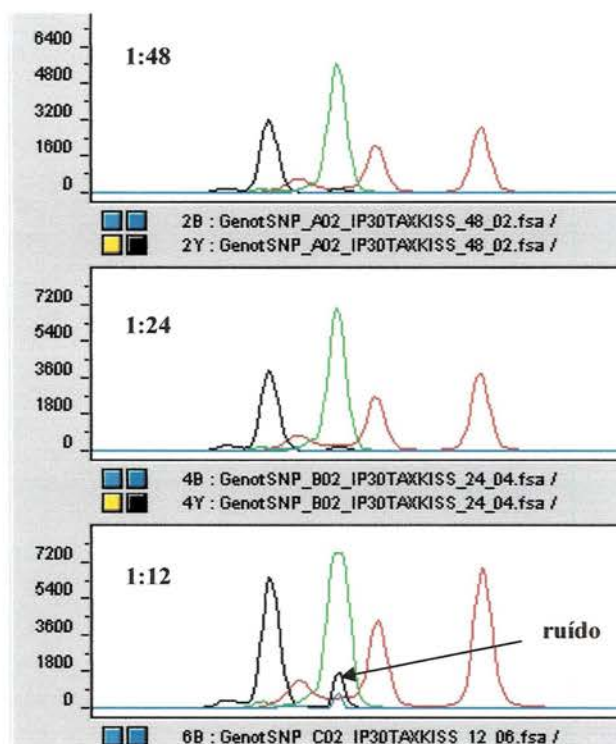


Figura 16 - Genotipagem com o kit *SNaPshot* em reação de triplex para os SNPs dos genes IP30, Tax1 e KiSS-1. Diluições 1:48, 1:24 e 1:12 respectivamente. Notar que com a diminuição das diluições, de cima para baixo, temos uma maior interferência da linha de base (ruído).

O passo seguinte foi tentar reduzir o volume de reação já que apenas 0,5µL de um volume de 10µL é utilizado e tentar uma reação de *pentaplex*.

A diluição do produto da PCR escolhida para ser utilizada nas reações de genotipagem foi de 1:30, um ponto intermediário entre 1:24 e 1:48. Novamente foi testado o triplex desta vez em um maior número de amostras e o protocolo seguido para o pentaplex foi basicamente o descrito acima. Também foram feitas reações de genotipagens (*triplex* e *pentaplex*) em volumes finais de 5µL e de 2,5µL obedecendo sempre às mesmas concentrações do protocolo padrão.

Mais uma vez os resultados obtidos foram muito satisfatórios, conseguimos reduzir a $\frac{1}{4}$ o volume de *kit* utilizado (Figura 17) e, embora o pentaplex não tenha sido obtido conseguimos bons resultados para quatro dos cinco SNPs testados (Figura 18). Isso quer dizer que podemos multiplicar cada reação inicial do *kit* por 16, pois podemos testar quatro SNPs ao mesmo tempo e multiplicar por quatro a rentabilidade do *kit*, ao usar 1,25µL do mesmo e não 5µL como sugerido. O mais interessante foi que não houve diferença entre os resultados obtidos com as reduções de volume realizadas com o *kit* (Figura 18), isso leva-nos a pensar em realizar um novo teste e reduzir ainda mais o volume de *kit* utilizado por reação.

Já que o *pentaplex* não funcionou para todos os SNPs algumas dúvidas ainda devem ser esclarecidas e novos testes deveram ser realizados para tentarmos atingir o limite sugerido do *kit* que é um decaplex. Caso o decaplex funcione associado a uma reação de 2,5µL o preço de cada genotipagem cai em 40 vezes o que torna o uso do *SNaPshot* extremamente atraente. Realizamos alguns testes com o primer que não havia funcionado para o sistema de *pentaplex* (LAD), e descobrimos que houve uma

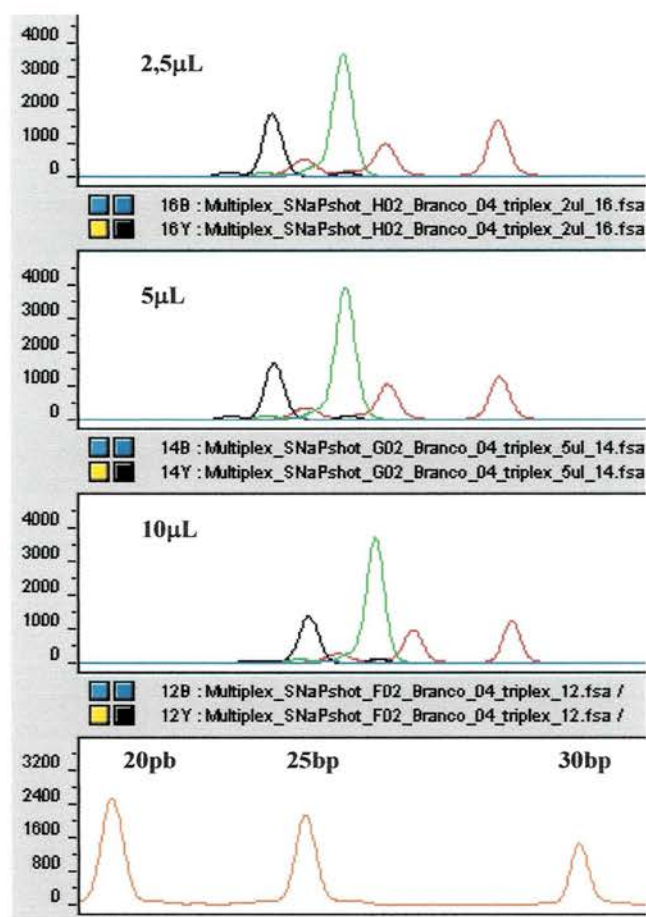


Figura 17 - Redução do volume de reação de genotipagem. Aqui foram utilizadas as condições indicadas no kit, ou seja, reação em 10µL e duas tentativas de redução de volume 5µL e 2,5µL. Notar que não existe nenhuma alteração detectável entre as três condições. O quadro inferior é o padrão de peso molecular LIZ-120.

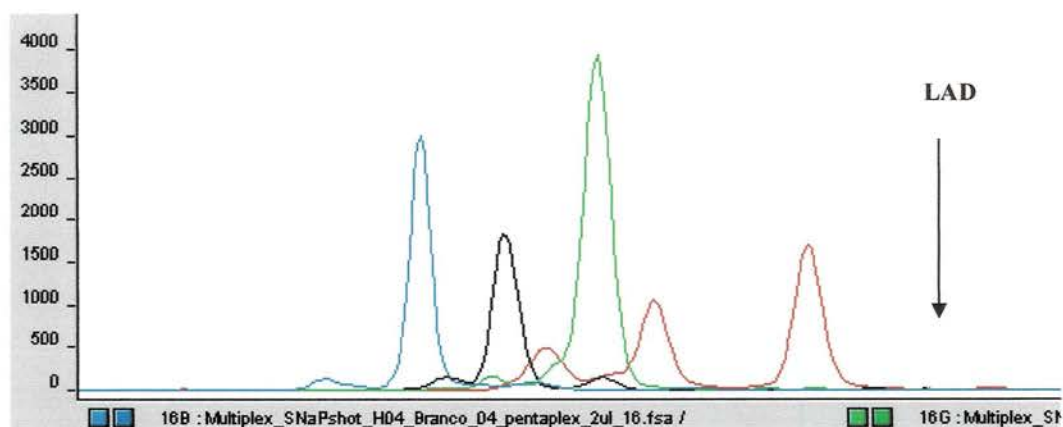


Figura 18 - Tentativa de genotipagem em sistema de pentaplex. Embora uma das genotipagens não tenha funcionado conseguimos obter resultados para quatro SNPs. Na figura esta marcado o ponto onde deveria aparecer um dos alelos do gene Ladinina.

falha na síntese deste iniciador. Foi solicitada uma resíntese deste iniciador para podermos continuar com os experimentos.

Um outro ponto a ser testado é a diluição do LIZ-120. Utilizando 0,5µL deste marcador estamos obtendo cerca de 900 pontos de intensidade de fluorescência e este valor pode ser reduzido, com segurança até 200 pontos de fluorescência, ou seja, podemos tentar novas corridas usando de 0,2 a 0,1µL do marcador, mais uma vez diminuindo os custos do procedimento.

Estes resultados nos permitem vislumbrar uma maior capacidade de genotipagens com um custo operacional extremamente atraente.

I.4.7.1 PCR *multiplex*

Outro ponto a ser trabalhado é a realização das reações de PCR no sistema de *multiplex*. Atualmente cada PCR é feita em separado, e somente na reação de genotipagem os produtos são reunidos. Caso consigamos padronizar todas as PCRs em uma mesma reação isto poupa tempo, diminui a manipulação, evitando possíveis erros, e diminuindo os custos.

Entretanto, os experimentos conduzidos até o momento foram totalmente insatisfatórios. Foram realizadas várias tentativas de se padronizar a PCR em *multiplex* para os 5 polimorfismos para os quais o *SNaPshot* foi trabalhado (LAD, SerProt, IP30, Tax1 e KiSS-1). Em um primeiro momento foram trabalhadas as concentrações individuais dos pares de iniciadores utilizados para a amplificação dos fragmentos contendo as variações. Iniciamos com a concentração utilizada nas PCR convencionais, ou seja 0,2µM.

Fundação Antonio Prudente

Prof.ª Maria Rodrigues Alves Kuninari
Coordenadora Pós-Graduação

Conseguimos ótimos resultados com os amplicons de KiSS-1 e Tax1, e resultados não tão animadores para LAD e IP30, já SerProt interfere na reação em multiplex e seus iniciadores foram retirados do sistema, provavelmente por interagir com os outros iniciadores.

Chegamos a trabalhar com 0,05 μ M dos iniciadores de Tax1, 0,1 μ M dos iniciadores de Kiss-1, 0,33 μ M dos iniciadores de IP30 e 0,5 μ M dos iniciadores de LAD onde obtivemos os melhores resultados, utilizando o programa 94°C por 5', 45x (94°C por 30'', 62°C por 30'', 1°C/s até 72°C, 72°C por 10''), 72°C por 5' (Figura 19). Não conseguimos uma uniformidade na intensidade das bandas, mas todas as quatro amplificaram o suficiente para os experimentos no *SNaPshot*. Infelizmente ao transpor este protocolo para nossas amostras populacionais os resultados não foram reprodutíveis inviabilizando por agora, o prosseguimento das PCRs em *multiplex*.

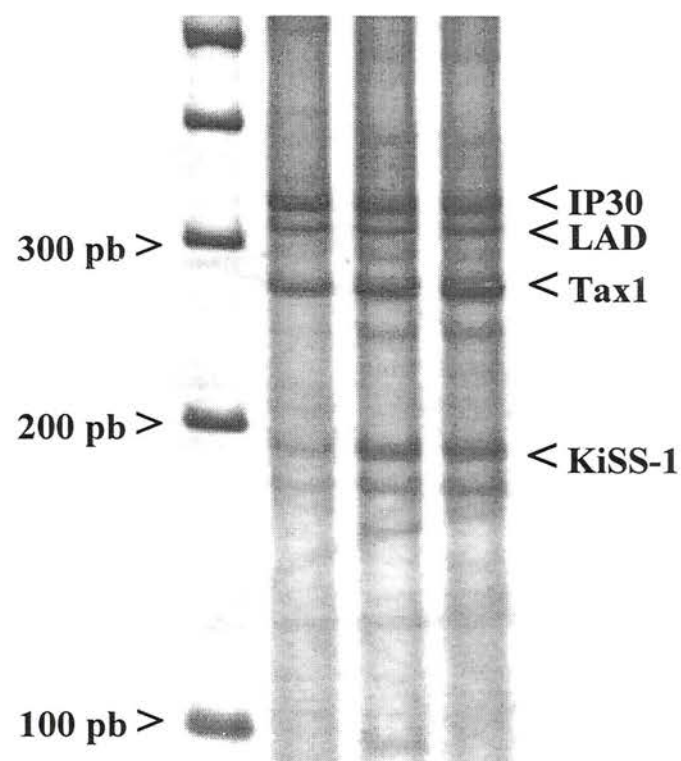


Figura 19 - Melhor resultado obtido para o sistema *multiplex* utilizando os iniciadores de IP30, Tax1, KiSS-1 e LAD. Infelizmente este resultado não foi reproduzido nas amostras populacionais.

Conclusões

- ORESTES é uma valiosa fonte para o encontro de SNPs não sinônimos;
- Apesar dos problemas inerentes a técnica, alguns passos na rotina de bioinformática podem ser aplicados para a identificação de polimorfismos reais a partir de ORESTES;
- Utilizando as seqüências do HCGP e a rotina de bioinformática do Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto foi possível sugerir 275 SNPs não sinônimos;
- Dos 275 SNPs não sinônimos identificados, 51 ou 18,5% foram validados eletronicamente no banco de dados dbSNP, os demais 81,5% constituem um *set* de novos possíveis polimorfismos humanos;
- A partir de um grupo de 22 SNPs foi possível validar 13 deles, 60%, em amostras populacionais;
- Identificamos e validamos os primeiros SNPs descritos para *Schistosoma mansoni*, demonstrando que os candidatos vacinais analisados são polimórficos;
- O *kit SNaPshot* foi implementado com sucesso, e alterações no protocolo original podem ser realizadas diminuindo em muito os custos operacionais.

Perspectivas

Fundação Antonio Prudente

Ana Maria Rodrigues Alves Kuninari
Coordenadora Pós-Graduação

Alguns pontos merecem ser ressaltados para este primeiro bloco.

Apesar da disponibilidade, agilidade e dedicação de toda a equipe de bioinformática, do Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto, envolvida na identificação dos SNPs humanos, melhorias no *pipeline* ainda podem ser implementadas. A primeira delas será a utilização do *RefSeq* como banco de âncoras, com isto estaremos triplicando ou mais, o número de polimorfismos identificados, outra atualização será elevar o valor de *Phred* para 20 aumentando ainda mais a estringência das análises.

Comparações entre os resultados obtidos utilizando somente seqüências derivadas de amostras de tecidos normais e somente seqüências derivadas de tecidos tumorais, podem ajudar a esclarecer as taxas observadas de alterações não sinônimas superiores a sinônimas, ao contrário do esperado. Uma hipótese seria que em uma célula tumoral alterações não sinônimas estariam sendo selecionadas positivamente para a progressão do tumor. Embora o HCGP tenha gerado mais de 1 milhão de seqüências, sua maioria é constituída de ORESTES derivadas de tecidos tumorais, assim somente uma análise individualizando seqüências normais e tumorais resolverá esta questão.

Pretendemos avaliar alguns dos polimorfismos não validados, desta feita investigando sua presença em tecidos tumorais a fim de averiguar a possibilidade dos mesmos serem usados como marcadores tumorais, já que existe a possibilidade destas alterações serem importantes para o desenvolvimento e progressão de alguns tipos de câncer.

A utilização dos eletroferogramas das seqüências geradas pelo CGAP poderão ajudar nas análises de identificação de mais SNPs. Ao unirmos os dois bancos de dados, CGAP e HCGP, poderemos considerar as seqüências do CGAP com sendo uma nova biblioteca para a qual possuímos os dados sobre a qualidade de cada seqüência e assim inferir sobre a qualidade dos dados gerados.

Para o sistema *SNaPshot* pretendemos ainda genotipar 10 polimorfismos em uma única reação e tentar diluir os reagentes do *kit*.

Novas tentativas de padronização da PCR em *multiplex* serão realizadas.

Artigos relacionados a tese

- SNPs: Sutis diferenças de um código – Publicado
- Identification of Single Nucleotide Polymorphisms in the Coding Region of Human Genes Using Open Reading frame ESTs (ORESTES) – Submetido
- Insights into the biology of *Schistosoma mansoni* parasite through transcriptome analysis – Trabalho em fase final de redação
- ORESTES, CGAP and the Cancer ORFeome – Trabalho em fase final de redação



SNPs:

Sutis diferenças de um código

O genoma sob um novo ponto de vista

A medida que a sequência nucleotídica do genoma humano foi sendo desvendada, uma constatação evidente foi o grande número de variações de ponto encontradas ao se comparar segmentos correspondentes do genoma (figura 1). Antes mesmo do primeiro rascunho da sequência completa ser divulgado, uma grande parcela da comunidade científica já voltava sua atenção para essas pequenas e abundantes variações dispersas por todo o código genético (figura 2). As mais comuns (ocorrem, aproximadamente, a cada 600 bases) são denominadas polimorfismos de nucleotídeos únicos ou SNPs (pronuncia-se 'S' 'N' 'Ps' ou SNiPs), e correspondem a posições onde existe uma alternância dos nucleotídeos A, C, G e T, em uma frequência alélica mínima de 1% em uma dada população (Brookes, 1999).

Os SNPs ocorrem tanto em regiões codificadoras como em não codificadoras dos genomas. Em regiões codificadoras, quando resultam em uma substituição de aminoácido na sequência protéica, são denominados não sinônimos, podendo a substituição ser conservativa ou não conservativa em função das características dos aminoácidos envolvidos na troca. Nesses casos, podem haver modificações estruturais e funcionais na proteína. Embora SNPs sinônimos não alterem a sequência protéica, eles podem modificar a estrutura e a estabilidade do RNA mensageiro,

e, conseqüentemente afetar a quantidade de proteína produzida. Além disso, SNPs podem promover: *splicing* alternativo, alterações no padrão de expressão de genes (como no caso de alterações em sequências de promotores), geração ou supressão de códons de terminação ou poliadenilação na molécula de RNA mensageiro e alteração nos códons de iniciação de tradução.

As substituições mais frequentes que ocorrem no DNA são as que

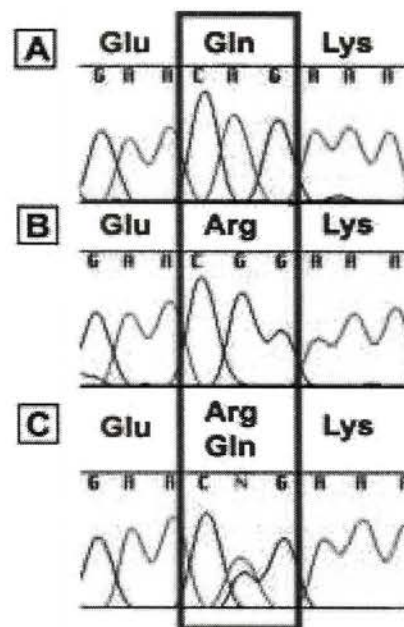


Figura 1 - Alinhamento de sequências correspondentes ao mesmo segmento genômico de três indivíduos, mostrando um SNP não sinônimo A/G, que causa a substituição de glutamina para arginina na sequência protéica.

A) Indivíduo homozigoto AA.
B) Indivíduo homozigoto GG.
C) Indivíduo heterozigoto AG

Pedro Edson Moreira Guimarães

Orientação do Prof. Dr. Emmanuel Dias Neto
Laboratório de Neurociências – LIM27
Departamento e Instituto de Psiquiatria
Universidade de São Paulo
pedson@usp.br

Maria Cristina Ramos Costa, PhD

Coordenação do Prof. Dr. Marco Antonio Zago
Centro de Terapia Celular – Hemocentro de Ribeirão Preto
Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo
mrcosta@usp.br

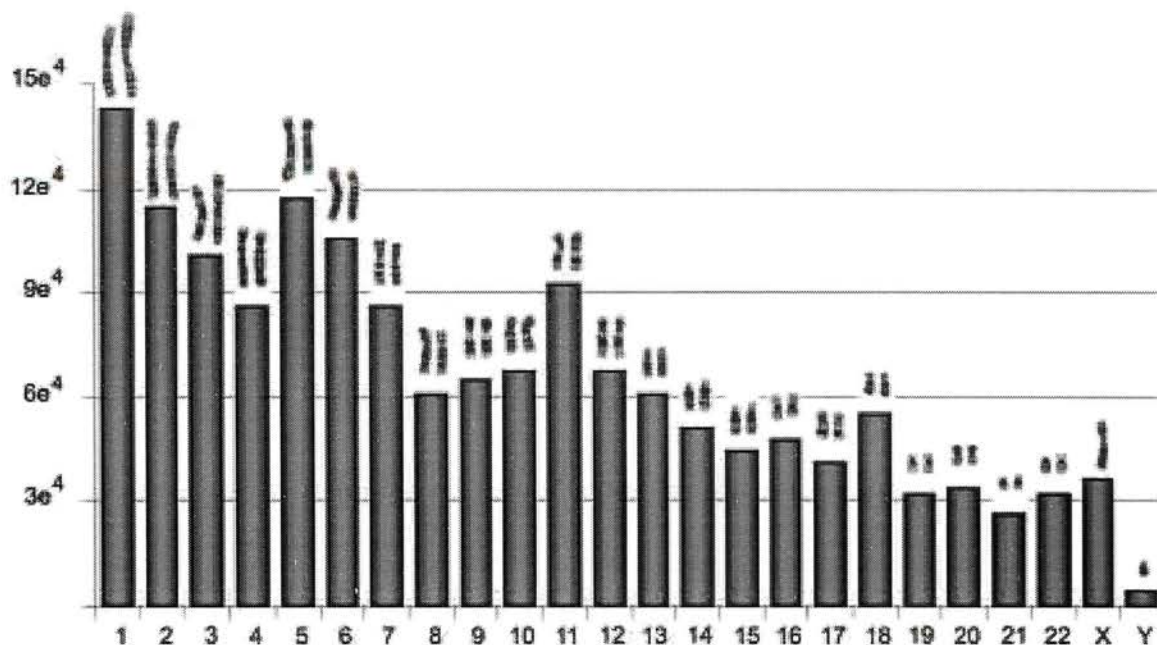


Figura 2 - Distribuição por cromossomo dos SNPs depositados no banco de dados dbSNP (<http://www.ncbi.nlm.nih.gov/SNP>). Aqui podemos observar, como mencionado no texto, que os polimorfismos humanos estão dispersos por todo o genoma

envolvem bases nitrogenadas de mesma característica estrutural, ou seja, trocas entre duas purinas (A/G ou G/A) ou duas pirimidinas (C/T ou T/C) e são denominadas transições. As transversões são substituições de uma purina por uma pirimidina ou vice-versa. Essas alterações, algumas vezes, têm origem em erros de incorporação de bases durante a replicação do DNA; em outros casos, são causadas por lesões no DNA por agentes ambientais. Caso essas mutações ocorram em células germinativas, sejam transmitidas às gerações seguintes e se fixem na população em uma frequência mínima de 1%, passam a ser denominadas de polimorfismos (Kwok & Gu, 1999).

Atualmente, os SNPs podem ser empregados nas mais diversas áreas, como medicina forense, antropologia molecular, evolução, definição de marcadores de predisposição a determinadas patologias e de prognóstico a diferentes tratamentos, genética de populações, conservação e manejo de fauna, farmacogenética, desenvolvimento de vacinas, entre outras. Como exemplo, podemos citar o mecanis-

mo de resistência a uma droga recentemente desenvolvida para o tratamento da leucemia mieloide crônica, o STI571, que é um inibidor específico da atividade tirosina quinase da proteína híbrida BCR-ABL, responsável pelo desenvolvimento dessa neoplasia. Apesar dos resultados favoráveis obtidos com essa nova droga, um grupo de pacientes mostrou-se resistente a ela. Em um estudo recente, foi demonstrado que a resistência à droga está relacionada com um polimorfismo no resíduo 315 da proteína. A substituição de treonina por isoleucina nessa posição impede a formação de uma ponte de hidrogênio essencial para a ligação do inibidor. Dessa forma, uma análise prévia desse polimorfismo em pacientes pode indicar quais devem ser tratados com o STI571 (Gorre *et al.*, 2001).

Um outro estudo recente, conduzido por um grupo de pesquisadores brasileiros, identificou um SNP não sinônimo no gene COL18A1, que eleva em duas vezes e meia o risco de se desenvolver câncer de próstata (Iughetti *et al.*, 2001). A substituição Asp104Asn na endostatina, um produto do

colágeno XVIII e potente inibidor da angiogênese, prejudica a função dessa proteína, o que explica a maior predisposição ao desenvolvimento do tumor.

Em busca de outros polimorfismos que pudessem contribuir para o desenvolvimento de abordagens farmacológicas individualizadas, ou seja, que atendessem ao perfil genético de cada paciente, foi criado, em 1999, um consórcio envolvendo onze gigantes da indústria farmacêutica e de biotecnologia, denominado *The SNP Consortium*. O projeto se estendeu até o final de 2001 e utilizou um orçamento estimado em 50 milhões de dólares, tendo sido capaz de identificar, juntamente com o Projeto Genoma Humano, 1,42 milhão de SNPs no genoma humano (Sachidanandam *et al.*, 2001).

Uma iniciativa anterior, lançada em 1996 pelo *National Cancer Institute* dos USA e denominada *Cancer Genome Anatomy Project* (CGAP), visa gerar informações e tecnologias necessárias para decifrar a anatomia molecular das células tumorais. Dentre as ferramentas desenvolvidas, algumas são volta-

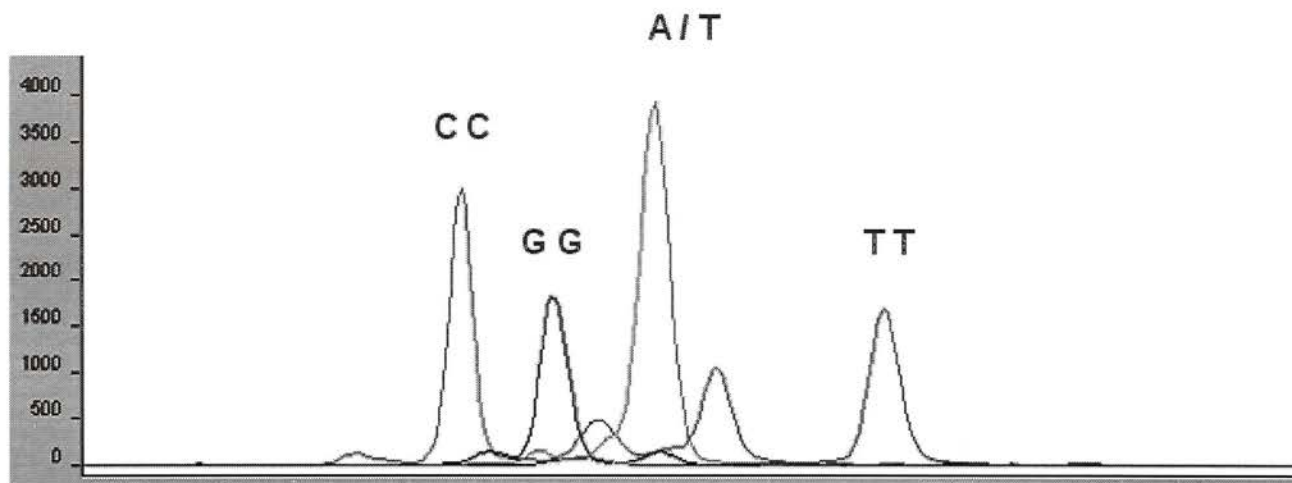


Figura 3 - Genotipagem em sistema de multiplex. Aqui um único indivíduo foi genotipado para quatro diferentes SNPs. Podemos ver que o indivíduo genotipado é homozigoto CC para o SNP1, homozigoto GG para o SNP2, heterozigoto AT para o SNP3 e homozigoto TT para o SNP4

das para a identificação e mapeamento de SNPs. Como resultado desse projeto, foram identificados, até o momento, mais de 20 mil SNPs humanos. As informações de mapeamento e caracterização de cada polimorfismo estão disponibilizadas em bancos de dados públicos (<http://gai.nci.nih.gov>, <http://www.ncbi.nlm.nih.gov/SNP>).

No Brasil, utilizando as seqüências geradas pelo Projeto Genoma

do Câncer Humano (FAPESP/LICR-HCGP), encontra-se em andamento uma iniciativa de identificação de SNPs na região codificadora de genes humanos (HCGP-SNP, http://bit.fmrp.usp.br/hcgp_snp). Os dados produzidos até o momento resultaram na identificação e validação experimental de novos SNPs humanos, que estão sendo avaliados em população normal para determinação de frequência alélica e em amostras derivadas de diferentes tumores.

É importante ressaltar que nos projetos de identificação de SNPs, principalmente aqueles em larga escala, a participação de uma equipe de bioinformática é fundamental. Graças ao desenvolvimento tecnológico de *softwares* e de equipamentos cada vez mais robustos e sofisticados, vem sendo possível gerar e analisar um grande número de dados de seqüências de DNA. Inicialmente, as abordagens de detecção de SNPs consistiam em amplificar fragmentos genômicos equivalentes, de regiões gênicas candidatas a determinado fenótipo (predisposição a uma doença neuropsiquiátrica, por exemplo), do DNA de vários indivíduos, entre grupos teste e controle, e comparar suas seqüências buscando variações. Atualmente, a maior parte dos estudos desenvolvidos visan-

do a identificação de SNPs em larga escala tem explorado as milhares de seqüências presentes nos bancos de dados, incluindo clones genômicos e, principalmente, seqüências de cDNA ou ESTs (*Expressed Sequence Tags*). Os bancos de dados de ESTs apresentam um elevado grau de redundância, e o fato de as seqüências serem derivadas de tecidos de vários indivíduos torna a utilização desses bancos muito promissora para a identificação de polimorfismos.

As ferramentas de bioinformática são utilizadas em várias etapas do estudo de polimorfismos, desde a identificação das variações até a predição do efeito delas. Outra aplicação importante é a construção e manutenção de bancos de dados e de ferramentas que podem ser acessadas pela internet. Esses bancos atuam como sedes de referência para a deposição de SNPs, sendo que no maior banco público de polimorfismos, que é mantido pelo *National Center for Biotechnology Information* (dbSNP, <http://www.ncbi.nlm.nih.gov/SNP>), já estão depositados mais de 4,1 milhões de SNPs humanos e um número considerável de polimorfismos de outras espécies.

Apesar da indiscutível contribuição da bioinformática na identificação de SNPs, a necessidade de

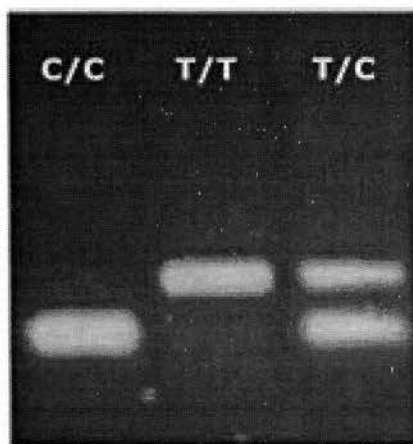


Figura 4 - Exemplo de análise de um polimorfismo por RFLP, onde a presença do nucleotídeo C cria um sítio de restrição para a enzima de restrição. Gel de agarose contendo na raia 1 um indivíduo homozigoto CC, na raia 2 um homozigoto TT e na raia 3 um heterozigoto TC (apenas o alelo C é digerido)

validação experimental dos dados e estudos de associação são indispensáveis para se avaliar o potencial informativo de cada polimorfismo. Dessa forma, o surgimento e a evolução de métodos de validação e de genotipagem de amostras para SNPs confirmados também merecem reconhecimento. Hoje, é possível genotipar vários SNPs em várias amostras em um mesmo tubo de reação; como exemplo podemos citar o sistema de multiplex desenvolvido pela *Applied Biosystems™*, que pode genotipar até 10 SNPs em uma mesma reação (figura 3). Outros métodos robustos de genotipagem podem ser citados: *chips* de DNA ou *microarrays*, cromatografia líquida de alta pressão desnaturante (DHPLC), espectrometria de massa, PCR em tempo real, mini-sequenciamento e sequenciamento de uma única base. Entretanto, em análises de menor escala, métodos como sequenciamento (figura 1) e digestão por enzimas de restrição – RFLP (*Restriction Fragment Length Polymorphism*) são rotineiramente empregados. Na figura 4, é mostrado um exemplo de SNP genotipado utilizando uma enzima de restrição, onde a presença do polimorfismo cria um sítio para a enzima, podendo os diferentes alelos ser identificados em um simples gel.

Perspectivas

Um tema ainda bastante polêmico que envolve polimorfismos de DNA é a legalidade de se patentear dados de SNPs. De um lado, há o interesse econômico de grandes empresas farmacêuticas. Um dos objetivos é o de conhecer o genoma humano, de modo a associar SNPs com resposta a drogas. De outro lado, os pesquisadores e as instituições envolvidos na geração de dados públicos. Muitas vezes, as empresas farmacêuticas utilizam os dados públicos em seus projetos, e deles retiram os resultados que serão transformados em patentes privadas sem retorno aos pes-

quisadores e às instituições responsáveis pela produção desses dados. Essa discussão ainda está longe de alcançar um ponto final, mas uma tendência que vem-se tornando consenso entre os envolvidos é a possibilidade de se patentear produtos industrializados específicos criados a partir de dados de SNPs, mas não os polimorfismos em si.

Até o momento, grande parte dos estudos tem relatado a associação entre um único SNP e uma determinada patologia, ou mesmo definido um SNP como um marcador para predisposição a uma doença. Entretanto, visando a compreensão das bases moleculares de doenças de características genéticas mais complexas, pesquisas têm sido intensificadas no sentido de tentar associar, ao fenótipo alterado, conjuntos de determinados SNPs, denominados haplótipos (Brookes, 2002). Embora esse tipo de enfoque necessite de maiores investimentos tanto tecnológicos como financeiros e envolva um número maior de amostras, ele constitui uma abordagem muito promissora nas patologias humanas mais comuns, como distúrbios neuropsiquiátricos, doenças autoimunes, doenças cardiovasculares e câncer.

Podemos vislumbrar para um futuro não muito distante a compreensão e subseqüentes tratamentos para essas patologias. Teremos uma medicina pessoal, voltada para as características genéticas de cada indivíduo. A indústria farmacêutica já busca, através de drogas mais específicas, suprir essas necessidades individuais, conseguindo, com algum sucesso, reduzir os efeitos colaterais de vários medicamentos. Testes mais acurados estão sendo desenvolvidos e será possível definir, antes mesmo do nascimento, um mapa clínico do genoma de cada pessoa, contendo informações como as patologias a que será mais susceptível e a quais tratamentos responderá melhor. Sob essa nova ótica, não é arriscado dizer que as

pessoas poderão viver mais e com melhor qualidade de vida.

Agradecimentos

Somos especialmente gratos ao Prof. Dr. Emmanuel Dias Neto pelo apoio e pela revisão do manuscrito. Agradecemos à FAPESP as bolsas de pós-doutorado de Maria Cristina Ramos Costa e a bolsa de mestrado de Pedro Edson Moreira Guimarães.

Referências

- Brookes AJ. 1999. The essence of SNPs. *Gene* **234**(2):177-186.
- Brookes AJ. 2002. 4th International meeting on Single Nucleotide Polymorphism and complex genome analysis various uses for DNA variations. *Eur J Hum Genet* **10**(2):153-155
- Gorre ME, Mohammed M, Ellwood K, Hsu N, Paquette R, Rao PN, Sawyers CL. 2001. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**(5531):876-880.
- Iughetti P, Suzuki O, Godoi PH, Alves VA, Sertie AL, Zorick T, Soares F, Camargo A, Moreira ES, di Loreto C, Moreira-Filho CA, Simpson A, Oliva G, Passos-Bueno MR. 2001. A polymorphism in endostatin, an angiogenesis inhibitor, predisposes for the development of prostatic adenocarcinoma. *Cancer Res.* **61**(20):7375-7378.
- Kwok PY, Gu Z. 1999. Single nucleotide polymorphism libraries: why and how are we building them? *Mol Med Today* **5**(12):538-543.
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409** (6822): 928-933. 

Identification of Single Nucleotide Polymorphisms in the Coding Region of Human Genes Using Tumor and Normal Tissues Derived Sequences

M.C.R. Costa^{1,2}, W.A. Silva Jr¹, A.J. Holanda^{1,5}, I. Tojal-Silva¹, K.S. Ferreira¹, R. Azevedo¹, A.J.G. Simpson³, P.E.M. Guimarães^{3,4}, E. Dias-Neto^{3,4}, M.A. Zago^{1*}.

¹Center for Cell-Based Therapy – Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil; ²Faculty of Medicine, University of Ribeirão Preto (UNAERP), Ribeirão Preto, Brazil; ³Ludwig Institute for Cancer Research, São Paulo, Brazil; ⁴Laboratory of Neurosciences (LIM27) – Department & Institute of Psychiatry, University of São Paulo, São Paulo, Brazil; ⁵Department of Physics and Mathematics – Faculty of Philosophy, Science and Letters of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil.

* Correspondence to:

Marco Antonio Zago

Centro de Terapia Celular - Hemocentro de Ribeirão Preto

Rua Tenente Catão Roxo, 2501, CEP 14051-140, Ribeirão Preto SP, Brazil.

E-mail: marazago@usp.br; Fax (55-16) 3963-9309.

ABSTRACT

SNPs are the markers of choice to build high resolution genetic maps owing to their huge number in the human genome and multiple detection approaches. The informative potential of SNPs lies on association studies to correlate different genotypes with susceptibility to develop diseases and to respond to treatment. The identification of SNPs in the coding region of genes (coding SNPs or cSNPs) may help in understanding the molecular basis of complex multifactorial diseases such as cancer, atherosclerosis, psychiatric disorders and autoimmune diseases, as well as the clinical response to specific drugs thus tailoring pharmacological treatment. In this study, we evaluate the suitability of identifying novel cSNPs from a special set of ESTs, termed ORESTES, derived from the coding region of genes. These sequences were generated from tumor and normal tissues by the FAPESP/LICR - Human Cancer Genome Project, an initiative that produced and submitted to GenBank over than 1.2 million ORESTES. An automated pipeline was built to mine cSNPs from 167,200 ORESTES that matched a panel of 7,950 full-length human genes. Applying stringent bioinformatics parameters, we identified 538 candidate cSNPs, of which 341 are non synonymous and 197 synonymous. From these, 80 (14.9%) have already been reported to dbSNP. Focusing on the non synonymous cSNPs, that may cause biological effects on protein function, we selected a sample of 22 putative cSNPs to perform experimental validation. We achieved an experimental validation rate of 60%, in agreement with the value expected for *in silico* SNP discovery from ESTs databases. For the confirmed cSNPs, we have determined the frequency in individuals from three ethnic groups.

Key Words: SNP discovery, non synonymous SNP, ORESTES, cancer, ESTs, database, bioinformatics

Databases: <http://www.ncbi.nlm.nih.gov/dbSNP> (dbSNP)

http://ctc.fmrp.usp.br/bit/snp_hcgp (HCGP-SNP)

<http://www.ncbi.nlm.nih.gov/genome/guide/human/> (Human Genome Resources at NCBI)

PARTE II

Análise de novos SNPs não sinônimos

utilizando ORESTES

Resumo

Neste projeto, nos propomos a analisar os novos SNPs não sinônimos validados a partir dos dados gerados pelo uso da técnica ORESTES em dois diferentes projetos genoma (HCGP e PGSM). Alguns SNPs selecionados foram avaliados dentro do contexto da importância destes genes dentro de questões de relevância médica. O potencial informativo dos SNPs está diretamente ligado a sua eventual associação com características de interesse. Neste sentido, em colaboração com diversos grupos de pesquisa, buscamos associar alguns dos novos SNPs humanos aqui descritos com doenças de interesse. A associação gene/doenças foi buscada levando-se em consideração a atividade biológica da proteína codificada e/ou o mapeamento genômico do gene. Deste modo, conduzimos estudos de associação entre os polimorfismos confirmados avaliando a sua frequência em indivíduos controles e em pacientes com Câncer de Cabeça e Pescoço ou com a Doença de Alzheimer.

Nossos dados preliminares, resultantes da análise do SNP no gene Ninjurina, em um grupo de 58 pacientes e 135 controles, sugerem uma associação deste polimorfismo com a Doença de Alzheimer ($p = 0,0133$, OR = 6,04).

Em amostras de pacientes com Câncer de Cabeça e Pescoço foram avaliados cinco polimorfismos e inicialmente foi encontrada uma associação significativa entre a presença do alelo variante de Tax1 e a patologia ($p = 0,0000095$, OR = 5,85). Entretanto novas análises necessitam ser feitas e um levantamento detalhado dos dados clínicos de cada paciente deve revelar novas informações.

Finalmente, a frequência dos SNPs não sinônimos validados nos genes candidatos à vacina pela OMS foi determinada em diferentes isolados de *S. mansoni*. Os alelos variantes chegaram a frequências de 42,3% (TPI) e 93,4% (GST28) nas populações de Senegal e Bahia respectivamente. Atualmente estamos avaliando a presença de SNPs em genes de *S. mansoni* como um dos critérios de exclusão ou inclusão de novos candidatos vacinais.

Summary

In this project we evaluated new non synonymous SNPs derived from ORESTES sequences from two distinct genome projects (HCGP and PGSM). some polymorphisms, that are likely to have an impact in medically relevant issues, were selected. The informative potential of SNPs is directly linked to its eventual association with characteristics of interest. In this way, in association with different research groups, we investigated the possible link of new human SNPs with diseases of interest. The association gene-disease was pursued after considering the biological activity of the protein and/or the genomic mapping of the gene to a region previously known to be associated with the disease. In this way, association studies were conducted by investigating the frequency of new human SNPs in controls and in individuals with Head & Neck tumours or with Alzheimer's Disease.

Our preliminary data, derived from the analysis of a new SNP in the gene of Ninjurin, in a group of 58 patients and 135 controls, suggested an association between this SNP and the disease ($p = 0,0133$, $OR = 6,04$).

The frequency of five SNPs was evaluated in DNA from patients with Head & Neck tumours. A possible association was found between a variant allele of Tax1 and the presence of this cancer ($p = 0,0000095$, $OR = 5,85$). Further analysis should be undertaken, including more clinical data from the patients, in order to enrich these findings.

Finally, the frequency of new non-synonymous SNPs located in vaccine-candidate genes recommended by the World Health Organisation (WHO) was determined in different isolates of *S. mansoni*. The variant alleles reached

frequencies as high as 42,3% (TPI - from Senegal isolates) and 93,4% (GST28 – from the isolate from Bahia). The presence/absence of SNPs is now being evaluated in new vaccine-candidate genes for this parasite.

Introdução

II.1.1 Importância dos SNPs em questões médicas relevantes

A cada dia, são cadastrados no NCBI de 5 a 10 artigos envolvendo o estudo de SNPs em patologias humanas (http://www.hgmp.mrc.ac.uk/embnet.news/vol6_1/pubcrawler.html); este número, que cresce cada vez mais, reflete a importância atual destas variações do nosso genoma. Como mencionado, estes polimorfismos podem não acarretar nenhum dano a célula, mas dependendo de sua localização no genoma, podem alterar ou abolir a expressão de um determinado gene ou gerar uma proteína com atividade biológica alterada ou ainda uma proteína truncada. Diante de seu potencial como marcador de uma série de importantes características biológicas, os SNPs são considerados hoje como uma das mais poderosas ferramentas para a pesquisa médica. O impacto das variações gênicas pode se dar em diversas doenças humanas, ou ainda, em uma larga série de organismos de importância médica. Diante da grande contribuição brasileira na geração de seqüências de cDNA, com uma nova tecnologia, temos a oportunidade de avaliar o impacto de uma grande quantidade de novos SNPs ainda não investigados por outros grupos, buscando o seu possível impacto em uma série de manifestações patológicas. Utilizando então os polimorfismo por nós identificados e validados (conforme descrito na Parte I desta tese) partimos para sua análise em questões médicas onde os respectivos genes estivessem envolvidos, assim, junto a nossos colaboradores foram obtidas amostras DNA para as quais seria relevante a investigação de frequência de alguns SNPs selecionados.

II.1.1.1 Câncer como uma doença genética: O impacto do Projeto Genoma Humano

O câncer é uma complexa família de doenças resultantes da aquisição e acúmulo de alterações no DNA. Múltiplas alterações genéticas são necessárias para que o fenótipo tumoral seja expresso. Estas alterações podem estar sob inúmeras formas, variando desde mutações de ponto a deleções, inserções, amplificações e translocações. Associadas a estas mudanças genômicas estão alterações nos padrões de expressão dos genes, tanto na transcrição quanto na tradução. Até recentemente, a biologia molecular dos tumores havia sido examinada para poucos genes; contudo, com a identificação de novos genes, a disponibilidade de novas técnicas e novas informações geradas dentro do Projeto Genoma Humano, ocorreu uma expansão no painel de genes que podem ser avaliados em relação ao seu papel no processo de carcinogênese (KAPRANOV et al. 2002).

A fim de identificar genes expressos em tumores humanos, o HCGP fez uma importante contribuição, gerando mais de 1 milhão de seqüências provenientes de diversas regiões topográficas, incluindo tecidos normais e tumorais.

II.1.1.1.1 Câncer de Cabeça e Pescoço no mundo

Carcinomas de células escamosas da cavidade oral, faringe e laringe são responsáveis por 270 mil mortes por ano no mundo, das quais cerca de 75% ocorrem em países em desenvolvimento (PISANI et al. 1999). É estimado que 41.000 novos casos de Câncer de Cabeça e Pescoço ocorram anualmente nos Estados Unidos (SEER 1994) e cerca de 500.000 novos casos ocorram mundialmente (PARKIN et al. 1999). Em termos de mortalidade, estes tumores representam a terceira maior causa

de morte por câncer em países em desenvolvimento, atrás somente de câncer de pulmão e câncer gástrico (PISANI et al. 1993; PARKIN et al. 1993). Tumores orais e orofaríngeais têm uma particular importância entre as minorias Afro-americanas e Hispânicas. Estes grupos possuem taxas de mortalidade, associadas a câncer oral, superiores à média nacional (EUA), sendo comparáveis ao grau de risco observado em alguns países em desenvolvimento. É notório que o aumento, tanto na incidência quanto na mortalidade por este tipo de câncer, tem sido observado em populações nas quais o tabaco não pode ser claramente identificado como o maior agente causador, especificamente em jovens adultos com câncer oral (SHEMEN et al. 1984; DEPUE 1986; DAVIS e SEVERSON 1987; SCHANTZ et al. 1988; FRANCHESCI et al. 1994).

A prevenção do câncer pela detecção e tratamento de lesões pré-malignas, é atualmente restrito a tumores da cavidade oral e cérvix (PARKIN et al. 1999). Em teoria, o câncer oral pode ser reduzido por programas de prognóstico de lesões pré-cancerosas (Anonymus 1985); contudo, a eficiência destes programas poderia ser maior se indivíduos com alto risco pudessem ser previamente identificados e analisados com uma maior frequência.

II.1.1.1.2 Gerando ORESTES de amostras de cabeça e pescoço

Devido à especificidade da expressão gênica e a fim de possuir um completo panorama dos genes expressos em um tumor em particular, é necessário explorar tecidos específicos usando material derivado de uma diversidade de situações patológicas. Em tumores de cabeça e pescoço, ORESTES foram produzidas pelo HCGP a partir de 16 diferentes preparações de mRNA. As localizações dos tumores

incluem cavidade oral, laringe, hipofaringe e tireóide com diferentes níveis de diferenciação. A distribuição das seqüências por tecido está listada na Tabela 08.

Nosso banco de dados de genes expressos em cabeça e pescoço é 30 vezes mais extenso que outros bancos de dados de cabeça e pescoço do mundo, contendo aproximadamente 180 mil ESTs derivadas exclusivamente destas regiões topográficas. Atualmente, este banco de dados está auxiliando na determinação do painel completo de genes expressos em tumores e tecidos normais de cabeça e pescoço (o chamado “Transcriptoma de Cabeça e Pescoço”), permitindo a identificação de novos genes, o estudo de expressão diferencial, a identificação de novas formas de *splicing* alternativo e a avaliação de polimorfismos/mutações nesta população de genes. A obtenção do conjunto completo de genes envolvidos na progressão de tumores de cabeça e pescoço será um marco no estudo desta doença e nosso grupo, associado a outros que também trabalham com estes tumores, participa ativamente desta iniciativa (www.lge.ibi.unicamp.br/cancerhn). As ORESTES provenientes de cabeça e pescoço já estão sendo analisadas e os primeiros resultados estão disponíveis dentro do contexto do Grupo Colaborativo de Cabeça e Pescoço.

II.1.1.1.3 SNPs em Tumores de Cabeça e Pescoço

Hoje se sabe que existem diversas características clinicamente relevantes que são determinadas ou influenciadas por componentes genéticos. Em tumores de cabeça e pescoço estas características incluem, mas não são limitadas a, predisposição ao desenvolvimento do tumor, evolução da doença, resposta clínica à quimioterapia, radioterapia e intervenção cirúrgica. Nestes tumores, os polimorfismos em genes como a Glutathione S-Transferase e P450 têm sido

Tabela 8 - Distribuição de seqüências de cabeça e pescoço por localização topográfica e histológica.

	Tumor	Normal
Cavidade oral	37.884	0
Hipofaringe	24.627	1.642
Laringe	43.297	2.788
Tireoide	48.171	11.224
Bócio tireóide	3.864	
Localização indeterminada	323	

associados com algumas destas características (PARK et al. 2000; TOPIC et al. 2000). Contudo, é certo que polimorfismos em diversos outros genes relevantes para esta doença ainda não foram investigados até o momento. Isto se dá graças à falta de conhecimento do conjunto completo de genes humanos, além da pequena quantidade de informação sobre os genes expressos nestes tumores, sendo que o primeiro trabalho no sentido de realizar um inventário sobre os genes diferencialmente expressos em Câncer de Cabeça e Pescoço foi publicado somente em abril de 2002 (AL MOUSTAFA et al.).

Neste projeto, confirmamos alguns SNPs e caracterizamos sua frequência em indivíduos normais e em pacientes com Câncer de Cabeça e Pescoço. Para esta etapa do trabalho, contamos com 271 amostras de DNA de pacientes, obtidas por colaboração com Dr. Luiz Paulo Kowalski, cirurgião chefe do Depto de Cirurgia em Cabeça e Pescoço e Otorrinolaringologia do Hospital do Câncer/Fundação Antônio Prudente.

II.1.1.2 Distúrbios psiquiátricos como doenças genéticas

Epilepsia, Doença de Alzheimer, Esquizofrenia, depressão e outras doenças do cérebro foram responsáveis por 15% de todas as patologias humanas em 1998, constituindo a maior causa de invalidez humana, acima do câncer e das doenças cardíacas (<http://www.iom.edu>). A maioria destes distúrbios neuropsiquiátricos são desordens complexas onde o componente genético está fortemente associado (LEVINSON et al. 1998; REES 2002; STROHMAN 2002; SAWA e SNYDER 2002). Para a Doença de Alzheimer, por exemplo, já é sabido que polimorfismos em alguns genes específicos, envolvidos com resposta inflamatória, promovem uma

predisposição do indivíduo ao desenvolvimento desta patologia (MCGEER e MCGEER 2001). Para esta mesma patologia, um polimorfismo situado no primeiro códon da enzima Acetil-colinesterase, envolvida no sistema colinérgico, está associado aos distúrbios cognitivos observados (MUBUMBILA et al. 2002), polimorfismos na região do gene da apolipoproteína E estão associados com a susceptibilidade e progressão de Esclerose Múltipla (SCHMIDT et al. 2002), alterações no gene DJ-1 estão associadas a formas precoces da Doença de Parkinson (BONIFATI et al. 2003), entre outros inúmeros outros achados. Está mais que evidente a participação de vários genes para o desenvolvimento e progressão das neuropatologias, entretanto, estes achados isolados são indispensáveis para a compreensão dos mecanismos envolvidos além de futuras abordagens preventivas, curativas e/ou paleativas.

Nosso grupo encontra-se hoje sediado no Laboratório de Neurociências (LIM-27), da Faculdade de Medicina da Universidade de São Paulo, sendo extremamente atraente a possibilidade de identificar e validar SNPs que possam estar associados a distúrbios neuropsiquiátricos. Apesar de não existir um projeto que utilize ORESTES para a análise de expressão gênica em distúrbios neuropsiquiátricos, no contexto do HCGP foram geradas dezenas de milhares de ESTs obtidas a partir de córtex cerebral humano. Além disto, a expressão da maioria dos genes é ubíqua, o que permite o estudo de genes importantes para o metabolismo cerebral sem que um projeto específico tenha sido desenhado para este fim.

II.1.1.3 SNPs em genes de *Schistosoma mansoni* candidatos à vacina

Como já mencionado, a esquistossomose é um dos maiores problemas de saúde humana em regiões tropicais e sub-tropicais, sendo que cerca de 600 milhões de pessoas vivem em regiões endêmicas. Tendo em vista os baixos níveis de proteção têm sido observados com os candidatos vacinais atuais e possuindo os polimorfismos validados para *S. mansoni* torna-se extremamente atraente a possibilidade de avaliar estes polimorfismos por nós encontrados em diferentes isolados geográficos do parasita. Uma análise neste sentido teria o importante papel de demonstrar a utilidade das ORESTES produzidas pelo PGSM, para a detecção de polimorfismos genéticos neste importante parasita.

Objetivos

II.2.1 Objetivo geral

Avaliar possíveis correlações entre os SNPs por nós validados em questões médicas relevantes onde seus respectivos genes pudessem estar associados.

II.2.2 Objetivos específicos

- Avaliar possíveis associações entre o SNP de Ninjurina 1 e a Doença de Alzheimer;
- Avaliar possíveis associações entre os SNPs de IP30, Tax-1 e KiSS-1 e Tumores de Cabeça e Pescoço;
- Avaliar a frequência dos SNPs validados de *S. mansoni* em amostras de diferentes isolados geográficos do parasita;
- Mapear os SNPs por nós validados em modelos proteicos tridimensionais e avaliar as possíveis alterações pertinentes a alteração de aminoácido promovida pelo SNP

Material e Métodos

II.3.1 Escolha dos SNPs a serem avaliados

Utilizando o painel de polimorfismos por nós validados, 16 SNPs humanos e 2 de *S. mansoni*, iniciamos colaborações para obter amostras onde os genes referentes a estes polimorfismos pudessem possivelmente estar associados. Segue-se então a descrição do painel de amostras obtido junto a nossos colaboradores.

II.3.2 Amostras humanas

Para as validações, foram utilizadas amostras de diversos grupos de pacientes e controles na tentativa de associar alguns dos polimorfismos a patologias onde os respectivos genes pudessem estar envolvidos.

II.3.2.1 Amostras de pacientes com Câncer de Cabeça e Pescoço

As amostras de pacientes com Câncer de Cabeça e Pescoço utilizadas neste trabalho, foram selecionadas no Departamento de Cabeça e Pescoço do Hospital do Câncer (sob responsabilidade de nosso colaborador, Dr. Luiz Paulo Kowalski). Todos os pacientes do estudo assinaram um consentimento informado, aprovado pelo Comitê de Ética do Hospital do Câncer AC Camargo, contendo as propostas gerais e natureza dos estudos desenvolvidos. Os participantes foram informados sobre sua liberdade de recusa para este estudo e que, durante todo o tempo, todas as informações pessoais seriam mantidas em absoluto sigilo.

Sempre que possível foram utilizadas amostras tumorais e de sangue periférico de cada paciente a fim de verificar se havia alguma discordância entre os

dados, refletindo possivelmente alterações genéticas surgidas durante o desenvolvimento do tumor.

II.3.2.2 Amostras de pacientes com a Doença de Alzheimer

As amostras de pacientes com a Doença de Alzheimer foram obtidas junto ao Laboratório de Neurociências do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC/FM/USP), sob a coordenação geral do Prof. Dr. Wagner F. Gattaz. Os pacientes passaram por exames clínicos e psicológicos realizados por dois psicólogos. Para a avaliação de cada caso foi preenchida uma ficha clínica, o que permitiu análises e controles posteriores. Os pacientes foram informados sobre os estudos que estavam em andamento no laboratório e convidados a participar; em caso de aceite, foi preenchido e assinado um termo de consentimento informado, previamente aprovado pelo Comitê de Ética da Faculdade de Medicina da USP, e o sangue foi coletado. Também no Laboratório de Neurociências está sendo construído um Banco de DNA onde estas amostras e todas as informações pertinentes a cada paciente estão sendo armazenadas em bancos de dados informatizados.

Nosso banco de amostras de pacientes, assim como um banco de DNA controle, pareado por sexo, idade e estadiamento, estão sendo expandidos junto a nossos colaboradores.

II.3.3 Amostras de *Schistosoma mansoni*

Junto ao Dr. David Rollison, do Museu de História Natural de Londres, foram obtidas amostras de DNA de *Schistosoma mansoni* de diversos isolados da África

(Senegal, Egito, Mali e Quênia) e do Oriente Médio (Líbano). Contamos ainda com amostras de três diferentes isolados brasileiros, oriundos dos Estados de Minas Gerais, Pernambuco e Bahia (estes dois últimos respectivamente cedidos pela Dra. Norma Lucena e pelo Dr. Mittermayer Reis, FIOCRUZ).

II.3.4 Predição de alteração da estrutura protéica

Para alguns SNPs, foram feitas tentativas de prever se a alteração de aminoácido na proteína poderia ou não modificar sua estrutura tridimensional e, conseqüentemente, sua função. Já para os SNPs de *S. mansoni*, se a alteração de aminoácido estiver localizada na superfície da molécula isto poderá alterar possíveis epítopos com prováveis efeitos no caso da utilização das respectivas proteínas em ensaios de imunoproteção. Para tanto, foi inicialmente utilizado um programa disponível na internet (<http://www.expasy.ch/swissmod/SWISS-MODEL.html>), que utiliza dados de proteínas cujas estruturas já foram resolvidas, para predizer, por homologia, a estrutura da sequência protéica de interesse. Inicialmente foi feita uma busca por similaridade de sequência dentro deste banco de estruturas, e a partir daí foi feita uma montagem da provável estrutura da proteína em questão.

Como há a necessidade de que exista no banco de dados uma proteína de sequência homóloga nem todas as nossas proteínas foram modeladas por esta estratégia.

II.3.5 Análises estatísticas

Todas as análises estatísticas foram feitas utilizando o programa Statcalc do pacote Epi Info do Centro de Controle e Prevenção de Doenças dos EUA (<http://www.cdc.gov/epiinfo>).

Resultados e Discussão

II.4.1 Avaliação dos SNPs em pacientes com Câncer de Cabeça e Pescoço

O primeiro polimorfismo a ser avaliado em alguns de nossos pacientes de cabeça e pescoço não havia sido validado nas amostras populacionais mas, mesmo assim, foram feitas algumas tentativas de identificar este SNP em pacientes com câncer de cabeça e pescoço, na expectativa de que este polimorfismo fosse específico deste grupo.

O polimorfismo em questão é uma variação de C para T no gene da CLK2 (LocusLink: 1196) uma quinase humana que gera um códon de terminação prematuro na sequência de aminoácidos desta proteína. Após tentarmos validar, sem sucesso, este polimorfismo em nossas amostras populacionais iniciamos os experimentos de validação nos pacientes dos quais se originaram as sequências que sugeriam a existência do SNP. Paralelamente foram realizados os mesmos experimentos em 32 amostras de pacientes com câncer de cabeça e pescoço provenientes do Hospital do Câncer. As amostras dos dois grupos de pacientes foram sequenciadas em conjunto. Selecionamos estas amostras tendo em vista que a sugestão da ocorrência deste SNP surgiu de bibliotecas ORESTES construídas a partir de mRNA derivado de pacientes com tumores de cabeça e pescoço.

Infelizmente, apesar de todo o zelo com que os SNPs foram selecionados para os experimentos de validação em bancada, este polimorfismo não foi encontrado em linfócitos do sangue periférico de nenhum dos pacientes dos dois grupos analisados. Não podemos no entanto, descartar a possibilidade de termos neste caso a ocorrência de uma mutação no tumor ao invés de um SNP no genoma do indivíduo.

Vale lembrar que nenhum dos SNPs não validados em nossas amostras populacionais foi encontrado nos pacientes dos quais foi extraído material para gerar as ORETES que sugeriam o polimorfismo.

Apesar do enfoque deste projeto ser em “novos SNPs” humanos, foram feitos alguns experimentos iniciais para tentar avaliar a frequência de dois SNPs, previamente descritos na literatura como associados à câncer de próstata, mas que apresentavam indícios de estarem correlacionados a outros tipos de tumor (TAVTIGIAN et al. 2001; WANG et al. 2001). Estes dois polimorfismos estão presentes no mesmo gene ELAC2/HPC2 e 32 de nossas amostras de pacientes tumorais foram genotipadas para os alelos deste gene. Todas as amostras seqüenciadas são homozigotas apresentando apenas os alelos normais.

Recentemente foi publicado um trabalho no qual foi feito o primeiro inventário dos genes que são diferencialmente expressos em Carcinomas de Células Escamosas de Cabeça e Pescoço (AL MOUSTAFA et al. 2002). Neste trabalho foi constatada uma diminuição de cerca de 50 vezes da expressão do gene Ladinina nas amostras celulares tumorais quando comparadas com as amostras de tecido normal, retiradas dos mesmos pacientes. Um dos SNPs por nós identificado e validado está justamente dentro do gene Ladinina mas foi justamente o SNP cujo iniciador apresentou problemas dentro do sistema de genotipagem em *SNaPshot* por nós padronizado. Posteriormente avaliaremos a frequência deste polimorfismo em nossas amostras.

Utilizando o sistema de genotipagem *SNaPshot* proseguimos com os experimentos de análise dos SNPs de KiSS, Tax1 e IP30 em nossas amostras de paciente com câncer de cabeça e pescoço. Uma boa parte das amostras de sangue

periférico dos pacientes já foi genotipada (Tabela 09) e atualmente estamos organizando os dados clínicos referentes a estes pacientes para uma análise mais detalhada dos resultados obtidos.

Por não termos ainda em mãos os dados completos referentes a frequência dos polimorfismos de IP30 e KiSS-1 nas amostras populacionais utilizadas (Tabela 06) análises estatísticas confiáveis somente puderam ser realizadas para o SNP de Tax1 (Tabela 09). No grupo de pacientes a frequência do alelo variante é de 15,9% enquanto no conjunto total das amostras populacionais esta frequência se reduz para 4,4%, esta diferença observada é estatisticamente significativa com $p < 1 \times 10^{-4}$ e *Odds Ratio* = 4,15 ($1,91 < OR < 9,11$). A diferença na frequência de indivíduos portadores do alelo variante (31,8% em pacientes e 7,3% no conjunto total das amostras populacionais) também é estatisticamente significativa entre os dois grupos com $p < 1 \times 10^{-5}$ e *Odds Ratio* = 5,85 ($2,46 < OR < 14,15$). Entretanto ao avaliarmos com mais cuidado os dados percebemos que a única população na qual o alelo variante foi encontrado foi a população branca. Ao tentarmos correlacionar os achados anteriores, utilizando somente esta população, perdemos as significâncias estatísticas anteriormente observadas ($p = 0,71$ e $p = 0,36$ respectivamente).

Estes resultados nos sugerem a necessidade do aumento e adequação de nossas amostras controles para pacientes com câncer de cabeça e pescoço para averiguarmos se as correlações observadas são reais ou artefactuais.

II.4.2 SNPs em pacientes com a Doença de Alzheimer

A fim de identificar SNPs em genes que pudessem estar associados à Doença de Alzheimer, iniciamos uma busca dentro do nosso painel de SNPs validados. Um

Tabela 9 - Genótipos dos pacientes com câncer de cabeça para IP30, Tax1 e KiSS-1

	IP30			Tax1			KiSS-1		
	TT	TC	CC	AA	AT	TT	TT	TC	CC
Pacientes	7	32	27	45	21	-	62	7	-
	10,6%	48,4%	40,9%	68,1%	31,8%	-	89,8%	10,2%	-
	C = 65,1% T = 34,8%			A = 84,1% T = 15,9%			T = 94,9% C = 5%		
Total	n = 66			n = 66			n = 69		
Amostras populacionais	1	3	15	138	9	2	17	1	1
	5,2%	15,7%	78,9%	92,6%	6%	1,3%	89,4%	5,2%	5,2%
	C = 86,8% T = 13,2%			A = 95,6% T = 4,4%			T = 92,1% C = 7,9%		
Total	n = 19			n = 149			n = 19		
População branca	1	3	8	38	9	2	17	1	1
	8,3%	25%	66,6%	%	%	%	89,4%	5,2%	5,2%
	C = 79,2% T = 20,8%			A = 86,8% T = 13,2%			T = 92,1% C = 7,9%		
Total	n = 12			n = 49			n = 19		

deles nos chamou especial atenção, e leva a uma troca de um Ácido Aspártico por uma Alanina na proteína Ninjurina 1. Este gene codifica para uma proteína de adesão que está envolvida com regeneração axonal após dano neuronal (ARAKI e MILBRANDT 1996). Como já mencionado, em pacientes com a Doença de Alzheimer o dano mais visível e complicador da patologia é a grande neurodegeneração, chegando a cerca de 10% de perda, por ano, de massa encefálica.

Possuindo em mãos o sistema de genotipagem para o SNP da Ninjurina 1 padronizado e um banco com 58 amostras de pacientes com Doença de Alzheimer e 135 controles, procedemos com as análises deste polimorfismo nestas amostras.

Surpreendentemente, dos 58 pacientes, 7 são homozigotos para o alelo variante e dos 135 controles somente 3 são homozigotos para o alelo variante, ou seja, 12% nos pacientes e 2,2% nos controles com $p = 0,0133$ e Odds ratio = 6,04. Embora um dos limites do intervalo de confiança tenha sido ultrapassado, $1.34 < OR < 30,83^*$, estes dados apontam para uma associação entre a homozigose do alelo variante e a Doença de Alzheimer.

Atualmente estamos aumentando nosso número de amostras, tanto de pacientes quanto de controles, para obtermos uma maior coesão dos resultados.

Em colaboração com a Dra. Maria Cristina Ramos Costa e com a aluna de mestrado Karine Sá Ferreira iremos em breve iniciar estudos funcionais com as duas variantes da proteína Ninjurina 1

II.4.3 Outras patologias onde os SNPs validados podem estar associados

A gama de colaboradores e de projetos que surgiram em decorrência do estudo dos polimorfismos por nós validados é um dos exemplos da importância do

extenso número de aplicações que pode ser dado a este tipo de informação. O próprio polimorfismo em Ninjurina 1 está atualmente sendo avaliado em três outras patologias (Paralisia Cerebral, Trauma Raquimedular e Neurite associada à Hanseníase) alguns dos resultados obtidos até o momento estão nos Anexos 2 e 3.

II.4.4 SNPs em genes de *Schistosoma mansoni* candidatos a vacina

Com a ajuda do Dr. David Rollison (Depto de Zoologia, Museu de História Natural de Londres) foram obtidas amostras de DNA de diversos isolados de *S. mansoni*, originados da África (Senegal, Egito, Mali e Kenia) e do Oriente Médio (Líbano). Contamos ainda com amostras de três diferentes isolados Brasileiros (Minas Gerais, Pernambuco e Bahia).

Para os experimentos de validação o segmento genômico contendo a possível variação foi amplificado nos diferentes isolados, clonado e sequenciado. É importante ressaltar que a PCR foi feita utilizando-se enzimas que possuíam atividade corretora a fim de evitar possíveis erros inseridos pela. Para acelerar os experimentos de validação e de determinação das frequências alélicas nos diferentes isolados, contamos com a colaboração de todos os grupos de sequenciamento envolvidos no PGSM. Foram sequenciadas no mínimo 25 colônias correspondentes a cada amplificação a fim de possuímos um panorama das frequências dos diferentes alelos encontrados. As seqüências geradas foram enviadas para a bioinformática do PGSM onde foram alinhadas e as discrepâncias de alta qualidade foram identificadas. Foram geradas até o momento mais de 300 seqüências correspondentes aos fragmentos onde estão presentes as possíveis alterações. Nas amostras sequenciadas até o momento o alelo polimorfo de TPI foi mais freqüente no isolado

de Senegal (42,3%), enquanto o alelo polimorfo de GST foi observado com uma maior frequência no isolado do Líbano (50%, Figura 20).

Estes estudos devem prosseguir, tendo implicações em aspectos importantes da biologia do parasita, tais como escape à resposta imune do hospedeiro, mecanismos de resistência a drogas e outros. Além disto, a maior investigação destes polimorfismos em um maior leque amostral de parasitas, deve fornecer novos dados sobre a entrada do parasita no continente Americano trazidos junto do tráfico de escravos a partir de diferentes regiões da África.

II.4.5 Predição de estrutura das proteínas onde foram validados SNPs

As análises de predição de estrutura foram iniciadas concomitantemente aos experimentos de validação em bancada e como já mencionado anteriormente, nem todas as proteínas são passíveis de serem modeladas pela estratégia que estamos adotando. Infelizmente, não foi possível fazer a modelagem para nenhum dos SNPs humanos que foram validados, entretanto vamos demonstrar alguns resultados obtidos para que alguns pontos possam ser discutidos.

A modelagem que é apresentada corresponde ao gene RAB1, um membro da família RAS de oncogenes. Este polimorfismo não foi validado em amostras populacionais e nem mesmo nas amostras que pacientes as quais originaram as seqüências que sugeriam a existência deste SNP. A alteração esperada seria de A para C o que levaria a uma troca de uma Lisina por uma Glutamina na seqüência de aminoácidos na posição 118 da proteína. Tanto a seqüência contendo a Lisina quanto a seqüência contendo a Glutamina foram modeladas e os resultados sobrepostos e analisados utilizando o programa *WebLab Viewer*. O resultado obtido pode ser

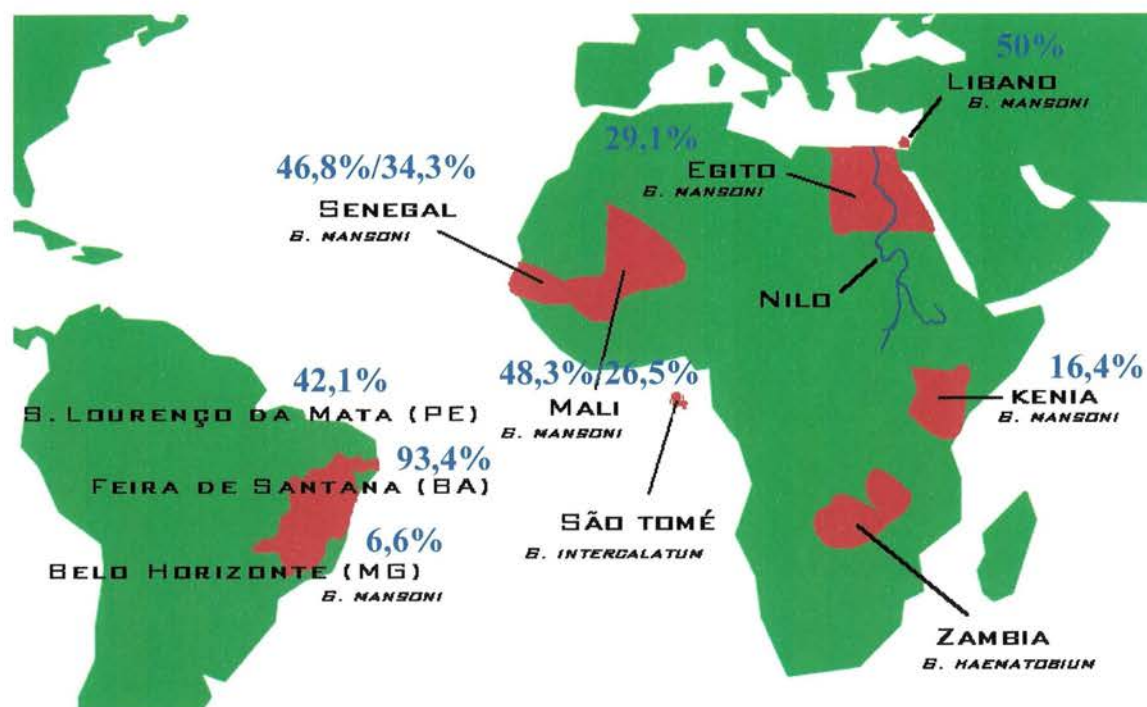


Figura 20 - Distribuição da frequência do alelo variante de GST nos diferentes isolados sequenciados. Podemos ver que a frequência mínima observada foi de 6,6%, sendo este alelo mais freqüente no isolado proveniente de Feira de Santana 93,4%.

visualizado na Figura 21. Pode-se constatar que não há nenhuma modificação na estrutura da proteína que contém a Glutamina. O que podemos concluir é que, ou não há realmente nenhuma modificação estrutural inserida na proteína, pela alteração do aminoácido, ou o programa de modelagem utilizado não é sensível, o bastante, para detectar alterações tão sutis. O que pode ser feito para contornar este problema é tentar utilizar programas de modelagens mais acurados, o que não caberia dentro do contexto desta tese.

Ao avaliar os polimorfismos validados nos genes de *S. mansoni* candidatos à vacina em modelos tridimensionais, encontramos alguns resultados interessantes já que para as duas proteínas os aminoácidos alterados encontram-se na superfície das moléculas (Figura 22) o que poderia alterar possíveis epitopos comprometendo a eficácia das vacinas produzidas já que os parasitas que possuíssem as alterações poderiam escapar do sistema imune do hospedeiro.

Fundação Antonio Prudente
Ina Maria Rodrigues Alves Kunhert
Coordenadora Pós-Graduação

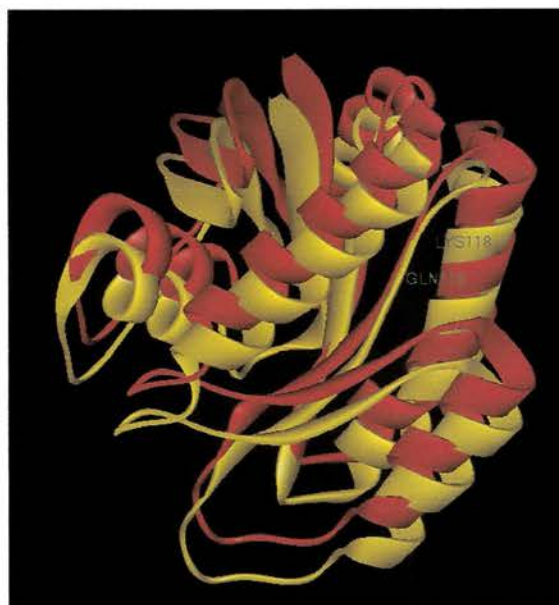


Figura 21 - Estruturas preditas pelo SWISS-MODEL da proteína RAB1 com e sem o SNP. Podemos ver que não existe nenhuma modificação aparente na estrutura das duas proteínas. Em amarelo podemos ver a proteína “selvagem” (Lys118) e em vermelho a proteína com a modificação de aminoácido Lys118Gln decorrente do SNP.

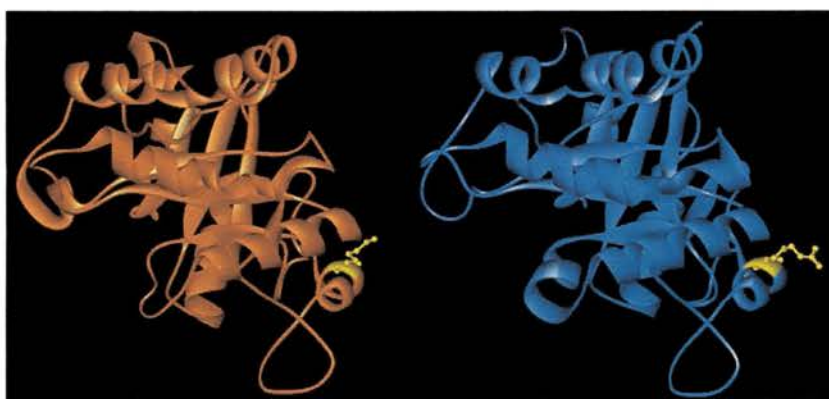


Figura 22 - Alteração de aminoácido promovida pela troca de G>A (ou C>T na fita complementar) na sequência do gene TIM de *Schistosoma mansoni*. Notar que pela predição do SWISS-MODEL o aminoácido se encontra na superfície da molécula podendo modular o potencial antigênico desta proteína.

Conclusões

- A análise do polimorfismo de Ninjurina 1 em pacientes portadores da Doença de Alzheimer sugere uma associação significativa ($p = 0,0133$, $OR = 6,04$), entre o alelo variante em homozigose e a patologia.

- Uma forte associação entre alelo variante de Tax1 e pacientes com Câncer de Cabeça e Pescoço foi encontrada ($p = 0,0000095$, $OR = 5,85$) e posteriormente será confirmada com novas amostras.

- Os alelos variantes de ELAC2/HPC2 não foram encontrados após a avaliação de 64 alelos de pacientes com Câncer de Cabeça e Pescoço.

- Até o momento não foi encontrada nenhuma associação entre os SNPs de IP30 e KiSS-1 em nossas amostras de pacientes com Câncer de Cabeça e Pescoço

- Os polimorfismos validados para *S. mansoni* se encontram localizados em regiões expostas das respectivas proteínas e eles estão presentes em todos os isolados analisados deste parasita. Atualmente estamos avaliando a presença de SNPs em genes de *S. mansoni* como um dos critérios de exclusão ou inclusão de novos candidatos vacinais.

Perspectivas

Embora o projeto de mestrado esteja em sua fase final podemos desde já imaginar futuras possibilidades para a continuidade deste estudo. Será muito interessante avaliar o comportamento e distribuição dos polimorfismos humanos, por nós descritos, em outros animais principalmente em outros primatas. A partir de alguns dos resultados encontrados poderemos compreender melhor o papel de cada um destes polimorfismos dentro de nosso próprio genoma. Para tanto estamos desde já estruturando uma pequena coleção de DNA de espécies animais que já conta com alguns primatas, outras ordens de mamíferos e exemplares de aves, répteis, peixes e invertebrata. Não é, a princípio, nossa intenção criar um grande banco de DNA de amostras animais mas sim possuímos espécies chaves para alguns estudos que desenvolvemos, como o de análise de polimorfismos.

Para alguns dos polimorfismos validados pretendemos também iniciar estudos funcionais com as proteínas envolvidas, como no caso da Ninjurina 1. É extremamente atraente a possibilidade de avaliar a expressão dos diferentes alelos em sistemas controlados (culturas celulares) e em diferentes órgãos, tanto em estado normal quanto patológico. Nosso laboratório já domina técnicas como o cultivo de linhagens neuronais e dosagem de proteínas em soro, estas metodologias serão de grande ajuda para os próximos passos a serem trabalhados.

Caso a associação de Ninjurina 1 e Alzheimer seja realmente confirmada com o aumento do número de casos e controles, será interessante realizar estudos de desequilíbrio de ligação do polimorfismo avaliado em relação aos demais existentes neste gene. Esta estratégia teria como finalidade identificar possíveis haplótipos

ligados a esta patologia e tentar definir se o polimorfismo em questão é um dos fatores causais da DA.

Estudos envolvendo o polimorfismo em Ninjurina 1 estão sendo conduzidos por colaboradores em pacientes portadores de hanseníase que desenvolveram lesão neuronal periférica, e os resultados iniciais corroboram a hipótese de que a homozigose do alelo variante pode estar associada a uma perda ou déficit neuronal em patologias onde este tipo de dano ocorre.

Será muito interessante avaliarmos os dados clínicos disponíveis referentes aos pacientes com câncer de cabeça e pescoço. Um aumento e adequação das amostras controles também será necessário para avaliarmos com segurança a possível associação existente com o polimorfismo de Tax1.

A re-síntese do iniciador do gene Ladinina permitirá a avaliação do polimorfismo dentro de nossas amostras de pacientes com câncer de cabeça e pescoço.

Futuras colaborações permitirão que análises mais detalhadas possam ser realizadas sobre influência do SNP dentro da estrutura tridimensional da proteína em questão.

Artigos relacionados a tese

- Association of a new polymorphism in Alox12 gene with Bipolar Disorder – European Archives of Psychiatry and Clinical Neurosciences, 2003, 253(1): 40-43
- In the genetics way of spinal cord injury evolution – Trabalho em fase final de redação
- Identification of SNPs in *Schistosoma mansoni* and evaluation of frequency in isolates from Middle East, Africa and South-America – Em preparação
- In the way of cerebral palsy evolution genetics basis – Em preparação

ORIGINAL PAPER

Cintia Fridman · Éliada P. B. Ojopi · Sheila P. Gregório · Elisa H. Ikenaga ·
Doris H. Moreno · Frederico N. Demetrio · Pedro E. M. Guimarães ·
Homero P. Vallada · Wagner F. Gattaz · Emmanuel Dias Neto

Association of a new polymorphism in ALOX12 gene with bipolar disorder*

Received: 19 September 2002 / Accepted: 24 September 2002

Abstract Bipolar disorder (BPD) is characterised by episodes of excitement interspersed with periods of depression. The role of genetic factors in BPD is indicated by studies in monozygotic twins showing 40–70% of concordance. Studies using genetic markers showed linkage of genes for affective disorders in different chromosome regions, emphasising the polygenic and multifactorial traits. The main goal of our research is to search non-synonymous SNPs (those that result in modifications in protein sequence) in genes that can be associated with psychiatric diseases as suggested by genomic mapping and/or by physiological function of the protein. Using DNA sequencing we could confirm a new non-synonymous SNP in the conservative domain of the ALOX12 gene (17p13.1), suggested by EST alignment. This SNP is an alteration from G to A that leads to a change of an arginine (A) to a glutamine in one of the most important domains of the protein. This SNP was evaluated by DNA sequencing in 182 patients with BPD

and 160 control individuals. An increased presence of allele A among patients (60% in controls and 73.1% in cases; $\chi^2=6.581$, $P=0.010$; $OR=1.8095$, 95% $CI=1.1477-2.853$) was found, suggesting an association of this polymorphism with the BPD in this Brazilian sample.

Key words ALOX12 · SNP · bipolar disorder · polymorphism · lithium

Introduction

The heterogeneity of mood disorders suggests that different combinations of alleles in different families can produce the deleterious phenotype (Avissar et al. 2002). The role of genetic factors in bipolar disorder (BPD) is established, as indicated by studies in which monozygotic twins showing a concordance between 40 and 70% (Craddock and Jones, 1999) as compared with a prevalence of 1–2% in the general population.

Lithium salts are the most effective long-term preventive treatment for BPD. It was recently proposed that the neuro-protective propriety of lithium against glutamate excitotoxicity results of its capacity of inhibiting GSK-3 β expression, blocking the facilitation of apoptosis produced by this enzyme (Chuang et al. 2002; Kozlovsky et al. 2002; Li et al. 2002). The lipoxygenase pathway also plays an important role in apoptosis (Tang et al. 1996) since the glutathione (GSH) depletion, caused by glutamate toxicity, triggers ALOX12 (arachidonate 12-lipoxygenase) activation, which leads to the production of peroxides and the influx of Ca²⁺, resulting in cell death (Li et al. 1997) (Fig. 1).

Studies using genetic markers have claimed linkage of genes for affective disorders in different chromosome regions such as 4p (Blackwood et al. 1996), 5q (Coon et al. 1993), 6 (Stancer et al. 1988), 11p (Egeland et al. 1987), 12q (Ewald et al. 1998), 13, 15 (Risch and Botstein, 1996), 18 (Berrettini et al. 1994), 21q (Aita et al. 1999), 22q (Kelsoe et al. 2001) and Xq (Baron et al. 1994), em-

C. Fridman · É. P. B. Ojopi · S. P. Gregório · E. H. Ikenaga ·
P. E. M. Guimarães · H. P. Vallada · W. F. Gattaz · E. Dias Neto
Laboratory of Neuroscience (LIM-27)
Department and Institute of Psychiatry
Faculty of Medicine
University of São Paulo
São Paulo, Brazil

D. H. Moreno · F. N. Demetrio
GRUDA, Institute of Psychiatry
Faculty of Medicine
University of São Paulo
São Paulo, Brazil

Cintia Fridman (✉)
Departamento e Instituto de Psiquiatria, FMUSP
R. Dr. Ovídio Pires de Campos, s/n, 3º andar (LIM27)
CEP: 05403-010, São Paulo, SP, Brasil
Tel.: +55-11/30 69-72 67
Fax: +55-11/30 62-48 51
E-Mail: cintiafri@hotmail.com

* Financial Support: Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), Conselho Nacional de Pesquisas (CNPq) and Associação Beneficente Alzira Denise Hertzog da Silva (ABADHS).

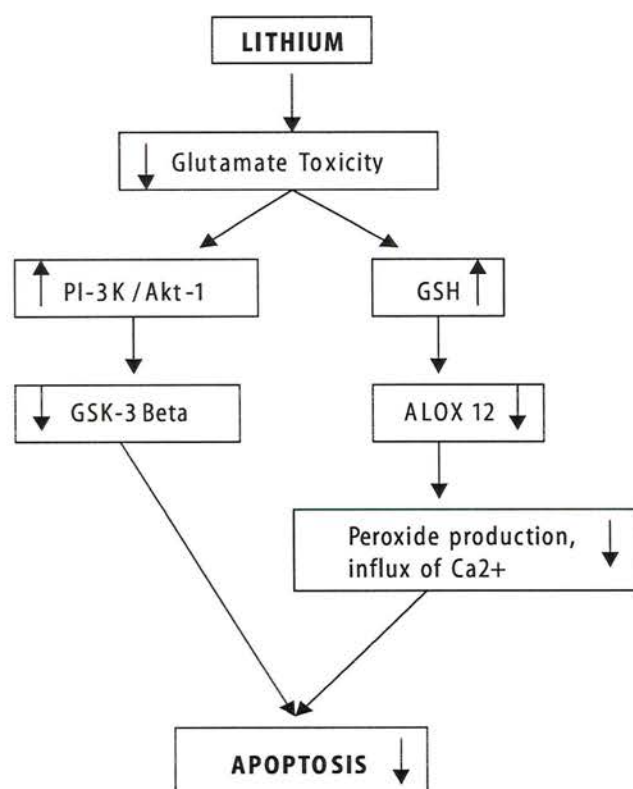


Fig. 1 A hypothetical neuro-protective action of lithium. Lithium intake protects against glutamate toxicity, which activates the PI-3K/Akt-1 expression, leading to an inactivation of GSK-3Beta, and blocking the apoptosis process. Lithium also avoids GSH depletion, which down-regulates ALOX12 expression reducing the levels of reactive oxygen species in the cell and a consequent decrease in apoptosis

phasing the polygenic and multifactorial traits of the disease.

Single nucleotide polymorphisms (SNPs) are the most frequent variations found in the human genome and can be defined as punctual regions of DNA where nucleotides can vary among individuals. SNPs have become very popular in the last few years as one of the most informative genetic markers used in the research of complex diseases. According to theoretical models, if genotypes of patients and controls are compared, a consistent association with affected individuals may be observed for certain alleles.

The main goal of our research is to search non-synonymous SNPs (those that result in modifications in protein sequence) in genes that can be associated with psychiatric diseases by mapping or by physiological

function of the protein. While investigating putative polymorphisms in the ALOX12 gene we could confirm a new non-synonymous SNP and associate it with BPD in a sample of Brazilian patients.

Material and methods

Sample

Patients with bipolar disorder selected for this study were recruited at the Institute of Psychiatry, Hospital das Clínicas, FMUSP, Brazil. Diagnosis was made according to DSM-III-R criteria (APA, 1987). Control individuals were selected among accompanying persons (not genetically related with patients) and people without psychiatric symptoms attending the same hospital. All patients and controls signed an informed consent form, which was previously approved by the ethics committee of the Faculty of Medicine, University of São Paulo. We deliberately decided to include a higher proportion of elderly people among the controls, in order to be certain of selecting patients that, even at older age, have not developed any psychiatric disorder. A total of 182 BPD patients (60 males and 122 females) and 160 controls (50 males and 110 females) were included in the present investigation. The mean age was 44.33 ± 14.33 (ranging from 18 to 83 years old) for patients and 65.35 ± 18.59 (ranging from 20 to 98 years old) for controls (Table 1).

Polymorphism search

The complete cDNA sequence of the ALOX12 gene was used in the BLAST analysis against a database of human Expressed Sequence Tags (ESTs) available in the GenBank (www.ncbi.nlm.nih.gov/dbEST). Putative polymorphic sites were selected for further analysis only when confirmed by ESTs from different cDNA libraries and when an amino acid alteration resulted (non-synonymous SNP).

DNA extraction, amplification and analysis

DNA was extracted from peripheral blood leukocytes of patients and controls, according standard phenol-chloroform protocols. A fragment of 244 nt of the ALOX12 gene was amplified by PCR in a total volume of 10 μ l, containing 5 pmol of each primer, 250 nM dNTPs, 100 ng template genomic DNA with an initial denaturation for 4 min at 95 °C. Thirty-eight cycles were run, with denaturation for 45 s at 95 °C, annealing for 30 s at 60 °C and extension for 30 s at 72 °C. The final extension was for 5 min at 72 °C. The primers used for PCR were ALOX12F (5'-CTGGCAGGATGATGAGTTGTT-3') and ALOX12R (5'-AGTACTGCAGCCTTCCTCTG-3'). The resulting PCR products were inspected in 1% ethidium bromide stained agarose gels and sequenced in an ABI3100 DNA sequencer (Applied Biosystems, Foster City, USA), using standard protocols.

Statistical Analysis

The magnitude of association between the alleles and the occurrence of BPD was measured by the odds ratio (OR) and respective 95% confidence interval (CI), together with the χ^2 test using the level of confidence of $\alpha = 0.05$.

Table 1 Features of the sample

	Males	Females	Mean age	Freq of G/G	Freq of G/A	Freq of A/A
BP Patients N = 182	60	122	44.3 $\pm$ 14.3	26.9% (49/182)	46.2% (84/182)	26.9% (49/182)
Controls N = 160	50	110	65.3 $\pm$ 18.6	40.0% (64/160)	37.5% (60/160)	22.5% (36/160)

Results

Our database search using the ALOX12 mRNA sequence, against a set of 4.9 million human ESTs, showed four ESTs covering a putative polymorphic nucleotide. These four ESTs were derived from four different cDNA libraries and the alleles A and G were present in the same proportion (2 ESTs for each). After primer design and PCR amplification using a panel of DNA samples, we were able to confirm this SNP by DNA sequencing. The alteration is located in the nucleotide 815 of the 6th exon of ALOX12, and results in a substitution of an arginine to a glutamine in the aminoacid 261 (R261Q).

DNA sequencing of 684 alleles has consistently demonstrated an increased presence of allele A among cases (60 % in controls and 73.1 % in cases; $\chi^2=6.581$, $P=0.010$; OR=1.8095, 95 % CI=1.1477–2.853). When comparing the frequency of genotypes G/A and G/G we found the best association between the heterozygous genotype and bipolar patients ($\chi^2=5.683$, $P=0.017$; OR=1.8286, 95 % CI=1.1111–3.0094) (Table 1). Nevertheless no significant association was found when the number of G/A to A/A individuals was compared (data not show).

Discussion

The analysis of ESTs obtained by us during the Human Cancer Genome Project (Dias Neto et al. 2000), and other ESTs available in the GenBank (dbEST), suggested the existence of a new non-synonymous SNP in the conserved domain of the ALOX12 gene mapped at 17p13.1 (ALOX12, OMIM *152391). This polymorphism results in the substitution of an arginine to a glutamine at the aminoacid 261 (R261Q) of the protein. The variant aminoacid is part of the lipoxygenase domain, one of the most important and conserved regions of this protein.

Analysis of protein sequences in the GenBank shows that this region is highly conserved among mammals. A comparison of protein sequences of human, mouse, rat, bovine, rabbit and pig shows that few aminoacid positions were altered in this region. The aminoacid change at the position observed in our study (R/Q) can be seen in other species of mammals. In fact, the presence of a glutamine at this position seems to be ancestral, as it is

a consensus among the majority of these mammals. However the existence of an arginine in this same position can only be observed in humans and bovines (Fig. 2). Thus, a fixation of the allele coding for an arginine in this position seems to be a recent event in the speciation of mammals.

ALOX12 gene belongs to the lipoxygenase family, is involved in the leukotrienes pathway and is the predominant lipoxygenase expressed in the brain (Bendani et al. 1995; Brash, 1999). This gene is also able to sensitise the nuclear membrane, which is more susceptible to the release of arachidonic acid (a sub-product of phospholipase A2, an important signal transducer in nervous cells). In search of a regulatory molecule that links cPLA2 and sPLA2-IIA, Murakami et al. (2000) found the 12/15-LOX isozyme. cPLA2 activation leads to production of 12/15LOX metabolites, which sensitise the nuclear membrane to be more susceptible to the sPLA2-IIA-mediated arachidonic acid release (Murakami et al. 2000). Lipoxygenase reactions may initiate the synthesis of signalling molecules or be involved in the induction of structural or metabolic changes in the cell (Brash, 1999). Expression analysis of ALOX12 showed higher levels of this gene in the mammary gland, ovary, prostate, brain, lung and heart (www.ncbi.nlm.nih.gov/Unigene).

Rapoport and Bosetti (2002) demonstrated that feeding rats with lithium reduced arachidonic acid turnover within the brain, corresponding to a down-regulation of gene expression and enzyme activity of cytosolic phospholipase A2 (cPLA2), an enzyme that selectively liberates arachidonic acid. Lithium also reduced the brain protein level and activity of cyclooxygenase 2, as well as the brain concentration of prostaglandin E[2], an arachidonate metabolite produced via cyclooxygenase 2. Their results raised the hypothesis that lithium and antimanic anticonvulsants act by targeting parts of the „arachidonic acid cascade,” which may be functionally hyperactive in mania. Thus, drugs that target enzymes in this cascade might be candidates for treating mania.

Our results suggest that, in our sample, the presence of the A allele is correlated with BPD. However, we should also consider that this possible association with BPD might in fact be due to a haplotype involving this ALOX12 polymorphism, rather than the R261Q polymorphism itself.

Functional studies are necessary to evaluate the en-

Human - <i>Homo sapiens</i> (M58704)	247 LPSRLVLPSPGMEEL <u>RA</u> QLE 265
Mouse - <i>Mus musculus</i> (U04334)	247 LPSRLVLPSPGMEEL <u>QA</u> QLE 265
Bovine - <i>Bos taurus</i> (Y08829)	145 LPTRLVLPSPGMEEL <u>RA</u> QLE 163
Rat - <i>Rattus norvegicus</i> (L06040)	248 LPARLVFPSPGMEEL <u>QA</u> QLN 266
Rabbit - <i>Oryctolagus cuniculus</i> (Z97654)	248 LPARLVFPSPGMEEL <u>QA</u> QLE 266
Pig - <i>Sus scrofa</i> (D10621)	248 LPARL <u>K</u> FPPGMEEL <u>QA</u> QLE 266

Fig. 2 Alignment of a region of ALOX12 protein from diverse mammals. Sequences were obtained from the GenBank (accession codes are given in parenthesis) and manually aligned. Aminoacid coordinates are given at the beginning and at the end of the sequences. The amino acids corresponding to the polymorphic site observed in this study are underlined. Other variant positions, with respect to the human sequence, are in bold

zymatic activity of ALOX12 in heterozygous and homozygous cells, confirming a link of this polymorphism and altered enzyme activity.

■ **GenBank/EMBL accession number.** ALOX12-arachidonate 12-lipoxygenase, NM_000697

■ **Acknowledgements** We thank Paulo Alberto Otto and Maria do Rosário for helpful considerations in the statistical analysis. The authors wish to thank Applied Biosystems do Brasil for their continued support during the course of the present study.

References

1. Aita VM, Liu J, Knowles JA, Terwilliger JD, Baltazar R, Grunn A, Loth JE, Kanyas K, Lerer B, Endicott J, Wang Z, Penchaszadeh G, Gilliam TC, Baron M (1999) A comprehensive linkage analysis of chromosome 21q22 supports prior evidence for a putative bipolar affective disorder locus. *Am J Hum Genet* 64:210–217
2. American Psychiatry Association (1987) *Diagnostic and Statistical Manual of Mental Disorder*. 3rd Edition. Revised (DSM-III-R). Washington DC, American Psychiatry Association
3. Avissar S, Schreiber G (2002) Towards molecular diagnostics of mood disorders in psychiatry. *Trends Mol Med* 8:294–300
4. Baron M, Straub RE, Lehner T, Endicott J, Ott J, Gilliam TC, Lerer B (1994) Bipolar disorder and linkage to Xq28. *Nature Genet* 7:461
5. Bendani MK, Palluy O, Cook-Moreau J, Beneytout JL, Rigaud M, Vallat J-M (1995) Localization of 12-lipoxygenase mRNA in cultured oligodendrocytes and astrocytes by in situ reverse transcriptase and polymerase chain reaction. *Neurosci Lett* 189:159–162
6. Berrettini WH, Ferraro TN, Goldin LR, Weeks DE, Detera-Wadleigh S, Nurnberger JI Jr, Gershon ES (1994) Chromosome 18 DNA markers and manic-depressive illness: evidence for a susceptibility gene. *Proc Natl Acad Sci USA* 91:5918–5921
7. Blackwood DH, He L, Morris SW, McLean A, Whitton C, Thomson M, Walker MT, Woodburn K, Sharp CM, Wright AF, Shibasaki Y, St Clair DM, Porteous DJ, Muir WJ (1996) A locus for bipolar affective disorder on chromosome 4p. *Nature Genet* 12:427–430
8. Brash AR (1999) Lipoxygenases: occurrence, functions, catalysis, and acquisition of substrate. *J Biol Chem* 274:23679–23682
9. Chuang DM, Chen RW, Chalecka-Franaszek E, Ren M, Hashimoto R, Senatorov V, Kanai H, Hough C, Hiroi T, Leeds P (2002) Neuroprotective effects of lithium in cultured cells and animal models of diseases. *Bipolar Disord* 4:129–136
10. Coon H, Jensen S, Hoff M, Holik J, Plaetke R, Reimherr F, Wender P, Leppert M, Byerley W (1993) A genome-wide search for genes predisposing to manic depression, assuming autosomal dominant inheritance. *Am J Hum Genet* 52:1234–1249
11. Craddock N, Jones I (1999) Genetics of bipolar disorder. *J Med Genet* 36:585–594
12. Dias Neto E, Corrêa RG, Verjovski-Almeida S, Briones MRS, Nagai MA, Silva Jr W, et al. (2002) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci USA* 97:3491–3496
13. Egeland JA, Gerhard DS, Pauls DL, Sussex JN, Kidd KK, Allen CR, Hostetter AM, Houseman DE (1987) Bipolar affective disorders linked to DNA markers on chromosome 11. *Nature* 325:783–787
14. Ewald H, Degn B, Mors O, Kruse TA (1998) Significant linkage between bipolar affective disorder and chromosome 12q24. *Psychiat Genet* 8:131–140
15. Kelsoe JR, Spence MA, Loetscher E, Foguet M, Sadovnick AD, Remick RA, Flodman P, Khristich J, Mroczkowski-Parker Z, Brown JL, Masser D, Ungerleider S, Rapaport MH, Wishart WL, Luebbert H (2001) A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. *Proc Natl Acad Sci USA* 98:585–590
16. Kozlovsky N, Belmaker RH, Agam G (2002) GSK-3 and the neurodevelopmental hypothesis of schizophrenia. *Eur Neuropsychopharmacol* 12:13–25
17. Li X, Bijur GN, Jope RS (2002) Glycogen synthase kinase-3 β , mood stabilizers, and neuroprotection. *Bipolar Disord* 4:137–144
18. Li Y, Maher P, Schubert D (1997) A role for 12-lipoxygenase in nerve cell death caused by glutathione depletion. *Neuron* 19:453–463
19. Murakami M, Nakatani Y, Kuwata H, Kudo I (2000) Cellular components that functionally interact with signaling phospholipase A2s. *Biochim Biophys Acta* 1488:159–166
20. Rapoport SI, Bosetti F (2002) Do lithium and anticonvulsants target the brain arachidonic acid cascade in bipolar disorder? *Arch Gen Psychiatry* 59:592–596
21. Risch N, Botstein D (1996) A manic depressive history. *Nature Genet* 12:351–352
22. Stancer HC, Weitkamp LR, Persad E, Flood C, Jorna T, Guttormsen AS, Yagnow RL (1998) Confirmation of the relationship of HLA (chromosome 6) genes to depression and manic depression. II. The Ontario follow-up and analysis of 117 kindreds. *Ann Hum Genet* 52(Pt 4):279–298
23. Tang DG, Chen YQ, Honn KV (1996) Arachidonate lipoxygenases as essential regulators of cell survival and apoptosis. *Proc Natl Acad Sci USA* 93:5241–5246

Referências bibliográficas

Al Moustafa AE, Alaoui-Jamali MA, Batist G, et al. Identification of genes associated with head and neck carcinogenesis by cDNA microarray comparison between matched primary normal epithelial and squamous carcinoma cells. **Oncogene** 2002; 21:2634-40.

[Anonymus]. Control of cancer of the mouth in developing countries. **Bull World Health Organ** 1985; 63:27-42.

Araki T, Milbrandt J. Ninjurin, a novel adhesion molecule, is induced by nerve injury and promotes axonal growth. **Neuron** 1996; 17:353-61.

Bailey JA, Gu Z, Clark RA, et al. Recent segmental duplications in the human genome. **Science** 2002; 297:1003-7.

Bansal A, Van Den Boom D, Kammerer S, et al. Association testing by DNA pooling: an effective initial screen. **Proc Natl Acad Sci U S A**. 2002; 99:16871-4.

Bergquist R, Al-Sherbiny M, Barakat R, Olds R. Blueprint for schistosomiasis vaccine development. **Acta Trop** 2002; 82:183-92.

Blight AR. Miracles and molecules-progress in spinal cord repair. **Nat Neurosci** 2002; 5 Suppl:1051-4.

Bonifati V, Rizzu P, Van Baren MJ, et al. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. **Science** 2003; 299:256-9.

Boyer JC, Bebenek K, Kunkel TA. Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates. **Proc Natl Acad Sci U S A** 1992; 89:6919-23.

Brookes AJ. The essence of SNPs. **Gene** 1999; 234:177-86.

- Brookes AJ, Lehvaslaiho H, Siegfried M, et al. HGBASE a database of SNPs and other variations in and around human genes **Nucleic Acids Res** 2000; 28:356-60.
- Camargo AA, Samaia HP, Dias-Neto E, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci U S A** 2001; 98:12103-8.
- Chakravarti A. To a future of genetic medicine. **Nature** 2001; 409:822-3.
- Chang H, Fujita T. PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. **Biochem Biophys Res Commun** 2001; 287:288-91.
- Ching A, Caldwell KS, Jung M, et al. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genet** 2002; 3:19.
- Chitsulo L, Engels D, Montresor A, Savioli L. The global status of schistosomiasis and its control. **Acta Trop** 2000; 77:41-51.
- Couzin J. Human genome: HapMap launched with pledges of \$100 million. **Science** 2002; 298:941-2.
- Davis S, Severson RK. Increasing incidence of cancer of the tongue in the United States among young adults. **Lancet** 1987; II:910-1.
- Depue RH. Rising mortality from cancer of the tongue in young white males [Letter]. **N Engl J Med** 1986; 315:647.
- Dias-Neto E, Corrêa RG, Verjovski-Almeida S, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci U S A** 2000; 97:3491-6.

- Engels D, Chitsulo L, Montresor A, Savioli L. The global epidemiological situation of schistosomiasis and new approaches to control and research. **Acta Trop** 2002; 82:139-46.
- Ewing B, Green P. Base-calling of automated sequencer traces using phred. **Genome Res** 1998; 8:186-94.
- Franchesci S, Levi F, Luchini F, et al. Trends in cancer mortality in young adults in Europe, 1955-1989. **Eur J Cancer** 1994; 30A:2096-118.
- Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. **Science** 2002; 296:2225-9.
- Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH. Multiplex PCR: critical parameters and step-by-step protocol. **Biotechniques** 1997; 23:504-11.
- Kapranov P, Cawley SE, Drenkow J, et al. Large-scale transcriptional activity in chromosomes 21 and 22. **Science** 2002; 296:916-9.
- Kim T, Mudry RA Jr, Rexrode CA 2nd, Pathak VK. Retroviral mutation rates and A-to-G hypermutations during different stages of retroviral replication. **J Virol** 1996; 70:7594-602.
- Kruglyak L, Nickerson DA. Variation is the spice of life. **Nat Genet** 2001; 27:234-6.
- Lee JH, Welch DR. Suppression of metastasis in human breast carcinoma MDA-MB-435 cells after transfection with the metastasis suppressor gene, KiSS-1. **Cancer Res** 1997; 57:2384-7.
- Levinson DF, Mahtani MM, Nancarrow DJ, et al. Genome scan of schizophrenia. **Am J Psychiatry** 1998; 155:741-50.
- Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. **Nucleic Acids Res** 2000; 28:126-8.

- Marth GT, Korf I, Yandell MD, et al. A general approach to single-nucleotide polymorphism discovery. **Nat Gene**. 1999; 23:452-6.
- McGeer PL, McGeer EG. Polymorphisms in inflammatory genes and the risk of Alzheimer disease. **Arch Neurol** 2001; 58:1790-2.
- McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J. Linkage disequilibrium in domestic sheep. **Genetics** 2002; 160:1113-22.
- Meirelles Kalil Pessoa de Barros E, Rodrigues CJ, de Barros TE, Bevilacqua RG. Presence of apolipoprotein E epsilon4 allele in cerebral palsy. **J Pediatr Orthop** 2000; 20:786-9.
- Menendez-Arias L. Molecular basis of fidelity of DNA synthesis and nucleotide specificity of retroviral reverse transcriptases. **Prog Nucleic Acid Res Mol Biol** 2002; 71:91-147.
- Mu J, Duan J, Makova KD, et al. Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. **Nature** 2002; 418:323-6.
- Mubumbila V, Sutter A, Ptok U, Heun R, Quirin-Stricker C. Identification of a single nucleotide polymorphism in the choline acetyltransferase gene associated with Alzheimer's disease. **Neurosci Lett** 2002; 333:9-12.
- Nasu S, Suzuki J, Ohta R, et al. Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. **DNA Res** 2002; 9:163-71.
- Ohtaki T, Shintani Y, Honda S, et al. Metastasis suppressor gene KiSS-1 encodes peptide ligand of a G-protein-coupled receptor. **Nature** 2001; 411:613-7.

- Park JY, Muscat JE, Kaur T, et al. Comparison of GSTM polymorphisms and risk of oral cancer between African-American and caucasians. **Pharmacogenetics** 2000; 10:123-31.
- Parkin DM, Pisani P, Ferlay J. Estimates of worldwide incidence of eighteen major cancers in 1985. **Int J Cancer** 1993; 54:594-606.
- Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. **Int J Cancer** 1999; 80:827-41.
- Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM, Pena SD. Color and genomic ancestry in Brazilians. **Proc Natl Acad Sci U S A** 2003; 100:177-82.
- Patil N, Berno AJ, Hinds DA, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. **Science** 2001; 294:1719-23.
- Pharoah PO. Twins and cerebral palsy. **Acta Paediatr Suppl** 2001; 90:6-10.
- Pharoah PO. Neurological outcome in twins. **Semin Neonatol** 2002; 7:223-30.
- Pisani P, Parkin DM, Ferlay J. Estimates of worldwide mortality from eighteen major cancers in 1985: implications for prevention and projections of future burden. **Int J Cancer** 1993; 55:891-903.
- Pisani P, Parkin DM, Bray F, Ferlay J. Estimates of worldwide mortality from 25 cancers in 1990. **Int J Cancer** 1999; 83:18-29.
- Rees J. Complex disease and the new clinical sciences. **Science** 2002; 296:698-700
- Riggins GJ, Strausberg RL. Genome and genetic resources from the Cancer Genome Anatomy Project. **Hum Mol Genet** 2001; 10:663-7.
- Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. **Nature** 2001; 409:928-33.

- Sanson GF, Kawashita SY, Brunstein A, Briones MR. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. **Mol Biol Evol** 2002; 19:170-8.
- Sawa A, Snyder SH. Schizophrenia: diverse approaches to a complex disease. **Science** 2002; 296:692-5.
- Schantz SP, Byers RM, Goepfert H. Tobacco and cancer of the tongue in young adults. **JAMA** 1988; 259:1943-4.
- Schechtman D, Tarrab-Hazdai R, Arnon R. The 14-3-3 protein as a vaccine candidate against schistosomiasis. **Parasite Immunol** 2001; 23:213-7.
- Schmidt S, Barcellos LF, DeSombre K, et al. Multiple Sclerosis Genetics Group: association of polymorphisms in the apolipoprotein E region with susceptibility to and progression of multiple sclerosis. **Am J Hum Genet** 2002; 70:708-17.
- SEER Cancer Statistics Review, 1973-1992. U.S. Bethesda: Department of Health and Human Services; 1994. [NIH Publication No. 96-2789].
- Shemen LJ, Klotz J, Schottenfeld D, Strong E. Increase of tongue cancer in young men [Letter]. **JAMA** 1984; 252:1857.
- Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. **Genome Res** 1999; 9:677-9.
- Shubitowski DM, Venta PJ, Douglass CL, Zhou RX, Ewart SL. Polymorphism identification within 50 equine gene-specific sequence tagged sites. **Anim Genet** 2001; 32:78-88.
- Strausberg RL, Dahl CA, Klausner RD. New opportunities for uncovering the molecular basis of cancer. **Nat Genet** 1997; 15 Spec No:415-6.



- Strohman R. Maneuvering in the complex path from genotype to phenotype. **Science** 2002; 296:701-3.
- Szala S, Kasai Y, Steplewski Z, Rodeck U, Koprowski H, Linnenbach AJ. Molecular cloning of cDNA for the human tumor-associated antigen CO-029 and identification of related transmembrane antigens. **Proc Natl Acad Sci U S A** 1990; 87:6833-7.
- Tavtigian SV, Simard J, Teng DH, et al. A candidate prostate cancer susceptibility gene at chromosome 17p. **Nat Genet** 2001; 27:172-80.
- Topic, E, Stefanovic, M, Ivanisevic, AM, Petrinovic, R, Curcic, I. The cytochrome p450 2D6 (CYP2D6) gene polymorphism among breast and head and neck cancer patients. **Clin Chim Acta** 2000; 296:101-9.
- Volkman SK, Hartl DL, Wirth DF, et al. Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. **Science** 2002; 298:216-8.
- Wang L, McDonnell SK, Elkins DA, et al. Role of HPC2/ELAC2 in hereditary prostate cancer. **Cancer Res** 2001; 61:6494-9.
- Welch DR, Chen P, Miele ME, et al. Microcell-mediated transfer of chromosome 6 into metastatic human C8161 melanoma cells suppresses metastasis but does not inhibit tumorigenicity. **Oncogene** 1994; 9:255-62.
- Yang W, Jackson DC, Zeng Q, McManus DP. Multi-epitope schistosome vaccine candidates tested for protective immunogenicity in mice. **Vaccine** 2000; 19:103-13.
- Zhang G, Zhang S, Chen W, et al. Go!Poly: A gene-oriented polymorphism database. **Hum Mutat** 2001; 18:382-7.

Endereços eletrônicos

Banco de SNPs da Celera (está associado a Applied Biosystems)

<http://www.celeradiscoverysystem.com>

BLAT

<http://genome.ucsc.edu/cgi-bin/hgBlat?command=start>

ClustalW Multiple Sequence Alignment

<http://searchlauncher.bcm.tmc.edu:9331/multi-align/Options/clustalw.html>

dbSNP

<http://www.ncbi.nlm.nih.gov/SNP>

Enzimas de restrição

<http://darwin.bio.geneseo.edu/~yin/WebGene/RE.html>

Epi Info

<http://www.cdc.gov/epiinfo>

HCGP – SNP

<http://bit.fmrp.usp.br/bit/projects.php?ptitle=Projects>

Informações sobre os dados de SNPs no CGAP

<http://lpg.nci.nih.gov/cgi-bin/tally.pl?tt=0&ht=0&f=h>

Modelagem de estrutura por homologia

<http://www.expasy.ch/swissmod/SWISS-MODEL.html>

NCBI e ferramentas

<http://www.ncbi.nlm.nih.gov/>

Primer 3

http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi

Projeto Genoma de *Schistosoma mansoni*

<http://bioinfo.iq.usp.br/schisto>

Protein Data Bank

<http://www.rcsb.org/pdb>

PubCrawler

http://www.hgmp.mrc.ac.uk/embnet.news/vol6_1/pubcrawler.html

Repeat Masker

<http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker>

Anexos

Anexo 1 - Exemplo de arquivo montado com todas as informações de cada

SNP: do alinhamento das ORESTES até o desenho dos oligos.

NM_004161
HSD00365

```

HSD00365 - TCGTTATGCCAGTGAAAAATGTCAACAAATTGTTGGTAGGGAACAAATGTGA
PM2-BT0723-090201-010-c07 - .....C.....
PM2-BT0723-130900-002-c07 - .....C.....
MR3-GN0190-301100-004-e08 - .....C.....C.....
MR4-ET0140-220101-004-d02 - .....C.....
MR4-EN0075-220101-006-d02 - .....C.....
RC5-BT0377-030200-012-A06 - .....a.....
PM2-BT0723-100900-003-h04 - .....a.*.....C.....
IL2-FT0160-070800-121-C02 - .....a.....
MR4-ET0140-190201-007-h04 - .....a.....
MR0-RT0037-121200-004-d02 - .....a.....
CM1-HN0016-161100-568-c06 - .....a.....
QV3-BN0046-150300-121-a12 - .....a.....
QV3-DT0045-210100-063-f03 - .....a.....
QV2-NN0045-220800-323-d03 - .....a.....
IL5-UM0067-240300-051-g06 - .....ggg.....a.....
CM4-HN0021-241100-457-h02 - .....a.....
MR0-RT0037-011200-002-a07 - .....a.....
MR2-UM0060-030400-103-g02 - .....a.....g.....
PM0-IT0018-091100-001-e02 - .....a.....
PM1-MT0143-101100-003-a06 - .....*.a.....
PM1-MT0143-101100-003-f11 - .....a.....
```

Query = gi|4758987|ref|NM_004161.1| Homo sapiens RAB1, member RAS oncogene family
(RAB1), mRNA (723 letters)

>gi|13176657|gb|AC084278.2|AC084278 Homo sapiens chromosome RPCI-11 clone RP11-
350H15, WORKING DRAFT

SEQUENCE, 11 unordered pieces
Length = 164811

Score = 484 bits (244), Expect = e-134
Identities = 254/256 (99%), Gaps = 1/256 (0%)
Strand = Plus / Plus

```

Query: 469 aggaatttgctgattcccttggaattccggtttttggaaccagtgctaagaatgcaacga 528
          |||
Sbjct: 57067 aggaatttgctgattcccttggaattccggtttttggaaccagtgctaagaatgcaacga 57126
```

```

Query: 529 acgtagaacagctctttcatgacgatggcagctgagattaaaaagcgaatgggtcccgag 588
          |
Sbjct: 57127 atgtagaacagctctttcatgacgatggcagctgagattaaaaagcgaatgggtcccgag 57186
```

```

Query: 589 caacagctggtggtgctgagaagtccaatgttaaaattcagagcactccagtcagcagt 648
          |||
Sbjct: 57187 caacagctggtggtgctgagaagtccaatgttaaaattcagagcactccagtcagcagt 57246
```

```

Query: 649 caggtggaggttgctgct-aaatttgctccatccttttctcacagcaatgaatttgcaa 707
          |||
Sbjct: 57247 caggtggaggttgctgctaaaatttgctccatccttttctcacagcaatgaatttgcaa 57306
```

```

Query: 708 tctgaacccaagtga 723
          |||
Sbjct: 57307 tctgaacccaagtga 57322
Score = 268 bits (135), Expect = 5e-69
```

Identities = 135/135 (100%)
Strand = Plus / Plus

Query: 337 aggagtccttcaataatgttaaacagtggtgcaggaaatagatcgttatgccagtgaag 396
|||||
Sbjct: 56687 aggagtccttcaataatgttaaacagtggtgcaggaaatagatcgttatgccagtgaag 56746

Query: 397 atgtcaacaaattgttggtaggaacaaatgtgatctgaccacaaagaaagtagtagact 456
|||||
Sbjct: 56747 atgtcaacaaattgttggtaggaacaaatgtgatctgaccacaaagaaagtagtagact 56806

Query: 457 acacaacagcgaagg 471
|||||
Sbjct: 56807 acacaacagcgaagg 56821

Score = 192 bits (97), Expect = 2e-46
Identities = 97/97 (100%)
Strand = Plus / Plus

Query: 243 tgggacacagcaggccaggaaagatttcgaacaatcacctccagttattacagaggagcc 302
|||||
Sbjct: 54681 tgggacacagcaggccaggaaagatttcgaacaatcacctccagttattacagaggagcc 54740

Query: 303 catggcatcatagttgtgtatgatgtgacagatcagg 339
|||||
Sbjct: 54741 catggcatcatagttgtgtatgatgtgacagatcagg 54777

Score = 190 bits (96), Expect = 9e-46
Identities = 96/96 (100%)
Strand = Plus / Minus

Query: 147 gatgatacatatacagaaagctacatcagcacaattggtgtggatttcaaaataagaact 206
|||||
Sbjct: 158765 gatgatacatatacagaaagctacatcagcacaattggtgtggatttcaaaataagaact 158706

Query: 207 atagagttagacgggaaaacaatcaagcttcaaata 242
|||||
Sbjct: 158705 atagagttagacgggaaaacaatcaagcttcaaata 158670

Score = 147 bits (74), Expect = 1e-32
Identities = 74/74 (100%)
Strand = Plus / Plus

Query: 74 tgattatttattcaagttacttctgattggcgactcaggggttgaaagtcttgccttct 133
|||||
Sbjct: 150310 tgattatttattcaagttacttctgattggcgactcaggggttgaaagtcttgccttct 150369

Query: 134 tcttaggtttgcag 147
|||||
Sbjct: 150370 tcttaggtttgcag 150383

Score = 73.8 bits (37), Expect = 1e-10
Identities = 37/37 (100%)
Strand = Plus / Minus

Query: 37 gcagctgcagtgacatgtccagcatgaatcccgaata 73
|||||
Sbjct: 30340 gcagctgcagtgacatgtccagcatgaatcccgaata 30304

Sequence 1	lcl seq_1	Length	51	(1 .. 51)
Sequence 2	gi 4758987	Homo sapiens RAB1, member RAS oncogene family		
(RAB1), mRNA		Length	723	(1 .. 723)

Anexo 2 - Análise de Ninjurina 1 em pacientes com Paralisia Cerebral

Componentes genéticos e Paralisia Cerebral

A maioria dos casos de paralisia cerebral está associada ao período perinatal ou no máximo até dois anos de desenvolvimento do indivíduo, sendo que em grande parcela dos casos o componente causal não pode ser identificado. Historicamente, a paralisia cerebral sempre esteve associada ao momento do parto e conseqüentemente à habilidade e procedimentos do obstetra. Atualmente esta versão já não é tão considerada.

Estudo envolvendo marcadores moleculares e paralisia cerebral são escassos, dentre eles podemos mencionar a associação entre a presença do alelo $\epsilon 4$ da apolipoproteína E em pacientes com paralisia cerebral (MEIRELLES KALIL PESSOA DE BARROS et al. 2000)

Estudos com gêmeos também têm ajudado a esclarecer os processos envolvidos e os não envolvidos com o desenvolvimento de paralisia cerebral. Ao se comparar a prevalência de paralisia cerebral em gêmeos do mesmo sexo (grupo onde encontramos os monozigóticos) e em gêmeos de sexos distintos (todos dizigóticos), notamos que uma diferença estatisticamente significativa separa os casos onde somente um dos indivíduos é afetado, ou seja, para gêmeos do mesmo sexo existe uma probabilidade maior dos dois indivíduos serem afetados do que quando comparados a gêmeos de sexos distintos (PHAROAH 2001, 2002). Infelizmente, estudos envolvendo gêmeos dizigóticos e monozigóticos em separado ainda não foram realizados, mas os resultados obtidos até o momento, com relação ao sexo dos indivíduos, têm demonstrado que o componente genético está mais fortemente

associado do que o fator ambiental (neste caso características do parto, procedimentos do obstetra, cuidados pré e pós-natais, etc). Mais uma vez este é um campo que necessita de estudos sobre os componentes genéticos envolvidos com a patologia.

Amostras de pacientes com Paralisia Cerebral

As amostras de pacientes com Paralisia Cerebral e seus respectivos controles foram obtidas no Instituto de Ortopedia e Traumatologia do HC/FM/USP junto a nossos colaboradores Dra. Erika Meirelles Kalil Pessoa de Barros e Dr. Tarcísio Eloy Pessoa de Barros Filho.

Nosso grupo de estudo foi composto por 33 indivíduos que apresentavam paralisia cerebral e 36 indivíduos apresentando exame neurológico normal.

Os critérios de inclusão para os pacientes com paralisia cerebral foram:

i) Serem portadores de paralisia cerebral sem causa estabelecida, ou seja, indivíduos que apesar de terem história clínica de gestação, parto e período neonatal normais desenvolveram a síndrome. A mesma, no acompanhamento clínico, foi considerada decorrente de hipóxia perinatal, mesmo que não houvesse registro médico ao nascimento que comprovasse esta hipótese;

ii) Ausência de diagnóstico de qualquer síndrome genética, história de trauma cranioencefálico ou de qualquer situação que sabidamente resulte em comprometimento de aporte de oxigênio ao sistema nervoso central;

iii) Indivíduos cujas mães não fizeram uso de substâncias sabidamente nocivas ao feto durante a gestação, como uso de drogas estimulantes ou depressoras,

permitidas ou não, ingestão de bebidas alcoólicas ou uso de drogas com finalidade abortiva. Não foi possível esclarecer o uso de tabaco em todos os casos, não sendo este, portanto, considerado um critério de exclusão.

O grupo formado por portadores de paralisia cerebral foi subdividido em dois.

O primeiro, portadores de paralisia cerebral leve a moderada, foi composto por 21 indivíduos ao todo. São indivíduos socialmente integrados, freqüentando escolas de acordo com sua capacidade intelectual individual. Todos estes indivíduos encontravam-se inseridos em um programa de terapia esportiva, mantido pelo Instituto de Ortopedia e Traumatologia do HC/FM/USP. Todos apresentavam marcha independente.

O segundo subgrupo, portadores de paralisia cerebral grave, foi composto por 12 indivíduos internados em instituição atendida pelo Instituto de Ortopedia e Traumatologia. Eram indivíduos com importante comprometimento das funções cognitivas, todos com rebaixamento intelectual extremo, incapazes de reconhecer pessoas, restritos aos seus leitos, sentando-se apenas com apoio e supervisão, sem qualquer capacidade de integração social, sem controles esfínterianos e sem desenvolvimento de fala.

SNPs em pacientes com Paralisia Cerebral

Os pacientes portadores de paralisia cerebral e seus respectivos controles foram genotipados para o SNP de Ninjurina 1 não tendo sido observado nenhuma associação nem entre os alelos nem entre os genótipos de Ninjurina 1 e a severidade

da patologia. Também não foi verificada nenhuma diferença significativa entre os alelos ou os genótipos de pacientes comparados a controles. Contudo o pequeno tamanho de nossa amostra não exclui a possibilidade de exista algum tipo de associação entre esta patologia e o gene Ninjurina 1.

- As amostras de pacientes com paralisia cerebral necessitam ser expandidas para confirmarmos os resultados negativos encontrados até o momento.

Meirelles Kalil Pessoa de Barros E, Rodrigues CJ, de Barros TE, Bevilacqua RG. Presence of apolipoprotein E epsilon4 allele in cerebral palsy. **J Pediatr Orthop** 2000; 20:786-9.

Pharoah PO. Twins and cerebral palsy. **Acta Paediatr Suppl** 2001; 90:6-10.

Pharoah PO. Neurological outcome in twins. **Semin Neonatol** 2002; 7:223-30.

Anexo 3 - Análise de Ninjurina 1 em pacientes com Trauma Raquimedular

Danos pós-trauma e os mecanismos envolvidos

Danos na coluna vertebral (SCI) promovem uma perda considerável das funções neurológicas abaixo do local da lesão e múltiplos efeitos adversos em vários órgãos. Pesquisas em neurotrauma têm surgido como um campo novo e evoluído significativamente nos últimos anos. Em SCI, existem linhas de tratamento em três diferentes frentes: neuroproteção aguda, estímulo à regeneração e tratamento da desmielinização (BLIGHT 2002).

O conceito de injúria secundária tratável vem de evidências histopatológicas que sugerem que nem todo o dano é causado no momento do impacto na coluna vertebral, mas que danos adicionais são produzidos e se acumulam por uma variedade de outros processos. Inicialmente, a coluna vertebral pode não apresentar, ou apresentar de forma moderada, destruição tecidual e perda funcional permanente, mas em poucas horas, a parte central da medula apresenta uma necrose hemorrágica, iniciando do centro da injúria (BLIGHT 2002).

O sistema nervoso central (SNC) dos mamíferos possui limitada habilidade de regeneração de axônios danificados. A relutância das fibras nervosas do SNC em se reparar, aparentemente é resultado de um número de mecanismos que limitam a plasticidade geral do SNC, mecanismos estes que envolvem proteínas que controlam a extensão e a regeneração axonal.

Também é observada uma escassa mielinização das fibras nervosas sobreviventes ao trauma, que pode estar envolvida nas perdas funcionais permanentes. Danos na mielina são provavelmente devidos a processos inflamatórios

no tecido lesado e podem envolver morte de oligodendrócitos por apoptose (BLIGHT 2002).

Apesar dos mecanismos fisiológicos envolvidos na SCI estarem sendo estudados e compreendidos, pouco ou quase nada é sabido sobre os componentes genéticos associados, principalmente aos danos observados após trauma. Existem casos em que um trauma leve desencadeia uma lesão gravíssima e existem outros casos em que um trauma severo não desencadeia um dano tão extenso. Neste campo, muitas questões ainda necessitam de resposta e provavelmente as diferentes evoluções clínicas individuais estão associadas a componentes genéticos.

Amostras de pacientes com Trauma Raquimedular

Assim como as amostras de Paralisia Cerebral, as amostras de Trauma Raquimedular e seus respectivos controles foram obtidas no Instituto de Ortopedia e Traumatologia do HC/FM/USP junto a nossos colaboradores Dra. Erika Meirelles Kalil Pessoa de Barros e Dr. Tarcísio Eloy Pessoa de Barros Filho.

Nosso grupo de estudo foi composto por 27 indivíduos que apresentavam lesão neurológica completa após trauma raquimedular e 36 não vítimas de trauma raquimedular e não apresentando qualquer patologia de origem neurológica, tratados para lesões traumáticas diversas no Instituto de Ortopedia e Traumatologia do HC/FM/USP,.

Os critérios de inclusão para os pacientes foram:

- i) Serem portadores de lesão neurológica completa após trauma raquimedular há pelo menos dois anos;

ii) Indivíduos que não apresentavam alteração de seus quadros neurológicos no último ano precedente à coleta de sangue;

iii) Fratura vertebral do tipo explosiva, sempre na ausência de luxação facetária na presença de medula no canal vertebral sem cone medular ou raízes nervosas;

iv) Idade inferior a 35 anos na ocasião do trauma;

v) Não ter o estado neurológico alterado por tratamento cirúrgico;

vi) Indivíduos tratados clínica ou cirurgicamente pela mesma equipe do Grupo de Trauma Raquimedular do Instituto de Ortopedia e Traumatologia do HC/FM/USP, seguindo-se os mesmos protocolos em todos os casos.

SNPs em pacientes com Trauma Raquimedular

Para os pacientes com trauma raquimedular, e controles, além da genotipagem do polimorfismo de Ninjurina 1, um SNP presente em um gene relacionado a resposta inflamatória também foi genotipado pela pós doutoranda Cíntia Fridman aluna de nosso laboratório (Tabela 10). O gene em questão é o Alox12 (LucusLink: 239).

Em nossa amostra foi encontrado um aumento significativo do número de homozigotos AA para Ninjurina 1 ligados a presença do alelo A para Alox 12 em pacientes quando comparados a seus controles (33,3% em controles e 59,2% em pacientes; $p = 0,012$ e *Odds Ratio* = 4.36, $1.32 < OR < 14.92$), contudo nenhuma associação significativa foi encontrada quando comparamos os dados de Ninjurina 1 ou Alox12 separadamente. Estes resultados corroboram a hipótese de que a

Tabela 10 - Genótipos dos pacientes com trauma raquimedular e controles para Ninjurina 1 e Alox12

Grupos	Sujeitos	Genótipo de Ninjurina 1	Genótipo de ALOX12
Controles	1	AC	-
	2	AC	AA
	3	AC	GA
	4	AC	AA
	5	AA	AA
	6	AC	GA
	7	AA	GA
	8	CC	GG
	9	AC	GA
	10	AC	-
	11	AC	GA
	12	AA	GG
	13	AA	GG
	14	AC	GA
	15	AA	GA
	16	AA	GA
	17	AA	GG
	18	AC	AA
	19	AA	GG
	20	AA	GG
	21	AC	GA
	22	AA	GA
	23	AA	GG
	24	AA	AA
	25	AA	GG
	26	AA	GG
	27	AA	GG
	28	AA	GG
	29	AA	GA
	30	AC	GA
	31	AA	GA
	32	AC	GA
	33	AC	GA
	34	AA	GG
	35	AC	GG
	36	CC	AA
	37	AA	GA
	38	AC	GG
Trauma raquimedular	S1	AA	AA
	S2	AA	GA
	S3	AA	GA
	S4	AC	GG
	S5	AA	AA
	S6	AA	GA
	S7	AA	GA
	S8	AA	GG
	S9	AC	GA
	S10	AA	AA
	S11	AA	GG
	S12	AA	GG
	S13	AA	GG
	S14	AC	GA
	S15	AA	GG
	S16	AC	GA
	S17	AA	GA
	S18	AA	AA
	S19	AA	AA
	S20	AA	GA
	S21	AC	GG
	S22	AC	GG
	S23	AA	GA
	S24	AA	GA
	S25	AA	-
	S26	AA	GA
	S27	AA	GA
	S28	AA	GA

combinação de processos inflamatórios e mecanismos de reparo axonal podem estar envolvidos na evolução da lesão nos casos de trauma raquimedular.

- Foi encontrada uma associação significativa entre o genótipo AA de Nijurina 1 e o alelo A de Alox12 em pacientes com Trauma Raquimedular ($p = 0,0127$, OR = 4,36).

- As amostras de pacientes com trauma raquimedular necessitam ser expandidas para confirmarmos os resultados positivos encontrados até o momento.

- Para as análises dos casos de trauma raquimedular, amostras de indivíduos portadores de lesões leves geradas por traumas severos devem ser adquiridas.

Blight AR. Miracles and molecules-progress in spinal cord repair. **Nat Neurosci** 2002; 5 Suppl:1051-4.

Colaborações

Este trabalho contou com colaborações que foram de grande importância para o bom andamento e sucesso do projeto.

Dr. Marco Antônio Zago, Dr. Wilson Araújo da Silva Jr. e Dra Maria Cristina Costa – Coordenadores do grupo de seqüenciamento RC e Coordenadores da iniciativa de análise de SNPs dentro do Projeto Genoma do Câncer Ludwig/FAPESP.

Adriano Holanda, Israel Tojal, Karine Ferreira e Rochele Azevedo – Integrantes do Laboratório de Biologia Molecular e Bioinformática do Hemocentro de Ribeirão Preto.

Dr. Luiz Paulo Kowalski – Cirurgião Chefe do Depto de Cabeça e Pescoço e Otorrinolaringologia do Hospital do Câncer/Fundação Antônio Prudente.

Dra. Erika Meirelles Kalil Pessoa de Barros e Dr. Tarcísio Eloy Pessoa de Barros Filho – Instituto de Ortopedia e Traumatologia do HC/FM/USP.

Dr. Milton Moraes – FIOCRUZ, Rio de Janeiro.

Dr. Wagner Farid Gattaz, Dr. Homero Vallada e Dr. Orestes Forlenza – Laboratório de Neurociências, do Departamento de Psiquiatria, da Faculdade de Medicina da Universidade de São Paulo.

Dr. David Rollinson – Museu de História Natural de Londres.

Projeto Genoma de *Schistosoma mansoni* – Grupo coordenado pelo Dr. Sergio Verjovski-Almeida.

Curriculum vitae (resumo)

Resumos em congressos

-Marcos, C.; **Guimarães, P.E.M.** e Melo-Júnior, T.A. 1996. Utilização do ninho de *Phacelodomus rufifrons* por *Falco femoralis*.; VI Congresso Brasileiro de Ornitologia. Belo Horizonte, MG.

- Penaforte, C.L.; Prado, V.F.; Beirão, P.S.L.; Gomez, M.G.; Diniz, C.R.; Prado, M.A.M.; Romano-Silva, M.A.; Cruz, J.S.; **Guimarães, P.E.M.** and Kalapothakis, E. 1997. Molecular cloning of Tx2-6: A neurotoxic peptide from *Phoneutria nigriventer* spider. XXVI^a Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular, Caxambú MG.

- **Guimarães, P.E.M.** 1998. Parasitismo de *Dione juno juno* (Lepidoptera: Nymphalidae) por *Microcharops peronota* (Hymenoptera: Ichneumonidae). XXII Congresso Brasileiro de Zoologia, Recife PE.

- **Guimarães, P.E.M.**; Santos, M.S.; Alves-Silva, J.; Pena, S.D.J. and Prado, V.F. 1998. Sequence diversity of mtDNA hypervariable segment II in Brazilian Caucasinans. XXVIIa Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular, Caxambú MG.

- Penaforte, C.L.; Jardim, S.N., **Guimarães, P.E.M.**, Romano-Silva, M.A.; Prado, M.A.M.; Prado, V.F.; Gomez, M.G.; and Kalapothakis, E. 1998. Molecular cloning of cDNA encoding insecticidal neurotoxic peptides from the spider *Phoneutria nigriventer*. V Simpósio da Sociedade Brasileira de Toxinologia, Angra dos Reis RJ.

- Prado V.F., Barbosa Jr J., Santos M.S., **Guimarães P.E.M.**, Gomez M.V., Romano-Silva M.A. & Prado M.A.M. 1999. Visualization of the vesicular acetylcholine transporter in SN56 cells. Joint Meeting of the International Society for Neurochemistry and the European Society for Neurochemistry, Berlin, Alemanha.

- **Guimarães P.E.M.**, Barbosa Jr J., Gomez M.V., Prado M.A.M. & Prado V.F., 1999. Expression of VACHT in a cholinergic cell line detected with an antibody against a recombinant C-terminal peptide. 45° Congresso Nacional de Genética, Gramado RS.

- **Guimarães P.E.M.** & Santos F.R. 2000. Banco de DNA. 46° Congresso Nacional de Genética, Águas de Lindoia, SP.

- Ojopi E.P.B., **Guimarães P.E.M.**, Rogato S.R. & Dias-Neto E. 2001. Head and Neck transcriptional map of 22q11.2-12.2: gene identification and SNP analysis based in 200,000 Head and Neck ORESTES. I Brazilian International Genome Conference, Angra dos Reis, RJ.

- Costa, MCR; Silva Jr, WA; Holanda, AJ; Tojal, I; Dias-Neto, E; **Guimarães PEM**; Ojopi, EPB; Souza, SJ; Simpson, AJG; Zago, MA. 2001 Identification of coding single nucleotide polymorphisms (cSNPs) in genes expressed in tumoral cells using the HCGP– ORESTES database. 51st American Society of Human Genetics Annual Meeting, San Diego, Califórnia, EUA.

- **Guimarães PEM**; Aires, KA; Ho, PL; Martins, EA; Leite, RA; Marques, RC; Ribeiro, MA; Gregório SP; Paquola, AP; Nishiyama-Jr, MY; Reis, EM; Piazza, JP; Kitajima, JP; Ojopi, EPB; Rodrigues, V; Gargione, C; Kawano, T; Madeira, AM; Leite, LC; Verjovski-Almeida, S and Dias-Neto, E. 2002. *The Schistosoma mansoni*

Genome Project reveals that candidate vaccine antigens recommended by WHO are polymorphic. 48º Congresso Nacional de Genética, Águas de Lindoia, SP.

- Fridman, C; Ojopi, EP; Gregório, SP; Ikenaga, EH; **Guimarães PEM**; Vallada, HP; Gattaz, WF and Dias-Neto, E. 2002. *Association of a new polymorphism in 12-LOX gene with Bipolar Disorder.* 48º Congresso Nacional de Genética, Águas de Lindoia, SP.

- Ferreira, KS; Costa MCR; Azevedo, R; **Guimarães PEM**; Dias-Neto, E; Greene, LJ e Zago, MA. 2002. *Expressão do gene Ninjurina 1 em células neoplásicas e avaliação da frequência do polimorfismo D110A na população Brasileira.* 48º Congresso Nacional de Genética, Águas de Lindoia, SP.

Palestras e cursos proferidos

- Criação de bancos de DNA e genética da conservação.

11 de Setembro de 2000, Universidade de Franca, Franca SP.

- Projeto Genoma do Câncer Humano.

06 de Setembro de 2002, Universidade Estadual de Minas Gerais, Divinópolis MG.

- Biologia Molecular e Conservação da Biodiversidade

23 a 25 de Outubro de 2002, Universidade de Franca, Franca SP.

Artigos científicos

- Kalapothakis, E.; Penaforte, C.L.; Beirão, P.S.L.; Romano-Silva, M.A.; Cruz, J.S.; Gomez, M.G.; Prado, M.A.M.; **Guimarães, P.E.M.** and Prado, V.F. Cloning of cDNA encoding neurotoxic peptides from the spider *Phoneutria nigriventer*. *Toxicon* 1998 Dec;36(12):1843-50.

- Alves-Silva, J.; **Guimarães, P.E.M.**; Rocha, J.M.; Pena, S.D.J. & Prado, V.F. Identification in Portugal and Brazil of a mtDNA lineage containing a 9-bp triplication in the intergenic COII/tRNA^{Lys} region. *Hum Hered* 1999 Jan;49(1):56-8.

- Penaforte, C.L.; Prado, V.F.; Prado, M.A.M.; Romano-Silva, M.A.; **Guimarães, P.E.M.**; Gomez, M.G.; and Kalapothakis. Molecular cloning of cDNA encoding insecticidal neurotoxic peptides from the *Phoneutria nigriventer* spider. *Toxicon* 2000 Oct;38(10):1443-9.

- Alves-Silva, J.; Santos, M.S.; **Guimarães, P.E.M.**; Ferreira, A.C.S.; Bandelt, H.-J.; Pena, S.D.J. & Prado, V.F. The ancestry of Brazilian mitochondrial DNA lineages. *Am J Hum Genet* 2000 Aug;67(2):444-61.

- Bandelt H.-J.; Alves-Silva, J.; **Guimarães, P.E.M.**; Santos, M.S.; Brehm, A.; Pereira, L.; Coppa, A.; Larruga, J.M.; Scozzari, R.; Torroni, A.; Prata, M.J.; Amorim, A.; Prado, V.F. & Pena, S.D.J. Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *An Hum Gen* 2001 Nov 65(06): 549-563.

- **Guimarães PEM** e Costa, MCR. SNPs: Sutis diferenças de um código. *Biotechnologia*. 2002 Maio/Junho Ano IV N26: 24-27.

- Santos, FR; Redondo, R; **Guimarães, PEM**. Bancos de DNA: Coleções estratégicas para estudos da biodiversidade. *Lundiana*, 2002 3 (2): 93-98.
- Costa, MCR; Silva Jr, WA; Holanda, AJ; Tojal, I; Ferreira, KS; Azevedo, R; Simpson, AJG; **Guimarães PEM**; Dias-Neto, E and Zago, MA. Identification of Single Nucleotide Polymorphisms in the Coding Region of Human Genes Using Open Reading frame ESTs (ORESTES). Trabalho submetido.
- Fridman, C; Ojopi, EP; Gregório, SP; Ikenaga, EH; **Guimarães PEM**; Vallada, HP; Gattaz, WF and Dias-Neto, E. Association of a new polymorphism in 12-LOX gene with Bipolar Disorder. *in press*.
- Bandelt H.-J.; **Guimarães, P.E.M.** & Santos, F.R. Inferring low-resolution intraspecific phylogenies in the presence of homoplasy: some caveats. Em preparação.