UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

THIAGO YUKIO KIKUCHI OLIVEIRA

Aspectos moleculares da evolução do gene
DARC em primatas

Ribeirão Preto – SP

2008

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Catálogo na Publicação
Serviço de Documentação
Faculdade de Medicina de Ribeirão Preto

Oliveira, Thiago Yukio Kikuchi

Aspectos moleculares da evolução do gene *DARC* em primatas / Thiago Yukio Kikuchi Oliveira; orientador: Wilson Araújo da Silva Jr. – Ribeirão Preto -SP, 2008.
90 f. : il.

Dissertação (Mestrado – Programa de Pós-Graduação em Genética. Área de Concentração: Genética) – Faculdade de Medicina de Ribeirão Preto.

1. Evolução – Gene *DARC*. 2. Sequência – Gene *DARC*.

THIAGO YUKIO KIKUCHI OLIVEIRA

Aspectos moleculares da evolução do gene *DARC* em primatas

Dissertação apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Mestre em Ciências.

Área de Concentração: Genética
Orientador: Prof. Dr. Wilson Araújo da Silva
Júnior

Ribeirão Preto – SP

2008

Dedicatória

*Dedico esta dissertação aos meus pais
Antonio de Oliveira e Luiza Kimi Kikuchi de Oliveira
cujo exemplo de honestidade, trabalho e força de vontade
tem sido um norteador para a minha vida.
É por vocês que eu cheguei até aqui e é por vocês que continuo em frente.
Dedico, também, à minha namorada Lilian Nogueira que sempre
me apoiou nos momentos mais difíceis.*

Agradecimentos

Ao Prof. Dr. Wilson Araújo da Silva Junior pela orientação, incentivo, paciência e principalmente por acreditar em mim.

Ao Prof. Dr. Eugene E. Harris do Departamento de Ciências Biológicas e Geologia da Queensborough Community College - USA, por ceder parte de suas amostras de primatas do Velho Mundo, pelo convite à Nova York e principalmente pela confiança depositada.

Ao Prof. Dr. Diogo Meyer do Instituto de Biologia da Universidade de São Paulo, pela enorme atenção e apoio sempre manifestado.

À doutoranda e “quase irmã” Alynne Oya e Chiromatzo pela ajuda, companhia e paciência em todos os momentos.

A todos da equipe do Laboratório de Genética Molecular pela alegre convivência e sábios conselhos durante a elaboração do projeto. Em especial, Cristiane Ayres Ferreira, Carla Martins Kaneto, Patrícia Santos Pereira Lima e Greice Molfetta.

Ao pessoal do Laboratório de Bioinformática (BiT) pela companhia e incentivo durante esses anos. Em especial, Brown (vulgo Rodrigo Martins Brandão), Israel Tojal da Silva, Daniel G. Pinheiro, Gislaine da Silva P. Pereira, Rodrigo Lucena Borges e Olsen Rodrigo.

À Meire Tarla pela paciência, ajuda e dedicação nas tarefas burocráticas e administrativas.

À Dalvinha pela ajuda nas tarefas burocráticas e principalmente pela alegria transmitida durante nossas conversas.

Aos meus pais, Antonio de Oliveira e Luiza Kimi Kikuchi de Oliveira, e à minha namorada, Lilian Nogueira, por estarem sempre ao meu lado mesmo à distância.

Aos empregados encarregados da limpeza do laboratório, pela paciência exercida nas várias noites em que atrapalhei o seu trabalho.

A todos que não citei por motivo de esquecimento mas que têm o meu profundo agradecimento por terem contribuído direta ou indiretamente com este projeto.

*“Ele faria da queda um passo de dança
do medo uma escada, do sono uma ponte
da procura um encontro.”*

Fernando Sabino - O Encontro Marcado

Resumo

Genes envolvidos com a interação hospedeiro-patógeno são fortemente afetados pela seleção natural positiva. O gene codificante do antígeno sanguíneo Duffy, também conhecido como *DARC* (*Duffy Antigen Receptor for Chemokines*), tem um importante papel na invasão dos eritrócitos pelos parasitas causadores da malária, *Plasmodium vivax* em humanos e *Plasmodium knowlesi* em outros primatas. A estrutura do gene *DARC* já é conhecida, estando este presente na região 1q22–q23 do cromossomo 1, e sendo composto por dois éxons separados por um grande íntron. Em uma população africana a deleção de um nucleotídeo no domínio GATA-1 da região promotora do gene é responsável pela não expressão de *DARC* nos eritrócitos e pela resistência à invasão pelo *P. Vivax*. Além disso, o antígeno *DARC* age como um receptor promíscuo de quimiocinas, sendo expresso nos eritrócitos, células endoteliais de vênulas e outros tecidos. Devido a esse papel dual, no presente estudo seqüenciou-se regiões homólogas do gene *DARC* em macacos do Novo e Velho Mundo e utilizando métodos estatísticos procurou-se indícios da seleção natural positiva na sua história evolutiva. Nenhuma nova mutação foi encontrada no promotor ou na região codificante. As árvores filogenéticas pelos métodos de máxima parcimônia, máxima verossimilhança e *neighbor-join* apresentaram topologias semelhantes com três grande clados monofiléticos reconhecíveis e com a espécie *Macaca fascicularis* apresentando um perfil polifilético. O teste de seleção positiva pelos métodos de Nei-Gojobori, máxima verossimilhança por ramos e máxima verossimilhança por sítios não demonstraram, estatisticamente, a ação da seleção positiva sobre o gene *DARC*. Porém, o teste de máxima verossimilhança por sítio em domínios demonstrou que existem regiões do gene *DARC* sujeitas à diferentes pressões seletivas, mas também falhou em detectar a assinatura da seleção positiva. Os resultados indicam a presença da seleção darwiniana sobre a região de ligação do *P. vivax*, porém os testes de máxima verossimilhança utilizados, aparentemente, não possuem poder suficiente para detectar a sua assinatura. Além disso, os resultados sugerem que a região de ligação do *P. vivax* está sob influência de duas pressões seletivas antagônicas (seleção positiva exercida pelo parasita e seleção purificadora exercida pelas quimiocinas) o que pode, também, explicar a não detecção da seleção positiva.

Abstract

Molecular aspects of *DARC* gene evolution in primates

Genes involved in pathogen-host interactions are strongly affected by positive natural selection. The gene of blood Duffy antigen, also known as *DARC* (Duffy Antigen Receptor for Chemokines), has an important role in the invasion of red blood cells by the parasites that cause malaria, *Plasmodium vivax* in humans and *Plasmodium knowlesi* in other primates. The structure of the *DARC* gene is known, it was mapped in 1q22-q23 region of chromosome 1, and is composed by two exons separated by a large intron. In an African population a nucleic acid deletion in GATA-1 of the gene promoter is responsible for the non-expression of *DARC* on red blood cells and the resistance to invasion by *P. vivax*. Moreover, the *DARC* antigen acts as a promiscuous receptor for chemokines and is expressed in red blood cells, endothelial venules cells and other tissues. Because of this dual role, in this study we sequenced homologous regions of the *DARC* gene in monkeys of the New and Old World and using statistical methods we tried to detect positive natural selection in their evolutionary history. New mutations were not found at promoter or in coding region. The phylogenetic trees by the methods of maximum parsimony, maximum likelihood and neighbor-join showed similar topologies with three large monophyletic clades recognizable and with the *Macaca fascicularis* showing a polyphyletic profile. The test of positive selection by the methods of Nei-Gojobori, maximum likelihood by branches and maximum likelihood by sites not shown, statistically, the action of positive selection on the *DARC* gene. But the maximum likelihood test using sites divided in domains showed that some regions of the *DARC* gene are subject to different selective pressures, but also failed to detect the signature of positive selection. The results indicate the presence of darwinian selection on *P. vivax* binding region, but the maximum likelihood tests used, apparently, do not have enough power to detect its signature. Moreover, the results suggest that *P. vivax* binding region is under the influence of two opposing selective pressures (positive selection exerted by the parasite and purifying selection exerted by chemokines) that can also explain the non-detection of positive selection.

Sumário

Lista de Figuras

Lista de Tabelas

1	Introdução	p. 13
1.1	Grupo Sangüíneo Duffy	p. 13
1.1.1	Proteína Duffy	p. 14
1.1.2	Polimorfismo do gene <i>Duffy</i> em Humanos	p. 15
1.1.3	Aspectos Funcionais	p. 17
1.1.4	Receptor Antigênico Duffy para Quimiocinas	p. 18
1.1.5	Polimorfismo Duffy e Malária em Humanos	p. 19
1.2	Evolução Molecular	p. 20
1.2.1	Mutação	p. 20
1.2.2	Substituição de Nucleotídeos	p. 20
1.2.3	Modelos de Substituição	p. 20
1.2.3.1	Modelo JC69 ou modelo de equivalência de substituição (Jukes and Cantor, 1969)	p. 22
1.2.3.2	Modelo K80 ou modelo de dois parâmetros de Kimura (Kimura, 1980)	p. 22
1.2.3.3	Modelo F81 ou modelo proporcional	p. 22
1.2.3.4	Modelo HKY85 e F84	p. 23
1.2.3.5	Modelo TN93	p. 24
1.2.3.6	Modelo geral de reversão ao longo do tempo (GTR) . .	p. 24

1.2.4	Testes para detecção de seleção positiva	p. 25
1.2.4.1	Método de Nei-Gojobori	p. 25
1.2.4.2	Métodos por máxima verossimilhança	p. 26
1.2.4.3	Método por ramos	p. 27
1.2.4.4	Método por sítios	p. 27
2	Objetivos	p. 28
3	Material e métodos	p. 29
3.1	Obtenção das amostras	p. 29
3.2	Análise <i>in silico</i> do gene <i>DARC</i> e desenho dos <i>primers</i>	p. 30
3.2.1	Obtenção das regiões conservadas	p. 31
3.2.2	Desenho dos <i>primers</i>	p. 31
3.3	Amplificação genômica das amostras	p. 31
3.4	Amplificação por PCR	p. 33
3.5	Seqüenciamento de nucleotídeos	p. 33
3.6	Montagem das seqüências	p. 34
3.7	Identificação de sítios heterozigotos	p. 34
3.8	Alinhamento múltiplo global	p. 35
3.9	Análise Filogenética	p. 35
3.10	Testes para detecção de seleção positiva	p. 35
3.11	Visualização das estruturas protéicas 3-D	p. 36
4	Resultados	p. 37
4.1	Amplificação por PCR	p. 37
4.2	Seqüenciamento e montagem das seqüências	p. 37
4.3	Comparação das seqüências do gene <i>DARC</i> e estudos filogenéticos . . .	p. 39
4.3.1	Alinhamento de nucleotídeos	p. 39

4.3.2	Alinhamento de aminoácidos	p. 39
4.3.3	Árvore Filogenética	p. 41
4.3.4	Método de Nei-Gojobori	p. 47
4.3.5	Método de ramo por máxima verossimilhança	p. 47
4.3.6	Método de sítio por máxima verossimilhança	p. 50
4.3.7	Análise de domínios pelo método de máxima verossimilhança . .	p. 52
5	Discussão	p. 55
5.1	Comparações entre as seqüências	p. 55
5.1.1	Região GATA-1	p. 55
5.1.2	Cisteínas	p. 55
5.1.3	Fy ^{a/b}	p. 56
5.1.4	Fy ⁶	p. 56
5.2	Testes de seleção positiva	p. 56
5.2.1	Método de Nei-Gojobori	p. 56
5.2.2	Método de ramo por máxima verossimilhança	p. 57
5.2.3	Método de sítio por máxima verossimilhança	p. 57
5.2.4	Análise de domínios pelo método de máxima verossimilhança . .	p. 58
6	Conclusões	p. 60
	Referências	p. 61
	Apêndice A – Alinhamento global de nucleotídeos de todas as seqüências de primatas	p. 66
	Apêndice B – Alinhamento global de aminoácidos de todas as seqüências de primatas	p. 87

Lista de Figuras

1	Topologia de sete α -hélices da proteína Duffy.	p. 14
2	Representação esquemática do gene <i>Duffy</i> , seu RNA mensageiro e a glicoproteína Fy (gp-Fy).	p. 15
3	Alelos do gene <i>Duffy</i> responsáveis pelo fenótipo Fy(a - b -).	p. 16
4	Alelos funcionais do gene <i>Duffy</i> responsáveis pelo fenótipo Fy(a - b -) .	p. 17
5	Esquema representando as possíveis transições e transversões entre as quatro bases nitrogenadas da cadeia de DNA.	p. 21
6	Fluxograma mostrando as etapas realizadas na execução deste projeto.	p. 29
7	Representação esquemática da estratégia de bioinformática utilizada para o desenho dos <i>primers</i>	p. 32
8	Visualização da montagem realizada pelo software Phrap das seqüências de <i>Cebbus Apella</i>	p. 37
9	Gel de agarose mostrando a amplificação por PCR das amostras de primatas.	p. 38
10	Alinhamento gerado pelo BLAT de todas as seqüências de primatas em relação ao genoma humano.	p. 39
11	Alinhamento de nucleotídeos flanqueando a região do primeiro éxon do gene <i>DARC</i>	p. 40
12	Alinhamento dos primeiros 73 aminoácidos preditos a partir da seqüência nucleotídica dos primatas mostrando as regiões Fy ₆ , Fy _b e parte do domínio transmembrana (DT 1).	p. 41
13	Alinhamento do tipo <i>fingerprint</i> dos aminoácidos preditos a partir da seqüência nucleotídica dos primatas.	p. 42
14	Topologia da árvore filogenética gerada pelo método de máxima parcimônia.	p. 43

15	Topologia da árvore filogenética gerada pelo método de <i>neighbor-join</i> . .	p. 44
16	Topologia da árvore filogenética gerada pelo método de máxima verossimilhança.	p. 45
17	Árvore filogenética gerada pelo método de máxima verossimilhança mostrando o comprimento dos ramos.	p. 46
18	Árvore filogenética utilizada na análise de ramo por máxima verossimilhança.	p. 48
19	Estrutura tridimensional predita da proteína DARC humana. Os sítios sob seleção positiva preditos pelo modelo M3 estão em vermelho e rosa. A estrutura em branco indica sítios conservados.	p. 51
20	Estrutura tridimensional predita da proteína DARC humana. Os sítios mais conservados estão em azul escuro em direção gradativa aos sítios sob seleção positiva em vermelho.	p. 52
21	Esquema mostrando a divisão dos sítios (códon) do gene <i>DARC</i> em domínios para análise de seleção positiva pelo método de sítio.	p. 53

Lista de Tabelas

1	Amostras de primatas não humanos utilizados.	p. 30
2	Lista de <i>primers</i> desenhados.	p. 31
3	Protocolo utilizado na amplificação do gene <i>DARC</i>	p. 33
4	Programa de termociclagem utilizado na reação de PCR.	p. 33
5	Temperatura de anelamento de cada par de <i>primers</i>	p. 34
6	Códigos de ambigüidade utilizados para representar sítios heterozigotos.	p. 35
7	Resultados do teste de seleção positiva (d_N/d_S) pelo método de Nei-Gojobori.	p. 47
8	Valores do log de verossimilhança, estimativa dos parâmetros e estatística da relação de verossimilhança ($2\Delta\ln L$) para o modelo M0 (d_N/d_S único para todos os ramos) e para diferentes ramos da árvore filogenética (hipótese alternativa ou H_1).	p. 48
9	Valores do log de verossimilhança e estimativa dos parâmetros sob diferentes ramos da árvore filogenética com $\hat{\omega}_1 = 1$ (modelo neutro ou H_2).	p. 49
10	Valores do log de verossimilhança, estimativa dos parâmetros e estatística da relação de verossimilhança ($2\Delta\ln L$) para os modelos de sítio (M1a, M2a, M0, M3, M7 e M8).	p. 50
11	Valores do log de verossimilhança e estimativa dos parâmetros para os modelos de sítios fixos (análise por domínios da proteína)	p. 54

1 *Introdução*

Genes envolvidos em interações hospedeiro-patógeno são inferidos como fortemente afetados pela seleção natural positiva durante a evolução humana (NIELSEN et al., 2005). Um desses genes codifica o Receptor Antigênico Duffy para Quimiocinas (*Duffy Antigen Receptor for Chemokines*), conhecido também como DARC. Este antígeno é expresso em vários tecidos incluindo eritrócitos, células endoteliais de vênulas e células de Purkinje do cerebelo (HADLEY; PEIPER, 1997). Nos eritrócitos, ele serve como um receptor usado pelo *Plasmodium vivax* para localização e invasão (TOURNAMILLE et al., 1995). Em indivíduos não-imunes, a invasão por *P. vivax* leva à malária-vivax, que causa significativa morbidade sendo um problema de saúde mundial (MENDIS et al., 2001). O *P. vivax* é amplamente encontrado em regiões tropicais da Ásia, Centro-Oeste, ao Leste do Pacífico, América Central, América do Sul e na África (com baixa incidência na região central do cinturão tropical africano) (MENDIS et al., 2001). Em populações africanas sub-Saara encontra-se em alta frequência uma mutação T > C no motivo GATA da região promotora que impede a transcrição do antígeno causando uma ausência eritrócito-específica (TOURNAMILLE et al., 1995).

1.1 Grupo Sangüíneo Duffy

O produto do gene *Duffy* (denominando *FY*) é uma glicoproteína (gp-Fy) que atravessa a membrana plasmática sete vezes, possui um domínio N-terminal extracelular e um domínio C-terminal intracelular. O sistema consiste em quatro alelos, cinco fenótipos e cinco antígenos (LANGHI; BORDIN, 2006).

O nome do grupo sangüíneo Duffy foi dado por Cutbush e Mollison (1950) em homenagem ao homem hemofílico multi-transfusionado, no qual encontraram o novo anticorpo sérico (CHAUDHURI et al., 1995). Inicialmente, dois antígenos co-dominantes foram descritos e designados Fy^a e Fy^b , sendo estes codificados por dois alelos co-dominantes, nomeados *FYA* e *FYB* (CUTBUSH; MOLLISON, 1950).

Entre caucasianos os anticorpos anti-Fy^a e anti-Fy^b definem os fenótipos Fy(a + b -), Fy(a + b +) e Fy(a - b +) que geralmente representam os genótipos *FYA/FYA*, *FYA/FYB* e *FYB/FYB*, respectivamente. O fenótipo Fy(a - b -) é mais comum entre indivíduos negros (afro-americanos ou africanos ocidentais) (CHAUDHURI et al., 1995), sendo o resultado de um alelo silenciado associado com uma mutação -33 T > C (anteriormente localizado na região -46 por Tournamille et al. (1995)) no motivo GATA da região promotora que impede a transcrição do *FYB* em células vermelhas do sangue (POGO; CHAUDHURI, 2000).

1.1.1 Proteína Duffy

A glicoproteína (gp-FY) que expressa os antígenos Duffy foi inicialmente descrita por Moore, Woodrow e McClelland (1982). Hadley et al. (1984) demonstraram por técnicas de imunodeteção que o componente das células vermelhas do sangue que carrega os antígenos Duffy é uma proteína com um peso molecular de 35-43kDa, sensível a tratamento com quimotripsina e resistente ao tratamento com tripsina. O mapa de hidropatia prediz um domínio N-terminal extracelular de 60 resíduos, sete α -hélices transmembrana, pequenos “loops” hidrofílicos protraídos, e um domínio C-terminal de 28 resíduos (Figura 1) (CHAUDHURI et al., 1997).

Figura 1: Topologia de sete α -hélices da proteína Duffy. **A)** Modelagem molecular prediz que Duffy tem uma topologia similar a outros membros da família de receptores de quimiocinas. O domínio amino-terminal extracelular contém o sítio de ligação para anti-Fy⁶ e o polimorfismo que resulta nos grupos sangüíneos Fy^a e Fy^b. Este domínio também contém sequências necessárias para ligações multi-específicas de quimiocinas. **B)** Sequência de aminoácidos deduzida de Duffy; os primeiros sete aminoácidos N-terminais pertencem ao primeiro éxon e estão presentes no transcrito maior de Duffy; os aminoácidos restantes pertencem ao segundo éxon. Adaptado de Hadley e Peiper (1997).

A gp-FY é composta pelos epítomos Fy^a e Fy^b, em adição aos epítomos Fy³, Fy⁴,

Fy⁵ e Fy⁶. Os epítomos Fy^a e Fy^b estão localizados no domínio extracelular, no resíduo 42 e o epítomo Fy³ está localizado no terceiro “loop” extracelular. O epítomo Fy⁶ foi cuidadosamente mapeado em um hepta-peptídio englobando os resíduos 19–25 e está localizado entre dois sítios glicosilados (Figura 1) (WASNIOWSKA et al., 1996).

A proteína Duffy é codificada por um pequeno gene de cópia única (1,5kb, 2 éxons) localizado no cromossomo 1 e exibe uma homologia com o receptor IL-8 de humanos e coelhos. Duffy está organizada em sete domínios transmembrana, similar a outros membros de receptores de quimiocinas acoplados à proteína G (NEOTE et al., 1994). A maior (338 aa) e menor (336 aa) isoforma do polipeptídeo Duffy são codificadas por “*splicing*” (encadeamento) alternativo com presença ou perda de seis e oito aminoácidos NH₂-terminais, mas ligam-se à anticorpos anti-Fy e quimiocinas igualmente bem (LANGHI; BORDIN, 2006).

1.1.2 Polimorfismo do gene *Duffy* em Humanos

A base molecular dos polimorfismos do grupo sangüíneo Duffy já está compreendida. O gene *Duffy* (*FY*) está localizado na região 1q22-q23 do cromossomo 1 (MATHEW et al., 1994). O seu cDNA foi clonado originalmente de um RNA mensageiro da isoforma maior, e assumiu-se, a princípio, que *Duffy* era um gene sem íntrons (CHAUDHURI et al., 1993). Posteriormente descobriu-se que a unidade de transcrição do gene *Duffy* compreende 1781 nucleotídeos (nt), incluindo um primeiro éxon de 21 nt, um único íntron de 480 nt, e um segundo éxon de 990 nt (IWAMOTO et al., 1996) (Figura 2).

Figura 2: Representação esquemática do gene *Duffy* (topo), seu RNA mensageiro (meio), e gp-Fy (abaixo), com o polimorfismo *FYA/FYB* no nucleotídeo 125. Adaptado de Langhi e Bordin (2006).

O polimorfismo antigênico Fy^{a/b} deve-se a uma mutação pontual no nucleotídeo 125, no gene *FY*, que muda uma única base 125 G > A. Esta mudança resulta em uma diferença de um único aminoácido, mudando uma glicina por um ácido aspártico na posição 42 da

gp-Fy. Uma glicina na posição 42 (Gly⁴²) define o antígeno Fy^a enquanto o ácido aspártico (Asp⁴²) define o antígeno Fy^b (CHAUDHURI et al., 1993) (Figura 2).

A mutação que possibilita o fenótipo Fy(a - b -) em indivíduos negros está localizada na posição -33T > C da região promotora do alelo *FYB* que interrompe o sitio de ligação do fator de transcrição GATA-1, afetando totalmente a atividade transcricional em células eritróides com GATA box de indivíduos que expressam essa mutação em homozigose (POGO; CHAUDHURI, 2000) (Figura 3A).

Figura 3: Alelos do gene *Duffy* responsáveis pelo fenótipo Fy(a - b -). **A)** Alelo não transcrito encontrado em populações negras. **B)** Alelos transcritos mas não funcionais encontrados em populações caucasianas. Adaptado de Castilho (2007).

O mesmo fenótipo Fy(a - b -) identificado em caucasianos é causado por mutações pontuais do tipo substituição e deleção que produzem quatro alelos distintos (MALLINSON et al., 1995; RIOS et al., 2000) (Figura 3B).

Foi identificado em caucasianos o alelo *FYX* que também pode apresentar o fenótipo Fy(a - b -). Neste caso, o fenótipo silencioso é provocado por duas mutações no éxon 2 nas posições 265 (C > T) e 298 (G > A) que codificam, respectivamente, a mudança dos aminoácidos Arg89Cys e Ala100Thr (Figura 4). Quando a mutação Ala100Thr ocorre sozinha, não é observado nenhuma deficiência do gene *FY*. Porém, quando a mesma está acompanhada da mutação Arg89Cys observa-se uma redução da proteína na superfície celular levando a um fenótipo Fy(a - b -) atenuado (PARASOL et al., 1998; TOURNAMILLE et al., 1998; OLSSON et al., 1998). Esta redução atinge cerca de 90% da proteína, dificultando

a detecção do alelo Fy^x nos ensaios de aglutinação rotineiros (YAZDANBAKHS et al., 2000). Na população brasileira, uma terceira mutação foi encontrada na posição 145 G > T do alelo *FYB* que coexiste com as mutações do alelo *FYX*. Esta resulta uma substituição de Ala para Ser no resíduo 49 que mostra uma redução marcante da expressão de Fy^b , Fy^3 e Fy^6 (CASTILHO et al., 2004). A combinação de polimorfismos do gene *FY* determina três genótipos diferentes: *FYB-298A*, *FYB-265T* (*FYB*) e *FYB-145T*. A substituição 298A foi encontrada sem os outros SNPs, mas a substituição 265T ainda não foi encontrada sozinha (Figura 4B).

Figura 4: Alelos funcionais do gene *Duffy*. A) Alelos *FYA/FYB*. B) Alelos responsáveis pelo fenótipo $Fy(a-b-)$. Adaptado de Castilho (2007).

1.1.3 Aspectos Funcionais

Embora todos os grupos de antígenos sangüíneos sejam serologicamente detectáveis em células vermelhas do sangue, a maioria também é expressa em tecidos não eritróides, levantando questões sobre suas funções fisiológicas em condições normais e patológicas. Em adição a sua diversidade estrutural, grupos de antígenos sangüíneos também possuem ampla diversidade funcional, e podem se divididos em cinco classes: (a) transportadores e canais; (b) receptores para ligantes, vírus, bactérias e parasitas; (c) moléculas de adesão; (d) enzimas; e (e) proteínas estruturais (LANGHI; BORDIN, 2006).

Muitas moléculas presentes na superfície de células vermelhas do sangue, entre elas as que definem grupos sangüíneos, são receptores para vírus, bactérias e parasitas sugerindo que estes antígenos podem exercer um papel direto na patogênese de doenças infecciosas (GARRATTY, 1995; MOULDS; MOULDS, 2000).

Uma variedade de microorganismos reconhecem estruturas de carboidratos presentes

em glicolipídeos e glicoproteínas. Por exemplo, os ácidos siálico, que são abundantemente representados na glicoforina A (GPA), ou o motivo Cal α 1-4Gal partilhado pelos glicolipídeos P, Pk e P1 usados por várias cepas de bactérias e toxinas responsáveis por infecções do trato urinário superior (KARLSSON, 1989).

Os antígenos Duffy são carregados por uma glicoproteína de membrana exibindo duas propriedades biológicas interessantes:

- como um promíscuo receptor para quimiocinas de sub-famílias de peptídeos pró-inflamatórios nomeados de acordo com a estrutura do motivo conservado de cisteína: CC (RANTES, MCP-1) e CXC (IL8, mgSA). Por isso a proteína Duffy foi renomeada como DARC (“*Duffy Antigen/Receptor for Chemokine*”).
- como um receptor para *Plasmodium vivax* e *Plasmodium knowlesi* nos eritrócitos.

1.1.4 Receptor Antigênico Duffy para Quimiocinas

Foi estabelecido por Darbonne et al. (1991) que as células vermelhas do sangue possuem receptores para quimiocinas.

As quimiocinas constituem uma família de citocinas pró-inflamatórias capazes de ativar leucócitos e causar quimiotaxia, além das atividades angiogênicas e angiostáticas (ROSSI; ZLOTNIK, 2000). De acordo com o número e a localização de cisteínas aminoterminalis as quimiocinas foram classificadas em quatro famílias C, CC, CXC e CXXXXC. O efeito biológico de quimiocinas é mediado pela ligação e ativação de proteínas G acopladas (WELLS; PROUDFOOT; POWER, 1999). Algumas quimiocinas (classe CC e CXC) também ligam-se à glicoproteína do grupo sangüíneo Duffy, que agora é denominada DARC (LANGHI; BORDIN, 2006).

Dentre os receptores celulares de quimiocinas, CXCR4, CCR5 e DARC são os únicos que provaram agir como co-receptores de alguns de patógenos. O DARC para a entrada do parasita causador da malária *Plasmodium vivax* e o CCR5 para a entrada do vírus da imunodeficiência adquirida (HIV) (TAMASAUSKAS et al., 2001).

Horuk et al. (1996) observaram, após experimentos com receptores de quimiocinas em eritrócitos, uma perfeita correlação entre a ligação de quimiocinas eritrocitárias com a presença de antígenos DARC. Foi demonstrado que anticorpos anti-Fy inibiram a habilidade de ligação de quimiocinas em eritrócitos DARC-positivos. As quimiocinas uma vez ligadas a eritrócitos bloqueavam a invasão de parasitas maláricos nas células vermelhas

do sangue por usarem o antígeno DARC como receptores.

Quimiocinas CC e CXC ligam-se com DARC com alta afinidade, sugerindo algum papel em reações inflamatórias (HADLEY; PEIPER, 1997). Provavelmente o antígeno DARC pode desempenhar a função de “vassoura” na superfície das células sangüíneas vermelhas para eliminar o excesso de quimiocinas tóxicas produzidas em algumas situações patológicas (DARBONNE et al., 1991).

1.1.5 Polimorfismo Duffy e Malária em Humanos

A malária causada pelo *Plasmodium vivax* é a mais prevalente no mundo e atinge cerca de 70 a 80 milhões de casos por ano (MENDIS et al., 2001).

Miller et al. (1975) demonstraram que eritrócitos DARC-negativos Fy(a - b-) de humanos eram resistentes a invasão por *Plasmodium knowlesi*, um parasita malárico de macacos que sabe-se ser capaz de invadir células vermelhas de humanos e, em raros casos, infectar humanos. Mais adiante, anticorpos anti-Fy^a e anti-Fy^b bloquearam, respectivamente, a invasão de *P. knowlesi* em eritrócitos Fy(a + b -) e Fy(a - b +). A remoção dos determinantes antigênicos Fy^a e Fy^b da superfície das células vermelhas através do tratamento, respectivamente, com quimotripsina e pronase as tornaram resistentes à invasão (MILLER et al., 1975).

Estudos subseqüentes foram feitos com *P. vivax* e *Plasmodium falcíparum*. Embora, *P. vivax* seja amplamente espalhado pelo mundo tropical e subtropical, sendo o maior problema de saúde da Índia e sudeste asiático, o mesmo é ausente no Oeste Africano, onde mais de 95% da população é negativa para DARC. Miller et al. (1979) demonstraram que a resistência ao *P. vivax* correlaciona-se com o fenótipo DARC-negativo (Fy(a - b -)). Recentemente, Barnwell, Nichols e Rubinstein (1989) demonstraram que merozoítos de *P. vivax* eram incapazes de invadir células vermelhas do sangue DARC-negativas. Portanto, tanto o *P. vivax* como o *P. knowlesi* interagem com o antígeno DARC para penetrar nos eritrócitos. Deve-se notar que o *P. vivax* invade preferencialmente reticulócitos (GALINSKI et al., 1992), apesar de eritrócitos maduros, bem como os reticulócitos, expressarem um co-receptor, que junto com o antígeno DARC, é necessário para uma invasão por *P. vivax*. O agente ligante do parasita *P. vivax*, que se liga especificamente no reticulócito foi identificado (GALINSKI et al., 1992).

No final da década de 80 um novo antígeno, denominado de Fy⁶, foi identificado em pacientes infectados por *P. vivax* (BARNWELL; NICHOLS; RUBINSTEIN, 1989).

Em humanos, a presença ou ausência de Fy^6 correlaciona-se com a presença e ausência de Fy^a e Fy^b . Em outras palavras, células vermelhas do sangue $Fy(a +)$ ou $Fy(b +)$ são sempre Fy^6 positivas e as $Fy(a - b -)$ Fy^6 negativas. A invasão de células $Fy(a +)$ por *P. vivax* pode ser parcialmente bloqueada cobrindo-as com anti- Fy^a (MILLER et al., 1975), e totalmente bloqueada cobrindo-as com anti- Fy^6 (HADLEY; MILLER; HAYNES, 1991).

Análises funcionais e estruturais identificaram os domínios da proteína DARC que estão envolvidos na interação com diferentes ligantes. Foi demonstrado que a região amino-terminal extracelular de DARC, a qual carrega os grupos de antígenos $Fy^{a/b}$ e Fy^6 (WASNIOWSKA et al., 1996), é suficiente para interagir com as proteínas de ligação de *P. vivax* e de *P. knowlesi* (CHITNIS et al., 1996) e é responsável pelo perfil de receptor promíscuo de quimiocinas da proteína DARC. No entanto, a região amino-terminal não é suficiente para a ligação de quimiocinas, pois já foi previamente demonstrado que uma íntima associação entre os quatros domínios extracelulares de DARC através de pontes de dissulfeto é necessária para a criação de um envelope ativo para a ligação com quimiocinas (TOURNAMILLE et al., 1997).

Devido a sua propriedade biológica dual, DARC é potencialmente submetido à duas pressões seletivas: uma seleção purificadora interna para manutenção das propriedades de ligação de quimiocinas (tendo as quimiocinas endógenas como fator seletivo) e uma seleção positiva externa (sendo os parasitas o fator seletivo).

1.2 Evolução Molecular

1.2.1 Mutação

Mutações são erros na replicação ou reparo do DNA. De acordo com Li (1997) elas podem ser classificadas em função do tamanho da sequência de DNA afetada. Por exemplo, uma mutação pode afetar um único nucleotídeo (mutação pontual), ou vários nucleotídeos adjacentes. Mutações podem também ser classificadas pelo tipo de mudança causada pelo evento mutacional como (1) **substituição**, a troca de um nucleotídeo por outro; (2) **recombinação**, que inclui *crossing-over* e conversão gênica; (3) **deleção**, a remoção de um ou mais nucleotídeos; (4) **inserção**, adições de nucleotídeos à sequência; e (5) **inversão**, rotações de 180° de um segmento de DNA consistindo em dois ou mais pares de base.

1.2.2 Substituição de Nucleotídeos

A substituição de nucleotídeos pode ser classificada em **transições** e **transversões**. Transições são substituições entre bases A e G (purinas) ou entre bases C e T (pirimidinas) (Figura 5). Transversões são substituições entre purinas e pirimidinas (Figura 5).

Figura 5: Esquema representando as possíveis transições e transversões entre as quatro bases nitrogenadas da cadeia de DNA.

1.2.3 Modelos de Substituição

Um modelo de substituição de nucleotídeos descreve um processo pelo qual uma base ou sequência de bases é alterada para outra base ou sequência de bases. Uma vez que a taxa de substituições nucleotídicas e a história evolutiva são inferidas a partir deste processo, modelá-lo permite compreender melhor como ocorrem as substituições ao longo do tempo.

Diversos modelos de substituição de nucleotídeos foram postulados e novos modelos serão propostos devido ao crescente conhecimento sobre a evolução molecular. Aqui serão apresentados os princípios básicos de alguns modelos e de outros que foram estendidos a modelos mais complexos, alguns dos quais foram utilizados na elaboração desta dissertação.

A maioria desses modelos possuem as seguintes características em comum: são neutros, independentes e apresentam um modelo de sítios finito. A neutralidade significa que a seleção natural não opera nas substituições. A independência indica que a mudança em um sítio não afeta a probabilidade da mudança em outro sítio (Cadeia de Markov com Tempo Contínuo) e a presença de sítios finitos indica a possibilidade de substituições múltiplas no mesmo sítio.

Por ser um processo Markoviano com Tempo Contínuo os modelos de substituição podem ser representados por matrizes de transição que são parametrizadas pelo tempo

(t). Se denominarmos $E_1, \dots, E_4$ como estados, então temos a matriz de transição $P(t) = (P_{ij}(t))$, onde entrada individual, $P_{ij}(t)$ refere à probabilidade de um estado E_i mudar para o estado E_j no tempo t . A matriz de transição para seqüências de DNA pode ser representada como:

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{TT}(t) \end{pmatrix} \quad (1.1)$$

1.2.3.1 Modelo JC69 ou modelo de equivalência de substituição (Jukes and Cantor, 1969)

Jukes e Cantor (1969) propuseram um modelo bastante simples onde os nucleotídeos A, C, G e T em uma seqüência de DNA ocorrem em freqüências iguais, e a probabilidade de substituição de um nucleotídeo E_i para um nucleotídeo E_j depende simplesmente da taxa de substituição instantânea u , estimada a partir dos dados. Matematicamente, podemos representar este modelo de maneira simplificada como:

$$P_{ij}(dt) = udt \quad (1.2)$$

Onde dt representa o intervalo de tempo decorrido. Assim, todos os elementos da matriz Pt de substituição dada acima seriam calculados como u .

1.2.3.2 Modelo K80 ou modelo de dois parâmetros de Kimura (Kimura, 1980)

A observância de que transições ocorrem mais freqüentemente do que transversões, permitiu Kimura (1980) desenvolver um modelo onde a taxa u de substituição é convertida nas taxas de transição α e de transversão β . Desta maneira:

$$P_{ij}(dt) = \begin{cases} \alpha & \text{para transição} \\ \beta & \text{para transversão} \end{cases} \quad (1.3)$$

E sua matriz:

$$Pt = \begin{pmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{pmatrix} \quad (1.4)$$

1.2.3.3 Modelo F81 ou modelo proporcional

A frequência dos quatro nucleotídeos, contudo, nem sempre é similar. O genoma mitocondrial de vertebrados por exemplo, apresenta uma redução na proporção de G em relação às demais bases. Assim, Felsenstein (1981) elaborou o seguinte modelo que leva em conta a frequência dos nucleotídeos:

$$P_{ij}(dt) = u\pi_j dt \quad (1.5)$$

Onde π_j representa a frequência de nucleotídeo j . E sua matriz pode ser representada como:

$$Pt = \begin{pmatrix} . & u\pi C & u\pi G & u\pi T \\ u\pi A & . & u\pi G & u\pi T \\ u\pi A & u\pi C & . & u\pi T \\ u\pi A & u\pi C & u\pi G & . \end{pmatrix} \quad (1.6)$$

1.2.3.4 Modelo HKY85 e F84

Os modelos K80 e F81 foram combinados concomitantemente por Hasegawa, Kishino e Yano (1984, 1985) e por Felsenstein (em seu pacote de softwares PHYLIP) em outros dois modelos denominados, respectivamente, HKY85 e F84. Esses modelos combinam as diferenças nas taxas de transição e transversão com a frequência desigual das bases, tornando estes modelos mais realísticos. Pode-se representar o modelo HKY85 como:

$$P_{ij}(dt) = \begin{cases} \alpha\pi_j dt & \text{para transição} \\ \beta\pi_j dt & \text{para transversão} \end{cases} \quad (1.7)$$

E sua matriz:

$$Pt = \begin{pmatrix} . & \beta\pi C & \alpha\pi G & \beta\pi T \\ \beta\pi A & . & \beta\pi G & \alpha\pi T \\ \alpha\pi A & \beta\pi C & . & \beta\pi T \\ \beta\pi A & \alpha\pi C & \beta\pi G & . \end{pmatrix} \quad (1.8)$$

A diferença entre este modelo e o F84, é o emprego do parâmetro k que determina a razão transição/transversão, e do parâmetro $\prod j$ representado pela frequência de pirimidinas ($\pi_C + \pi_T$) se j for C ou T e purinas ($\pi_A + \pi_G$) se j for A ou G. Matematicamente esse modelo é representado:

$$P_{ij}(dt) = \begin{cases} \left(\frac{k}{\prod j} + 1 \right) & \text{para transição} \\ u\pi_j dt & \text{para transversão} \end{cases} \quad (1.9)$$

1.2.3.5 Modelo TN93

A diferença na composição de bases reflete diferenças não apenas nas taxas de transições e transições, mas também na taxa de transição entre purinas e na taxa de transição entre pirimidinas. Para adequar essa diferença a modelos de evolução de seqüências, Tamura e Nei (1993) propuseram o modelo:

$$P_{ij}(dt) = \begin{cases} \alpha_R \pi_j dt & \text{para transição entre purinas} \\ \alpha_Y \pi_j dt & \text{para transição entre pirimidinas} \\ \beta \pi_j dt & \text{para transversão} \end{cases} \quad (1.10)$$

Onde α_R e α_Y representam, respectivamente, a taxa de transição entre purinas e a taxa de transição entre pirimidina. Sua matriz é representada como:

$$Pt = \begin{pmatrix} . & \beta\pi C & \alpha_R \pi G & \beta\pi T \\ \beta\pi A & . & \beta\pi G & \alpha_Y \pi T \\ \alpha_R \pi A & \beta\pi C & . & \beta\pi T \\ \beta\pi A & \alpha_Y \pi C & \beta\pi G & . \end{pmatrix} \quad (1.11)$$

1.2.3.6 Modelo geral de reversão ao longo do tempo (GTR)

Um modelo geral de substituição foi desenvolvido por Rodríguez et al. (1990), permitindo que a taxa de substituição de um nucleotídeo para cada um dos demais pudesse ser

considerada, levando-se em conta a sua frequência e a taxa de substituição.

Nesse caso, a matriz de substituição é construída considerando as frequências dos nucleotídeos e a probabilidade de uma base ser substituída por outra. Para cada uma das possíveis mudanças de nucleotídeos, há uma probabilidade específica associada:

$$Pt = \begin{pmatrix} . & a\pi C & b\pi G & c\pi T \\ a\pi A & . & d\pi G & e\pi T \\ b\pi A & d\pi C & . & f\pi T \\ c\beta\pi A & e\pi C & f\pi G & . \end{pmatrix} \quad (1.12)$$

Onde a, b, c, d, e e f são as probabilidades específicas para cada mudança de nucleotídeo.

1.2.4 Testes para detecção de seleção positiva

Detectar a seleção positiva Darwiniana é um aspecto crítico para entender os mecanismos da evolução molecular. Os testes existentes propostos pela genética de populações são poderosos o suficiente para rejeitar o modelo neutro. No entanto, tais testes geralmente não são suficientes para discernir diferentes formas de seleção natural ou detectar evolução molecular adaptativa (YANG; BIELAWSKI, 2000). Um método poderoso para detectar a seleção positiva é através da comparação das taxas de substituições sinônimas e não sinônimas. A pressão seletiva ao nível protéico é mensurada por $\omega = d_N/d_S$, onde d_N e d_S são, respectivamente, as taxas de substituições não sinônimas e sinônimas. Se as mudanças de aminoácidos são vantajosas, eles serão fixados em uma taxa maior que as mudanças sinônimas, com $d_N > d_S$. Então, uma alta taxa de substituições não sinônimas ($\omega > 1$) é evidência de evolução molecular adaptativa. Se as mudanças de aminoácidos são deletérias, a seleção purificadora reduzirá sua taxa de fixação, assim $d_N < d_S$ e $\omega < 1$. Mutações neutras resultam em $\omega = 1$, pois a seleção sobre a proteína não tem efeito adaptativo.

1.2.4.1 Método de Nei-Gojobori

No método de Nei-Gojobori, d_S e d_N são estimados computando o número de substituições sinônimas e não sinônimas e o número de sítio potencialmente sinônimos e não sinônimos. Os sítios potencialmente sinônimos e não sinônimos são calculados para cada códon sob a premissa de igual probabilidade de mudança para todos os nucleotídeos. Denota-se por f_i a proporção de mudanças sinônimas (a razão do número de mudanças

sinônimas pela soma de mudanças sinônimas e não sinônimas, excluindo mutações *non-sense*) na i -ésima posição nucleotídica de um códon ($i = 1, 2, 3$). O número de potenciais sítios sinônimos (s) e não sinônimos (n) para este códon é dado por $s = \sum_{i=1}^3 f_i$ e $n = 3 - s$, respectivamente.

Para obter o número total de sítios sinônimos (S) e não sinônimos (N) para toda a seqüência, utiliza-se a fórmula $S = \sum_{j=1}^C s_j$ e $N = 3C - S$, onde s_j é o valor de S para o j -ésimo códon e C é o número total de códons. Na prática se compara duas seqüências, então o valores médios de S e N para as duas seqüências são utilizados no cálculo. Note que $S + N = 3C$ é igual ao número total de nucleotídeos comparados.

1.2.4.2 Métodos por máxima verossimilhança

Goldman e Yang (1994) sugeriram um modelo de substituição de códons que é similar aos modelos de substituição nucleotídica (especialmente o HKY85) mas considerando o códon como a unidade de evolução. O modelo original usava distâncias químicas dos aminoácidos para modificar sua taxa de substituição, mas verificou-se que o modelo não representava bem os dados. Assim, o modelo Markoviano de substituições de códons propostos por Goldman e Yang (1994) foi modificado por Nielsen e Yang (1998) permitindo razões ω heterogêneas entre os sítios e ignorando o fato de que alguns aminoácidos são quimicamente próximos enquanto outros são muito diferentes. Abaixo encontra-se uma visão geral do método.

Seja h denotando um sítio em uma seqüência e N denotando o número de códons na seqüência ($h = 1, 2, \dots, N$). A taxa de substituição instantânea relativa do códon i para o códon j ($i \neq j$) no sítio h é dado por:

$$q_{ij}^{(h)} = \begin{cases} 0, & \text{se } i \text{ e } j \text{ diferem em duas ou três posições nucleotídicas} \\ \pi_j, & \text{se } i \text{ e } j \text{ diferem em uma transversão sinônima} \\ \kappa\pi_j, & \text{se } i \text{ e } j \text{ diferem em uma transição sinônima} \\ \omega_{(h)}\pi_j, & \text{se } i \text{ e } j \text{ diferem em uma transversão não sinônima} \\ \omega_{(h)}\kappa\pi_j, & \text{se } i \text{ e } j \text{ diferem em uma transição não sinônima} \end{cases} \quad (1.13)$$

Onde, π_j é a frequência de equilíbrio do códon j , κ é a razão da taxa de transição/transversão, e $\omega^{(h)}$ é a razão d_N/d_S no sítio h . A matriz com as probabilidades de transição sobre o tempo t é dada por $P(t) = e^{Qt}$, onde $Q = \{q_{ij}^{(h)}\}$.

A frequência de equilíbrio do códon j (π_j) pode ser considerada um parâmetro livre, mas também pode ser calculada a partir das frequências nucleotídicas nas três posições do códon. Sob este modelo, a relação assegura que $\omega = d_N/d_S$, a razão das taxas de substituições não sinônimas/sinônimas.

A razão ω é a mensuração da seleção natural agindo sobre a proteína. Simplificando, valores de $\omega < 1$, $= 1$, e > 1 significam, respectivamente, seleção purificadora negativa, evolução neutra e seleção positiva.

1.2.4.3 Método por ramos

Os modelos por ramos (*branch models*) permitem a razão ω variar entre os ramos na filogenia e são úteis para detectar seleção positiva agindo em linhagens particulares.

1.2.4.4 Método por sítios

Os modelos de sítio (*site models*) permite a razão ω variar entre os sítios (entre códon ou aminoácidos na proteína) e são úteis para detectar a seleção positiva agindo sobre sítos específicos.

2 *Objetivos*

Objetivou-se seqüenciar a região promotora e toda a região codificante do gene *DARC* em diversas espécies de macacos do novo e velho mundo. A partir disso, definiu-se mais três objetivos.

Primeiramente, descobrir possíveis novas mutações que possam afetar a expressão de *DARC* analisando à região promotora.

Em segundo lugar, descobrir mutações na região codificadora do gene *Duffy* que alterem ou removam aminoácidos de regiões críticas da proteína DARC.

Por último, as informações coletadas permitirão examinar se a seleção natural positiva agiu sobre a história evolutiva do antígeno DARC usando métodos estatísticos que comparem mutações sinônimas e não sinônimas entre as espécies.

3 *Material e métodos*

A Figura 6 descreve a seqüência dos métodos empregados neste projeto.

Figura 6: Fluxograma mostrando as etapas realizadas na execução deste projeto.

3.1 Obtenção das amostras

Um total de 11 amostras de DNA de primatas não humanos utilizadas fazem parte de nosso estoque de DNA e foram obtidas em estudos anteriores (Tabela 1). Outras 11 amostras de DNA foram cedidas pelo nosso colaborador Prof. Dr. Eugene E. Harris da Queensborough Community College de Nova Iorque, Estados Unidos (Tabela 1). Essas 22 amostras foram amplificadas e seqüenciadas no Laboratório de Genética Molecular e Bioinformática (LGMB).

Outras 23 amostras foram seqüenciadas pelo Prof. Dr. Eugene E. Harris (Tabela 1). O protocolo para a obtenção das seqüências foi o mesmo utilizado pelo LGMB. Todas as seqüências foram utilizados em nossas análises.

Tabela 1: Amostras de primatas não humanos utilizados. As duas primeiras colunas denominam, respectivamente, o local onde foram seqüenciadas e a proveniência das amostras.

Seqüenciamento	Proveniência	Espécie	N.º de amostras
LGMB	LGMB	<i>Macaca mulatta</i>	2
		<i>Macaca fascicularis</i>	3
		<i>Saimiri ustis</i>	2
		<i>Saimiri sciureus</i>	2
		<i>Cebbus apella</i>	1
		<i>Gorilla gorilla</i>	1
		Subtotal	11
	Prof. Dr. E. Harris	<i>Cercocebus torquatus lunatus</i>	2
		<i>Lophocebus aterrimus</i>	1
		<i>Cercocebus galeritus chrysogaleus</i>	1
		<i>Lophocebus albigena</i>	1
		<i>Mandrillus leucophaeus</i>	1
		<i>Macaca sp.</i>	1
		<i>Macaca mulatta</i>	1
		<i>Cercopithecus mitis</i>	2
		<i>Mandrillus sphinx</i>	1
		Subtotal	11
Prof. Dr. E. Harris	Prof. Dr. E. Harris	<i>Macaca mulatta</i>	4
		<i>Macaca fascicularis</i>	11
		<i>Macaca nemestrina</i>	4
		<i>Macaca nigra</i>	3
		<i>Macaca thibetana</i>	1
	Subtotal	23	
Total		45	

3.2 Análise *in silico* do gene *DARC* e desenho dos primers

O desenho dos *primers* foi feito a partir de regiões conservadas do gene *DARC* devido a diversidade de espécies utilizadas no estudo. Para definir as regiões conservadas foram utilizadas as seqüências cromossômicas referentes ao gene *DARC* dos seguintes primatas presentes no Genbank: *Homo sapiens*, *Macaca mulatta* e *Pan troglodytes*. Existem seqüências de vários outros macacos para o gene *DARC* no Genbank, porém os mesmos possuem somente a região codificante do gene. Como objetivou-se seqüenciar o gene inteiro, incluindo as regiões 5' *upstream*, promotor e 3' *downstream*, foram utilizadas somente as seqüências dos organismos que possuíam o genoma completamente seqüenciado.

3.2.1 Obtenção das regiões conservadas

As seqüências de *H. sapiens*, *P. troglodytes* e *M. mulatta* foram alinhadas utilizando o software de alinhamento múltiplo ClustalW (LARKIN et al., 2007) com os parâmetros padrões. Foi desenvolvido um script na linguagem Perl para encontrar as regiões conservadas a partir do alinhamento realizado pelo ClustalW. O resultado gerado pelo script é uma única seqüência com as regiões conservadas, tendo as regiões não conservadas mascaradas com a letra “N” (Figura 7).

3.2.2 Desenho dos *primers*

Os *primers* foram desenhados com o software Primer3, utilizando a seqüência conservada do gene *DARC* como entrada. Foram utilizados os parâmetros padrões do Primer3. O tamanho máximo dos fragmentos estabelecido foi de 500 à 550 nucleotídeos. Foram desenhados oito pares de *primers* internos de aproximadamente 500 nucleotídeos (Figura 7 e Tabela 2).

Tabela 2: Lista de *primers* desenhados.

Primer	Posição	Seqüência	Primer	Posição	Seqüência
DARCP1F	1	GCTGTCCCATTGTCCCCTAG	DARCP1R	458	ATAAGGGAGGTGCTGGGAAA
DARCP2F	367	TGAACTGCCTTTCTTGGAT	DARCP2R	807	CTGTCAGCGCCTGTGCTT
DARCP3F	761	CTTCCTGATGCCCCCTGT	DARCP3R	1209	CAGGAACAGTCTGGGGAAG
DARCP4FSaimiri	1151	CCTGAGCTTGCATGGCTCCC	DARCP4RSaimiri	1709	GGCACCACGATGCTGAAGA
DARCP4F	1151	TCTGACCTTGCACTGCTCCT	DARCP4R	1709	GGCACCACAATGCTGAAGAG
DARCP5F	1550	CACTGCCCTTCTTCATCCTC	DARCP5R	2081	CCCATACCCAATGCCTTCTT
DARCP6F	1972	GAGCTGAAGGCTTTCAGG	DARCP6R	2446	AATAAGAAACCACCGCTTCA
DARCP7F	2384	CTTCCCACCTGTCAACCTGA	DARCP7R	2832	CAAAGCCAGAGAGGAGACAGAA
DARCP8F	2750	GATGACAAAGCCCAGAGCAA	DARCP8R	2955	GGAACCTGGGAGAGATGAC

3.3 Amplificação genômica das amostras

Devido à pouca quantidade de DNA das amostras, todas foram submetidas à amplificação genômica total (“*Whole Amplification Genome*”) pelo kit GenomiPhi da GE Healthcare Bio-Sciences. Para cada amostra utilizou-se 10 ng de DNA original seguindo o protocolo do fabricante.

Desenho dos primers do gene DARC

Alinhamento múltiplo das seqüências de *H. sapiens*, *P. troglodytes* e *M. mulatta* para encontrar regiões conservadas.

ClustalW

[illegible]

Script em linguagem Perl retira uma sequência consenso a partir do alinhamento múltiplo. Regiões não conservadas são substituídas pela letra "N"

ClustalParser.pl

[illegible]

Primers com aproximadamente 500pb são desenhados a partir da sequência consenso.

Primer3

Figura 7: Representação esquemática da estratégia de bioinformática utilizada para o desenho dos *primers*.

3.4 Amplificação por PCR

O gene *DARC* foi amplificado através da reação em cadeia da polimerase (*Poly Chain Reaction* ou PCR) utilizando os pares de *primers* descritos na Tabela 2. A reação de PCR foi realizada usando o protocolo presente na Tabela 3.

Tabela 3: Protocolo utilizado na amplificação do gene *DARC*.

Reagentes	Volume (μL)
Buffer Biotools	2,5
dNTP	2,0
Primer F	1,5
Primer R	1,5
TAQ Biotools	1
DNA	1
H ₂ O MilliQ	15,5
Volume final	25

A reações fora submetidas ao programa de termociclagem descrito na Tabela 4. As temperaturas de anelamento de cada *primer* encontram-se na Tabela 5.

Tabela 4: Programa de termociclagem utilizado na reação de PCR.

Temperatura ($^{\circ}\text{C}$)	Tempo (s)	Repetições
94	240	1
TA*	120	
72	50	35
94	40	
TA*	40	
72	600	1
4	∞	

* TA é a temperatura de anelamento de cada par de *primers* presente na Tabela 5

3.5 Seqüenciamento de nucleotídeos

O método de seqüenciamento de DNA foi realizado em um seqüenciador automatizado da marca MEGABACE 1000 da Amersham/GE Healthcare Bio-Sciences.

Tabela 5: Temperatura de anelamento de cada par de *primers*.

<i>Primer</i>	Temperatura (°C)
DARCp1	59
DARCp2	58
DARCp3	59
DARCp4	57
DARCp4Saimiri	57
DARCp5	58
DARCp6	58
DARCp7	61
DARCp8	58

3.6 Montagem das seqüências

A montagem das seqüências foi feita utilizando os softwares Phred (EWING; GREEN, 1998; EWING et al., 1998), Phrap (GORDON; DESMARAIS; GREEN, 2001) e Consed (GORDON; ABAJIAN; GREEN, 1998). O programa Phred é responsável pelo “*base-calling*”, isto é, analisar os cromatogramas e verificar se a base atribuída está correta. O programa Phrap é responsável pela montagem das seqüências. O programa Consed é utilizado como visualizador e editor das seqüências geradas pelo Phrap. Para orientar a montagem foram utilizadas seqüências de referência (método também conhecido como montagem ancorada) retiradas do NCBI no formato FASTA. Durante a montagem foi utilizada a seqüência genômica de *M. mullata* como referência para todos os macacos do velho mundo (*Macaca sp.*, *Cercocebus sp.*, *Mandrillus sp.*, *Lophocebus sp.*, *Cercopithecus sp.*) e a seqüência de *Saimiri boliviensis* (acession number: AF3119180) como referência para as amostras do gênero *Saimiri*. As espécies *Cebbus apella* e *Gorilla gorilla* foram montadas sem seqüências de referência.

Foram desenvolvidos alguns scripts em linguagem Perl para auxiliar na montagem das seqüências. Estes foram responsáveis por separar os cromatogramas em diretórios com o nome de cada amostra, executar o phred, analisar a qualidade dos cromatogramas e executar o phrap. A importância desses scripts residiu na automação e na rapidez com as quais as seqüências puderam ser analisadas.

3.7 Identificação de sítios heterozigotos

A identificação de sítios heterozigotos à partir dos cromatogramas foi feita utilizando-se os softwares PolyPhred 6.11 (NICKERSON; TOBE; TAYLOR, 1997) e Polyscan 3.0 (CHEN et al., 2007). Os resultados de ambos foram inspecionados manualmente através do programa Consed e os sítios que apresentavam perfis cromatográficos compatíveis com heterozigose foram substituídos por símbolos de ambigüidade definidos pela IUPAC (Tabela 6).

Tabela 6: Códigos de ambigüidade utilizados para representar sítios heterozigotos.

Código	Descrição	Código	Descrição
R	Purina (A ou G)	K	T,U ou G
Y	Pirimidina (C,T ou U)	W	T, U ou A
M	C ou A	S	C ou G

3.8 Alinhamento múltiplo global

O alinhamento múltiplo das seqüências foi obtido utilizando o software ClustalW (LARKIN et al., 2007) com os parâmetros pré definidos e considerando o alinhamento dos códons. Todos os alinhamentos mostrados nessa dissertação foram formatados utilizando o pacote $\text{\LaTeX}$ shade (BEITZ, 2000) para $\text{\LaTeX}$. O mesmo foi utilizado para gerar o alinhamento do tipo *fingerprint*, assim como a hidropatia dos aminoácidos.

3.9 Análise Filogenética

As árvores filogenéticas pelo método de máxima verossimilhança foram construídas utilizando o software PHYML (GUINDON; GASCUEL, 2003). Para o método de máxima parcimônia utilizou-se o software MEGA 4.1 (TAMURA et al., 2007). A filogenia pelo método de *neighbor-join* foi feita com o software NEIGHBOR integrante do pacote PHYLIP (FELSENSTEIN, 2005). Todos os métodos utilizaram um *bootstrap* com 1000 réplicas. O modelo de substituição nucleotídica (HKY85 para máxima verossimilhança e F84 para *neighbor-join*) e os demais parâmetros (α , κ , etc..) foram computados a partir das seqüências pelo software ModelGenerator v0.84 (KEANE et al., 2006). Utilizou-se o software MEGA 4.1 para gerar as Figuras 14, 15, 16, 17 e 18 a partir dos arquivos em formato Newick.

3.10 Testes para detecção de seleção positiva

Todos os testes de seleção positiva (inclusive o método de Nei-Gojobori) foram feitos utilizando o software codeml integrante do pacote PAML v.4 (YANG, 2007). Os *gaps* e deleções foram tratados como códigos de ambigüidade (opção *cleandata=0*).

Todas as análises foram feitas em um servidor Dell com 2 processadores Xeon de 3GHz e 2GB de memória RAM. As análises, de acordo com sua complexidade, demoraram desde alguns minutos à mais de 4 horas de processamento.

3.11 Visualização das estruturas protéicas 3-D

A visualização dos aminoácidos selecionados positivamente na estrutura tridimensional do gene *DARC* foi feita com o software Rasmol 2.7.4.1 (SAYLE; MILNER-WHITE, 1995; BERNSTEIN, 2000). Utilizou-se o arquivo RasMol.txt gerado pelo software PAML (método de sítio M3) contendo as probabilidades posteriores de ω para cada códon. O gene DARC não possui sua estrutura protéica tridimensional cristalografada, por isso utilizamos sua estrutura predita (UniProt *accession number* Q5Y7A4).

4 Resultados

4.1 Amplificação por PCR

Os oito pares de *primers* desenhados com base nas regiões conservadas funcionaram para todas as amostras de primatas exceto para o gênero *Saimiri*. Para este, o quarto par de *primers* (DARCp4) foi redesenhado a partir da sequência de *Saimiri boliviensis* retirada do Genbank. A Figura 9 mostra o resultado da amplificação por PCR de todas as amostras de primatas. O oitavo par de *primers* não foi usado pois adiciona somente 100pb à região a ser seqüenciada, sendo sua utilização dispensável pois abrange uma região muito posterior ao último éxon do gene *DARC*.

4.2 Seqüenciamento e montagem das seqüências

Foram seqüenciadas pelo menos três vezes cada par de *primers*, esta abordagem foi necessária para garantir uma alta resolução das seqüências (Figura 8). Ao final, gerou-se mais de 1000 cromatogramas e o gene *DARC* inteiro foi montado para as 45 amostras seqüenciadas. As seqüências das amostras foram alinhadas contra o genoma humano e confirmaram sua homologia com o gene *DARC* (Figura 10).

Figura 8: Visualização da montagem realizada pelo software Phrap das seqüências de *Cebbus Apella*. Observa-se uma alta definição devido a sobreposição das seqüências.

Figura 9: Gel de agarose mostrando a amplificação por PCR das amostras de primatas utilizando os sete pares de *primers* internos.

Figura 10: Alinhamento gerado pelo BLAT de todas as seqüências de primatas em relação ao genoma humano. Todas as seqüências alinharam na região cromossômica correspondente ao gene *DARC* em humanos.

4.3 Comparação das seqüências do gene *DARC* e estudos filogenéticos

4.3.1 Alinhamento de nucleotídeos

A Figura 11 mostra parte do alinhamento das seqüências flanqueando parte da região promotora, o primeiro éxon e parte do íntron do gene *DARC*. No alinhamento os Homínidae e os macacos do velho mundo compartilham, na região 5' UTR, deleções de um e sete nucleotídeos, respectivamente. Os macacos do velho mundo compartilham o códon CCG (prolina) na sétima posição do primeiro éxon, os macacos do novo mundo possuem o códon CAG (glutamina), *G. gorilla* possui o códon ACG (treonina) enquanto *H. sapiens* apresenta AGG (arginina).

O alinhamento completo das seqüências pode ser visto no Apêndice A.

4.3.2 Alinhamento de aminoácidos

Todas as seqüências nucleotídicas foram traduzidas e suas seqüências protéicas alinhadas. O alinhamento dos primeiros 73 aminoácidos do gene *DARC* mostrando as regiões Fy₆, Fy_{a/b} e o primeiro domínio transmembrana encontra-se na Figura 12. A região Fy₆ possui várias mudanças de aminoácidos entre as espécies. Existe uma mudança de ácido aspártico (21D) presente nos Homínidae para asparagina (21N) nos demais primatas. O mesmo ocorre na posição 22 dos Homínidae, com F > S nos macacos do velho mundo e

Figura 11: Alinhamento de nucleotídeos flanqueando a região do primeiro éxon do gene *DARC*.

F > L nos macacos do novo mundo. Alguns macacos do velho mundo apresentam uma deleção de ácido glutâmico na posição 23. Na posição 25 observam-se várias mudanças de aminoácidos: valina (V) em *H. sapiens*, alanina (A) em *G. gorilla*, serina (S) em *M. nigra* e leucina (L) nos demais primatas.

A região Fy_{a/b} mostra a substituição de um ácido aspártico por uma glicina na posição 42 em humanos, caracterizando o alelo Fy^a. As posições com a letra "X" no alinhamento indicam códons com códigos de ambigüidade, impossibilitado a sua tradução. O alinhamento completo das seqüências de aminoácidos encontra-se no Apêndice B.

A Figura 13 mostra o alinhamento do tipo *fingerprint* exibindo a hidropatia dos aminoácidos. Este tipo de alinhamento é interessante para uma visão geral da similaridade entre as seqüências. Observa-se que apesar de várias mudanças de aminoácidos, as propriedades químicas dos mesmos encontram-se conservadas entre as seqüências.

A filogenia pelo método de máxima parcimônia diferiu das demais, mostrando que *Cercocebus galeritus chrysogaleus* está mais próximo de *Lophocebus aterrimus* do que *Cercocebus torquatus lunatus* (Figura 14). Todos os métodos de construção filogenética agruparam os primatas em 3 grandes grupos monofiléticos: Platyrrhini (macacos do Novo Mundo), Cercopithecidae (macacos do Velho Mundo) e Hominidae (*H. sapiens*, *G. gorilla*)(Figuras 14, 15 e 16). Além disso, em todas as árvores a espécie *M. fascicularis* formou grupos polifiléticos com alguns indivíduos mais próximos de *M. mulatta*. Os demais táxons formaram grupos monofiléticos de acordo com seu gênero (Figuras 14, 15 e 16).

A Figura 17 mostra a árvore filogenética por máxima verossimilhança mostrando, além da topologia, o comprimento dos ramos. Observa-se uma alta proximidade entre os táxons do clado Cercopithecidae o que pode explicar, em parte, a polifilia de *M. fascicularis*.

Figura 13: Alinhamento do tipo *fingerprint* dos aminoácidos preditos a partir da seqüência nucleotídica dos primatas. Os sete domínios transmembrana são mostrados na caixa vermelha no topo. As caixas verdes e pretas no topo indicam, respectivamente, as regiões externas e as regiões citoplasmáticas.

Figura 14: Topologia da árvore filogenética gerada pelo método de máxima parcimônia.

Figura 15: Topologia da árvore filogenética gerada pelo método de *neighbor-join*.

Figura 16: Topologia da árvore filogenética gerada pelo método de máxima verossimilhança.

Figura 17: Árvore filogenética gerada pelo método de máxima verossimilhança mostrando o comprimento dos ramos.

4.3.4 Método de Nei-Gojobori

A Tabela 7 mostra a razão d_N/d_S par à par entre todas as espécies de primatas obtida pelo método de Nei-Gojobori. O valor da razão d_N/d_S para cada espécie é a média aritmética obtida da razão d_N/d_S de todos os indivíduos pertencentes à mesma.

Levando em conta a região codificante completa do gene *DARC* a razão d_N/d_S é relativamente baixa (média 0,38 e desvio padrão 0,2) com valores entre 0 e 1,13 (Tabela 7). Algumas comparações par à par apresentam valores maiores que 1, porém não são discrepantes e não indicam uma forte ação da seleção positiva.

Tabela 7: Resultados do teste de seleção positiva (d_N/d_S) pelo método de Nei-Gojobori.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1																
2	1,0035															
3	0,4305	0,5008														
4	0,4216	0,4902	0,5960													
5	0,4250	0,4970	0,4264	0,5886												
6	0,3858	0,4501	0,3112	0,3120	0,2400											
7	0,4366	0,5108	1,0759	0,3589	0,1790	0,4055										
8	0,4391	0,5134	1,1274	0,4103	0,2046	0,4128	0,4830									
9	0,4366	0,5108	1,0759	0,3589	0,1790	0,4055	0,4744	0,3576								
10	0,4656	0,4681	0,7158	0,4776	0,3573	0,4328	0,4882	0,4462	0,3559							
11	0,4366	0,5108	1,0759	0,3589	0,1790	0,4055	0,4744	0,3576	0,2378	-1,0000						
12	0,4158	0,4859	0,5374	0,1793	0,1192	0,3545	0,4062	0,2382	0,1781	0,0000	0,0000					
13	0,4366	0,5108	1,0759	0,3589	0,1790	0,4055	0,4744	0,3576	0,2378	-1,0000	-1,0000	-1,0000				
14	0,3736	0,4238	0,3319	0,3329	0,3231	0,3044	0,3134	0,3134	0,3044	0,3337	0,3337	0,3337	0,3337			
15	0,3034	0,3281	0,2560	0,2568	0,2499	0,2349	0,2413	0,2413	0,2349	0,2556	0,2556	0,2556	0,2556	0,2556		
16	0,3125	0,3376	0,2635	0,2644	0,2571	0,2418	0,2484	0,2484	0,2418	0,2633	0,2633	0,2633	0,2633	0,2633	0,2633	0,2633

Nota: 1-*Homo sapiens*, 2-*Gorilla gorilla*, 3-*Cercopithecus galeries chrysogaleus*, 4-*Cercopithecus torquatus lunatus*, 5-*Cercopithecus mitis*, 6-*Lophocebus aterrimus*, 7-*Macaca fascicularis*, 8-*Macaca mulatta*, 9-*Macaca nemestrina*, 10-*Macaca nigra*, 11-*Macaca thibetana*, 12-*Mandrillus leucophaeus*, 13-*Mandrillus sphinx*, 14-*Cebus apella*, 15-*Saimiri sciureus*, 16-*Saimiri ustus*.

4.3.5 Método de ramo por máxima verossimilhança

O método de ramo por máxima verossimilhança é utilizado para encontrar indícios de seleção positiva em linhagens específicas. Um representante de cada espécie foi escolhido aleatoriamente e uma árvore filogenética foi gerada por máxima verossimilhança (Figura 18). Esta árvore foi utilizada para todas as análises de máxima verossimilhança. Para a análise de ramos usamos toda a região codificante do gene *DARC*.

Utilizou-se um modelo com duas razões d_N/d_S (*two-ratio model*) sendo, respectivamente, $\hat{\omega}_0$ e $\hat{\omega}_1$ os valores para a razão d_N/d_S dos ramos de fundo (*background*) e do ramo em evidência (*foreground*). Os ramos marcados como *foreground* e utilizados em cada análise foram marcados em ordem alfabética (Figura 18). Em cada análise um ramo era evidenciado, isto é, uma razão d_N/d_S diferente era calculada ($\hat{\omega}_1$) e os demais ramos se comportavam como plano de fundo, portanto todos possuíam a mesma razão d_N/d_S ($\hat{\omega}_0$). Para cada análise, além das razões $\hat{\omega}_0$ e $\hat{\omega}_1$, a verossimilhança (*likelihood* ou ℓ) e a taxa de transição-transversão (κ) era calculada e os resultados encontram-se na Tabela 8 (H_1).

Figura 18: Árvore filogenética utilizada na análise de ramo por máxima verossimilhança.

Tabela 8: Valores do log de verossimilhança, estimativa dos parâmetros e estatística da relação de verossimilhança ($2\Delta\ln L$) para o modelo M0 (d_N/d_S único para todos os ramos) e para diferentes ramos da árvore filogenética (hipótese alternativa ou H_1).

Modelo	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_1$	$2\Delta\ln L$	
M0: razão única	31	-2365,5986	4,8496	0,5421	-	-	-
Ramo	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_1$	I	II
A	32	-2365,5986	4,8495	0,5421	1,0014	0,0001	0,0000
B	32	-2363,2089	4,8512	0,6192	0,2016	4,7794*	6,4074*
C	32	-2365,1262	4,8574	0,5872	0,4007	0,9449	5,7923
D	32	-2365,2982	4,8480	0,5678	0,3802	0,6010	3,2712
E	32	-2365,5899	4,8493	0,5443	0,4786	0,0175	0,5628
F	32	-2364,9590	4,8498	0,5359	∞	1,2793	0,4899
G	32	-2364,8553	4,8489	0,5490	0,0001	1,4866	0,7246
H	32	-2365,0128	4,8502	0,5363	∞	1,1716	0,4358
I	32	-2364,9783	4,8489	0,5490	0,0001	1,2407	0,6416
J	32	-2365,5986	4,8496	0,5421	0,9760	0,0000	0,0000
K	32	-2365,5986	4,8496	0,5421	0,4390	0,0000	0,0000
L	32	-2365,5986	4,8496	0,5421	1,0272	0,0000	0,0000
M	32	-2365,3808	4,8485	0,5496	0,2490	0,4356	0,9502
N	32	-2365,4864	4,8507	0,5315	0,7273	0,2245	0,0127
O	32	-2365,2309	4,8483	0,5162	0,8472	0,7354	0,6506
P	32	-2365,5858	4,8497	0,5473	0,5037	0,0256	1,8727
Q	32	-2364,9666	4,8497	0,5360	∞	1,2641	0,4868
R	32	-2364,7723	4,8505	0,5325	∞	1,6527	0,5874

Nota: p é o número de parâmetros incluídos no modelo. Os parâmetros $\hat{\omega}_0$ e $\hat{\omega}_1$ são, respectivamente, a razão d_N/d_S para as linhagens de fundo (*background*) e linhagens em evidência (*foreground*). I e II são as estatísticas da relação de verossimilhança ($2\Delta\ln L$) entre M0 x H_1 e M0 x H_2 (os dados de H_2 estão na Tabela 9) respectivamente.

* Valores significativos ($P < 5\%$; $x_1^2 = 3,84$).

A verossimilhança calculada pelo modelo com duas razões d_N/d_S precisa ser compa-

rada com a verossimilhança de um modelo definido como hipótese nula (H_0). Utilizou-se como hipótese nula o modelo de taxa única (*one-ratio model* ou M0), onde uma única razão d_N/d_S é calculada para toda a árvore ($\hat{\omega}_0$). Assim, podemos utilizar um Teste de Razão de Verossimilhança (*Likelihood Ratio Test* ou LRT) para aceitar ou rejeitar a hipótese nula. Um LRT é calculado como duas vezes a diferença da verossimilhança entre os modelos ($2(\ell_1 - \ell_0)$ ou $2\Delta\ln L$) e sua distribuição pode ser aproximada à distribuição qui-quadrada (χ^2) com o grau de liberdade igual a diferença entre os número de parâmetros livres dos modelos envolvidos.

Calculamos novamente um modelo com duas razões (H_2) para cada ramo marcado, porém os ramos em evidência foram definidos como neutros ($\hat{\omega}_1=1$) e os resultados podem ser vistos na Tabela 9. Esse modelo permite descobrir se o resultado encontrado no primeiro LRT difere do esperado se o ramo em evidência evoluísse de forma neutra. Os resultados do LRT entre $H_1 \times M0$ (I) e $H_2 \times M0$ (II) estão na Tabela 8.

Tabela 9: Valores do log de verossimilhança e estimativa dos parâmetros sob diferentes ramos da árvore filogenética com $\hat{\omega}_1 = 1$ (modelo neutro ou H_2).

Ramo	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_1$
A	31	-2365,5986	4,8496	0,5421	1,0000
B	31	-2368,8023	4,9757	0,6068	1,0000
C	31	-2368,4948	4,9687	0,5816	1,0000
D	31	-2367,2343	4,9231	0,5595	1,0000
E	31	-2365,8800	4,8644	0,5424	1,0000
F	31	-2365,3537	4,8440	0,5359	1,0000
G	31	-2365,9609	4,8615	0,5492	1,0000
H	31	-2365,3808	4,8445	0,5365	1,0000
I	31	-2365,9194	4,8603	0,5479	1,0000
J	31	-2365,5986	4,8495	0,5421	1,0000
K	31	-2365,5986	4,8496	0,5421	1,0000
L	31	-2365,5986	4,8496	0,5421	1,0000
M	31	-2366,0737	4,8680	0,5498	1,0000
N	31	-2365,6050	4,8649	0,5310	1,0000
O	31	-2365,2733	4,8573	0,5148	1,0000
P	31	-2366,5350	4,8978	0,5418	1,0000
Q	31	-2365,3552	4,8439	0,5359	1,0000
R	31	-2365,3049	4,8427	0,5345	1,0000

Nota: p é o número de parâmetros incluídos no modelo. Os parâmetros $\hat{\omega}_0$ e $\hat{\omega}_1$ são, respectivamente, as taxas d_N/d_S para as linhagens de fundo (*background*) e linhagens em evidência (*foreground*).

As análises pelo método de ramo mostram que os ramos A e L possuem um $\hat{\omega}_1 > 1$, porém a estatística LRT não mostra que o modelo de duas taxas (H_1) representa melhor os dados que o modelo de taxa única (Tabela 8). Além disso, a verossimilhança entre o modelo de taxa única (M0) e o modelo de neutralidade (H_2) é a mesma para ambos os ramos ($\ell=-2365,5986$), demonstrando que os mesmo não estão evoluindo sob seleção positiva.

O ramo B é o único que apresenta uma estatística LRT significativa ($P < 5\%$), porém apresenta $\hat{\omega}_1=0,2016$ (Tabela 8), o que sugere a ação da seleção purificadora nessa linha-

gem. Portanto, os dados pela análise de ramos por máxima verossimilhança não indicam a assinatura da seleção positiva em nenhum ramo da árvore.

4.3.6 Método de sítio por máxima verossimilhança

O método de sítio por máxima verossimilhança permite a razão d_N/d_S (ω) variar entre os sítios (neste caso, cada códon é considerado um sítio) mas não entre ramos da árvore (NIELSEN; YANG, 1998). Foram utilizados três pares de modelos na análise de sítio. O primeiro par inclui os modelos M1a (Aproximadamente Neutro) e M2a (Seleção Positiva), o segundo par inclui M0 (razão única) e M3 (Discreto) e o terceiro par inclui M7 (beta) e M8 (beta& ω). Para a análise de sítios usamos toda a região codificante do gene *DARC*. Os resultados encontram-se na Tabela 10.

Tabela 10: Valores do log de verossimilhança, estimativa dos parâmetros e estatística da relação de verossimilhança ($2\Delta\ln L$) para os modelos de sítio (M1a, M2a, M0, M3, M7 e M8).

Modelo	p	ℓ	Distribuição $\hat{\omega}$		$2\Delta\ln L$	Sítios selecionados positivamente
M1a: neutro	1	-2356,1195	$\hat{\omega}_0=0,0000$ $\hat{\omega}_1=1,0000$	$\hat{p}_0=0,5175$ $\hat{p}_1=0,4825$	4,532	não permitido
M2a: seleção	3	-2353,8535	$\hat{\omega}_0=0,2353$ $\hat{\omega}_1=1,0000$ $\hat{\omega}_2=2,6261$	$\hat{p}_0=0,8509$ $\hat{p}_1=0,0000$ $\hat{p}_2=0,1492$		7R 22F 25V 28S 31G 34D 39G 211T 332L ($0,5 < P < 0,87$)
M0: razão única	1	-2365,5986	$\hat{\omega}_0=0,5421$		23,4909*	
M3: discreto	5	-2353,8532	$\hat{\omega}_0=0,0000$ $\hat{\omega}_1=0,2835$ $\hat{\omega}_2=2,6929$	$\hat{p}_0=0,1087$ $\hat{p}_1=0,7504$ $\hat{p}_2=0,1410$		7R 22F 25V 28S 39G 332L ($P > 0,95$)
M7: beta	2	-2356,1580	$\hat{p}=0,0050$	$\hat{q}=0,0050$	4,609	não permitido
M8: beta& ω	4	-2353,8535	$\hat{p}_0=0,8535$ $\hat{p}=11,5305$ $\hat{q}=36,5561$	$\hat{p}_1=0,1465$ $\hat{\omega}=2,6462$		7R 22F 25V 28S 31G 34D 39G 122S 124R 211T 329L 332L ($0,53 < P < 0,93$)

Nota: p é o número de parâmetros livres incluídos no modelo. Os parâmetros $\hat{\omega}_x$ ($x = 0,1$ ou 2) são as razões d_N/d_S incluídas nos respectivos modelos de sítio.

* Valor extremamente significativo ($P < 1\%$; $x_4^2 = 13,48$).

O modelo M1a (aproximadamente neutro) assume duas classes de sítio com ω_0 podendo variar entre 0 e 1 ($0 < \omega_0 < 1$), sendo estimado a partir dos dados e ω_1 fixado em 1 ($\omega_1=1$), com as proporções p_0 e p_1 estimadas a partir dos dados. O modelo M2a (seleção positiva) adiciona uma terceira classe com a razão ω_2 estimada. Este modelo sugere que aproximadamente 15% dos sítios estão sob seleção positiva com $\hat{\omega}_2=2,6261$ (Tabela 10). Devido ao modelo M2a (seleção) ser uma extensão do modelo M1a (aproximadamente neutro), os dois modelos podem ser comparados usando um LRT. A estatística do teste é $2\Delta\ln L = 2 \times ((-2353,8535) - (-2356,1195)) = 2 \times 2,266 = 4,532$, com $P=0,1$ e grau de liberdade (gl) = 2. Portanto o modelo M2a não é significantemente melhor que o modelo M1a para os dados.

O modelo discreto (M3) com 3 classes ($K=3$) sugere que 14% dos sítios estão sob seleção positiva com $\hat{\omega}_2=2,6929$ e identificou seis aminoácidos sob seleção positiva com $P > 95\%$. O modelo M3 foi significativamente melhor que o modelo de razão única M0; o teste estatístico é $2\Delta\ln L = 2 \times ((-2353,8532) - (-2365,5986)) = 2 \times 11,7454 = 23,4909$ com $P=0,0001$ e $gl=4$.

Os aminoácidos identificados sob seleção diversificadora no modelo M3 com $P > 70\%$ foram combinados com a estrutura tridimensional predita do gene *DARC* (Figuras 19 e 20). A estrutura tridimensional não possui os primeiros 26 aminoácidos do gene, portanto os sítios 7R, 22F e 25V não são mostrados. Observa-se nas figuras que dos 14 sítios sob seleção, quatro estão presentes nos primeiros 60 aminoácidos (região extracelular) da proteína. De acordo com Chitnis et al. (1996) a região de ligação do *P. vivax* estende-se do aminoácido 8 ao 42, indicando que essa região pode realmente estar sob forte influência da seleção diversificadora devido à ação do *P. vivax*.

Figura 19: Estrutura tridimensional predita da proteína DARC humana. Os sítios sob seleção positiva preditos pelo modelo M3 estão em vermelho e rosa. A estrutura em branco indica sítios conservados.

O modelo M7 (beta) assume uma distribuição beta para ω nos sítios. A distribuição beta é limitada ao intervalo (0, 1) e providencia uma hipótese nula flexível para testar a seleção positiva. Cálculos sugerem que esta distribuição se reduz ao modelo aproximadamente neutro (M1a). O modelo M8 (beta& ω) adiciona uma outra classe de sítio ao modelo M7 (beta), com a razão ω estimada a partir dos dados. Este modelo sugere que aproximadamente 15% dos sítios estão sob seleção positiva com $\hat{\omega}=2,6462$ e identificou 12

Figura 20: Estrutura tridimensional predita da proteína DARC humana. Os sítios mais conservados estão em azul escuro em direção gradativa aos sítios sob seleção positiva em vermelho.

sítios sob seleção positiva (os mesmo seis sítios sugeridos por M3 mais 31G, 34D, 122S, 124R, 211T e 329L). No entanto as diferenças entre M7 e M8 não são estatisticamente significativas com $2\Delta\ln L = 4,609$ com $P=0,1$ e $gl=4$.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob o gene *DARC* parece ser muito variável ao longo de sua seqüência.

4.3.7 Análise de domínios pelo método de máxima verossimilhança

Pode-se utilizar o método de sítio por máxima verossimilhança dividindo-se os sítios em classes (ou domínios) fixos, os quais podem estar sujeitos a diferentes pressões seletivas e, portanto, diferentes razões ω . Os sítios do gene *DARC* foram divididos em quatro domínios a partir de sua estrutura protéica: Domínio 1 - a região extracelular de ligação do *P. vivax* (os primeiros 62 aminioácidos), Domínio 2 - regiões transmembrana, Domínio 3 - regiões citoplasmáticas e Domínio 4 - demais regiões extracelulares (Figura 21).

A partir da definição dos domínios é necessário utilizar modelos para dados particionados. Quando aplica-se o modelo de sítio por máxima verossimilhança em dados particionados, utiliza-se diferentes razões ω para cada partição. De maneira similar po-

Figura 21: Esquema mostrando a divisão dos sítios do gene *DARC* em domínios para análise de seleção positiva pelo método de sítio. O Domínio 1 é a região extracelular de ligação do *P. vivax*. O Domínio 2 é composto pelas regiões transmembrana. O Domínio 3 é composto pelas regiões citoplasmáticas. O Domínio 4 é composto pelas regiões extracelulares, exceto pela região definida como Domínio 1.

demos permitir que outros parâmetros difiram entre os sítios particionados.

Utilizamos quatro modelos para acomodar varios níveis de heterogeneidade entre os sítios. O modelo mais simples assume que todos os sítios da seqüência têm o mesmo padrão de substituição com parâmetros idênticos (o modelo A na Tabela 11). Os parâmetros neste modelo incluem o comprimento dos ramos filogenéticos (b), a razão de transições/transversões (κ), a razão d_N/d_S (ω), e nove parâmetros para a frequência dos códons, com $b + 11$ parâmetros no total. O modelo mais complexo (modelo D na Tabela 11) assume comprimento dos ramos filogenéticos (b) proporcionais entre as partições, porém com diferentes frequências de códons (rs), diferentes razões de transições/transversões (κ), diferentes razões de sítio não-sinônimos/sinônimo (ω) e diferente frequência nucleotídica de equilíbrio (πS). Os modelos B e C estão entre estes dois extremos.

O resultados sob cada modelo de sítios fixos encontram-se na Tabela 11. O modelo A, que permite somente as taxas (rs) diferirem entre os domínios, mostrou $\ell_A = -2359,8080$. O modelo B permite, além disso, diferentes frequências de códons para cada domínio, as quais são determinada pela frequência nucleotídica nas três posições do códon, e representa

Tabela 11: Valores do log de verossimilhança e estimativa dos parâmetros para os modelos de sítios fixos (análise por domínios da proteína) .

Modelo	ℓ	$\hat{r}_2$	$\hat{r}_3$	$\hat{r}_4$	$\hat{\kappa}$		$\hat{\omega}$		$2\Delta\ln L$
A ($\neq rs$)	-2359,8080	0,5606	0,4154	0,7030	4,8255		0,5370		-
B ($\neq rs$ e πS)	-2333,2395	0,6736	0,4253	0,7462	4,8318		0,5662		-
C ($\neq rs, \kappa$ e ω)	-2348,2742	0,5658	0,4191	0,7271	$\hat{\kappa}_1=2,8024$ $\hat{\kappa}_2=5,0033$	$\hat{\kappa}_3=2,9001$ $\hat{\kappa}_4=12,3866$	$\hat{\omega}_1=\mathbf{1,3244}$ $\hat{\omega}_2=0,3083$	$\hat{\omega}_3=0,5442$ $\hat{\omega}_4=0,5911$	2,89
D ($\neq rs, \kappa, \omega$ e πS)	-2322,0869	0,6750	0,4257	0,7513	$\hat{\kappa}_1=2,9775$ $\hat{\kappa}_2=5,0090$	$\hat{\kappa}_3=3,0493$ $\hat{\kappa}_4=12,4181$	$\hat{\omega}_1=\mathbf{1,4049}$ $\hat{\omega}_2=0,3401$	$\hat{\omega}_3=0,5442$ $\hat{\omega}_4=0,5690$	3,32

Nota: Os parâmetros $\hat{\omega}_x$ e $\hat{\kappa}_x$ (onde $x = 1, 2, 3$ e 4) são, respectivamente, as razões d_N/d_S e a razão de transição/transversão para cada um dos quatro domínios definidos. Os parâmetros $\hat{r}_2$, $\hat{r}_3$ e $\hat{r}_4$ são as taxas para cada domínio ($\hat{r}_1 = 1$).

melhor os dados que o modelo A; a estatística LRT é $2\Delta\ln L = 2((-2333,2395)-(-2359,8080)) = 26,5685$, comparado com a distribuição χ^2 com g.l. = 9. Este resultado sugere que o padrão no uso dos códons é realmente diferente entre os domínios. O modelo C assume a mesma frequência de códons mas diferentes razões de transição/transversão (κ) e d_N/d_S (ω). Este modelo representa melhor os dados que o modelo A ($2\Delta\ln L = 11,5338$ g.l. = 2). Os parâmetros κ_2 e κ_4 variam muito em relação aos outros dois domínios. A razão d_N/d_S da região de ligação do *P. vivax* ($\omega_1 = 1,3244$) sugere a ação de seleção diversificadora enquanto as demais regiões ($\omega_2, \omega_3, \omega_4 < 1$) parecem estar sob seleção purificadora. O modelo D, que permite diferentes parâmetros κ , ω e diferentes frequências de códons, representou melhor os dados em relação aos demais modelos. Este modelo apresentou resultados similares aos do modelo C.

Para testar se a razão ω sobre o Domínio 1 é significativamente maior do que 1 calculou-se o log da verossimilhança para os modelos C e D fixando $\omega_1 = 1$. Assim, a estatística LRT para a hipótese nula $\omega_1 = 1$ para C e D é, respectivamente, 2,89 e 3,32 (Tabela 11) com g.l.=1. Portanto as razões d_N/d_S para o Domínio 1 nos modelos C e D não são significativamente maiores que 1.

5 *Discussão*

O antígeno/receptor para quimiocinas Duff (DARC) é amplamente expresso e age tanto como receptor promíscuo de quimiocinas quanto como receptor eritrocitário para o *Plasmodium vivax*. Devido a sua propriedade biológica dual, DARC é potencialmente submetido a duas pressões seletivas: uma seleção purificadora interna para manutenção das propriedades de ligação de quimiocinas (tendo as quimiocinas endógenas como fator seletivo) e uma seleção positiva externa (sendo os parasitas o fator seletivo). A discussão dos resultados abaixo apóia-se nessa hipótese.

5.1 Comparações entre as seqüências

5.1.1 Região GATA-1

A resistência à invasão por *P. vivax* em humanos é, geralmente, causada pela substituição de uma base no sítio de ligação do fator de transcrição GATA-1 determinando a não expressão do gene *DARC* nos eritrócitos. O motivo GATA-1 não apresentou nenhuma diferença entre as espécies seqüenciadas (alinhamento no Apêndice A), indicando, aparentemente, que esse polimorfismo ocorre somente em humanos, confirmando os resultados obtidos por Tournamille et al. (2004).

5.1.2 Cisteínas

De acordo com Tournamille et al. (1997) a conservação das cisteínas extracelulares é importante pois acredita-se que pelo menos quatro delas desempenham um importante papel na estrutura do receptor de quimiocinas. O gene *DARC* apresenta 12 cisteínas. Os resíduos C4, C51, C54, C129, C195 e C276 são preditos como expostos na região extracelular e C93, C134, C149, C214, C297 e C308 localizam-se, aparentemente, na região transmembrana ou intracelular.

Todos os resíduos de cisteínas encontram-se conservados entre as espécies de primatas estudadas, com exceção do resíduo C308 que foi substituído por um tirosina em gorila, confirmado os achados de (TOURNAMILLE et al., 2004).

5.1.3 $Fy^{a/b}$

A expressão de Fy^a e Fy^b em primatas não-humanos foi analisada por Palatnik e Rowe (1984). Tournamille et al. (2004) confirmaram por análise de citometria de fluxo que a expressão de Fy^a restringe-se à humanos e que a diferença é a mudança de um ácido aspártico por uma glicina no resíduo 42. Este resultado reforça a hipótese de Chaudhuri et al. (1995) que *FYB* é o alelo ancestral.

Nossos resultados reconfirmam os achados de Tournamille et al. (2004).

5.1.4 Fy^6

Foi demonstrado por Chitnis et al. (1996) que a sequência com 35 aminoácidos de comprimento (Ala8-Asp42) do domínio amino-terminal de DARC possui sítios de ligação para anti- Fy^6 e malária. Foi confirmado que alguns macacos do novo mundo (*Saimiri boliviensis* e *Aotus trivirgatus*) podem ser infectados por *P. vivax* porém outras espécies (*Cebus Apella*) não apresentaram sinais de infecção (TOURNAMILLE et al., 2004). Além disso, estudos mostraram que apenas o mapeamento da sequência primária correspondente ao domínio Fy^6 não é suficiente para determinar a susceptibilidade à invasão por *P. vivax* (TOURNAMILLE et al., 2004).

5.2 Testes de seleção positiva

5.2.1 Método de Nei-Gojobori

O teste de seleção positiva pelo método de Nei-Gojobori não indicou a ação de seleção positiva (Tabela 7). Isto pode ser explicado pelo próprio método. De acordo com Yang e Nielsen (2000) esta abordagem utiliza a média da taxa de substituição sobre todos os aminoácidos da sequência e sobre o intervalo de tempo separando as mesmas. Como espera-se que a maioria dos sítios de aminoácidos seja conservada e a evolução adaptativa afete somente alguns sítios em pequenos espaços temporais este método, teoricamente, não possui poder para detectar a ação da seleção positiva.

A falta de indícios de seleção natural pelo método de Nei-Gojobori pode, portanto, indicar duas explicações. A primeira é a não existência da seleção Darwiniana sobre o gene *DARC* e a segunda é a presença da mesma em apenas poucos sítios ou partes do gene, dificultando a detecção da mesma por esse método. Assim, outros testes de seleção positiva por métodos de máxima verossimilhança foram aplicados aos dados para esclarecer qual das duas explicações é verdadeira.

5.2.2 Método de ramo por máxima verossimilhança

A análise dos ramos por máxima verossimilhança da árvore filogenética representando a relação entre as espécies de primatas (Figura 18) também não apresentou sinais estatisticamente significativos de seleção positiva em nenhum ramo da árvore (Tabela 8). Ao contrário, mostrou a presença de uma seleção purificadora no ramo B (Figura 18 e Tabela 8) que separa o gênero *Saimiri* dos demais primatas. Este resultado não difere do encontrado na comparação par à par pelo método de Nei-Gojobori.

Na verdade, o método de ramo por máxima verossimilhança permite a razão d_N/d_S (ω) variar entre os ramos da árvore mas não permite, assim como o método de Nei-Gojobori, variar a razão d_N/d_S entre os sítios da sequência (YANG, 1998). Portanto, estes dois métodos devem apresentar resultados semelhantes. A diferença reside no fato do método de ramos permitir detectar as seleções positiva e purificadora em ramos específicos da árvore filogenética. Sendo assim, apesar de ter acrescentado novas informações, o método de ramos não teve poder para descartar a hipótese de seleção positiva atuando em poucos sítios.

5.2.3 Método de sítio por máxima verossimilhança

A análise de sítios por máxima verossimilhança permite a razão d_N/d_S variar entre os sítios mas não entre os ramos da árvore. Esta abordagem permite detectar a presença da seleção positiva em sítios específicos da sequência (YANG et al., 2000). Nos três modelos que permitem sítios com seleção positiva (M2a, M3 e M8) a razão d_N/d_S (ω_2) foi maior que 1, porém somente M3 obteve LRT (*Likelihood Ratio Test*) estatisticamente significativa (Tabela 10). De acordo com Yang e Nielsen (2002) o LRT entre os modelos M0 e M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção propriamente. O próprio manual do software PAML aconselha a utilização de dois pares de modelos (M1a-M2a e M7-M8) para detecção da seleção positiva.

Assim, o método de sítio falhou em detectar a seleção positiva nos dois pares de modelos propostos (M1a-M2a e M7-M8), porém indica uma alta variabilidade da razão ω entre os sítios (M0-M3). Os sítios indicados por M3 como sob influência da seleção positiva foram marcados na estrutura tridimensional da proteína DARC (Figuras 19 e 20) e a sua distribuição não mostrou ser aleatória, estando a maioria concentrada na região de ligação do *Plasmodium vivax*.

O conhecimento da distribuição dos sítios selecionados positivamente, somado as informações prévias da estrutura do gene *DARC* em relação a sua proteína, permitiu dividir a sequência do gene em domínios ou classes. Assim pudemos realizar análises com o método de sítio utilizando os domínios ao invés da sequência como um todo.

5.2.4 Análise de domínios pelo método de máxima verossimilhança

A análise de sítio utilizando domínios ou classes, permite uma razão ω diferente para cada domínio. Além disso outros parâmetros, como a taxa de transição/transversão (κ) e a frequência de equilíbrio dos códons (π) também podem variar entre os domínios. Os resultados dessa análise mostraram que a região amino-terminal de ligação do *P. vivax* (denominado Domínio 1 ou) possui a razão ω maior do que 1, porém não é estatisticamente diferente de 1. Os demais domínios mostraram-se conservados, com a região transmembrana (Domínio 2) sendo a mais conservada ($\omega < 0,5$).

Esses resultados confirmam que a conservação dos sítios do gene *DARC* não é homogênea ao longo de sua sequência. Devido a isso, análises que utilizam a média da razão d_N/d_S para representar a pressão seletiva sobre todos os sítios da sequência falharam em detectar a assinatura da seleção positiva.

Tournamille et al. (2004) realizaram um estudo semelhante com o gene *DARC* utilizando a comparação par à par de sequências pelo método de Nei-Gojobori modificado. Uma das análises foi feita utilizando toda a sequência codificante do gene e outra utilizando somente a região amino-terminal (Domínio 1). Os resultados obtidos por Tournamille et al. (2004) foram semelhantes aos do presente trabalho e a conclusão foi que não existiam argumentos a favor da seleção positiva sobre o gene *DARC*.

Contudo, Tournamille et al. (2004) não realizaram os testes de seleção positiva nas regiões que denominamos respectivamente domínios 2, 3 e 4. Nossos resultados mostram que essas regiões estão sob forte seleção purificadora, enquanto a região amino-terminal,

apesar de apresentar razões ω maiores que 1, está estatisticamente neutra. Essa discrepância em si mostra que existem diferentes forças seletivas atuando sobre os domínios.

A região de ligação do *P. vivax* (Domínio 1) foi analisada separadamente das demais regiões extracelulares (Domínio 4) por motivos de estudo, mas teoricamente, ambas estão sob a mesma pressão seletiva. Sendo assim a região amino-terminal deve estar sob influência da seleção purificadora como as demais regiões extracelulares. A diferença entre as duas regiões no entanto indica a existência de uma segunda força seletiva atuando sobre a região de ligação do *P. vivax*.

O modelo de duas pressões seletivas (seleção negativa e seleção positiva) atuando sobre uma mesma região do gene enquanto as demais regiões estão sob influência de apenas uma delas é uma explicação razoável para os resultados. Assim, não podemos descartar a existência da seleção positiva atuando no gene *DARC*. No entanto, não existem testes estatísticos que possam ser aplicados para confirmar este modelo. Além disso, apesar dos testes de seleção positiva por máxima verossimilhança serem os melhores disponíveis no momento, Yang e Nielsen (2002) argumentam que os testes LRT podem falhar em prover suporte estatístico significativo para a seleção positiva, principalmente quando as regiões analisadas são curtas. Isto pode explicar a falta de significância estatística para a seleção positiva na região de ligação do *P. vivax*, pois a mesma possui somente 60 aminoácidos.

Além disso, sabe-se que *P. vivax* infecta humanos, enquanto os demais primatas principalmente do gênero *Macaca* são infectados por *P. knowlesi*. De acordo com Iyer et al. (2007) *P. vivax* é altamente dependente da proteína *DARC* para a infecção dos eritrócitos, enquanto *P. knowlesi* possui vias alternativas desconhecidas. Isso indica, possivelmente, que a proteína *DARC* não sofre uma forte pressão seletiva em macacos, pois sua presença não é um fator determinante para a infecção, o que também pode explicar a falta de significância no teste de seleção positiva.

A proteína *DARC* possui uma complexa relação entre malária e sua função de receptor promíscuo de quimiocinas. Tudo isso, somado à sua expressão em diferentes tecidos, dificulta o estudo da relação e identificação dos fatores que contribuíram para sua evolução.

Assim sendo, os testes de seleção positiva não apresentam argumentos estatísticos a favor da seleção positiva atuando sobre o gene *DARC*. Porém, os resultados discrepantes indicam um cenário onde forças seletivas contra-balanceiam-se, não permitindo aos atuais testes estatísticos detectar a assinatura da seleção positiva. Portanto, conclui-se que há indícios da presença da seleção positiva na região amino-terminal do gene *DARC* mas novas abordagens de análise devem ser desenvolvidas e utilizadas para confirmar sua

existência.

6 Conclusões

Os dados do presente trabalho permite-nos concluir que:

- A substituição de uma base no sítio de ligação do fator de transcrição GATA-1, responsável pela não expressão da glicoproteína DARC nos eritrócitos, não foi encontrada em outros primatas.
- Os resíduos de cisteínas presentes no gene *DARC*, responsáveis pela estrutura do receptor de quimiocinas, são conservados entre os primatas.
- A sequência do gene *DARC* possui uma alta variabilidade da razão d_N/d_S entre os sítios.
- Toda a sequência do gene *DARC* está sob influência da seleção purificadora, com excessão da região amino-terminal que apresenta sinais de seleção positiva mas não com suporte estatístico.
- A hipótese da região amino-terminal estar sob influência de duas pressões seletivas antagônicas (seleção positiva e seleção negativa) pode explicar a discrepância das razões d_N/d_S encontradas.
- Os testes de taxa de verossimilhança (LRTs ou *Likelihood Ratio Tests*) podem ter falhado em detectar a seleção positiva devido ao pequeno tamanho da região amino-terminal (aproximadamente 60 aminoácidos).
- A influência da seleção positiva sob o gene *DARC* não pode ser descartada e o desenvolvimento de novos testes estatísticos são necessários para comprovar a sua assinatura.

Referências

- BARNWELL, J. W.; NICHOLS, M. E.; RUBINSTEIN, P. In vitro evaluation of the role of the Duffy blood group in erythrocyte invasion by *Plasmodium vivax*. *J Exp Med*, v. 169, n. 5, p. 1795–802, maio 1989. ISSN 0022-1007.
- BEITZ, E. TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon. *Bioinformatics*, v. 16, n. 2, p. 135–9, fev. 2000. ISSN 1367-4803.
- BERNSTEIN, H. J. Recent changes to RasMol, recombining the variants. *Trends Biochem Sci*, v. 25, n. 9, p. 453–5, set. 2000. ISSN 0968-0004.
- CASTILHO, L. The value of DNA analysis for antigens in the Duffy blood group system. *Transfusion*, v. 47, n. 1 Suppl, p. 28S–31S, jul. 2007. ISSN 0041-1132.
- CASTILHO, L. et al. A novel FY allele in Brazilians. *Vox Sang*, v. 87, n. 3, p. 190–5, out. 2004. ISSN 0042-9007.
- CHAUDHURI, A. et al. Detection of Duffy antigen in the plasma membranes and caveolae of vascular endothelial and epithelial cells of nonerythroid organs. *Blood*, v. 89, n. 2, p. 701–12, jan. 1997. ISSN 0006-4971.
- CHAUDHURI, A. et al. Cloning of glycoprotein D cDNA, which encodes the major subunit of the Duffy blood group system and the receptor for the *Plasmodium vivax* malaria parasite. *Proc Natl Acad Sci U S A*, v. 90, n. 22, p. 10793–7, nov. 1993. ISSN 0027-8424.
- CHAUDHURI, A. et al. The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissues in Duffy-negative individuals. *Blood*, v. 85, n. 3, p. 615–21, fev. 1995. ISSN 0006-4971.
- CHEN, K. et al. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res*, v. 17, n. 5, p. 659–66, maio 2007. ISSN 1088-9051.
- CHITNIS, C. E. et al. The domain on the Duffy blood group antigen for binding *Plasmodium vivax* and *P. knowlesi* malarial parasites to erythrocytes. *J Exp Med*, v. 184, n. 4, p. 1531–6, out. 1996. ISSN 0022-1007.
- CUTBUSH, M.; MOLLISON, P. L. The Duffy blood group system. *Heredity*, v. 4, n. 3, p. 383–9, dez. 1950. ISSN 0018-067X.
- DARBONNE, W. C. et al. Red blood cells are a sink for interleukin 8, a leukocyte chemotaxin. *J Clin Invest*, v. 88, n. 4, p. 1362–9, out. 1991. ISSN 0021-9738.

- EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, v. 8, n. 3, p. 186–94, mar. 1998. ISSN 1088-9051.
- EWING, B. et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, v. 8, n. 3, p. 175–85, mar. 1998. ISSN 1088-9051.
- FELSENSTEIN, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, v. 17, n. 6, p. 368–76, 1981. ISSN 0022-2844.
- FELSENSTEIN, J. *PHYLIP (Phylogeny Inference Package) version 3.6*. 2005. Distribuído pelo autor. Department of Genome Sciences, University of Washington, Seattle.
- GALINSKI, M. R. et al. A reticulocyte-binding protein complex of *Plasmodium vivax* merozoites. *Cell*, v. 69, n. 7, p. 1213–26, jun. 1992. ISSN 0092-8674.
- GARRATTY, G. Blood group antigens as tumor markers, parasitic/bacterial/viral receptors, and their association with immunologically important proteins. *Immunol Invest*, v. 24, n. 1-2, p. 213–32, 1995. ISSN 0882-0139.
- GOLDMAN, N.; YANG, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, v. 11, n. 5, p. 725–36, set. 1994. ISSN 0737-4038.
- GORDON, D.; ABAJIAN, C.; GREEN, P. Consed: a graphical tool for sequence finishing. *Genome Res*, v. 8, n. 3, p. 195–202, mar. 1998. ISSN 1088-9051.
- GORDON, D.; DESMARAIS, C.; GREEN, P. Automated finishing with autofinish. *Genome Res*, v. 11, n. 4, p. 614–25, abr. 2001. ISSN 1088-9051.
- GUINDON, S.; GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, v. 52, n. 5, p. 696–704, out. 2003. ISSN 1063-5157.
- HADLEY, T. J. et al. Identification of an erythrocyte component carrying the Duffy blood group Fya antigen. *Science*, v. 223, n. 4636, p. 597–9, fev. 1984. ISSN 0036-8075.
- HADLEY, T. J.; MILLER, L. H.; HAYNES, J. D. Recognition of red cells by malaria parasites: the role of erythrocyte-binding proteins. *Transfus Med Rev*, v. 5, n. 2, p. 108–22, abr. 1991. ISSN 0887-7963.
- HADLEY, T. J.; PEIPER, S. C. From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. *Blood*, v. 89, n. 9, p. 3077–91, maio 1997. ISSN 0006-4971.
- HASEGAWA, M.; KISHINO, H.; YANO, T. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. *Proc Jpn Acad*, v. 60, n. 4, p. 95–98, 1984.
- HASEGAWA, M.; KISHINO, H.; YANO, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, v. 22, n. 2, p. 160–74, 1985. ISSN 0022-2844.
- HORUK, R. et al. The Duffy antigen receptor for chemokines: structural analysis and expression in the brain. *J Leukoc Biol*, v. 59, n. 1, p. 29–38, jan. 1996. ISSN 0741-5400.

- IWAMOTO, S. et al. Identification of a novel exon and spliced form of Duffy mRNA that is the predominant transcript in both erythroid and postcapillary venule endothelium. *Blood*, v. 87, n. 1, p. 378–85, jan. 1996. ISSN 0006-4971.
- IYER, J. et al. Invasion of host cells by malaria parasites: a tale of two protein families. *Mol Microbiol*, v. 65, n. 2, p. 231–49, jul. 2007. ISSN 0950-382X.
- KARLSSON, K. A. Animal glycosphingolipids as membrane attachment sites for bacteria. *Annu Rev Biochem*, v. 58, p. 309–50, 1989. ISSN 0066-4154.
- KEANE, T. M. et al. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol*, v. 6, p. 29, 2006. ISSN 1471-2148.
- KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, v. 16, n. 2, p. 111–20, dez. 1980. ISSN 0022-2844.
- LANGHI, D. M.; BORDIN, J. O. Duffy blood group and malaria. *Hematology*, v. 11, n. 5, p. 389–98, out. 2006. ISSN 1607-8454.
- LARKIN, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, v. 23, n. 21, p. 2947–8, nov. 2007. ISSN 1460-2059.
- LI, W.-H. *Molecular Evolution*. [S.l.]: Sinauer, 1997. ISBN 0878934804.
- MALLINSON, G. et al. Mutations in the erythrocyte chemokine receptor (Duffy) gene: the molecular basis of the Fya/Fyb antigens and identification of a deletion in the Duffy gene of an apparently healthy individual with the Fy(a-b-) phenotype. *Br J Haematol*, v. 90, n. 4, p. 823–9, ago. 1995. ISSN 0007-1048.
- MATHEW, S. et al. Confirmation of Duffy blood group antigen locus (FY) at 1q22→q23 by fluorescence in situ hybridization. *Cytogenet Cell Genet*, v. 67, n. 1, p. 68, 1994. ISSN 0301-0171.
- MENDIS, K. et al. The neglected burden of *Plasmodium vivax* malaria. *Am J Trop Med Hyg*, v. 64, n. 1-2 Suppl, p. 97–106, 2001. ISSN 0002-9637.
- MILLER, L. H. et al. Interaction between cytochalasin B-treated malarial parasites and erythrocytes. Attachment and junction formation. *J Exp Med*, v. 149, n. 1, p. 172–84, jan. 1979. ISSN 0022-1007.
- MILLER, L. H. et al. Erythrocyte receptors for (*Plasmodium knowlesi*) malaria: Duffy blood group determinants. *Science*, v. 189, n. 4202, p. 561–3, ago. 1975. ISSN 0036-8075.
- MOORE, S.; WOODROW, C. F.; MCCLELLAND, D. B. Isolation of membrane components associated with human red cell antigens Rh(D), (c), (E) and Fy. *Nature*, v. 295, n. 5849, p. 529–31, fev. 1982. ISSN 0028-0836.
- MOULDS, J. M.; MOULDS, J. J. Blood group associations with parasites, bacteria, and viruses. *Transfus Med Rev*, v. 14, n. 4, p. 302–11, out. 2000. ISSN 0887-7963.

- NEOTE, K. et al. Functional and biochemical analysis of the cloned Duffy antigen: identity with the red blood cell chemokine receptor. *Blood*, v. 84, n. 1, p. 44–52, jul. 1994. ISSN 0006-4971.
- NICKERSON, D. A.; TOBE, V. O.; TAYLOR, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*, v. 25, n. 14, p. 2745–51, jul. 1997. ISSN 0305-1048.
- NIELSEN, R. et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, v. 3, n. 6, p. e170, jun. 2005. ISSN 1545-7885.
- NIELSEN, R.; YANG, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, v. 148, n. 3, p. 929–36, mar. 1998. ISSN 0016-6731.
- OLSSON, M. L. et al. The Fy(x) phenotype is associated with a missense mutation in the Fy(b) allele predicting Arg89Cys in the Duffy glycoprotein. *Br J Haematol*, v. 103, n. 4, p. 1184–91, dez. 1998. ISSN 0007-1048.
- PALATNIK, M.; ROWE, A. W. Duffy and duffy-related human antigens in primates. *J Hum Evol*, v. 13, n. 2, p. 173–179, fev. 1984. ISSN 0047-2484.
- PARASOL, N. et al. A novel mutation in the coding sequence of the FY*B allele of the Duffy chemokine receptor gene is associated with an altered erythrocyte phenotype. *Blood*, v. 92, n. 7, p. 2237–43, out. 1998. ISSN 0006-4971.
- POGO, A. O.; CHAUDHURI, A. The Duffy protein: a malarial and chemokine receptor. *Semin Hematol*, v. 37, n. 2, p. 122–9, abr. 2000. ISSN 0037-1963.
- RIOS, M. et al. New genotypes in Fy(a-b-) individuals: nonsense mutations (Trp to stop) in the coding sequence of either FY A or FY B. *Br J Haematol*, v. 108, n. 2, p. 448–54, fev. 2000. ISSN 0007-1048.
- RODRÍGUEZ, F. et al. The general stochastic model of nucleotide substitution. *J Theor Biol*, v. 142, n. 4, p. 485–501, fev. 1990. ISSN 0022-5193.
- ROSSI, D.; ZLOTNIK, A. The biology of chemokines and their receptors. *Annu Rev Immunol*, v. 18, p. 217–42, 2000. ISSN 0732-0582.
- SAYLE, R. A.; MILNER-WHITE, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, v. 20, n. 9, p. 374, set. 1995. ISSN 0968-0004.
- TAMASAUŠKAS, D. et al. A homologous naturally occurring mutation in Duffy and CCR5 leading to reduced receptor expression. *Blood*, v. 97, n. 11, p. 3651–4, jun. 2001. ISSN 0006-4971.
- TAMURA, K. et al. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*, v. 24, n. 8, p. 1596–9, ago. 2007. ISSN 0737-4038.
- TAMURA, K.; NEI, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, v. 10, n. 3, p. 512–26, maio 1993. ISSN 0737-4038.

- TOURNAMILLE, C. et al. Sequence, evolution and ligand binding properties of mammalian Duffy antigen/receptor for chemokines. *Immunogenetics*, v. 55, n. 10, p. 682–94, jan. 2004. ISSN 0093-7711.
- TOURNAMILLE, C. et al. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet*, v. 10, n. 2, p. 224–8, jun. 1995. ISSN 1061-4036.
- TOURNAMILLE, C. et al. Close association of the first and fourth extracellular domains of the Duffy antigen/receptor for chemokines by a disulfide bond is required for ligand binding. *J Biol Chem*, v. 272, n. 26, p. 16274–80, jun. 1997. ISSN 0021-9258.
- TOURNAMILLE, C. et al. Arg89Cys substitution results in very low membrane expression of the Duffy antigen/receptor for chemokines in Fy(x) individuals. *Blood*, v. 92, n. 6, p. 2147–56, set. 1998. ISSN 0006-4971.
- WASNIOWSKA, K. et al. Identification of the Fy6 epitope recognized by two monoclonal antibodies in the N-terminal extracellular portion of the Duffy antigen receptor for chemokines. *Mol Immunol*, v. 33, n. 11-12, p. 917–23, ago. 1996. ISSN 0161-5890.
- WELLS, T. N.; PROUDFOOT, A. E.; POWER, C. A. Chemokine receptors and their role in leukocyte activation. *Immunol Lett*, v. 65, n. 1-2, p. 35–40, jan. 1999. ISSN 0165-2478.
- YANG; BIELAWSKI. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, v. 15, n. 12, p. 496–503, dez. 2000. ISSN 0169-5347.
- YANG, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, v. 15, n. 5, p. 568–73, maio 1998. ISSN 0737-4038.
- YANG, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, v. 24, n. 8, p. 1586–91, ago. 2007. ISSN 0737-4038.
- YANG, Z.; NIELSEN, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, v. 17, n. 1, p. 32–43, jan. 2000. ISSN 0737-4038.
- YANG, Z.; NIELSEN, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, v. 19, n. 6, p. 908–17, jun. 2002. ISSN 0737-4038.
- YANG, Z. et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, v. 155, n. 1, p. 431–49, maio 2000. ISSN 0016-6731.
- YAZDANBAKHSH, K. et al. Molecular mechanisms that lead to reduced expression of Duffy antigens. *Transfusion*, v. 40, n. 3, p. 310–20, mar. 2000. ISSN 0041-1132.

APÊNDICE A – Alinhamento global de nucleotídeos de todas as seqüências de primatas

		Intron										
		1120	1130	1140	1150	1160	1170	1180				
logo		$\begin{matrix} 2 \\ 1 \end{matrix} \left[\begin{matrix} \text{AGTTCCATGTTGGTCCCTTGGTGCCTTTCCTTCTGACCTTGCACTGCTCCTCCAGCCC} & \text{CAGC} & \text{TGCCCT} \end{matrix} \right] \begin{matrix} 2 \\ 1 \end{matrix}$										
Homo sapiens		AGTTCCATCCTGGTCTCTTGGTGCCTTTTCCTTCTGACCTTGCACTGCTCCTCCAGCCC--CAGC---TGCCCT										1187
Gorilla gorilla (00G0)		T.....C.....C.....G.....										1186
Cercopithecus galeritus chrysogaleus (00E7)		T.....C.....C.....										1115
Cercopithecus torquatus lunatus (00E1)		T.....C.....C.....										1122
Cercopithecus torquatus lunatus (00E5)		T.....C.....C.....										1122
Cercopithecus mitis (0E18)		T.....C.....C.....										1119
Cercopithecus mitis (0E21)		T.....C.....C.....										1121
Lophocebus aterrimus (00E2)		T.....C.....C.....										1122
Lophocebus aterrimus (0E11)		T.....C.....C.....										1122
Macaca fascicularis (MF11)		T.....C.....C.....										1122
Macaca fascicularis (MF14)		T.....C.....C.....										1122
Macaca fascicularis (MF16)		T.....C.....C.....										1122
Macaca fascicularis (MF18)		T.....CY.....C.....Y.....										1122
Macaca fascicularis (MF2)		T.....C.....C.....										1122
Macaca fascicularis (MF20)		T.....C.....C.....										1122
Macaca fascicularis (MF22)		T.....C.....C.....										1122
Macaca fascicularis (MF4)		T.....C.....C.....										1122
Macaca fascicularis (MF5)		T.....C.....C.....										1122
Macaca fascicularis (MF7)		T.....C.....C.....										1122
Macaca fascicularis (MF9)		T.....C.....C.....										1122
Macaca fascicularis (MFG2)		T.....C.....C.....										1122
Macaca fascicularis (MFJ1)		T.....C.....C.....										1122
Macaca fascicularis (MFN2)		T.....C.....C.....C.....										1123
Macaca mulatta (0E19)		T.....C.....C.....										1122
Macaca mulatta (MM1)		T.....C.....C.....										1122
Macaca mulatta (MM24)		T.....C.....C.....										1122
Macaca mulatta (MM27)		T.....C.....C.....										1122
Macaca mulatta (MM30)		T.....C.....C.....										1120
Macaca mulatta (MMD20)		T.....C.....C.....										1122
Macaca mulatta (MMS2)		T.....C.....C.....										1122
Macaca sp. (0E16)		T.....C.....C.....										1122
Macaca nemestrina (NG07921)		T.....C.....C.....										1122
Macaca nemestrina (NG08452)		T.....C.....C.....										1122
Macaca nemestrina (NG08610)		T.....C.....C.....										1122
Macaca nemestrina (PRO0024)		T.....C.....C.....										1122
Macaca nigra (NG7101)	C	T.....C.....C.....TC.....CAGC.....										1128
Macaca nigra (NG7102)	C	T.....T.....C.....TC.....CAGC.....										1128
Macaca nigra (PRO0726)	C	T.....T.....C.....TC.....CAGC.....										1128
Macaca thibetana (PRO0711)		T.....C.....C.....										1122
Mandrillus leucophaeus (0E14)		T.....C.....C.....										1122
Mandrillus Sphinx (0E20)		T.....C.....C.....										1122
Cebus apella (00CA)		G.C.C.C.....C.....GA.....T.....										1108
Saimiri sciureus (SSH29)		C.C.CA.....C-T.....C.....TG.....CA.....T.....										1118
Saimiri sciureus (SSH9)		C.C.CA.....C-T.....C.....TG.....CA.....T.....										1123
Saimiri ustus (SUH31)		C.C.CA.....C-T.....C.....TG.....CA.....T.....										1122
Saimiri ustus (SUI5)		C.C.CA.....C-T.....C.....TG.....CA.....T.....										1122

		Intron										
		1190	1200	1210	1220	1230	1240	1250	1260			
logo		<div><div>2</div><div>1</div><div>GGCTTCCCCAGGACTGTTCTGCTCCGGCTCTTCAGGCTCCCTGCTTTGCTCTTTCCACTGTCGGCACTGCA</div><div>2</div><div>1</div></div>										
Homo sapiens		GGCTTCCCC--AGGACTGTTCTGCTCCGGCTCTTCAGGCTCCCTGCTTTGCTCTTTCCACTGTCGGCACTGCA										1260
Gorilla gorilla (00G0)		G.....										1255
Cercopithecus galeritus chrysogaleus (00E7)		A.....C.....T.....										1188
Cercopithecus torquatus lunatus (00E1)		A.....C.....T.....										1195
Cercopithecus torquatus lunatus (00E5)		A.....C.....T.....										1195
Cercopithecus mitis (0E18)		A.....C.....T.....										1192
Cercopithecus mitis (0E21)		A.....C.....RT.....										1194
Lophocebus aterrimus (00E2)		AA.....C.....T.....										1195
Lophocebus aterrimus (0E11)		A.....C.....T.....										1195
Macaca fascicularis (MF11)		A.....C.....T.....										1195
Macaca fascicularis (MF14)		A.....C.....T.....										1195
Macaca fascicularis (MF16)		A.....C.....T.....										1195
Macaca fascicularis (MF18)		A.....C.....T.....										1195
Macaca fascicularis (MF2)		A.....C.....T.....										1195
Macaca fascicularis (MF20)		A.....C.....T.....										1195
Macaca fascicularis (MF22)		A.....C.....T.....										1195
Macaca fascicularis (MF4)		A.....C.....T.....										1195
Macaca fascicularis (MF5)		A.....C.....T.....										1195
Macaca fascicularis (MF7)		A.....C.....T.....										1195
Macaca fascicularis (MF9)		A.....C.....T.....										1195
Macaca fascicularis (MFG2)		A.....C.....T.....										1195
Macaca fascicularis (MFJ1)		A.....C.....T.....T.....										1195
Macaca fascicularis (MFN2)		C.....C.....T.....T.....										1197
Macaca mulatta (0E19)		A.....C.....T.....										1195
Macaca mulatta (MM1)		A.....C.....T.....										1195
Macaca mulatta (MM24)		A.....C.....T.....										1195
Macaca mulatta (MM27)		A.....C.....T.....										1195
Macaca mulatta (MM30)		A.....C.....T.....										1193
Macaca mulatta (MMD20)		A.....C.....T.....										1195
Macaca mulatta (MMS2)		A.....C.....T.....										1195
Macaca sp. (0E16)		A.....C.....T.....										1195
Macaca nemestrina (NG07921)		A.....C.....T.....										1195
Macaca nemestrina (NG08452)		A.....C.....T.....										1195
Macaca nemestrina (NG08610)		A.....C.....T.....										1195
Macaca nemestrina (PRO0024)		A.....C.....T.....										1195
Macaca nigra (NG7101)		A.....C.....T.....										1201
Macaca nigra (NG7102)		A.....C.....T.....										1201
Macaca nigra (PRO0726)		A.....C.....T.....										1201
Macaca thibetana (PRO0711)		A.....C.....T.....										1195
Mandrillus leucophaeus (0E14)		A.....C.....T.....										1195
Mandrillus Sphinx (0E20)		A.....C.....T.....										1195
Cebus apella (00CA)		CC.C.....A..C.....A.....C.....CA.....T..GC										1183
Saimiri sciureus (SSH29)		CC.C.....C.....A..C.....A.....C.....CA.....T..GC										1193
Saimiri sciureus (SSH9)		C.....C.....A..C.....A.....C.....CA.....T..GC										1196
Saimiri ustus (SUH31)		C.....C.....C.....A..C.....C.....CA.....T..GC										1195
Saimiri ustus (SUI5)		C.....C.....A..C.....A.....C.....CA.....T..GC										1195