IGOR RODRIGUES DA COSTA

ALFie: um programa para a busca exaustiva de locos anônimos E SUA VALIDAÇÃO EM GENOMAS DE *hOMINÍDEOS*

Rio de Janeiro

2015

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE BIOQUÍMICA MÉDICA LEOPOLDO DE MEIS

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA BIOLÓGICA

IGOR RODRIGUES DA COSTA

ALFIE: UM PROGRAMA PARA A BUSCA EXAUSTIVA DE LOCOS ANÔNIMOS E SUA VALIDAÇÃO EM GENOMAS DE HOMINÍDEOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Química Biológica, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre.

Orientador:Prof. Francisco Prosdocimi,Ph.D

Co-orientador: Prof. William Bryan Jennings, Ph.D

Rio de Janeiro

2015

|  |
| --- |
| Costa, Igor Rodrigues da  Alfie: um programa para a busca exaustiva de locos anônimos e sua validação em genomas de *hominídeos* / Igor Rodrigues da Costa. – Rio de Janeiro: UFRJ/IBqM, 2015.  XXII, 132 p.: il.; 29,7 cm.  Orientadores: Francisco Prosdocimi, William Bryan Jennings  Dissertação (mestrado) – UFRJ/IBqM/Programa de Pós-graduação em Química Biológica, 2015.  1. Locos Anônimos. 2. Bioinformática. 3. Genômica Comparativa. II. UniversidadeFederal do Rio de Janeiro, Instituto de Bioquímica Médica Leopoldo de Meis, Programa de Pós-graduação em Química Biológica. III. Alfie: um programa para a busca exaustiva de locos anônimos e sua validação em genomas de *hominídeos*. |

IGOR RODRIGUES DA COSTA

ALFIE: UM PROGRAMA PARA A BUSCA EXAUSTIVA DE LOCOS ANÔNIMOS E SUA VALIDAÇÃO EM GENOMAS DE hOMINÍDEOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Química Biológica, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre.

Aprovada em / /

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Francisco Prosdocimi, Ph.D, UFRJ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. William Bryan Jennings, Ph.D, UFRJ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. <Nome do membro da banca>, <título ex.: Ph.D, DSc>, UFRJ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. <Nome do membro da banca>, <título ex.: Ph.D, DSc>, FIOCRUZ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. <Nome do membro da banca>, <título ex.: Ph.D, DSc>, USP

A você, caro leitor

Agradecimentos

<Epígrafe>

Resumo

COSTA, Igor Rodrigues da. **Alfie: um programa para a busca exaustiva de locos anônimos e sua validação em genomas de hominídeos**. 2015. Tese (Mestrado em Química Biológica) – Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

As novas tecnologias de sequenciamento provocaram um aumento exponencial na quantidade de genomas completos e parciais disponíveis. Estes genomas podem ser utilizados para a predição de marcadores genéticos ideais para genética de populações, em regiões sob seleção neutra e com segregação independente, chamados de locos anônimos. Este trabalho descreve a criação de um método computacional de alta eficiência para a predição de locos anônimos em genomas completos ou parciais. O programa, chamado *alfie,* foi criado usando a linguagem de programação python em conjunto com a biblioteca biophyton e contém 2 500 linhas de código. Para encontrar os locos anônimos, o *alfie* seleciona regiões com distância maior que 200 kb de genes ou de outros locos e procura estas regiões em outros genomas, filtrando apenas os locos com cópia única. O *alfie* foi aplicado em quatro genomas de Hominídeos: *Homo* sapiens (humano), *Gorilla gorilla* (gorila), *Pan troglodytes* (chimpanzé) e *Pongo abelii* (orangotango), gerando um conjunto de 300 locos anônimos. Estes locos anônimos foram submetidos à análise filogenética, predição de modelos de substituição e estimativa de população ancestral efetiva e tempo de divergência. Os parâmetros de tamanho efetivo de população ancestral foram estimados entre 37 000 e 50 000, 40 000 e 43 000, 72 000 e 95 000 para as linhagens humano-chimpanzé, gorila e orangotango, respectivamente. Já os parâmetros de tempo de divergência foram estimados em 4,3; 6,1 e 12,3 milhões de anos, respectivamente para os mesmos clados. Esses parâmetros foram obtidos com um intervalo de confiança de 95% consideravelmente menor do que os dados publicados previamente, devido a quantidade imprecedente de locos não ligados obtidos neste trabalho.

Palavras-chave: Bioinformática, genômica, locos anônimos, genética populacional

Abstract

COSTA, Igor Rodrigues da. **Alfie: um programa para a busca exaustiva de locos anônimos e sua validação em genomas de hominídeos**. 2015. Tese (Mestrado em Química Biológica) – Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

New sequencing tecnologies have caused an exponential increase in complete and parcial genome availability. These genomes can be useful for ideal genetic marker prediction, to be used in population genetics. Such markers, called anonymous loci, should be under neutral selection, single copy and be independently segregated. This work describes the development of a new high efficience computational method for anonymous loci prediction in complete or partial genomes. The program, named *alfie* (anonymous loci finder), was created using python programming language along with biopython library and contains 2 500 lines of code. In order to find anonymous loci, *alfie* selects regions further than 200 kb from genes and other loci and search for homologous regions in other genomes, filtering only single copy loci. *Alfie* was validated using four *hominidae* genomes: *Homo* sapiens (human), *Gorilla gorilla* (gorilla), *Pan troglodytes* (chimpanzee) e *Pongo abelii* (orangutan), generating a 300 anonymous loci dataset. These loci were submited to phylogenetic analysis, substitution model prediction and population genetic parameter estimation. The ancestral effective populations sizes were estimated to be between 37 000 and 50 000, 40 000 and 43 000, 72 000 and 95 000 for the human-chimp, gorilla and orangutan lineages, respectively. Likewise, the divergence time parameters were estimated to be 4,3; 6,1 and 12,3 millions of years, respectively for the same clades. These parameters had a considerably shorter 95% confidence interval than previously published data, a direct consequence of the unprecedent large number of unlinked loci studied.

Keywords: Bioinformatics, genomics, anonymous loci, population genetics

Lista de Figuras

[Figura 1: Algoritmo do *alfie* 21](#_Toc428155929)

[Figura 2: Distribuição dos ALs putativos e regiões anônimas 25](#_Toc428155930)

[Figura 3: Distribuição dos ALs encontrados 25](#_Toc428155931)

[Figura 4: Modelos de substituição preditos 26](#_Toc428155932)

[Figura 5: Estimativa do tempo de divergência dos *Hominidae* 29](#_Toc428155933)

[Figura 6: Estimativa do tamanho efetivo de população 29](#_Toc428155934)

[Figura 7: Histograma dos valores de Ti/Tv encontrados 30](#_Toc428155935)

Lista de Tabelas

[Tabela1: Sumário dos resultados encontrados 33](#_Toc428155939)

Lista de Siglas

|  |  |
| --- | --- |
| AL | Loco Anônimo *(Anonymous Loci)* |
| SNP | Polimorfismo de Nucleotídeo Único *(Single Nucleotide Polymorphism)* |
| GFF | *General Feature Format* |
| NGS | Sequenciamento de Nova Geração *(Next Generation Sequencing)* |
| bp | Pares de Base *(Base Pair)* |
| PCR | Reação de Polimerase em Cadeia *(Polymerase Chain Reaction)* |

Sumário

[1 Introdução 13](#_Toc428155988)

[1.1 Genômica 13](#_Toc428155989)

[1.1.1 Sequenciamento de nova geração 13](#_Toc428155990)

[1.1.2 Genomas completos e parciais 13](#_Toc428155991)

[1.2 Genômica Computacional 14](#_Toc428155992)

[1.2.1 Programas para análise de dados 14](#_Toc428155993)

[1.2.2 Bancos de dados 14](#_Toc428155994)

[1.3 Genética de População 14](#_Toc428155995)

[1.3.1 O modelo dos *Hominidae* na genética de população 15](#_Toc428155996)

[1.4 Marcadores genéticos 15](#_Toc428155997)

[1.4.1 Utilidade dos marcadores genéticos 16](#_Toc428155998)

[1.4.2 Microssatélites 16](#_Toc428155999)

[1.4.3 SNPs 16](#_Toc428156000)

[1.4.4 Locos anônimos 17](#_Toc428156001)

[2 Objetivo 19](#_Toc428156002)

[2.1 Objetivos específicos 19](#_Toc428156003)

[3 Desenvolvimento 20](#_Toc428156004)

[3.1 Programas, bibliotecas e linguagem 20](#_Toc428156005)

[3.2 Alfie 20](#_Toc428156006)

[3.3 Busca de regiões anônimas 20](#_Toc428156007)

[3.3.1 Locos putativos 21](#_Toc428156008)

[3.4 Busca por homólogos 22](#_Toc428156009)

[3.5 Alinhamento múltiplo 22](#_Toc428156010)

[3.6 Outros recursos do pacote *alfie* 22](#_Toc428156011)

[4 Resultados e Discussão 24](#_Toc428156012)

[4.1 Busca de locos anônimos em genomas completos de *Hominidae* 24](#_Toc428156013)

[4.1.1 Busca por regiões anônimas 24](#_Toc428156014)

[4.1.2 Filtragem por conservação e unicidade 24](#_Toc428156015)

[4.2 Análise dos locos anônimos de *Hominidae* 24](#_Toc428156016)

[4.2.1 Mapeamento dos locos anônimos por cromossomo 25](#_Toc428156017)

[4.2.2 Modelos de substituição 26](#_Toc428156018)

[4.2.3 Filogenia 27](#_Toc428156019)

[4.2.4 Estimativa da população efetiva e tempo de divergência 27](#_Toc428156020)

[4.2.5 Teste de neutralidade 30](#_Toc428156021)

[5 Conclusão 31](#_Toc428156022)

[Anexos 32](#_Toc428156023)

[ANEXO A – DECAIMENTO exponencial do custo de sequenciamento 32](#_Toc428156024)

[Apêndices 33](#_Toc428156025)

[APÊNDICE A – Sumário dos resultados encontrados 33](#_Toc428156026)

[Referências 42](#_Toc428156027)

1 Introdução

## Genômica

A palavra “genômica” surgiu em uma discussão de bar para decidir o nome de uma nova revista científica em 1986 [[1](#_ENREF_1)], focada em dados gerados pela exploração e comparação de genomas. Esta nova área nasceu junto com a era do “projeto genôma humano”, e trouxe inovações para a ciência e a sociedade, criando novos campos como a medicina personalizada e a filogenômica. O campo da genômica hoje está em franca expansão, uma projeção atual indica que, se a taxa de crescimento se mantiver, são esperados que 10.000 genomas de vertebrados sejam sequenciados nos próximos 10 anos [[2](#_ENREF_2)].

### Sequenciamento de nova geração

A nova geração de técnicas para sequenciamento de DNA tem feito com que este seja cada vez mais barato, eficiente e livre de erros [[3](#_ENREF_3)]. A última década viu um decaimento exponencial do preço de sequenciamento (ANEXO A) que influenciou diretamente no crescimento exponencial da quantidade de genomas completos e parciais disponíveis e publicados [[3](#_ENREF_3), [4](#_ENREF_4)]. As técnicas hoje chamadas de *Next Generation Sequencing* (NGS) incluem diversas plataformas e tecnologias, tendo como fator comum o baixo preço e a enorme quantidade de dados gerados por corrida, superando em muitas ordens de grandeza (por volta de 108) o método de sequenciamento de Sanger.

### Genomas completos e parciais

A proliferação destas novas técnicas gerou uma avalanche de projetos que buscam a descrição inicial dos genomas de espécies modelo e também, cada vez mais, de qualquer organismo de interesse. O sequenciamento completo de um genoma, ainda hoje, não é uma tarefa trivial, pois regiões altamente repetitivas são uma barreira para os algoritmos de montagem, já que as técnicas de NGS produzem sequências curtas [[5](#_ENREF_5)]. A anotação deste genoma, que envolve a indentificação de regiões codificantes e regulatórias, é um processo contínuo e iterativo, novas versões de genomas modelo, como o do *Homo Sapiens,* são lançadas regularmente. Portanto, é muito comum que genomas sejam disponibilizados em estágios iniciais de montagem e anotação, como genomas parciais.

## Genômica Computacional

Da necessidade de explorar essa grande coleção de dados biológicos, normalmente medidos em gigabytes, que se acumula, surgiu a genômica computacional. Esta área se concentra na criação e aplicação de ferramentas para análise, comparação e descrição de genes, genomas e sistemas.

### Programas para análise de dados

A magnitude da produção de dados torna inviável a análise manual para a maioria dos casos, por isso o uso de programas de computador em laboratórios de genômica é essencial. Predição de genes, montagem de genomas, reconstrução filogenética e alinhamento múltiplo de sequências são algumas das tarefas realizadas hoje quase que exclusivamente por programas de computador especializados. O presente trabalho conta com a utilização intensiva, e principalmente, a criação de ferramentas computacionais, o que permite um ganho de produtividade e rendimento relativo a técnicas manipulação *in vitro*.

### Bancos de dados

Bancos de dados são a matéria-prima dos estudos computacionais em problemas que envolvem “*Big Data*”. Na biologia computacional, os bancos de dados são gerados para organizar o conhecimento de modo que futuros estudos sejam facilitados. Estes bancos são disponibilizados na internet em grandes plataformas, como a do NCBI para genes [[6](#_ENREF_6)], taxonomias [[7](#_ENREF_7)], proteínas [[8](#_ENREF_8)], ou o ENSEMBL [[9](#_ENREF_9)], que é um banco de dados de genomas completos para organismos modelo.

## Genética de População

Para entender a dinâmica da evolução, migração e a história de diversos indivíduos de uma mesma espécie, é preciso observá-los no nível molecular. As pistas de eventos passados podem ser observadas no DNA, que é muitas vezes a única maneira de reconstruir os passos das populações e entender as pressões seletivas atuantes.

### O modelo dos *Hominidae* na genética de população

The divergence dates between human and other hominoids also have received much attention [[10](#_ENREF_10), [11](#_ENREF_11)]. However, there is still much uncertainty about these dates, especially the internodal time span between the human-chimpanzee divergence and the branch node of the gorilla lineage. For example, according to Horai et al.’s study of mitochondrial DNA sequences [[12](#_ENREF_12)], the divergence between human and chimpanzee occurred 4.7 ± 0.5 million years ago, whereas the gorilla lineage branched off 7.7 ± 0.7 million years ago, so the time span between the two speciation events was as long as 3 million years, ∼60% of the Homo-Pan divergence time. In contrast, in Bailey et al.’s study of the η-globin pseudogene region [[13](#_ENREF_13)], the internodal time span was only ∼10% of the divergence time between human and chimpanzee. Which is closer to the truth? As will be explained later, this internodal time span is useful for estimating the effective size of the ancestral population before the human-chimpanzee divergence.

There has also been considerable interest in the demographic history of hominoids [[11](#_ENREF_11), [14](#_ENREF_14), [15](#_ENREF_15)]. Of particular interest is the effective size (Ne) of the ancestral population before the human-chimpanzee divergence, because it may tell us whether there has been a significant reduction in population size in the human lineage since its separation from the chimpanzee lineage. Current data are not sufficient for a reliable estimate of Ne; for this purpose, a fairly large number of independent loci from the human, chimpanzee, and gorilla genomes are needed. Falar do artigo Science [[16](#_ENREF_16)].

## Marcadores genéticos

Marcadores genéticos são sequências de DNA que, por apresentarem variações, podem ser usadas para diferenciar organismos ou espécies. O primeiro marcador genético a ser desenvolvido, 30 anos atrás, se aproveitava de variações nos sítios alvo de enzimas de restrição para criar *fingerprints* genéticos e são usados até hoje na ciência forense e em testes de paternidade [[17](#_ENREF_17)]. Avanços nas técnicas de sequenciamento nos últimos 30 anos permitiram o desenvolvimento de diversos novos marcadores genéticos, como microssatélites, polimorfismos de nucleotídeo único (SNP, *single nucleotide polymorphism*) e os locos anônimos (AL, *anonymous loci*).

### Utilidade dos marcadores genéticos

Os marcadores genéticos são ferramentas versáteis, sendo úteis em campos como a ciência forense, testes de paternidade, medicina personalizada e diagnóstica, genética de população, filogenia e agropecuária, entre outros.

### Microssatélites

Um dos marcadores genéticos mais usados são os microssatélites, que são regiões hiper-variáveis compostas por repetições de pares de bases ou de motivos simples (como di, tri, tetra, penta e até hexa-nucleotídeos). A diferenciação de organismos e populações é feita pela variação do tamanho desta região, que é normalmente obtida por meio de amplificação por PCR ou sequenciamento direto. Locos de microssatélites são comumente conservados em espécies relacionadas, tornando possível a utilização do mesmo marcador para espécies do mesmo gênero.

### SNPs

Outra classe de marcadores genéticos, os polimorfismos de nucleotídeos únicos (SNP, *Single Nucleotide Polymorphism*) são variações de uma única basepresentes em pelo menos 1% da população. Estes marcadores são muito usados em estudos de associação genômica, onde se busca correlacionar a variação genotípica com características fenotípicas. Além disso, são também usados para estudar a migração e a filogeografia de populações, incluindo a dos seres humanos. O baixo custo para predição de SNPs, decorrente das técnicas de pirosequenciamento [[18](#_ENREF_18)] facilitou o desenvolvimento de bibliotecas contendo milhões de SNPs.

### Locos anônimos

Locos anônimos (ALs), os marcadores de interesse neste trabalho, são regiões não codificantes com características ideais para estudos de genética de população e filogenia. Essas características são:

1. Apresentar cópia única no genoma;
2. Estar sob seleção neutra;
3. Segregar independentemente.

As técnicas mais comumente usadas para descrição de ALs são baseadas na amplificação por PCR e sequenciamento de regiões aleatórias do genoma e, portanto, produzem marcadores que possivelmente estão associados a genes. Considerando que aproximadamente 90%-95% do genoma humano é não codificante [[19](#_ENREF_19)], é esperado que por volta de 10% dos locos retirados de regiões amplificadas aleatoriamente estejam em regiões gênicas. Uma fração ainda maior pode estar em regiões regulatórias próximas a estes genes, sujeita a seleção não-neutra. Estes marcadores requerem curadoria manual para filtragem dos fragmentos amplificados localizados em regiões codificantes após seu sequenciamento, o que traz uma dificuldade maior para o desenvolvimento de um número expressivo de marcadores. Por ser uma técnica árdua, muitos estudos usam por volta de 10 a 50 locos [[20-22](#_ENREF_20)], levando meses ou semanas para a obtenção destes locos.

Locos anônimos apresentam diversas vantagens sobre marcadores convencionais como SNPs e microssatélites. Primeiramente, são marcadores mais informativos, pelo simples fato de serem maiores em número de bases. Os fragmentos sequenciados costumam ter em torno de 1 kb, muito maiores que os 100-500 bp dos microssatélites ou 1 bp para os SNPs. Esta vantagem também faz com que este marcador possa ser usado numa extensão maior de tempo de divergência entre espécies enquanto que técnicas baseadas em microssatélites são muito efetivas para análise populacional, mas falham em estudos que envolvem espécies diferentes por variações na região de ligação do *primer* [[23](#_ENREF_23)].

Outra vantagem é a presença de diversos tipos de variação, como substituição, deleção e inserção, no mesmo marcador, permitindo a estimativa mais precisa para parâmetros de taxa de mutação. Por outro lado, ALs são mais difíceis de obter, são menos eficientes que marcadores gênicos para espécies distantes e não tem utilização tão difundida quanto SNPs ou microssatélites.

#### Estado da arte na descoberta e descrição de locos anônimos

Técnicas *in silico* para descoberta de locos anônimos em dados de NGS são um desenvolvimento recente [[24](#_ENREF_24)], porém ainda não existem técnicas que permitam garantir que estes marcadores não estejam associados a regiões sob seleção ou de cópia única. O presente trabalho apresenta uma nova solução para estes problemas, se aproveitando da disponibilidade de genomas completos em bancos de dados na internet.

# Objetivo

Desenvolver uma metodologia *in silico* para descoberta de locos anônimos em genomas completos e aplicá-la no modelo dos *Hominidae*.

## Objetivos específicos

* Produzir um programa para obter locos distantes de genes em arquivos completos de genoma com anotação de regiões regulatórias;
* Encontrar os ortólogos dos locos nas espécies estudadas;
* Verificar a distribuição dos locos encontrados ao longo do genoma de referência;
* Reconstruir árvores filogenéticas e estimar o modelo de substituição para cada loco encontrado em todos os genomas;
* Estimar os parâmetros de população ancestral efetiva e tempo de divergência para o grupo dos hominídeos.

# Desenvolvimento

## Programas, bibliotecas e linguagem

Este programa foi desenvolvido na linguagem Python, versão 2.7, com o auxílio da biblioteca Biopython [[25](#_ENREF_25)] e de programas auxiliares: BLAST+ [[26](#_ENREF_26)] para busca de sequências e PhyML [[27](#_ENREF_27)] para reconstrução de árvores filogenéticas. O programa foi chamado de *alfie* (*anonymous loci finder*) e disponibilizado com licença de código aberto GPL. O *alfie* pode ser obtido a partir do repositório git hospedado no *github*: https://github.com/igorrcosta/alfie. Neste site também está um manual completo de uso do programa (em inglês).

## Alfie

O*alfie* é um programa que efetua os diversos passos e filtros para busca e obtenção exaustiva de locos anônimos em genomas completos. A interface atual do programa é por linha de comando, usando *flags* para configurar as opções de análise. O *alfie* recebe, como arquivos de entrada, dois ou mais genomas em formato FASTA, sendo um deles o genoma de referência, e um arquivo GFF de anotação do genoma de referência, retornando os ALs encontrados em todos os genomas.

## Busca de regiões anônimas

O primeiro passo do programa é, a partir de um genoma referência e um arquivo contendo a anotação das regiões funcionais (GFF), selecionar todas as regiões que estão a uma distancia superior a 200 kb de qualquer gene (em uma análise mais conservadora, podem ser excluídas também as regiões próximas a pseudogenes). Regiões próximas (10 kb) dos telômeros, que são as regiões terminais dos cromossomos, também são excluídas. (Fig. 1 A)

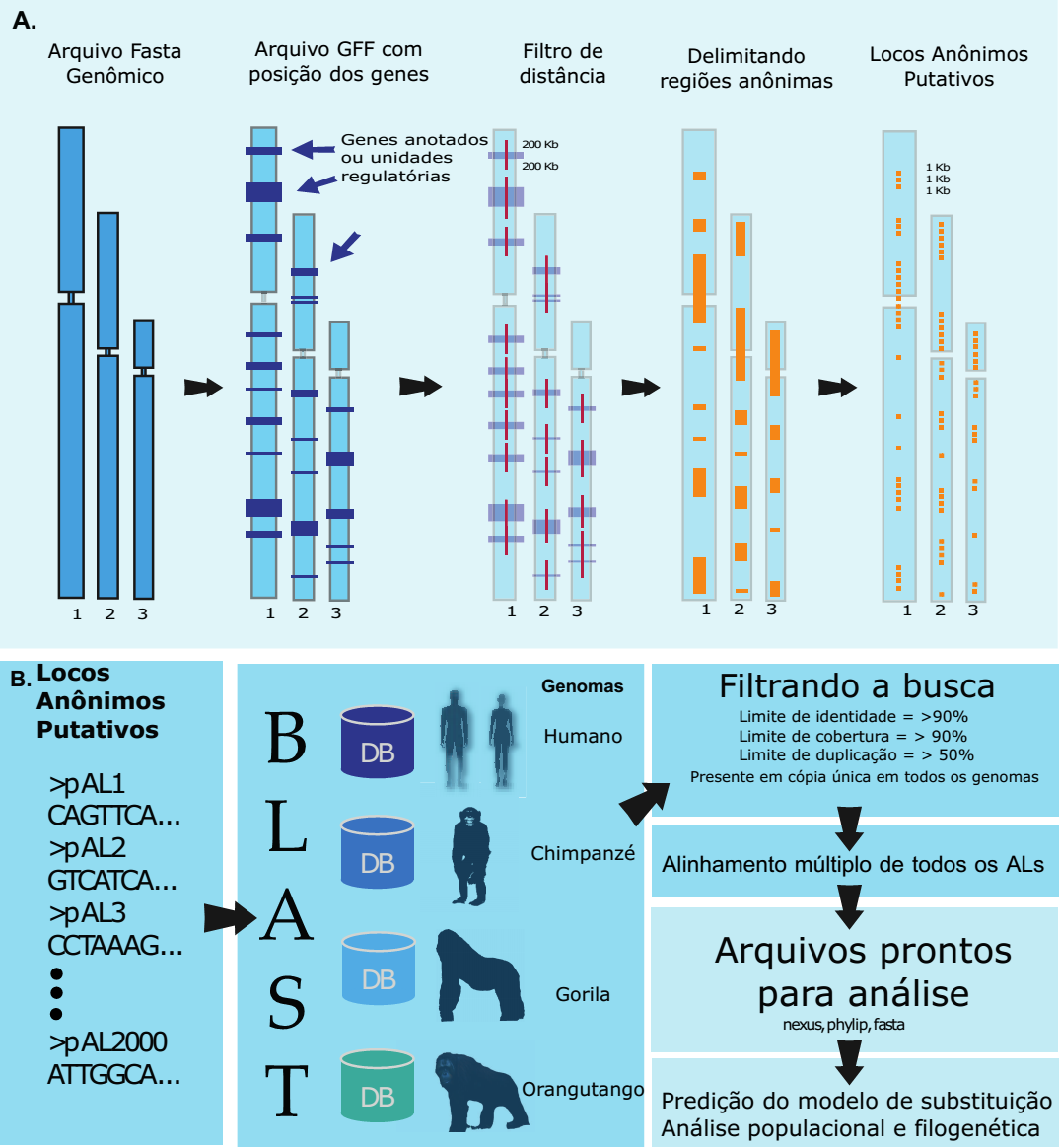


Figura 1: Algoritmo do *alfie*

### Locos putativos

As regiões anônimas do genoma de referência são recortadas em fragmentos de 1 kb, de modo diminuir a chance de que eventos de recombinação tenham ocorrido dentro do próprio loco. Estes fragmentos são recortados apenas de regiões anônimas sem bases não identificadas (“N”s). Ao fim, esses fragmentos, chamados de locos anônimos putativos, são salvos em um arquivo FASTA (Fig. 1 A).

## Busca por homólogos

Para gerar um banco de dados útil para análise filogenética o programa busca locos homólogos aos locos putativos do genoma de referência (Fig. 1 B). Os locos anônimos putativos são usados como *query* para busca em todos os genomas estudados. Usando os parâmetros padrões ao executar o programa, uma cópia homóloga é definida por no mínimo 90% de cobertura e 90% de identidade com a sequência de referência. Do mesmo modo, é considerada repetição qualquer sequência que possua mais de uma cópia com mais de 50% de identidade e 90% de cobertura.

Apenas os locos encontrados sem repetições em todos os genomas são selecionados para o próximo passo da análise.

## Alinhamento múltiplo

As sequências homólogas encontradas na busca por BLAST são extraídas dos genomas e salvas em arquivos FASTA. Cada arquivo é alinhado separadamente pelo programa ClustalW [[28](#_ENREF_28)]. São selecionados apenas as sequências que possuam acima de um certo número de bases alinhadas (por padrão, 500 bp). Destes locos são escolhidos o número máximo de locos de modo que todos estejam a uma distância superior a 200 kb do loco mais próximo, garantindo a segregação independente destes locos, que agora já podem ser considerados verdadeiros locos anônimos.

Os alinhamentos são salvos em diversos formatos, como FASTA, NEXUS e PHILIP, facilitando análises posteriores.

## Outros recursos do pacote *alfie*

Os passos realizados pelo programa *alfie* também podem ser realizados independentemente e com a adição de novos parâmetros e etapas. Um preditor de *primer* baseado no programa *primer3* está incluído, tornando possível, por exemplo,construir *primers* de locos putativos, retirados de um genoma referência, para amplificar e sequenciar em espécies próximas que não possuam o genoma completo ou parcial disponíveis

Outras etapas que podem ser realizadas ao final da análise são a reconstrução filogenética e predição dos modelos de substituição, realizados por *scripts* que executam os programas *PhyML* [[27](#_ENREF_27)] e *modeltest* [[29](#_ENREF_29)], respectivamente.

O *alfie* também pode ser configurado para encontrar regiões próximas a genes para comparação, utilizando distâncias negativas no parâmetro da distância gênica. Por exemplo, definindo o parâmetro de distância gênica em −2 000 bp, os locus serão extraídos a uma distância máxima de 2 kb dos genes.

# Resultados e Discussão

Após o desenvolvimento do *alfie*, este foi aplicado para a busca de locos anônimos nos genomas de Humano, Chimpanzé, Gorila e Orangotango. Os ALs encontrados foram analisados para obtenção de uma nova estimativa dos parâmetros populacionais destas espécies.

## Busca de locos anônimos em genomas completos de *Hominidae*

Os genomas utilizados foram os seguintes: *Homo sapiens* versão 38, *Pan troglodytes* versão 2.1.4, *Gorilla gorilla* versão 3.1 e *Pongo abelii* versão 2, todos eles com repetições e regiões de baixa complexidade mascaradas.

### Busca por regiões anônimas

Foram encontrados aproximadamente 247 Mb em regiões anônimas no genoma humano, equivalente a 8% do genoma. Destas regiões, 228 Mb (92,5%) foram mascaradas por serem de baixa complexidade ou representavam bases não identificadas (“N”s). A partir das regiões anônimas restantes (18 Mb, ~0,6% do genoma total) foram extraídos 4 233 locos anônimos putativos, cada um deles com 1 kb de extensão. Os locos putativos encontrados apresentaram uma distribuição cromossomal proporcional ao tamanho da região anônima não mascarada de cada cromossomo (Fig. 2).

### Filtragem por conservação e unicidade

Os 4 233 locos anônimos putativos encontrados no genoma humano foram usados como *query* na busca por homólogos e duplicações em todos os genomas estudados. Nesta busca foram encontrados apenas 304 locos anônimos que apresentam cópia única e alta identidade em cada um dos genomas.

## Análise dos locos anônimos de *Hominidae*

Do conjunto de 304 ALs gerados foram aleatoriamente selecionados 300 ALs para análise detalhada.

### Mapeamento dos locos anônimos por cromossomo

A distribuição de ALs no genoma foi similar a distribuição de ALs putativos (Fig. 3) apresentando uma relação linear entre o tamanho e o número de ALs para a maioria dos cromossomos. As exeções podem ser explicadas por variações nos tamanhos de regiões anônimas não repetitivas de cada cromossomo (Fig. 2).

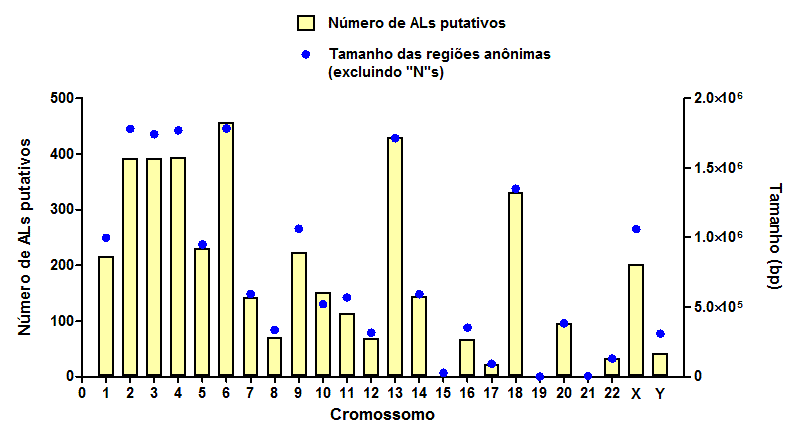


Figura 2: Distribuição dos ALs putativos e regiões anônimas

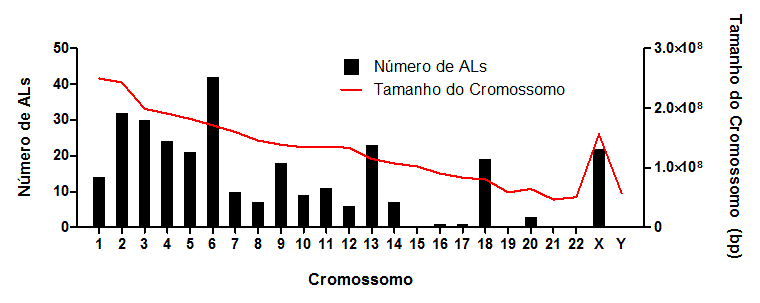


Figura 3: Distribuição dos ALs encontrados

### Modelos de substituição

Os 300 ALs foram submetidos a predição de modelos de substituição usando o programa *modeltest* (Fig. 4). Dentre os 88 modelos possíveis [[30](#_ENREF_30)], a grande maioria dos ALs (210/300) foi mais bem representada pelo modelo HKY (Hasegawa, Kishino e Yano) [[31](#_ENREF_31)] que consiste em apenas cinco parâmetros, um para cada base e um para a taxa de transição/transversão (Ti/Tv). Este resultado demonstra que as regiões anônimas do genoma de *Hominidae* estão sujeitas, basicamente, a um modo de evolução simples, compatível com a hipótese de evolução neutra.

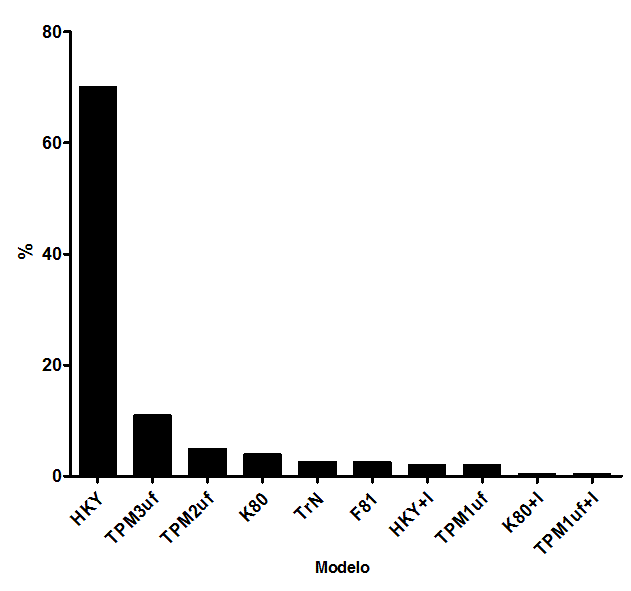


Figura 4: Modelos de substituição preditos

### Filogenia

Na análise filogenética dos 300 ALs, gerando uma árvore enraizada para cada loco, 64% das árvores exibiram a relação esperada entre humanos e chimpanzé (193/300), e por volta de 20% apresentaram as duas topologias alternativas ((H,G),C) e ((C,G),H) (Tabela 1).

Estes resultados concordam com a frequência de topologias esperadas pela teoria da coalescência [[32](#_ENREF_32)] e resultados empiricos publicados [[16](#_ENREF_16)].

### Estimativa da população efetiva e tempo de divergência

A estimativa dos parâmetros populacionais obteve resultados semelhantes aos já publicados [[16](#_ENREF_16), [33](#_ENREF_33)]. Os resultados foram comparados com os de Chen e Li de 2001 [[16](#_ENREF_16)], que obteve estas estimativas a partir de um conjunto de 53 locos nucleares e é um dos trabalhos mais citados da área.

Foi observado que o intervalo de confiança de 95% das estimativas de tempo de divergência foram reduzidos entre duas a três vezes (Fig. 5) nas estimativas geradas a partir do conjunto de 300 locos. Foram estimados em 4,3 milhões de anos o tempo de divergência entre humanos e chimpanzés e em 12,3 milhões de anos o tempo de divergência da linhagem do orangotango, 0,5 e 1,8 milhões de anos mais recente do que o estimado utilizando o conjunto de 53 locos respectivamente. Ambos conjuntos concordaram no tempo de divergência do gorila em valores próximo a 6 milhões de anos (5,9 utilizando 53 locos e 6,1 utilizando 300).

Para os parâmetros de tamanho efetivo de população ancestral as análises foram feitas usando tempo de geração (TG) variável ou fixo em 20 anos. Analizando o conjunto de 300 locos, as estimativas para o ancestral humano-chimapanzé variaram entre 37 000 e 50 000, dependendo do TG utilizado, enquanto que as estimativas usando 53 locos variaram entre 17 000 e 23 000 (Fig. 6). De modo similar, o tamanho de população efetiva para o ancestral humano, chimpanzé, gorila foi estimada entre 40 000 e 43 000 para todos os conjuntos de dados.

Houve, porém, uma diferença considerável nas estimativas para o tamanho de população efetiva da linhagem hominóide ancestral. O *dataset* de 53 locos forneceu uma estimativa variando entre 21 000 e 33 000 enquanto que os valores estimados a partir do conjunto de 300 locos foi bem maior, entre 72 000 e 95 000. Não foram observadas diferenças no intervalo de confiança para a linhagem ancestral humano-chimpanzé, mas as outras duas estimativas (ancestral humano-chimpanzé-gorila e hominóide) tiveram seus intervalos reduzidos pela metade.

Todas estas análises foram repetidas e resultados semelhantes foram obtidos utilizando priores bayesianos exponenciais, demonstrando robustez para a escolha de priores.

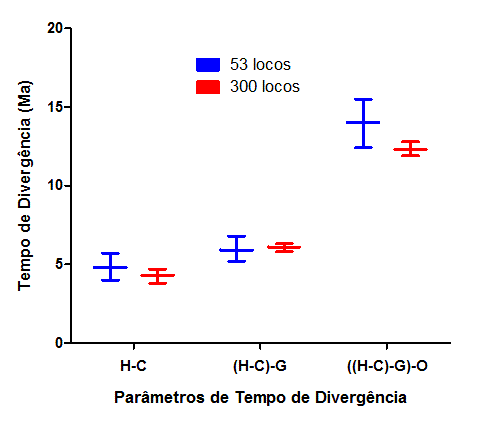


Figura 5: Estimativa do tempo de divergência dos *Hominidae*

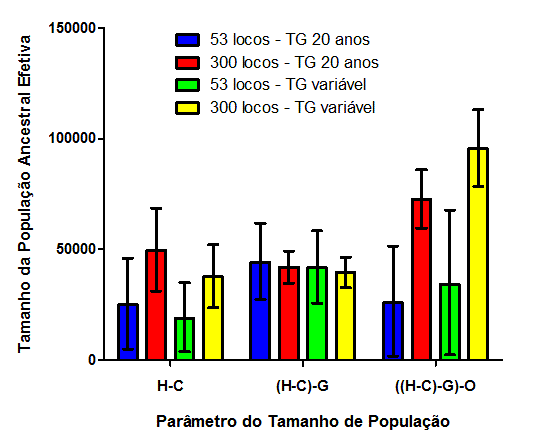


Figura 6: Estimativa do tamanho efetivo de população

### Teste de neutralidade

Os valores da razão transição/transverção (Ti/Tv), que é um dos parâmetros estimados pelo programa *modeltest*, apresentaram uma distribuição unimodal de média 2,3 ± 0,9 (Fig. 7). Esta distribuição é uma evidência de que sitios intergênicos neutros seguem um único modo de evolução.

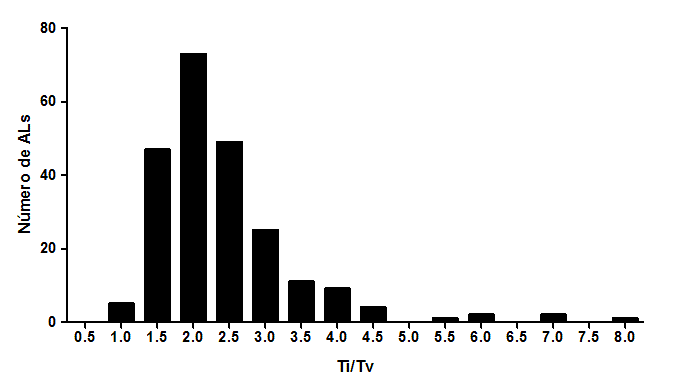


Figura 7: Histograma dos valores de Ti/Tv encontrados

5 Conclusão

Neste trabalho foi desenvolvido um programa que sugere uma solução para o problema da predição de locos anônimos em genomas completos ou parciais usando uma abordagem totalmente computacional (*in silico*). Embora o tempo de desenvolvimento da ferramenta tenha sido de cerca de 3 anos, agora esse problema pode ser resolvido com extrema facilidade em algumas horas, facilitando enormemente o trabalho de busca por locos anônimos. Este programa foi capaz de gerar 300 locos anônimos conservados para o grupo dos *Hominidae*, utilizando quatro genomas completos. Estes 300 locos foram analisados e foram obtidos resultados compatíveis com trabalhos publicados, validando o programa. Além disso, foi obtido um valor de Ti/Tv médio para as regiões não codificantes de hominóides. Finalmente, foram obtidas estimativas de parâmetros populacionais de hominídeos mais precisos que os já publicados. Esperamos que o nosso trabalho se demonstre como o novo padrão a ser comparado para as estimativas de parâmetros populacionais dos *Hominidae*.

Anexos

ANEXO A – DECAIMENTO exponencial do custo de sequenciamento



Extraído de <http://www.genome.gov/sequencingcosts/>

Apêndices

APÊNDICE A – Sumário dos resultados encontrados

Tabela1: Sumário dos resultados encontrados

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Cromossomo | Posição Inicial | Posição Final | Topologia | Ti/Tv | Modelo de Substituição |
| 1 | 1 | 68744514 | 68745513 | ((C,G),H) | 2.77 | HKY+I |
| 5 | 1 | 72507917 | 72508916 | ((H,C),G) | N.A. | TPM3uf |
| 9 | 1 | 79725749 | 79726748 | ((H,C),G) | N.A. | TPM3uf |
| 37 | 1 | 88108901 | 88109900 | ((H,G),C) | 1.94 | HKY |
| 50 | 1 | 104474803 | 104475802 | ((H,C),G) | 2.04 | HKY |
| 78 | 1 | 104827235 | 104828234 | ((H,C),G) | N.A. | TPM2uf |
| 89 | 1 | 105036129 | 105037128 | ((H,C),G) | 1.93 | HKY |
| 114 | 1 | 106289093 | 106290092 | ((H,G),C) | N.A. | TPM3uf |
| 132 | 1 | 164141460 | 164142459 | ((H,C),G) | 2.91 | HKY |
| 133 | 1 | 189343013 | 189344012 | ((H,G),C) | 2.39 | HKY |
| 146 | 1 | 191440464 | 191441463 | ((C,G),H) | 1.52 | HKY |
| 172 | 1 | 195326778 | 195327777 | ((H,C),G) | 2.62 | HKY |
| 191 | 1 | 195534484 | 195535483 | ((H,C),G) | 2.07 | HKY |
| 210 | 1 | 195752207 | 195753206 | ((H,C),G) | 0.97 | HKY |
| 219 | 10 | 9508281 | 9509280 | ((H,C),G) | 2.42 | HKY |
| 224 | 10 | 55958449 | 55959448 | ((H,C),G) | 1.78 | HKY |
| 253 | 10 | 56838042 | 56839041 | ((H,G),C) | 1.87 | HKY |
| 259 | 10 | 57715895 | 57716894 | ((H,C),G) | 2.99 | HKY |
| 271 | 10 | 81424322 | 81425321 | ((H,C),G) | 2.12 | HKY |
| 286 | 10 | 81658264 | 81659263 | ((C,G),H) | 2.10 | HKY |
| 289 | 10 | 111724861 | 111725860 | ((H,C),G) | N.A. | TPM2uf |
| 294 | 10 | 128524321 | 128525320 | ((H,C),G) | 2.86 | K80 |
| 347 | 10 | 130727396 | 130728395 | ((H,C),G) | 3.70 | HKY |
| 374 | 11 | 21845057 | 21846056 | ((H,C),G) | N.A. | TPM2uf |
| 383 | 11 | 36910115 | 36911114 | ((H,C),G) | 1.46 | HKY |
| 397 | 11 | 37495524 | 37496523 | ((H,C),G) | N.A. | TrN |
| 398 | 11 | 38884861 | 38885860 | ((H,G),C) | 2.19 | HKY |
| 405 | 11 | 39465300 | 39466299 | ((C,G),H) | 2.00 | HKY+I |
| 410 | 11 | 42463677 | 42464676 | ((H,C),G) | 2.17 | HKY |
| 434 | 11 | 42665603 | 42666602 | ((H,C),G) | 2.35 | HKY |
| 439 | 11 | 81337312 | 81338311 | ((H,C),G) | 2.26 | HKY |
| 444 | 11 | 91385993 | 91386992 | ((H,C),G) | 2.27 | HKY |
| 463 | 11 | 114918045 | 114919044 | ((H,C),G) | 3.49 | K80 |
| 465 | 11 | 116164046 | 116165045 | ((H,C),G) | 3.19 | K80 |
| 484 | 12 | 60626720 | 60627719 | ((H,C),G) | 3.38 | HKY |
| 498 | 12 | 60835126 | 60836125 | ((H,C),G) | 2.38 | HKY |
| 519 | 12 | 73426393 | 73427392 | ((H,G),C) | 1.87 | HKY |
| 526 | 12 | 83386945 | 83387944 | ((C,G),H) | 3.26 | HKY |
| 540 | 12 | 83931107 | 83932106 | ((C,G),H) | 2.75 | HKY |
| 546 | 12 | 87382934 | 87383933 | ((H,C),G) | N.A. | TPM3uf |
| 554 | 13 | 22482509 | 22483508 | ((H,C),G) | 2.25 | HKY |
| 560 | 13 | 55793867 | 55794866 | ((C,G),H) | N.A. | TPM3uf |
| 567 | 13 | 56484628 | 56485627 | ((C,G),H) | N.A. | F81 |
| 585 | 13 | 64296706 | 64297705 | ((H,C),G) | 2.82 | HKY |
| 614 | 13 | 65510955 | 65511954 | ((H,C),G) | 1.21 | HKY |
| 621 | 13 | 72179360 | 72180359 | ((H,C),G) | 4.68 | HKY |
| 631 | 13 | 76351053 | 76352052 | ((C,G),H) | 2.85 | HKY |
| 664 | 13 | 76597135 | 76598134 | ((C,G),H) | 1.47 | HKY |
| 679 | 13 | 80132695 | 80133694 | ((H,C),G) | 1.63 | K80 |
| 683 | 13 | 81931238 | 81932237 | ((C,G),H) | 2.65 | HKY |
| 704 | 13 | 82143317 | 82144316 | ((H,C),G) | N.A. | TPM2uf |
| 727 | 13 | 82490030 | 82491029 | ((C,G),H) | 2.05 | HKY |
| 742 | 13 | 83507578 | 83508577 | ((H,C),G) | 1.16 | HKY |
| 760 | 13 | 84227523 | 84228522 | ((C,G),H) | 4.08 | HKY |
| 773 | 13 | 84799327 | 84800326 | ((H,C),G) | 1.61 | HKY |
| 782 | 13 | 86488112 | 86489111 | ((H,C),G) | 2.12 | HKY |
| 784 | 13 | 103650785 | 103651784 | ((H,G),C) | 2.62 | HKY |
| 821 | 13 | 103849493 | 103850492 | ((C,G),H) | 2.27 | HKY |
| 857 | 13 | 104061299 | 104062298 | ((H,C),G) | 1.01 | HKY |
| 892 | 13 | 104295286 | 104296285 | ((H,C),G) | N.A. | TPM3uf |
| 923 | 13 | 104498915 | 104499914 | ((H,C),G) | 1.84 | HKY |
| 944 | 13 | 105015903 | 105016902 | ((H,C),G) | 2.26 | HKY |
| 979 | 13 | 105238455 | 105239454 | ((C,G),H) | 3.24 | HKY |
| 996 | 14 | 39737967 | 39738966 | ((H,C),G) | N.A. | TPM2uf |
| 1023 | 14 | 39973727 | 39974726 | ((H,C),G) | 2.03 | HKY |
| 1050 | 14 | 40602287 | 40603286 | ((H,G),C) | N.A. | TPM3uf |
| 1067 | 14 | 83271762 | 83272761 | ((C,G),H) | 1.98 | HKY |
| 1086 | 14 | 83474393 | 83475392 | ((C,G),H) | 2.05 | HKY |
| 1098 | 14 | 84399968 | 84400967 | ((H,C),G) | 1.87 | HKY |
| 1122 | 14 | 86632388 | 86633387 | ((C,G),H) | 1.37 | HKY |
| 1182 | 16 | 61262633 | 61263632 | ((H,G),C) | N.A. | TPM2uf |
| 1219 | 17 | 54255285 | 54256284 | ((C,G),H) | N.A. | TPM2uf |
| 1222 | 18 | 25563441 | 25564440 | ((C,G),H) | 3.17 | K80 |
| 1233 | 18 | 28395826 | 28396825 | ((H,C),G) | 2.73 | HKY |
| 1240 | 18 | 30143273 | 30144272 | ((H,C),G) | 2.09 | HKY+I |
| 1272 | 18 | 30353880 | 30354879 | ((H,G),C) | 1.50 | HKY |
| 1290 | 18 | 37895673 | 37896672 | ((H,C),G) | 4.63 | HKY |
| 1323 | 18 | 38095162 | 38096161 | ((H,C),G) | 2.15 | HKY |
| 1340 | 18 | 38297158 | 38298157 | ((H,G),C) | 3.19 | HKY |
| 1364 | 18 | 40349006 | 40350005 | ((H,C),G) | 2.94 | HKY |
| 1405 | 18 | 40556148 | 40557147 | ((H,G),C) | 1.78 | HKY |
| 1428 | 18 | 40741304 | 40742303 | ((H,C),G) | N.A. | TPM2uf |
| 1465 | 18 | 40949070 | 40950069 | ((H,C),G) | 1.47 | K80 |
| 1493 | 18 | 41226829 | 41227828 | ((H,G),C) | 2.67 | HKY |
| 1501 | 18 | 43540271 | 43541270 | ((H,C),G) | 2.04 | HKY |
| 1517 | 18 | 52113203 | 52114202 | ((H,C),G) | 1.82 | HKY |
| 1519 | 18 | 53837849 | 53838848 | ((H,C),G) | 2.40 | HKY |
| 1528 | 18 | 61127278 | 61128277 | ((H,C),G) | N.A. | TPM3uf |
| 1531 | 18 | 64883353 | 64884352 | ((H,C),G) | 7.76 | HKY |
| 1533 | 18 | 66281153 | 66282152 | ((H,C),G) | 2.34 | HKY |
| 1537 | 18 | 72159816 | 72160815 | ((H,C),G) | 1.62 | HKY |
| 1553 | 2 | 13219366 | 13220365 | ((H,G),C) | 1.66 | HKY |
| 1564 | 2 | 35997192 | 35998191 | ((H,C),G) | 2.77 | HKY |
| 1583 | 2 | 41375562 | 41376561 | ((H,C),G) | 3.29 | HKY |
| 1591 | 2 | 49498606 | 49499605 | ((H,C),G) | N.A. | TPM3uf |
| 1618 | 2 | 53114436 | 53115435 | ((H,G),C) | N.A. | TPM2uf |
| 1642 | 2 | 116115936 | 116116935 | ((H,C),G) | 1.14 | HKY |
| 1655 | 2 | 116330528 | 116331527 | ((H,C),G) | 2.63 | HKY |
| 1670 | 2 | 117411742 | 117412741 | ((H,G),C) | 1.46 | HKY |
| 1685 | 2 | 118431064 | 118432063 | ((H,C),G) | 1.93 | HKY |
| 1711 | 2 | 118610736 | 118611735 | ((C,G),H) | 3.26 | K80 |
| 1715 | 2 | 122000479 | 122001478 | ((H,C),G) | 2.13 | HKY |
| 1728 | 2 | 122203140 | 122204139 | ((H,C),G) | 2.18 | HKY |
| 1750 | 2 | 122422345 | 122423344 | ((H,C),G) | 3.79 | HKY |
| 1772 | 2 | 122652643 | 122653642 | ((H,G),C) | 3.43 | HKY |
| 1783 | 2 | 122861233 | 122862232 | ((C,G),H) | N.A. | TPM3uf |
| 1785 | 2 | 125207247 | 125208246 | ((H,C),G) | 2.10 | HKY |
| 1805 | 2 | 125423956 | 125424955 | ((H,C),G) | 1.42 | HKY |
| 1812 | 2 | 128787322 | 128788321 | ((H,G),C) | 2.41 | HKY |
| 1823 | 2 | 133802185 | 133803184 | ((H,C),G) | 1.33 | K80 |
| 1832 | 2 | 139144319 | 139145318 | ((H,G),C) | 2.10 | HKY |
| 1834 | 2 | 147101848 | 147102847 | ((H,C),G) | 2.24 | HKY |
| 1847 | 2 | 155543931 | 155544930 | ((H,C),G) | 2.27 | HKY |
| 1882 | 2 | 155755037 | 155756036 | ((H,C),G) | 2.43 | HKY |
| 1891 | 2 | 180240600 | 180241599 | ((H,C),G) | 1.38 | HKY |
| 1894 | 2 | 184174075 | 184175074 | ((H,C),G) | N.A. | TPM3uf |
| 1898 | 2 | 185164829 | 185165828 | ((H,C),G) | N.A. | F81 |
| 1915 | 2 | 192405374 | 192406373 | ((H,C),G) | 2.03 | HKY |
| 1924 | 2 | 192995547 | 192996546 | ((C,G),H) | 2.50 | HKY |
| 1930 | 2 | 193482992 | 193483991 | ((H,C),G) | 1.97 | HKY |
| 1936 | 2 | 195241727 | 195242726 | ((H,C),G) | 1.56 | HKY |
| 1937 | 2 | 210880132 | 210881131 | ((H,C),G) | N.A. | TPM3uf |
| 1940 | 2 | 226422985 | 226423984 | ((H,C),G) | 2.17 | HKY |
| 1945 | 20 | 12581284 | 12582283 | ((H,G),C) | 1.91 | HKY |
| 2035 | 20 | 39552847 | 39553846 | ((H,C),G) | N.A. | TPM3uf |
| 2041 | 20 | 55782618 | 55783617 | ((H,C),G) | 1.40 | HKY |
| 2075 | 3 | 5463141 | 5464140 | ((H,C),G) | N.A. | TrN |
| 2089 | 3 | 5677111 | 5678110 | ((H,C),G) | 2.83 | K80+I |
| 2095 | 3 | 20723443 | 20724442 | ((H,G),C) | 1.67 | HKY |
| 2110 | 3 | 20930151 | 20931150 | ((H,G),C) | 2.97 | HKY |
| 2112 | 3 | 70512728 | 70513727 | ((H,C),G) | 3.28 | HKY |
| 2118 | 3 | 74764327 | 74765326 | ((H,C),G) | 6.75 | HKY |
| 2131 | 3 | 83010946 | 83011945 | ((H,C),G) | 1.81 | HKY |
| 2159 | 3 | 83211631 | 83212630 | ((H,G),C) | 2.34 | HKY |
| 2178 | 3 | 83415546 | 83416545 | ((H,C),G) | 2.11 | HKY |
| 2196 | 3 | 83620650 | 83621649 | ((C,G),H) | 1.67 | HKY |
| 2211 | 3 | 89792169 | 89793168 | ((H,C),G) | N.A. | TrN |
| 2239 | 3 | 94708276 | 94709275 | ((H,C),G) | N.A. | TPM3uf |
| 2246 | 3 | 95396204 | 95397203 | ((H,C),G) | 1.44 | HKY |
| 2256 | 3 | 95897114 | 95898113 | ((C,G),H) | N.A. | TPM3uf |
| 2265 | 3 | 103847850 | 103848849 | ((H,G),C) | N.A. | TPM3uf |
| 2267 | 3 | 106069554 | 106070553 | ((H,C),G) | N.A. | TrN |
| 2272 | 3 | 110185193 | 110186192 | ((H,C),G) | 2.73 | HKY |
| 2287 | 3 | 117339812 | 117340811 | ((C,G),H) | 1.34 | HKY |
| 2289 | 3 | 135672541 | 135673540 | ((C,G),H) | 3.68 | K80 |
| 2293 | 3 | 137217173 | 137218172 | ((H,C),G) | 2.23 | HKY |
| 2308 | 3 | 144404890 | 144405889 | ((H,C),G) | N.A. | TPM1uf |
| 2328 | 3 | 144605905 | 144606904 | ((H,C),G) | N.A. | TPM3uf |
| 2352 | 3 | 144823203 | 144824202 | ((H,C),G) | 1.46 | HKY |
| 2378 | 3 | 145091788 | 145092787 | ((H,C),G) | 2.77 | HKY |
| 2402 | 3 | 145304167 | 145305166 | ((H,G),C) | N.A. | TPM3uf |
| 2406 | 3 | 162026067 | 162027066 | ((C,G),H) | 1.91 | HKY |
| 2437 | 3 | 162235127 | 162236126 | ((H,G),C) | N.A. | TPM2uf |
| 2450 | 3 | 163764888 | 163765887 | ((H,G),C) | 2.21 | HKY |
| 2454 | 3 | 164405624 | 164406623 | ((H,C),G) | 1.61 | HKY |
| 2464 | 3 | 176351409 | 176352408 | ((H,C),G) | 2.63 | HKY |
| 2473 | 4 | 10943393 | 10944392 | ((H,C),G) | 2.21 | K80 |
| 2483 | 4 | 11138281 | 11139280 | ((H,C),G) | 1.27 | HKY |
| 2484 | 4 | 18247486 | 18248485 | ((H,G),C) | 2.37 | HKY |
| 2487 | 4 | 18689510 | 18690509 | ((H,C),G) | N.A. | TrN |
| 2497 | 4 | 18894844 | 18895843 | ((H,C),G) | N.A. | TPM2uf |
| 2507 | 4 | 24114292 | 24115291 | ((H,C),G) | 2.47 | HKY |
| 2539 | 4 | 30434197 | 30435196 | ((C,G),H) | 1.97 | HKY |
| 2552 | 4 | 32564007 | 32565006 | ((C,G),H) | 1.63 | HKY |
| 2574 | 4 | 32765487 | 32766486 | ((H,G),C) | N.A. | TPM1uf |
| 2582 | 4 | 35175679 | 35176678 | ((H,C),G) | 1.91 | HKY |
| 2595 | 4 | 45732389 | 45733388 | ((C,G),H) | 4.08 | HKY |
| 2605 | 4 | 60037725 | 60038724 | ((H,C),G) | 2.50 | HKY |
| 2625 | 4 | 60258127 | 60259126 | ((H,C),G) | N.A. | TPM3uf |
| 2648 | 4 | 63695828 | 63696827 | ((H,C),G) | 1.93 | HKY |
| 2665 | 4 | 63893286 | 63894285 | ((H,C),G) | 2.02 | HKY |
| 2686 | 4 | 67106369 | 67107368 | ((H,C),G) | 2.46 | HKY |
| 2700 | 4 | 114310805 | 114311804 | ((H,C),G) | 2.02 | HKY |
| 2715 | 4 | 115347616 | 115348615 | ((H,C),G) | N.A. | TPM3uf |
| 2725 | 4 | 124935714 | 124936713 | ((H,G),C) | 2.38 | HKY |
| 2746 | 4 | 126314160 | 126315159 | ((H,G),C) | 2.00 | HKY |
| 2761 | 4 | 129505308 | 129506307 | ((C,G),H) | 2.88 | HKY |
| 2768 | 4 | 130169222 | 130170221 | ((H,C),G) | N.A. | TPM2uf |
| 2784 | 4 | 160183891 | 160184890 | ((H,C),G) | N.A. | TPM1uf |
| 2835 | 4 | 180211863 | 180212862 | ((C,G),H) | 2.41 | HKY |
| 2879 | 5 | 2514990 | 2515989 | ((H,C),G) | 2.62 | HKY+I |
| 2886 | 5 | 3803204 | 3804203 | ((H,C),G) | 2.07 | HKY |
| 2900 | 5 | 5807182 | 5808181 | ((H,C),G) | 2.85 | K80 |
| 2906 | 5 | 27780120 | 27781119 | ((H,C),G) | N.A. | TPM3uf |
| 2916 | 5 | 30631523 | 30632522 | ((H,C),G) | 1.42 | HKY |
| 2939 | 5 | 30850307 | 30851306 | ((H,C),G) | 2.37 | HKY |
| 2950 | 5 | 51684176 | 51685175 | ((H,C),G) | N.A. | F81 |
| 2954 | 5 | 63548670 | 63549669 | ((H,C),G) | 2.09 | HKY |
| 2975 | 5 | 63751670 | 63752669 | ((C,G),H) | 1.88 | HKY |
| 2976 | 5 | 84706636 | 84707635 | ((H,C),G) | 2.08 | HKY |
| 2995 | 5 | 87949953 | 87950952 | ((H,C),G) | N.A. | F81 |
| 3003 | 5 | 99241732 | 99242731 | ((H,G),C) | N.A. | TPM3uf |
| 3015 | 5 | 103813829 | 103814828 | ((H,C),G) | 4.25 | HKY |
| 3023 | 5 | 105633320 | 105634319 | ((C,G),H) | 1.75 | HKY |
| 3032 | 5 | 106129658 | 106130657 | ((H,C),G) | 1.65 | HKY |
| 3040 | 5 | 110086427 | 110087426 | ((H,G),C) | N.A. | TPM2uf |
| 3041 | 5 | 113824045 | 113825044 | ((H,G),C) | 1.70 | HKY |
| 3046 | 5 | 119917448 | 119918447 | ((H,C),G) | 1.44 | HKY |
| 3060 | 5 | 123845234 | 123846233 | ((H,C),G) | 2.54 | HKY |
| 3077 | 5 | 144751618 | 144752617 | ((H,C),G) | 1.73 | HKY |
| 3082 | 5 | 166598000 | 166598999 | ((C,G),H) | 3.46 | HKY |
| 3093 | 6 | 9378117 | 9379116 | ((H,C),G) | 2.78 | HKY |
| 3097 | 6 | 18780959 | 18781958 | ((C,G),H) | 1.65 | HKY |
| 3117 | 6 | 48415235 | 48416234 | ((H,C),G) | 1.34 | HKY |
| 3138 | 6 | 66315472 | 66316471 | ((C,G),H) | 3.97 | HKY |
| 3160 | 6 | 66933522 | 66934521 | ((H,C),G) | 2.22 | HKY |
| 3163 | 6 | 76986504 | 76987503 | ((C,G),H) | 2.79 | HKY |
| 3183 | 6 | 78164459 | 78165458 | ((H,C),G) | N.A. | F81 |
| 3200 | 6 | 78390820 | 78391819 | ((C,G),H) | 2.80 | HKY |
| 3204 | 6 | 80763831 | 80764830 | ((C,G),H) | 1.49 | HKY |
| 3221 | 6 | 80988597 | 80989596 | ((H,C),G) | 2.04 | HKY |
| 3243 | 6 | 81229063 | 81230062 | ((H,G),C) | 2.20 | HKY |
| 3249 | 6 | 82578541 | 82579540 | ((H,G),C) | 2.26 | HKY |
| 3263 | 6 | 90793821 | 90794820 | ((H,C),G) | 2.02 | HKY |
| 3305 | 6 | 90999441 | 91000440 | ((C,G),H) | 2.28 | HKY |
| 3339 | 6 | 93652222 | 93653221 | ((H,C),G) | N.A. | TPM3uf |
| 3342 | 6 | 94649346 | 94650345 | ((H,G),C) | 1.62 | HKY |
| 3358 | 6 | 94858271 | 94859270 | ((H,C),G) | 1.93 | HKY |
| 3371 | 6 | 95006333 | 95007332 | ((H,C),G) | 3.97 | HKY |
| 3377 | 6 | 95096367 | 95097366 | ((H,C),G) | N.A. | TPM3uf |
| 3380 | 6 | 95106016 | 95107015 | ((H,G),C) | 2.30 | HKY |
| 3384 | 6 | 95120342 | 95121341 | ((H,G),C) | N.A. | TPM3uf |
| 3385 | 6 | 95122099 | 95123098 | ((H,C),G) | 2.48 | HKY |
| 3386 | 6 | 95126619 | 95127618 | ((H,C),G) | N.A. | F81 |
| 3387 | 6 | 95128722 | 95129721 | ((H,C),G) | 6.10 | HKY |
| 3393 | 6 | 95174381 | 95175380 | ((H,C),G) | 1.59 | HKY+I |
| 3395 | 6 | 95186508 | 95187507 | ((H,C),G) | 6.07 | HKY |
| 3396 | 6 | 95206851 | 95207850 | ((H,G),C) | N.A. | TPM1uf |
| 3397 | 6 | 95217561 | 95218560 | ((H,C),G) | 1.55 | HKY |
| 3401 | 6 | 95231796 | 95232795 | ((H,C),G) | 4.05 | HKY |
| 3410 | 6 | 95276070 | 95277069 | ((H,C),G) | 3.14 | HKY |
| 3415 | 6 | 101107652 | 101108651 | ((H,C),G) | 1.93 | HKY |
| 3425 | 6 | 103207185 | 103208184 | ((H,C),G) | 2.58 | HKY |
| 3443 | 6 | 103798086 | 103799085 | ((H,C),G) | 3.04 | HKY |
| 3446 | 6 | 104226765 | 104227764 | ((H,G),C) | 2.15 | HKY |
| 3473 | 6 | 104433238 | 104434237 | ((H,G),C) | N.A. | TrN |
| 3481 | 6 | 115094761 | 115095760 | ((H,C),G) | N.A. | TPM3uf |
| 3487 | 6 | 119549763 | 119550762 | ((H,C),G) | 2.13 | HKY |
| 3515 | 6 | 119738103 | 119739102 | ((C,G),H) | 4.06 | HKY |
| 3523 | 6 | 126885903 | 126886902 | ((H,G),C) | 1.56 | HKY |
| 3530 | 6 | 141846722 | 141847721 | ((H,C),G) | N.A. | TPM1uf |
| 3534 | 6 | 145061120 | 145062119 | ((H,G),C) | 2.10 | HKY |
| 3546 | 6 | 156146881 | 156147880 | ((H,G),C) | 2.18 | K80 |
| 3556 | 7 | 41359684 | 41360683 | ((H,C),G) | 2.83 | HKY |
| 3573 | 7 | 42451678 | 42452677 | ((H,C),G) | 2.20 | HKY |
| 3577 | 7 | 49468643 | 49469642 | ((H,C),G) | 1.63 | HKY |
| 3589 | 7 | 52474207 | 52475206 | ((H,C),G) | 2.70 | HKY |
| 3625 | 7 | 52690817 | 52691816 | ((H,C),G) | N.A. | TPM3uf |
| 3661 | 7 | 85888782 | 85889781 | ((H,G),C) | N.A. | F81 |
| 3677 | 7 | 113632664 | 113633663 | ((H,C),G) | 3.41 | HKY |
| 3681 | 7 | 118671413 | 118672412 | ((H,C),G) | 2.53 | HKY |
| 3683 | 7 | 125669627 | 125670626 | ((H,C),G) | N.A. | TPM2uf |
| 3686 | 7 | 145883492 | 145884491 | ((C,G),H) | 1.55 | HKY |
| 3694 | 8 | 76000407 | 76001406 | ((H,C),G) | 2.94 | HKY |
| 3707 | 8 | 77781402 | 77782401 | ((H,C),G) | 1.49 | HKY |
| 3727 | 8 | 83659161 | 83660160 | ((H,C),G) | N.A. | TPM1uf |
| 3737 | 8 | 89010429 | 89011428 | ((C,G),H) | 2.28 | HKY |
| 3741 | 8 | 114534601 | 114535600 | ((C,G),H) | N.A. | TPM1uf+I |
| 3749 | 8 | 114992073 | 114993072 | ((H,G),C) | 6.80 | HKY+I |
| 3761 | 8 | 141723200 | 141724199 | ((C,G),H) | 2.23 | HKY |
| 3766 | 9 | 1536428 | 1537427 | ((H,C),G) | 4.46 | HKY |
| 3783 | 9 | 1743015 | 1744014 | ((H,G),C) | 1.71 | HKY |
| 3789 | 9 | 11850793 | 11851792 | ((H,C),G) | 2.10 | HKY |
| 3794 | 9 | 13705376 | 13706375 | ((H,C),G) | 4.05 | HKY |
| 3800 | 9 | 18009516 | 18010515 | ((H,C),G) | N.A. | TPM3uf |
| 3830 | 9 | 23189453 | 23190452 | ((H,C),G) | 1.59 | HKY |
| 3860 | 9 | 23412105 | 23413104 | ((H,C),G) | N.A. | F81 |
| 3863 | 9 | 24157509 | 24158508 | ((H,G),C) | N.A. | TPM3uf |
| 3882 | 9 | 25351936 | 25352935 | ((H,G),C) | 1.76 | HKY |
| 3890 | 9 | 26382945 | 26383944 | ((H,C),G) | N.A. | TrN |
| 3896 | 9 | 30093346 | 30094345 | ((C,G),H) | 2.05 | HKY |
| 3922 | 9 | 30356760 | 30357759 | ((C,G),H) | N.A. | TPM3uf |
| 3923 | 9 | 31849777 | 31850776 | ((H,C),G) | 1.65 | HKY |
| 3937 | 9 | 32044149 | 32045148 | ((H,C),G) | N.A. | TPM3uf |
| 3955 | 9 | 74083427 | 74084426 | ((H,C),G) | 1.84 | HKY |
| 3964 | 9 | 81075848 | 81076847 | ((C,G),H) | 2.61 | HKY |
| 3975 | 9 | 118356705 | 118357704 | ((H,C),G) | N.A. | TPM2uf |
| 3989 | 9 | 119729136 | 119730135 | ((C,G),H) | 2.30 | HKY |
| 4019 | X | 7624243 | 7625242 | ((H,G),C) | 2.10 | HKY |
| 4024 | X | 20680056 | 20681055 | ((H,C),G) | 2.14 | HKY |
| 4036 | X | 20896970 | 20897969 | ((H,C),G) | 1.92 | HKY |
| 4043 | X | 35320455 | 35321454 | ((H,C),G) | 1.35 | HKY |
| 4050 | X | 66898623 | 66899622 | ((H,C),G) | 1.90 | HKY |
| 4058 | X | 67116446 | 67117445 | ((H,C),G) | N.A. | TrN |
| 4062 | X | 79697749 | 79698748 | ((H,G),C) | 2.10 | HKY |
| 4066 | X | 82207815 | 82208814 | ((H,C),G) | 2.33 | HKY |
| 4072 | X | 83274809 | 83275808 | ((H,C),G) | 2.72 | HKY |
| 4075 | X | 85589655 | 85590654 | ((H,C),G) | 3.80 | HKY |
| 4087 | X | 91096066 | 91097065 | ((H,G),C) | 1.68 | HKY |
| 4110 | X | 98888588 | 98889587 | ((H,C),G) | 2.28 | HKY |
| 4135 | X | 99170074 | 99171073 | ((C,G),H) | 2.54 | HKY |
| 4138 | X | 117612134 | 117613133 | ((H,C),G) | 2.46 | HKY |
| 4143 | X | 121580764 | 121581763 | ((H,C),G) | 1.52 | HKY |
| 4149 | X | 122213105 | 122214104 | ((H,C),G) | N.A. | TPM3uf |
| 4153 | X | 125866857 | 125867856 | ((H,C),G) | 1.63 | HKY |
| 4155 | X | 127094708 | 127095707 | ((H,C),G) | 1.78 | HKY |
| 4160 | X | 137895295 | 137896294 | ((H,C),G) | 5.25 | HKY |
| 4177 | X | 138145219 | 138146218 | ((H,G),C) | 2.53 | HKY |
| 4181 | X | 142823291 | 142824290 | ((C,G),H) | N.A. | TPM3uf |
| 4184 | X | 144511738 | 144512737 | ((H,C),G) | 1.53 | HKY |

Referências

1. Kuska, B., *Beer, Bethesda, and biology: how "genomics" came into being.* J Natl Cancer Inst, 1998. **90**(2): p. 93.

2. *Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species.* J Hered, 2009. **100**(6): p. 659-74.

3. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.

4. Mardis, E.R., *A decade's perspective on DNA sequencing technology.* Nature, 2011. **470**(7333): p. 198-203.

5. Koboldt, D.C., et al., *Challenges of sequencing human genomes.* Brief Bioinform, 2010. **11**(5): p. 484-98.

6. Brown, G.R., et al., *Gene: a gene-centered information resource at NCBI.* Nucleic Acids Res, 2015. **43**(Database issue): p. D36-42.

7. Federhen, S., *The NCBI Taxonomy database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D136-43.

8. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**(Database issue): p. D130-5.

9. Flicek, P., et al., *Ensembl 2014.* Nucleic Acids Res, 2014. **42**(Database issue): p. D749-55.

10. Takahata, N., Y. Satta, and J. Klein, *Divergence time and population size in the lineage leading to modern humans.* Theor Popul Biol, 1995. **48**(2): p. 198-221.

11. Ruvolo, M., *Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets.* Mol Biol Evol, 1997. **14**(3): p. 248-65.

12. Horai, S., et al., *Man's place in Hominoidea revealed by mitochondrial DNA genealogy.* J Mol Evol, 1992. **35**(1): p. 32-43.

13. Bailey, W.J., et al., *Molecular evolution of the psi eta-globin gene locus: gibbon phylogeny and the hominoid slowdown.* Mol Biol Evol, 1991. **8**(2): p. 155-84.

14. Takahata, N., *A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism.* Proc Natl Acad Sci U S A, 1990. **87**(7): p. 2419-23.

15. Rogers, A.R. and H. Harpending, *Population growth makes waves in the distribution of pairwise genetic differences.* Mol Biol Evol, 1992. **9**(3): p. 552-69.

16. Chen, F.C. and W.H. Li, *Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.* Am J Hum Genet, 2001. **68**(2): p. 444-56.

17. Jeffreys, A.J., V. Wilson, and S.L. Thein, *Hypervariable 'minisatellite' regions in human DNA.* Nature, 1985. **314**(6006): p. 67-73.

18. Ahmadian, A., et al., *Single-nucleotide polymorphism analysis by pyrosequencing.* Anal Biochem, 2000. **280**(1): p. 103-10.

19. Rands, C.M., et al., *8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage.* PLoS Genet, 2014. **10**(7): p. e1004525.

20. Jennings, W.B. and S.V. Edwards, *Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees.* Evolution, 2005. **59**(9): p. 2033-47.

21. Lee, J.Y. and S.V. Edwards, *Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (Malurus melanocephalus).* Evolution, 2008. **62**(12): p. 3117-34.

22. Gottscho, A.D., S.B. Marks, and W.B. Jennings, *Speciation, population structure, and demographic history of the Mojave Fringe-toed Lizard (Uma scoparia), a species of conservation concern.* Ecol Evol, 2014. **4**(12): p. 2546-62.

23. Jarne, P. and P.J. Lagoda, *Microsatellites, from molecules to populations and back.* Trends Ecol Evol, 1996. **11**(10): p. 424-9.

24. Bertozzi, T., et al., *Anonymous nuclear loci in non-model organisms: making the most of high-throughput genome surveys.* Bioinformatics, 2012. **28**(14): p. 1807-10.

25. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.

26. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

27. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol, 2003. **52**(5): p. 696-704.

28. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.

29. Posada, D. and K.A. Crandall, *MODELTEST: testing the model of DNA substitution.* Bioinformatics, 1998. **14**(9): p. 817-8.

30. Posada, D., *jModelTest: phylogenetic model averaging.* Mol Biol Evol, 2008. **25**(7): p. 1253-6.

31. Hasegawa, M., H. Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* J Mol Evol, 1985. **22**(2): p. 160-74.

32. Degnan, J.H. and N.A. Rosenberg, *Gene tree discordance, phylogenetic inference and the multispecies coalescent.* Trends Ecol Evol, 2009. **24**(6): p. 332-40.

33. Hobolth, A., et al., *Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model.* PLoS Genet, 2007. **3**(2): p. e7.