IGOR RODRIGUES DA COSTA

Locos anônimos

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

LOCOS ANÔNIMOS

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: William Bryan Jennings, Ph.D

Rio de Janeiro

2015

|  |
| --- |
| Costa, Igor Rodrigues da  Locos Anônimos/ Igor Rodrigues da Costa. – Rio de Janeiro: UFRJ/IBqM, 2015.  XXII, 132 p.: il.; 29,7 cm.  Orientador: Francisco Prosdocimi  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. Título. |

IGOR RODRIGUES DA COSTA

LOCOS ANÔNIMOS

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

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

Bryan Jennings, Ph.D, UFRJ

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

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

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

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

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

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

A você, caro leitor

Agradecimentos

<Epígrafe>

Resumo

COSTA, Igor Rodrigues da. **Locos Anônimos**. 2015. <número de folhas>. Tese (Mestrado em Química Biológica) – Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

O resumo deve ser redigido em só parágrafo, de preferência, na 3ª pessoa do singular e o verbo na voz ativa com, no máximo, 500 palavras e no mínimo 150 palavras. Sugere-se que o resumo venha antecedido por uma referência bibliográfica do trabalho.

Palavras-chave: <lista de palavras-chave>

Abstract

COSTA, Igor Rodrigues da. **Locos Anônimos**. 2015. <número de folhas>. Tese (Mestrado em Química Biológica) – Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

O resumo deve ser redigido em só parágrafo, de preferência, na 3ª pessoa do singular e o verbo na voz ativa com, no máximo, 500 palavras e no mínimo 150 palavras.Sugere-se que o resumo venha antecedido por uma referência bibliográfica do trabalho.

Keywords: <lista de palavras-chave>

Lista de Figuras

[Figura 1: Frequência das topologias observadas 23](#_Toc426371352)

Lista de Tabelas

[Quadro 1: <descrição da tabela> 3](#_Toc338858942)

[Quadro 2: <descrição da tabela> 3](#_Toc338858943)

Lista de Siglas

|  |  |
| --- | --- |
| AL | Loco Anônimo |
| SNP | *Single Nugleotide Polymorphism* |
| <sigla 3> | <palavras e expressões correspondentes grafadas por extenso> |
| <sigla 4> | <palavras e expressões correspondentes grafadas por extenso> |
| <sigla 5> | <palavras e expressões correspondentes grafadas por extenso> |

Sumário

[1 Introdução 15](#_Toc426371675)

[1.1 Genômica 15](#_Toc426371676)

[1.1.1 Sequenciamento de nova geração 15](#_Toc426371677)

[1.1.2 Genomas completos e parciais 15](#_Toc426371678)

[1.2 Genômica Computacional 15](#_Toc426371679)

[1.2.1 Bancos de dados 16](#_Toc426371680)

[1.2.2 Softwares para análise de dados 16](#_Toc426371681)

[1.3 Genética de População 16](#_Toc426371682)

[1.3.1 Parâmetros de interesse para estudo 16](#_Toc426371683)

[1.3.2 O modelo dos *Hominidae* na genética de população 16](#_Toc426371684)

[1.4 Marcadores genéticos 16](#_Toc426371685)

[1.4.1 Utilidade dos marcadores genéticos 17](#_Toc426371686)

[1.4.2 Microssatélites 17](#_Toc426371687)

[1.4.3 SNPs 17](#_Toc426371688)

[1.4.4 Locos Anônimos 17](#_Toc426371689)

[1.5 Resumo 18](#_Toc426371690)

[2 Objetivo 19](#_Toc426371691)

[3 Desenvolvimento 20](#_Toc426371692)

[3.1 Programas, bibliotecas e linguagem 20](#_Toc426371693)

[3.2 Busca de regiões anônimas e locos putativos 20](#_Toc426371694)

[3.3 Busca de homólogos e filtro de cópia única 20](#_Toc426371695)

[4 Resultados 21](#_Toc426371696)

[4.1 Busca de Locos Anônimos em genomas completos de *Hominidae* 22](#_Toc426371697)

[4.1.1 Busca por regiões anônimas 22](#_Toc426371698)

[4.1.2 Filtragem por conservação e unicidade 22](#_Toc426371699)

[4.1.3 Alinhamento dos Locos Anônimos 22](#_Toc426371700)

[4.2 Análise dos Locos Anônimos de *Hominidae* 22](#_Toc426371701)

[4.2.1 Mapeamento dos Locos Anônimos por cromossomo 22](#_Toc426371702)

[4.2.2 Modelos de substituição 22](#_Toc426371703)

[4.2.3 Filogenia 22](#_Toc426371704)

[4.2.4 Estimativa da população efetiva e tempo de divergência 23](#_Toc426371705)

[4.2.5 CpG e taxa de mutação 23](#_Toc426371706)

[4.2.6 Teste de neutralidade 23](#_Toc426371707)

[Conclusão 24](#_Toc426371708)

[Referências 25](#_Toc426371709)

[Glossário 26](#_Toc426371710)

[Anexos 27](#_Toc426371711)

[ANEXO A –Decaimento exponencial do custo de sequenciamento 27](#_Toc426371712)

[Apêndices 28](#_Toc426371713)

[APÊNDICE A – Código fonte dos programas utilizados 28](#_Toc426371714)

[Índice 29](#_Toc426371715)

1 Introdução

## Genômica

O Dicionário Oxford de Inglês possui a seguinte definição para o sufixo “*-ome*”:

“Forming nouns with the sense ’all of the specified constituents of a cell, considered collectively or in total’.”

A palavra “genômica” surgiu em uma discussão de bar para decidir o nome de uma nova revista científica em 1986 [[1](#_ENREF_1)]. TERMINAR

### 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 [[2](#_ENREF_2)]. 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[[2](#_ENREF_2),[3](#_ENREF_3)]. TERMINAR (descrever técnicas)

### 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[[4](#_ENREF_4)]. A anotação deste genoma, que envolve a anotação 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 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.

### 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 [[5](#_ENREF_5)], taxonomias[[6](#_ENREF_6)], proteínas [[7](#_ENREF_7)], ou o ENSEMBL[[8](#_ENREF_8)], um banco de dados de genomas completos para organismos modelo.

### Softwares para análise de dados

## 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, sendo esta muitas vezes a única maneira de reconstruir os passos das populações e entender as pressões seletivas atuantes.

### Parâmetros de interesse para estudo

Quando

#### Estimativa de tempo de divergência

#### Estimativas de população efetiva

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

## 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 [[9](#_ENREF_9)].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

### SNPs

Outra classe de marcadores genéticos, os polimorfismos de nucleotídeos únicos (SNPs) são variações de uma única base.

### 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. Cópia única no genoma;
2. Sob seleção neutra;
3. Segregação independente.

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 [[10](#_ENREF_10)], é esperado que 1 em 10 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 amplicons gênicos 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 [[11-13](#_ENREF_11)].

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 kbp, muito maiores que os 100-500 bp dos microssatélites ou 1bp dos 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* [[14](#_ENREF_14)].

Outra vantagem é TERMINAR

#### 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 [[15](#_ENREF_15)], 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.

TERMINAR (Falar do [[16](#_ENREF_16)]primer do model -> non-model 2008)

## Resumo

Neste trabalho é descrito um novo programa criado para predição *in* silicode marcadores moleculares ideais para genética de população. Além disso, são apresentadas análises de marcadores criados a partir da aplicação deste programa no genoma dos *Hominidae*. Por fim, os parâmetros estimados são comparados favoravelmente com o estado da arte.

# Objetivo

Desenvolver uma metodologia para descoberta de Locos Anônimos em genomas completos e aplicá-la no modelo dos *Hominidae*.

# Desenvolvimento

## Programas, bibliotecas e linguagem

Este programa foi desenvolvido na linguagem Python, versão 2.7, com o auxílio da biblioteca Biopython [[17](#_ENREF_17)] e de programas auxiliares: BLAST+ [[18](#_ENREF_18)] para busca de sequências e PhyML [[19](#_ENREF_19)]para reconstrução de árvores filogenéticas.

## Busca de regiões anônimas e locos putativos

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 200kbp de qualquer gene (conservadoramente, podem ser excluídas as regiões próximas a pseudogenes). Regiões próximas (10kbp) dos telômeros também são excluídas. As regiões anônimas do genoma de referência são recortadas em fragmentos de 1kbp, 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.

## Busca de homólogos e filtro de cópia única

Os locos anônimos putativos são TERMINAR

# Resultados

Após o desenvolvimento do programa alfie foi realizada 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*

### Busca por regiões anônimas

### Filtragem por conservação e unicidade

### Alinhamento dos Locos Anônimos

## Análise dos Locos Anônimos de *Hominidae*

### Mapeamento dos Locos Anônimos por cromossomo

### Modelos de substituição

### Filogenia

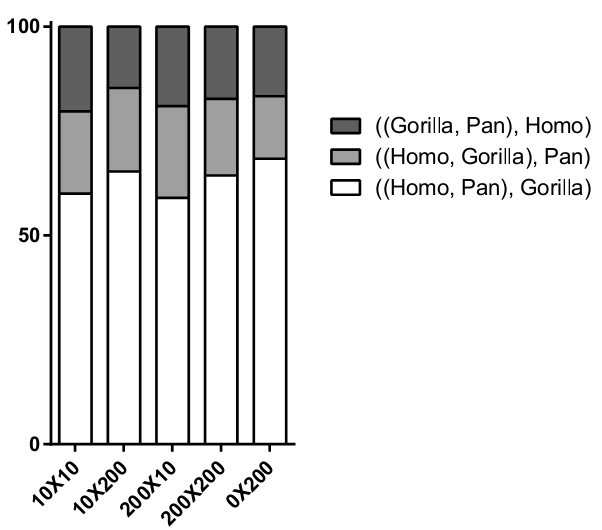


Figura 1: Frequência das topologias observadas

### Estimativa da população efetiva e tempo de divergência

### CpG e taxa de mutação

CpG

Close: 0,490479667 +- 0,094620764

### Teste de neutralidade

Normal normal normal

Nota[[1]](#footnote-2)

Quadro 1: <descrição da tabela>

|  |  |  |
| --- | --- | --- |
| Coluna 1 | Coluna 2 | Coluna 3 |
| A | B | C |
| B | C | A |

5 Conclusão

Normal normal normal

Referências

Glossário

Relação de palavras ou expressões técnicas de uso restrito ou de sentido obscuro, utilizadas no texto, acompanhadas das respectivas definições. É um elemento opcional, elaborado em ordem alfabética.

Anexos

ANEXO A –Decaimento exponencial do custo de sequenciamento



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

Apêndices

Elemento opcional que consiste em um texto ou documento elaborado pelo autor, a fim de complementar sua argumentação,sem prejuízo da unidade nuclear do trabalho.

APÊNDICE A – Código fonte dos programas utilizados

Normal normal normal

Índice

É a lista de palavras ou frases, ordenadas segundo um determinado critério, que localiza e remete para as informações contidas no texto. O índice aparece no final da publicação.

1. Kuska, B. Beer, bethesda, and biology: How "genomics" came into being. *J Natl Cancer Inst* **1998**, *90*, 93.

2. Schuster, S.C. Next-generation sequencing transforms today's biology. *Nat Methods* **2008**, *5*, 16-18.

3. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **2011**, *470*, 198-203.

4. Koboldt, D.C.; Ding, L.; Mardis, E.R.; Wilson, R.K. Challenges of sequencing human genomes. *Brief Bioinform* **2010**, *11*, 484-498.

5. Brown, G.R.; Hem, V.; Katz, K.S.; Ovetsky, M.; Wallin, C.; Ermolaeva, O.; Tolstoy, I.; Tatusova, T.; Pruitt, K.D.; Maglott, D.R.*, et al.* Gene: A gene-centered information resource at ncbi. *Nucleic Acids Res* **2015**, *43*, D36-42.

6. Federhen, S. The ncbi taxonomy database. *Nucleic Acids Res* **2012**, *40*, D136-143.

7. Pruitt, K.D.; Tatusova, T.; Brown, G.R.; Maglott, D.R. Ncbi reference sequences (refseq): Current status, new features and genome annotation policy. *Nucleic Acids Res* **2012**, *40*, D130-135.

8. Flicek, P.; Amode, M.R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.*, et al.* Ensembl 2014. *Nucleic Acids Res* **2014**, *42*, D749-755.

9. Jeffreys, A.J.; Wilson, V.; Thein, S.L. Hypervariable 'minisatellite' regions in human DNA. *Nature* **1985**, *314*, 67-73.

10. Rands, C.M.; Meader, S.; Ponting, C.P.; Lunter, G. 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*, e1004525.

11. Jennings, W.B.; Edwards, S.V. Speciational history of australian grass finches (poephila) inferred from thirty gene trees. *Evolution* **2005**, *59*, 2033-2047.

12. Lee, J.Y.; Edwards, S.V. Divergence across australia's carpentarian barrier: Statistical phylogeography of the red-backed fairy wren (malurus melanocephalus). *Evolution* **2008**, *62*, 3117-3134.

13. Gottscho, A.D.; Marks, S.B.; Jennings, W.B. Speciation, population structure, and demographic history of the mojave fringe-toed lizard (uma scoparia), a species of conservation concern. *Ecol Evol* **2014**, *4*, 2546-2562.

14. Jarne, P.; Lagoda, P.J. Microsatellites, from molecules to populations and back. *Trends Ecol Evol* **1996**, *11*, 424-429.

15. Bertozzi, T.; Sanders, K.L.; Sistrom, M.J.; Gardner, M.G. Anonymous nuclear loci in non-model organisms: Making the most of high-throughput genome surveys. *Bioinformatics* **2012**, *28*, 1807-1810.

16. Thomson, R.C.; Shedlock, A.M.; Edwards, S.V.; Shaffer, H.B. Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles. *Mol Phylogenet Evol* **2008**, *49*, 514-525.

17. Cock, P.J.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.*, et al.* Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422-1423.

18. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. Blast+: Architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421.

19. Guindon, S.; Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **2003**, *52*, 696-704.

1. Descrição da nota de rodapé [↑](#footnote-ref-2)