UNIVERSIDADE FEDERAL DO PARANÁ

ELIZA CRISTIANE DE CARVALHO

USO DE ELEMENTOS ULTRACONSERVADOS (UCEs) COMO FERRAMENTA PARA ANÁLISE FILOGENÉTICA

CURITIBA

2018

ELIZA CRISTIANE DE CARVALHO

USO DE ELEMENTOS ULTRACONSERVADOS (UCEs) COMO FERRAMENTA PARA ANÁLISE FILOGENÉTICA

Projeto de pesquisa apresentado à disciplina "Trabalho de Conclusão de Curso II – TCC II", como requisito parcial à obtenção do grau de Bacharel em Biomedicina no curso de graduação em Biomedicina, Setor de Ciências Biológicas da Universidade Federal do Paraná.

Orientador: Prof. Dr. Marcos Soares Barbeitos

Co-orientadora: Ms. Carolina de Lima Adam

CURITIBA

2018

... Dedico esse trabalho aos meus pais Geraldo e Felícia!

AGRADECIMENTOS

```
#!/usr/bin/perl
# Author: Eliza Carvalho <eliza.carvalho01@gmail.com>
# Date: 05 May 2018
# Description: AGRADECIMENTOS

use strict;

# Institution
my $UFPR = 'Universidade Federal do Paraná' ;

open ( UFPR, $UFPR ) || die " Cannot find $UFPR:  $! " ;
while ( <UFPR>) {
        if ( /Marcos Barbeitos/ || /Carolina Adam/ ) {
        print " Meus sinceros agradecimentos ao meu orientador Prof. Marcos pelos ensinamentos em programação, por toda paciência, competência e carinho durantes todos esses anos no LEOM. Agradeço também a Carol pela dedicação e empenho dedicado à elaboração deste trabalho e pela revisão minuciosa do texto. " ;
        }
        elsif  ( /Família/ || /Amigos/ || /Mozão/) {
        print " Agradecimento especial à minha família e amigos, por toda a alegria, apoio incondicional e compreensão ao longo desta jornada. Sem vocês nada disso seria possível! " ;
        }
        elsif ( / Brant Faircloth/ ) {
        print " Agradecimentos também ao Dr. Fairtcloth pela assistência nesse projeto, inspiração e pelo brilhante trabalho com as UCEs. " ;
        }
}

# Conclunding
close ( UFPR ) ;
print " \nJob Done ! \n "
```

RESUMO

Todos os organismos são filogenéticamente relacionados de acordo com evidências obtidas a partir de dados bioquímicos, morfológicos e de sequências genéticas. Estas relações podem ser representadas por uma vasta árvore filogenética, a Árvore da Vida, que não só revela a hierarquia das relações entre os táxons, mas também fornecem contexto evolutivo para a compreensão do tempo e origem de diversos sistemas funcionais. Neste trabalho, avaliamos a utilidade de uma nova classe de marcadores moleculares chamada Elementos Ultraconservados (UCEs), cuja utilização na sistemática filogenética tem se tornado popular e fornecido dados promissores. UCEs são regiões altamente conservadas do genoma (sequências de DNA com mais de 80% de correspondência e ao menos 100 pares de bases) entre organismos de diferentes táxons e que ja foram reportadas em 481 sequências maiores que 200 pares de base compartilhadas entre o genoma humano, de camundongos e de ratos. Como marcadores moleculares, UCEs apresentam características que os tornam particularmente desejáveis para estudos filogenéticos e, também, já foram utilizadas para reconstruir filogenias de táxons tão divergentes quanto mamíferos, peixes, aves, répteis, anfíbios e até mesmo em plantas. Este estudo, portanto, propõe identificar UCEs no genoma de uma ampla amostragem dos metazoários e estimar sua utilidade para inferência filogenética de alto nível entre metazoários utilizando como base organismos com genoma completo sequenciado e disponível em bancos de dados de livre acesso. Nossos resultados indicam baixo compartilhamento entre filos muito distintos, entretanto ressaltando potencial utilidade como marcadores para filogenias em menor escala.

Palavras-chave: Elementos Ultraconservados. Metazoa. Filogenia. Sistemática Molecular. Filogenômica.

ABSTRACT

All organisms are phylogenetically related according to evidence obtained from biochemical, morphological and genetic sequences. These relationships can be represented by a vast phylogenetic tree, the Tree of Life, which not only reveals the hierarchy of relations between taxa but also provides evolutionary context for understanding the time and origin of various functional systems. In this work, we evaluate the utility of a new class of molecular markers called Ultraconserved Elements (UCEs), whose use in phylogenetic systematics has become popular and provides promising data. UCEs are highly conserved regions of the genome (DNA sequences with more than 80% correspondence and at least 100 base pairs) between organisms of different taxa and have already been reported in 481 sequences greater than 200 base pairs shared between the human genome, mouse and mice. As molecular markers, UCEs have characteristics that make them particularly desirable for phylogenetic studies and have also been used to reconstruct phylogenies of as divergent taxa as mammals, fish, birds, reptiles, amphibians and even plants. This study therefore proposes to identify UCEs in the genome of a broad sampling of metazoans and to estimate their usefulness for high level phylogenetic inference between metazoa using as a basis organisms with complete genome sequenced and available in free access databases. Our results indicate low sharing between very different phyla, however highlighting potential utility as markers for phylogenies on a smaller scale.

Keywords: Ultraconserved Elements. Metazoa. Phylogeny. Molecular Systematics. Phylogenomics.

LISTA DE ABREVIATURAS E/OU SIGLAS

BLAST – *Basic local alignment search tool*

BLOSUM – *Block substitution matrix*

csv – *Comma separated values*

DNA – Ácido desoxirribonucleico

DNAr – Ácido desoxirribonucleico ribossômico

EF-2 – Fator de alongamento 2

EMBL – *European molecular biology laboratory*

EVEs – Elementos virais endógenos

GO – *Gene Ontology*

MAF – *Mutation annotation format*

Mpb – Milhões de pares de bases

NCBI – *National center for biotechnology information*

nr – Proteína não redundante

OSC – *Ohio supercomputer center*

PAM – *Percent accepted mutations*

pb – Pares de base

PCR – Reação em cadeia da polimerase

PDB – *Protein database bank*

RNA – Ácido ribonucleico

RNAr – Ácido ribonucleico ribossômico

SNPs – Polimorfismos de nucleotídeo único

THG – Transferência horizontal de genes

UCEs – Elementos ultraconservados

α – Alfa

β – Beta

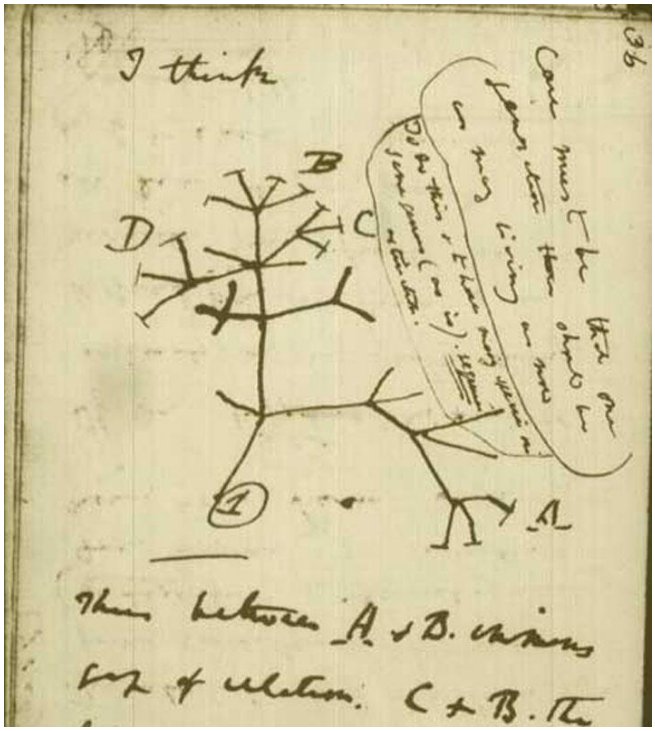
SUMÁRIO

1 INTRODUÇÃO

O avanço de técnicas mais modernas de biologia molecular, mais acentuadamente da reação em cadeia da polimerase (PCR), permitiu o sequenciamento rápido e barato de DNA e afetou nossa compreensão da história evolutiva e relações filogenéticas entre espécies ou grupos de espécies (PATWARDHAN et al., 2014).

Atualmente, aproximadamente cerca de 9 milhões de espécies habitam a Terra (MORA et al., 2011). Todos os organismos são geneticamente relacionados de acordo com evidências obtidas a partir de dados bioquímicos, morfológicos e de sequências genéticas. Suas relações genealógicas podem ser representadas por uma vasta árvore filogenética, a Árvore da Vida - termo introduzido por Darwin como metáfora para evolução na obra “A Origem das Espécies” (DARWIN, 1859; SADAVA et al., 2009; MINDELL, 2013), cujo método de obtenção foi posteriormente formalizado por Willi Hennig (HENNING et al., 1999). Um ancestral hipotético simboliza a raiz da árvore da qual todas as outras formas de vida surgiram (FIGURA 1). Desde então, as relações entre as espécies vêm sendo estudadas por meio de estruturas de árvores semelhantes às propostas por Darwin em sua obra.

FIGURA 1 – DIAGRAMA DA ÁRVORE DA VIDA PROPOSTO POR DARWIN.



FONTE: THE NATURE INSTITUTE (2009).

Essas relações são chamadas de filogenia ou história evolutiva das espécies e são construídas a partir de semelhanças compartilhadas em suas características morfofisiológicas, comportamentais e/ou genéticas (DOWELL, 2008). As filogenias não só revelam a hierarquia das relações entre os táxons (unidades taxonômicas), mas também fornecem contexto evolutivo para a compreensão do tempo e origem de diversos sistemas e traços complexos das espécies (LIN et al., 2016; PALMER e JIGGINS, 2015; ARENDT et al., 2016).

Neste trabalho, utilizamos uma nova classe de marcadores moleculares chamada Elementos Ultraconservados (UCEs da sigla inglesa *ultraconserved elements*), cuja utilização na sistemática filogenética tem se tornado popular e fornecido dados promissores. UCEs são regiões altamente conservadas do genoma (sequências de DNA com mais de 80% de correspondência idêntica e ao menos 100 pares de bases) entre organismos de diferentes táxons (BEJERANO et al., 2004; FAIRCLOTH et al., 2012). Em estudos anteriores, UCEs se mostraram eficientes para resolver relações filogenéticas entre mamíferos (MCCORMACK et al., 2012), aves (FAIRCLOTH et al., 2012), répteis (CRAWFORD et al., 2012), peixes (FAIRCLOTH et al., 2013) e insetos (FAIRCLOTH et al., 2015). Esses trabalhos contribuíram substancialmente para uma melhor compreensão das relações ainda não elucidadas entre grupos taxonômicos complexos.

Portanto, a investigação de sequências conservadas de DNA em outras linhagens de vertebrados e invertebrados tem potencial para a reconstrução da história evolutiva entre grupos de organismos de diferentes níveis taxonômicos. Além disso, fornece informações úteis sobre a origem e o papel desses elementos na evolução de organismos de diferentes distâncias evolutivas.

1.1 OBJETIVO GERAL

Este trabalho visa avaliar a possível utilidade das UCEs como marcadores moleculares para reconstrução das relações filogenéticas de alto nível entre metazoários.

1.2 OBJETIVOS ESPECÍFICOS

- Desenvolver uma *pipeline* para identificação e processamento de UCEs em genomas de metazoários;
- Verificar a sua utilidade para a inferência filogenética a partir do grau de compartilhamento destes marcadores entre os grandes grupos de metazoários;
- Confirmar os intrigantes padrões reportados por Ryu e colaboradores (2012) de alto grau de compartilhamento de UCEs entre táxons tão díspares quanto anêmonas e estrelas do mar, cujo tempo de separação é possivelmente pré-Cambriano.

2 REVISÃO BIBLIOGRÁFICA

2.1 FILOGENIAS

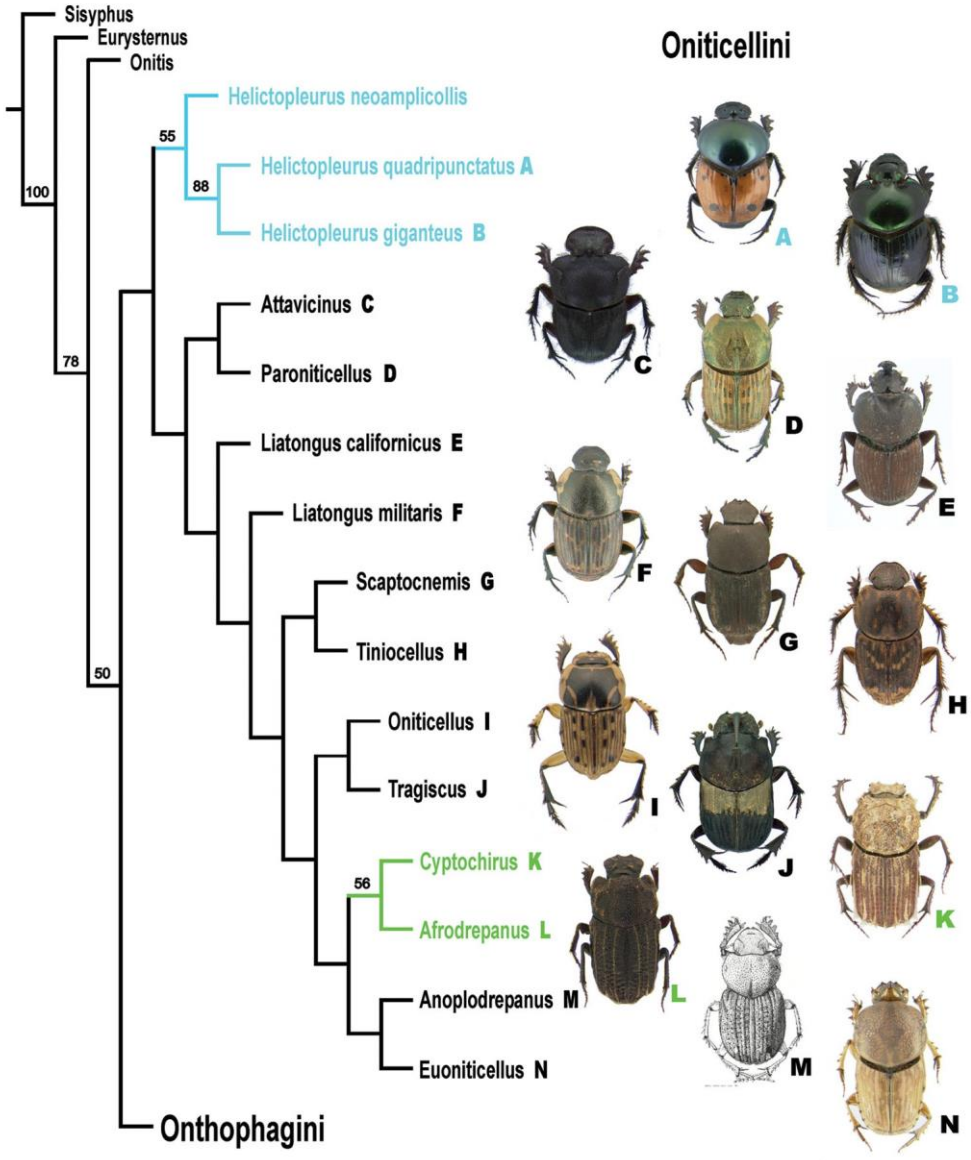
Filogenia é um ramo da ciência que busca estabelecer e analisar as relações evolutivas entre táxons. Seu objetivo é estimar uma árvore evolutiva entre várias espécies de seres vivos, assumindo que diferentes espécies derivam de um ancestral comum (TELFORD et al., 2015; PACE et al., 2012). A primeira descrição formal filogenética dos Metazoários, e a origem do próprio termo “filogenia”, foi publicada por Haeckel em 1866 (PACE et al., 2012). Taxonomia, identificação, classificação e nomeação de organismos, são genericamente designadas por análises filogenéticas. O termo também pode ser aplicado à genealogia de genes derivados de um gene ancestral comum. Na filogenia molecular, as relações entre organismos ou genes são estudadas pela comparação de homólogos de DNA ou sequências proteicas (TELFORD et al., 2015; PACE et al., 2012).

Até a década de 70 as reconstruções filogenéticas se baseavam quase exclusivamente em análises de caracteres morfológicos (SAVOLAINEN e CHASE, 2003). Por esse método, as relações entre diferentes espécies eram estabelecidas pela distinção de um conjunto de características observáveis, tais como: planos de simetria (radial ou bilateral), sistemas de células (organizados como tecidos ou órgãos mais especializados), osteologia, entre outros (SAVOLAINEN e CHASE, 2003; TELFORD et al., 2015).

Com base na semelhança dessas características morfológicas, duas espécies poderiam ser agrupadas, juntamente com outros organismos que compartilhavam a mesma característica (FIGURA 2) (TELFORD et al., 2015). Entretanto, essa metodologia pode levar a erros, considerando-se que duas espécies sem um ancestral comum podem desenvolver características similares. Caracteres morfológicas podem evoluir várias vezes, e de forma independente, através da história evolutiva (FUTUYMA, 1998; STAYTON, 2008).

Atualmente, a classificação taxonômica utiliza além dos dados morfológicos e dos registros fósseis, dados bioquímicos, fisiológicos e moleculares, os quais têm crescido proporcionalmente ao desenvolvimento tecnológico (MINDELL, 2013; TELFORD et al., 2015).

FIGURA 2 – FILOGENIA DOS BESOUROS ONIOPHELLINI OBTIDA A PARTIR DE EVIDÊNCIAS MORFOLÓGICAS.



FONTE: PHILIPS (2016).

Além de representar as relações entre as espécies na Árvore da Vida (DARWIN, 1859), análises filogenéticas também são usadas para explicar a origem de sistemas orgânicos e outros traços complexos, descrever relações entre parálogos em uma família de genes, histórias de populações, dinâmica evolutiva e epidemiológica de agentes patogênicos, relacionamento genealógico de células somáticas durante a diferenciação e desenvolvimento de câncer, entre outros (SADAVA et al., 2009; FREEMAN e HERRON, 2009). Filogenias também são a base de programas de pesquisas em biogeografia histórica, filogeografia, ecologia

histórica e coevolução (ECKERT e HALL, 2006; EMERSON et al., 2011; HARDY e LINDER, 2007; LAURON et al., 2015).

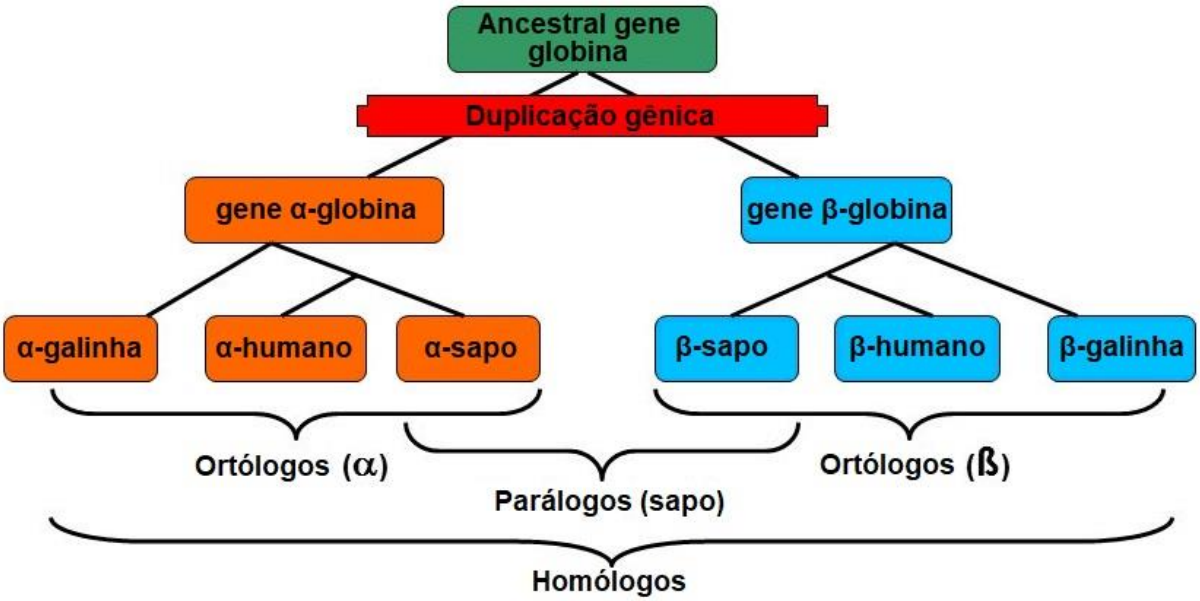
2.1.1 Homologia e conceitos

Genômica comparativa é o uso da evolução molecular como ferramenta na investigação de processos biológicos (MILLER et al., 2004). As sequências de nucleotídeos comuns aos genomas de várias espécies divergentes são indicativas da biologia compartilhada, ao passo que diferenças na sequência e estrutura genômica podem esclarecer o que diferencia espécies. A identificação de elementos genômicos que foram conservados ao longo do tempo permite que pesquisadores concentrem suas experiências nas partes do genoma que são fundamentais para grande parte de processos vitais (MILLER et al., 2004).

O conceito de homologia está frequentemente implícito nas discussões de sequências comuns ou conservadas (MILLER et al., 2004). Homologia é definida como a relação que existe entre duas estruturas (por exemplo, duas sequências genômicas e/ou proteicas ou dois caracteres morfológicos) derivadas de um ancestral comum e refere-se à correspondência de estruturas e caracteres entre diferentes espécies (MILLER et al., 2004; GABALDÓN, 2008). Para classificar os diferentes tipos de homologia, Fitch introduziu em 1970 os termos 'ortologia' e 'paralogia' (GABALDÓN, 2008). Duas estruturas ou sequências são consideradas ortólogas se derivadas de um evento de especiação a partir de um ancestral comum, e parálogas se derivadas de um evento de duplicação na mesma espécie (SADAVA et al., 2009). Por exemplo, o gene β-globina humana é um parálogo do gene α-globina humana correspondente e ortóloga do gene β-globina do sapo (FIGURA 3).

De acordo com a definição original de ortologia, o tipo de relação homóloga entre qualquer par de genes ou características relacionadas pode ser estabelecida com base na história de divergência de seus ancestrais comuns (SADAVA et al., 2009). Assim sendo, em sequências genômicas de espécies diferentes, os genes ortólogos são identificáveis como sequências similares, que codificam produtos também similares e com funções semelhantes em cada uma das espécies. Quanto maior a presença de genes ortólogos na comparação de genomas de duas espécies, maior será sua proximidade na escala evolutiva (MILLER et al., 2004; GABALDÓN, 2008).

FIGURA 3 – DIAGRAMA MOSTRANDO HISTÓRIA EVOLUTIVA HIPOTÉTICA DO GENE GLOBINA.



FONTE: A AUTORA (2018).

Com isso, o agrupamento de genes ortólogos fornece uma ferramenta valiosa para a genômica comparativa na anotação de genes e integração de informações na análise de múltiplos genomas, destacando a divergência e a conservação de famílias de genes e processos biológicos (APPEL e FEYTMANS, 2009; SADAVA et al., 2009).

2.2 FILOGENIA MOLECULAR E A CLASSIFICAÇÃO TAXONÔMICA

Nos últimos anos, o rápido desenvolvimento e barateamento das técnicas de sequenciamento de DNA possibilitaram o sequenciamento completo de genomas simples, como de bactérias e organismos unicelulares, e complexos, como de animais, plantas e fungos. Além de fornecerem uma nova estratégia para reconstruções filogenéticas e contribuírem para a composição e atual concepção da classificação taxonômica das espécies (JENNINGS, 2016).

Pela abordagem da filogenia molecular, a relação entre grupos de genes ou espécies é estudada pela comparação de conteúdos de sequências gênicas ou proteicas. Dissimilaridades entre essas sequências indicam divergência genética decorrentes da evolução molecular ao longo do tempo (FREEMAN e HERRON, 2009; JENNINGS, 2016).

Em resumo, ao passo que a abordagem filogenética clássica depende das características morfológicas de um organismo, as abordagens moleculares utilizam sequências de nucleotídeos de RNA e DNA e sequências de aminoácidos de uma proteína que são determinadas usando técnicas de biologia molecular (JENNINGS, 2016).

Os primeiros passos para reconstrução da filogenia de metazoários a partir de dados moleculares ocorreram nos anos 80 e foram baseados principalmente nos genes ribossomais nucleares 18S e 28S (FIELD et al.,1988; LAKE,1990; CHRISTEN et al., 1991). Nessa filogenia, Cnidaria, Ctenophora, Placozoa, Porifera constituem o grupo basal de metazoários, acompanhados de Bilateria numa posição mais derivada.

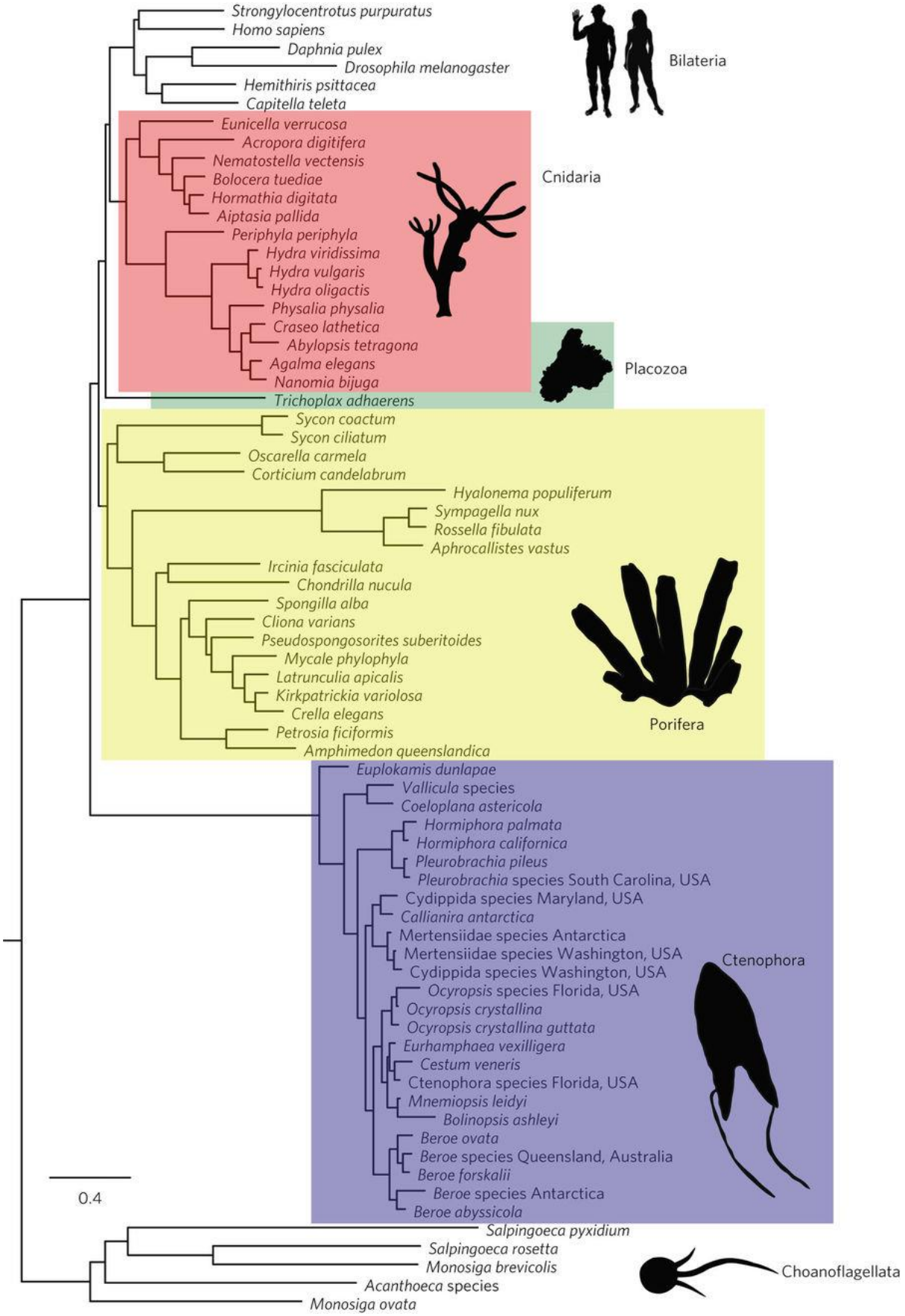
Desde então, uma grande variedade de marcadores moleculares têm sido utilizada pela biologia molecular, como por exemplo espaçadores intergênicos (e.g. AMBROSE e CREASE, 2011), ordem dos genes no genoma mitocondrial (e.g. CHAN et al., 2010), sequências de genes codificadores de proteínas (CHATRACHATCHAYA et al., 2016), genes codificadores de DNA ribossomal (e.g. SONG et al., 2018), polimorfismos de DNA (SNPs) (e.g. LEACHÉ e OAKS, 2017), UCEs (e.g. MCCORMACK et al., 2012), transcripoma (ESTs) (e.g. WEY-FABRIZIUS et al., 2014; GUZMAN e CONACO, 2016) entre outros.

Os resultados obtidos com esses marcadores e outras análises filogenéticas dos metazoários deram suporte para algumas das visões tradicionais difundidas pelas pesquisas iniciais com genes ribossomais nucleares por FIELD (1988), como a posição basal de Porifera, Placozoa, Ctenophora e Cnidaria em relação aos Bilateria (DELLAPORTA et al., 2006; GIRIBET et al., 2007; MEDINA et al., 2001).

Por outro lado, novas hipóteses a respeito das relações dos metazoários também vêm sendo propostas, como o já bem resolvido status monofilético de Bilateria que dá origem a duas linhagens principais, Protostomia e Deuterostomia (HEJNOL et al., 2009; PHILIPPE et al., 2009). Os animais protostomados, por sua vez, seriam divididos nos Lophotrochozoa e Ecdysozoa (PHILIPPE et al., 2005).

Atualmente, a filogenia metazoária recente mais aceita – compilada com base na congruência de inúmeros estudos e novos dados de transcriptomas – é representada na Figura 4. O acúmulo de dados moleculares e análises, entretanto, ainda não trouxe uma filogenia consenso robusta e resolvida para os metazoários.

FIGURA 4 – RECONSTRUÇÃO FILOGENÉTICA MAIS RECENTE DOS METAZOÁRIOS.



FONTE: WHELAN et al. (2017).

Nesse contexto, examinamos o potencial dos UCEs (FAIRCLOTH et al., 2012; MCCORMACK et al., 2012) como marcadores moleculares para resolver relações filogenéticos desafiantes.

2.3 DADOS GENÔMICOS E FILOGENÔMICA

Atualmente, grande parte das análises filogenéticas moleculares baseiam-se na análise comparativa de genomas como, por exemplo, a comparação de genes ortólogos ou de sequências gênicas ou proteicas (SACCONE e PESOLE, 2003). Nesse contexto, a genômica comparativa explora genomas por diferentes estratégias, variando desde a análise de rearranjos genômicos em grande escala à comparação de características genômicas como genes, RNAs, sequências de codificação, regiões reguladoras e, por fim, polimorfismos de nucleotídeo único (SNPs) (TOUCHMAN, 2010).

Essa técnica está relacionada com a evolução de organismos, sendo amplamente utilizado para estabelecer relações evolutivas entre diferentes espécies (BROWN, 2008). As relações entre os táxons são inferidas com base na homologia (padrões de similaridade de sequências biológicas) em genomas inteiros, seja na comparação gene a gene (SZÖLLOSI et al., 2012), múltiplos genes (YUTIN et al., 2012) ou pela abordagem de genoma completo (RANNALA e YANG, 2008).

Fora do contexto filogenético, diversas aplicações potenciais da genômica comparativa têm sido reportadas. Por exemplo, na área de medicina molecular, a identificação de alvos de drogas para doenças infecciosas (análises comparativas de genomas podem levar à identificação de diversos alvos putativos para novos fármacos), reconhecimento de vias metabólicas, identificação de genes, mecanismo de resistências bacterianas, prognóstico de câncer, etc (ODDS, 2005; RODIONOV et al., 2004; BATZOGLOU et al., 2000; JEUKENS et al., 2017; WALDRON et al., 2014).

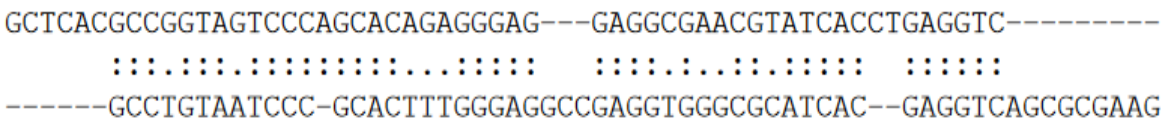
Um alinhamento é definido como um pareamento de sequências gênicas ou proteicas de modo a identificar similaridades presentes em ambas as sequências (MOUNT, 2004). Atualmente, existem principalmente dois tipos de algoritmos de alinhamentos utilizados por ferramentas de busca por similaridade em sequências biológicas: o alinhamento global e o alinhamento local (SACCONE e PESOLE, 2003).

No alinhamento global, duas sequências são comparadas por todo seu comprimento, potencializando as regiões de similaridade e reduzindo os espaçamentos (*gaps*), logo esta é a metodologia apropriada para comparar sequências cujas semelhanças ocorrem em toda sua extensão. Para se alcançar o melhor alinhamento global entre duas sequências biológicas, o algoritmo de Needleman-Wunsch é utilizado (NEEDLEMAN e WUNSCH, 1970). O método de alinhamento local, por sua vez, permite a pesquisa de segmentos da sequência biológica principal sem considerar todo o seu comprimento. Essa técnica é utilizada para comparar sequências que possuem geralmente tamanhos desiguais nas quais suspeita-se que hajam regiões de similaridade descontínuas. O algoritmo de alinhamento local ótimo é o Smith-Waterman (SMITH e WATERMAN, 1981).

Para cada alinhamento realizado é atribuída uma pontuação ou escore total, que consiste na soma de cada par de resíduo alinhado mais as inserções e remoções de resíduos, conhecidos como *gaps* (GUSFIELD, 2010). A similaridade entre um par de sequências é definida como o maior escore de um alinhamento. Para calcular o escore do alinhamento, resíduos entre um par de sequências recebem pontuação positiva se equivalentes (*matches*) ou negativa se divergentes (*mismatches*). *Gaps* são espaçamentos (lacunas) inseridos pelo algoritmo de modo a maximizar o número possível de caracteres idênticos para sequências alinhadas (BROWN, 2008; GUSFIELD, 2010).

A Figura 5 apresenta um exemplo de alinhamento entre duas sequências, onde os *matches* estão representados pelo símbolo ":", *mismatches* pelo símbolo "." e *gaps* pelo símbolo "–".

FIGURA 5 – EXEMPLO DE ALINHAMENTO ENTRE DUAS SEQUÊNCIAS BIOLÓGICAS.



FONTE: A AUTORA (2018).

Com esses valores, uma matriz de substituição é produzida utilizando a projeção de uma sequência contra outra contendo uma linha inicial em branco e uma coluna inicial também em branco (BROWN, 2008; GUSFIELD, 2010).

2.3.1 Bancos de dados e ferramentas computacionais

Como consequência do crescimento exponencial de sequências biológicas disponibilizadas nos últimos anos, principalmente desde o Projeto Genoma, Proteoma e Transcriptoma, surgiu a necessidade da criação de redes em biologia molecular e bancos de dados biológicos para atender à demanda em pesquisas nestas áreas (PEVSNER, 2009; PEARSON, 2014).

Um repositório importante para os estudos de filogenética é o banco de dados de genomas do *National Center for Biotechnology Information* (NCBI), que mantém um website (disponível em www.ncbi.nlm.nih.gov) com diversas ferramentas de bioinformática. O GenBank (BENSON et al., 2015) é o principal banco de dados do NCBI e armazena todas as sequências gênicas e proteicas disponíveis publicamente (PEVSNER, 2009; MOUNT, 2004). Atualmente o NCBI armazena genomas de cerca de 34 mil espécies (consultado em fevereiro de 2018), sendo a maior parte de bactérias (21.198 espécies com genoma sequenciado), seguido de vírus (9.526 espécies) e eucariotos (2.732 espécies). Dados genômicos também podem ser obtidos em outros repositórios, como no EMBL (*European Molecular Biology Laboratory*) (STOESSER et al., 1997) e no ENSEMBL (HERRERO et al., 2016). Além de disponibilizar as sequências genômicas, esses bancos ainda concedem as anotações dos genes, que são dados bastante empregados em análises filogenéticas.

O conjunto de proteínas de um organismo também pode ser adquirido nestes bancos ou, ainda, em bancos de dados especializados em sequências proteicas como o UniProt (CHEN et al., 2017) e o *Protein Database Bank* (*PDB*) (BERMAN et al., 2003). Outra ferramenta importante disponível para a manipulação de dados é o algoritmo BLAST (*Basic Local Alignment Search Tool*), que permite análises, buscas e comparações entre sequências biológicas (ALTSCHUL et al., 1997). O programa BLAST foi desenvolvido para realizar buscas comparando sequências gênicas ou proteicas contra um banco de dados de domínio público, retornando sequências com maior similaridade e de maior significância estatística. O algoritmo BLAST é seguramente a ferramenta computacional de referência para a busca de similaridade, análises filogenéticas e identificação de membros de famílias gênicas (PEVSNER, 2009; PEARSON, 2014).

2.4 ELEMENTOS ULTRACONSERVADOS

Por convenção, UCEs foram definidos como regiões do genoma altamente conservadas (sequências de DNA com mais de 80% de correspondência idêntica e ao menos 100 pares de bases) entre organismos de diferentes táxons (BEJERANO et al., 2004; FAIRCLOTH et al., 2012). Essa categoria de sequências conservadas de DNA foi reportada em 481 sequências maiores que 200 pb compartilhadas entre o genoma humano, de camundongos e de ratos, dos quais mais da metade não mostram evidências de transcrição (BEJERANO, 2004).

Elas também são altamente conservadas em outros grupos de mamíferos e aves (99% e 95% de correspondência idêntica nos genomas homem/cão e homem/galinha, respectivamente). UCEs mais curtas também já foram documentadas em mamíferos (MCCORMACK et al., 2012), insetos (FAIRCLOTH et al., 2015), répteis (CRAWFORD et al., 2012), peixes (FAIRCLOTH et al., 2013) e até mesmo em plantas (ZHENG e ZHANG, 2008). Em contraste, elas apresentam-se pouco conservadas entre grandes distâncias evolutivas: apenas 5% (24 das 481) foram parcialmente identificadas em genomas de invertebrados como urocordados, moscas e vermes (SANGES et al., 2013; MAKUNIN et al., 2013; BEJERANO et al., 2005).

Apesar de suas funções ainda serem pouco esclarecidas, estudos sugerem que alguns desses elementos ultraconservados não-codificantes parecem desempenhar papel como acentuadores de longo alcance de genes reguladores do desenvolvimento, reguladores epigenéticos ou outras funções na regulação genes (VISEL et al., 2013; LINDBLAD-TOH et al., 2011; PENNACCHIO et al., 2006; BERNSTEIN et al., 2006; WOOLFE et al., 2004). Polimorfismos de nucleotídeo único (SNPs) em UCEs também foram associados como fator de risco para desenvolvimento de câncer (YANG et al., 2008; LIN et al., 2012).

Experimentos *knockout* realizados por Ahituv e colaboradores (2007) indicam que as UCEs são dispensáveis para a viabilidade de camundongos. Apesar disso, foi proposto que sua remoção *in vivo* pode levar a um impacto fenotípico significativo. Em estudo, Dickel e colaboradores (2018) demonstraram que UCEs que atuam como acentuadores são necessárias para o desenvolvimento cerebral normal de ratos.

2.4.1 Análise filogenética inferida a partir de elementos ultraconservados

Como marcadores moleculares, UCEs apresentam características que os tornam particularmente desejáveis para estudos de filogenia e contexto evolutivo. A presença de sequências altamente conservadas facilita sua extração e identificação através de ferramentas da genômica comparativa (BEJERANO, 2004); são encontradas em números elevados em todo genoma (STEPHEN et al., 2008); UCEs também raramente são encontradas em regiões duplicadas (DERTI et al., 2006) e tendem a ser ortólogos com poucos retroelementos de inserção (MCCORMACK et al., 2012).

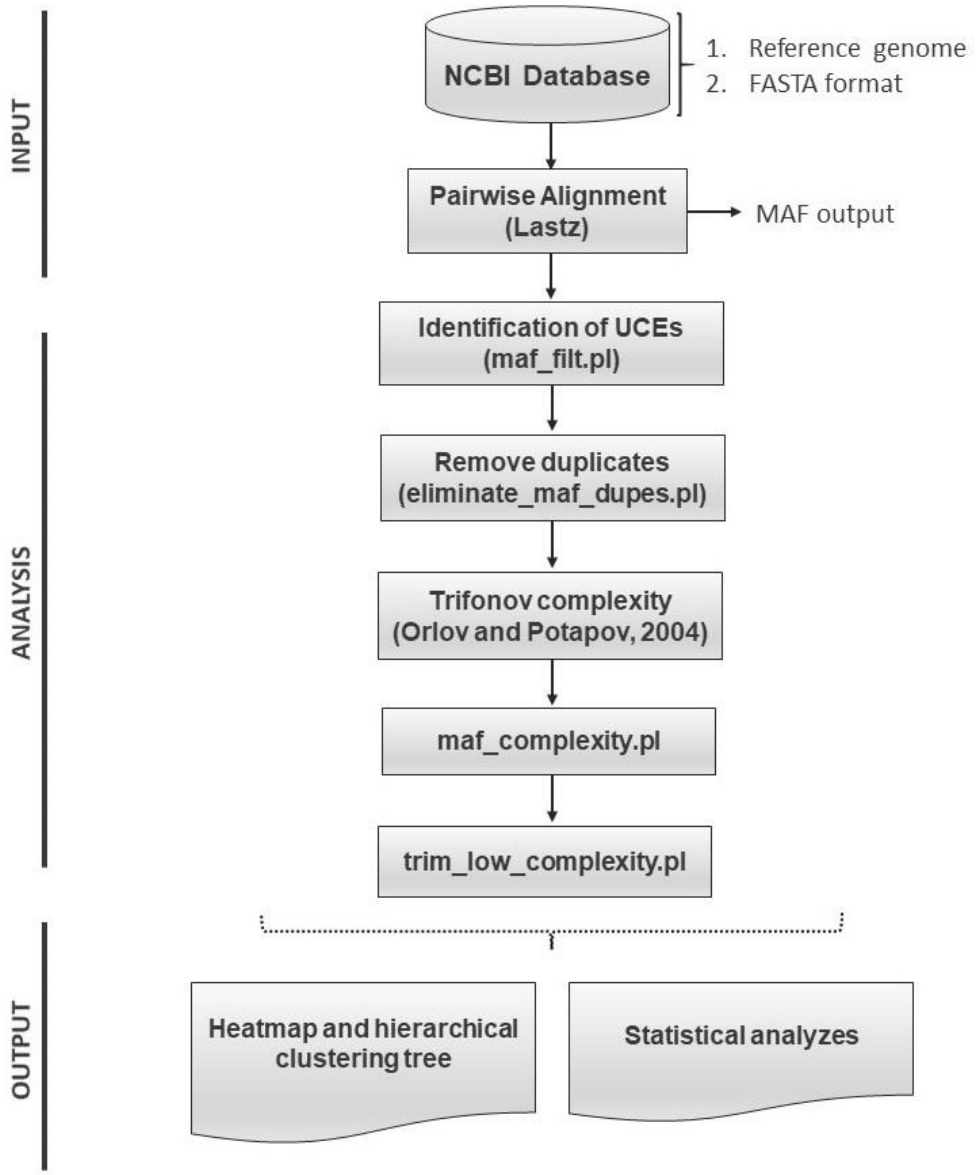
Adicionalmente, esses elementos conservados são flanqueados por regiões de maior variabilidade, o que os tornam úteis tanto para pesquisas com foco em divergências filogenéticas antigas (CRAWFORD et al., 2012) como em contextos mais recentes (SMITH et al., 2014). A premissa de aumento na variabilidade de sequências flanqueadoras de UCEs sugere que elas possam reter sinal evolutivo de diferentes épocas dependendo da distância da UCE principal, como uma espécie de “fóssil molecular” (FAIRCLOTH et al., 2012). Esta característica permite que os pesquisadores adaptem o uso das UCEs, selecionando aquelas com taxas evolutivas semelhantes ou selecionando regiões flanqueadoras que otimizem suas análises.

Esta técnica já foi aplicada em estudos com himenópteros (FAIRCLOTH et al., 2015; BLAIMER et al., 2015) e utilizada para reconstruir filogenias de táxons tão divergentes quanto mamíferos (BEJERANO et al., 2004), peixes (FAIRCLOTH et al., 2013), aves (WHITE et al., 2017; MCCORMACK et al., 2013), répteis (STREICHER e WIENS, 2017; CRAWFORD et al., 2012) e anfíbios (STREICHER et al., 2018).

3 METODOLOGIA

Este estudo propõe o desenvolvimento de uma ferramenta personalizada para a identificação de UCEs no genoma de metazoários vertebrados e invertebrados de diferentes distâncias evolutivas. O trabalho foi desenvolvido utilizando o comando de linha no sistema operacional Linux. As principais linguagens de programação utilizadas foram R, PERL e Python. Os códigos-fonte encontram-se nos apêndices desse trabalho. Segue abaixo um fluxograma representando os processos de execução deste estudo (FIGURA 6).

FIGURA 6 – PIPELINE DOS PROCESSOS DE EXECUÇÃO DO ESTUDO.



FONTE: A AUTORA (2018).

3.1 ALINHAMENTO E COMPOSIÇÃO DA MATRIZ DE DADOS

Neste estudo, definimos UCEs como sequências de DNA de pelo menos 40 pares de base (pb) e com 100% de similaridade nos genomas de metazoários (FAIRCLOTH et al., 2015). Selecionamos 90 organismos referências que incluem representantes em diferentes filos de invertebrados (incluindo Porifera, Cnidaria, Arthropoda, Echinodermata, Placozoa, entre outros) e principais classes de vertebrados (Mammalia, Amphibia, Reptilia, Aves e Peixes). A lista completa das espécies encontra-se no APÊNDICE A. As sequências completas dos genomas das espécies foram obtidas do NCBI (www.ncbi.nlm.nih.gov), disponíveis em formato FASTA de sequências de nucleotídeos (genomic.fna.gz). Os genomas foram alinhados par a par utilizando a ferramenta LASTZ (HARRIS, 2007), com os parâmentos “--notransition --step=20 --nogapped --format=maf” (script em APÊNDICE B).

As análises foram executadas de forma remota no Ohio Supercomputer Center (OSC) (https://www.osc.edu), nas clusters Oakley e Owens, a partir da estação de trabalho Linux local. O programa “lastz_maf.pl” foi usado para o processamento do arquivo MAF resultante contendo os alinhamentos para computar e explorar o grau conservação entre as sequências (script em APÊNDICE C).

As UCEs foram identificadas utilizando o programa “maf_filt.pl” (script em APÊNDICE D), retendo apenas aquelas com 100% de similaridade e no mínimo 40pb. UCEs duplicadas foram removidas pelo programa “eliminate_maf_dupes.pl” (script em APÊNDICE E). Para análise estatística e quantificação das UCEs os arquivos maf foram convertidos para o formato csv (script em APÊNDICE F).

3.1.1 Filtragem

A filtragem de cópias múltiplas foi feita através da localização de endereços genômicos repetidos nos arquivos MAF, i.e., se a mesma região de um *scaffold* ou cromossomo aparecia alinhada mais de uma vez com o outro genoma. Não foi feita nenhuma tentativa de procura de duplicações por similaridade de sequências. A fim de se eliminarem as sequências com motivos reptitivos, computou-se a complexidade de Trifonov como descrita em Orlov e Potapov (2004). Resumidamente, esta métrica avalia a riqueza do vocabulário de cada UCE, isto é,

quantas palavras de comprimento *i* aparecem em sua sequência, onde *i* varia de 1 ao comprimento da UCE. Estes valores são normalizados em relação ao número máximo de palavras que poderiam ser encontrados em cada janela, de forma que o valor final varia entre 0 e 1. Não há um valor crítico abaixo do qual a sequência é considerada repetitiva.

A filtragem destas sequências foi feita em duas etapas. O script escrito em Perl “maf_complexity.pl” (script em APÊNDICE G) foi usado para gerar um arquivo csv em que se registrou a complexidade das UCEs de cada alinhamento MAF. Em seguida utilizou-se um segundo script “trim_low_complexity.pl” (script em APÊNDICE H) para filtrar as sequências abaixo de um valor crítico (padronizado nesse estudo em 0,95) destes arquivos, gerando um segundo conjunto que foi utilizado em análises subsequentes.

4 RESULTADOS E DISCUSSÃO

Foram selecionados 90 genomas de espécies diferentes, abrangendo 18 filos, 43 classes e 74 ordens de metazoários (TABELA 1). O tamanho dos genomas variou de 26,4 milhões de pares de bases (Mpb) no caso do cnidário parasita *Kudoa iwatai* (Cnidaria:Myxozoa:Multivalvulida – YAHALOMI et al., 2017) até um tamanho superior a 32 bilhões de pares de bases no anfíbio *Ambystoma mexicanum* (Chordata:Amphibia:Caudata – NOWOSHILOW et al., 2018). A distribuição dos tamanhos dos genomas é aproximadamente log-normal com mediana de 454,1 Mpb, sendo o genoma do axolote mais de 10 vezes maior que o segundo colocado, o da aranha marrom *Loxosceles reclusa* (Arthropoda:Arachnida:Aranae – POELCHAU et al., 2015), que possui 3.262,4 Mpb (FIGURA 7).

TABELA 1 – DADOS DE IDENTIFICAÇÃO DAS ESPÉCIES SELECIONADAS

Espécie	Genoma	Identificação	Tamanho (MB)	Filo/Classe	Cód. Genbank
Acanthaster planci	completo	acanthaster	383.7	Echinodermata	OKI-Apl_1.0
Acropora digitifera	completo	acropora	447.5	Cnidaria	Adig_1.1
Adineta vaga	completo	adineta	216.2	Rotifera	PRJEB1171
Aedes aegypti	completo	aedes	1278.7	Arthopoda	AaegL3
Alatina alata	parcial	alatina	125.6	Cnidaria	Não consta
Ambystoma mexicanum	completo	ambystoma	32393.6	Amphibia	AmbMex13_14
Amphimedon queenslandica	completo	amphimedon	166.7	Porifera	v1.0
Amplexidiscus fenestrafer	parcial	amplexidiscus	370.0	Cnidaria	Não consta
Anolis caroliniensis	completo	anolis	1799.1	Reptilia	AnoCar2.0
Apis mellifera	completo	apismellifera	235.3	Arthopoda	Amel_4.5
Aplysia californica	completo	aplysia	927.3	Mollusca	AplCal3.0
Ascaris lumbricoides	completo	ascaris	316.9	Nematoda	A_lumbricoides
Asymmetron lucayanum	completo	asymmetron	460.6	Cephalochordata	Asyluc0.1
Bombyx mori	completo	bombys	397.7	Arthropoda	ASM15162v1
Branchiostoma belcheri	completo	branchiostoma	426.1	Cephalochordata	Haploidv18h27
Caenorhabditis elegans	completo	celegans	101.2	Nematoda	WBcel235
Capitella teleta	completo	capitella	333.2	Annelida	Capca1
Centruroides exilicauda	completo	centruroides	925.5	Arthopoda	Cexi_1.0
Chelonia mydas	completo	chelonia	2208.4	Reptilia	CheMyd_1.0
Ciona intestinalis	completo	cintestinalis	115.9	Tunicata	KH
Crassostrea gigas	completo	crassostrea	557.7	Mollusca	oyster_v9
Crocodylus porosus	completo	crocodylus	2085.1	Reptilia	CroPor_comp1
Crotalus horridus	completo	crotalus	1520.3	Reptilia	ASM162548v1
Culex quinquefasciatus	completo	culexq	579.0	Arthropoda	CulPip1.0
Danio rerio	completo	daniorerio	1427.3	Peixe	GRCz10
Daphnia pulex	completo	daphnia	197.2	Crustacea	V1.0
Dendroctonus ponderosae	completo	dendroctonus	257.1	Arthropoda	DendPond
Drosophila melanogaster	completo	drosophila	137.7	Arthropoda	Release 6
Enterobius vermicularis	completo	enterobius	150.0	Nematoda	E_vermicularis
Enteromyxum leei	completo	enteromyxum	67.9	Cnidaria	ASM145529v1
Exaiptasia pallida	completo	aiptasia	256.1	Cnidaria	Aiptasia
Fasciola hepatica	completo	fasciola	1236.4	Platyhelminthes	Fhepatica_v1
Gallus gallus	completo	gallus	1043.2	Ave	Gallus_gallus
Geospiza fortis	completo	geospiza	1065.3	Ave	GeoFor_1.0
Helobdella robusta	completo	helobdella	235.4	Annelida	Helobdella
Homo sapiens	completo	homosapiens	2995.7	Mammalia	GRCh38.p10
Hyalella azteca	completo	hyalella	550.9	Crustacea	Hazt_2.0
Hydra vulgaris	completo	hydra	1055.6	Cnidaria	Hydra_RP_1.0
Hypochthonius rufulus	completo	hypochthonius	172.4	Arthropoda	ASM98884v1
Hypsibius dujardini	completo	hypsibius	182.2	Tardigrada	nHd_3.1
Intoshia linei	completo	intoshia	41.6	Orthonectida	JCVI_ISG_i3_1
Ixodes scapularis	completo	ixodes	1765.4	Arthropoda	IntLin1.0
Kudoa iwatai	completo	kudoaiwatai	26.4	Cnidaria	Kudoa1
Latimeria chalumnae	completo	latimeria	2798.5	Peixe	LatCha1

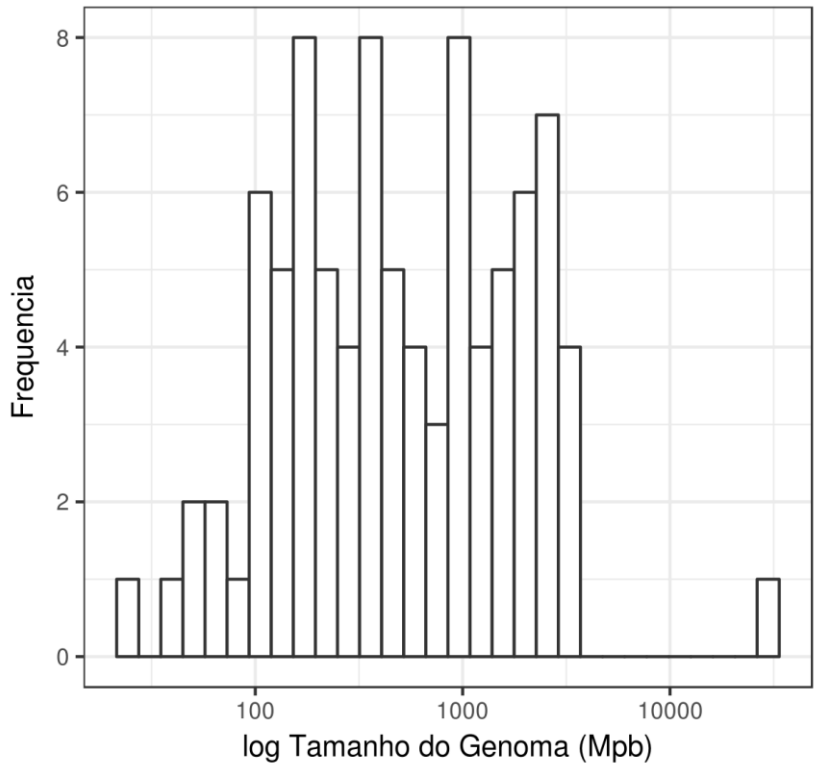
continua

Lepeophtheirus salmonis	completo	lsalmonis	665.3	Crustacea	lsal_atl_
Limulus polyphemus	completo	limulus	1828.3	Arthropoda	Limulus_
Lingula anatina	completo	lingula	406.3	Brachiopoda	LinAna1.0
Loa loa	completo	loaloa	93.9	Nematoda	Loa_loa_V3.1
Loxosceles reclusa	completo	loxosceles	3262.4	Arthropoda	Lrec_1.0
Mnemiopsis leidyi	completo	mnemiopsis	155.9	Ctenophora	MneLei_
Musca domestica	completo	musca	636.3	Arthropoda	Musca_
Mus musculus	completo	musm	2671.8	Mammalia	GRCm38.p6
Mytillus galloprovincialis	completo	mytillus	1561.4	Mollusca	ASM167691v1
Nanorana parkeri	completo	nanorena	2053.8	Amphibia	ASM93562v1
Nasonia vitripennis	completo	nasonia	295.8	Arthopoda	Nvit_2.1
Necator americanus	completo	necator	244.1	Nematoda	N_ americanus
Nematostella vectensis	completo	nematostella	356.6	Cnidaria	ASM20922v1
Notamacropus eugenii	completo	notamacropus	3075.2	Mammalia	Meug_1.1
Octopus bimaculoides	completo	octopus	2338.2	Mollusca	Octopus_
Oikopleura dioica	completo	oikopleura	57.8	Tunicata	ASM20953v1
Ophiothrix spiculata	completo	ophiothrix	2764.3	Echinodermata	Ospi.un_1.0
Orbicella faveolata	completo	ofaveolata	486.5	Cnidaria	ofav_dov_v1
Orcinus orca	completo	orcinusorca	2372.9	Mammalia	Oorc_1.1
Ornithorhynchus anatinus	completo	ornithorhynchus	1993.0	Mammalia	Ornithorhynchu
Pan troglodytes	completo	pantroglodytes	3050.4	Mammalia	Pan_tro 3.0
Parastichopus parvimensis	completo	parastichopus	873.0	Echinodermata	Ppar_1.0
Pediculus humanus	completo	pediculus	110.7	Arthopoda	JCVI_LOUSE_
Petromyzon marinus	completo	petromyzon	1007.9	Peixe	Petromyzon_
Pleurobrachia bachei	completo	pleurobrachia	156.1	Ctenophora	P.bachei_
Poecilia reticulata	completo	poecilia	731.6	Peixe	Guppy
Priapulus caudatus	completo	priapulus	511.7	Priapulida	Priapulus
Ptychodera flava	completo	ptychodera	1228.6	Hemichordata	ptychodera
Rattus norvegicus	completo	rattus	2743.3	Mammalia	Rnor_6.0
Renilla reniformis	completo	renilla	131.5	Cnidaria	Renilla
Saccoglossus kowalevskii	completo	sacoglossus	775.8	Hemichordata	Skow_1.1
Sarcoptes scabei	completo	sarcoptes	56.2	Arthropoda	SarSca1.0
Schistosoma mansoni	completo	mansoni	364.5	Platyhelmintes	ASM23792v2
Sphaeromyxa zaharoni	completo	sphaeromyxa	173.5	Cnidaria	ASM145528v1
Strigamia maritima	completo	strigamia	176.2	Arthropoda	Smar_1.0
Strongylocentrotus purpuratus	completo	strongylocentrotus	990.9	Echinodermata	Spur_4.2
Strongyloides venezuelensis	completo	strongyloides	52.1	Nematoda	S_venezuelensi
Taenia saginata	completo	taenia	169.1	Platyhelminthes	ASM169307v2
Tetraodon nigroviridis	completo	tetraodon	342.4	Peixe	ASM18073v1
Thelohanellus kitauei	completo	thelohanellus	150.3	Cnidaria	ASM82789v1
Trichoplax adhaerens	completo	trichoplax	105.6	Placozoa	v1.0
Triops cancriformis	completo	triops	109.1	Crustacea	tcf_1.0
Tyto alba	completo	tytoalba	1120.1	Ave	ASM68720v1
Ursus maritimus	completo	ursus	2301.3	Mammalia	UrsMar_1.0
Wuchereria bancrofti	completo	wuchereria	85.9	Nematoda	Wb_PNG_
Xenopus tropicalis	completo	xenopus	1440.4	Amphibia	Xenopus

Fonte: *National center for biotechnology information* (2018).

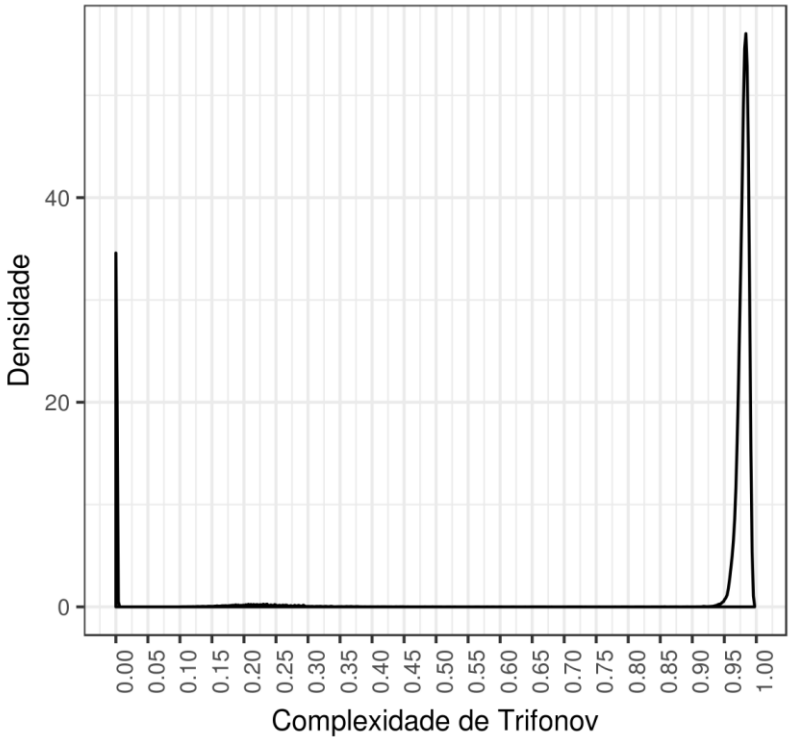
Foram realizados todos os 4.005 alinhamentos pareados possíveis para os 90 genomas elencados. Em aproximadamente 14% dos alinhamentos (548), nenhuma UCE foi recuperada; após a eliminação das cópias múltiplas, este número subiu para 19% (766). Foram subsequentemente eliminadas as UCEs com valor de complexidade de Trifonov inferior a 0,95. Este valor foi selecionado arbitrariamente após o exame visual da distribuição de complexidade através dos genomas (FIGURA 8) que é fortemente bimodal. Aproximadamente 85% da densidade é igual ou superior a 0,95, ao passo que pouco mais de 12% está abaixo de 0,05, o que corresponde a sequências altamente repetitivas. Pouco menos de 3% das sequências situa-se entre 0,05 e 0,95 (FIGURA 8).

FIGURA 7 – DISTRIBUIÇÃO EM ESCALA LOGARÍTMICA (BASE 10) DO TAMANHO DOS GENOMAS EM MILHÕES DE PARES DE BASES.



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

FIGURA 8 – DISTRIBUIÇÃO DE DENSIDADE DA COMPLEXIDADE DE TRIFONOV ENTRE AS UCES DE CÓPIA SIMPLES.

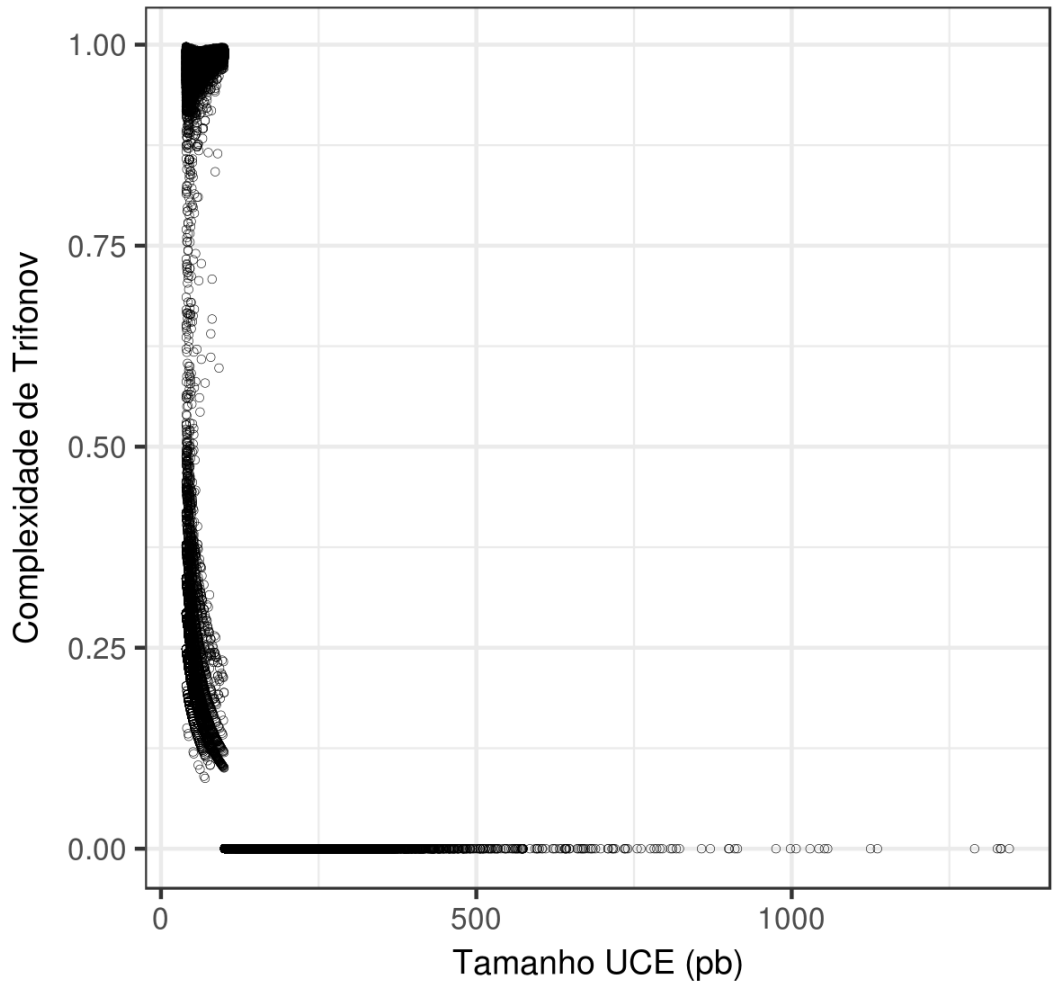


FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

Portanto, a opção por um valor de corte conservador removeu uma fração reduzida das sequências das análises subsequentes. O número máximo de UCEs (3.679.712) foi encontrado entre *Homo sapiens* (Choradata:Mammalia:Primates – INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001) e o chimpanzé *Pan troglotydes* (Choradata:Mammalia:Primates – MIKKELSEN et al., 2004) e é bem superior à soma de todas as UCEs encontradas em outros alinhamentos (320.140). Sendo assim, este alinhamento foi excluído do cômputo do valor de corte de complexidade e dos resultados nas Figuras 9, 10 e 11.

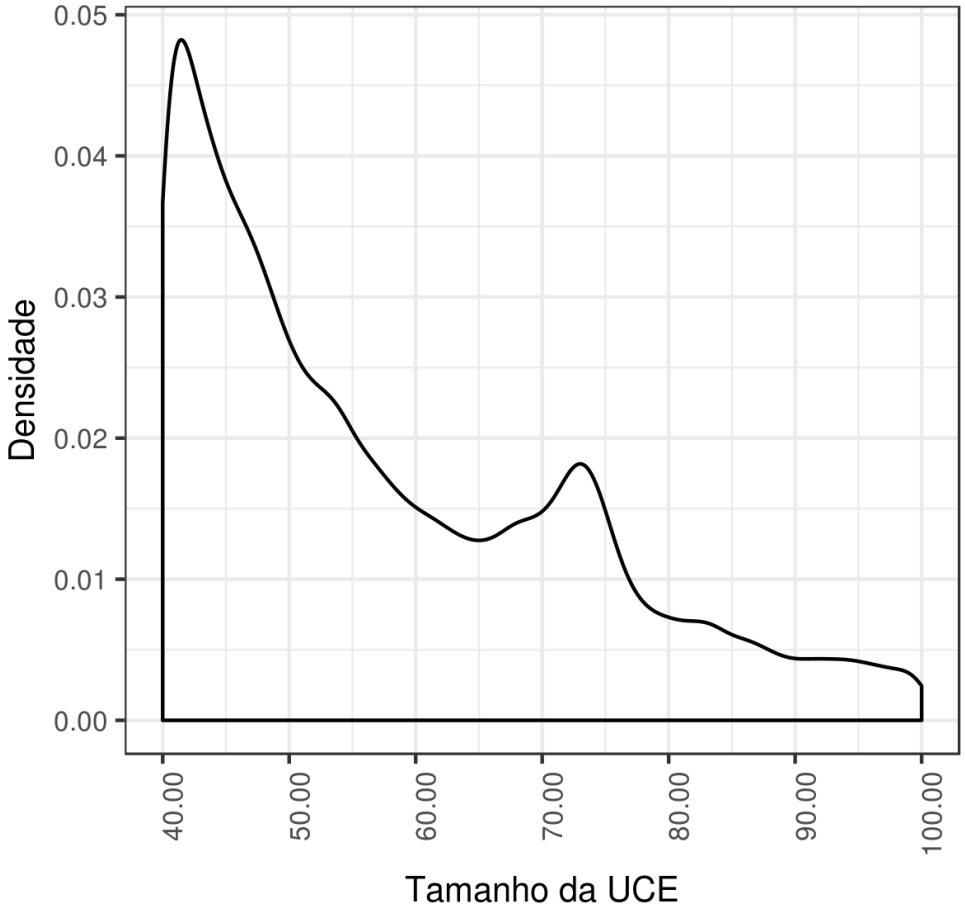
A Figura 9 mostra que a grande maioria das UCEs com mais de 100 pb são formadas por sequências de complexidade próxima a 0. Após a sua eliminação, a distribuição de tamanho é fortemente unicaudal e enviesada à esquerda, embora haja uma segunda moda por volta de 75 pb (FIGURA 10).

FIGURA 9 **–** DISPERSÃO DA COMPLEXIDADE DE TRIFONOV VS. TAMANHO DAS UCES EM PARES DE BASES.



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

FIGURA 10 – DISTRIBUIÇÃO DE TAMANHO DAS UCES APÓS A ELIMINAÇÃO DAS SEQUÊNCIAS COM COMPLEXIDADE DE TRIFONOV INFERIOR A 0.95.



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

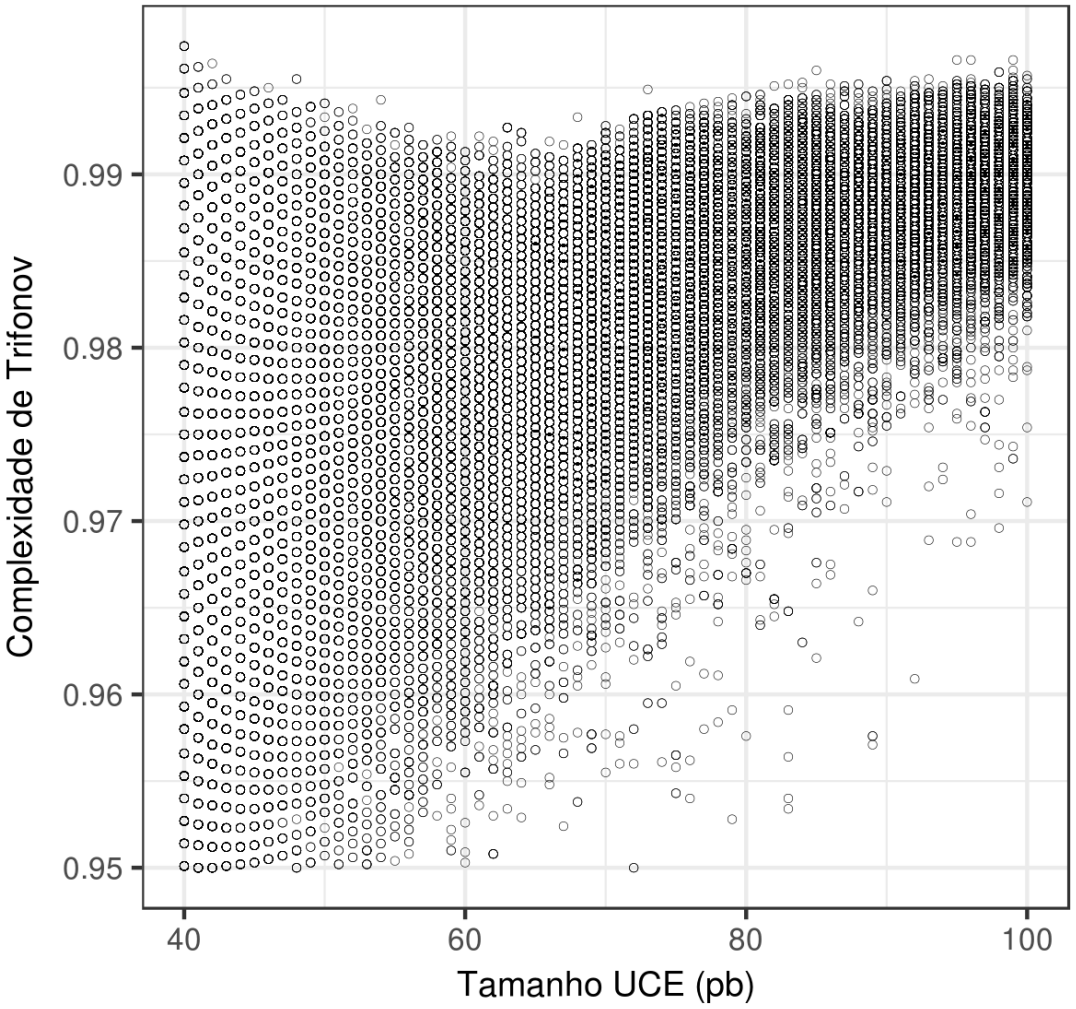
É interessante notar que, quando consideramos apenas as sequências acima do valor de corte, existe uma clara tendência de aumento da complexidade com o tamanho das UCEs (FIGURA 11), portanto a eliminação das sequências maiores não é função da escolha da métrica de Trifonov.

O número total de alinhamentos com 0 UCES subiu para cerca de 23% (905) após a eliminação das sequências com complexidade de Trifonov inferior a 0,95. A Figura 12a mostra o decaimento no número de UCEs a partir das filtragens sucessivas. A linha representando o maior número de UCEs corresponde ao alinhamento de *H. sapiens x P. troglotydes*. A eliminação das cópias múltiplas e sequências de menor complexidade eliminou cerca de 40% das UCEs, mas ainda assim este alinhamento reteve 2.188.334 sequências.

A Figura 12b mostra que o maior número de alinhamentos com 0 UCEs após a eliminação das cópias múltiplas e sequências de menor complexidade já

apresentavam um número inicial de UCEs relativamente reduzido (na ordem das dezenas ou inferior).

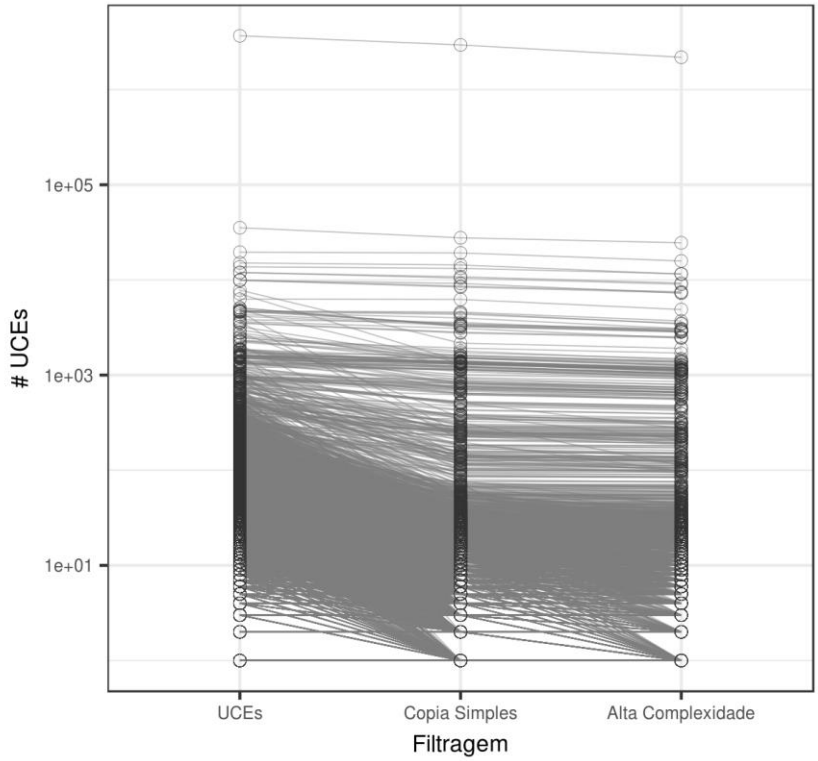
FIGURA 11 – COMPLEXIDADE DE TRIFONOV VS. TAMANHO DAS UCES PARA SEQUÊNCIAS DE COMPLEXIDADE ACIMA DO VALOR DE CORTE (0,95).



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

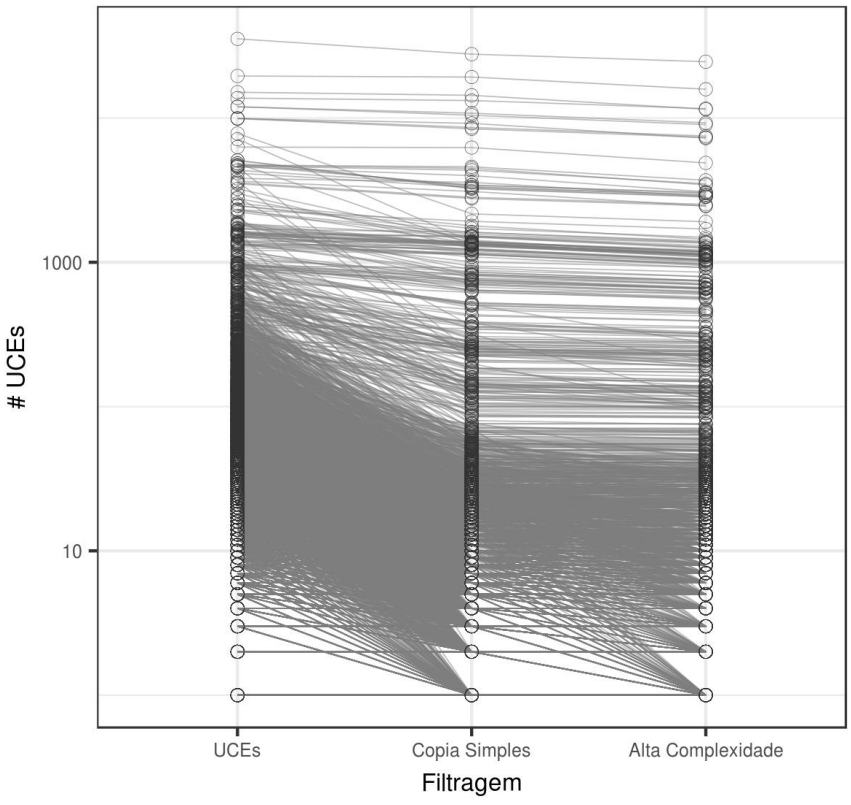
Note também que a etapa principal de remoção de UCEs parece ser a eliminação de sequências de cópias múltiplas e não a de sequências contendo motivos repetitivos (FIGURA 12B). Isso é melhor evidenciado pela Figura 13: em relação ao número original de UCEs encontradas por LASTZ, a maior parte dos alinhamentos apresentou uma queda superior a 75% (mediana ~ 78,6%) após a eliminação das cópias múltiplas, mas uma redução subsequente inferior a 5% (mediana ~ 83,3%) após a eliminação de sequências de menor complexidade.

FIGURA 12A – DECAIMENTO DO NÚMERO DE UCES LOCALIZADAS USANDO LASTZ, APÓS A ELIMINAÇÃO DAS SEQUÊNCIAS DE CÓPIAS SIMPLES E DE MENOR COMPLEXIDADE.



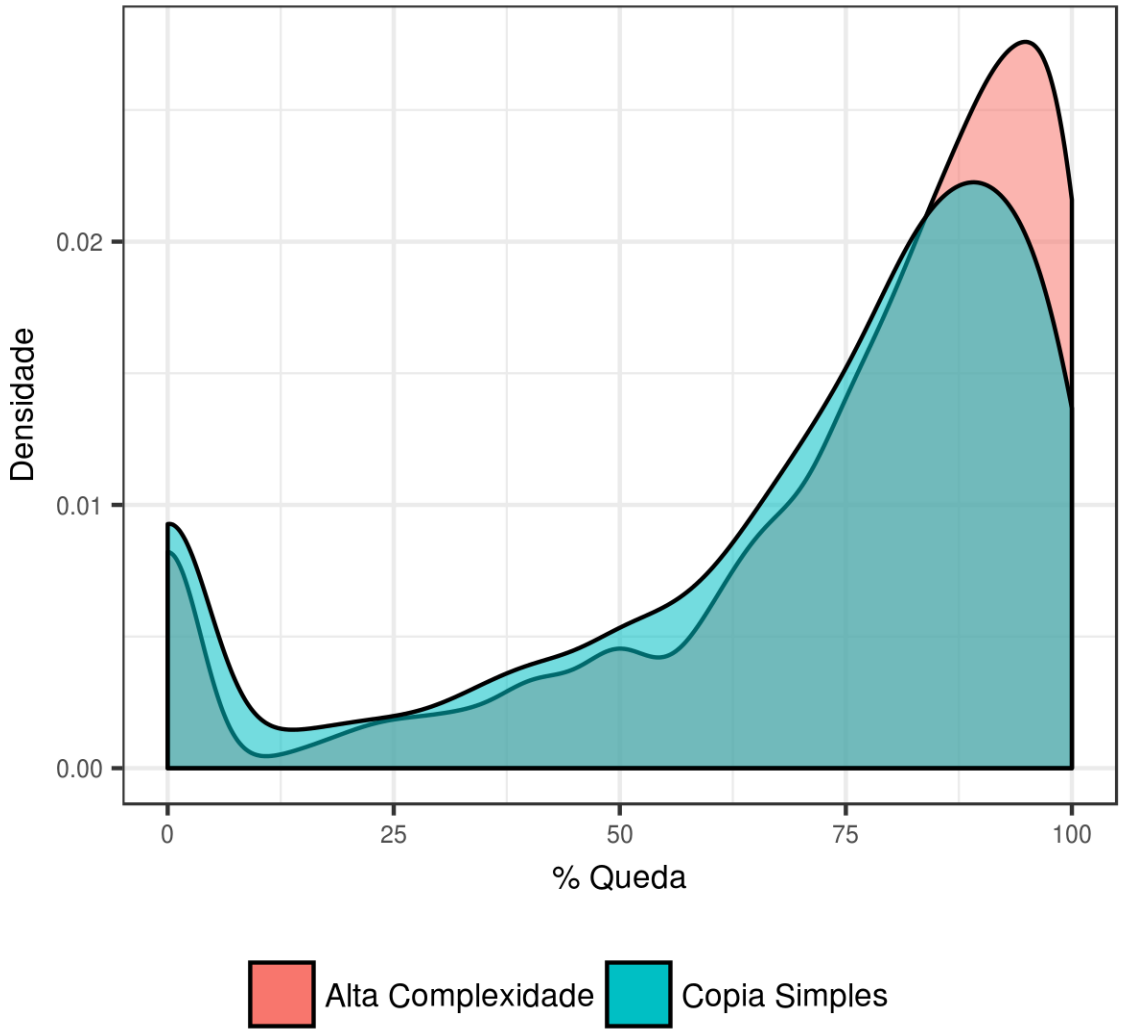
FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

FIGURA 12B – MESMOS RESULTADOS EXCLUINDO ALINHAMENTO DE *H. SAPIENS* X *P. TROGLOTYDES.*



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

FIGURA 13 – QUEDA PERCENTUAL EM RELAÇÃO AO NO NÚMERO INICIAL DE UCES APÓS A ELIMINAÇÃO DAS SEQUÊNCIAS COM CÓPIAS MÚLTIPLAS (EM AZUL) E DE SEQUÊNCIAS DE MENOR COMPLEXIDADE (EM VERMELHO).

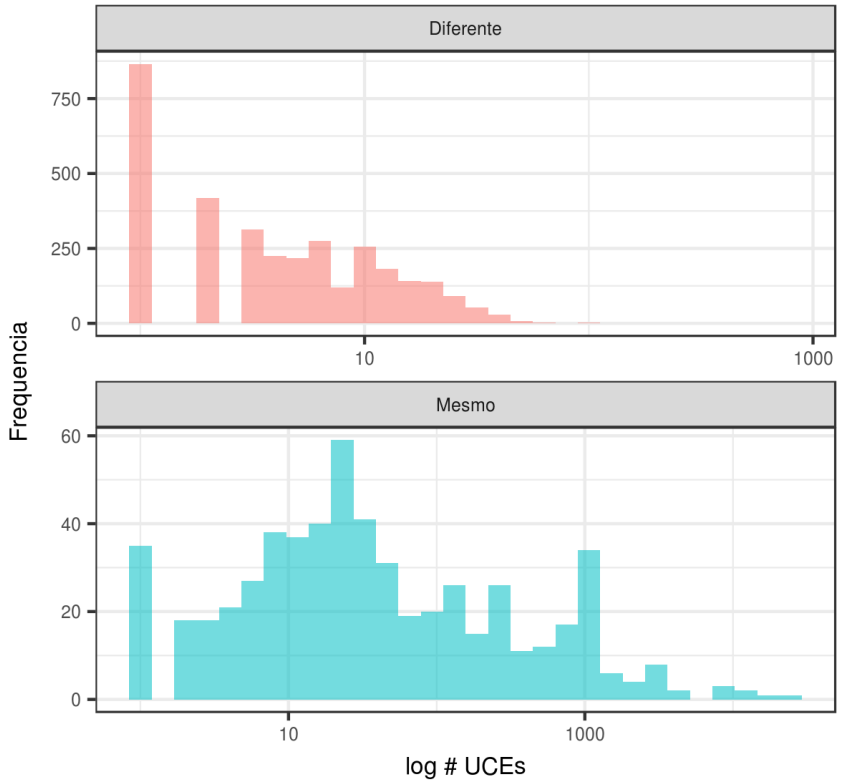


FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

A Figura 14 mostra que o número total de UCEs compartilhadas entre duas espécies é função da proximidade filogenética entre as mesmas. Quando considerados os números de UCEs em alinhamentos entre filos diferentes (vermelho – FIGURA 14A) em oposição alinhamentos entre pares de genomas de espécies de um mesmo filo (azul – FIGURA 14A), nota-se um considerável desvio para a esquerda, sendo a classe modal a de 0 UCEs.

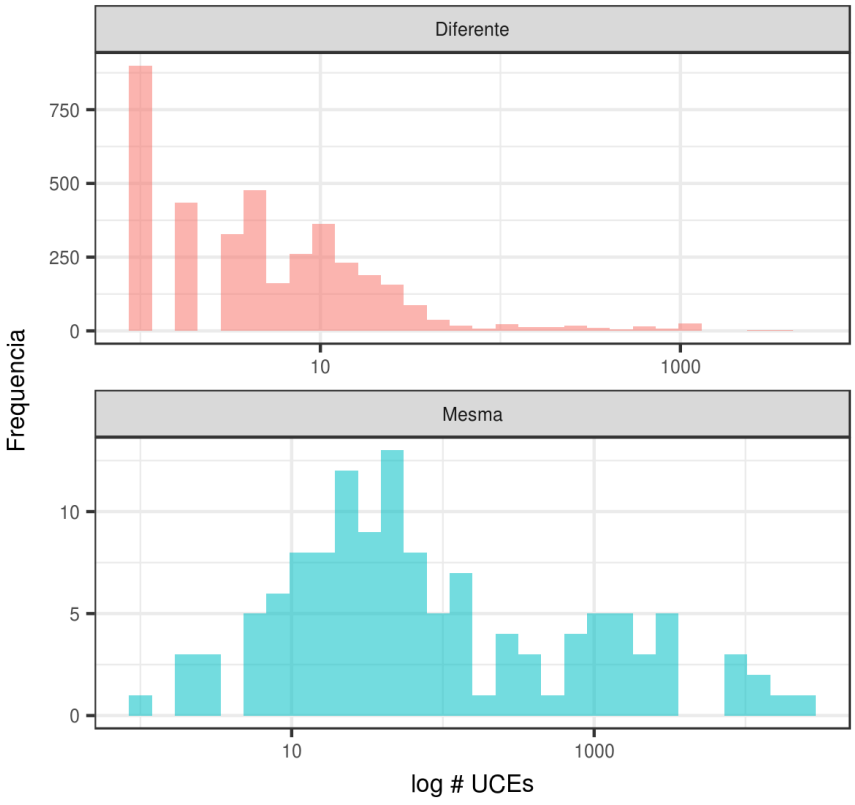
Naturalmente, um maior número de UCEs é encontrado entre espécies pertencentes à mesma classe (azul – FIGURA 14B), mas é interessante notar que, mesmo dentre estas, o número de alinhamentos com 10 ou menos UCEs não é desprezível.

FIGURA 14A – COMPARAÇÃO ENTRE OS NÚMEROS DE UCES ENCONTRADOS EM ALINHAMENTOS ENTRE ESPÉCIES DE DIFERENTES FILOS X MESMO FILO.



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

FIGURA 14B – COMPARAÇÃO ENTRE OS NÚMEROS DE UCES ENCONTRADOS EM ALINHAMENTOS ENTRE ESPÉCIES DE CLASSES DIFERENTES X MESMA CLASSE.



FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

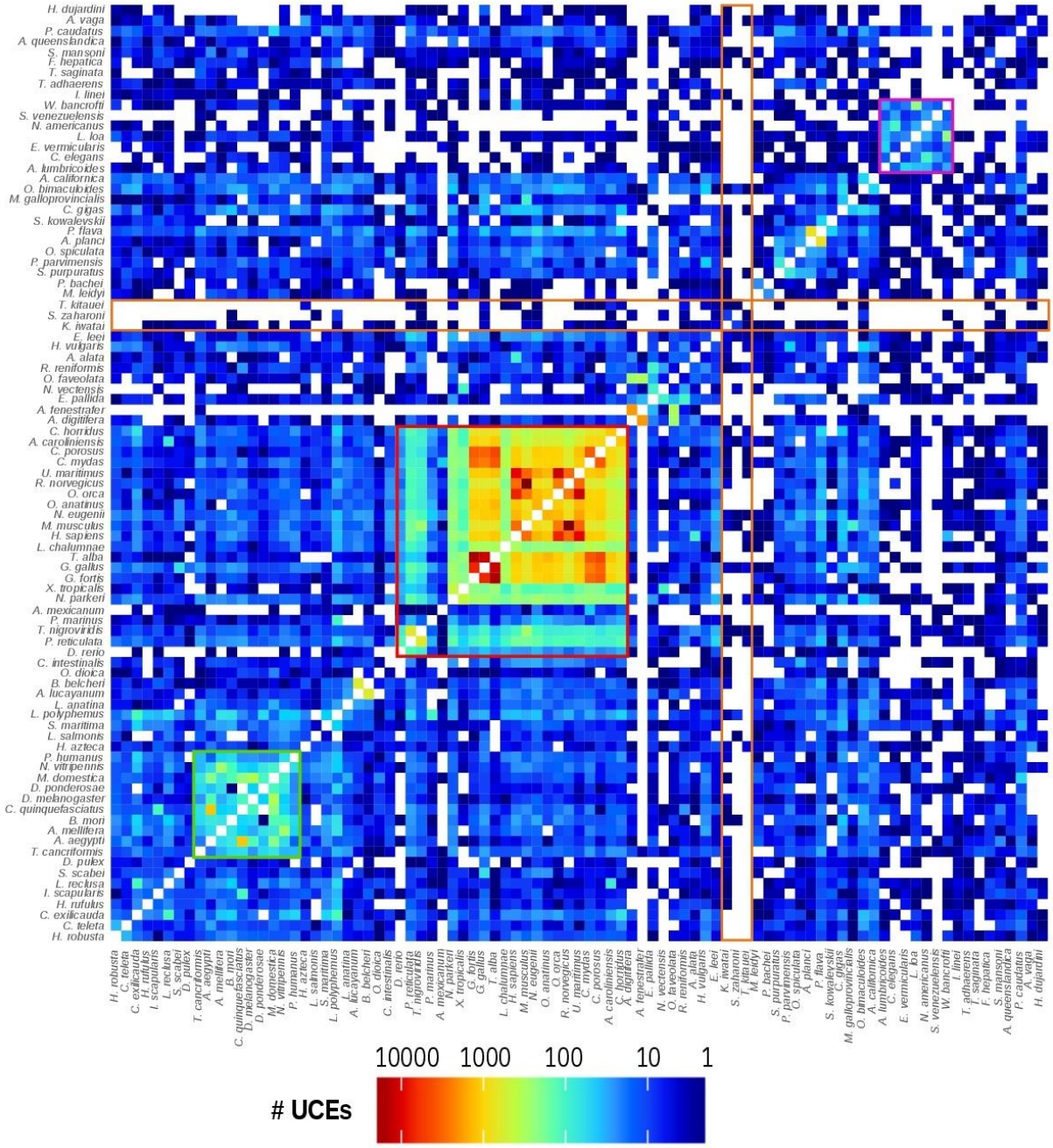
A relação entre proximidade filogenética e número de UCEs é também evidenciada na Figura 15, uma matriz representando todos os alinhamentos pareados, onde a temperatura da cor aumenta com número correspondente de UCEs, representados em escala logarítmica. Devido ao número muito superior de UCEs quando alinhado a *Homo sapiens*, *Pan troglotydes* foi excluído da matriz a fim de aumentar a separação de cores entre os demais alinhamentos.

De uma forma geral, nota-se uma certa limitação na utilidade das UCEs para filogenia em larga escala dos Metazoa, visto que muitos grupos compartilham poucas (células em azul escuro) ou nenhuma UCE (células brancas).

O maior número de UCEs é compartilhado pelos Vertebrata (destacados em vermelho), seguido pelos insetos (destacados em verde) e pelos Nematoda (destacados em lilás). Com exceção de *Caenorhabditis elegans* (Nematoda:Cromadorea:Rhabditida – HILLIER et al., 2007), os demais nematódeos sequenciados são endoparasitas e compartilham poucos ou nenhum UCEs com os demais metazoários. O mesmo é verdade para os Myxozoa (destacados em laranja), que compartilham poucos UCEs mesmo dentro da própria classe. Estes últimos organismos também são endoparasitas obrigatórios de peixes de água doce e possuem genomas relativamente pequenos (variando entre aproximadamente 27 e 173 Mpb).

Um típico genoma de invertebrado (por exemplo, *Nematostella vectensis*, *Caenorhabditis elegans*, *Apis mellifera*) possui aproximadamente 15 mil genes codificadores de proteínas (ELSIK et al., 2015; HILLIER et al., 2007; PUTNAM et al., 2007). Os vertebrados, por sua vez, possuem um número superior, com uma média de cerca de 25 mil genes codificadores. Em vários momentos da evolução dos metazoários, genes inteiros foram duplicados principalmente a partir de genes antigos presentes em invertebrados. Por exemplo, os invertebrados normalmente contêm uma cópia do cluster HOX que participa do desenvolvimento regulando aspectos da embriogênese, morfogênese e diferenciação celular em metazoários, enquanto os vertebrados possuem vários clusters HOX que diferem entre os difenrentes táxons (MYERS, 2008; CROW e WAGNER, 2006; MAZET e SHIMELD, 2002). Comparações do número de genes previstos entre vertebrados e invertebrados sugerem que os vertebrados possuem mais famílias gênicas que codificam para produtos que participam no desenvolvimento do que qualquer outro táxon animal (CROW e WAGNER, 2006; MAZET e SHIMELD, 2002).

FIGURA 15 – HEATMAP REPRESENTANDO, EM ESCALA LOGARÍTMICA, O NÚMERO DE UCES DE CÓPIA SIMPLES E ALTA COMPLEXIDADE ENCONTRADOS NOS GENOMAS DE METAZOÁRIOS ALINHADOS.



Legenda: O quadrado vermelho destaca os vertebrados, o verde destaca os insetos, o quadrado lilás destaca os nematódeos e os retângulos alaranjados destacam os mixozoários. FONTE: Laboratório de Evolução de Organismos Marinho – LEOM (2018).

Pertinentemente, muito dessas famílias gênicas importantes para desenvolvimento de vertebrados, como os genes HMX1, ARX, MEIS1 e SHH; estão imersos dentro de regiões ricas em elementos não codificantes altamente

conservadas (assim como as UCEs), corroborando os dados encontrados de superioridade numérica de UCEs compartilhadas nos genomas de vertebrados (DICKEL et al., 2018; ROYO et al., 2005; ANDERSON et al., 2014; CALLE-MUSTIENES et al., 2005).

Adicionalmente, estudos atuais têm ressaltado que a aquisição acidental de material genético de outras espécies por genomas de invertebrados é muito mais comum do que se inicialmente era cogitado. Os DNAs transferidos são de origem bacteriana ou eucarótica e, em ambos os casos, as espécies receptoras podem acabar utilizando os genes transferidos em seu benefício próprio. Por último, elementos virais endógenos (EVEs), também podem ser encontrados em genomas de invertebrados, entretanto os mecanismos envolvidos para essa integração são geralmente desconhecidos (DUNNING HOTOPP et al., 2007; FESCHOTTE e GILBERT, 2012; KATZOURAKIS e GIFFORD, 2010).

Os fatos citados tendem a aumentar a variabilidade genética nos genomas de invertebrados e diminuir a ocorrência de regiões altamente conservadas compartilhadas mesmo em grupos próximos, como no caso dos Hymenoptera *Apis mellifera* e *Nasonia vitripennis* que divergiram na mesma época há 50 e 60 milhões de anos (PETERS et al., 2017), respectivamente, e recuperam apenas 277 UCEs em nossos alinhamentos genômicos. Em contrapartida em vertebrados distante filogeneticamente como o réptil *Crocodylus porosus* e o mamífero *Pan troglodytes* (separados em 220 milhões de anos – KUMAR e HEDGES, 1998) nosso alinhamento recuperou o número extraordinário de 1128 UCEs.

Vale apontar também que o padrão de baixos números de UCEs compartilhadas entre metazoários e endoparasitas como os nematódeos e o myxozoários parece ser explicado pelo processo de transferência horizontal de genes nessas espécies parasitas que levaria a eventos de perda de conservação genômica (FIGURA 15). A transferência horizontal de genes (THG) é um fenômeno de aumento da variabilidade genética que ocorre frequentemente em procariontes, mas parece ser rara em espécies eucariontes. *Meloidogyne incognita*, um nematoide parasita de plantas, parece ter ganhado 7 genes por THG de bactérias que ocupam nichos semelhantes no solo e nas raízes (HAEGEMAN et al., 2011). Os genes adquiridos são importantes para o ciclo biológico parasítico do nematoide, como celulases para digerir material vegetal e moléculas sinalizadoras que induzem

mudanças morfológicas na planta, facilitando a invasão (HAEGEMAN et al., 2011; WEERASINGHE et al., 2005).

Outros grupos de nematoides parasitas vivem em simbiose com bactérias específicas transportadas pelos nematoides como a caso da Wolbachia que vive *Brugia malayi* e outros nematoides filariais (BLAXTER et al., 1998). A captação do conjunto de genes da Wolbachia parece ter sido adaptativa para os nematoides filariais, uma vez que Foster e colaboradores (2005) revelaram que tratar Wolbachia com antibióticos reduz o crescimento e a fecundidade dos nematoides (FOSTER et al., 2005). Fenômenos de transferência horizontal de genes também parecem ocorrer em mixozoários, embora suas causas ainda não estejam totalmente esclarecidas (SHOSTAK, 1993).

Finalmente, nossos métodos de filtragem e eliminação de sequências repetidas nas etapas metodológicas de eliminação de sequências de baixa complexidade, parecem ter reduzido o número de UCEs maiores recuperadas e que foram reportadas por Bejerano e colaboradores (2004). As 481 sequências reportadas e compartilhadas entre o genoma humano, de camundongos e de ratos parecem ser de sequências repetitivas, mesmo após etapa de filtragem realizada pelo software *Repeat Masker* (BEJERANO et al., 2004).

O mesmo parece ocorrer na comparação das UCEs obtidas por Ryu e colaboradores (2012) em seu estudo das relações evolutivas entre espécies de metazoários de diferentes distâncias evolutivas. Nesse estudo são reportadas 381 UCEs compartilhadas entre o mamífero *Homo sapiens* e a anêmona-do-mar *Nematostella vectensis* (nossas analises recuperaram 10 UCEs no mesmo alinhamento genômico). Ryu e colaboradores (2012) reportaram ainda, 5.303 UCEs compartilhadas entre os Cnidaria *Nematostella vectensis* e *Amphimedon queenslandica;* nossos alinhamentos não recuperaram nenhuma posição de destaque para esse alinhamento o que pode sugerir que o número extraordinariamente elevado de UCEs reportadas nesse estudo pode ser resultado de artefatos e falhas na etapa de eliminação de sequências de baixa complexidade (RYU et al., 2012).

5 CONSIDERAÇÕES FINAIS

Nesse estudo, apresentamos uma nova *pipeline* para identificação de sequências de DNA altamente conservadas em genomas de metazoários e nossos resultados indicam que o baixo compartilhamento entre filos muito distintos sugere limitada aplicação para filogenia em larga escala dos Metazoa, entretanto ressaltando sua potencial utilidade como marcadores para filogenias em menor escala.

REFERÊNCIAS

AHITUV, N.; ZHU, Y.; VISEL, A.; HOLT, A.; AFZAL, V.; PENNACCHIO, L. A.; RUBIN, E. M. Deletion of ultraconserved elements yields viable mice. **PLoS biology**, v. 5, n 9, 2007.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389-3402, 1997.

ANDERSON, E.; DEVENNEY, P. S.; HILL, R. E.; LETTICE, L. A. Mapping the Shh long-range regulatory domain. **Development,** v. 141, p. 3934–3943, 2014.

AMBROSE, C. D.; CREASE, T. J. Evolution of the nuclear ribosomal DNA intergenic spacer in four species of the Daphnia pulex complex. **BMC Genetics**, 2011.

APPEL, R. D.; FEYTMANS, E. **Bioinformatics**: a Swiss perspective. Singapore; World Scientific Pub., 2009.

ARENDT, D. et al. The origin and evolution of cell types. **Nature Reviews Genetics**, v. 17, p. 744–757, 2016.

BATZOGLOU, S.; PACHTER, L.; MESIROV, J. P.; BERGER, B.; LANDER, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. **Genome Research**, v. 10, n. 7, p. 950-8, 2000.

BEJERANO, G. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. **Genome Research**, v.15, n. 8, p. 1034–1050, 2005.

BEJERANO, G. Ultraconserved elements in the human genome. **Science**, v. 304, n. 5675, p. 1321–1325, 2004.

BENSON, D. A.; CLARK, K.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; SAYERS, E. W. GenBank. **Nucleic Acids Research**, v. 43, 2015.

BERMAN, H.; HENRICK, K.; NAKAMURA, H. Announcing the worldwide Protein Data Bank. **Nature Structural Biology**, v.10, n. 12, 2003.

BERNSTEIN, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. **Cell**, v. 125, p. 315-326, 2006.

BLAIMER, B. B.; LLOYD, M. W.; GUILLORY, W. X.; BRADY, S. G. Sequence Capture and Phylogenetic Utility of Genomic Ultraconserved Elements Obtained from Pinned Insect Specimens. **PLoS One**, v.11, n. 8, 2016.

BLAXTER, M. L. et al. A molecular evolutionary framework for the Phylum Nematoda. **Nature**, v. 392, n. 6671, p. 71-5, 1998.

BROWN, J. R. **Comparative genomics**: basic and applied research. Boca Raton, FL: CRC Press, 2008.

CALLE-MUSTIENES, E. T al. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. **Genome Research**, v. 15, n. 8, p. 1061-72, 2005.

CHAN, Y. C.; ROOS, C; INOUE-MURAYAMA, M.; INOUE, E. et al. Mitochondrial Genome Sequences Effectively Reveal the Phylogeny of Hylobates Gibbons**. PLoS ONE**, v. 5, n.12, 2010.

CHATRACHATCHAYA, C. et al. cDNA Sequence Analysis and Structural Phylogenetic Tree of Novel Myoglobin from Striped Snake-Head Fish (Ophicephalusstriatus). **Oriental Journal Of Chemistry**, v. 32, n. 1, p. 09-28, 2016.

CHEN, C.; HUANG, H.; WU, C. H. Protein Bioinformatics Databases and Resources. **Methods in Molecular Biology**, v. 1558, p. 3-39, 2017.

CHRISTEN, R. et al. An analysis of the origin of metazoans, using comparisons of partial sequences of the 28S RNA, reveals an early emergence of triploblasts. **Embo Journal**, v. 10, n. 3, p. 499-503, 1991.

CONESA, A. et al. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674-3676, 2005.

CRAWFORD, N. G.; FAIRCLOTH, B. C.; MCCORMACK, J. E. et al. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. **Biology Letters**, v. 8, n. 5, p. 783–786, 2012.

CROW, K. D.; WAGNER, G. What is the role of genome duplication in the Evolution of complexity and diversity? **Molecular Biology and Evolution**, v. 23, n. 5, p. 887-892, 2006.

DARWIN, C. **On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.** London, UK: Murray; 1859.

DELLAPORTA, S. L. et al. Mitochondrial genome of Trichoplax adhaerens supports Placozoa as the basal lower metazoan phylum. **Proceedings of the National Academy of Sciences**, v. 103, n. 23, p. 8751–8756, 2006.

DERTI, A.; ROTH, F. P.; CHURCH, G. M.; WU, C. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. **Nature Genetics**, v. 38, n. 10, p. 1216–1220, 2006.

DICKEL, D. E. et al. Ultraconserved Enhancers Are Required for Normal Development. **Cell**, v. 172, n. 3, p. 491–499, 2018.

DOWELL, K. **Molecular Phylogenetics**: An introduction to computational methods and tools for analyzing evolutionary relationships, 2008. Disponível em: <goo.gl/K4bHNX>. Acesso em: 16 de dezembro de 2017.

DUNNING HOTOPP, J. C. et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. **Science**, v. 317, p. 1753-6, 2007.

ECKERT, A. J.; HALL, B. D. Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): phylogenetic tests of fossil-based hypotheses. **Molecular Phylogenetics and Evolution**, v. 40, p. 166–182, 2006.

ELSIK, C. G., et al. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. **Nucleic Acids Research**, v. 4, n. 44, p. 793-800, 2016.

EMERSON, B. C.; CICCONARDI, F., FANCIULLI, P. P.; SHAW, P. J. Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. **Philosophical Transactions of the Royal Society of London B**, v. 366, p. 2391–2402, 2011.

FAIRCLOTH, B. C.; BRANSTETTER, M. G.; WHITE, N. D.; BRADY, S. G. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. **Molecular Ecology Resources**, v. 15, n. 3, p. 489–501, 2015.

FAIRCLOTH, B. C.; MCCORMACK, J. E.; CRAWFORD, N. G. et al. Ultraconserved elements anchor Thousands of genetic markers spanning multiple evolutionary Timescales. **Systematic Biology**, v. 61, n. 5, p. 717–726, 2012.

FAIRCLOTH, B. C.; SORENSON, L.; SANTINI, F.; ALFARO, M. E. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes ased upon Targeted Sequencing of Ultraconserved Elements (UCEs). **PLoS One**, v. 8, n. 6, 2013.

FESCHOTTE, C.; GILBERT, C. Endogenous viruses: insights into viral evolution and impact on host biology. **Nature Reviews Genetics**, v. 13, p. 283-96, 2012.

FIELD, K. G. et al. Molecular phylogeny of the animal kingdom. **Science**, v. 239, n. 4841, p. 748-53, 1988.

FOSTER, J. et al. The Wolbachia genome of Brugia malayi: endosymbiont evolution within a human pathogenic nematode. **PLOS Biology**, v. 3, n. 4, 2005.

FREEMAN, S.; HERRON, J. C. **Análise Evolutiva**. Editora, Artmed Editora, 2009.

FUTUYMA, D. J. **Evolutionary Biology**. 3ª ed. Sinauer Associates Inc., Sunderland, 1998.

GABALDÓN, T. Large-scale assignment of orthology: back to phylogenetics? **Genome Biology**, v. 9, n.10:235, 2008.

GIRIBET, G.; DUNN, C. W.; EDGECOMBE, G. D.; ROUSE, G. W. A modern look at the animal tree of life. **Zootaxa**, 2007.

GUSFIELD, D. **Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology**. Cambridge University Press, 2010.

GUZMAN, C.; CONACO, C. Comparative transcriptome analysis reveals insights into the streamlined genomes of haplosclerid demosponges. **Scientific Reports**, v. 6, 2016.

HAEGEMAN, A.; JONES, J. T.; DANCHIN, E. G. J. Horizontal Gene Transfer in Nematodes: A Catalyst for Plant Parasitism? **CURRENT REVIEW**, v. 24, n. 8, p. 879–887, 2011.

HARDY, C. R.; LINDER, H. P. Phylogeny and Historical Ecology of Rhodocoma *(*Restionaceae) from the Cape Floristic Region. Aliso: **A Journal of Systematic and Evolutionary Botany:** v. 23, n. 1, p. 213–226, 2007.

HARRIS, R. S. **Improved pairwise alignment of genomic DNA**. The Pennsylvania State University, 2007.

HEJNOL, A. et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. **Proceedings. Biological Sciences**, v. 276, n. 1677, p. 4261-70, 2009.

HENNIG, W.; DWIGHT, D.; ZANGERL, R.Phylogenetic Systematics. University of Illinois Press, 1999.

HERRERO, J. et al. Ensembl comparative genomics resources. **Database (Oxford)**, 2016.

HILLIER, L. W., et al. Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny. **PLoS Biology**, v. 5, p. 167-167, 2007.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860-921, 2001.

JENNINGS, W. B. **Phylogenomic Data Acquisition**: Principles and Practice. 1ª Edição. Boca Raton: CRC Press/Taylor & Francis, 2016.

JEUKENS, J. et al. Comparative genomics of a drug-resistant Pseudomonas aeruginosa panel and the challenges of antimicrobial resistance prediction from genomes. **FEMS Microbiology Letters**, v. 364, n. 18, 2017.

KATZOURAKIS, A., GIFFORD, R. J., Endogenous viral elements in animal genomes. **PLoS Genetics**, v. 6, p. 10, 2010.

KUMAR, S.; HEDGES, B. Amolecular timescale for vertebrate Evolution. **Nature**, v. 392, n. 6679, p. 917-20, 1998.

LAKE, J. A. Origin of the Metazoa**. Proceedings of the National Academy of Sciences of the United States of America**, v. 87, n. 2, p. 763-766, 1990.

LAURON, E. J.; LOISEAU, C.; BOWIE, R. C. K.; SPICER, G. S.; SMITH, T. B.; MELO, M.; SEHGAL, R. N. Coevolutionary patterns and diversification of avian malaria parasites in African sunbirds (Family Nectariniidae). **Parasitology**, v. 142, n. 5, p. 635-647, 2015.

LEACHÉ, A. D.; OAKS, J. R. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. **Annual Review of Ecology, Evolution, and Systematics**, v. 48, p. 69-84, 2017.

LIN, M. et al. Identification of polymorphisms in ultraconserved elements associated with clinical outcomes in locally advanced colorectal adenocarcinoma. **Cancer**, v. 118, n. 24, p. 6188–6198, 2012.

LIN, Q. et al. The seahorse genome and the evolution of its specialized morphology. **Nature**, v. 540, n. 7633, p. 395–399, 2016.

LINDBLAD-TOH, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. **Nature**, v. 478, p. 476–482, 2011.

MAKUNIN, I. V. et al. Comparison of Ultra-Conserved Elements in Drosophilids and Vertebrates. **PLoS One**., v. 8, n. 12, 2013.

MAZET, F.; SHIMELD, S. M. Gene duplication and divergence in the early evolution of vertebrates. Current opinions in genetics and development, v.12, p. 393-396, 2002.

MCCORMACK, J. E. et al. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. **Genome Research**, v. 22, n. 4, p. 746–754, 2012.

MEDINA, M.; COLLINS, A. G; SILBERMAN, J. D.; SOGIN, M. L. Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98: p. 9707-9712, 2001.

MIKKELSEN, T.S., et al. Initial sequence of the chimpanzee genome and comparison with the human genome. **Nature**, v. 437, p. 69–87, 2005.

MILLER, W.; MAKOVA, K. D.; NEKRUTENKO, A.; HARDISON, R. C. Comparative genomics. **Annual Review of Genomics and Human Genetics**, v. 5, p. 15–56, 2004.

MINDELL, D. P. The Tree of Life: Metaphor, Model, and Heuristic Device. **Systematic Biology**, v. 62, n. 3, p. 479–489, 2013.

MORA, C. et al. How Many Species Are There on Earth and in the Ocean? **PLoS Biology,** v. 9, n. 8, p. 1-8, 2011.

MOUNT, D. W. **Bioinformatics**: sequence and genome analysis. Cold Spring Harbor Laboratory Press, 2004.

MYERS, P. Z. et al. Hox Genes in Development: The Hox Code. **Nature Education**, v.1, n.1, p. 2, 2008.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. **Journal of Molecular Biology**, v. 48, p. 443-453, 1970.

NOWOSHILOW, S. et al. The Axolotl Genome and the Evolution of Key Tissue Formation Regulators. **Nature, v.** 554, n. 7690, p. 50-55, 2018.

ODDS, F. C. Genomics, molecular targets and the discovery of antifungal drugs. **Revista Iberoamericana de Micología**, v. 22, n. 4, p. 229-37, 2005.

ORLOV, Y. L.; POTAPOV V. N. Complexity: An Internet Resource for Analysis of DNA Sequence Complexity. **Nucleic Acids Research, v.** 1, n. 32, p. W628-33, 2004.

PACE, N. R.; SAPP, J.; GOLDENFELD, N. Phylogeny and beyond*:* scientific*,* historical, and conceptual significance of the first tree of life*.* **Proceedings of the National Academy of Sciences of the United States of America**, v. 104, n. 4, p. 1011–1018, 2012.

PALMER, W. J.; JIGGINS, F. M. Comparative Genomics Reveals the Origins and Diversity of Arthropod Immune Systems. **Molecular Biology and Evolution**, v. 32, n. 8, p. 2111–2129, 2015.

PATWARDHAN, A.; RAY, S.; ROY, A. Molecular Markers in Phylogenetic Studies – A Review. **Journal of Phylogenetics and Evolutionary Biology**, v. 2, n. 2, p. 1–9, 2014.

PEARSON, W. R. An Introduction to Sequence Similarity ("Homology") Searching. **Current Protocols in Bioinformatics**, 2014.

PENNACCHIO. L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. **Nature**, v. 444, p. 499–502, 2006.

PETERS, R. S. et al. Evolutionary History of the Hymenoptera. **Current Biology**, v. 27, n. 7, p. 1013-1018, 2017.

PEVSNER, J. **Bioinformatics and Functional Genomics**. 2ª Ed. Wiley-Blackwell, 2009.

PHILIPPE, H. et al. Phylogenomics revives traditional views on deep animal relationships. **Curr. Biol**, v. 19, p.706–712, 2009.

PHILIPPE, H.; LARTILLOT, N.; BRINKMANN, H. Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. **Molecular Biology and Evolution**, v. 22, p.1246–1253, 2005.

PHILIPS, T. K. Phylogeny of the Oniticellini and Onthophagini dung beetles (Scarabaeidae, Scarabaeinae) from morphological evidence. **Zookeys**, v. 579, p. 9–57, 2016.

POELCHAU, M. et al. The i5k orkspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. **Nucleic Acids Research**, v. 43, p. 28, p. 714–719, 2015.

PUTNAM, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. **Science**, v. 6, n. 317, p ;86-94, 2007.

R CORE TEAM. R: A Language and Environment for Statistical Computing, 2017. Disponível em: <https://www.R-project.org/>. Acesso em: 07 de janeiro de 2018.

RANNALA, B.; YANG, Z. Phylogenetic inference using whole genomes. **Annual Review of Genomics and Human Genetics**, v. 9, p.217-31, 2008.

RAY, P.; SHRINGARPURE, S.; KOLAR, M.; XING, E. P. CSMET: Comparative Genomic Motif. Detection via Multi-Resolution Phylogenetic Shadowing". **PLoS Computational Biology**, v. 4, n. 6, 2008.

REIS, M. et al. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. **Proceedings of the Royal Society of London B**, v. 279, p 3491–3500, 2012.

RODIONOV, D. A.; VITRESCHAK, A. G.; MIRONOV, A. A.; GELFAND, M. S. Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. **Nucleic Acids Research**, v. 32 n. 11, p. 3340-53, 2004.

ROYO, J. L. et al. Identification and analysis of conserved cis-regulatory regions of the MEIS1 gene. **PLoS One**, v. 7, n. 3, 2012.

RYU, T.; SERIDI, L.; RAVASI, T. The evolution of ultraconserved elements with different phylogenetic origins. **BMC Evolutionary Biology**, v. 12, n. 236, 2012.

SACCONE, C.; PESOLE, G. **Handbook of Comparative Genomics**: Principles and Methodology. Hoboken (New Jersey), 2003.

SADAVA, D. et al. **Vida**: a ciência da biologia. 8. ed. Porto Alegre: Artmed, 2009.

SANGES, R. et al. Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. **Nucleic Acids Research**, p. 1–19, 2013.

SAVOLAINEN, V; CHASE, M. W. A decade of progress in plant molecular phylogenetics. **TRENDS in Genetics**, v. 19, n. 12, p. 717-724, 2003.

SHOSTAK, S. A symbiogenetic theory for the origin of cnidocysts in Cnidaria. **BioSystems**, v. 29, p. 49–58, 1993.

SMITH, B. T.; HARVEY, M. G.; FAIRCLOTH, B. C.; GLENN, T. C.; BRUMFIELD, R. T. Target capture and massively parallel Sequencing of Ultraconserved elements for comparative studies at shallow evolutionary time scales. **Systematic Biology**, v. 63, n. 1, p. 83–95, 2013.

SMITH, T. F.; WATERMAN, M. S. Identification of Common Molecular subsequences. **Journal of Molecular Biology**, v. 147, p. 195-197, 1981.

SONG, N.; LIN, A.; ZHAO, X. Insight into higher-level phylogeny of Neuropterida: Evidence from secondary structures of mitochondrial rRNA genes and mitogenomic data. **PLoS ONE**, v. 3, n. 1, 2018.

STAYTON, C. T. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. **Journal of Theoretical Biology**, v. 252, p. 1–14, 2008.

STEPHEN, S.; PHEASAN, T. M.; MAKUNIN, I. V.; MATTICK, J. S. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. **Molecular Biology and Evolution**, v. 25, n. 2, p. 402-408, 2008.

STOESSER, G.; STERK, P.; TULI, M.; STOEHR, P.; CAMERON, G. The EMBL Nucleotide Sequence Database. **Nucleic Acids Research**, v. 25, n. 1, p. 7–14, 1997.

STREICHER, J. W.; MILLER, E. C.; GUERRERO, P. C.; CORREA, C.; ORTIZ, J. C.; CRAWFORD, A. J.; PIE, M. R.; WIENS, J. J. Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2214 loci. **Molecular Phylogenetics and Evolution**, v. 119, p. 128-143, 2018.

STREICHER, J. W.; WIENS, J. J. Phylogenomic analyses of more than 4000 nuclear loci resolve the origin of snakes among lizard families. **Biology Letters**, v. 13, n. 9, 2017.

SZÖLLOSI, G. J. et al. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. **PNAS**, v. 109, n. 43, p.17513–17518, 2012.

TELFORD, M. J.; BUDD, G. E.; PHILIPPE, H. Phylogenomic Insights into Animal Evolution**. Current Biology**, v. 25, n. 19, p. 876-887, 2015.

THE NATURE INSTITUTE. Evolution Evolving. Disponível em: < http://natureinstitute.org/pub/ic/ic21/darwin.htm>. Acesso em: 02 de janeiro de 2019.

TOUCHMAN, J. Comparative genomics. **Nature Education Knowledge**, v. 3, n. 10, p. 13, 2010.

VISEL, A. et al. A high-resolution enhancer atlas of the developing telencephalon. **Cell**, v.152, n. 4, p. 895-908, 2013.

WALDRON, L. et al. Comparative Meta-analysis of Prognostic Gene Signatures for Late-Stage Ovarian Cancer. **Journal of the National Cancer Institute**, v. 106 n. 5, 2014.

WEERASINGHE, R. R.; BIRD, D. M.; ALLEN, N. S. Root-knot nematodes and bacterial Nod factors elicit common signal transduction events in Lotus japonicus. **Proceedings of the National Academy of Sciences**, v. 102, n. 8, p. 3147-3152, 2005.

WEY-FABRIZIUS, A. R. et al. Transcriptome Data Reveal Syndermatan Relationships and Suggest the Evolution of Endoparasitism in Acanthocephala via an Epizoic Stage. **PLoS One**, v. 9, n. 2, 2014.
WHELAN, N. T. et al. Ctenophore relationships and their placement as the sister group to all other animals. **Nature Ecology & Evolution**, v. 1, p. 1737–1746, 2017.

WHITE, N. D. et al. Ultraconserved elements resolve the phylogeny of potoos (Aves: Nyctibiidae). **Journal of Avian Biology**, 2017.

WOOLFE, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. **PLoS Biology**, v. 3, n. 1, 2005.

YAHALOMI, D. et al. The Multipartite Mitochondrial Genome of Enteromyxum leei (Myxozoa): Eight Fast-Evolving Megacircles. **Molecular Biology and Evolution**, v. 34, p. 1551-1556, 2017.

YANG, R. et al. SNPs in ultraconserved elements and familial breast cancer risk. **Carcinogenesis**, v. 29, n. 2, p. 351-355, 2008.

YUTIN, N.; PUIGBÒ, P.; KOONIN, E. V.; WOLF, Y. I. Phylogenomics of prokaryotic ribosomal proteins. **PLoS One**, v. 7, n. 5, 2012.

ZHENG, W. X.; ZHANG, C. T. Ultraconserved elements between the genomes of the plants Arabidopsis thaliana and rice. **Journal of Biomolecular Structure and Dynamics**, v. 26, p. 1-8, 2008.

APÊNDICE A – ORGANISMOS REFERÊNCIA UTILIZADOS NO ESTUDO

Espécie	Classificação	Filo/Classe	Código Genbank
Acanthaster planci	Invertebrado	Echinodermata	GCA_001949145.1
Acropora digitifera	Invertebrado	Cnidaria	GCA_000222465.2
Adineta vaga	Invertebrado	Rotifera	GCA_000513175.1
Aedes aegypti	Invertebrado	Arthopoda	GCF_000004015.4
Alatina alata	Invertebrado	Cnidaria	Não consta
Amphimedon queenslandica	Invertebrado	Porifera	GCF_000090795.1
Amplexidiscus fenestrafer	Invertebrado	Cnidaria	Não consta
Apis melífera	Invertebrado	Arthopoda	GCF_000002195.4
Aplysia californica	Invertebrado	Mollusca	GCF_000002075.1
Ascaris lumbricoides	Invertebrado	Nematoda	GCA_000951055.1
Asymmetron lucayanum	Invertebrado	Cephalochordata	GCA_001663935.1
Bombyx mori	Invertebrado	Arthropoda	GCF_000151625.1
Branchiostoma belcheri	Invertebrado	Cephalochordata	GCF_001625305.1
Caenorhabditis elegans	Invertebrado	Nematoda	GCF_000002985.6
Capitella teleta	Invertebrado	Annelida	GCA_000328365.1
Centruroides exilicauda	Invertebrado	Arthopoda	GCA_000671375.1
Ciona intestinalis	Invertebrado	Tunicata	GCF_000224145.2
Crassostrea gigas	Invertebrado	Mollusca	GCF_000297895.1
Culex quinquefasciatus	Invertebrado	Arthropoda	GCF_000209185.1
Daphnia pulex	Invertebrado	Crustacea	GCA_000187875.1
Dendroctonus ponderosae	Invertebrado	Arthropoda	GCF_000355655.1
Drosophila melanogaster	Invertebrado	Arthropoda	GCF_000001215.4
Enterobius vermicularis	Invertebrado	Nematoda	GCA_000951215.1
Enteromyxum leei	Invertebrado	Cnidaria	GCA_001455295.1
Exaiptasia pallida	Invertebrado	Cnidaria	GCA_001417965.1
Fasciola hepatica	Invertebrado	Platyhelminthes	GCA_000947175.1
Helobdella robusta	Invertebrado	Annelida	GCF_000326865.1
Hyalella azteca	Invertebrado	Crustacea	GCF_000764305.1
Hypochthonius rufulus	Invertebrado	Arthropoda	GCA_000988845.1
Hypsibius dujardini	Invertebrado	Tardigrada	GCA_002082055.1
Intoshia linei	Invertebrado	Orthonectida	GCF_000208615.1
Ixodes scapularis	Invertebrado	Arthropoda	GCA_001642005.1
Kudoa iwatai	Invertebrado	Cnidaria	GCA_001407235.1
Lepeophtheirus salmonis	Invertebrado	Crustacea	GCA_001005205.1
Limulus polyphemus	Invertebrado	Arthropoda	GCF_000517525.1
Lingula anatina	Invertebrado	Brachiopoda	GCF_001039355.1
Loa loa	Invertebrado	Nematoda	GCF_000183805.2
Loxosceles reclusa	Invertebrado	Arthropoda	GCA_001188405.1
Mnemiopsis leidyi	Invertebrado	Ctenophora	GCA_000226015.1
Musca domestica	Invertebrado	Arthropoda	GCF_000371365.1
Mytillus galloprovincialis	Invertebrado	Mollusca	GCA_001676915.1
Nasonia vitripennis	Invertebrado	Arthopoda	GCF_000002325.3
Necator americanus	Invertebrado	Nematoda	GCF_000507365.1

Nematostella vectensis	Invertebrado	Cnidaria	GCA_000209225,2
Octopus bimaculoides	Invertebrado	Mollusca	GCF_001194135.1
Oikopleura dioica	Invertebrado	Tunicata	GCA_000209535.1
Ophiothrix spiculata	Invertebrado	Echinodermata	GCA_000969725.1
Orbicella faveolata	Invertebrado	Cnidaria	GCF_002042975.1
Parastichopus parvimensis	Invertebrado	Echinodermata	GCA_000934455.1
Pediculus humanus	Invertebrado	Arthopoda	GCF_000006295.1
Pleurobrachia bachei	Invertebrado	Ctenophora	GCA_000695325.1
Priapulus caudatus	Invertebrado	Priapulida	GCF_000485595.1
Ptychodera flava	Invertebrado	Hemichordata	GCA_900177555.1
Renilla reniformis	Invertebrado	Cnidaria	GCA_001465055.1
Saccoglossus kowalevskii	Invertebrado	Hemichordata	GCF_000003605.2
Sarcoptes scabei	Invertebrado	Arthropoda	GCA_000828355.1
Schistosoma mansoni	Invertebrado	Platyhelmintes	GCA_000237925.2
Sphaeromyxa zaharoni	Invertebrado	Cnidaria	GCA_001455285.1
Strigamia maritima	Invertebrado	Arthropoda	GCA_000239455.1
Strongylocentrotus purpuratus	Invertebrado	Echinodermata	GCF_000002235.4
Strongyloides venezuelensis	Invertebrado	Nematoda	GCA_001028725.1
Taenia saginata	Invertebrado	Platyhelminthes	GCA_001693075.2
Thelohanellus kitauei	Invertebrado	Cnidaria	GCA_000827895.1
Trichoplax adhaerens	Invertebrado	Placozoa	GCF_000150275.1
Triops cancriformis	Invertebrado	Crustacea	GCA_000981345.1
Wuchereria bancrofti	Invertebrado	Nematoda	GCA_001555675.1
Ambystoma mexicanum	Vertebrado	Amphibia	GCA_001455525.1
Anolis carolinensis	Vertebrado	Reptilia	GCF_000090745.1
Chelonia mydas	Vertebrado	Reptilia	GCF_000344595.1
Crocodylus porosus	Vertebrado	Reptilia	GCF_001723895.1
Crotalus horridus	Vertebrado	Reptilia	GCA_001625485.1
Danio rerio	Vertebrado	Peixe	GCF_000002035.5
Gallus gallus	Vertebrado	Ave	GCA_000002315.3
Geospiza fortis	Vertebrado	Ave	GCF_000277835.1
Homo sapiens	Vertebrado	Mammalia	GCF_000001405.36
Latimeria chalumnae	Vertebrado	Peixe	GCF_000225785.1
Mus musculus	Vertebrado	Mammalia	GCA_000001635.8
Nanorana parkeri	Vertebrado	Amphibia	GCF_000935625.1
Notamacropus eugenii	Vertebrado	Mammalia	GCA_000004035.1
Orcinus orca	Vertebrado	Mammalia	GCF_000331955.2
Ornithorhynchus anatinus	Vertebrado	Mammalia	GCF_000002275.2
Pan troglodytes	Vertebrado	Mammalia	GCF_000001515.7
Petromyzon marinus	Vertebrado	Peixe	GCA_000148955.1
Poecilia reticulata	Vertebrado	Peixe	GCF_000633615.1
Rattus norvegicus	Vertebrado	Mammalia	GCA_000001895.4
Tetraodon nigroviridis	Vertebrado	Peixe	GCA_000180735.1
Tyto alba	Vertebrado	Ave	GCF_000687205.1
Ursus maritimus	Vertebrado	Mammalia	GCF_000687225.1
Xenopus tropicalis	Vertebrado	Amphibia	GCF_000004195.3

APÊNDICE B – SCRIPT PARA O ALINHAMENTO DE GENOMAS PAR A PAR

```
#!/bin/bash
function trim_path {

        echo $1 | sed 's/\/$//'
}

function get_prefix {

    echo $1 | grep -o '[^\/]*\.' | sed 's/\.//'
}

function get_maf_filename {

    local IFS="_"
    echo "$*"
}

bash ~/scripts/pbs_header.sh -m 4GB -w 00:15:00 -j genome1_genome2_updp -e
msbarbeitos@gmail.com -a PAS1032

path=$( trim_path $1 )

files=$path/*.fa
echo "cp ${files[@]}"' $TMPDIR'

for (( i=0; i<$(( ${#files[@]} - 1 )); i++ ))
do
    prefix=$( get_prefix "${files[$i]}" )

    j=$(($i + 1))

    while [ $j -lt "${#files[@]}" ]
    do
        prefix2=$( get_prefix "${files[$j]}" )

        maf_filename=$( get_maf_filename $prefix $prefix2 "updp" ).maf

        bash ~/scripts/pbs_tail.sh -x "/nfs/08/osu9668/lastz/lastz
$prefix.fa[multiple] $prefix2.fa\
        --notransition --step=20 --nogapped --format=maf > $maf_filename"\
        -o "*.maf"

        (( j += 1 ))
    done
done

echo "rm *.fa"
```

APÊNDICE C – SCRIPT PARA PROCESSAMENTO DO ARQUIVO MAF

```
#!/usr/bin/perl -w
# lastz_maf.pl --- A perl script to parse MAF alignments
# Author: Marcos <msbarbeitos@gmail.com>
# Created: 22 Oct 2014
# Version: 0.01

use warnings;
use strict;

use Bio::AlignIO;
use Data::Dumper;

my $maf = Bio::AlignIO->new( -file => $ARGV[0], -format => 'maf' );

print "length,percent_identity\n";
my $uce_count;

while ( my $aln = $maf->next_aln() )
{
    $uce_count++;

    print $aln->length, ',';
    printf '%.2f', $aln->overall_percentage_identity();
    print "\n";
}

print "\nTotal de UCEs: $uce_count\n";

__END__

=head1 NAME

lastz_maf.pl - Experimental. So far, it outputs only the length and
percent identity of MAF alignments in CSV format.

=head1 SYNOPSIS

lastz_maf.pl input_file.maf > data.csv

=head1 DESCRIPTION

See  SYNOPSIS

=head1 AUTHOR

Marcos Barbeitos E<lt>msbarbeitos@gmail.com<gt>

=head1 COPYRIGHT AND LICENSE

Copyright (C) 2014 by Marcos Barbeitos

This program is free software; you can redistribute it and/or modify
it under the same terms as Perl itself, either Perl version 5.8.2 or,
at your option, any later version of Perl 5 you may have available.
```

APÊNDICE D – SCRIPT PARA FILTRAR AS UCEs

```
#!/usr/bin/perl -w
# trim_maf.pl --- A perl script to parse MAF alignments
# Author: Marcos <msbarbeitos@gmail.com>
# Created: 28 Nov 2017
# Version: 0.01
use warnings;
use strict;

use Carp;
use Bio::AlignIO;
use File::Basename;

use Data::Dumper;

my $INMAF = _get_inmaf();
my $INFH = _get_infh();
my $OUTFH = _get_outfh();

my %KEEP;
my $align_num = 0;

while ( my $aln = $INMAF->next_aln() )
{
    $align_num++;
    if ( $aln->length() >= 40 && $aln->overall_percentage_identity() ==
100 )
    {
        $KEEP{ $align_num } = 1;
    }
}

$align_num = 0;

while ( my $line = <$INFH> )
{
    chomp $line;

    if ( $line =~ '#' )
    {
        print $OUTFH "$line\n";
        next;
    }

    if ( $line =~ 'score' )
    {
        $align_num++;
    }

    if ( $KEEP{ $align_num } )
    {
        print $OUTFH "$line\n";
    }
}

close $INFH;
close $OUTFH;
```

```
sub _get_inmaf
{
    $ARGV[0] || confess "Script requires input MAF file as argument";

    return Bio::AlignIO->new( -file => $ARGV[0], -format => 'maf' );
}

sub _get_infh
{
    open( my $infh, '<:encoding(UTF-8)', $ARGV[0] );

    return $infh;
}

sub _get_outfh
{
    my $basename = basename( $ARGV[0], ( '.maf' ) );
    my $outfilename = $basename . '_filt.maf';

    open( my $outfh, '>:encoding(UTF-8)', $outfilename );

    return $outfh;
}

__END__

=head1 NAME

trim_maf.pl - Experimental. So far, it selects UCES with 100% similarity
and size over 40 bp from maf files.

=head1 SYNOPSIS

trim_maf.pl input_file.maf

=head1 DESCRIPTION

See  SYNOPSIS

=head1 AUTHOR

Marcos Barbeitos E<lt>msbarbeitos@gmail.com<gt>

=head1 COPYRIGHT AND LICENSE

Copyright (C) 2017 by Marcos Barbeitos

This program is free software; you can redistribute it and/or modify
it under the same terms as Perl itself, either Perl version 5.8.2 or,
at your option, any later version of Perl 5 you may have available.

=head1 BUGS

None reported... yet.

=cut
```

APÊNDICE E – SCRIPT PARA A REMOÇÃO DE UCEs DUPLICADAS

```
#!/usr/bin/perl
# eliminate_maf_dupes.pl --- A perl script to parse MAF alignments
# Author: Marcos <msbarbeitos@gmail.com>
# Created: 11 Jan 2016
# Version: 0.01

use warnings;
use strict;

use Bio::AlignIO;
use Carp;
use Carp::Assert;
use Data::Dumper;
use File::Basename;

use Check;

$ARGV[0] || _confess_usage( "No file path provided" );

my $inmaf = Bio::AlignIO->new( -file => $ARGV[0], -format => 'maf' );

my $uce_index = 0;
my %MAFSEQ;
my %MAFCOUNT;

while ( my $uce = $inmaf->next_aln() )
{
    $uce_index++;

    for my $seq ( $uce->each_seq )
    {
    my $seq_address = _get_seq_address( $seq );

    $MAFSEQ{ $uce_index }{ $seq_address } = 1;
        $MAFCOUNT{ $seq_address }++;
    }
}

my $duplicate_counts = 0;
for my $uce_index ( keys %MAFSEQ )
{
    my $dupe_flag = 0;

    for my $seq_address ( keys %{ $MAFSEQ{ $uce_index } } )
    {
    $dupe_flag++ if ( $MAFCOUNT{ $seq_address } > 1 );
    }

    if ( $dupe_flag )
    {
    $duplicate_counts++;
        delete $MAFSEQ{ $uce_index };
    }
}
print $ARGV[0] . ",$uce_index,$duplicate_counts\n";

exit unless ( keys %MAFSEQ );
```

```
open ( my $infile, '<:encoding(UTF-8)', $ARGV[0] )
    || die "Can't create " . $ARGV[0] . "\n";

open ( my $outfile, '>:encoding(UTF-8)', _get_outname() )
    || die "Can't create " . _get_outname() . "\n";

$uce_index = 0;
my $print_flag = 0;

while( my $line = <$infile> )
{
    chomp $line;

    if ( $line =~ /#/o )
    {
        print $outfile $line, "\n";
    next;
    }

    if ( $line =~ /score/o )
    {
        $uce_index++;
        $print_flag = $MAFSEQ{ $uce_index } ? 1 : 0;
    }

    print $outfile "$line\n" if $print_flag;
}
close $infile;
close $outfile;

sub _get_seq_address
{
    my $seq = shift;
    object_isa( $seq, 'Bio::LocatableSeq' ) if ( DEBUG );

    return join "_", map { $seq->$_ } qw( display_id start length );
}

sub _get_outname
{
    my $in_name = basename( $ARGV[0], ( '.maf' ) );

    return $in_name . '_no_dupes.maf';
}

sub _confess_usage
{
    my $message = shift;
    assert_scalar( $message ) if DEBUG;

    confess $message .  ' - try something like - perl eliminate_maf_dupes
/path/to/maf/file';
}
```

APÊNDICE F – SCRIPT DE CONVERSÃO PARA O FORMATO CSV

```
#! /bin/bash
for file in *.maf; do

    prefix=$( echo $file | sed -r 's/\.maf//' )
        maf_file=$prefix.maf
    csv_file=$prefix.csv

    if [ ! -f $csv_file ]
    then

        echo "Converting $maf_file"

        perl /home/alunos/uces/lastz/lastz_maf.pl $maf_file > $csv_file

        total=$( grep 'Total' $csv_file )

        echo "$maf_file $total"

        sed -i '$d' $csv_file
    fi
done
```

APÊNDICE G – SCRIPT DE IDENTIFICAÇÃO DA COMPLEXIDADE DAS UCES

```
#!/usr/bin/perl -w
# maf_complexity.pl --- A Perl script to compute DNA Trifonov complexity
for MAF files
# Author:  <taioba@marcos-OptiPlex-755>
# Created: 19 Mar 2018
# Version: 0.01
use warnings;
use strict;

use Bio::AlignIO;
use Bio::Tools::SeqWords;
use Carp;
use File::Basename;
use List::Util qw( min );

use Check;
use Data::Dumper;


$ARGV[0] || confess "Script requires MAF file path as argument";
-e $ARGV[0] || confess "MAF file path does not exist - $ARGV[0]";

my $MAF = Bio::AlignIO->new( -file => $ARGV[0], -format => 'maf' );
my $file_name = basename( $ARGV[0], ( '.maf' ) );

my %MAX_WORD_HASH;
my $uce_count = 0;

while ( my $aln = $MAF->next_aln() )
{
    $uce_count++;

    my $window_size = $aln->length;

    $MAX_WORD_HASH{ $window_size }
        or $MAX_WORD_HASH{ $window_size } = _max_word_count( $window_size
);

    my $obs_word_count = _obs_word_count( $aln, $window_size, $uce_count
);
    my $cl = sprintf( '%.4f', $obs_word_count / $MAX_WORD_HASH{
$window_size } );

    print join ",", ( $file_name, $uce_count, $window_size, $cl );
    print "\n";
}

sub _max_word_count
{
    my $window_size = shift;
    assert_number( $window_size );

    my $alphabet_size = 4;

    # Trivial words: length one (alphabet size) and equal to window size
    my $max_word_count = $alphabet_size + 1;
```

```
    for my $word_length ( 2 .. $window_size - 1 )
    {
        $max_word_count += min( $alphabet_size ** $word_length,
$window_size - $word_length + 1 );
    }

    return $max_word_count;
}

sub _obs_word_count
{
    my ( $aln, $window_size, $uce_count ) = @_;
    object_isa( $aln, 'Bio::SimpleAlign' );
    assert_number( $window_size );
    assert_number( $uce_count );

    my $word_obj = Bio::PrimarySeq->new
        ( -seq => $aln->consensus_string( $window_size ), -id =>
$uce_count, -alphabet => 'dna' );

    my $total_word_count = 1;
    for my $word_length ( 1 .. $window_size - 1 )
    {
        my $word_count_hash = Bio::Tools::SeqWords->count_overlap_words(
$word_obj, $word_length );
        $total_word_count += scalar( keys %{ $word_count_hash } );
    }

    return $total_word_count;
}
```

APÊNDICE H – SCRIPT PARA FILTRAR SEQUÊNCIAS DE VALOR CRÍTICO

```
#!/bin/perl
# maf_complexity.pl --- A Perl script to compute DNA Trifonov complexity
for MAF files
# Author:  <taioba@marcos-OptiPlex-755>
# Created: 29 Oct 2018
# Version: 0.01
use warnings;
use strict;

use Bio::AlignIO;
use Storable;
use Text::CSV;
use File::Basename;

use Check;

use Data::Dumper;

my %COMPLEXITY = _get_complexity();
my @FILES = <$ARGV[1]/*filt*.maf>;

for my $infile ( @FILES )
{
    my $outfile = _get_outfile_name( $infile );

    my $in_fh = _get_fh( '<', $infile );
    my $out_fh = _get_fh( '>', $outfile );

    my ( $genome1, $genome2 ) = _get_genomes( $infile );

    my $uce_count = 0;
    my $print_flag = 1;

    while ( my $line = <$in_fh> )
    {
      chomp $line;

      if ( $line =~ /^a/ )
      {
          $uce_count++;

          if ( $COMPLEXITY{ $genome1 }{ $genome2 }{ $uce_count } )
          {
                $print_flag = 1
            }
            else
            {
            $print_flag = 0
            }
      }

        print $out_fh "$line\n" if ( $print_flag );
    }
    close $in_fh;
    close $out_fh;
}
```

```
sub _get_complexity
{
    $ARGV[0] || _confess_usage( "Script requires path to CSV file as
argument" );

    return %{ retrieve( $ARGV[0] ) } if $ARGV[0] =~ '\.sto';

    my $fh = _get_fh( '<', $ARGV[0] );

    my $csv = Text::CSV->new( { binary => 1, sep_char => ",", auto_diag =>
1, diag_verbose => 1 } )
        or die "Cannot use CSV: " . Text::CSV->error_diag();

    my @cols = @{ $csv->getline( $fh ) };
    my $row = {};

    $csv->bind_columns( \@{$row}{@cols} );

    my %complexity;

    while ( $csv->getline( $fh ) )
    {
        $complexity{ $row->{ 'genome1' } }{ $row->{ 'genome2' } }{ $row->{
'uce' } } = 1
            unless ( $row->{ 'complex' } < _get_threshold_complexity() );
    }

    store \%complexity, './complexity.sto';

    return %complexity;
}

sub _get_outfile_name
{
    my $infile = shift;
    assert_scalar( $infile );

    $infile =~ s/filt/trim/g;

    return $infile;
}

sub _get_fh
{
    my ( $tag, $filename ) = @_;
    assert_scalars( $tag, $filename );

    open( my $fh, $tag . ':encoding(UTF-8)', $filename )
        || _confess_usage( "Can't find $filename" );

    return $fh;
}

sub _get_genomes
{
    my $filename = shift;
    assert_scalar( $filename );

    my @tokens = ( split /_/, basename( $filename ) );

    return ( $tokens[0], $tokens[1] );
```

```
}

sub _confess_usage
{
    my $message = shift;
    assert_scalar( $message );

    confess_usage( $message, 'perl trim_low_comp.pl
/path/to/csv/or/sto/file /path/to/maf/folder' );
}

sub _get_threshold_complexity
{
    return 0.95;
}

=head1 NAME

trim_low_comp.pl - compute Trifonov complexity from alignments in MAF
files

=head1 SYNOPSIS

perl maf_complexity.pl /path/to/csv/file /path/to/maf/folder

=head1 DESCRIPTION

Eliminates low complexity sequences from MAF alignments according to info
stored in a CSV file
computed by I<maf_complexity.pl>. The threshold complexity is currenlty
hardcoded as 0.95.
The program will dump the internal hash representation of the CSV file as
a *.sto file in
the working directory. This file may be used instead of the CSV file to
expedite the execution
of the script.

=head1 AUTHOR

E<lt>msbarbeitos@gmail.com<gt>

=head1 COPYRIGHT AND LICENSE

Copyright (C) 2018 by Marcos S. Barbeitos

This program is free software; you can redistribute it and/or modify
it under the same terms as Perl itself, either Perl version 5.8.2 or,
at your option, any later version of Perl 5 you may have available.
```