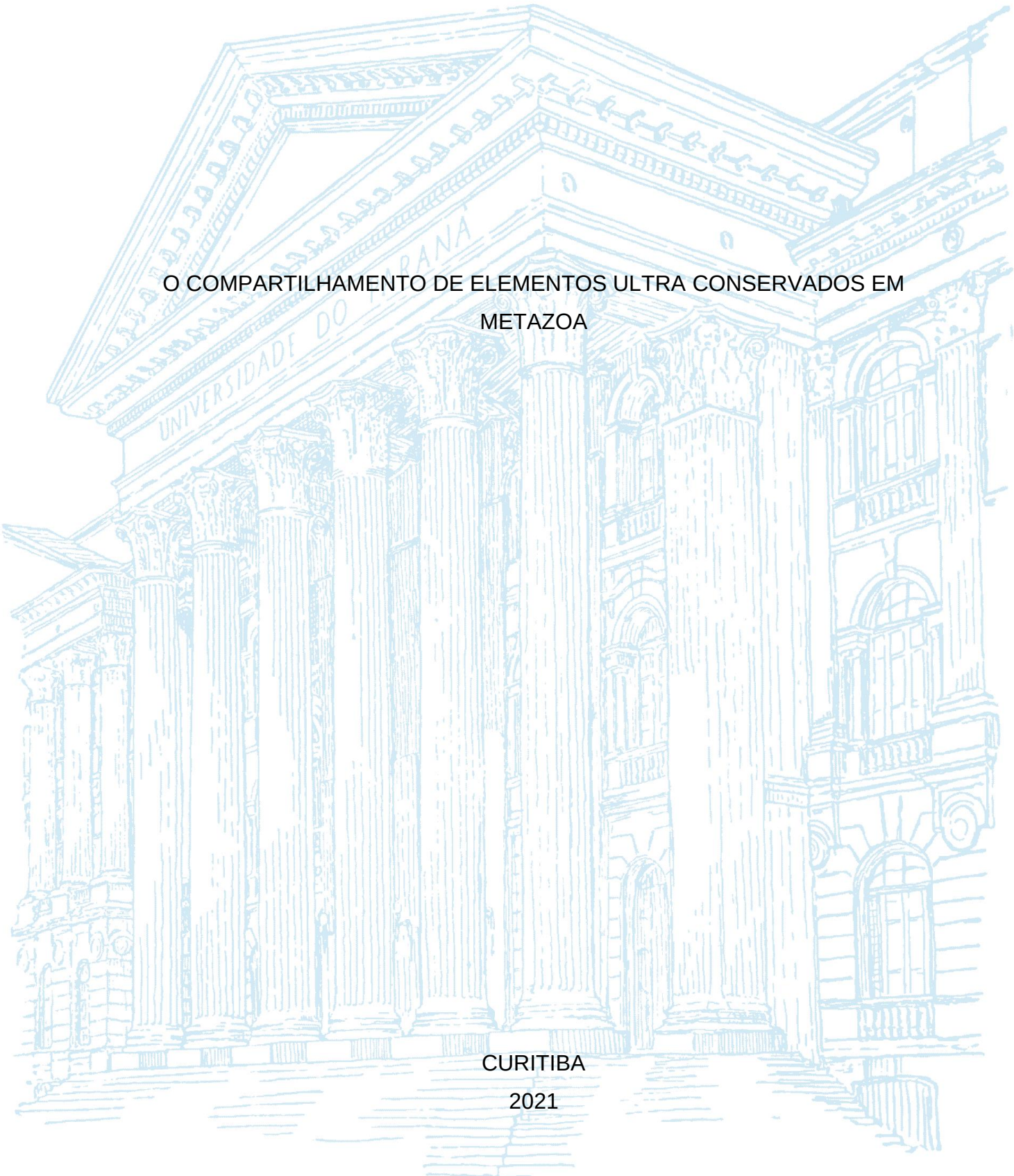
UNIVERSIDADE FEDERAL DO PARANÁ

FERNANDA CARNEIRO CANABARRO



O COMPARTILHAMENTO DE ELEMENTOS ULTRA CONSERVADOS EM METAZOA

CURITIBA

2021

FERNANDA CARNEIRO CANABARRO

O COMPARTILHAMENTO DE ELEMENTOS ULTRA CONSERVADOS EM METAZOA

Monografia apresentada como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas pela Universidade Federal do Paraná.

Orientador: Prof. Dr. Marcos Soares Barbeitos

Coorientadora: Ms. Carolina de Lima Adam

Título do Projeto: O Compartilhamento de Elementos Ultra Conservados em Metazoa

Curitiba

2021

AGRADECIMENTOS

O desenvolvimento e finalização deste trabalho de conclusão de curso contou com o apoio e ajuda de diversas pessoas, de forma direta ou indireta, dentre as quais eu agradeço imensamente:

À minha família que me proporcionou as oportunidades adequadas para chegar até o fim desta etapa, me motivando e apoiando em minhas escolhas e durante todo o percurso desta trajetória. Sou grata pelo acolhimento e força vinda de cada um de vocês, são a minha principal motivação para continuar.

Ao meu orientador Marcos Barbeitos que abriu as portas de seu laboratório para me receber, se dispondo a me orientar sobre um tema do qual eu desconhecia totalmente, se dedicando, entendendo minhas dificuldades e sendo o responsável por eu ter tido a oportunidade de adquirir muito conhecimento sobre Biologia e Bioinformática durante esse período de dois anos no LEOM.

À minha coorientadora Carolina Adam, agradeço pela paciência em me ensinar em todos os momentos, pelo interesse em transmitir conhecimento, pelas sugestões e correções sobre o trabalho, e que com sua dedicação e esforço se tornou um espelho para a minha trajetória como pesquisadora.

À Universidade Federal do Paraná que sempre possibilitou o pensamento crítico dos estudantes e que juntamente com o excelente trabalho e dedicação da Pró-Reitoria de Assuntos Estudantis (PRAE) ofereceu suporte assistencial para os alunos, garantindo minha permanência durante a graduação.

Ao LEOM, ao departamento de Zoologia e à Coordenação do Curso de Ciências Biológicas.

Aos meus amigos, inclusive aqueles em que conheci durante o curso, ainda que por vezes distantes, são parte importante deste ciclo que entre nós não se encerra agora. Afinal, o bar vai estar sempre lá!

RESUMO

O presente trabalho busca estimar a quantidade e o tamanho de Elementos Ultra Conservados (UCEs) em relação à distância entre os genomas de 95 representantes de Metazoa. Os UCEs foram identificados previamente e definidos como sequências com no mínimo 200 pares de bases (pb) e 100% de similaridade compartilhados entre os genomas de humanos, camundongos e ratos. A origem dessa conservação ainda é uma questão a ser solucionada, porém em função do alto grau de compartilhamento em determinadas espécies há autores que defendem a essencialidade dessas sequências para a viabilidade dos organismos. Além disso, os UCEs e suas regiões flanqueadoras têm sido úteis como âncoras para marcadores moleculares para inferência filogenética entre classes/ordens e também a nível de espécie; foram relacionados com o processamento de RNA e regulação da transcrição em genes ligados ao desenvolvimento. Sendo assim, esta pesquisa buscou estimar a quantidade de UCEs e sua relação com a distância entre os genomas. Para a detecção e contagem de UCEs foram utilizados *scripts* personalizados, e para o cálculo da disparidade entre genomas foram utilizados os programas BinDash e KSSD. Sendo assim, o trabalho foi capaz de refutar os resultados de Ryu et al. (2012), o qual encontrou uma expressiva quantidade de UCEs compartilhados entre táxons muito díspares na escala evolutiva. Portanto, os nossos dados sugerem que a distância entre os genomas é um dos fatores que influencia na conservação das sequências compartilhadas entre os diferentes genomas.

Palavras-chave: Elementos Ultra Conservados. Distância. Metazoa.

ABSTRACT

The present work seeks to estimate the quantity and size of Ultra Conserved Elements (UCEs) in relation to the distance between the genomes of 95 representatives of Metazoa. The UCEs were identified and defined as sequences with at least 200 base pairs (bp) and 100% shared similarity between the genomes of humans, mice and rats. The origin of this conservation is still an issue to be solved, however, due to the high degree of sharing in certain species, there are authors who defend the essentiality of these sequences for the viability of organisms. In addition, the UCEs and their flanking regions have been useful as anchors for molecular markers for phylogenetic inference between classes / orders and also at the species level; were related to RNA processing and regulation of transcription in genes linked to development. Therefore, this research sought to estimate the number of UCEs and their relationship with the distance between the genomes. For the detection and counting of UCEs, custom scripts were used, and for the calculation of the disparity between genomes, BinDash and KSSD programs were used. Thus, the work was able to refute the results of Ryu et al. (2012), who found a significant number of UCEs shared among taxa that are very different in the evolutionary scale. Therefore, our data shows that the distance between the genomes is one of the factors that influences the conservation of the shared sequences between the different genomes.

Keywords: Ultraconserved elements. Distance. Metazoa.

SUMÁRIO

1 INTRODUÇÃO

Os Elementos Ultra Conservados (da sigla em inglês UCEs) foram reconhecidos por Bejerano et al. (2004) como regiões que contém no mínimo 200 pares de bases (pb) e estão conservados com 100% de identidade, ou seja, sem inserções ou deleções nos genomas de humanos, camundongos e ratos. Essas sequências de DNA possuem regiões chamadas de “flanqueadoras” encontradas adjacentes às suas extremidades 5’ e 3’, caracterizadas por sua alta variabilidade. Faircloth et al. (2012) confirmaram esse resultado, reportando um aumento dessa variabilidade com o aumento da distância crescente do centro do UCE, além de confirmar o potencial de aplicação destas regiões em reconstruções filogenéticas que podem abranger diferentes escalas de tempo evolutivas. Desde então, houve um grande aumento na utilização de UCEs como âncoras para marcadores moleculares para estudos filogenéticos entre aves, répteis, mamíferos placentários e peixes (JARVIS et al. 2014; CRAWFORD et al. 2012; MCCORMACK et al. 2012 e BOSSERT et al. 2019). Esses elementos parecem ter um papel importante na regulação da expressão de genes ligados ao desenvolvimento, além de estarem associados com fatores de transcrição e com proteínas de ligação a RNA (BEJERANO et al. 2004).

Em um trabalho desenvolvido previamente pela autora deste estudo (CANABARRO et al. 2020) foram realizados alinhamentos pareados entre 95 representantes dos mais variados táxons de Metazoa (APÊNDICE 1) a fim de estimar a quantidade e o tamanho dos UCEs compartilhados entre eles. Os resultados demonstram que alguns grupos apresentam maior número de UCEs compartilhados entre si. Diante disso, o presente estudo busca estimar a taxa de decaimento do tamanho e da quantidade de UCEs em relação à disparidade entre os genomas, com o objetivo de identificar quais táxons possuem mais sequências compartilhadas, bem como inferir se a disparidade encontrada entre os genomas dos indivíduos se relaciona com o número e tamanho de UCEs recuperados.

Ryu et al. (2012) realizaram análises sobre a evolução dos elementos ultra conservados em grupos de metazoários com diferentes distâncias evolutivas. Esse estudo computou 43.707 sequências conservadas ≧30 pb entre *Nematostella vectensis* (Cnidário) e *Strongylocentrotus purpuratus* (Equinodermo), sugerindo que pode haver um alto grau de compartilhamento mesmo entre espécies com amplas distâncias evolutivas. Boffelli et al. (2004) relatam que o aumento da amostragem de

espécies em comparações de genomas torna menos provável que as sequências sejam conservadas por acaso, de forma que o presente estudo propõe ampliar a amostragem de metazoários a fim de verificar se é possível corroborar os resultados de Ryu et al. (2012) utilizando um número maior de táxons.

OBJETIVO GERAL

O objetivo geral deste trabalho consiste em identificar se a quantidade e o tamanho de UCEs se relaciona com a disparidade evolutiva entre os 95 genomas representantes de Metazoa e, com isso corroborar o padrão reportado por Ryu et al. (2012) de um alto grau de compartilhamento de UCEs entre táxons que divergiram há até 600 milhões de anos atrás (VAN ITEN et al. 2014).

OBJETIVOS ESPECÍFICOS

- Verificar quais táxons apresentam maior número de UCEs compartilhados entre seus representantes;
- Estimar a taxa de decaimento do número e tamanho dos UCEs em relação ao grau de divergência evolutiva entre os genomas de Metazoa.

2 A GENÔMICA E A BIOINFORMÁTICA

“A sequência de DNA representa um único formato no qual uma ampla gama de fenômenos biológicos pode ser projetada para a coleta de dados de alto desempenho” (SHENDURE & JI, 2008). Desde o seu desenvolvimento, o sequenciamento de Sanger et al. (1977) foi capaz de decifrar genes e até genomas inteiros. Após o seu uso em larga escala durante 40 anos tornou-se obsoleto com o surgimento das novas tecnologias, as quais inicialmente foram recebidas de forma reservada pela comunidade científica (SCHUSTER, 2007).

O sequenciamento de nova geração (da sigla em inglês NGS) permite o sequenciamento paralelo massivo de milhares de sequências de nucleotídeos, gerando dados genômicos em larga escala (GOODWIN et al. 2016). Ainda que a fidelidade do sequenciamento, bem como o comprimento de leitura e o custo de infraestrutura tenham sido questionados durante seu início, as vantagens da utilização desse novo tipo de tecnologia tornaram-se indispensáveis (SCHUSTER, 2007). Para Van Dijk (2014) há três componentes principais responsáveis pelo avanço das técnicas de NGS: não exige clonagem bacteriana de fragmentos de DNA, eliminando a possibilidade de haver viés de clonagem que possa impactar a representação do genoma alvo; têm a capacidade de produzir milhões de reações de sequenciamento em paralelo, ao invés de apenas 96 por vez como no sequenciamento de Sanger; e sua leitura é detectada diretamente sem a necessidade da realização de eletroforese. Outra vantagem da técnica é a possibilidade de gerar sequências a partir de poucos microgramas de DNA, além de produzir sequências com comprimentos de leitura mais curtos (35-250 pb), o que não era possível anteriormente (MARDIS, 2007).

Desta forma, o NGS trouxe à tona a possibilidade de reanimar o estudo de genomas já existentes ou complementar grandes estudos baseados em Sanger, e a redução dos custos possibilita estudar uma gama mais variada de organismos e ecossistemas (SCHUSTER, 2007).

A metagenômica permite sequenciar simultaneamente uma comunidade de organismos em uma mesma amostra, se tornando um campo fértil na era genômica. A atual eficácia e baixo custo da técnica possibilita sequenciar amostras de DNA ambiental (da sigla em inglês eDNA) e até mesmo de microbiotas inteiras presentes no corpo humano (HUGENHOLTZ & TYSON, 2008). Apesar das técnicas convencionais já serem capazes de cumprir esse objetivo, a utilização de técnicas de

nova geração, associadas a ferramentas de bioinformática e a grande disponibilidade de dados genômicos, atualmente permite responder quais organismos estão presentes em uma amostra utilizando menos recursos (MARDIS, 2007).

A disponibilidade e barateamento de experimentos utilizando sequenciamento de nova geração estão acarretando um aumento exponencial de dados gerados a partir do DNA, que são traduzidos para a linguagem computacional (STEPHENS et al. 2015), e com isso surge o desafio do correto processamento e análise desses novos elementos de pesquisa. Para Belcaid e Toonen (2015) existe a necessidade de um conhecimento básico da ciência da computação como forma de treinamento essencial para a formação de novos biólogos e "não apenas como uma ferramenta durante a inevitável integração da ciência da computação na biologia, mas também para promover interações produtivas na nova era da biologia multidisciplinar" (BELCAID & TOONEN, 2015).

Shendure e Ji (2008) ainda observam que a crescente variedade dos dados moleculares pode avaliar a variação genética, a expressão de RNA, interações entre proteína e DNA e estrutura do cromossomo. Segundo Wilhelm (2009) as técnicas de sequenciamento de alto rendimento permitem aplicações para a genômica pessoal com a análise detalhada dos trechos do genoma individual, análise precisa de trechos de RNA para expressão gênica, microbiologia e metagenômica. Além disso, as informações em larga escala do DNA têm sido essenciais para a resolução de relações de parentesco evolutivo entre os organismos (CRAWFORD et al. 2012) através, por exemplo, da utilização de elementos ultra conservados como âncoras para marcadores moleculares (STEPHEN et al. 2008).

2.2 OS ELEMENTOS ULTRA CONSERVADOS

Elementos Ultra Conservados foram identificados por Bejerano et al. (2004) e foram alvo de diversos estudos a partir de então. Caracterizados inicialmente como sequências de 200 pb com 100% de similaridade, sem inserções ou deleções, compartilhados entre humanos, camundongos e ratos. “Praticamente todos esses segmentos são conservados também em genomas de galináceos e cachorros, com uma média de 95 e 99% de identidade, respectivamente” (BEJERANO et al. 2004). Os UCEs também foram caracterizados em relação à sua disposição no cromossomo, classificados como exóticos ou parcialmente exóticos, não exóticos e possivelmente exóticos. As sequências não exóticas (intrônicas e intergênicas) estão mais próximas aos fatores de transcrição e genes de desenvolvimento, e os elementos exóticos e possivelmente exóticos dispõe-se de forma mais aleatória ao longo dos cromossomos (BEJERANO et al. 2004).

Bejerano et al. (2006) descobriram classes de regiões conservadas originadas de retrotransposons, sugerindo que estas possam ter passado por processos de exaptação, no qual caracteres que possam ter evoluído para outras funções, ou até mesmo sejam não funcionais, posteriormente tenham sido “alocados” para a função atual (GOULD & VRBA, 1982). Existem evidências de que quase metade de todo o DNA genômico de vertebrados é densamente carregado de elementos transponíveis, chamados de transposons (SIMONS et al. 2007). Os autores ainda pontuam que os transposons podem ter propriedades funcionais no genoma, porém identificaram também regiões livres de transposons (da sigla em inglês TFRs) em humanos, camundongos e gambás, consistindo em sua maioria DNA repetitivo, complexo e não codificador de proteínas, também associados a regiões reguladoras do desenvolvimento. Além disso, os UCEs foram frequentemente associados sobrepostos aos TFRs, de forma que também podem ser potenciadores de reguladores do desenvolvimento. Verificou-se também que os elementos ultra conservados não possuíam quaisquer transposons conhecidos, sugerindo que as TFRs são regiões que estão sob seleção negativa, onde mutações potencialmente deletérias ocorrem muito raramente (SIMONS et al. 2006).

Os UCEs são frequentemente encontrados em aglomerados, chamados de c*lusters* (BEJERANO et al. 2004) e têm sido chamados de âncoras para marcadores moleculares, visto que possuem regiões alta de variabilidade em suas extremidades

conhecidas como “flanqueadoras”. Faircloth et al. (2012) relatam que o aumento da variabilidade de pares de bases nessas regiões caracteriza um tipo de “fóssil molecular”, conservando sinal filogenético através de grandes escalas temporais. Os autores ainda afirmam que o aumento da variabilidade dessas regiões tende a aumentar com o aumento da distância do centro do UCE, sugerindo também que as regiões flanqueadoras sigam processos coalescentes. Este aspecto permite que os filogeneticistas adaptem o uso de UCEs escolhendo aqueles com taxas de evolução semelhantes ou selecionando uma subamostra de regiões de UCE cujas regiões flanqueadoras otimizam suas análises (GILBERT et al. 2015). Desde a sua identificação essas sequências foram utilizadas para resolver questões filogenéticas profundas entre aves modernas (JARVIS et al. 2014), répteis (CRAWFORD et al. 2012; WOOD et al. 2020), mamíferos placentários (MCCORMACK et al. 2012), coleópteros preservados em museus (BACA et al. 2017) e famílias de abelhas (BOSSERT et al. 2018).

Além disso, em uma análise comparativa Gilbert et al. (2015) observou que as regiões flanqueadoras e os núcleos dos UCEs possuem maior informação filogenética do que genes codificadores de proteínas. Este estudo sugere que os UCEs são úteis para resolver relações de parentesco entre percomorfos e são promissoras para serem utilizadas em outros vertebrados. Faircloth (2012) ressalta também que os UCEs não são úteis apenas para a resolução de relações filogenéticas profundas, mas também são ferramentas para solucionar relações superficiais a nível filogeográfico e individual. “Esse tipo de marcador genético e a estrutura analítica que descrevemos podem ser aplicados em toda a árvore da vida, potencialmente remodelando nossa compreensão da filogenia em muitos níveis taxonômicos” (FAIRCLOTH & CORMACK, 2012). Embora os UCEs tenham sido identificados em representantes de Porifera, Cnidaria (RYU et al. 2012) e Arthropoda (ZHANG et al. 2019; BOSSERT et al. 2018), Stephen et al. (2008) sugere que esses elementos tenham aparecido em larga escala durante a evolução dos amniotas ou tetrápodes, visto que o número de UCEs compartilhados entre humanos e peixes não é observado ao comparar os genomas de humanos e aves.

Desde a identificação desses elementos acredita-se que o seu compartilhamento e conservação durante milhões de anos ocorre devido à sua essencialidade para os organismos. A seleção negativa (purificadora) pode ser uma das principais razões para a conservação genética entre espécies (Siepel et al. 2005).

Segundo Simons et al. (2006), genomas maiores e de organismos com maior complexidade biológica geral possuem uma quantidade absoluta de DNA sob seleção negativa superior. Análises realizadas em leveduras e vertebrados mostram que a maior fração das sequências conservadas estão fora de éxons conhecidos ou suspeitos de serem genes codificadores de proteínas, sugerindo que esses elementos são importantes reguladores não codificadores em eucariotos complexos (Siepel et al. 2005).

Ahituv et al. (2007) realizaram um experimento em que deletaram 4 UCEs que se acreditava estarem associados a regiões importantes do genoma de camundongos e que, devido à essa deleção, eles não seriam capazes de gerar descendentes férteis. No entanto, todas as linhagens geraram indivíduos viáveis e férteis, indicando que essas sequências possam não ser tão essenciais quanto se pensava. Dickel et al. (2018) refizeram o experimento relatando que apesar dos ratos nascerem viáveis eles desenvolveram anormalidades neurológicas ou de crescimento, com alterações na quantidade de neurônios e defeitos estruturais no cérebro. Para os autores, os defeitos cognitivos resultantes da deleção dos UCEs colocariam os ratos em desvantagem no ambiente natural, portanto, as mutações iriam diminuir seu sucesso reprodutivo. Isto evidencia a importância da conservação desses elementos em largas escalas de tempo evolutivo. Os autores acreditam que esse trabalho pode servir como base para estudos futuros sobre doenças como Alzheimer e outros distúrbios neurológicos.

Devido ao seu alto grau de similaridade, elementos ultra conservados são regiões fáceis de serem alinhadas entre genomas extremamente divergentes. (SIEPEL et al. 2005). Usados para resolver questões de cunho filogenético e associação com deficiências neurais, os UCEs têm mostrado relação com a regulação da genes ligados ao desenvolvimento, fatores de transcrição e proteínas de ligação a RNA (BEJERANO et al. 2004), em conservação de espécies (OOSTERHOUT, 2020) e também são associados à regulação do processo de splicing (STEPHEN et al. 2008). No entanto, o alto grau de compartilhamento e conservação dos UCEs em alguns grupos ainda gera questões a serem solucionadas.

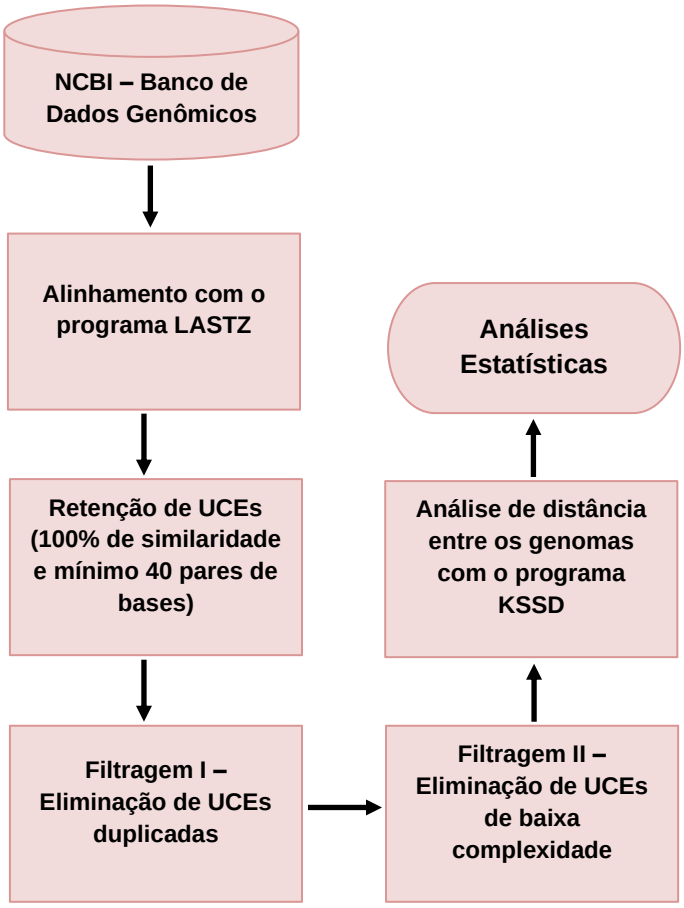
3 METODOLOGIA

3.1 OBTENÇÃO DOS UCEs

Os dados utilizados para a realização deste estudo são provenientes dos resultados da *pipeline* desenvolvida em um trabalho prévio por meio da linha de comando do sistema operacional Linux, o qual teve como objetivo a detecção dos elementos ultra conservados a partir dos genomas de 95 representantes de Metazoa, incluindo espécies de invertebrados representantes de Porifera, Cnidaria, Arthropoda, Echinodermata, Mollusca, Annelida, entre outros, e também de representantes das principais classes de vertebrados como Pisces, Amphibia, Reptilia, Aves e Mammalia (APÊNDICE 1).

A partir de genomas coletados no banco de dados National Center for Biotechnology Information (NCBI) foram realizados alinhamentos par a par através do programa LASTZ (HARRIS, 2007), totalizando 4.467 alinhamentos. Os alinhamentos foram então submetidos à sucessivas filtragens, sendo elas: I - retenção de UCEs com 100% de similaridade e no mínimo 40 pares de bases; II - eliminação de sequências duplicadas; III - eliminação de sequências de baixa complexidade e, por fim, os alinhamentos processados foram submetidos às análises dos dados. Todas essas etapas foram realizadas por meio de uma *pipeline* escrita nas linguagens de programação *Bash, PERL* e R com o objetivo de isolar os UCEs da forma mais acurada possível e assim analisar os possíveis padrões encontrados entre os genomas.

O fluxograma abaixo representa as etapas realizadas durante o alinhamento dos genomas.



FONTE: A autora (2021).

3.2 ANÁLISE DE DISTÂNCIA COM O SOFTWARE BINDASH

As análises realizadas com o software BinDash (ZHAO, 2019) tiveram como objetivo estimar a distância entre os 95 genomas, podendo assim inferir relações de proximidade filogenética entre diferentes táxons a partir do cálculo da dissimilaridade entre eles.

A análise de divergência entre os genomas funciona por meio de um método sem alinhamento que utiliza a técnica MinHash descrita por Broder (1997), a qual “transforma os genomas em conjuntos de *k-mers*” (ZHAO, 2019). Conforme o autor, esse conjunto de *k-mers* é uma representação adequada dos genomas, embora em ordens de magnitude menores, exigindo menor poder computacional em análises comparativas. A distância entre os genomas é estimada através do índice de Jaccard, uma medida da porcentagem da sobreposição entre dois conjuntos, resultando em valores de 0 (menor similaridade) até 1 (maior similaridade) (ARNABOLDI, 2015).

Segundo Ondov et al. (2016), a técnica MinHash é capaz de computar a similaridade entre as duas sequências com erro limitado, e o valor do índice de Jaccard calculado é similar entre as representações compactadas (chamados *sketches*) e os genomas originais.

Os dados dos UCEs foram convertidos do formato FASTA para a extensão *sketch* (responsável por transformar os genomas em frações representativas), através do *script* “bindash_run.sh” (APÊNDICE 2). Posteriormente, foram computados os valores de distância entre cada genoma a partir de seus s*ketches* através do *script* “dist_bindash_run.sh” (APÊNDICE 3). O resultado é um arquivo de saída em formato CSV que foi submetido ao *script* “merge_uce_dist.pl” (APÊNDICE 4) para a combinação dos valores de distâncias entre pares de genomas e as contagens de UCEs entre os mesmos genomas, gerando um novo arquivo CSV para a análise da correlação utilizando o R.

3.3 ANÁLISE DE DISTÂNCIA COM O SOFTWARE K-MER SUBSTRING SPACE DECOMPOSITION

O programa KSSD (YI; LIN & JIN, 2019) também foi utilizado para estimar a disparidade entre os genomas. Ele é mais robusto quando a comparação se dá entre genomas com alto grau de divergência, sendo mais adequado à diversidade de filos e classes abordados neste trabalho, bem como a diversidade de tamanhos de genomas.

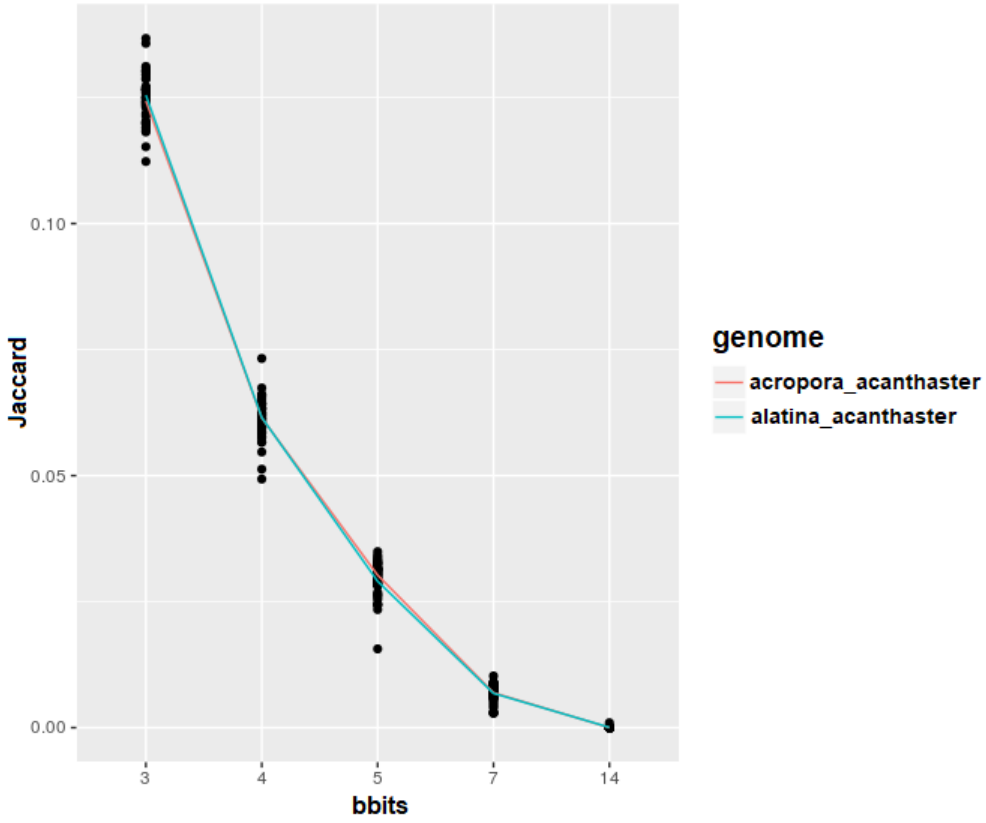
Essa ferramenta propõe uma nova técnica de sequência *sketch* para a redução da dimensionalidade dos genomas. O KSSD utiliza a amostragem aleatória de *k-mers*, a qual é usada para medir dois tipos de métricas: a semelhança e a contenção. A semelhança identifica a distância, ou seja, a similaridade de duas sequências de tamanhos aproximados, enquanto a contenção captura a distância/similaridade entre duas sequências de tamanhos muito diferentes. Métodos tradicionais baseados em minhash, como o BinDash, transformam em *sketches* apenas as sequências de interesse (*k-mer* menor), mantendo o banco de dados de comparação (*k-mer* maior) em sua configuração original, o que ocupa muita memória e exige alto poder computacional. Já a técnica KSSD transforma tanto as sequências de interesse quanto o banco de dados de comparação em um conjunto de *sketches*, que são então comparados entre si, reduzindo significativamente o poder computacional exigido.

Como todo o conjunto de dados é reduzido à *sketches*, a técnica possibilita a comparação de genomas de tamanhos muito distintos ao calcular sua semelhança e contenção diretamente do banco de dados de representação reduzida. Conforme Yi; Lin & Jin, 2019, o programa superou outras técnicas, como Mash e BinDash, consumindo menos espaço e menos tempo, e sendo capaz de analisar rapidamente o coeficiente de contenção e estimar com acurácia a distância (através do índice de Jaccard) entre conjuntos de genomas de tamanhos distintos.

4 RESULTADOS

A escolha inicial de parâmetros utilizados no software BinDash resultou em valores de Jaccard iguais a 0 para todos os genomas. A fim de escolher o parâmetro mais acurado para resultados consistentes foram feitas comparações de diferentes combinações de parâmetros no programa R utilizando apenas dois alinhamentos de referência: *Acropora digitifera* (Filo Cnidaria) com *Acanthaster planci* (Filo Echinodermata) e *Alatina alata* (Filo Cnidaria) com *Acanthaster planci*. Na Figura 1 foi comparado o parâmetro "--bbits=3" (reduz o tamanho do *sketch* e o tempo de execução para compará-los) associando-o com outras combinações de parâmetros, como por exemplo: "--kmerlen=21" (define o tamanho da subsequência de tamanho *k*) e "--sketchsize64=64" (define tamanho do *sketch* para representar os genomas). É possível observar que o parâmetro que mais influenciou os valores de Jaccard próximos a 1, ou seja, com maior similaridade, foi o "--bbits=3", o qual foi utilizado nas análises subsequentes.

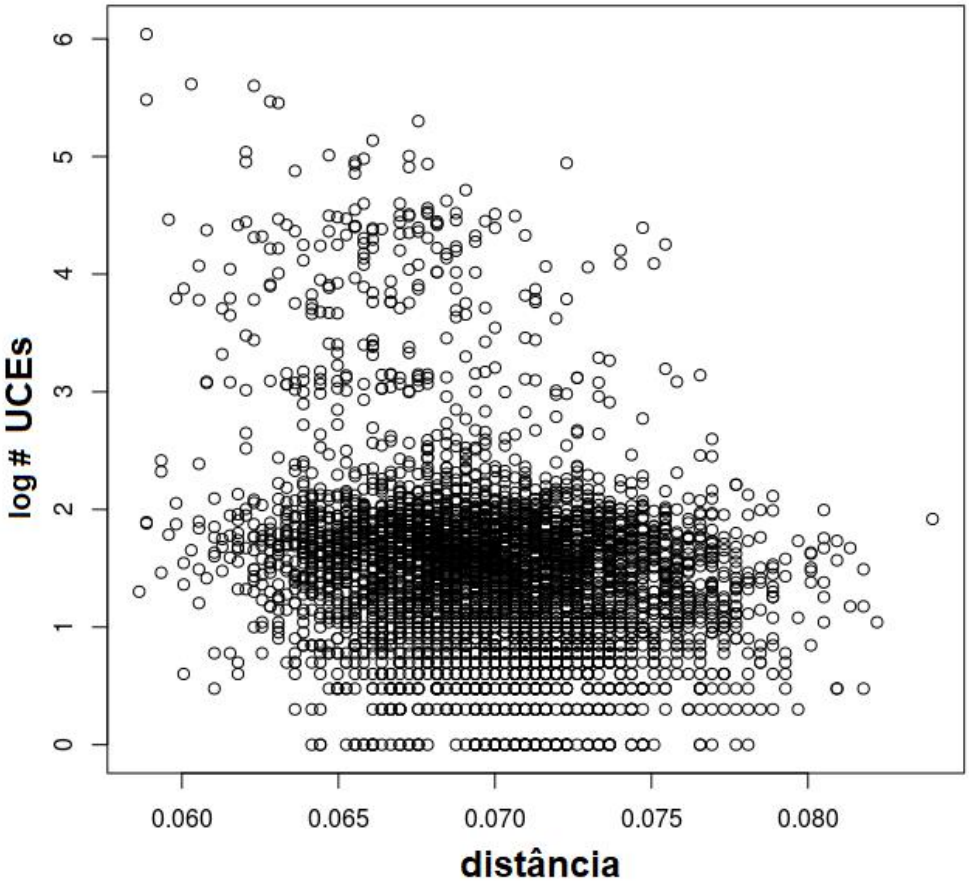
FIGURA 1 - COMPARAÇÃO DOS VALORES DO PARÂMETRO 'bbits'.



FONTE: Laboratório de Evolução de Organismos Marinhos - LEOM (2020).

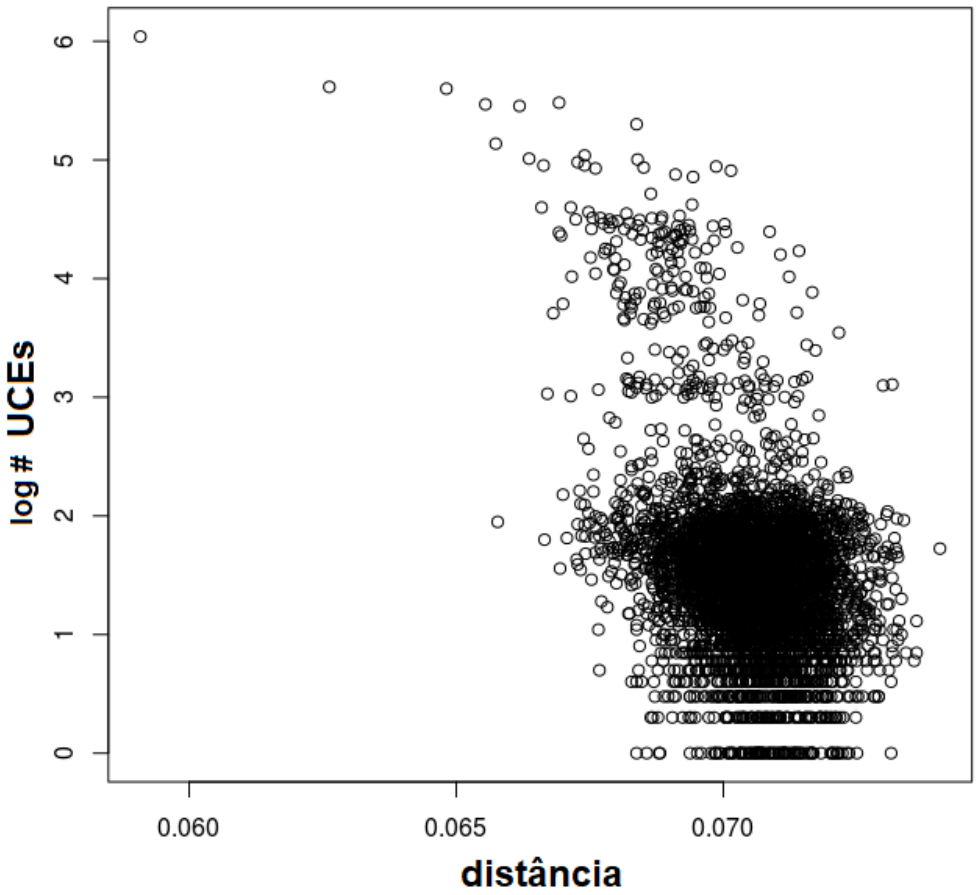
Foi realizada a análise de distância com os valores de *sketch* extremos como “--sketchsize64=32” e “--sketchsize64=256” para verificar qual teve mais influência nos resultados. Nas Figuras 2 e 3 nota-se que não houve muita discrepância entre os resultados mesmo com as alterações extremas do tamanho do sketch, e os valores de distância foram iguais ou muito semelhantes.

FIGURA 2 – GRÁFICO DE DISPERSÃO ENTRE DISTÂNCIA E O LOG DO NÚMERO UCES UTILIZANDO OS PARÂMETROS “--sketchsize64=32”, “--bbits=3” e “--kmerlen=18”.



FONTE: Laboratório de Evolução de Organismos Marinhos - LEOM (2020).

FIGURA 3 – GRÁFICO DE DISPERSÃO ENTRE DISTÂNCIA E O LOG DO NÚMERO DE UCES UTILIZANDO OS PARÂMETROS “--sketchsize64=256”, “--bbits=8” e “--kmerlen=7”.



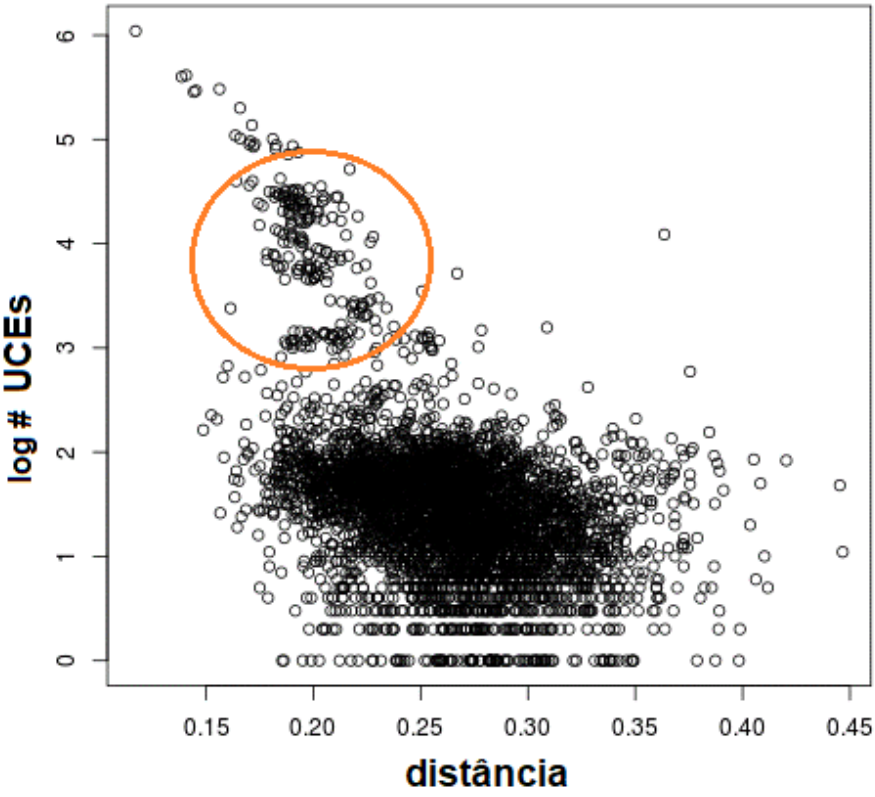
FONTE: Laboratório de Evolução de Organismos Marinhos - LEOM (2020).

O programa BinDash se mostrou favorável apenas para a comparação entre genomas compactos, sem regiões repetitivas, íntrons e transposons (XiaoFei Zhao, comunicação pessoal), o que não é o caso do presente estudo, visto que foram selecionados organismos com tamanhos de genomas distintos que variam entre 26 milhões de pares de bases (Mb) em *Kudoa iwatai* até entorno de 30 bilhões de pb em *Ambystoma mexicanum*. Isso se refletiu na falha em adquirir resultados consistentes utilizando uma ampla combinação de parâmetros. Em função da falta de precisão nos resultados foi decidido descartar o uso do programa nas análises subsequentes. Sendo assim, optou-se pela utilização do programa KSSD, o qual é capaz de analisar a divergência entre genomas de diferentes tamanhos.

A Figura 4 é resultado da utilização do programa KSSD, o círculo indicado em laranja identifica que o número de UCEs compartilhados entre organismos tende a decair em linhagens mais distantes, confirmando o observado por Stephen et al. (2008), que demonstrou que o número de UCEs, bem como seu grau de similaridade,

sofre redução em linhagens divergentes. Observaram, por exemplo, 94,7% de similaridade entre um grupo de sequências ortólogas de eutérios entre o genoma humano e de gambá, que divergiram há cerca de 180 milhões de anos, e uma redução para 74,1% de similaridade entre o genoma humano e do peixe baiacú, que divergiram há cerca de 450 milhões de anos. Harmston (2013) também observa que em distâncias evolutivas maiores, o número de elementos conservados não codificadores diminui rapidamente, sugerindo que a proximidade evolutiva está positivamente relacionada com a quantidade de UCEs compartilhadas entre as espécies. Levando em consideração que o número de elementos é significativamente maior entre organismos mais próximos evolutivamente, esta relação corrobora a ideia de Faircloth (2012) de que UCEs e suas regiões flanqueadoras podem ser mais informativas para solucionar relações a nível de classes/ordens e a nível de espécies.

FIGURA 4 - GRÁFICO DE DISPERSÃO DO LOG DO NÚMERO DE UCES EM COMPARAÇÃO À DIVERGÊNCIA ENTRE OS GENOMAS UTILIZANDO O PROGRAMA KSSD. O CÍRCULO LARANJA REPRESENTA O ALTO NÚMERO DE UCES COMPARTILHADOS ENTRE GENOMAS MAIS PRÓXIMOS.



FONTE: Laboratório de Evolução de Organismos Marinhos - LEOM (2020).

O *heatmap* apresentado na Figura 6 é uma representação gráfica da matriz que exibe a relação entre a quantidade de UCEs compartilhadas e a proximidade evolutiva

dos indivíduos por meio do aumento da temperatura da cor, indo do azul (menor número de UCEs compartilhados) ao vermelho (maior número de UCEs compartilhados). Os pontos em branco indicam nenhum UCE compartilhado entre o par de genomas alinhados. O *heatmap* evidencia a ocorrência de um maior número de UCEs compartilhados entre os genomas de vertebrados (*cluster* de temperatura mais quente observada no centro da figura, representada pelo círculo amarelo) e de artrópodes (*cluster* representada pelo círculo verde).

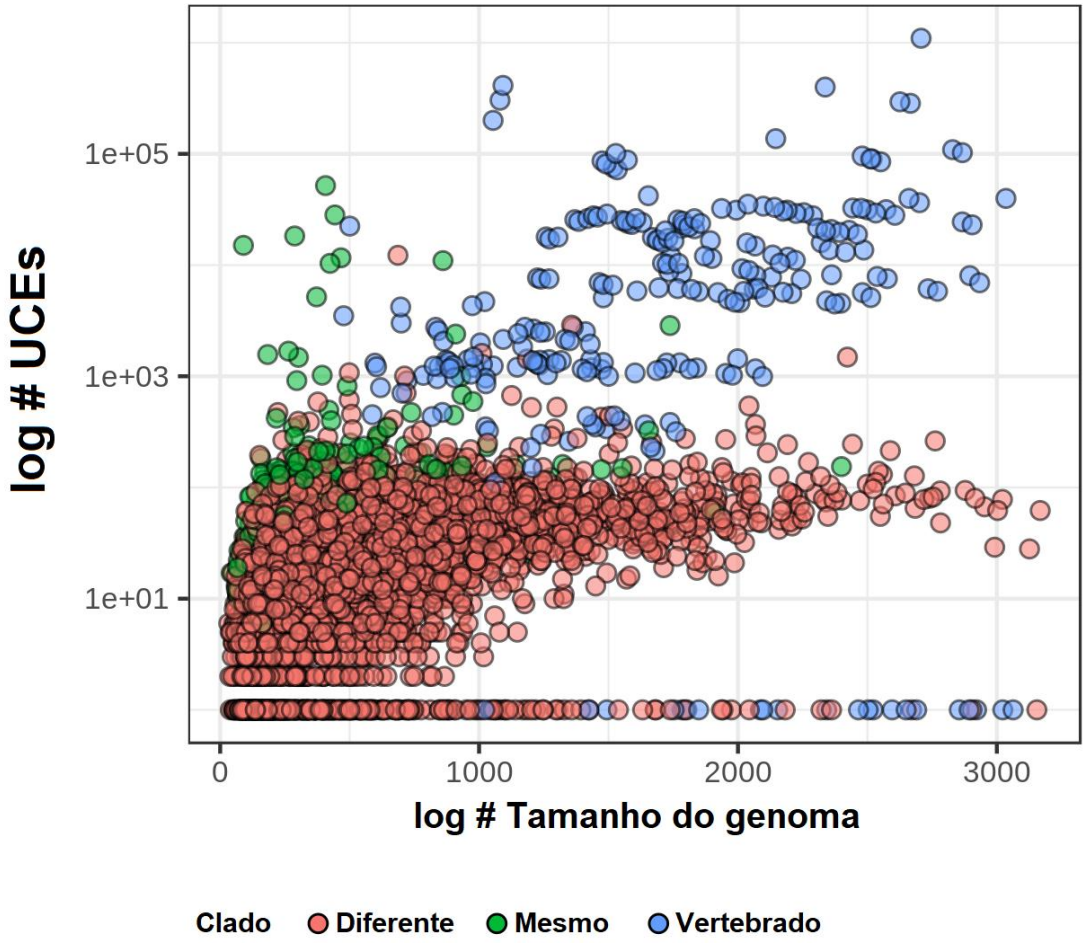
Em nossos dados, por exemplo, foram encontrados 293.958 UCEs com tamanhos de até 883 pb entre *Ursus maritimus* (urso polar) e *Homo Sapiens,* cujo ancestral comum mais recente data do início do Cretáceo Superior, cerca de 94 milhões de anos atrás (SPRINGER et al. 2003). Já *Homo sapiens* e *Geospiza fortis* (ave passeriforme), pertencentes a classes distintas e separados por >300 milhões de anos (SHEDLOCK & EDWARDS, 2009), compartilham 24.804 UCEs. Entre os genomas de *Danio rerio* (peixe ósseo) e *Homo sapiens,* espécies que divergiram há cerca de 430 milhões de anos (BLAIR & HEDGES, 2005) e que também pertencem a diferentes classes foram encontrados 1.165 UCEs com até 168 pb. Em contrapartida, nossos resultados recuperaram 343 UCEs entre os dípteros *Culex quinquefasciatus* e *Drosophila melanogaster*, que divergiram há cerca de 260 milhões de anos (CHEN et al. 2015), e 1.681 UCEs entre *Apis mellifera* (abelha) e *Nasonia vitripenis* (abelha), ambos pertencentes à classe Hymenoptera, separados por cerca de 124 milhões de anos (XIAO et al. 2013). Bejerano et al. (2004) também encontrou sequências conservadas compartilhadas entre representantes de artrópodes, porém em menores quantidades do que o reportado em vertebrados, o que foi corroborado por Faircloth (2012).

Em um estudo sobre elementos ultra conservados em Diptera, Glazov et al. (2005) também relatam que há menores e menos quantidades de UCEs entre insetos do que entre vertebrados, mesmo comparando grupos que divergiram em tempos evolutivos semelhantes. Os autores sugerem que a menor quantidade de UCEs encontrados pode estar relacionada às maiores taxas de substituição ocorridas nos genomas de Diptera, somado a seu maior número de gerações quando comparado a espécies de vertebrados. Nesse sentido, o relógio molecular pode influenciar nas taxas de substituição e conservação de sequências.

Na Figura 5 é possível observar que espécies pertencentes ao mesmo clado (i.e. Filo/Subfilo, ver listagem no Apêndice 1), apresentam maior compartilhamento de

UCEs, sendo que os vertebrados apresentam números ainda maiores, independente do tamanho do menor genoma utilizado na comparação (eixo X), sugerindo que o alto grau de compartilhamento nesse grupo não se deve ao tamanho do genoma de seus representantes.

FIGURA 5 - DISTRIBUIÇÃO DO LOG DO NÚMERO UCES EM FUNÇÃO DOS DIFERENTES CLADOS DE METAZOA.



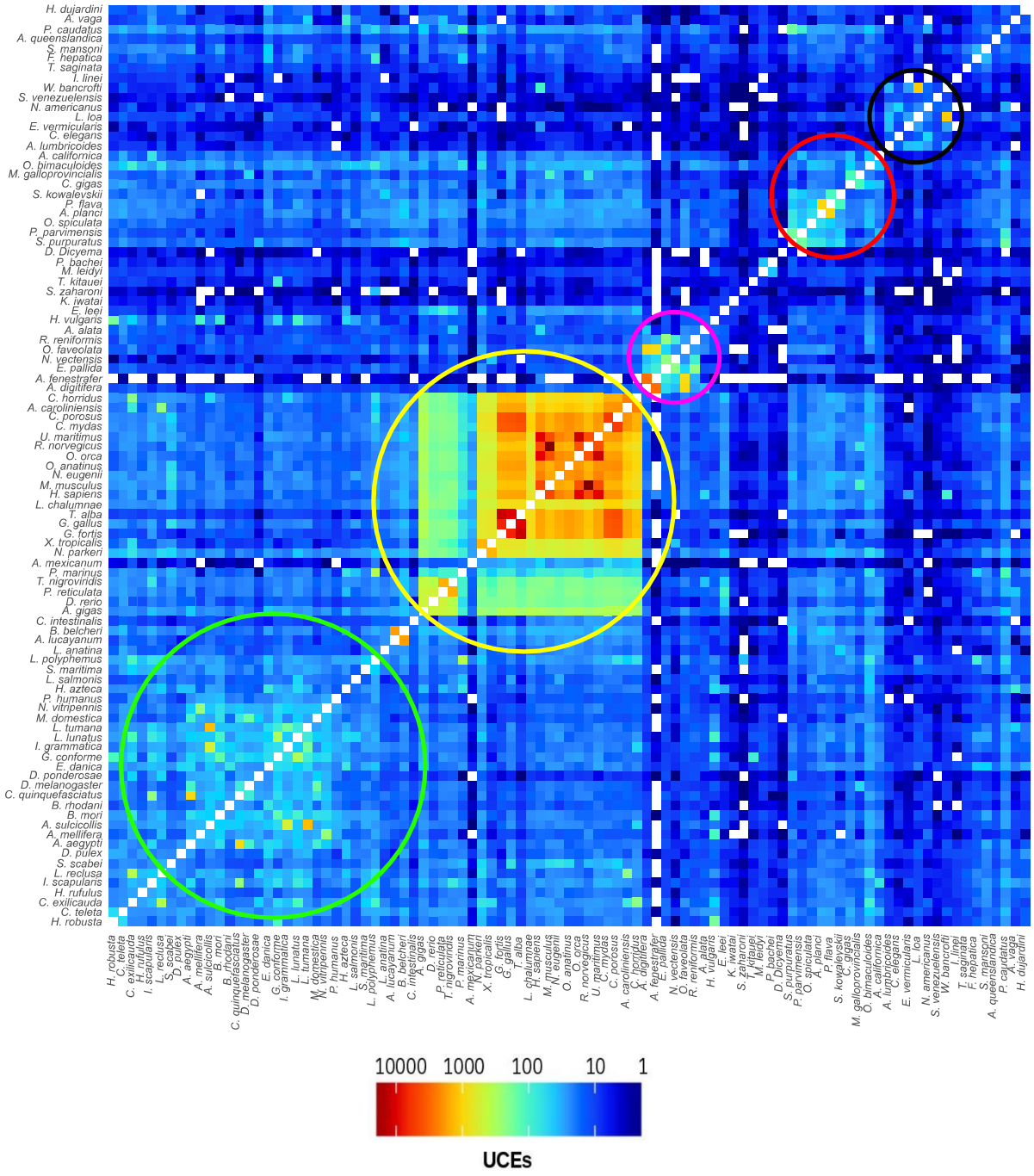
FONTE: Laboratório de Evolução de Organismos Marinhos - LEOM (2020).

O aumento em larga escala no número de UCEs ocorrido durante a especiação dos tetrápodes (STEPHEN et al. 2008) também pode explicar a maior incidência de sequências conservadas compartilhadas entre os genomas de Vertebrados, especialmente em grupos mais próximos (Figura 6). Os autores sugerem que esse possível aumento de UCEs compartilhadas possa estar acompanhado de uma desaceleração do relógio molecular, de forma que a taxa de substituição dos UCEs tenha mudado no curso da evolução, ou até mesmo que esses elementos tenham passado a ter novas funções em amniotas e vertebrados. Os seus resultados “sugerem que o aumento da complexidade do sistema regulatório dentro e ao redor

dos principais genes do desenvolvimento coincide com a diversificação da linhagem dos tetrápodes” (STEPHEN et al. 2008).

Outros grupos também apresentam números elevados de UCEs compartilhados, embora em menor escala do que entre os vertebrados. Ainda na Figura 6, os círculos rosa e preto representam cnidários e vermes nematódeos, respectivamente. O círculo vermelho representa moluscos e equinodermos. Estes resultados demonstram que a utilização de UCEs em estudos filogenéticos é provavelmente mais eficaz entre organismos evolutivamente mais próximos, e que sua aplicação para estudos filogenéticos de larga escala, como a nível de filos, pode ser dificultada dado o baixo número de elementos compartilhados entre organismos altamente divergentes.

FIGURA 6 – REPRESENTAÇÃO DA PROXIMIDADE FILOGENÉTICA EM FUNÇÃO DA QUANTIDADE DE UCES EM CADA TÁXON.



FONTE: Laboratório de Evolução de Organismos Marinhos - LEOM (2020).

Entretanto, é notável o compartilhamento de centenas de UCEs entre clados distantes, evidenciada pela abundância de "*pixels*" com tonalidades entre o azul e o verde (ver legenda da Fig. 6) através de todo o diagrama. Isto indica que o potencial de utilização destes marcadores para resolver relações entre filos de metazoários merece consideração em estudos futuros.

Para Boffelli et al. (2004) a comparação de genomas muito divergentes como

de humanos e peixes, é uma estratégia interessante para identificar sequências que provavelmente apresentam atividade funcional significativa. Essa abordagem baseia-se na ideia de que a conservação de sequências em genomas de organismos separados por distâncias evolutivas significativas indica seleção, e possivelmente localizada em uma região codificadora que desempenha uma função (SIMONS et al. 2005).

Glazov et al. (2005) acredita que a conservação de sequências de DNA entre espécies distantes na escala evolutiva pode indicar a conservação de estruturas proteicas ou até mesmo de importantes elementos reguladores de ação cis, os quais possuem relação com o desenvolvimento e fatores de transcrição (WITTKOPP & KALAY, 2012). Dickel (2018) sugere que a deleção de UCEs pode levar à diminuição da aptidão ao longo do tempo evolutivo, portanto, o aumento de UCEs durante o surgimento e diversificação dos Vertebrados, reportado por Stephen et al. (2008), pode estar relacionado com um aumento no sucesso reprodutivo desses organismos no ambiente.

Ao reproduzir os alinhamentos de Ryu et al. (2012), não foi possível recuperar o resultado obtido pelos autores (Tabela 1). Para o alinhamento entre *Nematostella vectensis* e *Drosophila melanogaster,* Ryu et al. (2012) encontraram 5440 ≧30 pb e 256 ≧50 pb, enquanto que nossos dados alcançaram valores de 281 pb sem as etapas de filtragem e 9 pb após o processamento. A maior UCE encontrada pelos autores foi de 796 pb entre *Nematostella vectensis* e *Homo sapiens*, enquanto que nossa *pipeline* recuperou sua maior sequência de 1244 pb para este mesmo alinhamento. Os autores reportaram 43.707 ≧30 pb e 5525 ≧50 pb no alinhamento entre *Nematostella vectensis* (cnidário) e *Strongylocentrotus purpuratus* (equinodermo), porém nossa *pipeline* recuperou 281 UCEs, as quais foram reduzidas a 9 após as etapas de filtragem. Para o alinhamento entre *Nematostella vectensis* (cnidário) e *Homo sapiens* Ryu et al. (2012) obteve 10 UCEs ≧50 pb e 381 UCEs ≧30 pb, porém nossas análises resultaram em 210 UCEs, reduzidas a 17 após o processamento. Ryu et al. (2012) obteve 967 UCEs ≧50 pb e 19 UCEs ≧30 pb no alinhamento entre *Strongylocentrotus purpuratus* (equinodermo) e *Homo sapiens*, e nossos resultados recuperaram 2.465 UCEs >40 pb e 57 UCEs após as etapas de filtragem.

TABELA 1 – COMPARAÇÃO DE UCES ENCONTRADAS NAS ANÁLISES DE RYU ET AL. (2012) E DE CANABARRO ET AL. (2020).

	UCEs – Ryu et al. (2012)	**UCEs – Canabarro et al. (2020)**
Cnidário X Equinodermo	47.707 / 5525	281 / 9
Cnidário X Humano	381 / 10	210 / 17
Equinodermo X Humano	917 / 19	2.465 / 57
Cnidário X Artrópode	5.440 / 256	62 / 12

FONTE: RYU ET AL. (2012); CANABARRO ET AL. (2020).
NOTA: * Os valores para cada alinhamento representam UCEs ≧ 30 e ≧ 50, respectivamente para Ryu et al. (2012). Para Canabarro et al. (2020) os valores referem-se às UCEs antes e depois do processamento.

Para o alinhamento entre *Nematostella vectensis* e *Drosophila melanogaster,* Ryu et al. (2012) encontraram 5440 ≧30 pb e 256 ≧50 pb, enquanto que nossos dados alcançaram valores de 281 pb sem as etapas de filtragem e 9 pb após o processamento. A maior UCE encontrada pelos autores foi de 796 pb entre *Nematostella vectensis* e *Homo sapiens*, enquanto que nossa *pipeline* recuperou sua maior sequência de 1244 pb para este mesmo alinhamento.

Os resultados de Ryu et al. (2012) não são corroborados pelas fortes evidências apresentadas de que o número de UCEs compartilhadas diminui proporcionalmente à distância evolutiva entre os organismos. Os autores obtiveram um número de UCEs compartilhadas entre o genoma de equinodermos e humanos, ambos deuterostômios, muito menor do que o número de UCEs compartilhadas entre os genomas de equinodermos e cnidários, cujo ancestral comum mais recente data de mais de 600 milhões de anos (VAN ITEN et al. 2014). Nossos resultados corroboram a hipótese de que o número de UCEs decai com a distância evolutiva, e sugerem que a metodologia empregada por Ryu et al. (2012) não reflete o número real de elementos ultra conservados compartilhados entre os organismos analisados. É provável que não foram utilizados parâmetros apropriados para a recuperação das UCEs, ou que não foram empregadas importantes etapas de filtragem e curagem dos dados.

5 CONSIDERAÇÕES FINAIS

Diante da modernização de técnicas de sequenciamento de nova geração, da ampliação na disponibilidade de dados genômicos em plataformas públicas e do rápido desenvolvimento de ferramentas de bioinformática para a análise rápida e acurada do enorme volume de dados disponíveis foi facilitado à comunidade acadêmica a descoberta de regiões genômicas antes desconhecidas. Em 2004 Bejerano e seus colaboradores identificaram UCEs que se mostraram demasiadamente úteis como marcadores moleculares potencialmente capazes de solucionar questões filogenéticas em diversas escalas evolutivas.

Utilizando elementos ultra conservados presentes nos genomas de 95 representantes de Metazoa, nossos resultados corroboram conclusões obtidas em estudos prévios de que o número de UCEs compartilhadas entre organismos é inversamente proporcional à sua distância evolutiva, e que os genomas de vertebrados apresentaram o maior número de UCEs compartilhadas, um resultado já observado em tetrápodes por Stephen et al. (2008). Com isso, refutamos os resultados obtidos no trabalho de Ryu et al. (2012), o qual recuperou um alto número de UCEs compartilhados entre indivíduos de amplas distâncias evolutivas, como cnidários e equinodermos. Sendo assim, os alinhamentos resultantes desse trabalho evidenciam que a aplicação de UCEs em análises filogenéticas de larga escala pode não ser tão eficaz, dado o baixo número de UCEs compartilhadas entre organismos altamente divergentes.

Por meio da combinação de *scripts* personalizados e *softwares* apropriados conseguimos atingir os objetivos desta pesquisa, visto que foi possível estimar se a quantidade e o tamanho de UCEs se relaciona com a disparidade entre os genomas dos 95 representantes de metazoários. A identificação dos UCEs contidos em uma ampla amostragem permite a realização de novos estudos promissores, bem como a detecção da posição destes UCEs nos genomas, verificando se estão próximos a íntrons; a transposons, retrotransposons; próximos a regiões reguladoras, intergênicas; sobrepostos a éxons ou fazem parte de regiões sob seleção purificadora. A localização dos UCEs nos cromossomos irá ajudar a compreender sobre a funcionalidade dessas sequências, bem como sobre a sua conservação ao longo tempo evolutivo.

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. 2007.

ANSORGE, W. J. **Next-generation DNA sequencing techniques**. New Biotechnology. 2009.

ARNABOLDI, V., CAMPANA M., DELMASTRO F., E. PAGANI. **Tag-based Recommender System for Context-Aware Content Dissemination in Opportunistic Networks**. Istituto di Informatica e Telematica. 2015.

BACA, S. M., ALEXANDER, A., GUSTAFSON, G. T. e SHORT, E. Z. **Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of 'Hydradephega'**.

BELCAID, M., TOONEN, R. J. **Desmystifying computer science for molecular ecologists**. The Hawai'i Institute of Marine Biology. 2015.

BEJERANO, G., PHEASANT, M. MAKUNIN, I. STEPHEN S, Kent W, Mattick J, Haussler D. **Ultraconserved elements in the human genome**, Science, 2004, vol. 304 p. 1321.

BLAIR, J. E., HEDGES, S. B. **Molecular phylogeny and divergence times of deuterostome animals**. Molecular biology and evolution, v. 22(11), p. 2275-2284. 2005.

BOFFELLI, D., NOBREGA, M. A., RUBIN, E. M. **Comparative Genomics At The Vertebrate Extremes**. Nature. 2004.

BOSSERT, S., MURRAY, E. A. et al. **Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae**. Molecular Phylogenetics and Evolution. v. 130 (2019). p. 121-131. 2019.

BRODER, A.Z. **On the resemblance and containment of documents**. Proceedings of the Compression and Complexity. 1997.

CANABARRO, F., BARBEITOS, M. S., ADAM, C. **A Detecção de Elementos Ultraconservados em Metazoa**. Iniciação Científica, curso de Ciências Biológicas. Curitiba, p. 1-26. 2020.

CHEN, X. G., JIANG, X., GU, J., XU, M., WU et al**. Genome sequence of the Asian Tiger mosquito, Aedesalbopictus, reveals insights into its biology, genetics, and evolution**. Proceedings of the National Academy ofSciences, v. 112(44), p.E5907-E5915. 2015.

CRAWFORD, N. G., FAIRCLOTH, J. E., MCCORMACK, BRUMFIELD, R. T., WINKER, K., GLENN, T. C. **More then 1000 ultraconserved elements provide evidence that turtles are sister group of archosaurs**. Biology Letters. 2012.

CRISCUOLO, A. **A Fast Alignment-free Bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies**. Research Idead and Outcomes. 2019.

DICKEL, D. E., YPSILANTI, A. R., PLA, R., RUBENSTEIN, J. L. R., PENNACCHIO, L. A., VISEL, A. **Ultraconserved Enhancers Are Required for Normal Development**. Cell. (2018).

DIJK, V. L. E., AUGER, H., JASZCYSZYN, Y. e THERMES, C. **Ten years of next-generation sequencing technology**. CellPress. v. 30, p. 418-426. 2014.

FAIRCLOTH, B. C., MCCORMACK, J. E., CRAWFORD, N. G., HARVEY, M. G., BRUMFIELD, R. T., GLENN, T. C., **Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales**, Systematic Biology, Volume 61., p. 717–726. 2012.

FAIRCLOTH, B. C., SORENSON, L. SANTINI, F., ALFARO, M. E. **A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements** (UCEs). 2013.

GLAZOV, A. E., PHEASANT, M., MCGRAW, E., BEJERANO, G. e MATTICK, J. S. **Ultraconserved elements in insect genomes**: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Research. v. 15, p. 800-808. 2005.

GILBERT, Princess. S., et al. **Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes**. Molecular Phylogenetics and Evolution. Los Angeles-CA. p. 140-146. 2015.

GOODWIN, S., MCPHERSON, J. D. e MCCOMBIE, W.R. **Coming of age**: ten years of next-generation sequencing technologies. Nature Reviews Genetics, v. 17(6), p. 333. 2016.

GOULD, S. J., VRBA, E. S. **Exaptation - A Missing Term in the Science of Form**. Paleobiology. v. 8. p. 6. 1982;

HARMSTON, N., BARESIC, A. e LENHARD, B**. The mystery of extreme non-coding conservation**. Philosophical Transactions of the Royal Society B: Biological Sciences. v. 368. p. 1-12. 2013.

HUGENHOLTZ, P. e TYSON, G. W. **Metagenomics**. Nature, v. 455(7212), p. 481-483. 2008.

JARVIS, E. D., MIRARAB, S. et al. **Whole-Genome Analyses Resolve Early Branches In The Tree Of Life Ofmodern Birds**. Science. v. 346, p. 1320–1331. 2014.

MARDIS, Elaine. **The impact of next-generation sequencing technology on genetics**. Cell Press. St Louis-MO. v. 24, p. 133-141. 2007.

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

National Center for Biotechnology Information (NCBI). Bethesda (MD): **National Library of Medicine** (US), National Center for Biotechnology Information; [1988] – Disponível em: https://www.ncbi.nlm.nih.gov/.

OOSTERHOUT, C. V. **Mutation load is the spectre of species conservation**. Nature Ecology e Evolution. 2020.

RYU, T., SERIDI, L., RAVASI, T. **The Evolution of Ultraconserved Elements With Different Phylogenetic Origins**. BMC Evolutionary Biology. 2012.

SANGER, F. et al. **DNA Sequencing With Chain-Terminating Inhibitors.** Proc. Nati. Acad. Sci. USA. v. 74. p. 5463-5467. 1977.

SCHUSTER, C. STEPHAN. **Next-Generation Sequencing Transforms Today's Biology.** Nature. Pennsylvania-EUA. v. 5. p. 16-18. 2007.

SHEDLOCK, A. M., EDWARDS, S. V. **Amniotes** (amniota). The timetree of life, v. 375, p. 379. 2009.

SHENDURE, J., JI, H. **Next-generation DNA sequencing**. Nature Biotechnology. 2008.

SIEPEL, A., BEJERANO, G., PEDERSEN, et al. **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. Center for Biomolecular Science and Engineering. 2005.

SIMONS, C., PHEASANT, M., MAKUNIN, I., MATTICK, J. S. **Transposon-free regions in mammalian genomes**. Genome Research. 2006.

SIMONS, C., MAKUNIN, I. V., PHEASANT, M. e MATTICK, J. S. **Maintenance of transposon-free regions throughout vertebrate**. Evolution. BMC Genomics, Australia, 2007.

SPRINGER, M. S., MURPHY, W. J., EIZIRIK, E. e O'BRIEN, S. J. **Placental mammal diversification and the Cretaceous**–Tertiary boundary. Proceedings of the National Academy of Sciences, v. 100(3), p. 1056-1061. 2003.

STEPHEN, S., PHEASANT, 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, p. 402–408. 2008.

STEPHENS, Z. D., LEE, S. Y. et al. **Big data**: astronomical or genomical? PLoS biology, v. 13(7), p. 1-11. 2015.

VAN ITEN, H., LEME, J.M. et al. **Origin and early diversification of phylum Cnidaria**: key macrofossils from the Ediacaran System of North and South America. In The Cnidaria, past, present and future. p. 31-40. 2016.

WILHELM, A. J. **Next-generation DNA sequencing techniques**. New Biotechnology. v. 25. p. 195-203. 2009.

WITTKOPP, P., KALAV, G. ***Cis*-regulatory elements**: molecular mechanisms and evolutionary processes underlying divergence. Nature Ver Genet. v. 13. p. 59-69. 2012.

WOOD JR, P. L., XIANGUANG, G. **Parachute geckos free fall into synonymy**: Gekko phylogeny, and a new subgeneric classification, inferred from thousands of ultraconserved elements. Molecular Phylogenetics and Evolution. v. 146 (2020). p. 1-11.

XIAO FEI, Z. **BinDash**, software for fast genome distance estimation on a typical personal laptop, Bioinformatics. 2019.

XIAO, J. H., YUE, Z. et al. **Obligate mutualism within a host drives the extremespecialization of a fig wasp genome.** Genome Biology, v. 14(12), p.1-18. 2013.

YI, H., LIN, Y., JIN, W. **Sequences Dimensionality-Reduction by K-mer Substring Space Sampling Enables Effective Resemblance and Containment Analysis for Large-Scale omics-data**. BioRxiv. 2019.

ZHANG, J., CHIODINI, R., AHMED, B., ZHANG, G. **The Impacto of Next Generation-Sequencing on Genomics**. Journal of Genetics and Genomics. v. XX. p. 1-15. 2011.

ZHANG, Y. M., WILLIAMS, J. L., LUCKY, A**. Understanding UCES: A comprehensive primer on using Ultraconserved Elements for Arthropod Phylogenomics**. Insect and Systematics and Diversity. 2019.

APÊNDICE 1 – TABELA DOS 95 GENOMAS DE METAZOÁRIOS

(continua)

Filo/Subfilo	Espécies	Assembly Level	Tamanho Mpb	Acesso GenBank
Echinodermata	*Acanthaster planci*	scaffold	383.7	GCA_001949145.1
Cnidaria	*Acropora digitifera*	scaffold	447.5	GCA_000222465.2
Rotifera	*Adineta vaga*	scaffold	216.2	GCA_000513175.1
Arthopoda	*Aedes aegypti*	scaffold	1278.7	GCF_000004015.4
Cnidaria	*Aiptasia pallida*	scaffold	256.1	GCF_001417965.1
Cnidaria	*Alatina alata*	scaffold	125.6	GCA_008930755.1
Craniata	*Ambystoma mexicanum*	scaffold	32393,6	GCA_001455525.1
Porifera	*Amphimedon queenslandica*	scaffold	166,7	GCF_000090795.1
Arthopoda	*Amphinemura sulcicollis*	contig	271,925	GCA_001676325.1
Cnidaria	*Amplexidiscus fenestrafer*	scaffold	370.0	NC_027101.1
Craniata	*Anolis carolinensis*	chromosome	1799.1	GCF_000090745.1
Arthropoda	*Apis mellifera*	chromosome	235.3	GCF_000002195.4
Mollusca	*Aplysia californica*	scaffold	927.3	GCF_000002075.1
Chordata	*Arapaima gigas*	scaffold	664,315	GCA_900497675.1
Nematoda	*Ascaris lumbricoides*	contig	316,9	GCA_000951055.1
Cephalochordata	*Asymmetron lucayanum*	scaffold	460,6	GCA_001663935.1
Arthropoda	*Baetis rhodani*	contig	174,177	GCA_001676355.1
Arthropoda	*Bombyx mori*	scaffold	397.7	GCF_000151625.1
Cephalochordata	*Branchiostoma belcheri*	scaffold	426.1	GCF_001625305.1
Nematoda	*Caenorhabditis elegans*	genome	101.2	GCF_000002985.6
Annelida	*Capitella teleta*	scaffold	333.2	GCA_000328365.1
Arthopoda	*Centruroides exilicauda*	scaffold	925.5	GCA_000671375.1
Craniata	*Chelonia mydas*	scaffold	2208.4	GCF_000344595.1
Tunicata	*Ciona intestinalis*	chromosome	115.9	GCF_000224145.2
Mollusca	*Crassostrea gigas*	scaffold	557.7	GCF_000297895.1
Craniata	*Crocodylus porosus*	scaffold	2085.1	GCF_001723895.1
Craniata	*Crotalus horridus*	scaffold	1520,3	GCA_001625485.1
Arthropoda	*Culex quinquefasciatus*	scaffold	579.0	GCF_000209185.1
Chordata	*Danio rerio*	chromosome	1411,76	GCF_000002035.5
Crustacea	*Daphnia pulex*	scaffold	193378	GCA_000187875.1
Arthropoda	*Dendroctonus ponderosae*	scaffold	257.1	GCF_000355655.1
Dicyemida	*Dicyema*	scaffold	67,5424	GCA_011109175.1
Arthropoda	*Drosophila melanogaster*	chromosome	137.7	GCF_000001215.4
Nematoda	*Enterobius vermicularis*	contig	150.0	GCA_000951215.1
Cnidaria	*Enteromyxum leei*	scaffold	67.9	GCA_001455295.1
Arthropoda	*Ephemera danica*	scaffold	474,347	GCA_000507165.2
Platyhelminthes	*Fasciola hepatica*	scaffold	1236.45	GCA_000947175.1
Craniata	*Gallus gallus*	chromosome	1043.2	GCA_000002315.3
Craniata	*Geospiza fortis*	scaffold	1065.3	GCF_000277835.1
Arthropoda	*Glossosoma conforme*	scaffold	604,294	GCA_003347265.1
Annelida	*Helobdella robusta*	scaffold	235.4	GCF_000326865.1
Craniata	*Homo sapiens*	chromosome	2995.7	GCF_000001405.36
Crustacea	*Hyalella azteca*	scaffold	550.9	GCF_000764305.1
Cnidaria	*Hydra vulgaris*	scaffold	1055.6	GCA_000004095.1

(continua)

Filo/Subfilo	Espécies	Assembly Level	Tamanho Mpb	Acesso GenBank
Arthropoda	*Hypochthonius rufulus*	scaffold	172.4	GCA_000988845.1
Tardigrada	*Hypsibius dujardini*	scaffold	182.2	GCA_002082055.1
Orthonectida	*Intoshia linei*	scaffold	41,6	GCF_000208615.1
Arthropoda	*Isoperla grammatica*	contig	509,523	GCA_001676475.1
Arthropoda	*Ixodes scapularis*	scaffold	1765.4	GCA_001642005.1
Cnidaria	*Kudoa iwatai*	scaffold	26.4	GCA_001407235.1
Craniata	*Latimeria chalumnae*	scaffold	2798.5	GCF_000225785.1
Arthropoda	*Lednia tumana*	scaffold	304,502	GCA_003287335.1
Crustacea	*Lepeophtheirus salmonis*	contig	665.3	GCA_001005205.1
Arthropoda	*Limnephilus lunatus*	scaffold	1369,18	GCA_000648945.2
Arthropoda	*Limulus polyphemus*	scaffold	1828,3	GCF_000517525.1
Brachiopoda	*Lingula anatina*	scaffold	406,3	GCF_001039355.1
Nematoda	*Loa loa*	scaffold	93,9	GCF_000183805.2
Arthropoda	*Loxosceles reclusa*	scaffold	3262,4	GCA_001188405.1
Ctenophora	*Mnemiopsis leidyi*	scaffold	155.9	GCA_000226015.1
Arthropoda	*Musca domestica*	scaffold	636.3	GCF_000371365.1
Chordata	*Mus musculus*	chromosome	2671,82	GCA_000001635.8
Mollusca	*Mytillus galloprovincialis*	scaffold	1561.4	GCA_001676915.1
Craniata	*Nanorana parkeri*	scaffold	2053,8	GCF_000935625.1
Arthopoda	*Nasonia vitripennis*	chromosome	295.8	GCF_000002325.3
Nematoda	*Necator americanus*	scaffold	244.1	GCF_000507365.1
Cnidaria	*Nematostella vectensis*	scaffold	356.6	GCA_000209225.1
Craniata	*Notamacropus eugenii*	scaffold	3075.2	GCA_000004035.1
Mollusca	*Octopus bimaculoides*	scaffold	2338.2	GCF_001194135.1
Echinodermata	*Ophiothrix spiculata*	scaffold	2764,3	GCA_000969725.1
Cnidaria	*Orbicella faveolata*	scaffold	486.578	GCA_001896105.1
Craniata	*Orcinus orca*	scaffold	2372,9	GCF_000331955.2
Craniata	*Ornithorhynchus anatinus*	chromosome	1993	GCF_000002275.2
Echinodermata	*Parastichopus parvimensis*	scaffold	873	GCA_000934455.1
Arthopoda	*Pediculus humanus*	scaffold	110,7	GCF_000006295.1
Craniata	*Petromyzon marinus*	scaffold	1007,9	GCA_000148955.1
Ctenophora	*Pleurobrachia bachei*	scaffold	156,1	GCA_000695325.1
Craniata	*Poecilia reticulata*	chromosome	731,6	GCF_000633615.1
Priapulida	*Priapulus caudatus*	scaffold	511,7	GCF_000485595.1
Hemichordata	*Ptychodera flava*	scaffold	1228,6	GCA_900177555.1
Chordata	*Rattus norvegicus*	chromosome	2743,3	GCA_000001895.4
Cnidaria	*Renilla reniformis*	scaffold	131,5	GCA_001465055.1
Hemichordata	*Saccoglossus kowalevskii*	scaffold	775,8	GCF_000003605.2
Arthropoda	*Sarcoptes scabei*	scaffold	56,2	GCA_000828355.1
Platyhelminthes	*Schistosoma mansoni*	chromosome	364,538	GCA_000237925.3
Cnidaria	*Sphaeromyxa zaharoni*	scaffold	173,5	GCA_001455285.1
Arthropoda	*Strigamia maritima*	scaffold	176,2	GCA_000239455.1
Echinodermata	*Strongylocentrotus purpuratus*	scaffold	990,9	GCF_000002235.4
Nematoda	*Strongyloides venezuelensis*	contig	52,1	GCA_001028725.1
Platyhelminthes	*Taenia saginata*	scaffold	169,1	GCA_001693075.2
Craniata	*Tetraodon nigroviridis*	scaffold	342,4	GCA_000180735.1

(conclusão)

Filo/Subfilo	Espécies	Assembly Level	Tamanho Mpb	Acesso GenBank
Cnidaria	*Thelohanellus kitauei*	scaffold	150,3	GCA_000827895.1
Craniata	*Tyto alba*	scaffold	1120,1	GCF_000687205.1
Craniata	*Ursus maritimus*	scaffold	2301,3	GCF_000687225.1
Nematoda	*Wuchereria bancrofti*	scaffold	85,9	GCA_001555675.1
Craniata	*Xenopus tropicalis*	chromosome	1440,4	GCF_000004195.3

FONTE: National Center for Biotechnology Information (2020).

NOTA: * Tamanho dos genomas em milhões de pares de bases.

APÊNDICE 2 – SCRIPT PARA TRANSFORMAR FORMATO FASTA DE GENOMAS EM SKETCH

```
#!/bin/bash

function trim_path {

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

function get_prefix {

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

genome_file=$1

sketchsizes=( 16 )
bbits=( 3 )
kmerlens=( 18 )

prefix=$( get_prefix $genome_file )

for sketchsize in "${sketchsizes[@]}"
 do
    for bbit in "${bbits[@]}"
    do
       for kmerlen in "${kmerlens[@]}"
       do
          outfile=$prefix'_'$sketchsize'_'$bbit'_'$kmerlen
            ~/fernanda/bindash-master/release/bindash sketch
-- sketchsize64=$sketchsize --bbits=$bbit --kmerlen=$kmerlen
--outfname=$outfile.sketch $genome_file
             done
    done
done
```

APÊNDICE 3 – SCRIPT PARA COMPUTAR OS VALORES DE DISTÂNCIA ENTRE OS GENOMAS

```
#!/bin/bash

function trim_path {

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

function get_prefix {

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

function get_suffix {

   echo $1 | grep -o '[0-9].*[0-9]'
}


path_1=$( trim_path $1 )
path_2=$( trim_path $2 )

files=( $path_1/*.sketch )

for file_1 in "${files[@]}"
do
   prefix=$( get_prefix $file_1 )
   suffix=$( get_suffix $prefix )
   echo $suffix
   file_2=( $path_2/*$suffix* )

   echo "~/fernanda/bindash-master/release/bindash
dist $file_1 $file_2"
   exit
done
```

APÊNDICE 4 – SCRIPT PARA UNIR INFORMAÇÕES SOBRE UCES E VALORES DE DISTÂNCIA

(continua)

```
#!/bin/perl

use strict;
use warnings;

use Data::Dumper;

my $success = _sanitize_args();
my $fh_distance = _get_filehandle( $ARGV[0] );

my %DATA_HASH;

 for my $line ( <$fh_distance> )
{
   chomp $line;

   my @token = split /,/, $line;

   next if $token[0] eq $token[1];

   my @sorted_token = sort { $a cmp $b } ( @token[0,1] );


$DATA_HASH{ $sorted_token[0] }{ $sorted_token[1] }{ distance }
= $token[2];
}

close $fh_distance;

my $fh_uces = _get_filehandle( $ARGV[1] );

 my $error_flag = 0;
 for my $line ( <$fh_uces> )
{
   chomp $line;

   my @token = split /,/, $line;

   unless ( $DATA_HASH{ $token[0] }{ $token[1] } )
   {
       warn "Genome pair $token[0]:$token[1] not found in the
distance file!\n";
    $error_flag++;
   }
   else
   {
```

(continua)

```
            $DATA_HASH{  $token[0]  }{  $token[1]  }{  uces  }  =
$token[2];
        }
}

close $fh_uces;

die "Found $error_flag errors\n" if ( $error_flag );

print qq/"genome_1","genome_2","distance","uces"\n/;

for my $genome_1 ( sort { $a cmp $b } keys %DATA_HASH )
{
    for   my   $genome_2   (   sort   {   $a   cmp   $b   }
keys %{ $DATA_HASH{ $genome_1 } } )
    {
        my @tokens = ( $genome_1, $genome_2 );

        for my $variable ( 'distance', 'uces' )
     {
            push                                          @tokens,
$DATA_HASH{ $genome_1 }{ $genome_2 }{ $variable };
     }

        unless ( defined $tokens[-1] )
     {
         warn "Genome pair $genome_1:$genome_2 not found in the
UCE file! Will assume 0 UCEs\n";
         $tokens[-1] = 0;
     }

     my @token_check = ( grep { defined $_ } @tokens );
     @token_check == 4
            ||   die   "Something   wrong   with   genome_pair
$genome_1:$genome_2 - @tokens";

     print join ",", map { _format_tokens( $_, $tokens[ $_ ] ) }
( 0 .. @tokens - 1 );
     print "\n";
    }
}

# Verifica se os argumentos estão corretos
sub _sanitize_args
{
    for my $i ( 0, 1 )
```

(conclusão)

```
{
    for my $i ( 0, 1 )
    {
        defined $ARGV[$i] or
            die "Script need two arguments: the path to the
'distance' and 'uces' CSV files";

        -e $ARGV[$i] or
            die "Invalid path to either the 'distance' or 'uces'
CSV file";
    }

    return 1;
}

sub _get_filehandle
{
    my $filename = shift;

    open( my $fh, "<:encoding(UTF-8)", $filename )
        or die "Can't get filehandle for $filename: $!";

    return $fh;
}

sub _format_tokens
{
    my ( $index, $value ) = @_;

    $index < 3
     ? return q/"/ . $value . q/"/
     : return $value;
}
```