UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

LUCAS MICHEL CANDIDO DE RAMOS

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE NOVOS lncRNAS NO GENOMA DE LÚPULO *(Humulus lupulus)*, USANDO FERRAMENTAS DE BIOINFORMÁTICA

DOIS VIZINHOS

2024

LUCAS MICHEL CANDIDO DE RAMOS

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE NOVOS lncRNAS NO GENOMA DE LÚPULO *(Humulus lupulus)*, USANDO FERRAMENTAS DE BIOINFORMÁTICA

Identification and characterization of novel lncRNAs in the hop *(Humulus lupulus)* genome, using bioinformatics

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Bioprocessos e Biotecnologia do Curso de Bacharelado em Engenharia de Bioprocessos e Biotecnologia da Universidade Tecnológica Federal do Paraná.

Orientador: Profª. Drª. Tatianne Costa Negri Rocha

DOIS VIZINHOS

2024

LUCAS MICHEL CANDIDO DE RAMOS

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE NOVOS lncRNAS NO GENOMA DE LÚPULO *(Humulus lupulus)*, USANDO FERRAMENTAS DE BIOINFORMÁTICA

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Bioprocessos e Biotecnologia do Curso de Bacharelado em Engenharia de Bioprocessos e Biotecnologia da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 16/Outubro/2024

Profª. Drª. Tatianne Costa Negri Rocha
Doutorado em Informática e Gestão do Conhecimento
Universidade Tecnológica Federal do Paraná

Profª. Drª. Flavia Regina Oliveira de Barros
Doutorado em Reprodução Animal
Universidade Tecnológica Federal do Paraná

Profª. Drª. Juliana Morini Küpper Cardoso
Doutorado em Genética e Biologia Molecular
Universidade Tecnológica Federal do Paraná

DOIS VIZINHOS

2024

AGRADECIMENTOS

A finalização deste trabalho representa a concretização de um importante ciclo acadêmico, que não seria possível sem o apoio de diversas pessoas e instituições. Assim, deixo aqui meu mais sincero agradecimento a todos que contribuíram de alguma forma para esta jornada.

Agradeço primeiro a Deus primeiramente a Deus, por me conceder saúde, força e perseverança para superar os desafios ao longo desta caminhada. Pela minha mãe Marisa, pelo suporte emocional, incentivo constante e compreensão em momentos difíceis, você sempre foi minha base. Ao meu primo Leandro por todo suporte técnico, auxílio que foi essencial para a conclusão deste trabalho

À minha orientadora, Tatianne, pela paciência, orientação e dedicação. Suas contribuições foram fundamentais para o desenvolvimento deste trabalho e para meu crescimento acadêmico e pessoal.

Agradeço também aos colegas e amigos que compartilharam comigo esta caminhada, oferecendo apoio, ideias e momentos de descontração que tornaram este percurso mais leve e significativo.

À UTFPR e professores que proporcionaram os conhecimentos necessários para a realização deste trabalho, minha gratidão pela dedicação e pela formação de qualidade.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a execução deste trabalho. Sem vocês, esta conquista não seria possível.

Muito obrigado.

"*Se cheguei até aqui foi porque me apoiei no ombro dos gigantes"*
-Isaac Newton

RESUMO

O lúpulo (*Humulus lupulus*) é usado na produção de cerveja há mais de 1.200 anos e também tem aplicações farmacêuticas, exibindo benefícios consideráveis para a saúde humana. Apesar de o Brasil ser o terceiro maior produtor de cerveja do mundo, cerca de 98% do lúpulo utilizado no país é importado, o que aumenta significativamente o custo final do produto. Conhecer essas sequências de lncRNAs é essencial para programas de melhoramento genético da espécie, com o objetivo de adaptar o lúpulo a diferentes climas, aumentar sua produtividade e melhorar sua resistência a patógenos. Nesse contexto, os RNAs longos não codificantes (lncRNAs) desempenham funções importantes no genoma de plantas, incluindo a regulação da expressão gênica, regulação epigenética e resposta a estresses abióticos e bióticos. Durante a análise do banco de dados AlnC, identificamos que a maior sequência de lncRNA, registrada como AlnC_91911, possuía 2.405 pares de bases. Esse dado foi crucial para definir o processo de preparo do genoma para análises subsequentes. Com base no maior tamanho de fragmento de lncRNA encontrado, desenvolvemos um script em Python para recortar o genoma em sequências sobrepostas, definindo o tamanho de 2.500 pb com sobreposições de 2.499 pb. Após a montagem da biblioteca genômica, composta por 8713 arquivos de aproximadamente 1,1 GB cada, utilizamos os softwares CPC2 e RNAplonc para a predição de lncRNAs. O CPC2 identificou 3.582.809 transcritos como lncRNAs, enquanto o RNAplonc catalogou 2.263.355 transcritos. Para remover sequências redundantes, geradas pelas sobreposições, e extrair apenas aquelas classificadas como lncRNAs ou *non-coding*, utilizamos a ferramenta de *Business Intelligence Pentaho*. Após a eliminação das redundâncias, as sequências foram analisadas em outros programas de predição para aumentar a confiabilidade dos resultados. As sequências não codificantes identificadas pelo CPC2 e RNAplonc foram validadas através de outros programas preditores. Assim, classificamos as sequências conforme o número de programas que confirmaram a predição. As sequências de lncRNA identificadas pelos preditores foram comparadas com aquelas do banco de dados AlnC para eliminar duplicatas e garantir a originalidade dos resultados. Essa abordagem foi otimizada devido ao grande volume de dados (aproximadamente 8,7 TB) e ao tempo de processamento necessário para cada arquivo de 1,1 GB. Além disso com o objetivo de validar e comparar a eficácia das diferentes ferramentas, reavaliamos as sequências encontradas no banco de dados AlnC com as mesmas ferramentas RNAplonc (99,69%), CPC2(99,62%), CNCI(98%) e PLEK(99%). Concluímos que o método de 'fatiamento genômico' empregado neste trabalho se mostra tão eficiente quanto

os métodos convencionais baseados em RNA-seq, o quais apresentam um número inferior de identificações quando comparadas com o método proposto neste trabalho. Além disso, através da validação cruzada e comparação de sequências já conhecidas, garantimos confiabilidade e inovação nas novas sequências preditas. Os lncRNAs identificados neste estudo representam um rico recurso para futuras pesquisas, abrindo novas perspectivas para o melhoramento genético do lúpulo. A compreensão da função desses elementos regulatórios pode levar ao desenvolvimento de variedades mais produtivas, resistentes a doenças e com características agronômicas superiores.

Palavras-chave: plantas; genômica; não-codificantes; fatiamento; melhoramento.

ABSTRACT

Hops (*Humulus lupulus*) have been used in beer production for over 1,200 years and also have pharmaceutical applications, exhibiting considerable benefits for human health. Although Brazil is the third largest beer producer in the world, approximately 98% of the hops used in the country are imported, which significantly increases the final cost of the product. Knowing these lncRNAs sequences is essential for breeding programs, with the aim of adapting hops to different climates, increasing their productivity and improving their resistance to pathogens. In this context, long non-coding RNAs (lncRNAs) plays important roles in the plant genome, including the regulation of gene expression, epigenetic regulation and response to abiotic and biotic stresses. During the analysis of the AlnC database, we identified that the largest lncRNA sequence, registered as AlnC_91911, had 2.405 base pairs. This data was crucial to define the process of preparing the genome for subsequent analyses. Based on the largest lncRNA fragment size found, we developed a Python script to cut the genome into overlapping sequences, defining the size of 2.500 bp with overlaps of 2.499 bp. After assembling the genomic library, composed of 8,713 files of approximately 1,1 GB each, we used the CPC2 and RNAplonc software to predict lncRNAs. CPC2 identified 3.582.809 transcripts as lncRNAs, while RNAplonc cataloged 2.263.355 transcripts. To remove redundant sequences generated by overlaps and extract only those classified as lncRNAs or non-coding, we used the Pentaho Business Intelligence tool. After eliminating redundancies, the sequences were analyzed in other prediction programs to increase the reliability of the results. The non-coding sequences identified by CPC2 and RNAplonc were validated using other prediction programs. Thus, we classified the sequences according to the number of programs that confirmed the prediction. The lncRNA sequences identified by the predictors were compared with those in the AlnC database to eliminate duplicates and ensure the originality of the results. This approach was optimized due to the large volume of data (approximately 8.7 TB) and the processing time required for each 1,1 GB file. In addition, in order to validate and compare the effectiveness of the different tools, we re-evaluated the sequences found in the AlnC database with the same tools RNAplonc (99.69%), CPC2 (99.62%), CNCI (98%) and PLEK (99%). We conclude that the ‘genomic slicing’ method used in this study is as efficient as conventional methods based on RNA-seq, which present a lower number of identifications when compared to the method proposed in this study. Furthermore, through cross-validation and comparison of already known sequences, we ensure reliability and innovation in the new predicted sequences. The lncRNAs

identified in this study represent a rich resource for future research, opening new perspectives for the genetic improvement of hops. Understanding the function of these regulatory elements can lead to the development of more productive varieties, resistant to diseases and with superior agronomic characteristics.

Keywords: plants; genomic; non-coding; slicing; breeding.

SUMÁRIO

1 INTRODUÇÃO

Os RNAs longos não codificantes (lncRNAs) são definidos como um grupo de RNAs não codificantes com mais de 200 nucleotídeos, incapazes de produzir uma proteína completa. A principal diferença entre os lncRNAs e os mRNAs (RNAs mensageiros) está na ausência de um potencial de codificação detectável (BUDAK; KAYA; CAGIRICI, 2020). Nos últimos anos, a importância dos lncRNAs na regulação de diversos processos biológicos, como desenvolvimento, diferenciação e metabolismo, tem sido cada vez mais reconhecida, exemplo disso é o trabalho de WANG & CHANG, 2011, que identificou que apenas 1,5% do genoma humano é responsável pela transcrição de proteínas, enquanto muitos elementos regulatórios não codificantes são transcritos em RNAs não codificantes. Isso implica que os ncRNAs desempenham importantes funções regulatórias em organismos complexos.

Tal importância não se restringe apenas aos animais. Diversos estudos têm demonstrado a relevância das funções dessas moléculas no genoma das plantas, incluindo a regulação da expressão gênica, regulação epigenética e resposta a estresses bióticos e abióticos. Portanto a compreensão dessas moléculas (lncRNAs), no contexto do lúpulo (*Humulus lupulus*) se torna pertinente, visto que essa planta é originária de climas temperados, o qual é muito distinto dos encontrados no Brasil.

O lúpulo tem sido utilizado na produção de cerveja há mais de 1200 anos e também é utilizado como fármaco, exibindo efeitos consideráveis na saúde humana. Ele possui propriedades antioxidantes e anti-inflamatórias, além de conter prenilflavonoides, que são fitoestrógenos altamente ativos (COPATTI *et al.*, 2019). O gênero Humulus é composto por três espécies: *H. japonicus, H. yunnanensis e H. lupulus*, sendo que apenas esta última possui importância na indústria cervejeira. O *Humulus lupulus* pertence à família Cannabaceae, que faz parte da ordem Rosales (SOUZA, 2020).

O lúpulo desempenha um papel fundamental na produção de cerveja, conferindo características sensoriais únicas e contribuindo para a estabilidade coloidal da bebida. No entanto, apesar do Brasil ser o terceiro maior produtor de cerveja do mundo, cerca de 98% do lúpulo utilizado é importado, o que eleva consideravelmente o custo do produto final (IBGE - Instituto Brasileiro de Geografia e Estatística, 2020).

Diante do sucesso alcançado em áreas complexas e dos grandes investimentos em estudos na área, tanto no meio acadêmico quanto no setor corporativo, surgem perspectivas de utilização de ferramentas de bioinformática para a busca de novas regiões de lncRNAs no genoma do lúpulo (*Humulus lupulus*). Conhecer essas sequências, sua localização e modo de atuação é um passo importante na caracterização dessas moléculas. Essas informações podem ser utilizadas em estudos de melhoramento genético, visando aprimorar características específicas do lúpulo para torná-lo mais adequado a diferentes climas, aumentar sua produção, melhorar sua resistência a patógenos e reduzir a dependência de importação.

1.1 Objetivos

1.1.1 Objetivo geral

Realizar a busca no genoma do lúpulo por novas regiões de lncRNAs, utilizando como auxílio quatro ferramentas de bioinformática (CPC2, CNCI, RNAplonc, PLEK) além de uma re-avaliação com as mesmas ferramentar nas regiões já conhecidas de lúpulo, obtidas de bancos de dados de lncRNAs, a fim de validar os métodos e as ferramentas.

1.1.2 Objetivos específicos

- Realizar a busca na literatura pelas sequências de lncRNAs *Humulus lupulus* já descritos.
- Aplicar ferramentas de Bioinformática (CPC2, CNCI, RNAplonc, PLEK), para identificar novos lncRNAs no genoma de lúpulo.
- Comparar resultados obtidos com sequências previamente depositadas em bancos de dados biológicos.

2 REFERENCIAL TEÓRICO

2.1 RNAs não-codificantes

De acordo com Nelson e Cox (2019), os nucleotídeos desempenham várias funções no metabolismo celular. Eles funcionam como uma fonte de energia nas transações metabólicas, são fundamentais para as respostas das células a hormônios e outros estímulos externos, e são componentes estruturais de cofatores enzimáticos e intermediários metabólicos organizados. Além disso, eles são os blocos de construção dos ácidos nucleicos - DNA e RNA - que são responsáveis por armazenar e transmitir informações genéticas. O RNA é a única macromolécula conhecida que tem um papel tanto no armazenamento da informação quanto na catálise, o que levou a muita especulação a respeito do seu possível papel como intermediário químico no desenvolvimento da vida na Terra.

Ingram (1957) foi um dos primeiros a propor a relação direta entre um gene e uma proteína, uma ideia que, segundo Crick (1970), constitui o 'dogma central' da biologia molecular. Essa visão, que considerava os RNAs apenas como moléculas intermediárias na síntese proteica, foi dominante por muitos anos. Contudo, como afirmam Nam, Choi e You (2016), 'a descoberta dos RNAs não codificadores revolucionou nossa compreensão da biologia molecular'. A compreensão moderna do gene transcende a simples codificação de proteínas. Pesquisas como as de Gerstein *et al.* (2007) revelaram que regiões regulatórias e sequências transcritas em RNAs não codificadores também desempenham papéis cruciais na expressão gênica. O advento das tecnologias de sequenciamento de nova geração, como destacado por Shendure *et al.* (2017), possibilitou uma análise mais profunda e abrangente dos genomas, desvendando complexidades genômicas antes desconhecidas. Conforme Griffiths *et al.* (2022), os RNAs podem ser divididos em duas categorias principais: aqueles que codificam proteínas, como o mRNA, e aqueles que não codificam, os ncRNAs. O mRNA é frequentemente descrito como um 'mensageiro' molecular, pois ele carrega as instruções genéticas do núcleo para o citoplasma, onde as proteínas são sintetizadas.

Nesse contexto, ainda há muito o que se descobrir, com o auxílio da computação, pois, novas descobertas e novos métodos vão alterando ou consolidando teorias e hipóteses existentes, que não puderam ser testadas ou observadas anteriormente. Um exemplo clássico desse processo é o próprio Dogma Central da Biologia Molecular, sedimentado por Watson e Crick (WATSON; CRICK, 1953) que, à época, estabelecia que a única função do DNA era transcrever RNA para a produção de proteínas. Hoje, sabe-se que existem RNAs que não traduzem proteínas, mas que atuam na regulação gênica das células (KUNG; COLOGNORI; LEE, 2013).

Os RNAs não codificantes (ncRNAs) são ácidos ribonucleicos (RNA) que são transcritos, mas não são traduzidos em proteínas. Com os avanços científicos na compreensão do transcriptoma, muitos desses RNAs que eram considerados "lixo" até então, revelaram terem papéis funcionais, mesmo que desconhecidos. Por exemplo, o projeto "*The Encyclopedia of DNA Ele-*

ments (ENCODE)", revelou que mais de 80% do genoma da humanidade é formado por RNAs transcritos associados à funções bioquímicas, enquanto somente de fato 1,5% codificam proteínas (ECKER *et al*., 2012). A elucidação deste "DNA lixo", permitiu que os ncRNAs fossem reconhecidos como cruciais componentes da complexidade celular e organização genômica (BIRNEY; AL., 2007).

É sabido que cerca de apenas 2% do genoma humano consiste em genes que codificam proteínas, resultando em aproximadamente 20 mil genes. Por causa disso, alguns pesquisadores usaram o termo "DNA lixo"para se referir às sequências não codificadoras de proteínas presentes no genoma (PONTING; BELGARD, 2010);(DOBBLER, 2015). No entanto, o conceito de "DNA lixo"é considerado desatualizado e impreciso atualmente. Embora a maior parte do genoma não codifique proteínas, a maioria das sequências não codificadoras desempenha funções importantes na regulação e modulação de processos biológicos complexos. Um pensamento comum era de que essas sequências "extras", como genes duplicados e sequências não codificantes, poderiam apresentar funcionalidade acumulando mutações sem que estas ocorressem em genes codificantes. No entanto, atualmente está claro que até 90% do genoma eucarioto é transcrito, gerando uma grande quantidade de RNAs sem capacidade codificadora (DOBBLER, 2015).

Embora não codifiquem proteínas, os ncRNAs desempenham papéis fundamentais nas células. Eles são responsáveis por muitas funções celulares, incluindo atuar como catalisadores em diversas reações bioquímicas e na regulação genética. Essa regulação pode ocorrer em vários níveis importantes para a função do genoma, como a arquitetura da cromatina, a segregação cromossômica, o processamento e splicing do RNA, a transcrição, a tradução e o turnover proteico. Portanto, é evidente que o RNA não codificante é fundamental para a regulação da expressão gênica e o funcionamento celular adequado.

Anteriormente considerado como "lixo evolutivo"devido à sua ineficácia na tradução em proteínas, o ncRNA é agora reconhecido como tendo um impacto em diversos mecanismos moleculares e funções biológicas relevantes (RIELLA, 2019). A versatilidade do ncRNA surge da sua capacidade de formar estruturas complexas que interagem com outras moléculas de RNA, DNA, proteínas e outras pequenas moléculas (DYKSTRA; KAPLAN; SMOLKE, 2022). Além disso, a quantidade de ncRNA presente em um organismo pode ser correlacionada com a sua complexidade, o que reforça a influência que essas moléculas têm no desenvolvimento e organização dos seres vivos (BEERMANN *et al*., 2016). O ncRNA pode ser classificado em dois grupos principais: *small noncoding RNA* (sncRNA) e *long noncoding RNA* (lncRNA).

Os ncRNAs surgiram como produtos importantes do transcriptoma eucariótico, com papel regulatório significativo (RYMARQUIS *et al*., 2008); (BROSNAN; VOINNET, 2009). Nos últimos dez anos, houve avanços significativos na compreensão das funções e mecanismos dos microRNAs (miRNAs), dos pequenos RNAs interferentes (siRNAs) e dos siRNAs antisense naturais (nat-siRNAs) na regulação da expressão gênica em nível transcricional e pós-transcricional (CUPERUS; FAHLGREN; CARRINGTON, 2011); (CHEN, 2012). Mais recente-

mente, foram catalogadas moléculas com mais de 200 nucleotídeos como RNAs não codificantes longos (lncRNAs), que atuam como novos elementos regulatórios envolvidos em muitos processos biológicos em mamíferos (CHEETHAM *et al.*, 2013); (LEUNG *et al.*, 2013); (NG *et al.*, 2013).

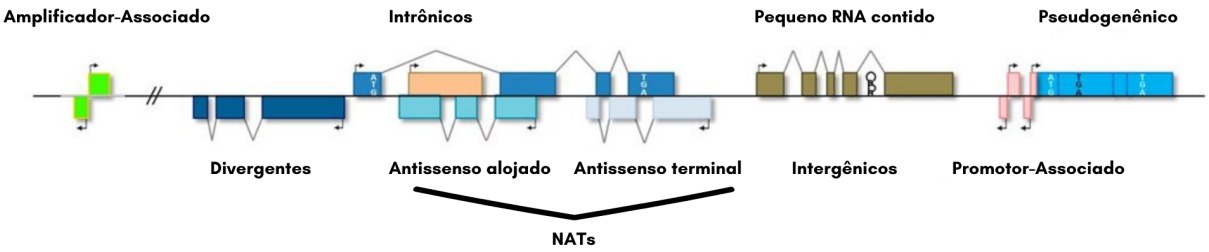
2.1.1 RNAs longos não–codificantes

Os lncRNAs são um grupo heterogêneo de RNAs não codificantes, geralmente com mais de 200 nucleotídeos e menos éxons do que os mRNAs (BUDAK; KAYA; CAGIRICI, 2020); (BHATIA *et al.*, 2017). Eles são distinguidos com base em sua posição em relação aos genes codificadores de proteínas vizinhos no genoma, sendo os (i) RNAs não codificantes intergênicos longos (lincRNAs) definidos como transcritos de RNA com mais de 200 nucleotídeos localizados entre dois genes na mesma fita. Por definição, esses RNAs não devem ter quadros de leitura abertos que codifiquem proteínas (RUIZ-ORERA *et al.*, 2014). A maioria desses transcritos é gerada pela RNA polimerase II, sendo submetidos ao processo de *splicing* e poliadenilação (ULITSKY; BARTEL, 2013)., (ii) Transcritos antisenso naturais (NATs) são moléculas de RNA que são transcritas a partir da cadeia de DNA oposta a outros transcritos e se sobrepõem em parte com o RNA senso. Tanto o RNA senso quanto o antisenso podem codificar proteínas ou ser transcritos não codificadores de proteínas (FAGHIHI; WAHLESTEDT, 2009). Os NATs têm um papel na regulação de respostas a diversos estresses abióticos e bióticos (LAVORGNA *et al.*, 2004). (iii) Há exemplos de RNAs intrônicos longos não codificantes (incRNAs) descritos na literatura (0,6-2 kb), que são lncRNAs derivados de um íntron na mesma fita e estão envolvidos em diferentes processos biológicos, incluindo o controle transcricional ou pós-transcricional da expressão gênica (GOODRICH; KUGEL, 2006);(KHOCHBIN; LAWRENCE, 1989), bem como na regulação do salto de exon (YAN *et al.*, 2005) e retenção de íntrons (KRYSTAL; ARMSTRONG; BATTEY, 1990)., (iv) pseudogenes, que são frequentemente gerados por meio de duplicação de genes em tandem ou retrotransposição de RNAs mensageiros, que geram cópias extras de genes que não estão mais sob pressão seletiva (KUNG; COLOGNORI; LEE, 2013). (v) Os lncRNAs sobrepostos são transcritos de lncRNA que contêm um gene codificador de proteína dentro do "íntron"do lncRNA ou são lncRNAs que se sobrepõem ao íntron de um gene codificador de proteína (MATTICK; RINN, 2015). (vi) Divergentes são lncRNAs que iniciam a transcrição na fita oposta de um gene. E (vii) transcritos promotores associados, e enhancers (intensificadores) RNAs (eRNAs), são transcritos curtos abundantes (variando de 20 a 2500 nt) que são produzidos a partir da circulação dos locais de início da transcrição nas direções sense e antissense (KUNG; COLOGNORI; LEE, 2013).

No entanto, outras classificações de lncRNAs de plantas surgiram nos últimos anos (Figura 1), incluindo lncRNAs nucleares, que controlam a expressão gênica por meio da remodelação da cromatina, *splicing* alternativo e modificações epigenéticas (NGUYEN; VU; CHEONG, 2022). Recentemente, com base em seu modo funcional ou tipo de interação, os lncRNAs foram

classificados naqueles que formam complexos ribonucleoproteicos, híbridos RNA-DNA, duplexes RNA-RNA e uma subclasse que regula genes codificadores de proteínas próximos por meio de mecanismos desconhecidos (LUCERO *et al.*, 2021). Esses lncRNAs podem regular a expressão gênica na modificação da cromatina, nível transcricional e pós-transcricional, atuando como sinais, *decoys*, guias e *scaffolds* (WANG; CHANG, 2011). Além disso, evidências emergentes sugerem que a expressão de alguns lncRNAs é altamente específica do tecido e muitos deles respondem a estresses bióticos e abióticos (ZHU *et al.*, 2014); (DI *et al.*, 2014) & (SHUAI *et al.*, 2014) . Os estudos dos papéis biológicos dos lncRNAs são desafiadores devido aos seus diversos padrões de expressão e regulação em uma ampla gama de células e tecidos (ØROM; SHIEKHATTAR, 2011).

Figura 1 – Classificação dos diferentes tipos de lncRNAs



Fonte: Adaptado de (KUNG; COLOGNORI; LEE, 2013). Os lncRNAs podem ser unidades de transcrição autônomas ou podem ser transcritos de amplificadores (eRNAs), promotores (TSSa-RNAs, uaRNAs, pasRNAs e PROMPTs) ou íntrons de outros genes (neste caso, um gene codificador de proteína, com códon de início ATG e códon de parada TGA, em branco); de pseudogenes (mostrado aqui com um códon de parada prematuro TGA em preto); ou antisense para outros genes (NATs) com vários graus de sobreposição, com nenhum grau (divergente), parcial (terminal) e completo (alojado). Os lncRNAs também podem hospedar um ou mais pequenos RNAs (*hairpin*) dentro de suas unidades de transcrição.

Estudos mostraram que a expressão do lncRNA é mais específica do tipo de célula do que os genes codificadores de proteínas. No nível do DNA e da cromatina, os *loci* do lncRNA são semelhantes aos *loci* do mRNA, mas os lncRNAs mostram um viés por ter apenas um íntron e uma tendência para *splicing* cotranscricional menos eficiente (BATISTA; CHANG, 2013).

2.1.2 RNAs longos não–codificantes em plantas

Os ncRNAs reguladores em plantas são um grupo principal de transcritos endógenos que não codificam proteínas, mas são capazes de regular a expressão gênica em diferentes níveis, incluindo os níveis transcricional, pós-transcricional e translacional. Nos últimos anos, milhares de longos (lnc)RNAs não codificantes foram identificados nos transcriptomas eucarióticos por meio de tecnologias de sequenciamento de próxima geração (LUCERO *et al.*, 2021). A caracterização funcional desses lncRNAs em plantas ainda é um grande desafio. Para superar essa barreira, novas análises estruturais e abordagens de biologia celular têm sido combinadas

com edição de genoma e deleções genômicas mediadas por tecnologias como CRISPR, com o objetivo de caracterizar as funções moleculares do ncRNA (LUCERO *et al.*, 2021).

Os longos RNAs não codificadores (lncRNAs) em vegetais exercem um papel crucial na regulação da expressão gênica, podendo modular tanto genes vizinhos (cis-regulação) quanto genes distantes (trans-regulação). Os lncRNAs de ação cis atuam próximo ao local de síntese e operam diretamente nas sequências nucleotídicas locais ou em regiões cromossômicas em um ou mais genes contíguos. Por outro lado, os lncRNAs de ação trans se dispersam do local de síntese e podem atuar em vários genes a grandes distâncias, inclusive em diferentes cromossomos (WAITITU *et al.*, 2020).

Vários lncRNAs na escala do genoma foram caracterizados em plantas como respostas bióticas diretas ou indiretas, por exemplo em arroz (*Oryza sativa*) (JAIN *et al.*, 2017) e *A. thaliana* (ZHU *et al.*, 2014) infectados com os patógenos fúngicos, *Magnaporthe oryzae* e *Fusarium oxysporum* , respectivamente. Já os pesquisadores, (ZHANG *et al.*, 2018), relataram que o silenciamento de dois lncRNAs o GhlncNAT-ANX2 e GhlncNAT-RLP7, que estão envolvidos na regulação da atividade da lipoxigenase (LOX1 e LOX2), resultou em maior resistência aos patógenos *Verticillium dahliae* e *Botrytis cinerea* no algodão. Da mesma forma, em *A. thaliana* , o lncRNA ELF18INDUCED LONGNONCODING RNA1 (ELENA1) demonstrou melhorar a resistência a *Pseudomonas syringae* interagindo com a subunidade 19a do mediador, que por sua vez regula positivamente a expressão do gene 1 (PR1) relacionado à patogênese (SEO *et al.*, 2017). No tomate, a superexpressão de lncRNA16397 demonstrou induzir a expressão de glutaredoxina, diminuir a produção de espécies reativas de oxigênio e conferir resistência a *Phytophthora infestans* (CUI *et al.*, 2017). Além disso, um conjunto abrangente de 529 lncRNAs putativos foi identificado usando métodos de bioinformática em tomate (WANG *et al.*, 2015).

Nath *et al.* (2021) identificou lncRNAs envolvidos na biossíntese do metabólitos secundários de lúpulo, como o ativador de transcrição de ligação à calmodulina e o fator de transcrição TCP4. Esses lncRNAs têm um papel potencialmente importante na regulação da biossíntese de metabólitos secundários e no desenvolvimento de tricomas glandulares. Além disso, foram encontrados diversos genes-alvo de lncRNA envolvidos na via de biossíntese dos fenilpropanóides, entre eles, foram encontrados também diversos fatores de transcrição como WRKY1, bZIP e bHLH2 e genes estruturais como a 4-cumarato-CoA ligase2 (4CL2) e fenil amônia-liase (PAL) os quais foram previstos como alvos putativos dos lncRNAs candidatos. Além disso, alguns lncRNAs de plantas respondem ao estresse abiótico por meio da remodelação da cromatina (WAITITU *et al.*, 2020). A expressão de COOLAIR e COLDAIR reprime o lócus FLC em *Arabidopsis* estressada pelo frio por meio de modificações na cromatina mediadas por lncRNA (lncR2Epi) (WAITITU *et al.*, 2020). O COOLAIR medeia a redução de H3K36me3 ou H3K4me2 no lócus FLC (WAITITU *et al.*, 2020), enquanto o COLDAIR se combina com o complexo repressivo polycomb2 (PRC2) para promover o acúmulo de H3K27me3 no FLC (WAITITU *et al.*, 2020). Alguns lncRNAs respondem ao estresse ambiental por meio de via RdDM. Em tomates, a supressão de SlAGO4A, que codifica um fator central da via RdDM, aumentou significativa-

mente a resistência ao sal e ao estresse hídrico em comparação com espécies selvagens e plantas transgênicas superexpressando SlAGO4A (WAITITU *et al*., 2020).

Os LncRNAs relatados em espécies de plantas são limitados a apenas algumas plantas modelo de angiospermas, como Arabidopsis, arroz, milho, trigo, painço e soja (LIU *et al*., 2015). A tarefa de descobrir novos lncRNAs vegetais ainda é muito árdua. Nos últimos anos, o sequenciamento do DNA genômico de plantas se desenvolveu rapidamente, e dados de sequenciamento do genoma de dezenas de espécies de plantas foram relatados (LIU *et al*., 2015).

2.2 O Lúpulo

O gênero Humulus, assim como o gênero Cannabis e os gêneros anteriormente classificados na família Celtidaceae, fazem parte da família Cannabaceae. Esse gênero é composto por três espécies principais: *H. lupulus L., H. scandens (Lourr.) Merr. e H. yunnanensis Hu*. O lúpulo comum é uma planta trepadeira perene e dióica, que cresce naturalmente em sebes e orlas de regiões temperadas da Europa, Ásia e América do Norte (BOCQUET *et al*., 2018). Essa planta apresenta forte dominância apical e o crescimento lateral, onde as flores se desenvolvem, é iniciado posteriormente (MARCOS *et al*., 2011). Além disso, o caule herbáceo do lúpulo é capaz de se enrolar em torno de um suporte e atingir até 10 metros de altura (BOCQUET *et al*., 2018).

Figura 2 – Corte transversal em inflorescências de lúpulo

Fonte: Adaptado de (SPóSITO RODRIGO VERALDI ISMAEL,). A inflorescência (cone) de plantas femininas (na esquerda). As flores protegidas pelas brácteas (no centro) e a raquis onde se fixam as flores e as brácteas (na direita).

Os cones de lúpulo utilizados na indústria cervejeira são formados pelas inflorescências das plantas femininas. Esses cones são compostos por brácteas e bractéolas semelhantes a pétalas, que cercam um eixo central ou estribo. À medida que o lúpulo amadurece, as glândulas

de lupulina são formadas na base das bractéolas. Somente as plantas femininas são capazes de produzir essas glândulas de lupulina, que consistem em um pó fino, resinoso e amarelo (ALMAGUER *et al.*, 2014). Já as plantas masculinas produzem panículas multiramificadas contendo muitas flores minúsculas, com cerca de 7,5 a 12,5 cm de comprimento (LIN *et al.*, 2019).

Nas glândulas de lupulina é onde os principais princípios cervejeiros do lúpulo, as resinas e os óleos essenciais, são sintetizados e acumulados. Foi em 1821 que Ives atribuiu o nome 'lupulina' a esse pó amarelo. Ele foi o primeiro a observar que é na lupulina onde as substâncias amargas e aromáticas do lúpulo são armazenadas (ALMAGUER *et al.*, 2014). Essas glândulas são compostas por grupos de compostos economicamente valiosos para a indústria cervejeira e para produtos de saúde (LIN *et al.*, 2019).

Os cones de lúpulo da espécie *Humulus lupulus L.*, cultivados principalmente para a indústria cervejeira, contêm vários elementos como resinas, óleos essenciais, proteínas, polifenóis, lipídios, ceras, celulose e aminoácidos. O valor cervejeiro do lúpulo está nas resinas secretadas pelas glândulas de lupulina, que contêm compostos ativos de sabor e amargor. Além disso, os óleos essenciais de lúpulo são importantes para fornecer características de sabor e aroma à cerveja. Aproximadamente 97% do lúpulo cultivado em todo o mundo é utilizado na fabricação de cerveja (ALMAGUER *et al.*, 2014).

2.2.1 A indústria cervejeira

Apesar de haver estudos que descrevem o uso do lúpulo na área da saúde, devido às suas propriedades antimicrobianas, antiinflamatórias, fitoestrogênicas e calmantes, é na indústria cervejeira que o seu uso é mais comum (DURELLO; SILVA; BOGUSZ, 2019). O lúpulo é um ingrediente importante na produção de cerveja, pois além de conferir características sensoriais, promove estabilidade coloidal à espuma e atua como antioxidante e antimicrobiano, protegendo a cerveja de processos oxidativos e contaminações microbiológicas (SCHÖNBERGER; KOSTELECKY, 2011). Tamanha é a importância do lúpulo para a indústria cervejeira, que assim como o malte, é muitas vezes chamado de "alma da cerveja". Variando-se apenas o tipo e/ou quantidade de lúpulo em uma mesma receita base, é possível fabricar cervejas com perfis sensoriais totalmente distintos em termos de amargor e aroma (INUI *et al.*, 2013) (NACHEL, 2011).

Os Estados Unidos e a Alemanha são os dois maiores produtores de lúpulo, sendo produzidas 48.191 e 41.556 toneladas anualmente, respectivamente (IHGC - International Hop Growers' Convention, 2018). O Brasil é o terceiro maior produtor de cerveja no mundo (CARVALHO *et al.*, 2018) com cerca de 15,3 bilhões de litros produzidos em 2023 (PECUáRIA, 2024). Esse grande volume de cerveja requer grandes quantidades de insumo, sendo os principais: malte de cevada e lúpulo, deste, pelo menos 98% do produto é importado (BERBERT, 2017).

2.2.2 Diferença climática e Melhoramento genético do Lúpulo

O lúpulo tem origem nas zonas temperadas do hemisfério Norte, com a China sendo o provável local de onde se originou e se espalhou para a Europa e América do Norte (NEVE, 1991). Geralmente, é cultivado em países com clima temperado, situados entre os paralelos 35 e 55 em ambos os hemisférios, e sua colheita ocorre anualmente (BOCQUET *et al.*, 2018). Lavouras de lúpulo são encontradas em locais de climas diversos (MAHAFFEE *et al.*, 2009), mas isso não significa que as plantas consigam se desenvolver de maneira adequada ou minimamente satisfatória em todas essas situações.

Para que o cultivo do lúpulo seja viável, a disponibilidade de luz durante o período de crescimento é a principal variável física a ser considerada, sendo influenciada pela latitude onde a área plantada está localizada. É essencial que durante a fase de crescimento vegetativo, haja dias longos com 16 a 18 horas de luminosidade, a fim de se obter os melhores resultados dessa cultura, conforme apontado por (ENGELHARD; LUTZ; SEIGNER, 2011).

O estudo de Wang *et al.* (2014), identificou e caracterizou um novo tipo de RNA não codificante (lncRNA) chamado HID1, que está envolvido no processo de fotomorfogênese. Esse RNA pode regular negativamente a expressão do gene PIF3, que é um dos principais repressores na resposta à luz. HID1 está localizado no núcleo e pode se associar à cromatina e à região promotora do PIF3 para reprimir sua expressão. Além disso, os homólogos de HID1 são encontrados em muitas espécies de plantas e podem ter funções semelhantes em diferentes espécies.

Com o objetivo de obter novas características, como aromas mais intensos e maior rendimento, são realizados programas de cruzamento de lúpulo para criar novas variedades (ou cultivares), proporcionando novas experiências para os cervejeiros ou melhorias para os produtores. Esse processo de desenvolvimento de novas variedades leva cerca de 10 a 15 anos de pesquisa e testes, até que a nova cultivar seja protegida. No entanto, a utilização de seleção assistida por marcadores moleculares (SAMM) pode acelerar esse processo (ČERENAK *et al.*, 2019). Portanto, o conhecimento desses mecanismos moleculares é importante para futuras aplicações em programas de melhoramento genético do lúpulo.

2.3 Bancos de dados de lncRNAs em plantas

Nos últimos anos, foram estabelecidos 20 bancos de dados de lncRNA relacionados a plantas, que são ferramentas úteis para permitir um estudo preciso e detalhado desses RNAs (LOU *et al.*, 2022). Embora milhares de lncRNAs tenham sido identificados em *Arabidopsis* e outras plantas, e sua expressão tenha sido caracterizada em todo o genoma, nem todos foram registrados e anotados em bancos de dados públicos (JIN *et al.*, 2013). Assim, o uso dos bancos de dados de lncRNAs é uma abordagem valiosa para a facilitação de pesquisas mais aprofundadas e precisas sobre esses RNAs. No entanto, a maioria desses bancos de dados

fornece informações básicas de lncRNAs em espécies e predição de genes-alvo de acordo com dados de transcriptoma, entre os quais 11 bancos de dados podem ser citados (Tabela 1), totalizando dados de 838 espécies de plantas.

Banco de dados	Número de espécies catalogadas	Link
AlnC (SINGH; VIVEK; KUMAR, 2021)	804	14.139.61.8/AlnC/
CantataDB (SZCZEŚNIAK; ROSIKIEWICZ; MAKAŁOWSKA, 2016)	39	cantata.amu.edu.pl/
EVLncRNAs (ZHOU *et al.*, 2023)	43	www.sdklab-biophysics-dzu.net/EVLncRNAs2/
GreenC (GALLART *et al.*, 2016)	94	greenc.sequentiabiotech.com/wiki2/Main$_{Page}$
LncPheDB (LOU *et al.*, 2022)	9	www.lncphedb.com/
NONCODE (LIU *et al.*, 2005)	23	v5.noncode.org/
PNRD (YI *et al.*, 2015)	150	tools4mirs.org/software/mirna$_{d}atabases/pnrd/$
PlncDB (JIN *et al.*, 2013)	80	www. tobaccodb.org/plncdb/
PlncRNADB (BAI *et al.*, 2019)	4	bis.zju.edu.cn/PlncRNADB/index.php
NPInter (WU *et al.*, 2006)	2	bigdata.ibp.ac.cn/npinter
lncRNAdb (AMARAL *et al.*, 2011)	10	lncrnadb.org/

TABELA 1: Tabela com os bancos de dados que catalogam lncRNAs de diferentes espécies de organismos vivos.

Das 838 espécies mapeadas conta-se com um total de 12.896.356 lncRNAS catalogados. Entretanto, apesar da grande quantidade de espécies registradas nesses bancos de dados, o lúpulo é encontrado apenas no banco AlnC, ainda assim com poucas sequências de lncRNAs identificadas, contando com 21888 catalogados.

AlnC está operando em uma pilha Linux, Apache, MySQL e PHP a partir de agora. A estrutura de banco de dados atual é construída no servidor Apache e a interface da web AlnC foi

projetada usando HTML, CSS e JavaScript. Todos os AlnC lncRNAs e outras anotações de dados relacionadas são manipuladas por um banco de dados relacional configurado com MySQL. Além disso, o AlnC é integrado ao BLAST (v2.11.0) autônomo para pesquisa de similaridade online, ViennaRNA (v2.4.16) para a visualização da estrutura secundária e ORFfinder (v0.4.3) para a exploração de lncRNA contendo sORFs (ROMBEL *et al.*, 2002); (ALTSCHUL *et al.*, 1990) & (LORENZ *et al.*, 2011).

Segundo Singh, Vivek e Kumar (2021), foi organizada e compilada uma coleção de 10,855,598 lncRNAs de 809 amostras disponíveis no projeto 1 KP, que funciona como um catálogo abrangente de lncRNA de 682 plantas com flores armazenadas na plataforma AlnC. Não existem outros repositórios de dados sobre lncRNAs nessa escala, e a maioria dos lncRNAs das espécies incluídas em AlnC pertencem a táxons pouco estudados, tornando AlnC de grande interesse entre os pesquisadores de plantas. Usando-se um fluxo de trabalho de análise adaptado para plantas, mas que também pode ser usado para identificação de lncRNA baseada em RNA–seq em espécies não vegetais. O fluxo de trabalho fornece um pipeline de bioinformática baseado em aprendizado de máquina (ML) para identificar lncRNAs de alta confiança em organismos angiospermas, o que difere de outros métodos de anotação usados em bancos de dados de lncRNA já disponíveis (como CATATAdb) (SZCZEŚNIAK; ROSIKIEWICZ; MAKAŁOWSKA, 2016). Este método produziu um grande número de transcritos de lncRNA putativos, que foram então organizados e catalogados na interface web AlnC. AlnC inclui lncRNAs com evidência de pontuação de probabilidade de RNA não codificante e permite exploração adicional de ORFs juntamente com outros recursos primários de lncRNA, fornecendo assim aos pesquisadores capacidade funcional para alavancar dados e informações de AlnC em seus projetos individuais por meio de sua interface web.

2.4 Ferramentas de busca de lncRNAs

A Bioinformática não possui métodos únicos para identificação e classificação de lncRNAs como pode ser visto na Tabela 1, temos uso de SVM (Suporte de máquina de vetor), LR (Regressão Logística), RF (Árvore de decisão Aleatória), DL (Deep Learning), DNN (Rede Neural Profunda), CNN (Rede Neural Convolucional), DBN (*Deep belief network*) e REPtree (*Reduced Error Pruning Tree*). Na Tabela 1, o SVM e o RF são os modelos mais usados, já as características para a construção dos modelos de classificação variam pouco, visto que o conhecimento a respeito dessas moléculas, suas funções e suas estruturas secundárias ainda é muito incipiente. A área de pesquisa de ncRNA está crescendo rapidamente. No entanto, ainda é um desafio distinguir lncRNAs de genes codificadores de proteínas, pois lncRNAs compartilham muitas características semelhantes aos mRNAs. Além disso, existem muitos transcritos ou genes incompletos mal anotados ou contendo erros de sequenciamento que também geram grandes erros de discriminação. Nos últimos dez anos, foram feitos muitos esforços na identificação de lncRNA e muitas abordagens foram desenvolvidas para fazer uma

discriminação mais precisa. Diferentes abordagens com o objetivo de detectar diferentes tipos de ncRNAs são apresentadas na Tabela 2 além de uma visão geral de algumas ferramentas.

Tabela 2 – Lista de ferramentas descritas na literatura

Ferramentas	Característica	Modelo	Espécie de treino/ Ano
CONC (LIU; GOUGH; ROST, 2006)	tamanho do peptídeo, composição de aminoácidos, hidrofobicidade, conteúdo da estrutura secundária, porcentagem de resíduos expostos ao solvente, entropia de composição da sequência, número de homólogos obtidos por PSI-BLAST, entropia de alinhamento.	SVM	Eucarioto/ 2006.
CPC (KONG; AL., 2007)	qualidade da ORF e alinhamento.	SVM	Eucarioto/ 2007.
PhyloCSF (LIN; JUNGREIS; KELLIS, 2011)	GC%, conservação do DNA, estrutura secundária de energia livre, conservação da estrutura secundária, conservação da sequência de proteína, Poly-A + RNA seq (max), pequenos RNA seq (max), Total RNA mantido array (max), Poly-A + RNA mantido array (max).	SVM	Drosófila/ 2011.
CPAT (WANG; AL., 2013)	Tamanho da ORF, cobertura da ORF, estatística de teste de codificação de Fickett, e hexâmero.	LR	Humano, camundongo, mosca, peixe-zebra/ 2013.
CNCI (SUN *et al.*, 2013)	Codon-ant matrix e bias, estrutura de sequência MLCDS.	SVM	Humano, camundongo, minhoca, Agrião/ 2013.
iSeeRNA (SUN; AL., 2013)	10 características para treinamento do modelo: uma característica de conservação do transcrito, duas características relacionadas a ORF (tamanho total e proporção do tamanho da ORF pelo tamanho do transcrito) e sete características de frequência de di- e trinucleotídeos.	SVM .	Humano, Camundongo/ 2013.
RF-based classifier (LERTAMPAIPORN; AL., 2014)	tamanho da sequência, modularidade, potencial de codificação, robustez da estrutura.	RF	Humano/ 2014.
PLEK (LI; ZHANG; ZHOU, 2014)	K-mer de 1 a 5.	SVM	Humano/ 2014.
LncRNA-MFDL (FAN; ZHANG, 2015)	Tamanho e cobertura da ORF, Estrutura secundária MLCDS.	DL	Humano/ 2015.

continua na próxima pagina

Tabela 2 continuação

Ferramentas	Característica	Modelo	Treinamento/ Ano
LncRNA-ID (ACHAWANANTAKUN; AL., 2015)	Tamanho e cobertura da ORF. Kozak motif. Sinal de liberação, alterações de ligação e alinhamento de energia do ribossomo. Alinhamento baseado em HMM. Conservação do alinhamento da proteína, comprimento da sequência, tamanho da região alinhada.	RF	Humano e camundongo/ 2015.
LncTar (LI, 2014)	energia livre, tamanho da sequência, alinhamento	PerlPrimer	Humano/ 2015.
LncRScan-SVM (SUN; AL., 2015)	Tamanho do stop códon Média do txcdspredict, comprimento da sequência, contagem dos éxons Alinhamento-Phylo-HMM. Conservação de Proteínas Pontuações médias de PhastCons.	SVM	Humano e camundongo/ 2015.
LncRNApred (PIAN; AL., 2016)	Tamanho e cobertura da ORF. Tamanho da sequência, Relação sinal ruído, esquema K-mer e conteúdo GC.	RF	Humano/ 2016.
DeepLNC (TRIPATHI; AL., 2016)	K-mer (2,3,4,5).	DNN	Humano/ 2016.
lncScore (ZHAO; SONG; WANG, 2016)	Pontuação do Hexâmero (HS), Distância de Pontuação do Hexâmero (HSD), Conteúdo de GC, comprimento da subsequência máxima de codificação, Pontuação de codificação, Porcentagem de pontuação de codificação, dados de ORF (tamanho, Pontuação de Fickett, Pontuação do Hexâmero, Distância de Pontuação do Hexâmero)	LR	Humano e camundongo/ 2016.
PlantRNA Sniffer (VIEIRA; AL., 2017)	ORF tamanho e proporção. e frequência de 10 nucleotídeos padrão.	SVM	lincRNA em cana de açúcar e milho/ 2016.
COME(HU; AL., 2016)	conteúdo do GC, pontuação de conservação de sequência, poli(A)com menos expressão, poli(A)com mais expressão, conservação de estrutura RNA, pequena expressão de RNA	RF	Humano, camundongo, abelha, minhoca e Arabidopsis/ 2016 .
CPC2(KANG *et al.*, 2017)	score Fickett, comprimento de ORF, integridade da ORF e ponto isoelétrico com base na maior ORF	SVM	Humano, camundongo,peixe-zebra, abelha, minhoca e Arabidopsis/ 2017
PlncPRO (SINGH *et al.*, 2017)	tamanho e cobertura da ORF, BLASTX para alinhamento, Swiss-PROT para saber se codifica.	RF	Plantas/ 2017.
FEELnc (WUCHER; AL., 2017)	k-mer 6, dados de ORF(tamanho, inicio, fim)	RF	Humano/ 2017.

continua na próxima pagina

Tabela 2 continuação

Ferramentas	Característica	Modelo	Treinamento/ Ano
RNAplonc (NEGRI *et al.*, 2018)	Características da ORF, K-mer, Conteúdo GC	REPTree	Plantas/ 2018.
LncRNAnet (BAEK; AL., 2018)	Características da ORF	CNN/RNN	Humano e Camundongo/ 2018.
BASiNET (ITO; AL., 2018)	número de nucleotídeos na sequência e k-mer 3	RN	Humano, camundongo,peixe-zebra, abelha, minhoca e Arabidopsis/ 2018.
LncFinder (HAN; AL., 2018)	Orf, hexâmeros, estrutura secundária, energia livre, escalas e das propriedades físico-químicas baseadas em EIIP	SVM	Humano, camundongo e trigo /2018.
LncADeep (YANG; AL., 2018)	Características da ORF, conteúdo GC, e HMMER index.	DBN	Humano e Camundongo/ 2018.
SEEKR (KIRK *et al.*, 2018)	k-mer = 6	LR	Humano/ 2018
CREMA (SIMOPOULOS; WERETILNYK; GOLDING, 2018)	comprimento das sequências, Comprimento ORF, conteúdo GC, Pontuação de Fickett, pontuação hexamérica, alinhamento com banco de dados Swiss-Prot, presença de elemento transponível.	RF	Humano e plantas/ 2018.
PLIT (DESHPANDE; AL., 2019)	tamanho da ORF, cobertura média da ORF, tamanho da sequência, conteúdo GC, pontuação de Fickett e pontuação de Hexamer	RF	plantas/ 2019.
LGC (WANG; AL., 2019)	tamanho da ORF e Conteúdo GC	RF	humano e plantas/ 2019.
lncRNA-LSTM (MENG; AL., 2019)	K-mer= 1 e 2	LSTM	Zea mays/ 2019
CPPred (TONG; LIU, 2019)	comprimento ORF, cobertura ORF, escore de Fickett e escore de Hexamer	SVM	Humano, camundongo,peixe-zebra, mosca da fruta e S. cerevisiae/ 2019
PredLnc-GFStack (LIU *et al.*, 2019)	k-mer, características de ORF, pontuação de Fickett	RF	Humano e camundongo/ 2019
DeepCNPP (ALAM; AL., 2019)	sequência fasta	CNN	Humano/ 2019

continua na próxima pagina

Tabela 2 continuação

Ferramentas	Característica	Modelo	Treinamento/ Ano
ltLnc-BXE (ZHANG; AL., 2019)	características baseadas em sequência, abertura do quadro de leitura (ORF), características baseadas em códons e características baseadas em alinhamento	XGBoost	Plantas / 2019
PLAIDOH (PYFROM; LERNER; PAYTON, 2019)	algoritmos modulares que calculam escores preditivos com base em várias medidas diferentes de controle regulatório da transcrição, interação de proteínas e localização subcelular.	MA	Humano/ 2019
NAMS (SUN; WANG; SUN, 2020)	txCDSpredict para tamanho de ORF.	SVM - DT	Plantas/ 2020
RNAsamba (CAMARGO; AL., 2020)	Dados da Sequência inteira e a maior ORF e assinaturas de codificação de proteína.	IGLOO	Humano/ 2020

FONTE: Adaptado de (NEGRI *et al.*, 2018). DNN: Rede Neural Profunda, CNN: Rede Neural Convolucional , DL: *Deep Learning*, DT: *Decision Tree*, IGLOO: *Slicing the Feature Space to Represent LongSequences*, DBN: Deep belief network, DP: Aprendizado Profundo, RN: Rede Neural, LR: Regressão Logística, LSTM:Redes de memória de longo prazo, RF: Árvore de decisão Aleatória , SVM: Suporte de máquina de vetor, XGBoost: eXtreme Gradient Boosting, REPTree(Reduced Error Pruning Tree) e MA: algoritmo modular.

O PLEK (LI; ZHANG; ZHOU, 2014) (202.200.112.245/plek/), usa o esquema k-mer e janela deslizante para analisar as transcrições. O modelo de classificação do PLEK é o SVM com um *kernel* de função de base radial. Foi utilizado para treinamento dados de Humanos, os transcritos de lncRNAs do GENCODE v17 e os mRNas RefSeq (versão 60). Portanto o PLEK é uma ferramenta valiosa para predição de dados de Humanos, não utiliza de alinhamento como o CPC e o CPC2, tornando-se mais rápida e mais eficaz para identificar lncRNAs em transcritos montados de novo sem genomas de referência.

O CPC2 (KANG *et al.*, 2017) (cpc2.gao-lab.org/), dos mesmos criadores da ferramenta CPC (KONG; AL., 2007), o CPC2 vem com a proposta de ser aproximadamente 1000 vezes mais rápido que o CPC, e com maior precisão, especialmente para lncRNAs. Além disso, o modelo do CPC2 é neutro em relação às espécies, tornando-o viável para transcriptomas de organismos não modelo. O CPC2 usa 4 características intrínsecas: como escore Fickett TESTCODE, comprimento do quadro de leitura aberta (ORF), integridade da ORF e ponto isoelétrico (pI). Embora a pontuação Fickett TESTCODE seja derivada da frequência ponderada de nucleotídeos da transcrição completa inserida, o restante das características (comprimento da ORF, integridade da ORF e ponto isoelétrico) é calculado com base na maior ORF putativa identifi-

cada em sílico. O modelo CPC2 é criado com base SVM, usando o pacote LIBSVM. O CPC2 também tem versão online para arquivos pequenos ou versão local para grandes arquivos.

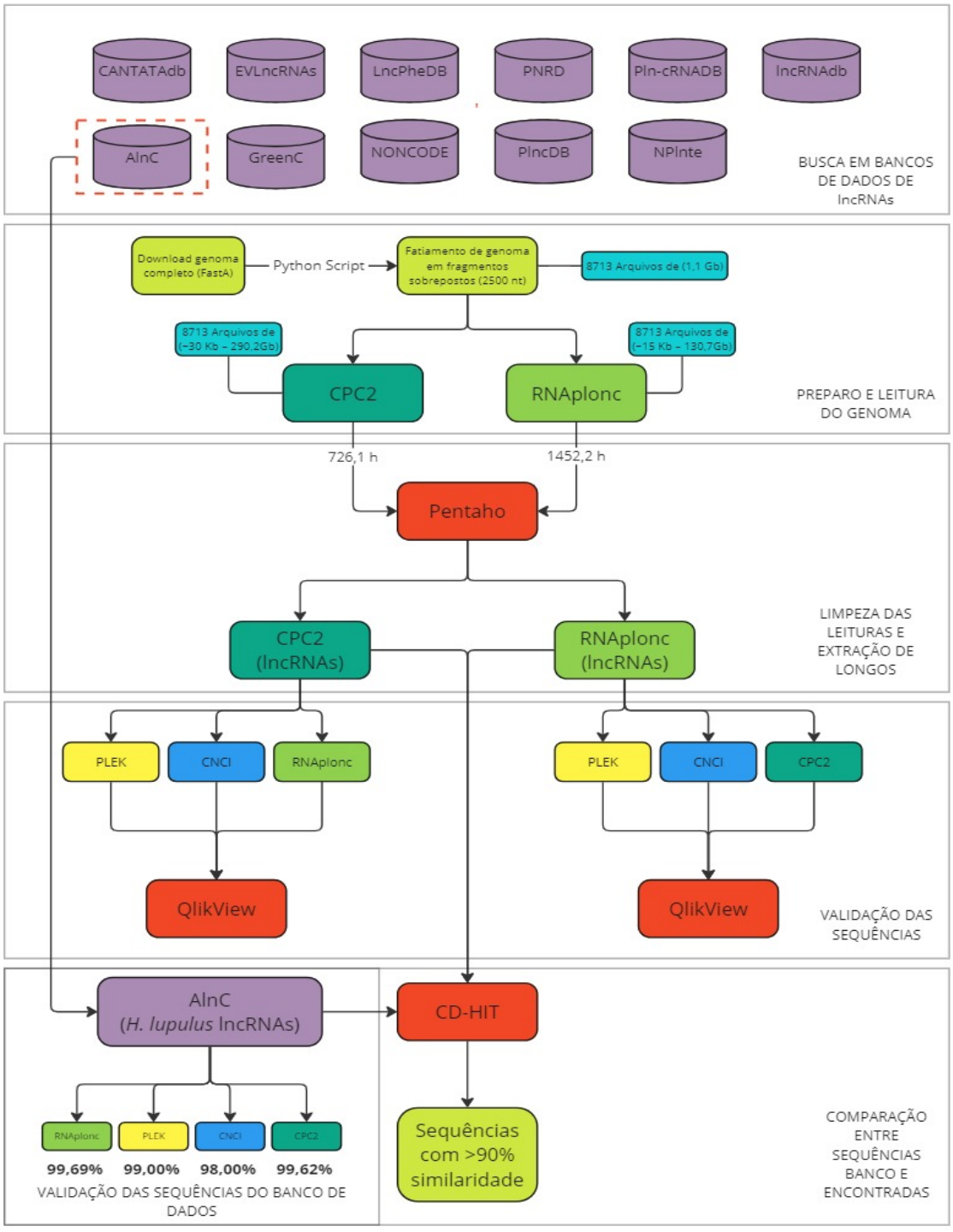
O RNAplonc (NEGRI *et al.*, 2018) (github.com/TatianneNegri/RNAplonc), essa ferramenta utiliza dezesseis características para a classificação, sendo divida em características da sequência (conteúdo GC e tamanho da sequência), características de ORF obtidas do programa txCdsPredict(score de orf, codon de parada, tamanho da orf, e porcentagem de orf) e K-mer (AACG, CCGT, CGCA, CGCT, CGGG, CGTA, TACC, TACG, TCCG, TCGC). O modelo Foi criado utilizando o REPTree. A ferramenta conta com um manual detalhado de instalação e de execução, e ainda tem uma velocidade de execução superior as outras ferramentas executadas.

O CNCI (SUN *et al.*, 2013) (https://github.com/www-bioinfo-org/CNCI) cria perfis baseados em tripletos de nucleotídeos adjacentes (ANT), permitindo uma classificação precisa das sequências RNA. Foi desenvolvido para superar desafios na distinção entre sequências codificadoras de proteínas (cDNA) e não codificadoras (lncRNA), sem depender de anotações previamente conhecidas. Uma característica notável do CNCI é sua capacidade de classificar transcrições como codificadoras ou não codificadoras de proteínas com base unicamente na composição intrínseca da sequência de nucleotídeos, sem a necessidade de anotações de genoma completo.

3 MATERIAIS E MÉTODOS

A Figura 3 abaixo, aprepresenta um esquema de fluxograma demonstrando as etapas as quais foram realizadas as análises até a obtenção das novas sequências de lncRNAs.

Figura 3 – Fluxograma da metodologia



Fonte: Autoria própria (2024).

3.1 Preparo e leitura do genoma

Inicialmente a identificação das novas regiões se deu pela obtenção do genoma de lúpulo, *Humulus lupulus L.*, através do banco de dados National Center for Biotechnology Information (NCBI) sendo fruto do artigo Padgitt-Cobb *et al.* (2023), pelo código de referência PRJNA562558 (ncbi.nlm.nih.gov/assembly/GCA_023660075.1#/def), contendo um total de 3,711,963,939 pares de bases. Além disso realizamos uma busca em bancos de dados de RNAs longos não codificantes em plantas, por lncRNAs já catalogados em lúpulo, sendo eles AlnC (SINGH; VIVEK; KUMAR, 2021), CantataDB (SZCZEŚNIAK; ROSIKIEWICZ; MAKAŁOWSKA, 2016), EVLncRNAs (ZHOU *et al.*, 2023), GreenC (GALLART *et al.*, 2016), LncPheDB (LOU *et al.*, 2022), NONCODE (LIU *et al.*, 2005), PNRD (YI *et al.*, 2015), PlncDB (JIN *et al.*, 2013), PlncRNADB (BAI *et al.*, 2019), NPInter (WU *et al.*, 2006) e lncRNAdb (AMARAL *et al.*, 2011) no entanto apenas o banco AlnC apresentou sequências de lúpulo, Figura 4.

Figura 4 – Sequências de lncRNAs obtidas do banco de dados AlnC

Fonte: Autoria própria (2024).

Posteriormente, ao avaliarmos o banco AlnC, verificamos que a sequência de maior tamanho foi a de acesso AlnC_91911, sendo essa composta por 2405 pares de bases. Esse dado foi importante na durante a tomada decisão para o processo preparo do genoma para análises subsequentes.

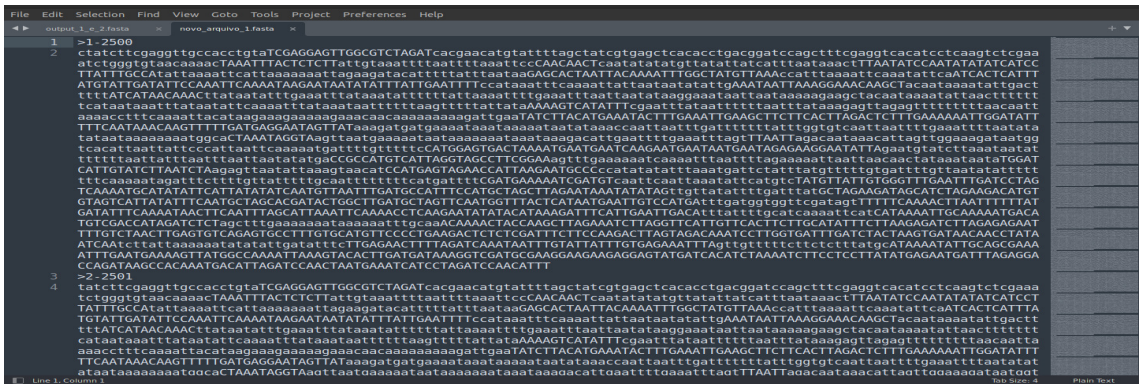
Com base no maior tamanho de fragmento de lncRNA encontrado no banco AlnC, foi desenvolvido, através da linguagem de programação Python um *script* que pudesse fragmentar o genoma em sequências sobrepostas, sendo determinado no *script* o tamanho de sequência de 2500 e as sobreposições de 2499 pb. Em virtude da limitação dos programas de predição escolhidos, os quais não conseguiram processar o genoma por completo ($\approx$ 3,8Gb). Através desse método foi possível ter uma cobertura completa e detalhada do genoma. Abaixo, na figura 2 é possível observar o recorte das posições sobrepostas em um dos arquivo de 1.1Gb.

A sequência AlnC_91911, com 2405 pb, foi a mais longa encontrada no banco AlnC e serviu como referência para determinar o tamanho ideal dos fragmentos a serem gerados para a análise genômica. Com base nesse dado, desenvolvemos um *script* em Python para fragmentar o genoma em sequências sobrepostas de 2500 pb, com uma janela deslizante de 1 pb (Figura 6.

Essa estratégia, denominada ’fatiamento’, foi adotada em virtude das limitações computacionais dos programas de predição, permitindo uma análise abrangente do genoma. A Figura 5 ilustra o processo de fatiamento, mostrando a sobreposição de 1 pb entre os fragmentos de 2500 pb em um arquivo de 1,1 Gb. Essa estratégia garantiu uma cobertura completa do genoma, sem perda de informação.

Figura 5 – Fatiamento do genoma em sobreposição de 2499 pb em duas diferentes visões: (a) Visão bloco, (b) Visão em linhas

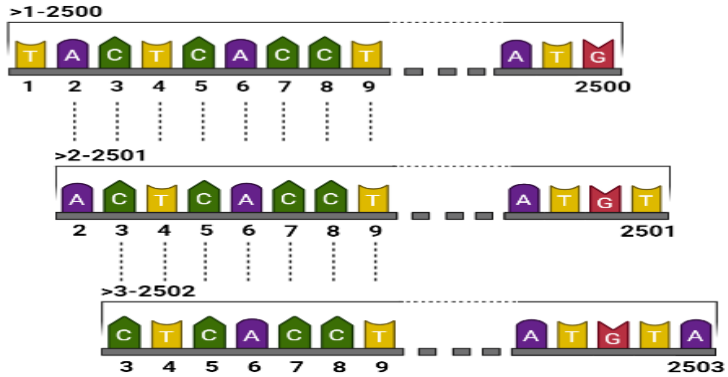
(a) Fatiamento visão em bloco



(b) Fatiamento visão em linhas

Fonte: Autoria própria (2024).

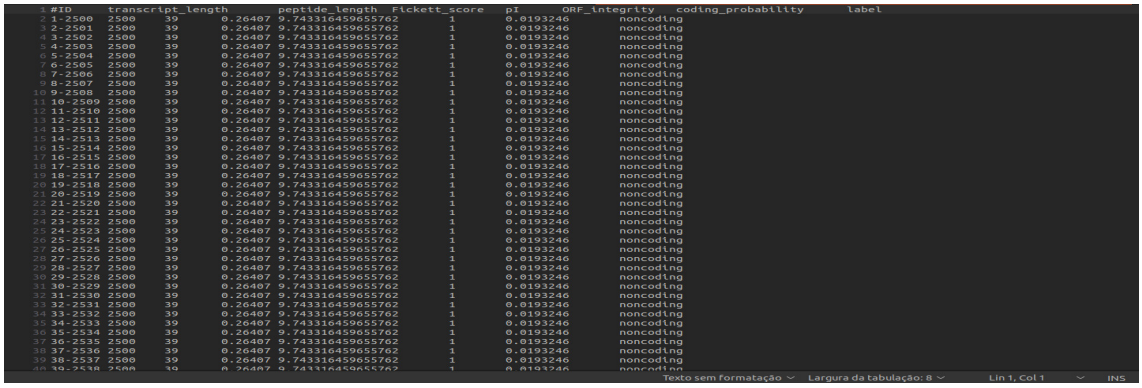
Figura 6 – Representação do fatiamento do genoma em sobreposição



Fonte: Autoria própria (2024).

Após a montagem da biblioteca de fragmentos para análise, composta por 8713 arquivos de 1.1Gb, utilizando os sistemas de predição CPC2 (KANG *et al.*, 2017) e RNAplonc (NEGRI *et al.*, 2018) com todas as configurações padrões, foram realizadas as análises de predição das sequências. Ao final, os sistemas retornavam, sequências catalogadas em lncRNA e mRNA (RNAplonc) e *coding* e *non-coding* (CPC2). As Figuras 7 e 8 ilustram o resultado de cada programa. Além disso, para otimizar o processamento, utilizamos 7 terminais de comando em paralelo, explorando ao máximo a capacidade computacional da máquina, configurada com um processador Intel Core i3-12100F de 4 núcleos e 8 threads, 16 GB de RAM DDR4 a 3200MHz e um armazenamento híbrido de 4,2 TB (HDD/SSD).

Figura 7 – Resultados após processamento do arquivos de sobreposição retornados por CPC2



#ID	transcript_length	peptide_length	Fickett_score	pI	ORF_integrity	coding_probability	label
1-2500	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
2-2501	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
3-2502	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
4-2503	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
5-2504	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
6-2505	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
7-2506	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
8-2507	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
9-2508	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
10-2509	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
11-2510	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
12-2511	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
13-2512	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
14-2513	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
15-2514	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
16-2515	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
17-2516	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
18-2517	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
19-2518	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
20-2519	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
21-2520	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
22-2521	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
23-2522	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
24-2523	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
25-2524	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
26-2525	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
27-2526	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
28-2527	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
29-2528	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
30-2529	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
31-2530	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
32-2531	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
33-2532	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
34-2533	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
35-2534	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
36-2535	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
37-2536	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding
38-2537	2500	39	0.26407	9.743316459655762	1	0.0193246	noncoding

Texto sem formatação | Largura da tabulação: 8 | Lin 1, Col 1 | INS

Fonte: Autoria própria (2024).

Figura 8 – Resultados após processamento do arquivos de sobreposição retornados por RNAplonc

	Seq	predicted	prediction
1-2500	1:lncRNA	0.920	
2-2501	1:lncRNA	0.920	
3-2502	1:lncRNA	0.920	
4-2503	1:lncRNA	0.920	
5-2504	1:lncRNA	0.920	
6-2505	1:lncRNA	0.920	
7-2506	1:lncRNA	0.920	
8-2507	1:lncRNA	0.920	
9-2508	1:lncRNA	0.920	
10-2509	1:lncRNA	0.920	
11-2510	1:lncRNA	0.920	
12-2511	1:lncRNA	0.920	
13-2512	1:lncRNA	0.920	
14-2513	1:lncRNA	0.920	
15-2514	1:lncRNA	0.920	
16-2515	1:lncRNA	0.920	
17-2516	1:lncRNA	0.920	
18-2517	1:lncRNA	0.920	
19-2518	1:lncRNA	0.920	
20-2519	1:lncRNA	0.920	
21-2520	1:lncRNA	0.920	
22-2521	1:lncRNA	0.920	
23-2522	1:lncRNA	0.920	
24-2523	1:lncRNA	0.920	
25-2524	1:lncRNA	0.920	
26-2525	1:lncRNA	0.920	
27-2526	1:lncRNA	0.920	
28-2527	1:lncRNA	0.920	
29-2528	1:lncRNA	0.920	
30-2529	1:lncRNA	0.920	
31-2530	1:lncRNA	0.920	
32-2531	1:lncRNA	0.920	
33-2532	1:lncRNA	0.920	
34-2533	1:lncRNA	0.920	
35-2534	1:lncRNA	0.920	
36-2535	1:lncRNA	0.920	
37-2536	1:lncRNA	0.920	
38-2537	1:lncRNA	0.920	

Texto sem formatação | Largura da tabulação: 8 | Lin 1, Col 1 | INS

Fonte: Autoria própria (2024).

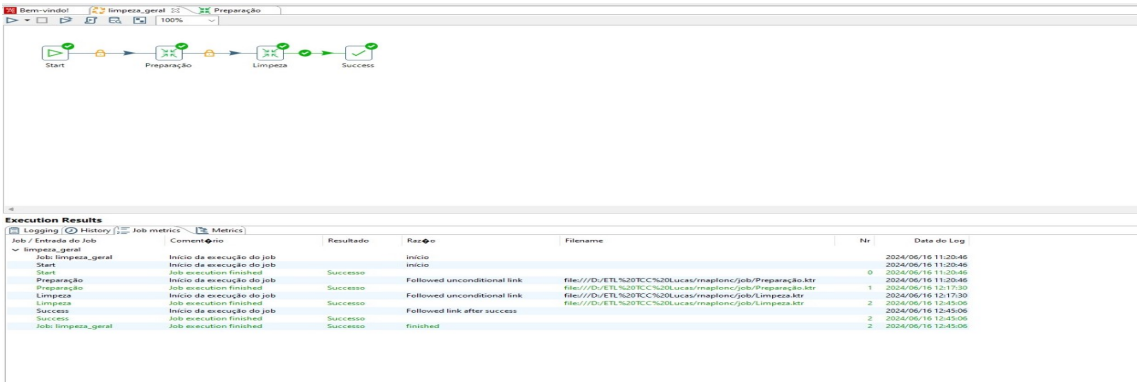
3.2 Limpeza das leituras e extração de longos

A fim de remover a redundância causada pela sobreposição de 2499 pares de bases e isolar as sequências classificadas como lncRNA e *non-coding* (Figura 9), empregamos a ferramenta de *Business Intelligence* Pentaho Data Integration CE (Hitachi® and Hitachi Vantara® Ltd. in the U.S.), através de um processo de agrupamento baseado em classificação e posição.

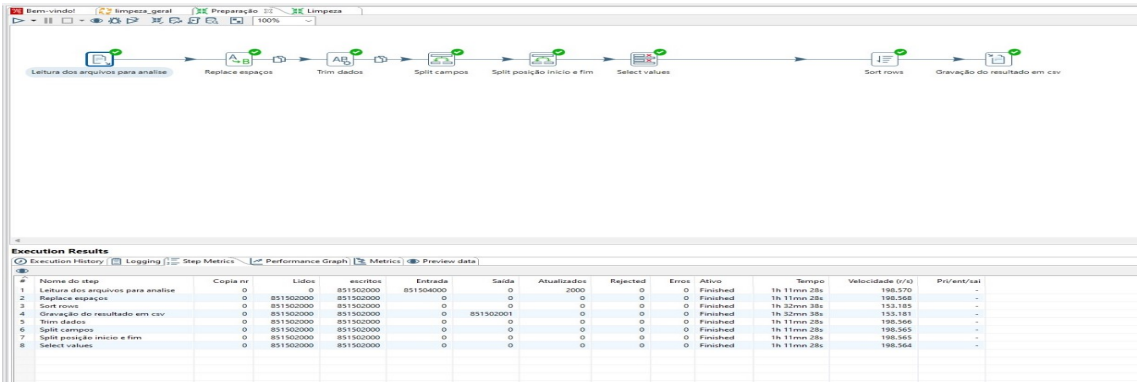
O esquema de ETL (*Extract, Transform, Load*) utilizado para agrupamento é demonstrado na Figura 10.

Figura 9 – ETL montada para agrupamento e filtragem dos resultados: (a) Primeira etapa, (b) Segunda etapa e (c) Terceira etapa

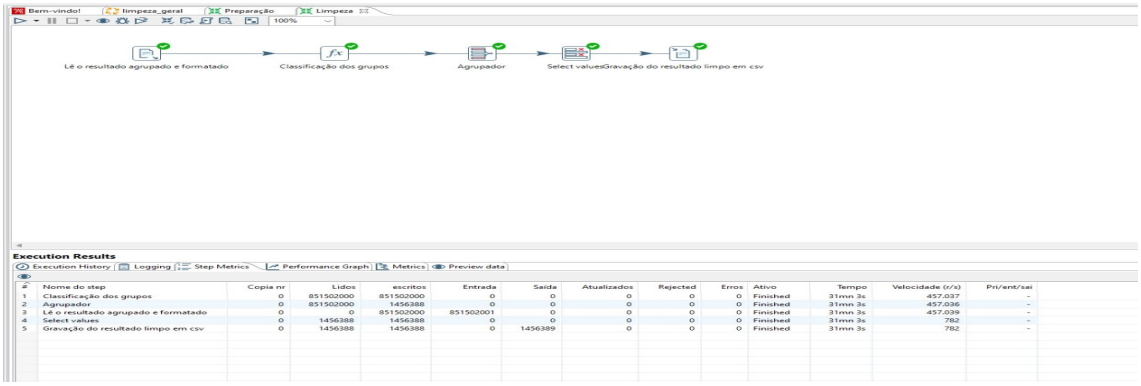
(a) Primeira etapa



(b) Segunda etapa



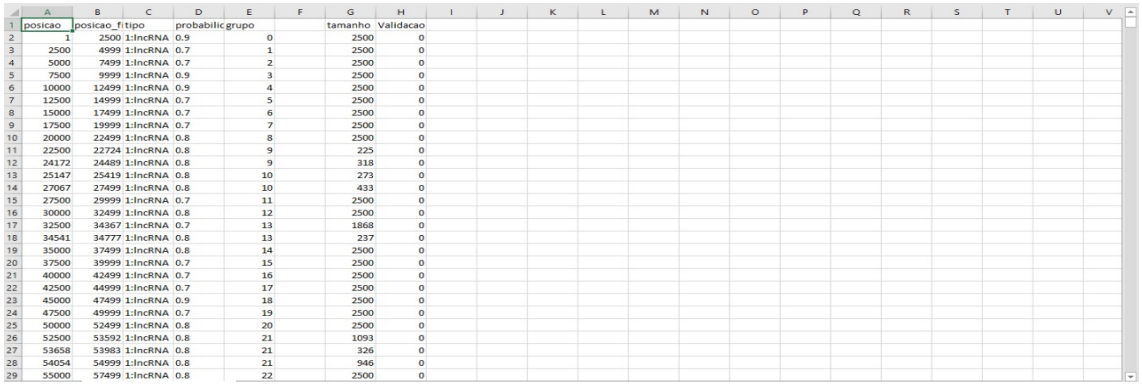
(c) Terceira etapa



Fonte: Autoria própria (2024).

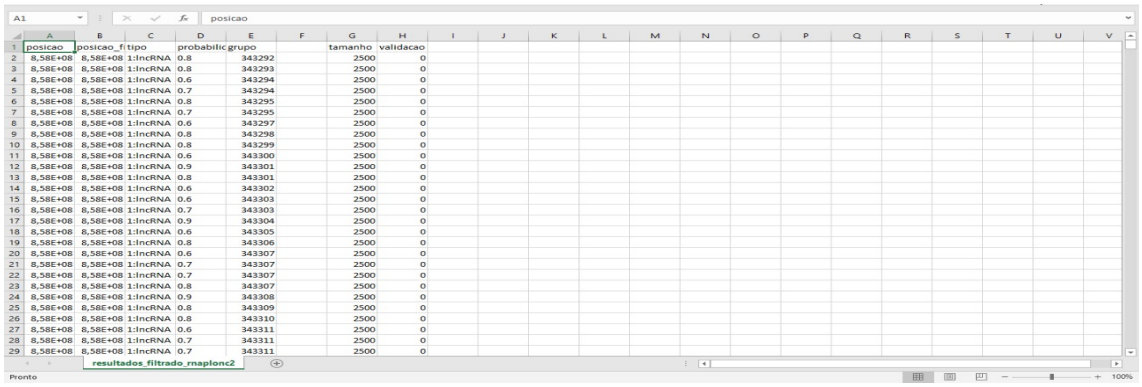
Figura 10 – Sequências sem redundantes, agrupadas e filtradas em relação a tamanho de fragmento e tipo: (a) Posições iniciais e (b) posições finais do mesmo documento

(a) Região das posições iniciais



posicao	posicao_f	tipo	probabilic	grupo	tamanho	Validacao
1	2500	1:lncRNA	0.9	0	2500	0
2500	4999	1:lncRNA	0.7	1	2500	0
5000	7499	1:lncRNA	0.7	2	2500	0
7500	9999	1:lncRNA	0.9	3	2500	0
10000	12499	1:lncRNA	0.9	4	2500	0
12500	14999	1:lncRNA	0.7	5	2500	0
15000	17499	1:lncRNA	0.7	6	2500	0
17500	19999	1:lncRNA	0.7	7	2500	0
20000	22499	1:lncRNA	0.8	8	2500	0
22500	22724	1:lncRNA	0.8	9	225	0
24172	24489	1:lncRNA	0.8	9	318	0
25147	25419	1:lncRNA	0.8	10	273	0
27067	27499	1:lncRNA	0.8	10	433	0
27500	29999	1:lncRNA	0.7	11	2500	0
30000	32499	1:lncRNA	0.8	12	2500	0
32500	34367	1:lncRNA	0.7	13	1868	0
34541	34777	1:lncRNA	0.8	13	237	0
35000	37499	1:lncRNA	0.8	14	2500	0
37500	39999	1:lncRNA	0.7	15	2500	0
40000	42499	1:lncRNA	0.7	16	2500	0
42500	44999	1:lncRNA	0.7	17	2500	0
45000	47499	1:lncRNA	0.9	18	2500	0
47500	49999	1:lncRNA	0.7	19	2500	0
50000	52499	1:lncRNA	0.8	20	2500	0
52500	53592	1:lncRNA	0.8	21	1093	0
53658	53983	1:lncRNA	0.8	21	326	0
54054	54999	1:lncRNA	0.8	21	946	0
55000	57499	1:lncRNA	0.8	22	2500	0

(b) Região das posições finais



posicao	posicao_f	tipo	probabilic	grupo	tamanho	validacao
8,58E+08	8,58E+08	1:lncRNA	0.8	343292	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343293	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343294	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343294	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343295	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343295	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343297	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343298	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343299	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343300	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.9	343301	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343301	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343302	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343303	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343303	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.9	343304	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343305	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343306	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343307	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343307	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343307	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343307	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.9	343308	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343309	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.8	343310	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.6	343311	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343311	2500	0
8,58E+08	8,58E+08	1:lncRNA	0.7	343311	2500	0

Fonte: Autoria própria (2024).

3.3 Validação das Sequências

Para avaliar a confiabilidade das sequências preditas como lncRNAs, adotamos uma abordagem de consenso, similar à proposta em estudos anteriores Nath *et al.* (2021); (BARUAH *et al.*, 2021), bem como validação de bancos de dados de lncRNAs (SINGH; VIVEK; KUMAR, 2021). Essa abordagem atribui níveis de confiança às predições com base no grau de concordância entre diferentes ferramentas de predição.

No nosso estudo, as sequências inicialmente identificadas como não codificantes pelo CPC2 foram submetidas a uma análise adicional nos programas RNAplonc, CNCI e PLEK. Analogamente, as sequências identificadas como lncRNAs pelo RNAplonc foram avaliadas pelos programas CPC2, CNCI e PLEK.

Baseando-se no método de classificação de Gallart *et al.* (2016), consideramos mais confiáveis as sequências identificadas por um maior número de ferramentas, indicando um consenso entre os métodos de predição. Por outro lado, sequências identificadas por apenas uma ferramenta foram classificadas como menos confiáveis. Essa abordagem em múltiplos passos

nos permitiu refinar a identificação de lncRNAs e fornecer uma estimativa mais precisa da confiabilidade das nossas predições.

3.4 Comparação entre sequências do banco e sequências encontradas

Por fim, através do CD-HIT (FU *et al.*, 2012), um clusterizador amplamente utilizado para comparar sequências biológicas, as agrupando sequências com determinada similaridade, foram comparadas as sequências de lncRNA obtidas pelos sistemas de predição com as sequências previamente encontradas no banco de dados AlnC, demonstrado na Figura 11.

Figura 11 – Sequências sem clusterizadas em função da similaridade com as sequências encontradas no banco de dados AlnC

```
>Cluster 1
0   2405nt, >AlnC_191911_Core... *
1   364nt, > 528570000-52857036... at +/98.63%
>Cluster 10
0   1954nt, >AlnC_192415_Core... *
1   284nt, > 216539716-21653999... at -/97.89%
>Cluster 32
0   1570nt, >AlnC_193133_Core... *
1   321nt, > 420757179-42075749... at -/99.69%
>Cluster 37
0   1523nt, >AlnC_192019_Core... *
1   242nt, > 142002500-14200274... at +/90.08%
2   224nt, > 154272500-15427272... at +/93.75%
3   255nt, > 186599604-18659985... at -/90.98%
4   500nt, > 381370000-38137049... at -/91.00%
5   210nt, > 412924790-41292499... at -/90.48%
6   291nt, > 554665000-55466529... at -/92.78%
7   264nt, > 559465000-55946526... at -/92.80%
8   253nt, > 781707197-78170744... at -/91.70%
9   400nt, > 781757500-78175789... at +/90.75%
10  971nt, > 827392500-82739347... at -/90.53%
>Cluster 45
0   1437nt, >AlnC_192015_Core... *
1   793nt, > 841268367-84126915... at +/97.60%
>Cluster 63
0   1314nt, >AlnC_192550_Core... *
1   318nt, > 161415000-16141531... at -/93.71%
>Cluster 78
0   1250nt, >AlnC_200111_Core... *
1   238nt, > 730044619-73004485... at -/100.00%
```

Fonte: Autoria própria (2024).

3.5 Validação das Sequências do Banco de Dados

Para confirmar a confiabilidade das predições iniciais, as sequências de lncRNAs de lúpulo identificadas no banco de dados AlnC foram reavaliadas utilizando os softwares CPC2, PLEK, RNAplonc e CNCI.

4 RESULTADOS E DISCUSSÃO

4.1 Longos identificados

Este estudo teve como objetivo identificar e caracterizar longos RNAs não codificantes (lncRNAs) no genoma de *Humulus lupulus*. A identificação foi realizada utilizando métodos quantitativos, com o intuito de explorar e caracterizar regiões genômicas ainda não descritas. Utilizando ferramentas de bioinformática, como os softwares CPC2 e RNAplonc – preditores de lncRNAs – realizamos uma análise abrangente do genoma de Humulus lupulus. A análise resultou na identificação de 3.582.809 transcritos classificados como lncRNAs pelo software CPC2, enquanto o RNAplonc catalogou 2.263.355. A discrepância entre os resultados pode ser explicada por diferenças nas metodologias e critérios de classificação empregados por cada software.

A maioria dos estudos anteriores (WANG *et al.*, 2024; ZHU *et al.*, 2014; NATH *et al.*, 2021; JAIN *et al.*, 2017; DI *et al.*, 2014; CUI *et al.*, 2017; BARUAH *et al.*, 2021), incluindo os protocolos da EMBRAPA (CAMPOS *et al.*, 2023), emprega dados de RNA-seq para a identificação de lncRNAs. No entanto, na ausência desses dados para o lúpulo, propusemos um novo método de análise. A partir do genoma completo, geramos sequências sintéticas que mimetizam os dados de RNA-seq (invertendo a metodologia convencional),permitindo assim a identificação de lncRNAs.

Comparada a estudos anteriores, nossa metodologia apresenta um fluxo de trabalho mais enxuto para a identificação de lncRNAs de alta confiabilidade. Ao reduzir o número de etapas subsequentes à identificação inicial, evitamos a perda de candidatos promissores e aumentamos a eficiência do processo.

4.2 Desempenho de Ferramentas

Inicialmente, planejamos avaliar as sequências em cada programa de forma individual para comparar os resultados. Entranto considerando o tempo de processamento individual de cada arquivo e o imenso volume de dados (8.713 arquivos de 1,1 Gb – 8,7 Tb), uma análise sequencial em todos os programas seria computacionalmente inviável. Para superar essa limitação, adotamos uma abordagem otimizada, priorizando as ferramentas mais rápidas e realizando análises cruzadas para garantir a cobertura de todos os dados. A Tabela 3 demonstra o tempo utilizado por cada durante a avaliação.

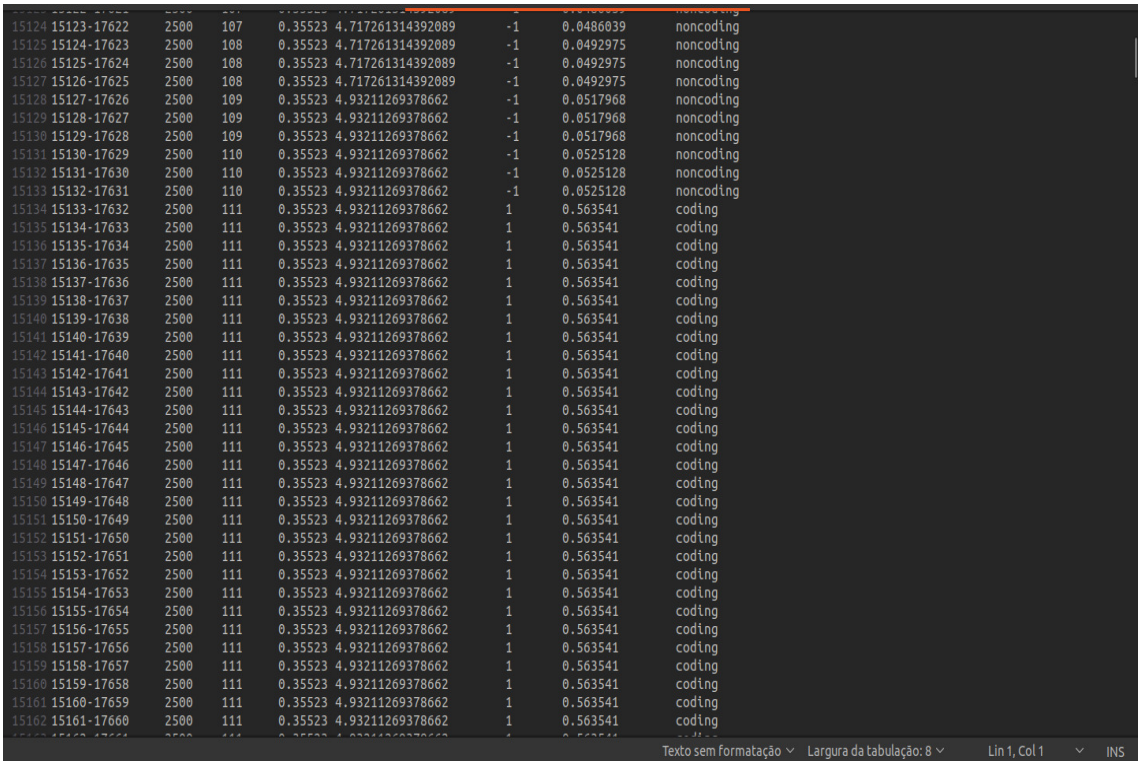
Tabela 3 – Tabela com os tempos necessário para processamento dos arquivos

Ferramenta	Tempo para processamento de um arquivo (1.1Gb)	Tempo total para toda análise
CPC2	5 min	726,1 h
RNAplonc	10 min	1452,2 h
PLEK	> 1 h	> 8713 h
CNCI	> 1 h	> 8713 h

Fonte: Autoria própria (2024).

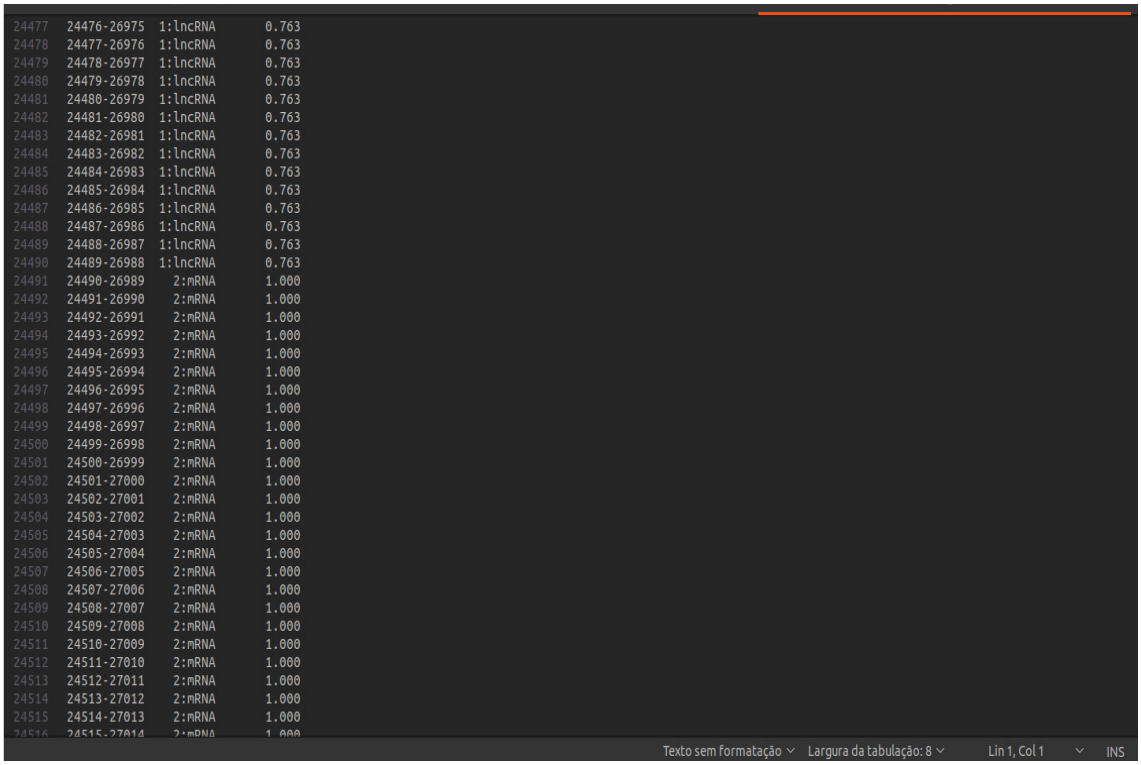
Em virtude do uso de diversos terminais em paralelo, houve uma redução drástica no número de horas necessárias para cada análise, sendo CPC2 um total de 103,7 h enquanto que para RNAplonc foram 209,1 h, entretanto, mesmo utilizando essa abordagem, não foi possível utimizar o processamento de PLEK e CNCI. Além disso, vale ressaltar que as análises foram realizadas em um ambiente *dual boot*, com sistemas operacionais Ubuntu 22.04 LTS e Windows 11 23H2 onde cada sistema operacional operou em diferentes partes do projeto. Diante da grande quantidade de dados gerados, dividimos o processamento em lotes de 2 Tb para otimizar o uso do espaço em disco. Após a conclusão das análises, obtivemos um volume total de dados brutos de 130,7 Gb para o RNAplonc e 290,2 Gb para o CPC2. A diferença no tamanho dos arquivos pode ser explicada pela maior quantidade de informações detalhadas geradas pelo CPC2 ao final da análise, uma maior quantidade de caracteres exige um nível de maior de armazenamento interno na máquina. As Figuras 12 e 13 exemplificam essa quantidade de caracteres.

Figura 12 – Resultados de CPC2



Fonte: Autoria própria (2024).

Figura 13 – Resultados de RNAplonc



24477	24476-26975	1:lncRNA	0.763
24478	24477-26976	1:lncRNA	0.763
24479	24478-26977	1:lncRNA	0.763
24480	24479-26978	1:lncRNA	0.763
24481	24480-26979	1:lncRNA	0.763
24482	24481-26980	1:lncRNA	0.763
24483	24482-26981	1:lncRNA	0.763
24484	24483-26982	1:lncRNA	0.763
24485	24484-26983	1:lncRNA	0.763
24486	24485-26984	1:lncRNA	0.763
24487	24486-26985	1:lncRNA	0.763
24488	24487-26986	1:lncRNA	0.763
24489	24488-26987	1:lncRNA	0.763
24490	24489-26988	1:lncRNA	0.763
24491	24490-26989	2:mRNA	1.000
24492	24491-26990	2:mRNA	1.000
24493	24492-26991	2:mRNA	1.000
24494	24493-26992	2:mRNA	1.000
24495	24494-26993	2:mRNA	1.000
24496	24495-26994	2:mRNA	1.000
24497	24496-26995	2:mRNA	1.000
24498	24497-26996	2:mRNA	1.000
24499	24498-26997	2:mRNA	1.000
24500	24499-26998	2:mRNA	1.000
24501	24500-26999	2:mRNA	1.000
24502	24501-27000	2:mRNA	1.000
24503	24502-27001	2:mRNA	1.000
24504	24503-27002	2:mRNA	1.000
24505	24504-27003	2:mRNA	1.000
24506	24505-27004	2:mRNA	1.000
24507	24506-27005	2:mRNA	1.000
24508	24507-27006	2:mRNA	1.000
24509	24508-27007	2:mRNA	1.000
24510	24509-27008	2:mRNA	1.000
24511	24510-27009	2:mRNA	1.000
24512	24511-27010	2:mRNA	1.000
24513	24512-27011	2:mRNA	1.000
24514	24513-27012	2:mRNA	1.000
24515	24514-27013	2:mRNA	1.000

Fonte: Autoria própria (2024).

Devido às limitações computacionais dos softwares empregados, a análise do genoma completo não era viável. Assim, optou-se por uma abordagem fragmentada, processando o genoma em partes menores, o que permitiu contornar as restrições de memória e tempo de processamento e garantindo uma análise confiável. Essa abordagem inovadora possibilitou a descoberta de um grande número de novos lncRNAs no genoma do lúpulo. A caracterização destes elementos genômicos pode fornecer insights importantes sobre os mecanismos regulatórios que controlam a expressão gênica e o desenvolvimento da planta.

Seguindo os resultados de Tian *et al.* (2024), que comparou o desempenho de RNAplonc e CPC2 e outros preditores de lncRNAs em 20 conjuntos de dados de plantas, observamos que RNAplonc apresentou um bom equilíbrio entre sensibilidade, especificidade e acurácia na identificação de lncRNAs. Embora CPC2 tenha se mostrado altamente específico, sua eficácia foi comprometida por pontuações mais baixas em outras métricas relevantes, como a sensibilidade. Diante disso, para validar os resultados obtidos com essas ferramentas, utilizamos CNCI e PLEK, que, de acordo com Tian *et al.* (2024), apresentam alta acurácia, especificidade e precisão, metodologia similar a usada por outros autores como Nath *et al.* (2021), Cui *et al.* (2017), Wang *et al.* (2024) e Baruah *et al.* (2021) o qual após obter sequência dos transcritos usou CPC2, PLEK e CNIT (GUO *et al.*, 2019) como método de filtragem. Essa etapa de validação foi fundamental para confirmar a confiabilidade das sequências de lncRNAs identificadas tanto por RNAplonc quanto por CPC2. Também sugerimos para análises posteriores o uso de

CPAT (WANG; AL., 2013) e LncFinder (HAN *et al.*, 2019), o quais, também apresentaram ótimos desempenhos em conjuntos de dados de plantas.

4.3 Fatiamento e Validação de Sequências

O método de fragmentação empregado, baseado em sobreposição, gerou um número significativo de sequências redundantes. Essa estratégia, embora essencial para garantir a cobertura completa do genoma, resultou em um conjunto de dados consideravelmente maior, demandando um tratamento especial para a identificação de lncRNAs únicos. Tradicionalmente, a ferramenta CD-HIT (FU *et al.*, 2012) é utilizada para clusterização de sequências redundantes. No entanto, dada a escala do nosso conjunto de dados, optamos pela ferramenta de *Business Intelligence* Pentaho Data Integration CE, que demonstrou maior eficiência e flexibilidade para lidar com grandes volumes de informações.

A remoção das redundâncias foi realizada em duas etapas:

- **Agrupamento por posição:** As sequências foram agrupadas em fragmentos de 2.500 bases, facilitando a identificação de regiões genômicas distintas.
- **Agrupamento por predição:** Dentro de cada fragmento, as sequências foram agrupadas com base na predição dos softwares (mRNA ou lncRNA), evitando a mistura de classes funcionais.

Para refinar o conjunto de fragmentos identificados, aplicamos um filtro rigoroso, considerando apenas aqueles com mais de 200 pares de bases e classificados como lncRNAs tanto pelo RNAplonc quanto pelo CPC2. Essa abordagem, em linha com estudos prévios de Baruah *et al.* (2021) e Wang *et al.* (2024), visa garantir que os lncRNAs selecionados atendam a critérios de qualidade amplamente aceitos na literatura. É importante destacar que, embora estudos como o de Wang *et al.* (2024) incluam etapas adicionais de classificação, nosso objetivo neste trabalho foi a identificação inicial de lncRNAs candidatos, sem aprofundar em análises funcionais ou utilizar dados de expressão gênica

Para validar a confiabilidade das sequências de lncRNA identificadas, submetemos os resultados dos softwares CPC2 e RNAplonc a uma análise de validação cruzada utilizando as ferramentas PLEK, CNCI e RNAplonc para os dados obtidos de CPC2 e PLEK, CNCI e CPC2 para os dados de RNAplonc. Essa abordagem permitiu avaliar a robustez das predições e estimar a proporção de falsos positivos. As sequências foram classificadas em quatro categorias de acordo com o número de ferramentas que confirmaram a predição:

- Três confirmações: Sequências validadas pelos três programas.
- Duas confirmações: Sequências validadas por dois dos três programas.
- Uma confirmação: Sequências validadas por apenas um programa.

- Nenhuma confirmação: Sequências não validadas por nenhum dos programas.

As Figuras 13 e 14 apresentam a distribuição das sequências nas diferentes categorias de validação, evidenciando a concordância entre os resultados dos diferentes softwares e a qualidade geral do conjunto de dados de lncRNAs identificados.

Figura 14 – Gráfico de validação CPC2. Eixo X representa a posição no genoma a cada 10 milhões de pb e eixo Y a quantidade de longos encontrados em cada região. As cores de cada linhas, azul, vermelho, verde e amarelo representam o número de lncRNAs validados pelos programas sendo eles zero, uma, duas e três validações respectivamente. Clique aqui para visualizar as imagens em tamanho em tamanho real

Fonte: Autoria própria (2024).

Figura 15 – Gráfico de validação RNAplonc. Eixo X representa a posição no genoma a cada 10 milhões de pb e eixo Y a quantidade de longos encontrados em cada região. As cores de cada linhas, azul, vermelho, verde e amarelo representam o número de lncRNAs validados pelos programas sendo eles zero, uma, duas e três validações respectivamente.Clique aqui para visualizar as imagens em tamanho em tamanho real

Fonte: Autoria própria (2024).

A fim de identificar similaridades entre as sequências geradas pelo fatiamento e as presentes no banco de dados AlnC, empregamos o CD-HIT-EST-2D com configuração padrão (90% de identidade). A análise resultou em 53 grupos de sequências (clusters), revelando a existência de sequências com diferentes níveis de similaridade dentro de cada cluster. A baixa quantidade de sequências similares ao banco pode ser explicada devido ao alto nível de similaridade que as configurações padrões trazem. A diminuição da porcentagem de lncRNAs identificados por CD-HIT pode ser explicada por diversos fatores. As diferenças genéticas entre a cultivar de lúpulo utilizada no banco de dados e a cultivar estudada neste trabalho podem influenciar a abundância e a distribuição dos lncRNAs. Além disso, a metodologia de fragmentação do genoma em segmentos de 2500 bases pode ter subestimado o número de lncRNAs menores, que podem estar contidos dentro desses fragmentos maiores assim, impedindo que CD-HIT alinhasse corretamente as sequências. A predição de lncRNAs dentro de um fragmento não implica necessariamente que todo o fragmento seja um lncRNA completo. E por último a diminuição nos parâmetros de identidade poderiam ter afetado positivamente o alinhamento das novas sequências.

Com o objetivo de validar e comparar a eficácia das diferentes ferramentas, reavaliamos as sequências encontradas no banco de dados AlnC. A Tabela 4 apresenta a porcentagem de sequências preditivas como lncRNAs, calculada pela razão entre o número de sequências classificadas como lncRNAs e o número total de sequências analisadas em cada ferramenta.

Tabela 4 – Tabela com os tempo necessário para processamento dos arquivos

Ferramenta	Confirmação de lncRNAs (%)
CPC2	99,62
RNAplonc	99,69
PLEK	99
CNCI	98

Fonte: Autoria própria (2024).

A presença de ruídos nas sequências impactou diretamente o desempenho dos programas CNCI e RNAplonc, que não conseguiram processar a totalidade dos dados (8.576 sequências), descartando as sequências com erros. PLEK e CPC2, por sua vez, mostraram-se mais vuneráveis, processando todas as sequências mesmo aquelas apresentando ruídos. Os erros presentes no sequenciamento dessas leituras podem levar a interpretações incorretas pelos programas de análise, comprometendo a precisão dos resultados.

5 CONCLUSÕES

Este estudo teve como objetivo identificar e caracterizar longos RNAs não codificantes (lncRNAs) no genoma do lúpulo, uma planta de grande importância econômica para a indústria cervejeira. Para tanto, desenvolvemos um *pipeline* computacional não convencional, baseado no método de 'fatiamento genômico', que demonstrou ser altamente eficiente na identificação de lncRNAs a partir de dados genômicos. Esse método mostrou-se ser tão efificiente quanto os métodos baseados em RNA-seq, o quais, apresentam um número baixo de identificações de lncRNAs quando comparados ao método proposto neste trabalho, em virtude do alto custo para o sequênciamento em RNA-seq de regiões genômicas.

Adicionalmente, após a comparação exaustiva com o banco de dados AlnC, identificamos apenas 53 sequências com similaridade significativa, o que representa uma pequena porcentagem do total de transcritos não codificantes identificados. Esse resultado sugere que a grande maioria das sequências identificadas neste estudo representa novos lncRNAs, não descritos anteriormente na literatura para a espécie *Humulus lupulus*. Essa descoberta amplia significativamente o nosso conhecimento sobre o transcriptoma não codificante do lúpulo. A validação cruzada das predições, utilizando diferentes ferramentas bioinformáticas, conferiu robustez e confiança aos nossos resultados.

Os lncRNAs identificados neste estudo representam um rico recurso para futuras pesquisas, abrindo novas perspectivas para o melhoramento genético do lúpulo. A compreensão da função desses elementos regulatórios pode levar ao desenvolvimento de variedades mais produtivas, resistentes a doenças e com características agronômicas superiores.

No entanto, é importante ressaltar que este estudo apresenta algumas limitações. A validação experimental dos lncRNAs identificados é fundamental para confirmar sua existência e função biológica, neste sentido, este trabalho pode servir como base para esses estudos futuros.

REFERÊNCIAS

ACHAWANANTAKUN, R.; AL. et. Lncrna-id: Long non-coding rna identification using balanced random forests. **Bioinformatics**, v. 31, n. 24, p. 3897–3905, Dec 2015.

ALAM, T.; AL. et. Deepcnpp: Deep learning architecture to distinguish the promoter of human long non-coding rna genes and protein-coding genes. **Studies in health technology and informatics**, v. 262, p. 232–235, Jul 2019.

ALMAGUER, C. *et al.* Humulus lupulus–a story that begs to be told. a review. **Journal of the Institute of Brewing**, Wiley Online Library, v. 120, n. 4, p. 289–314, 2014.

ALTSCHUL, S. F. *et al.* Basic local alignment search tool. **Journal of molecular biology**, Elsevier, v. 215, n. 3, p. 403–410, 1990.

AMARAL, P. P. *et al.* lncrnadb: a reference database for long noncoding rnas. **Nucleic acids research**, Oxford University Press, v. 39, n. suppl_1, p. D146–D151, 2011.

BAEK, J.; AL. et. lncrnanet: Long non-coding rna identification using deep learning. **Bioinformatics**, v. 1, p. 9, 2018.

BAI, Y. *et al.* Plncrnadb: a repository of plant lncrnas and lncrna-rbp protein interactions. **Current Bioinformatics**, Bentham Science Publishers, v. 14, n. 7, p. 621–627, 2019.

BARUAH, P. M. *et al.* Identification and functional analysis of drought responsive lncrnas in tea plant. **Plant gene**, Elsevier, v. 27, p. 100311, 2021.

BATISTA, P. J.; CHANG, H. Y. Long noncoding rnas: cellular address codes in development and disease. **Cell**, Elsevier, v. 152, n. 6, p. 1298–1307, 2013.

BEERMANN, J. *et al.* Non-coding rnas in development and disease: background, mechanisms, and therapeutic approaches. **Physiological reviews**, American Physiological Society Bethesda, MD, 2016.

BERBERT, S. Conheça a produção de lúpulo brasileiro. **Globo Rural**, fevereiro 2017. Disponível em: https://revistagloborural.globo.com/Noticias/Agricultura/noticia/2017/02/conheca-producao-de-lupulo-brasileiro.html.

BHATIA, G. *et al.* Present scenario of long non-coding rnas in plants. **Non-coding RNA**, MDPI, v. 3, n. 2, p. 16, 2017.

BIRNEY, E.; AL. et. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. **Nature**, v. 447, n. 7146, p. 799–816, Jun 2007.

BOCQUET, L. *et al.* Humulus lupulus l., a very popular beer ingredient and medicinal plant: Overview of its phytochemistry, its bioactivity, and its biotechnology. **Phytochemistry reviews**, Springer, v. 17, n. 5, p. 1047–1090, 2018.

BROSNAN, C. A.; VOINNET, O. The long and the short of noncoding rnas. **Current opinion in cell biology**, Elsevier, v. 21, n. 3, p. 416–425, 2009.

BUDAK, H.; KAYA, S. B.; CAGIRICI, H. B. Long non-coding rna in plants in the era of reference sequences. **Frontiers in Plant Science**, Frontiers Media SA, v. 11, p. 276, 2020.

CAMARGO, A. P.; AL. et. Rnasamba: neural network-based assessment of the protein-coding potential of rna sequences. **NAR Genomics and Bioinformatics**, v. 2, n. 1, Jan 2020.

CAMPOS, F. G. *et al.* **Manual para identificação de RNAs longos não codificantes (lncRNAs) por meio de ferramentas bioinformáticas na análise de transcriptomas**. Concórdia, SC: Embrapa Suínos e Aves, 2023.

CARVALHO, N. B. *et al.* Characterization of the consumer market and motivations for the consumption of craft beer. **British Food Journal**, Emerald Publishing Limited, v. 120, n. 2, p. 378–391, 2018.

ČERENAK, A. *et al.* New male specific markers for hop and application in breeding program. **Scientific Reports**, Springer, v. 9, n. 1, p. 1–9, 2019.

CHEETHAM, S. *et al.* Long noncoding rnas and the genetics of cancer. **British journal of cancer**, Nature Publishing Group, v. 108, n. 12, p. 2419–2425, 2013.

CHEN, X. Small rnas in development–insights from plants. **Current opinion in genetics & development**, Elsevier, v. 22, n. 4, p. 361–367, 2012.

COPATTI, A. S. *et al.* Bap no cultivo in vitro de lúpulo cascade. *In*: **ENPOS, XXI encontro de pós-graduação – UFPEL**. [*S.l.*: *s.n.*], 2019.

CRICK, F. Central dogma of molecular biology. **Nature**, Nature Publishing Group UK London, v. 227, n. 5258, p. 561–563, 1970.

CUI, J. *et al.* Comparative transcriptome analysis between resistant and susceptible tomato allows the identification of lnc rna 16397 conferring resistance to phytophthora infestans by co-expressing glutaredoxin. **The plant journal**, Wiley Online Library, v. 89, n. 3, p. 577–589, 2017.

CUPERUS, J. T.; FAHLGREN, N.; CARRINGTON, J. C. Evolution and functional diversification of mirna genes. **The Plant Cell**, American Society of Plant Biologists, v. 23, n. 2, p. 431–442, 2011.

DESHPANDE, S.; AL. et. Plit: An alignment-free computational tool for identification of long non-coding rnas in plant transcriptomic datasets. **Comput. Biol. Med.**, v. 105, p. 169–181, Feb 2019.

DI, C. *et al.* Characterization of stress-responsive lnc rna s in a rabidopsis thaliana by integrating expression, epigenetic and structural features. **The Plant Journal**, Wiley Online Library, v. 80, n. 5, p. 848–861, 2014.

DOBBLER, P. C. T. **RNA longo não-codificante: características, mecanismos e funcionalidade do DNA "lixo"**. jan. 2015. 37 p. Dissertação (Mestrado) — Universidade Federal do Pampa, Campus São Gabriel, jan. 2015.

DURELLO, R. S.; SILVA, L. M.; BOGUSZ, S. Química do lúpulo. **Química Nova**, SciELO Brasil, v. 42, p. 900–919, 2019.

DYKSTRA, P. B.; KAPLAN, M.; SMOLKE, C. D. Engineering synthetic rna devices for cell control. **Nature Reviews Genetics**, Nature Publishing Group UK London, v. 23, n. 4, p. 215–228, 2022.

ECKER, J. R. *et al.* Encode explained. **Nature**, Nature Publishing Group UK London, v. 489, n. 7414, p. 52–54, 2012.

ENGELHARD, B.; LUTZ, A.; SEIGNER, E. **Hopfen für alle Biere der Welt**. Vöttinger Straße 38, 85354 Freising-Weihenstephan, Institut für Pflanzenbau und Pflanzenzüchtung: Bayerische Landesanstalt für Landwirtschaft (LfL), 2011. ES-Druck, 85356 Freising-Tüntenhausen, März.

FAGHIHI, M. A.; WAHLESTEDT, C. Regulatory roles of natural antisense transcripts. **Nature reviews Molecular cell biology**, Nature Publishing Group UK London, v. 10, n. 9, p. 637–643, 2009.

FAN, X. N.; ZHANG, S. W. lncrna-mfdl: identification of human long non-coding rnas by fusing multiple features and using deep learning. **Molecular bioSystems**, v. 11, n. 3, p. 892–897, Mar 2015.

FU, L. *et al.* Cd-hit: accelerated for clustering the next-generation sequencing data. **Bioinformatics**, Oxford University Press, v. 28, n. 23, p. 3150–3152, 2012.

GALLART, A. P. *et al.* Greenc: a wiki-based database of plant lncrnas. **Nucleic acids research**, Oxford University Press, v. 44, n. Database issue, p. D1161, 2016.

GERSTEIN, M. B. *et al.* What is a gene, post-encode? history and updated definition. **Genome research**, Cold Spring Harbor Lab, v. 17, n. 6, p. 669–681, 2007.

GOODRICH, J. A.; KUGEL, J. F. Non-coding-rna regulators of rna polymerase ii transcription. **Nature reviews Molecular cell biology**, Nature Publishing Group UK London, v. 7, n. 8, p. 612–616, 2006.

GRIFFITHS, A. J. F. *et al.* **Introdução à Genética**. 12. ed. Rio de Janeiro: Guanabara Koogan, 2022.

GUO, J.-C. *et al.* Cnit: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. **Nucleic acids research**, Oxford University Press, v. 47, n. W1, p. W516–W522, 2019.

HAN, S.; AL. et. Lncfinder: an integrated platform for long non-coding rna identification utilizing sequence intrinsic composition, structural information and physicochemical property. **Briefings in Bioinformatics**, v. 20, n. 6, p. 2009–2027, Jul 2018.

HAN, S. *et al.* Lncfinder: an integrated platform for long non-coding rna identification utilizing sequence intrinsic composition, structural information and physicochemical property. **Briefings in bioinformatics**, Oxford University Press, v. 20, n. 6, p. 2009–2027, 2019.

HU, L.; AL. et. Come: a robust coding potential calculation tool for lncrna identification and characterization based on multiple features. **Nucleic Acids Research**, v. 45, n. 1, p. e2–e2, Sep 2016.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Pesquisa Industrial Anual – Produto**. 2020. Disponível em: https://sidra.ibge.gov.br/pesquisa/pia-produtos/quadros/brasil/2018.

IHGC - International Hop Growers' Convention. **Economic Commission: Summary Reports**. Paris: International Hop Growers' Convention, 2018. 55 p. Acesso em: 29 jan. 2019. Disponível em: http://www.hmelj-giz.si/ihgc/doc/2018%20MAY%20IHGC%20EC%20Reports.pdf.

INGRAM, V. M. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. **Nature**, Springer, v. 180, p. 326–328, 1957.

INUI, T. *et al.* Different beers with different hops. relevant compounds for their aroma characteristics. **Journal of agricultural and food chemistry**, ACS Publications, v. 61, n. 20, p. 4758–4764, 2013.

ITO, E. A.; AL. et. Basinet—biological sequences network: a case study on coding and non-coding rnas identification. **Nucleic Acids Research**, v. 46, n. 16, p. e96–e96, Jun 2018.

JAIN, P. *et al.* Identification of long non-coding rna in rice lines resistant to rice blast pathogen maganaporthe oryzae. **Bioinformation**, Biomedical Informatics Publishing Group, v. 13, n. 8, p. 249, 2017.

JIN, J. *et al.* Plncdb: plant long non-coding rna database. **Bioinformatics**, Oxford University Press, v. 29, n. 8, p. 1068–1071, 2013.

KANG, Y.-J. *et al.* Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. **Nucleic acids research**, Oxford University Press, v. 45, n. W1, p. W12–W16, 2017.

KHOCHBIN, S.; LAWRENCE, J.-J. An antisense rna involved in p53 mrna maturation in murine erythroleukemia cells induced to differentiate. **The EMBO Journal**, v. 8, n. 13, p. 4107–4114, 1989.

KIRK, J. M. *et al.* Functional classification of long non-coding rnas by k-mer content. **Nature genetics**, Nature Publishing Group US New York, v. 50, n. 10, p. 1474–1482, 2018.

KONG, L.; AL. et. Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. **Nucleic acids research**, v. 35, n. Web Server issue, p. W345–349, Jul 2007.

KRYSTAL, G. W.; ARMSTRONG, B. C.; BATTEY, J. F. N-myc mrna forms an rna-rna duplex with endogenous antisense transcripts. **Molecular and cellular biology**, Taylor & Francis, v. 10, n. 8, p. 4180–4191, 1990.

KUNG, J. T.; COLOGNORI, D.; LEE, J. T. Long noncoding rnas: past, present, and future. **Genetics**, Oxford University Press, v. 193, n. 3, p. 651–669, 2013.

LAVORGNA, G. *et al.* In search of antisense. **Trends in biochemical sciences**, Elsevier, v. 29, n. 2, p. 88–94, 2004.

LERTAMPAIPORN, S.; AL. et. Identification of non-coding rnas with a new composite feature in the hybrid random forest ensemble algorithm. **Nucleic acids research**, v. 42, n. 11, p. e93–e93, 2014.

LEUNG, A. *et al.* Novel long noncoding rnas are regulated by angiotensin ii in vascular smooth muscle cells. **Circulation research**, Am Heart Assoc, v. 113, n. 3, p. 266–278, 2013.

LI, A.; ZHANG, J.; ZHOU, Z. Plek: a tool for predicting long non-coding rnas and messenger rnas based on an improved k-mer scheme. **BMC bioinformatics**, Springer, v. 15, p. 1–10, 2014.

LI, J. e. a. Lnctar: a tool for predicting the rna targets of long noncoding rnas. **Briefings in Bioinformatics**, v. 16, n. 5, p. 806–812, 2014.

LIN, M. *et al.* Role of characteristic components of humulus lupulus in promoting human health. **Journal of agricultural and food chemistry**, ACS Publications, v. 67, n. 30, p. 8291–8302, 2019.

LIN, M. F.; JUNGREIS, I.; KELLIS, M. Phylocsf: a comparative genomics method to distinguish protein coding and non-coding regions. **Bioinformatics**, v. 27, n. 13, p. i275–282, Jul 2011.

LIU, C. *et al.* Noncode: an integrated knowledge database of non-coding rnas. **Nucleic acids research**, Oxford University Press, v. 33, n. suppl_1, p. D112–D115, 2005.

LIU, J.; GOUGH, J.; ROST, B. Distinguishing protein-coding from non-coding rnas through support vector machines. **PLoS Genetics**, v. 2, n. 4, p. e29, Apr 2006.

LIU, S. *et al.* Predlnc-gfstack: a global sequence feature based on a stacked ensemble learning method for predicting lncrnas from transcripts. **Genes**, MDPI, v. 10, n. 9, p. 672, 2019.

LIU, X. *et al.* Long non-coding rnas and their biological roles in plants. **Genomics, proteomics & bioinformatics**, Elsevier, v. 13, n. 3, p. 137–147, 2015.

LORENZ, R. *et al.* Viennarna package 2.0. **Algorithms for molecular biology**, Springer, v. 6, p. 1–14, 2011.

LOU, D. *et al.* Lncphedb: a genome-wide lncrnas regulated phenotypes database in plants. **Abiotech**, Springer, v. 3, n. 3, p. 169–177, 2022.

LUCERO, L. *et al.* Functional classification of plant long noncoding rnas: a transcript is known by the company it keeps. **New Phytologist**, Wiley Online Library, v. 229, n. 3, p. 1251–1260, 2021.

MAHAFFEE, W. F. *et al.* **Compendium of hop diseases and pests.** [*S.l.*]: American Phytopathological Society (APS Press), 2009.

MARCOS, J. A. M. *et al.* **GUIA DEL CULTIVO DEL LÚPULO**. [*S.l.*]: S.A.E. Fomento del Lúpulo, 2011.

MATTICK, J. S.; RINN, J. L. Discovery and annotation of long noncoding rnas. **Nature structural & molecular biology**, Nature Publishing Group US New York, v. 22, n. 1, p. 5–7, 2015.

MENG, J.; AL. et. lncrna-lstm: Prediction of plant long non-coding rnas using long short-term memory based on p-nts encoding. *In*: **Intelligent Computing Methodologies**. [*S.l.*]: Springer International Publishing, 2019. p. 347–357.

NACHEL, M. **Beer for dummies**. [*S.l.*]: John Wiley & Sons, 2011.

NAM, J.-W.; CHOI, S.-W.; YOU, B.-H. Incredible rna: dual functions of coding and noncoding. **Molecules and cells**, Korean Society for Molecular and Cellular Biology, v. 39, n. 5, p. 367, 2016.

NATH, V. S. *et al.* Identification and characterization of long non-coding rna and their response against citrus bark cracking viroid infection in humulus lupulus. **Genomics**, Elsevier, v. 113, n. 4, p. 2350–2364, 2021.

NEGRI, T. d. C. *et al.* Pattern recognition analysis on long noncoding rnas: a tool for prediction in plants. **Briefings in bioinformatics**, Oxford University Press, v. 20, n. 2, p. 682–689, 2018.

NELSON, D. L.; COX, M. M. **Princípios de Bioquímica de Lehninger**. 7. ed. Porto Alegre: Artmed, 2019.

NEVE, R. A. **Hops**. [*S.l.*]: Springer Science & Business Media, 1991.

NG, S.-Y. *et al.* The long noncoding rna rmst interacts with sox2 to regulate neurogenesis. **Molecular cell**, Elsevier, v. 51, n. 3, p. 349–359, 2013.

NGUYEN, N. H.; VU, N. T.; CHEONG, J.-J. Transcriptional stress memory and transgenerational inheritance of drought tolerance in plants. **International Journal of Molecular Sciences**, MDPI, v. 23, n. 21, p. 12918, 2022.

ØROM, U. A.; SHIEKHATTAR, R. Long non-coding rnas and enhancers. **Current opinion in genetics & development**, Elsevier, v. 21, n. 2, p. 194–198, 2011.

PADGITT-COBB, L. K. *et al.* An improved assembly of the “cascade” hop (humulus lupulus) genome uncovers signatures of molecular evolution and refines time of divergence estimates for the cannabaceae family. **Horticulture Research**, Oxford University Press, v. 10, n. 2, p. uhac281, 2023.

PECUáRIA, B. M. da Agricultura e. **Anuário da Cerveja 2024: ano de referência 2023**. Brasília: MAPA/SDA, 2024.

PIAN, C.; AL. et. Lncrnapred: Classification of long non-coding rnas and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. **PLoS ONE**, v. 11, n. 5, p. e0154567, 2016.

PONTING, C. P.; BELGARD, T. G. Transcribed dark matter: meaning or myth? **Human molecular genetics**, Oxford University Press, v. 19, n. R2, p. R162–R168, 2010.

PYFROM, S. C.; LERNER, L. H.; PAYTON, J. Plaidoh: a novel method for functional prediction of long non-coding rnas identifies cancer-specific lncrna activities. **BMC Genomics**, v. 20, n. 137, Feb 2019.

RIELLA, C. V. **Pequenos RNAs não codificantes: do lixo ao tesouro**. [*S.l.*]: SciELO Brasil, 2019. 168–169 p.

ROMBEL, I. T. *et al.* Orf-finder: a vector for high-throughput gene identification. **Gene**, Elsevier, v. 282, n. 1-2, p. 33–41, 2002.

RUIZ-ORERA, J. *et al.* Long non-coding rnas as a source of new peptides. **elife**, eLife Sciences Publications, Ltd, v. 3, p. e03523, 2014.

RYMARQUIS, L. A. *et al.* Diamonds in the rough: mrna-like non-coding rnas. **Trends in plant science**, Elsevier, v. 13, n. 7, p. 329–334, 2008.

SCHÖNBERGER, C.; KOSTELECKY, T. 125th anniversary review: The role of hops in brewing. **Journal of the Institute of Brewing**, Wiley Online Library, v. 117, n. 3, p. 259–267, 2011.

SEO, J. S. *et al.* Elf18-induced long-noncoding rna associates with mediator to enhance expression of innate immune response genes in arabidopsis. **The Plant Cell**, American Society of Plant Biologists, v. 29, n. 5, p. 1024–1038, 2017.

SHENDURE, J. *et al.* Dna sequencing at 40: past, present and future. **Nature**, Nature Publishing Group UK London, v. 550, n. 7676, p. 345–353, 2017.

SHUAI, P. *et al.* Genome-wide identification and functional prediction of novel and drought-responsive lincrnas in populus trichocarpa. **Journal of experimental botany**, Oxford University Press UK, v. 65, n. 17, p. 4975–4983, 2014.

SIMOPOULOS, C. M. A.; WERETILNYK, E. A.; GOLDING, G. B. Prediction of plant lncrna by ensemble machine learning classifiers. **BMC Genomics**, v. 19, n. 1, p. 316, May 2018.

SINGH, A.; VIVEK, A.; KUMAR, S. Alnc: An extensive database of long non-coding rnas in angiosperms. **Plos one**, Public Library of Science San Francisco, CA USA, v. 16, n. 4, p. e0247215, 2021.

SINGH, U. *et al.* Plncpro for prediction of long non-coding rnas (lncrnas) in plants and its application for discovery of abiotic stress-responsive lncrnas in rice and chickpea. **Nucleic acids research**, Oxford University Press, v. 45, n. 22, p. e183–e183, 2017.

SOUZA, R. D. Estabelecimento in vitro, micropropagação e variação somaclonal de lúpulo (humulus lupulus l.). **Dissertação (mestrado) - Universidade do Estado de Santa Catarina, Centro de Ciências Agroveterinárias**, Programa de Pós-Graduação em Produção Vegetal, Lages, SC, p. 80, 2020.

SPóSITO RODRIGO VERALDI ISMAEL, C. M. d. A. B. A. L. T. M. B. **A Cultura do Lúpulo**. Piracicaba, SP.

SUN, K.; AL. et. iseerna: identification of long intergenic non-coding rna transcripts from transcriptome sequencing data. **BMC Genomics**, v. 14, n. Suppl 2, p. S7, 2013.

SUN, K.; WANG, H.; SUN, H. Nams webserver: coding potential assessment and functional annotation of plant transcripts. **Briefings in Bioinformatics**, Sep 2020.

SUN, L.; AL. et. lncrscan-svm: A tool for predicting long non-coding rnas using support vector machine. **PLoS ONE**, v. 10, n. 10, p. e0139654, 2015.

SUN, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. **Nucleic acids research**, Oxford University Press, v. 41, n. 17, p. e166–e166, 2013.

SZCZEŚNIAK, M. W.; ROSIKIEWICZ, W.; MAKAŁOWSKA, I. Cantatadb: a collection of plant long non-coding rnas. **Plant and Cell Physiology**, Oxford University Press, v. 57, n. 1, p. e8–e8, 2016.

TIAN, X.-C. *et al.* Plant-lncpipe: a computational pipeline providing significant improvement in plant lncrna identification. **Horticulture Research**, Oxford University Press, v. 11, n. 4, p. uhae041, 2024.

TONG, X.; LIU, S. Cppred: coding potential prediction based on the global description of rna sequence. **Nucleic Acids Research**, v. 47, n. 8, p. e43–e43, Feb 2019.

TRIPATHI, R.; AL. et. Deeplnc, a long non-coding rna prediction tool using deep neural network. **Network Modeling and Analysis in Health Informatics and Bioinformatics**, v. 5, n. 21, p. 1–14, Dec 2016.

ULITSKY, I.; BARTEL, D. P. lincrnas: genomics, evolution, and mechanisms. **Cell**, Elsevier, v. 154, n. 1, p. 26–46, 2013.

VIEIRA, L. M.; AL. et. Plantrna_sniffer: A svm-based workflow to predict long intergenic non-coding rnas in plants. **non-coding RNA**, v. 3, n. 1, p. 1–11, Mar 2017.

WAITITU, J. K. *et al.* Plant non-coding rnas: Origin, biogenesis, mode of action and their roles in abiotic stress. **International Journal of Molecular Sciences**, MDPI, v. 21, n. 21, p. 8401, 2020.

WANG, G.; AL. et. Characterization and identification of long non-coding rnas based on feature relationship. **Bioinformatics**, v. 35, n. 17, p. 2949–2956, Jan 2019.

WANG, J. *et al.* Genome-wide analysis of tomato long non-coding rnas and identification as endogenous target mimic for microrna in response to tylcv infection. **Scientific reports**, Springer, v. 5, n. 1, p. 1–16, 2015.

WANG, K. C.; CHANG, H. Y. Molecular mechanisms of long noncoding rnas. **Molecular cell**, Elsevier, v. 43, n. 6, p. 904–914, 2011.

WANG, L.; AL. et. Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. **Nucleic Acids Res.**, v. 41, n. 6, p. e74, Apr 2013.

WANG, L. *et al.* Genome-wide identification and functional profile analysis of long non-coding rnas in avicennia marina. **The Plant Genome**, Wiley Online Library, p. e20450, 2024.

WANG, Y. *et al.* Arabidopsis noncoding rna mediates control of photomorphogenesis by red light. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 111, n. 28, p. 10359–10364, 2014.

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. **Nature**, Nature Publishing Group UK London, v. 171, n. 4356, p. 737–738, 1953.

WU, T. *et al.* Npinter: the noncoding rnas and protein related biomacromolecules interaction database. **Nucleic acids research**, Oxford University Press, v. 34, n. suppl_1, p. D150–D152, 2006.

WUCHER, V.; AL. et. Feelnc: a tool for long non-coding rna annotation and its application to the dog transcriptome. **Nucleic acids research**, v. 45, n. 8, p. e57–e57, 2017.

YAN, M.-D. *et al.* Identification and characterization of a novel gene saf transcribed from the opposite strand of fas. **Human molecular genetics**, Oxford University Press, v. 14, n. 11, p. 1465–1474, 2005.

YANG, C.; AL. et. Lncadeep: an ab initio lncrna identification and functional annotation tool based on deep learning. **Bioinformatics**, v. 34, n. 22, p. 3825–3834, Nov 2018.

YI, X. *et al.* Pnrd: a plant non-coding rna database. **Nucleic acids research**, Oxford University Press, v. 43, n. D1, p. D982–D989, 2015.

ZHANG, G.; AL. et. Ptlnc-bxe: Prediction of plant lncrnas using a bagging-xgboost-ensemble method with multiple features. 2019.

ZHANG, L. *et al.* Long noncoding rna s involve in resistance to verticillium dahliae, a fungal disease in cotton. **Plant biotechnology journal**, Wiley Online Library, v. 16, n. 6, p. 1172–1185, 2018.

ZHAO, J.; SONG, X.; WANG, K. lncscore: alignment-free identification of long noncoding rna from assembled novel transcripts. **Scientific reports**, Springer, v. 6, n. 1, p. 1–12, 2016.

ZHOU, B. *et al.* Evlncrna-dpred: improved prediction of experimentally validated lncrnas by deep learning. **Briefings in Bioinformatics**, Oxford University Press, v. 24, n. 1, p. bbac583, 2023.

ZHU, Q.-H. *et al.* Long noncoding rna s responsive to f usarium oxysporum infection in a rabidopsis thaliana. **New phytologist**, Wiley Online Library, v. 201, n. 2, p. 574–584, 2014.