UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS GRADUAÇÃO EM BIOINFORMÁTICA
DISSERTAÇÃO DE MESTRADO

MARCOS JOSÉ ANDRADE VIANA

Plant Co-expression Annotation Resource 2.0: uma ferramenta web para a associação de proteínas de função desconhecida a estresses abióticos em plantas.

Belo Horizonte

Janeiro

2021

MARCOS JOSÉ ANDRADE VIANA

Plant Co-expression Annotation Resource 2.0: uma ferramenta web para a associação de proteínas de função desconhecida a estresses abióticos em plantas.

Dissertação de Mestrado apresentada ao programa de interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Bioinformática

Orientador: Dr. Maurício de Alvarenga Mudadu
Coorientador: Prof. Dr. Francisco Pereira Lobo
Coorientador: Dr. Adhemar Zerlotini Neto

043 Viana, Marcos José Andrade.
Plant Co-expression Annotation Resource 2.0: uma ferramenta web para a associação de proteínas de função desconhecida a estresses abióticos em plantas [manuscrito] / Marcos José Andrade Viana. – 2021.
77 f. : il. ; 29,5 cm.

Orientador: Dr. Maurício de Alvarenga Mudadu. Coorientadores: Prof. Dr. Francisco Pereira Lobo e Dr. Adhemar Zerlotini Neto.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Biologia Computacional. 2. Sistemas Computacionais. 3. Organismos Geneticamente Modificados. 4. Estresse Fisiológico. 5. Plantas. I. Mudadu, Maurício de Alvarenga. II. Lobo, Francisco Pereira. III. Zerlotini Neto, Adhemar. IV. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. V. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Rosilene Moreira Coelho de Sá – CRB 6 - 2726

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-graduação em Bioinformática

ATA DA DEFESA DE DISSERTAÇÃO

MARCOS JOSÉ ANDRADE VIANA

Às nove horas do dia 25 de janeiro de 2021, reuniu-se, através do aplicativo Zoom, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho do Sr. Marcos José Andrade Viana intitulado: **"Plant Co-expression Annotation Resource 2.0: uma ferramenta web para a associação de proteínas de função desconhecida a estresses abióticos em plantas"**, requisito para obtenção do grau de Mestre em Bioinformática**.** Abrindo a sessão, o Presidente da Comissão, Dr. Mauricio de Alvarenga Mudadu , após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**Indicação**
Dr. Mauricio de Alvarenga Mudadu - Orientador	EMBRAPA	Aprovado
Dr. Francisco Pereira Lobo - Coorientador	UFMG	Aprovado
Dr. Adhemar Zerlotini Neto - Coorientador	EMBRAPA	Aprovado
Dr. José Miguel Ortega	UFMG	Aprovado
Dr. Roberto Willians Noda	EMBRAPA	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 25 de janeiro de 2021.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 25/01/2021, às 11:26,

conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 25/01/2021, às 11:42, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maurício de Alvarenga Mudadu**, **Usuário Externo**, em 25/01/2021, às 16:27, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adhemar Zerlotini Neto**, **Usuário Externo**, em 25/01/2021, às 16:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Roberto Willians Noda**, **Usuário Externo**, em 25/01/2021, às 16:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0533739** e o código CRC **D72C932B**.

Referência: Processo nº 23072.203733/2021-99 SEI nº 0533739

AGRADECIMENTOS

Primeiramente a Deus por sempre ter colocado em meu caminho pessoas especiais que ajudaram no meu desenvolvimento pessoal e profissional e também por sempre ter me proporcionado oportunidades incríveis, como foi essa!

A minha amada esposa Rossana e filhos, Benício e Mariana. A eles gostaria de primeiramente pedir desculpas pelos momentos de ausência dos últimos meses, e agradecer imensamente pela compreensão, incentivo e apoio que sempre me deram, sem eles eu não teria conseguido! Conseguimos juntos meus amores, vocês são tudo para mim!

A Embrapa, por ter me dado a oportunidade de cursar o mestrado, tenho muito orgulho de fazer parte dessa empresa! E espero dar muito retorno com bastante trabalho e dedicação! Aos colegas de Embrapa Milho e Sorgo, o primeiro, sem dúvidas, Dr. Roberto Noda, que foi o grande incentivador e responsável pela minha escolha do curso de Bioinformática, obrigado pelos ensinamentos, paciência, apoio e incentivo. Ao meu amigo Wanderley, pelo incentivo, apoio, amizade e companheirismo dispensados a mim desde que cheguei na empresa, uma pessoa de um coração gigante. As colegas Dra. Vera Alves e Adriana Noce, pelo apoio e incentivo, elas foram responsáveis por eu não desistir em uma fase crucial de todo esse processo. Aos pesquisadores Dr. Jurandir Magalhães e Dra. Claudia Teixeira pelos ensinamentos e apoio.

Aos meu orientador Dr. Maurício Mudadu, que aceitou o desafio de orientar um aluno da área de exatas, com conhecimento incipiente em Biologia e que estava há pelo menos 11 anos longe da vida acadêmica. Muito obrigado por todos os ensinamentos, parceria e pela paciência! Ao meu coorientador Dr. Adhemar Neto, pelos ensinamentos, paciência, atenção e incentivo. Tenho um time de excelentes Bioinformatas em quem me espelhar!

Ao meu coorientador Dr. Chico Lobo, que me acolheu em seu laboratório (LAB) na UFMG, muito obrigado pelos ensinamentos e incentivo. Acabei me transformando em um grande admirador do Chico, pelo seu caráter, profissionalismo, inteligência e o grande respeito, carinho e atenção que tem por todos seus orientandos e alunos. Chico tá formando uma nova geração de pesquisadores que terão bases sólidas não só de conhecimento, como de dignidade humana. Aos professores do PPG de Bioinformática pelos ensinamentos e profissionalismo, e aos secretários do curso Tiago e Sheyla que foram sempre muito solícitos e profissionais.

A todos os colegas do mestrado, e em especial a minha equipe nota 10 da disciplina de EGTP: Renato, João Paulo e Bruno, muito obrigado senhores, pelos ensinamentos e companheirismo!

Ao meu pai, em memória, que sempre acreditou em mim juntamente com meu irmão Sérgio, minha mãe Marinete e minha tia Alzerina. Meu muito obrigado por tudo, amo vocês! Enfim, a todos que de alguma forma contribuíram direta ou inteiramente para concretização desse projeto.

RESUMO

O desenvolvimento de culturas geneticamente modificadas (GM) inclui a descoberta de genes candidatos através de análises de bioinformática, usando dados genômicos, expressão gênica, entre outros. As proteínas de função desconhecida (PFD) são alvos interessantes para pipelines de culturas GM devido a novidade associada a esses alvos e também para evitar proteções de direitos autorais. Um método para inferir a possível função das PFD é relacioná-las a fatores de interesse, como estresses abióticos, usando redes de ortologia e coexpressão, aplicando a abordagem de culpa por associação. O objetivo desse trabalho é desenvolver e disponibilizar a versão 2 da ferramenta PlantAnnot (PlantAnnot2) e o objetivo do PlantAnnot2 é a descoberta de PFD envolvidas em respostas a estresses abióticos em plantas. Para isso, foram processados dados genômicos de 67 plantas, com algumas importantes espécies tolerantes a stresses abióticos. Os softwares Diamond e InterproScan foram usados para descobrir PFDs em todas as plantas. Dados de expressão gênica (RNA-seq) relacionados a estresses abióticos foram baixados do NCBI / GEO e usados no software LSTrAP para construir *clusters* de coexpressão cujos membros estão mais provavelmente relacionados aos mecanismos moleculares associados ao estresse abiótico em plantas. Grupos de ortólogos foram criados com OrthoFinder, depois foram buscadas PFDs associadas a esses grupos de ortólogos e também a clusters de coexpressão simultâneamente. Como resultado, foram armazenados no Machado 2.136.336 genes e 2.714.161 mRNA, juntamente com suas proteínas traduzidas. Recuperados 78.416 PFDs com as análises de Diamond e Interproscan, criados 91.172 grupos de ortólogos e 1.975 clusters de coexpressão. Foi desenvolvido um protocolo para busca de anotações de PFDs, que recupera PFDs que pertençam a algum grupo de ortólogos que contenha também proteínas cujo mRNA faz parte de *clusteres* de coexpressão relacionados à estresses abióticos. Dessa forma, foram recuperadas s 4.673 PFDs, na versão 1 do sistema tinha sido 1.364 PFDs. Foi realizado uma pesquisa bibliográfica sobre as proteínas que pertencem aos grupos ortólogos, para todos os PFDs pertencentes às espécies *Pearl millet(Cenchrus americanus), Populus simonii, Oropethium thomaeum* e *Boea hygrometrica*, todos conhecidos por serem tolerantes a estresses abiótico (517 PFDs). Encontrados estudos relacionados a estresses abióticos, em média, para 67,5% das PFDs, na versão 1 do sistema o percentual era de 35%. Um servidor web https://www.machado.cnptia.embrapa.br/plantannot2 está disponível gratuitamente e fornece consultas indexadas de PFDs possivelmente associadas a estresses abióticos. Espera-se que o PlantAnnot2 seja útil para pesquisadores que buscam obter genes relacionados a respostas a estresses abióticos para a produção de novos cultivos GM tolerantes aos riscos das mudanças climáticas.

ABSTRACT

The development of genetically modified (GM) crops includes the discovery of candidate genes through bioinformatics analysis, using genomic data, gene expression, among others. Proteins of unknown function (PUF) are interesting targets for GM crop pipelines due to the novelty associated with these targets and also to avoid copyright protections. One method to infer the possible function of PUF is to relate them to factors of interest, such as abiotic stresses, using orthology and coexpression networks, applying the guilt by association approach. The objective of this work is to develop and make available version 2 of the PlantAnnot tool (PlantAnnot2) and the objective of PlantAnnot2 is the discovery of PFD involved in responses to abiotic stresses in plants. For that, genomic data from 67 plants were processed, with some important species tolerant to abiotic stresses. Diamond and InterproScan software were used to discover PUF in all plants. Gene expression data (RNA-seq) related to abiotic stresses were downloaded from NCBI / GEO and used in the LSTrAP software to build coexpression clusters whose members are most likely related to the molecular mechanisms associated with abiotic stress in plants. Groups of orthologous were created with OrthoFinder, then PUF associated with these groups of orthologous and also with coexpression clusters were searched simultaneously. As a result, 2,136,336 genes and 2,714,161 mRNA were stored in Machado, along with their translated proteins. 78,416 PUF were recovered with Diamond and Interproscan analyzes, 91,172 groups of orthologous and 1,975 coexpression clusters were created. A protocol for searching PUF annotations was developed, which retrieves PUF that belong to a group of orthologous that also contain proteins whose mRNA is part of coexpression clusters related to abiotic stresses. Thus, 4,673 PUF were recovered, in version 1 of the system it had been 1,364 PUF. A bibliographic search was carried out on proteins belonging to orthologous groups, for all PUF belonging to the species Pearl millet (Cenchrus americanus), Populus simonii, Oropethium thomaeum and Boea hygrometrica, all known to be tolerant to abiotic stresses (517 PUF). Studies related to abiotic stresses were found, on average, for 67.5% of PUF, in version 1 of the system the percentage was 35%. A web server https://www.machado.cnptia.embrapa.br/plantannot2 is available for free and provides indexed queries for PUF possibly associated with abiotic stresses. PlantAnnot2 is expected to be useful for researchers looking to obtain genes related to responses to abiotic stresses for the production of new GM crops tolerant to the risks of climate change.

LISTA DE FIGURAS

Figura 1 – Sinalização do ABA envolve quinases e fosfatases.
Figura 2 – Visão geral do funcionamento do algoritmo do OrthoFinder.
Figura 3 – Fluxo de trabalho do LSTrAP.
Figura 4 – Algoritmo para anotação de PFDs.
Figura 5 – Busca por PFD usando ortologia e coexpressão gênica.
Figura 6 – Fluxo de trabalho do PlantAnnot2
Figura 7 – Filtros do PlantAnnot2.
Figura 8 – Uso de busca textual no PlantAnnot2.
Figura 9 – Baixar resultado de uma busca em arquivo formato .tsv.
Figura 10 - Redes de enriquecimento para Arabidopsis thaliana no Plantannot2 x PlantAnnot.
Figura 11 – Redes de enriquecimento para Oryza sativa no Plantannot2 x PlantAnnot.

LISTA DE TABELAS

Tabela 1 – Genomas de organismos do PlantAnnot2.
Tabela 2 – Protocolos de busca de PFDs.
Tabela 3 – Dados armazenados no PlantAnnot x PlantAnnot2
Tabela 4 – Comparativos de resultados de homologia entre OrthoMCL, OrthoFinder e OrthoFinder + cd-hit do PlantAnnot2.
Tabela 5 – Clusters de coexpressão gênica do PlantAnnot2.
Tabela 6 – Proteínas de Função Desconhecida recuperadas no PlantAnnot2 por protocolo de busca.
Tabela 7 – Comparação de quantidade de PFDs recuperadas por análise de homologia no PlantAnnot2
Tabela 8 – Enriquecimento de vias Arabidopsis thaliana, comparação PlantaAnnot x PlnatAnnot2
Tabela 9 – Enriquecimento de vias Oryza sativa, comparação PlantaAnnot x PlnatAnnot2

LISTA DE SIGLAS

FAO - Organização das Nações Unidas para a Alimentação e a Agricultura

FIDA - Fundo Internacional de Desenvolvimento Agrícola

OMS - Organização Mundial da Saúde

UNICEF - Fundo das Nações Unidas para a Infância

ONU - Organização das Nações Unidas

GM - Geneticamente Modificadas

DPI - Direitos de Propriedade Intelectual

CRISPR - *Clustered Regularly Interspaced Short Palindromic Repeats*

Cas - CRISPR-*associated*

TI - Tecnologia da Informação

LMB - Laboratório Multiusuário de Bioinformática

RAM - *Random Access Memory*

SGE - *Sun Grid Engine*

SSH - *Secure Socket Shell*

API - *Application Programming Interface*

IPMGSC - *International Pearl Millet Genome Sequencing Consortium*

GSA - *Genetics Society of America*

NCBI - *National Center for Biotechnology Information*

GFF3 - *Generic Feature Format Version 3*

NCBI - *National Center for Biotechnology Information*

NCBI-nr - Banco de dados *nonredundant* do NCBI

DIAMOND - *Double Index Alignment of Next-generation Sequencing Data*

TSV - *tab-separated values*

XML - *Extensible Markup Language*

DFD - Domínios de Função Desconhecida

PFD - Proteínas de Função Desconhecida

RBNH - *Reciprocal Best Length-Normalized hit*

SAM - *Sequence Alignment Map*

BAM - *Binary Alignment Map*

LSTrAP - *Large Scale Transcriptome Analysis Pipeline*

GEO - *Gene Expression Omnibus*

GMOD - *Generic Model Organism Database*

DIAMOND - *Double Index Alignment of Next-generation Sequencing Data*

PCC - *Pearson Cofficient Correlation*

ABA - Ácido abscísico

BLAST - *Basic Local Alignment Search Tool*

FDR - *False Discovery Rate*

SGBD - Sistema de Gerenciamento de Banco de Dados

TPM - Transcrito por milhão

Sumário

AGRADECIMENTOS V
RESUMO VII
ABSTRACT VIII
LISTA DE FIGURAS IX
LISTA DE TABELAS X
LISTA DE SIGLAS XI
Sumário XIII
CAPÍTULO 1: INTRODUÇÃO 1
1.1. Estresse abióticos em plantas 2
1.1.1. Fatores centrais de sinalização do ácido abscísico (ABA) incluem fosfatases e quinases 4
1.1.2. Plantas com importantes características de tolerância a estresses abióticos 6
1.2. Recuperação de proteínas de função desconhecidas 7
1.3. Homologia 7
1.3.1. OrthoFinder 9
1.4. Redes de coexpressão gênica 11
1.4.1. Large-scale Transcriptome Analysis Pipeline (LSTrAP) 12
CAPÍTULO 2: OBJETIVOS 14
2.1. Objetivo Geral 14
2.2. Objetivos específicos 14
CAPÍTULO 3: MATERIAIS E MÉTODOS 15
3.1. Infraestrutura 15
3.2. Algoritmo e fluxo de tarefas do projeto 16
3.3. Download e tratamento de dados genômicos 18
3.4. Recuperação de PFDs 21
3.5. Criação de grupos de homólogos 22
3.6. Redes de coexpressão gênica 25
3.7. Carregamento (*loading*) dos dados 26
3.8. Instalação e configuração da interface web 27
3.9. Protocolos de filtragem de PFDs 28
3.10. Enriquecimento de vias 31
CAPÍTULO 4: RESULTADOS 32
4.1. Dados armazenados 32
4.2. Homologia 32

4.3. Redes de coexpressão 34
4.4. Caracterização de PFDs 35
4.5. Anotação de PFDs 37
4.5.1. Estudos de caso 38
4.6. Enriquecimentos de vias 39
CAPÍTULO 5: DISCUSSÃO 44
5.1. Dados armazenados 44
5.2. Homologia 44
5.3. Redes de coexpressão 46
5.4. Caracterização de PFDs 47
5.5. Anotação de PFDs 47
5.6. Estudos de caso 50
5.7. Enriquecimentos de vias 51
5.8. Ferramentas semelhantes 52
CAPÍTULO 6: CONCLUSÃO 53
CAPÍTULO 7: PERSPECTIVAS 54

"Será que ficar em sua zona de conforto, sem arriscar, sem lutar, é vencer?"

Érico Macri

CAPÍTULO 1: INTRODUÇÃO

Segundo estimativa da Organização das Nações Unidas (ONU), a população mundial saltará de 7,7 bilhões de pessoas em 2019 para 9,7 bilhões em 2050 e a Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) afirma que a produção de alimentos deverá crescer em 70% para suprir esse aumento populacional (ONU, 2019). Outro ponto de alerta está no relatório "O Estado da segurança alimentar e nutrição no mundo 2020 (SOFI)**"** desenvolvido pela (FAO), pelo Fundo Internacional de Desenvolvimento Agrícola (FIDA), pela Organização Mundial da Saúde (OMS), pelo Programa Mundial de Alimentos (WFP), e o Fundo das Nações Unidas para a Infância (UNICEF), nesse importante documento, estima-se que a fome atualmente afeta 7,4% da população e deve aumentar para 9,5% até 2030 (FAO, 2020). O mundo enfrentará nas próximas décadas um enorme desafio, pois a combinação de uma população crescente e as mudanças ambientais tornam a segurança alimentar muito mais difícil de ser alcançada (CHALLINOR *et al.*, 2010).

Nos atuais cenários de mudanças climáticas, a exposição de plantas a estresses abióticos é mais frequente. Algumas espécies com grande importância agropecuária sentem os efeitos dos estresses abióticos que afetam dramaticamente seu crescimento e sua produtividade (FEDOROFF *et al.*, 2010)(ZHU, Jian Kang, 2016)(DOLFERUS, 2014)(SHAMEER *et al.*, 2019). Por exemplo, a seca e as altas temperaturas afetam negativamente a produção do trigo em até 56% (QASEEM; QURESHI; SHAHEEN, 2019).

Nesse sentido, plantas geneticamente modificadas (GM) podem nos ajudar a superar vários obstáculos que se colocam em nosso futuro. Nessas plantas podemos inserir características desejáveis *(traits*) importantes, como longevidade, adaptabilidade às mudanças climáticas ou tolerância a estresses abióticos, de forma a minimizar essas perdas de produtividade (BAILEY-SERRES *et al.*, 2019)(NUTAN *et al.*, 2020) .

Com o avanço mais recente da engenharia genética de precisão (EGP), tecnologias eficientes de edição gênica, como sistemas baseados em CRISPR/CRISPR-associated (Cas) (do inglês *Clustered Regularly Interspaced Short Palindromic Repeats*), têm gerado grandes oportunidades para a agricultura, permitindo a manipulação de genomas de espécies de plantas de grande interesse agropecuário com precisão, de forma rápida e com menor custo (CHEN, Kunling *et al.*, 2019)(MOLINARI *et al.*, 2020), inserindo *traits* diretamente em linhagens elite (GAO *et al.*, 2020). Os direitos de propriedade intelectual (DPI) são amplamente utilizados por empresas de biotecnologia para suas plantas GM, para permitir direitos exclusivos e fornecer melhores retornos financeiros para os altos investimentos em pesquisa e desenvolvimento (WOŹNIAK *et al.*, 2019).

Um dos primeiros passos na criação de organismos GM é a descoberta do gene candidato, que se baseia em análises de bioinformática que usam grandes volumes de dados biológicos e reduzem de forma significante o custo e o tempo para recuperar esse gene (SCHEBEN; EDWARDS, 2018)(PRADO *et al.*, 2014)(AMBROSINO *et al.*, 2020). Para evitar direitos de propriedade intelectual sobre genes já patenteados, seus mecanismos moleculares e produtos, pode ser desejável começar a pesquisar genes e proteínas sem funções ainda descritas (VIANA, Marcos José Andrade; ZERLOTINI; DE ALVARENGA MUDADU, 2021). Essas proteínas de função desconhecida (PFD) são muito prevalentes em genomas eucarióticos e podem desempenhar papéis na determinação de diferenças entre as espécies (GOLLERY *et al.*, 2006) e também podem estar relacionadas à resistência a estresses abióticos (LUHUA *et al.*, 2013).

A tolerância a estresses abióticos é uma característica complexa e poligênica. Durante o evento de estresse, vários genes são expressos na planta, desencadeando respostas desde a percepção do sinal de estresse até a ativação das respostas. Compreender os mecanismos pelos quais as plantas percebem esses sinais e, posteriormente, sua transmissão ao maquinário celular para ativar as respostas adaptativas é de importância crítica para o desenvolvimento de plantas GM com *traits* de tolerância a estresses abióticos (MANTRI *et al.*, 2013). Ferramentas e análises como QTL, GWAS, expressão gênica e redes regulatórias podem ser usadas para encontrar os genes e mecanismos moleculares que podem desempenhar papéis importantes durante o acontecimento de condições adversas ao desenvolvimento normal da planta (NOGUÉ *et al.*, 2016)(NUCCIO *et al.*, 2018)(PROOST; KRAWCZYK; MUTWIL, 2017).

Apresentamos aqui uma ferramenta web denominada *Plant Co-expression Annotation Resource 2.0* (*PlantAnnot2*) disponível publicamente e de forma gratuita pelo endereço web https://www.machado.cnptia.embrapa.br/plantannot2. O *PlantAnnot2* é a segunda versão do *PlantAnnot* (VIANA, Marcos José Andrade; ZERLOTINI; DE ALVARENGA MUDADU, 2021). Ela usa dados genômicos de plantas, dados de sequenciamento de RNA, ortologia e redes de coexpressão no intuito de selecionar PFD para participação em *pipelines* de melhoramento de culturas GM tolerantes a estresses abióticos.

1.1. Estresse abióticos em plantas

Estresses abióticos são aqueles ocasionados por qualquer condição ambiental, como falta ou excesso de água, calor ou frio intenso, salinidade ou metais pesados no solo, que impedem a planta de alcançar seu potencial genético pleno. As condições ideais para o desenvolvimento pleno de determinada planta são aquelas que permitem ela alcançar o máximo crescimento e potencial

reprodutivo. As plantas respondem e se adaptam a esses estresses nos níveis molecular, celular, fisiológico e bioquímico (YAMAGUCHI-SHINOZAKI; SHINOZAKI, 2006) e muitas vezes têm seu crescimento e/ou produtividade prejudicados devido a essas adaptações. Por exemplo, uma maneira da planta compensar a restrição de água é fechando seus estômatos para reduzir a perda do líquido por transpiração, no entanto, esse procedimento também diminui a absorção de CO_2 pela folha, reduzindo, assim, a fotossíntese e consequentemente reprimindo o seu crescimento (TAIZ *et al.*, 2017).

As plantas são sésseis e não podem fugir de estresses abióticos, como os animais, e assim desenvolveram, em milhões de anos de evolução, uma ampla gama de mecanismos moleculares para responder à complexa rede de sinais ambientais, que ativam múltiplas vias, moduladas por diferentes genes responsivos, para tentar conferir tolerância aos estresses abióticos (HIRAYAMA; SHINOZAKI, 2010)(YOU; CHAN, 2015). Todas as mudanças fisiológicas, morfológicas e de desenvolvimento em plantas têm uma base molecular/genética. Na era atual de mudanças climáticas contínuas, o entendimento dos aspectos moleculares envolvidos na resposta ao estresse abiótico em plantas é uma prioridade e possui fundamental importância para o desenvolvimento de plantas GM tolerantes a essas mudanças (AMBROSINO *et al.*, 2020).

Um dos mais importantes estresses abióticos é a seca. Nesse tipo de estresse as respostas fisiológicas da planta representam uma combinação de eventos moleculares a montante, que foram ativados pela detecção dos sinais de estresse. Portanto, também é muito importante compreender como esses eventos são ativados ou desativados e como interagem, para o desenvolvimento de novas cultivares mais tolerantes, com melhor estabilidade e rendimento em condições de estresse abióticos, como por exemplo a seca (NEPOMUCENO *et al.*, 2001). As abordagens genéticas direta e reversa revelaram genes e produtos gênicos que estão envolvidos na expressão gênica, transdução de sinal e tolerância ao estresse (YAMAGUCHI-SHINOZAKI; SHINOZAKI, 2006). Existem também estudos sobre fisiologia e os mecanismos moleculares de tolerância ao estresse abiótico que levaram à caracterização de vários genes associados à adaptação ao estresse (MANTRI *et al.*, 2013).

A tolerância à dessecação (TD) foi uma adaptação chave que permitiu a plantas sairem da água e sobreviverem em solo terrestre (VANBUREN *et al.*, 2015), há aproximadamente 470 milhões de anos (LENTON *et al.*, 2012). Essa característica foi uma das primeiras necessárias para as plantas colonizarem a terra e resultado de milhões de anos de evolução, as vias que controlam a TD são provavelmente ancestrais e conservadas na maioria das plantas angiospermas, então sugere-se que essa característica possa ter surgido a partir de reconexão dessas vias (VANBUREN *et al.*, 2017)(OLIVER; TUBA; MISHLER, 2000). Existem espécies de plantas que são naturalmente tolerantes a estresses abióticos, como a *Oropetium thomaeum*, que pode

sobreviver a perdas maiores que 85% da água celular em períodos de estresse hídrico (BARTELS, D; MATTAR, 2002). Tentar relacionar por meio de homologia e redes de coexpressão gênica essas plantas com modelos melhores descritos e pesquisados, como *Arabidopsis thaliana*, e procurar entender as redes regulatórias e os mecanismos biológicos relacionados a detecção, respostas, adaptação e tolerância ao estresse pode dar pistas valiosas para o melhoramento genético de outras espécies de maior valor agronômico e de grande importância para a segurança alimentar mundial.

1.1.1. Fatores centrais de sinalização do ácido abscísico (ABA) incluem fosfatases e quinases

Proteínas quinase e proteínas fosfatase desempenham papéis importantes nas rotas de transdução de sinal em plantas. Um exemplo de via que é modulada por eventos de estresse abiótico é a do ácido abscísico (ABA), e um exemplo bem descrito é a rota de transcrição de sinal do hormônio ABA, que é dependente de proteínas de resistência à pirabactina 1/tipo resistência à pirabactina 1/componente regulador do receptor ABA (PYR/PYL/RCAR). Será explicado aqui, de forma resumida, o funcionamento dos fatores centrais de sinalização do ácido abscísico (ABA), pois eles serão abordados como estudos de caso nesse trabalho.

O ABA é o hormônio geralmente associado às principais respostas das plantas ao estresse (CHEN, Kong *et al.*, 2020). Esse importante fitormônio é responsável pela regulação de vários efeitos fisiológicos em plantas quando submetidas a alguns tipos de estresses abióticos (ZHU, Jian Kang, 2016). Portanto, são importantes as estratégias que ampliam o sinal molecular da percepção do estresse, fazendo com que a planta se antecipe e acelere seus mecanismos de defesa (KASUGA *et al.*, 1999).

A interação da proteína PYR/PYL/RCAR com PP2C-fosfatases é dependente do ABA para regular a atividade de proteínas serinas/treoninas quinase da família *Sucrose non-Fermenting Related Kinase2* (SnRK2). Estando o ABA ausente, essas PP2Cs conectam-se a SnRK2s e impedem a atividade da SnRK2-quinase, removendo grupos fosfato de uma região chamada alça de ativação dentro do domínio quinase (Figura 1A). A interligação com o ABA muda a conformação dos receptores PYR/PYL/RCAR de forma a permitir sua interação com PP2C e reprimir a atividade da PP2C-fosfatase. Liberando de inibição as SnRK2-quinasese e permitindo que proteínas SnRK2 fiquem livres para fosforilar muitas proteínas-alvo e fatores de transcrição que ligam os elementos de resposta ao ABA (ABFs) aos promotores gênicos para ativar a expressão gênica responsiva ao ABA (Figura 1B). A transdução de sinal do ABA é baseada na inversão do balanço entre as atividades da proteína PP2C-fosfatase e da SnRK2-quinase. A rota de sinalização dependente de PYR/PYL/RCAR possui papel importante no fechamento estomático em resposta ao ABA (TAIZ et al., 2017).

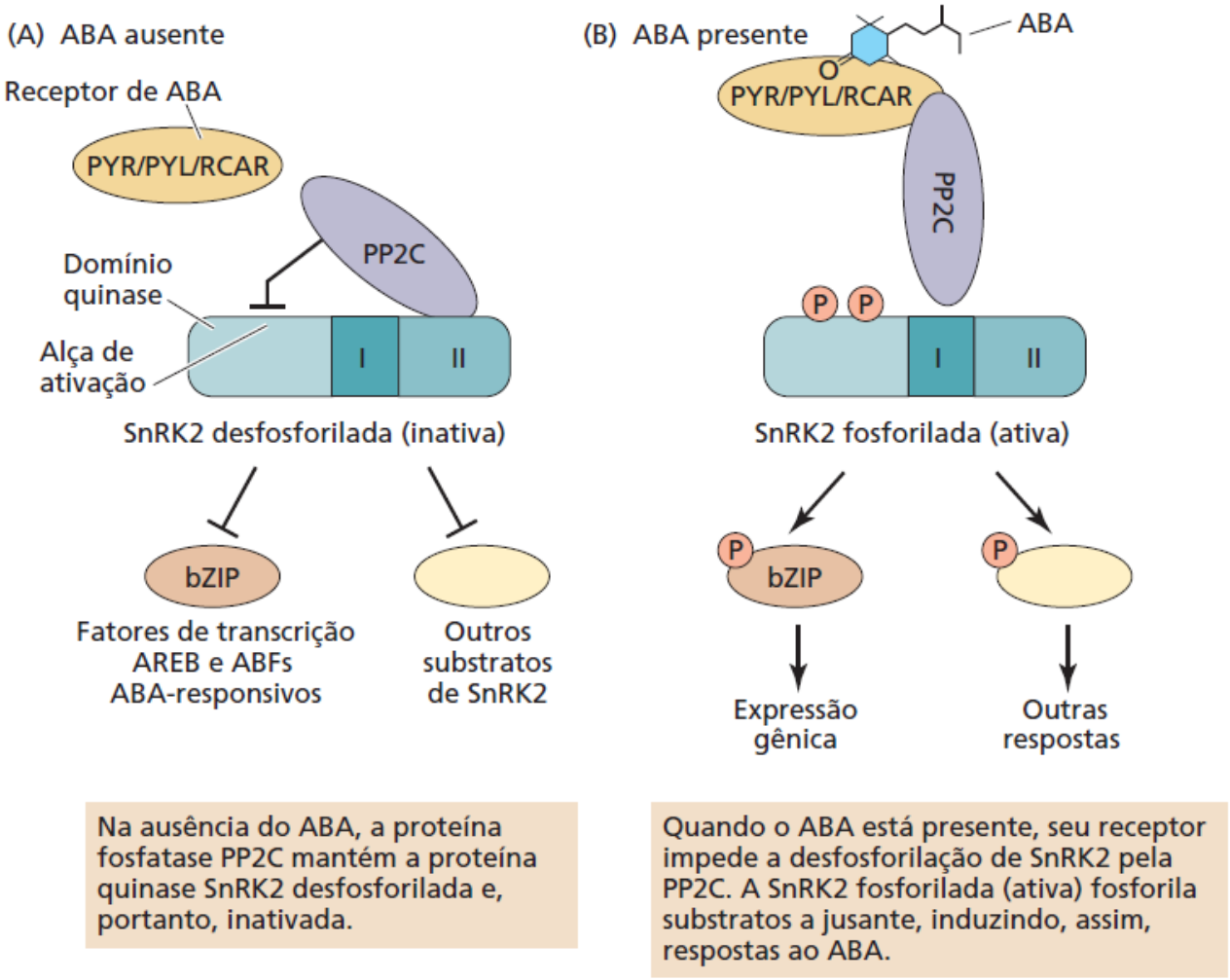

Figura 1 – Sinalização do ABA envolve quinases e fosfatases, retirado de (TAIZ *et al.*, 2017). A - Na ausência do ABA, PP2C desfosforila e inativa a SnRK2-quinase. B - Na presença do ABA, PYR/PYL/RCAR interage com PP2C, bloqueando a ação da fosfatase e liberando SnRK2 da regulação negativa. A SnRK2 ativada fosforila fatores de transcrição ABA-responsivos (bZIP) e outros substratos desconhecidos, para induzir uma resposta ao ABA.

O ABA também é denominado um importante mensageiro que atua como mediador de sinalização para regular a resposta adaptativa das plantas a diferentes condições de estresse ambiental (SAH; REDDY; LI, 2016). A regulação desses componentes é crítica para controlar as respostas de plantas sob condições de estresse abiótico e constituem um sistema de sinalização central que mantém o crescimento de plantas em ambientes não ideais ao seu desenvolvimento (CHEN, Kong *et al.*, 2020).

1.1.2. Plantas com importantes características de tolerância a estresses abióticos

Plantas com reconhecida tolerância a estresses abióticos podem ser alvos interessantes de estudos devido, por exemplo, a neofuncionalização de genes que podem potencializar as respostas/adaptação/tolerância dessas plantas diante de estresses (VANBUREN *et al.*, 2017). Serão apresentados aqui 5 organismos que fazem parte desse estudo e possuem importantes características de resistência a seca e/ou desidratação: *Boea hygrometrica, Populus simonii, Pearl millet(Cenchrus americanus), Oropetium thomaeum, e Sorghum bicolor.*

A *Boea hygrometrica* (XIAO *et al.*, 2015) possui a característica de "secar sem morrer", que é essencial na evolução das plantas terrestres, portanto é um importante modelo de planta com essa particularidade. As plantas de ressurreição, como podem ser chamadas, são excelentes sistemas modelo para o estudo dos mecanismos pelos quais as plantas sobrevivem à desidratação e para a identificação de genes que poderiam aumentar a tolerância à seca das safras por métodos biotecnológicos (BARTELS, Dorothea; SALAMINI, 2001).

A *Populus simonii* é uma importante árvore amplamente distribuída no Hemisfério Norte e com longa história de cultivo (WU, Hainan *et al.*, 2020). Ela se distribui pela China e comumente aparecem em regiões de desertificação, possuindo características de fácil enraizamento, tolerância à seca, resistência ao frio e ampla adaptação (ZHU, Jialei *et al.*, 2018).

O *Pearl millet (Cenchrus americanus)* (VARSHNEY *et al.*, 2017) é uma cultura de cereal resistente, cultivada principalmente em ambientes marginais nas regiões tropicais áridas e semiáridas da Ásia e da África, é cultivado em áreas com chuvas muito limitadas (300–500 mm na maioria dos casos), onde culturas como milho ou sorgo têm grande probabilidade de falhar na maioria dos anos (VADEZ *et al.*, 2012). Além disso, características reprodutivas, fisiológicas e adaptação climática inteligente do *Pearl millet* tornam esta cultura bem adequada para o crescimento em condições adversas, incluindo baixa fertilidade do solo, alto pH, alta saturação de Al3+, baixa umidade, alta temperatura, alta salinidade e precipitação limitada (VARSHNEY *et al.*, 2017)..

A gramínea *Oropetium thomaeum* (VANBUREN *et al.*, 2015) é um modelo emergente para tolerância à seca extrema (VANBUREN *et al.*, 2017). Ela é resiliente à secagem extrema e prolongada e deve ter genes envolvidos nos mecanismos moleculares relacionados ao controle deste fenótipo (VANBUREN *et al.*, 2018).

O *Sorghum bicolor* (MCCORMICK *et al.*, 2018) é uma grama C4 com características de tolerância à seca que pode ser usada para a produção de grãos, forragem, açúcar e biomassa lignocelulósica e um modelo genético para gramíneas C4 devido entre outras coisas ao seu genoma relativamente pequeno (aproximadamente 800 Mbp).

1.2. Recuperação de proteínas de função desconhecidas

A anotação funcional de proteínas visa identificar a função biológica de sequências através de análise de bioinformática e é uma importante forma de tentar inferir função nelas. Dois exemplos de softwares bastante utilizados para essa análise são o Diamond e o Interproscan. Pode-se considerar que sequencias de proteínas que não possuem correspondências (*match*) encontradas em nenhum dos dois softwares são caracterizadas como proteínas de função desconhecidas (PFD).

O Diamond é um software *open source* para alinhamento de sequências de proteínas contra um banco de dados de referência, como por exemplo o NCBI *nonredundant* (NCBI-nr). Ele usa indexação dupla e pode ser até 20.000 vezes mais rápido que o Blastx (ALTSCHUL et al., 1997), além disso, possui similar grau de sensibilidade (BUCHFINK; XIE; HUSON, 2014).

O Interproscan é um software que fornece análise funcional de sequências de proteínas e prediz a presença de domínios e sites importantes, ele foi criado para unir informações de múltiplos bancos de dados secundários de famílias de proteínas, domínios e *sites* funcionais (QUEVILLON *et al.*, 2005) e dessa forma tentar predizer a função de PFDs e/ou Domínios de Função Desconhecidas (DFDs). Cada um dos bancos de dados tem seu próprio foco biológico, método de produção de assinaturas e processamento das correspondências de assinatura. A diversidade de abordagens ajuda a garantir que as anotações sejam o mais abrangentes possível (FINN *et al.*, 2017).

1.3. Homologia

Nos últimos anos é cada vez maior o número de publicações de genomas completos de diversas espécies, isso abre grandes possibilidades para estudos funcionais em busca de características desejáveis, como a tolerância estresses abióticos, assim como analisar a dinâmica evolutiva entre as espécies, diante, por exemplo, da conservação de genes entre elas. Proteínas homólogas são aqueles que compartilham um ancestral em comum (KOONIN, 2005). A detecção e inferência de relações de homologia entre sequencias é fundamental para muitos aspectos da pesquisa biológica, e fornecem uma estrutura coerente para a extrapolação do conhecimento biológico entre os organismos, podendo sustentar, dessa forma, a anotação de genomas e transcriptomas (EMMS; KELLY, 2015)(DESSIMOZ *et al.*, 2012) (LI, Li; STOECKERT; ROOS, 2003).

As duas principais classes de homólogos são os ortólogos e os parálogos e eles são muito importantes para criação de grupos de ortólogos (LI, Li; STOECKERT; ROOS, 2003). Umas das primeira definições para essas classes foi dado por (FITCH, 1970), ortólogos ("orto" significando

"exato") são genes derivados por especiação, enquanto parálogos ("para" significando "ao lado" ou "próximo a") são genes que evoluíram por meio da duplicação. Então, ortólogos são sempre encontrados em espécies diferentes, enquanto que parálogos podem ser encontrados na mesma espécie ou em espécies diferentes, quando o evento de duplicação foi precedido de um evento de especiação. Um grupo de homólogos é o grupo de proteínas, sejam elas parálogas ou ortólogas, mas vale salientar, que esses grupos de homólogos serão referidos aqui, simplesmente, como grupo de ortólogos ou no inglês *orthogroup*. O aspecto da ortologia que é mais importante para pesquisadores do genoma e biólogos em geral é a expectativa de que os genes ortólogos sejam responsáveis por funções equivalentes em diferentes organismos(GABALDÓN; KOONIN, 2013).

A transferência de anotação funcional baseia-se na “conjectura da função de ortólogos”: ortólogos realizam funções idênticas, ou mais precisamente, biologicamente equivalentes em diferentes organismos, por outro lado parálogos normalmente divergem após a duplicação (GABALDÓN; KOONIN, 2013). No entanto, existem estudos afirmando que parálogos em um mesmo organismo estão mais próximos funcionalmente que ortólogos em diferentes organismos (NEHRT *et al.*, 2011) e inclusive sugerem que a inclusão de parálogos em grupos ortólogos pode oferecer mais e melhores previsões de funções de proteínas, pois a combinação de ortólogos e parálogos melhora consistentemente a anotação funcional em comparação com a predição baseada apenas em ortólogos (STAMBOULIAN *et al.*, 2020).

Podemos classificar em dois os príncipais métodos para inferencia de ortologia. Um deles, aborda o problema inferindo relações de pares entre genes em duas espécies e, em seguida, estendendo a ortologia a várias espécies, identificando conjuntos de genes abrangendo essas espécies em que cada par de genes é um ortólogo. Métodos populares que adotam essa abordagem incluem MultiParanoid (ALEXEYENKO *et al.*, 2006) e OMA (ALTENHOFF *et al.*, 2011). O outro método, tenta realizar a descoberta de grupos ortólogos completos. Ele é o utilizado pelo OrthoMCL (LI, Li; STOECKERT; ROOS, 2003), software mais utilizado para montagem de grupos ortólogos, e pelo OrthoFinder (EMMS; KELLY, 2015). Os dois softwares tentam calcular pontuações de similaridade de sequências, usando softwares como BLAST (ALTSCHUL *et al.*, 1997) ou Diamond (BUCHFINK; XIE; HUSON, 2014), em várias espécies e depois usam o algoritmo MCL (VAN DONGEN, 2000) para agrupamento de ortólogos.

No entanto, nesse segundo método existe um importante viés com relação as pontuações obtidas pelos programas de análise de similaridade de sequencias e o comprimento das sequencias. Pois sequências curtas não podem produzir grandes escores ou baixos *e-value*, enquanto sequências longas produzem muitos *hits* com pontuações melhores do que aqueles para os melhores hits de sequências curtas, dessa forma, sequências curtas sofrem com baixa taxa de *recall* (muitas sequências

curtas falham ao serem atribuídas a um *ortogroup*) e sequências longas sofrem com baixa precisão (muitas sequências longas são atribuídas ao *ortogroup* incorreto) (EMMS; KELLY, 2015).

Se não houver uma normalização dessa pontuação antes do agrupamento, é possível que exista um grande número de proteínas ausentes em grupos que contêm sequencias curtas e sequencias agrupadas incorretamente em grupos que contêm sequências longas. Além disso, as taxas de evolução de sequência é variável entre genes e freqüentemente levam a erros de falso positivo e falso negativo (LAFOND; MEGHDARI MIARDAN; SANKOFF, 2018)(FITCH, 1970). Esses erros podem ser atenuados pela análise de árvores filogenéticas de genes (DALQUEN; DESSIMOZ, 2013), pois elas podem distinguir taxas de evolução de sequência variável (comprimento dos ramos) da ordem em que as sequências divergiram (topologia da árvore) e, portanto, esclarecer as relações de ortologia e paralogia (EMMS; KELLY, 2019).

1.3.1. OrthoFinder

O OrthoFinder é uma plataforma rápida, precisa e abrangente para genômica comparativa, ele cria grupos de ortólogos, infere árvores gênicas e de espécies, além de fornecer estatísticas abrangentes para análises genômicas comparativas e é simples de usar. (EMMS; KELLY, 2015) .

Uma visão geral do funcionamento do algorítmo do OrthoFinder aplicado a um pequeno conjunto de genes é exibido na figura 2, nela representamos dois grupos de ortólogos, inicialmente desconhecidos (Figura 2A), que o algoritmo precisa reconhecer ao final da análise.

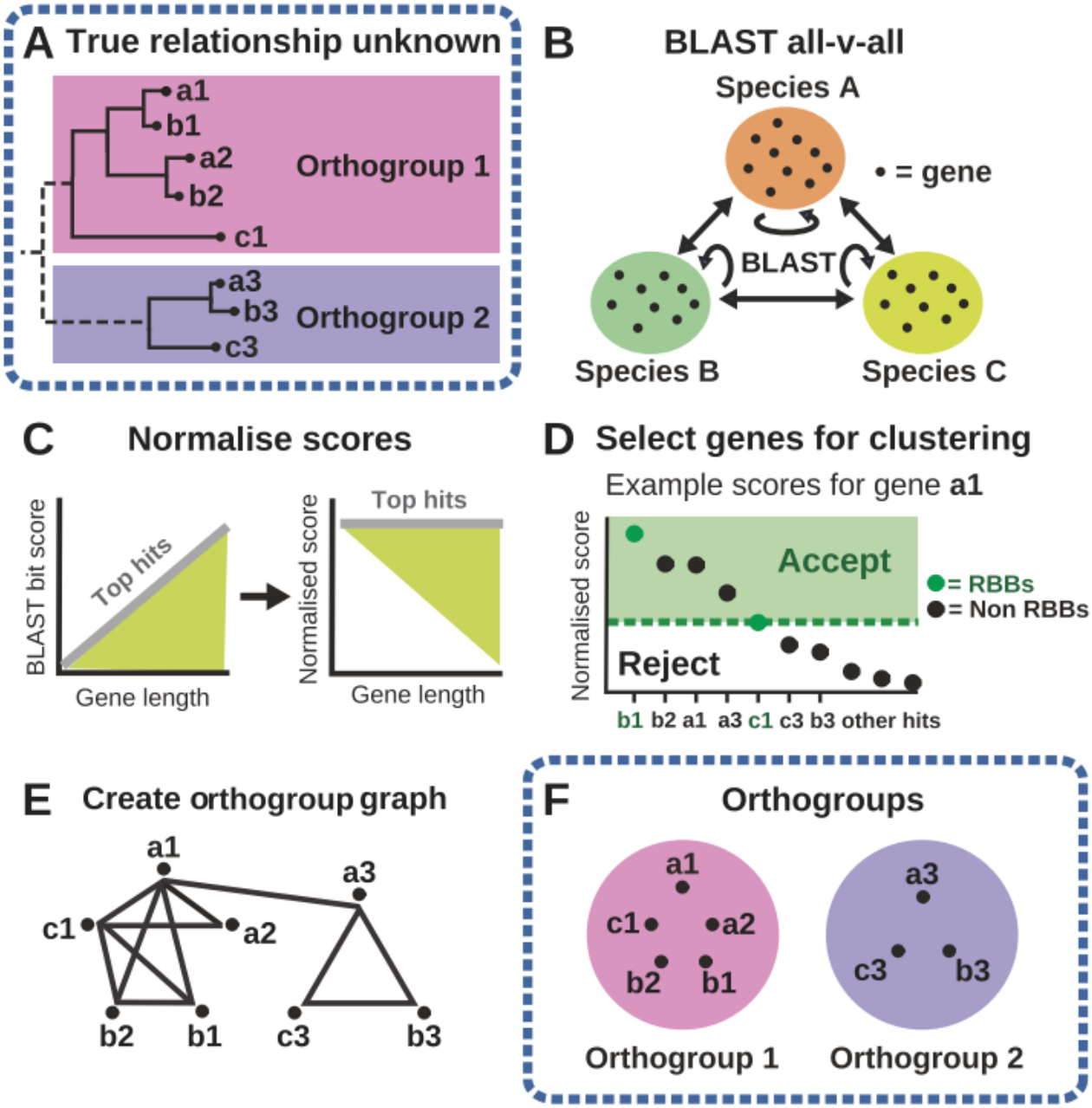

Figura 2 – Visão geral do funcionamento do algorítmo do OrthoFinder. Retirado de (EMMS; KELLY, 2015). A – Verdadeiras relações de homologia. B – Alinhamento de sequências todos contra todos. C – Normalização dos escores de alinhamento. D – Seleção dos genes para montagem do grafo. E – Criação do Grafo. F – Criação dos grupos de ortólogos.

O processamento dos dados inicia com uma análise de BLAST todos contra todos (Fig. 2B), na realidade a versão usada nas análises já usa o Diamond invés do BLAST. Na etapa seguinte (Fig. 2C) é realizada uma normalização para diminuir a dependência dos escores de similaridades (vide seção 1.3), encontrados na etapa anterior, com relação ao comprimento e distância filogenética dos genes, e então é criada uma nova pontuação chamada *Blast bit* que será utilizada para inferência

dos grupos de ortólogos. Em sequida (Fig. 2D) , de posse dos valores normalizados, são criados os limiares desses valores de similaridade de sequência normalizados de grupos de ortólogos usando RBNHs (*Reciprocal Best length-Normalized hit*) para definir o limite inferior de similaridade de sequência para possíveis genes homólogos de cada sequência de consulta (*query*). Na etapa seguinte (Fig. 2E) é construindo um grafo de grupo de ortólogos, os pares de possíveis genes homólogos são identificados e conectados no grafo com pesos dados pelas pontuações de BLAST *bit* normalizadas. Na última etapa do algoritmo (Fig. 2F) é usado o MCL para criar os grupos de ortólogos, usando o prâmetro de *infaltion index* padrão de 1,5. Os grupos de ortólogos criados pelo OrthoFinder possuem também parálogos, além dos ortólogos. (https://davidemms.github.io/OrthoFinder_tutorials/exploring-orthofinders-results.html).

1.4. Redes de coexpressão gênica

As consequências do estresse para o crescimento e produtividade da planta dependem da severidade (intensidade) do fator causador do estresse, bem como da duração da perturbação, da quantidade de vezes que a planta é submetida ao fator de estresse durante seu ciclo de cultivo, do estágio de desenvolvimento e do genótipo selecionado (MAIA; MORAES, 2015). Ademais, na prática, é comum a planta ser submetida a vários tipos de estresses combinados, portanto, pode ser interessante a avaliação de como os genes de comportam em diferentes tecidos e fases de desenvolvimento da planta quando submetidos a diferentes tipos de estresses, assim como, o estudo da interligação entre esses genes em rede através da correlação de expressão.

Atualmente, há uma grande disponibilidade de dados públicos de RNA-seq para perfis de expressão gênica (AMBROSINO *et al.*, 2020) (PROOST; KRAWCZYK; MUTWIL, 2017) como por exemplo, a tolerência a estresses abióticos. Aliado a isso, os avanços em poder computacional, abordagens estatísticas e ferramentas de bioinformática fornecem uma excelente plataforma para a descoberta de funções de genes em caminhos biológicos não resolvidos (TZFADIA *et al.*, 2016) (ALLEN *et al.*, 2012). Esses dados estão sendo explorados para tentar revelar respostas transcricionais a certos estímulos externos (ex.: estresses abióticos) e descobrir o perfil de expressão coordenada (coexpressão) de diferentes genes (USADEL *et al.*, 2009).

A análise da rede de coexpressão permite a identificação, agrupamento e exploração simultânea de múltiplos genes com padrões de expressão semelhantes em várias condições (genes coexpressos) (SERIN *et al.*, 2016) e é um dos métodos mais populares para lidar com conjuntos de dados de transcriptoma em grande escala (RAO; DIXON, 2019). Ao combinar abordagens genômicas tradicionais com redes de coexpressão, podemos expandir nosso conhecimento sobre como as vias

biológicas surgem ou são estendidas, e tentar transferir conhecimento funcional entre espécies (RUPRECHT *et al.*, 2017).

A análise de coexpressão assume que os genes cujos mRNAs são expressos de forma semelhante, isto é, genes que foram superexpressos (*upregulated*) ou subexpressos (*downregulated*), em muitas condições, por exemplo, diferentes tecidos, genótipos e durante estresses bióticos/abióticos, estão envolvidos em processos biológicos relacionados (USADEL *et al.*, 2009), portanto, usando a abordagem de culpa por associação (*guilt-by-association*), devem possuir funções biológicas equivalentes. Além disso, a análise de padrões de expressão em vários tecidos, estágios de desenvolvimento e condições podem lançar luz sobre quando e onde um gene é necessário, o que, por sua vez, também fornece pistas sobre a função do gene (PROOST; KRAWCZYK; MUTWIL, 2017).

Uma pontuação de correlação é definida para representar o nível de similaridade do padrão de expressão entre pares de genes, os que obtiverem uma pontuação acima de um determinado nível (*cutoff*) é considerado como coexpresso. Um dos métodos mais utilizados para mensurar essa correlação é o Coeficiente de Correlação de Pearson (do inglês *Pearson Cofficient Correlation*) (PCC) (STEUER *et al.*, 2002)(SONG; LANGFELDER; HORVATH, 2012)(RAO; DIXON, 2019). Depois que a pontuação de correlação de todos os pares de genes é medida, eles podem ser interligados em redes de acordo com os valores de coexpressão. Como essas redes costumam ficar muito grandes, é possível o uso de algoritmos de clusterização para formação de clusters de coexpressão menores e ainda mais interligados (correlacionados), o que pode facilitar sua visualização em ferramentas como, por exemplo, o software cytoscape (SMOOT *et al.*, 2011).

1.4.1. Large-scale Transcriptome Analysis Pipeline (LSTrAP)

O software *Large Scale Transcriptome Analysis Pipeline* (LSTrAP) v1.3 (PROOST; KRAWCZYK; MUTWIL, 2017) facilita e agiliza a construção da rede e *clusters* de coexpressão com os dados de RNA-seq. O LSTrAP é na realidade um um pipeline que envolve várias ferramentas de bioinformática conectadas e na figura 3 é apresentado de forma sumarizada o fluxo de trabalho desse pipeline.

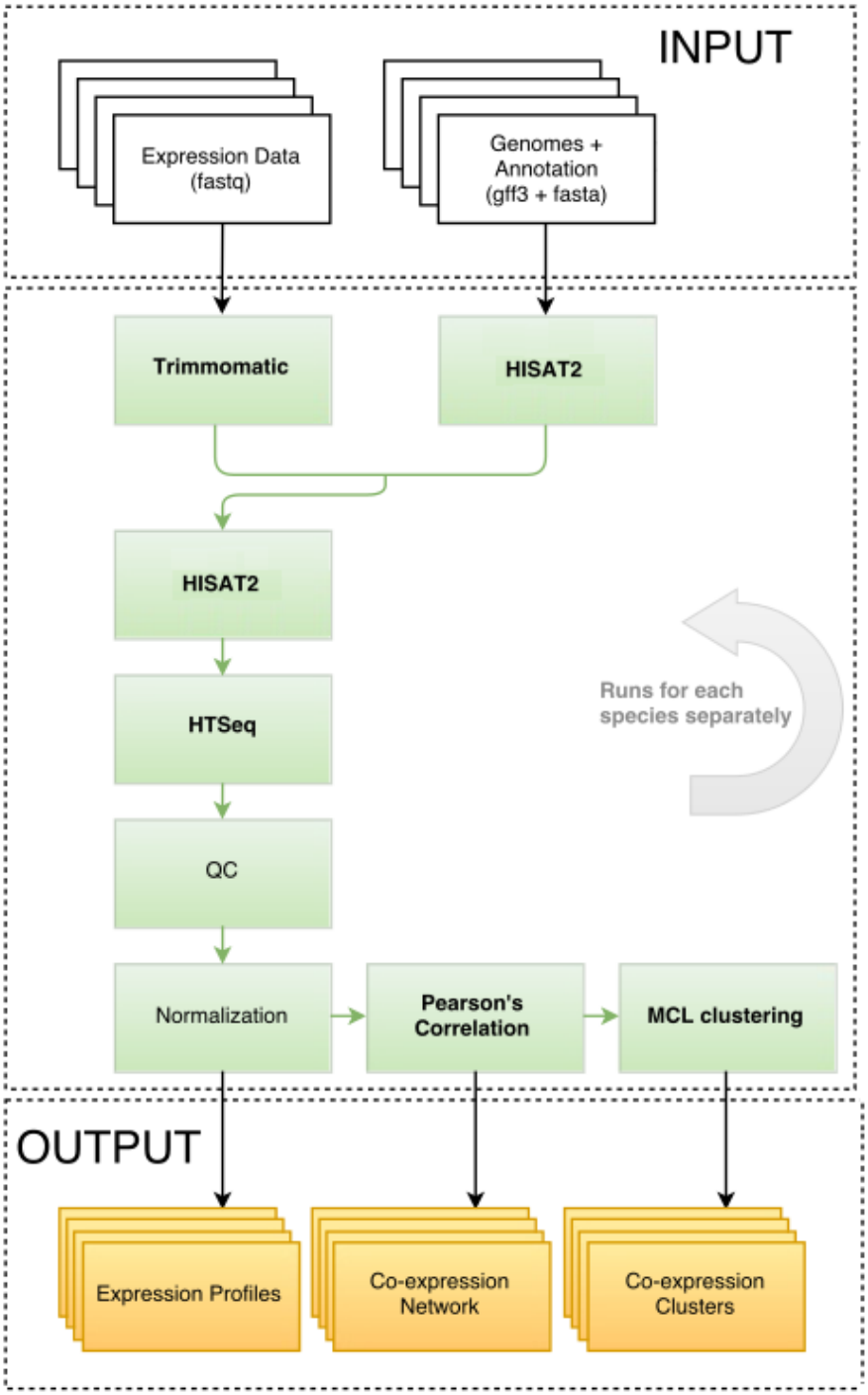

Figura 3 – Fluxo de trabalho do LSTrAP. Retirado de (PROOST; KRAWCZYK; MUTWIL, 2017)

Na primeira etapa do pipeline é criado um arquivo de índices para os dados do genoma de referência usando o hisat2-builder, isso vai tornar a etapa de mapeamento de *reads* mais eficiente. Na etapa seguinte é usado o Trimmomatic (BOLGER; LOHSE; USADEL, 2014) para tentar garantir que todas as amostras de RNA-seq sigam o mesmo padrão mínimo de qualidade, ele processa os fastq trimmando (cortando) as bases de baixa qualidade, e além disso, removendo as sequencias

adaptadores residuais do processo de sequenciamento. As *reads* trimmadas são mapeadas para o genoma indexado usando o HISAT2 (KIM; LANGMEAD; SALZBERG, 2015), que criará arquivos *Sequence Alignment Map* (SAM)/*Binary Alignment Map*(BAM) contendo o alinhamento de cada *read* com regiões do genoma. Para cada gene o número de *reads* mapeadas é contado usando o HTSeq-Count (ANDERS; PYL; HUBER, 2015), que produz para cada amostra um arquivo contendo as reads mapeadas por gene. O LSTrAP agrega esses arquivos em uma única matriz (m x n) contendo o valor de expressão para cada gene (m) em cada amostra (n). Uma etapa de controle de qualidade é inserida no pipeline para excluir amostras inadequadas que podem afetar negativamente a criação dos clusters de expressão. Os parâmetros de qualidade podem ser alterados dentro do arquivo data.ini, foram utilizados os valores padrões, que mais de 65% das *reads* devem ser mapeadas para o genoma e, dessas, pelo menos 40% sejam *reads* codificadoras de proteínas.

Antes de criar as redes e *clusters* de coexpressão é necessária normalização para diferenças na profundidade de sequenciamento entre as amostras e comprimento do gene. O LSTrAP normalizará a matriz de expressão usando duas abordagens muito comuns; Transcrições por quilobase por milhão (TPM) e leituras por quilobase por milhão (RPKM) (MORTAZAVI *et al.*, 2008). Então, finalmente finalmente serão criadas as redes e *clusters* de coexpressão gênica, utilizando operações de matrizes de NumPy (WALT; COLBERT; VAROQUAUX, 2011), para calcular coeficientes de correlação de Pearson (PCC), que foi considerado um dos mais eficazes para estudos de coexpressão baseados em RNA-Seq (BALLOUZ; VERLEYEN; GILLIS, 2015), com base no TPM matriz de expressão normalizada. O valor do PCC varia de -1,0 a 1,0, onde zero significa nenhuma correlação, valores positivos indicam correlação positiva e valores negativos correspondem à anticorrelação. O resultado é uma tabela que descreve para cada gene os 1000 genes coexpressos mais fortes no conjunto de dados. Todos os pares com um valor PCC > 0,7 (a configuração recomendada ao usar MCL em dados de coexpressão) são armazenados separadamente e representam a rede de coexpressão global, que é agrupada em grupos de genes coexpressos usando o algoritmo MCL (ENRIGHT; VAN DONGEN; OUZOUNIS, 2002).

CAPÍTULO 2: OBJETIVOS

2.1. Objetivo Geral

Desenvolver e disponibilizar a versão 2 da ferramenta PlantAnnot (PlantAnnot2).

2.2. Objetivos específicos

1) Desenvolver e disponibilizar uma segunda versão online do sistema web PlantAnnot com as seguintes novidades:

1.1) Inserir no sistema 14 novas espécies, incluindo 2 novos genomas com características de tolerância a estresses abióticos;

1.2) Inserir novos dados de RNA-seq incluindo diversos tecidos, níveis de desenvolvimento da planta submetidos a ensaios de estresses abióticos;

1.3) Uso do OrthoFinder para definição de homólogos.

2) Testar/validar o funcionamento do software:

2.1) Verificar/validar se PFDs recuperadas têm relação com estresses abióticos;

2.2) Verificar se o enriquecimento de vias corrobora os resultados relacionados a estresses abióticos;

2.3) Realizar e analisar estudos de caso com espécies sabidamente resistentes à estresses abióticos/ressurgentes;

2.4) comparar resultados com PlantAnnot.

CAPÍTULO 3: MATERIAIS E MÉTODOS

3.1. Infraestrutura

Para realizar todas as análises, processamento e armazenamento dos dados desse projeto, foi utilizada a infraestrutura de Tecnologia da Informação (TI) do Laboratório Multiusuário de Bioinformática (LMB) e do Núcleo de Tecnologia da Informação (NTI) da Embrapa Informática Agropecuária, que ficam localizados em Campinas-SP. O LMB disponibiliza recursos computacionais para armazenamento e processamento de alto desempenho para serem utilizados no âmbito de projetos da Embrapa e de parceiros priorizando, mas não se limitando, àqueles que exigem a análise de dados biológicos (CINTRA, 2016).

Esse laboratório possui uma robusta infraestrutura computacional composta por um cluster de servidores que juntos somam aproximadamente 664 núcleos de processamento, 7 TB de memória do tipo *Random Access Memory* (RAM), além de 44 TB de espaço bruto de armazenamento interno. Além disso, existem mais 210 TB úteis dedicados para armazenamento de dados em 2 equipamentos exclusivos para esse fim, *storages*. A interconexão desses equipamentos é realizada por meio 3 redes internas: uma rede ethernet 1Gbit/s para controle, uma rede ethernet de 10Gbit/s para tráfego de dados

e uma rede ótica de 8Gbit/s também para trafego de dados. O inventário completo da infraestrutura de TI existente no LMB está disponível através do link https://www.lmb.cnptia.embrapa.br/web/lmb/hardware. Foram utilizadas ainda 2(duas) máquinas virtuais disponibilizadas pelo NTI, em uma delas está hospedado o PlantAnnot2 disponibilizado publicamente (8 cores, 16GB de RAM e 3TB de armazenamento) e a outra foi usada como ambiente de desenvolvimento e testes dessa aplicação (24 cores, 32GB de RAM e 4TB para armazenamento).

O sistema de *batch Sun Grid Engine* (SGE) (GENTZSCH, 2001) é utilizado para o gerenciamento do processamento no cluster (CINTRA, 2016). Todos os softwares usados foram instalados localmente em máquinas virtuais com sistema operacional CentOS Linux release 7.8.2003 (https://www.centos.org/) para disparar as análises e Ubuntu 18.04.4 LTS (https://ubuntu.com/) para instalação da ferramenta web e banco de dados.

3.2. Algoritmo e fluxo de tarefas do projeto

Como não há informações sobre a função das PFDs, uma maneira de inferir função é vinculá-las a outras moléculas usando grupos de ortologia e a abordagem de culpa por associação (*Guilt-by-association*) (Figura 4). Desta forma, membros de um determinado grupo ortólogo que já possuem anotação e/ou têm domínios de proteínas caracterizados, podem ser usados como um *proxy* para inferir a função em PFDs por associação. Além disso, sempre que um determinado PFD faz parte de um grupo ortólogo no qual algum membro, necessariamente de pelo menos uma das plantas *Arabidopsis thaliana* (Ath), *Oryza sativa* (Osa), *Glycine max* (Gma) e *Zea mays* (Zma), tem seu mRNA compondo um grupo de coexpressão, então, por associação, a PFD pode supostamente também está relacionado com resposta a estresses abióticos em plantas (Figura 5) .

Figura 4 – Algoritmo para anotação de PFDs (VIANA, Marcos José Andrade; ZERLOTINI; DE ALVARENGA MUDADU, 2021). A PFD deve pertencer a um grupo de ortólogo que possua proteínas cujo(s) mRNA(s) faça(m) parte de grupo de coexpressão relacionado a estresse abiótico.

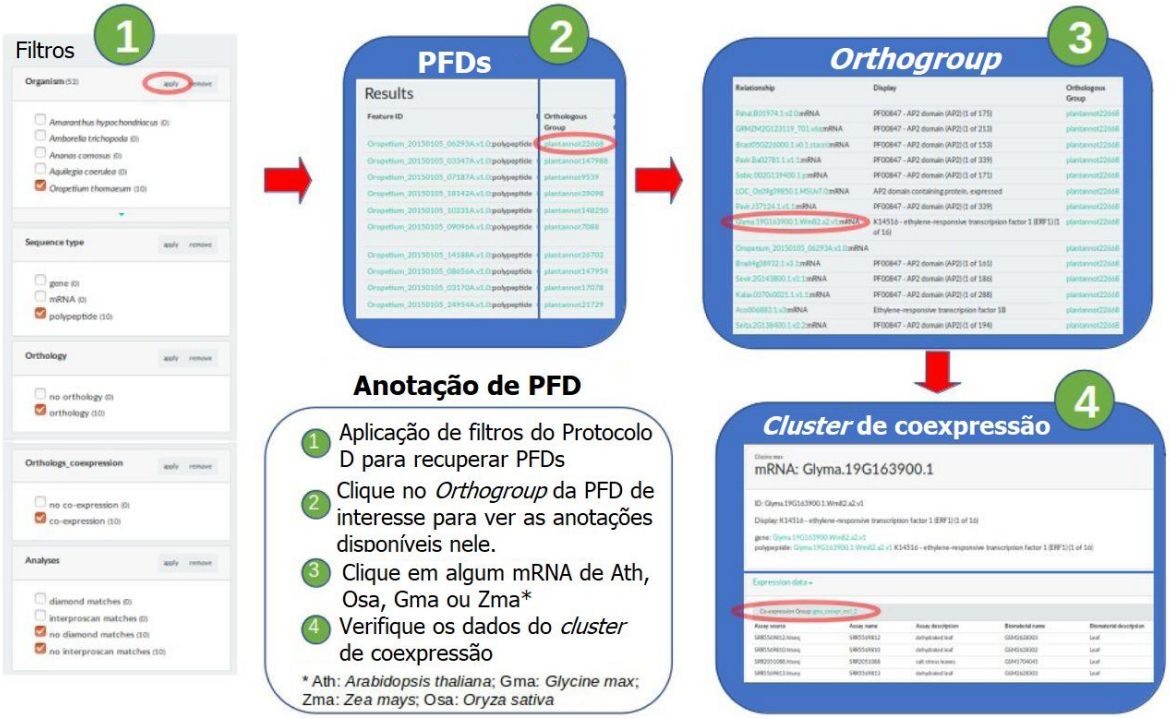

Figura 5 – Busca por PFD usando ortologia e coexpressão gênica. Retirado de (VIANA, Marcos José Andrade; ZERLOTINI; DE ALVARENGA MUDADU, 2021)

O fluxo de trabalho desse projeto (figura 6) foi composto basicamente por 6 etapas principais, são elas:

1 – Download e tratamento de dados de genômicos, esses dados serviram de entrada para as análises 2 e 3 da figura 6;

2 – Caracterização de Proteínas de Função Desconhecidas com softwares Diamond (BUCHFINK; XIE; HUSON, 2014) e Interproscan (JONES *et al.*, 2014), a ideia é que todas as proteínas que não tiveram correspondência (*no match*) no NCBI-nr e/ou nos bancos de dados do Interproscan, serão consideradas PFDs;

3 – Criação de grupos de homólogos com OrthoMCL (LI, Li; STOECKERT; ROOS, 2003) e OrthoFinder (EMMS; KELLY, 2019);

4 – Download de dados de perfis de expressão gênica (RNA-seq) do NCBI/GEO de ensaios em plantas sob estresses abióticos. Esses dados serviram de entrada para análise 5 da figura 6;

5 – Criação de rede e *clusters* de coexpressão com o software LSTrAP (PROOST; KRAWCZYK; MUTWIL, 2017);
6 – Carregamento dos dados e resultados das análises no Machado (DE ALVARENGA MUDADU; ZERLOTINI, 2020) e configuração da interface web para navegação e visualização desses dados.

Figura 6 – Fluxo de trabalho do PlantAnnot2, adaptado de (VIANA, Marcos José Andrade; ZERLOTINI; DE ALVARENGA MUDADU, 2021). 1 – Download e tratamento de dados de genômicos 2 – Caracterização de PFDs; 3 – Criação de grupos de homólogos; 4 – Download de dados de RNA-seq; 5 – Criação de redes e *clusters* de coexpressão; 6 – Carregamento dos dados e resultados das análises no Machado.

3.3. Download e tratamento de dados genômicos

Foram baixados 67 genomas, listados na Tabela 1 (Figura 6.1), dos quais, 64 foram baixados do Phytozome v12.1, mais os genomas da *Boea hygrométrica* (XIAO *et al.*, 2015) baixado do NCBI, do *Pearl millet (Cenchrus americanus)* (VARSHNEY *et al.*, 2017) baixado do *International Pearl Millet* do *Genome Sequencing Consortium* (IPMGSC) (ICRISAT, 2020) e do *Populus simonii* (WU, Hainan *et al.*, 2020) baixado diretamente da seção de material suplementar do artigo de divulgação da montagem de seu genoma na *Genetics Society of America* (GSA) *Journal* (GSA, [*s. d.*]).

Organismo	Versão do Genoma	Origem do Genoma
Amaranthus hypochondriacus	v1.0	Phytozome v12.1
Amborella trichopoda	v1.0	Phytozome v12.1
Ananas comosus	v3	Phytozome v12.1
Aquilegia coerulea	v3.1	Phytozome v12.1
Arabidopsis halleri	v1.1	Phytozome v12.1
Arabidopsis lyrata	v2.1	Phytozome v12.1
Arabidopsis thaliana	TAIR10	Phytozome v12.1
Boea hygrometrica	GCA_001598015.1	NCBI
Boechera stricta	v1.2	Phytozome v12.1
Brachypodium distachyon	v3.1	Phytozome v12.1
Brachypodium stacei	v1.1	Phytozome v12.1
Brassica oleracea	v1.0	Phytozome v12.1
Brassica rapa	FPsc v1.3	Phytozome v12.1
Capsella grandiflora	*v1.1*	Phytozome v12.1
Capsella rubella	*v1.0*	Phytozome v12.1
Carica papaya	*ASGPBv0.4*	Phytozome v12.1
Chlamydomonas reinhardtii	v5.5	Phytozome v12.1
Citrus clementenina	v1.0	Phytozome v12.1
Citrus sinensis	v1.1	Phytozome v12.1
Coccomyxa subellipsoidea	*C-169v2.0*	Phytozome v12.1
Cucumis sativus	v1.0	Phytozome v12.1
Daucus carota	v2.0	Phytozome v12.1
Dunaliella salina	v1.0	Phytozome v12.1
Eucalyptus grandis	v2.0	Phytozome v12.1
Eutrema salsugineum	v1.0	Phytozome v12.1
Fragaria vesca	v1.1	Phytozome v12.1
Glycine max	Wm82.a2.v1	Phytozome v12.1
Gossypium raimondii	v2.1	Phytozome v12.1
Kalanchoe fedtschenkoi	v1.1	Phytozome v12.1
Kalanchoe laxiflora	v1.1	Phytozome v12.1
Linum usitatissimum	v1.0	Phytozome v12.1
Malus domestica	v1.0	Phytozome v12.1
Manihot esculenta	v6.1	Phytozome v12.1
Marchantia polymorpha	v3.1	Phytozome v12.1
Medicago truncatula	Mt4.0v1	Phytozome v12.1
Micromonas pusilla	*CCMP1545v3.0*	Phytozome v12.1

Micromonas sp.	*RCC299v3.0*	Phytozome v12.1
Mimulus guttatus	v2.0	Phytozome v12.1
Musa acuminata	v1	Phytozome v12.1
Oropetium thomaeum	v1.0	Phytozome v12.1
Oryza sativa	v7_JGI	Phytozome v12.1
Ostreococcus lucimarinus	v2.0	Phytozome v12.1
Panicum hallii	v2.0	Phytozome v12.1
Panicum virgatum	v1.1	Phytozome v12.1
Pearl millet	v1.1	IPMGSC
Phaseolus vulgaris	v2.1	Phytozome v12.1
Physcomitrella patens	v3.3	Phytozome v12.1
Populus Simonii	v2.0	Material suplementar do artigo
Populus trichocarpa	v3.0	Phytozome v12.1
Prunus pérsica	v2.1	Phytozome v12.1
Ricinus communis	v0.1	Phytozome v12.1
Salix purpúrea	v1.0	Phytozome v12.1
Selaginella moellendorffii	v1.0	Phytozome v12.1
Setaria itálica	v2.2	Phytozome v12.1
Setaria viridis	v1.1	Phytozome v12.1
Solanum lycopersicum	iTAG2.4	Phytozome v12.1
Solanum tuberosum	v4.03	Phytozome v12.1
Sorghum bicolor	v3.1.1	Phytozome v12.1
Sphagnum fallax	v0.5	Phytozome v12.1
Spirodela polyrhiza	v2	Phytozome v12.1
Theobroma cacao	v1.1	Phytozome v12.1
Trifolium pratense	v2	Phytozome v12.1
Vitis vinífera	Genoscope.12X	Phytozome v12.1
Volvox carteri	v2.1	Phytozome v12.1
Zea mays	284_AGPv3	Phytozome v12.1
Zea maysPH207	v1.1	Phytozome v12.1
Zostera marina	v2.2	Phytozome v12.1

Tabela 1- Genomas de organismos do PlantAnnot2.

Foi realizado o *download* dos dados do Phytozome (GOODSTEIN *et al.*, 2012) usando a *Application Programming Interface* (API) disponibilizada pelo próprio site e seguindo os procedimentos descritos na seção "*Download with API*". Para todos os genomas, foram baixados

além dos arquivos *Generic Feature Format Version 3* (GFF3), os arquivos Fastas de montagem (*assembly*), de mRNA e de proteínas, entre outros arquivos. Antes de realizar as análises, foi necessário realizar um tratamento nos dados para normalizar o identificador único (ID) de todos os mRNAs existentes no arquivo *GFF3* de cada organismo para ser igual ao identificador da sequência, código alfa numérico que vem logo após o sinal ">", no arquivo fasta. Foi necessário realizar essas alterações para relacionar os dados genômicos e resultados das análises através de um identificador único (ID).

3.4. Recuperação de PFDs

Foram utilizados dois softwares para tentar recuperar todas as PFDs (Figura 6.2) dos organismos envolvidos nesse projeto: o *Double Index Alignment of Next-generation Sequencing Data* (Diamond) v0.9.24 e o Interproscan v5.36-75.0.

Diamond

Foi utilizado o Diamond v0.9.24 para alinhar as sequencias de consulta (*query*), que são as sequencias de todas as proteínas de todos os organismos dessa análise, contra as sequencias do banco de dados *nonredundant* do NCBI (NCBI-nr), que foi baixado em setembro de 2019 do endereço ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz. Depois de descompactado com *gunzip*, o NCBI-nr foi indexado usando o comando abaixo, e a saída desse comando foi o arquivo nr.dmnd que é o banco de dados NCBI-nr indexado.

"makeblastdb -dbtype prot -out nr -in nr"

Finalmente, foi executado o Diamond usando o comando abaixo.

"diamond blastp -k 1 -f 5 -q arquivo.protein.fa -d nr.dmnd -o arquivo.protein.fa.nr.dmnd.result"

Onde o parâmetro *"blastp"* informa as sequencias envolvidas nesse análise são de proteínas, "*-k 1"* que devem ser retornadas apenas a sequência do NCBI-nr que ficou melhor alinhada com a sequência de consulta (*query*), "*-f 5"* que o formato do arquivo de saída será o BLAST *Extensible Markup Language* (XML), "*-q arquivo.protein.fa"* o nome do arquivo fasta onde estão as sequencias que serão pesquisadas no NCBI-nr, as sequencias de consulta (*queries), "-d nr.dmnd"*

informa o banco de dados NCBI-nr indexado, "*-o arquivo.protein.fa.nr.dmnd.result"* informa o nome do arquivo de saída do comando. O parâmetro *expected-value* (*e-value)* não foi informado no comando, mas o valor padrão usado é 0,001.

O arquivo de saída trará todas as proteínas de consulta (*queries*) com informações que indicam se o alinhamento foi realizado, como o identificador da sequência melhor alinhada no NCBI-nr, o *e-value* e o *score* de alinhamento, ou se não houve alinhamento, nenhuma das informações anteriores será fornecida.

Interproscan

Foi utilizado o Interproscan v5.36-75.0, e o procedimento de instalação e configuração pode ser consultado diretamente na página oficial da ferramenta no *github* (https://github.com/ebi-pf-team/interproscan/wiki). Essa ferramenta realizou buscas sobre funções de proteínas nos seguintes bancos de dados: CDD-3.17 (MARCHLER-BAUER *et al.*, 2015), Coils-2.2.1, Gene3D-4.2.0 (LEES *et al.*, 2012), Hamap-2019_01 (PEDRUZZI *et al.*, 2015), MobiDBLite-2.0 (POTENZA *et al.*, 2015), PANTHER-14.1 (MI *et al.*, 2016), Pfam-32.0 (FINN *et al.*, 2016), PIRSF-3.02 (WU, Cathy H *et al.*, 2004), PRINTS-42.0 (ATTWOOD *et al.*, 2012), ProSitePatterns-2019_01 (HULO *et al.*, 2004), ProSiteProfiles-2019_01 (HULO *et al.*, 2004), SFLD-4,SMART-7.1 (LETUNIC; DOERKS; BORK, 2012), SUPERFAMILY-1.75 (OATES *et al.*, 2015), TIGRFAM-15.0 (HAFT *et al.*, 2013). O comando utilizado está abaixo e foi executado para cada fasta de proteínas dos 67 organismos.

"interproscan.sh -i fasta.protein.fa -dp -clusterrunid id_job_cluster"

Onde o parâmetro "*-i fasta.protein.fa"* informa o arquivo fasta de proteínas do organismo, o "*-dp"* desativa o serviço de pesquisa de *match* pré-calculado disponibilizada por meio de uma pesquisa web ao Interproscan, todas as pesquisas por *matches* foram realizadas localmente (JONES *et al.*, 2014) e "*-clusterrunid id_job_cluster"* é o parâmetro onde informamos o nome do *job* que será submetido ao cluster SGE. A saída do comando acima sé feita em 3 formatos diferentes de arquivos: gff3, *tab-separated values* **(**tsv) e xml, será utilizado apenas esse último nas análises. Essa saída informará se a sequência deu *hit* em algum dos bancos de dados pesquisado, informando o nome de cada banco e o domínio de proteína que teve correspondência com a sequência consultada.

3.5. Criação de grupos de homólogos

Foram realizados testes para verificar qual software seria melhor no agrupamento de proteínas homólogas (Figura 6.3) e que teria melhor resultado na recuperação de PFDs relacionadas a estresses abióticos. Nesse sentido, foram avaliados os softwares OrthoMCL (LI, Li; STOECKERT; ROOS, 2003) e OrthoFinder (EMMS; KELLY, 2015).

OrthoMCL

Foi realizada análise com o OrthoMCL v2.0.9 utilizando os arquivos fasta de proteínas dos 67 organismos (ver seção 3.3). A análise com o OrthoMCL é basicamente dividida em 5 fases:

1. Filtragem de sequências;
2. Alinhamento de todas as sequencias de todos os orgnismos par a par usando o Diamond;
3. Cálculo do percentual de correspondência para cada *hit* do Diamond;
4. Encontrar potenciais pares de inparalog, ortólogos e co-ortólogos;
5. Usar programa MCL para agrupar os pares em grupos.

Para verificar detalhes do funcionamento do algoritmo e todos os passos de instalação e execução do OrthoMCL acesse o endereço web https://orthomcl.org/common/downloads/software/v2.0/UserGuide.txt.

OrthoFinder

Foi utilizado o OrthoFinder v.2.3.12 e sua instalação foi realizada seguindo as orientações presentes no endereço web https://github.com/davidemms/OrthoFinder. Antes de executar a análise, foi necessário realizar uma mudança nos arquivos fasta de proteínas dos organismos, e foi aproveitado o script orthomclAdjustFasta do orthoMCL para inserir um código de 3 letras iniciais do organismo juntamente com o caractere "|" antes das identificações de todas as sequencias, esse código é chamado de *taxon_code* no orthomcl. Pois o script de carregamento (*loading*) do Machado é compatível só com saídas do OrthoMCL. Abaixo temos o exemplo do comando utilizado para ajustar o fasta de proteínas do organismo *Arabidopsis lyrata*:

"orthomclAdjustFasta Aly Alyrata.protein.fa 1"

Onde, "*Aly"* é o *taxon_code*, "*Alyrata.protein.fa"* é o nome do arquivo fasta e "*1"* é o número do campo que quero alterar, ele indica que é o campo logo após o sinal ">". Para ficar mais claro,

aqui temos um exemplo de identificação de uma sequência antes de executar o script "AL706U10010.t1" e após execução do script "Aly|AL706U10010.t1".

Feito esse tratamento nos arquivos fastas, foi finalmente executado o OrthoFinder com o único comando abaixo, em que o parâmetro "*-f fasta_files_directory"* indica o diretório onde estão os fastas de proteínas dos organismos envolvidos na análise e "*-t 30"* o número de *threads* utilizada na análise para deixá-la mais rápida.

"OrthoFinder -f fasta_files_directory -t 30"

Embora o OrthoFinder retorne muitas saídas, será usada, a princípio, apenas o arquivo de *orthogroups.txt*, essa saída segue o mesmo padrão do orthomcl por questões de compatibilidade. Cada linha é composta por uma identificação do grupo de ortólogos e em seguida a identificação de todas as sequencias que pertence a esse grupo separadas por um espaço em branco, um exemplo está na linha abaixo.

OG0007423: Aha|Araha.40054s0003.1.v1.1 Aly|AL706U10010.t1 .....

Cd-hit

Foi realizado um teste também usando o software CD-HIT v4.7 (LI, Weizhong; GODZIK, 2006) nos fastas de proteínas antes de usá-los como entrada para o OrthoFinder. O manual de instalação e configuração do cd-hit está disponível em https://github.com/weizhongli/cdhit/wiki/3.-User's-Guide#CDHIT . O Cd-hit vai remover as variantes de *splicing* e deixar apenas a maior isoforma existente. O Comando abaixo foi utilizado para executar o programa em cada arquivo fasta de proteína.

cdhit -i ArquivoFasta.fasta -o ArquivoFasta.fasta.cdhit

Onde o parâmetro "*-i ArquivoFasta.fasta"* indica o arquivo fasta que será utilizado como entrada e "*-o ArquivoFasta.fasta.cdhit",* o arquivo fasta de saída da análise do cd-hit. O padrão do programa é excluir as variantes que possuem no mínimo 90% de identidade de sequência global (número de aminoácidos idênticos em alinhamento dividido pelo comprimento total da sequência mais curta), deixando as maiores.

3.6. Redes de coexpressão gênica

Nesse passo foram criados as redes e clusters de coexpressão usando o software *Large Scale Transcriptome Analysis Pipeline* (LSTrAP) v1.3 (PROOST; KRAWCZYK; MUTWIL, 2017) .

Download dados RNA-seq

Foram buscados manualmente por análises de RNA-seq no Gene Expression Omnibus (GEO) NCBI (https://www.ncbi.nlm.nih.gov/gds/) (Figura 6.4) referentes a ensaios de qualquer tipo de estresse abiótico em plantas em estágios de desenvolvimento diferentes e utilizando qualquer tipo de tecido. Foram encontrados dados suficientes para montagem das redes de coexpressão apenas para 4 (quatro) espécies: *Arabidopsis thaliana*, *Oryza sativa*, *Glycine max* e *Zea mays*, a pesquisa foi realizada em dezembro de 2019. Um exemplo de como foram os filtros utilizados nas buscas de dados no GEO/NCBI está abaixo para *Arabidopsis thaliana*.

Na caixa de pesquisa textual escrevemos: "(*arabidopsis thaliana*[*Organism*]) *AND stress*"
Tipo de estudo: *"Expression profiling by high throughput sequencing"*

Dessa forma, foram recuperadas 44 diferentes séries do GEO, 154 amostras e 184 SRA *short read files* para os 4 organismos pesquisados, para saber dados detalhados de cada amostra veja a tabela suplementar 1 no apêndice dessa dissertação. Os *raw data* (dados brutos), correspondentes às amostras GEO encontradas, foram obtidas do NCBI/SRA e depois extraídas localmente usando o programa *fastq-dump* do sratoolkit v2.9.2 (https://github.com/ncbi/sra-tools) que transformou os arquivos .sra em arquivos .fastq.

Montagem das redes e *clusters* de coexpressão

Para criar as redes e clusters de coexpressão gênica foi utilizado o software LSTrAP v1.3 (Figura 6.5) que é um pipeline e engloba várias ferramentas de bioinformática conectadas para construção da rede e clusters de coexpressão. A instalação e configuração é bastante simples e foi realizado seguindo as orientações dos criadores da ferramenta existentes no endereço web https://github.molgen.mpg.de/proost/LSTrAP. O sofware é executado com um único comando, que gerencia o funcionamento de todas as ferramentas envolvidas na construção da rede e *clusters* de coexpressão usando os dados de expressão de RNA-Seq. As etapas do pipeline, já explicadas na seção

1.4.2, incluem trimmagem (corte) de *reads* e adaptadores, mapeamento de *reads*, geração de perfis de expressão normalizados, a construção de redes de coexpressão e posteriormente de clusters de coexpressão (figura 3). Além disso, as métricas de controle de qualidade incluídas indicam quais amostras de RNA-seq são potencialmente inadequadas ou de baixa qualidade (PROOST; KRAWCZYK; MUTWIL, 2017).

Foram usados dois tipos de dados como entrada para o LSTrAP, os dados de RNA-seq e o genoma de referência para cada um dos 4 organismos envolvidos na análise. Vale ressaltar, que foram utilizados os mesmos genomas das análises de Diamond e Interproscan para estes organismos. Antes de executar a análise, foi seguida uma recomendação feita no artigo do LSTrAP, dessa forma, para todos os organismos que possuem múltiplas variantes de *splicing*, foi deixado apenas o transcrito primário, ou a maior variante de *splicing* de um determinado gene. Como HTSeq-Count (ANDERS; PYL; HUBER, 2015) e, portanto, por extensão LSTrAP, considera apenas *reads* que mapeiam de forma inequívoca para um único gene, a inclusão de variantes de *splicing* resultaria na perda de *reads* que mapeiam para partes compartilhadas de isoformas. O script *parse_gff.py* disponibilizado pelo LSTrAP foi usado para extrair apenas a variante de *splicing* mais longa dos arquivos *gff3*.

Foram utilizados os parâmetros padrões para todos os programas do pipeline do LSTrAP e o comando utilizado para executá-lo é simples e está logo abaixo, ele executará todo o pipeline sem necessidade de nenhuma intervenção manual. O script em python *"run.py"* é o executável, o parâmetro "*--use-hisat2"* é o comando para usar o hisat2 invés do bowtie2/tophat2 para indexação do genoma e mapeamento de *reads*, "*./config.ini* " informa o arquivo que contém os parâmetros de configuração do LSTrAP e de todos os programas que compõem o pipeline, como por exemplo o caminho dos executáveis, "*./data.ini* " é o arquivo que indica o caminho de todos os dados que serão utilizados nas analises, como os de RNA-seq e o genoma de referência.

./run.py --use-hisat2 ./config.ini ./data.ini

O LSTrAP tem várias saídas, mas as que foram carregadas no banco de dados do projeto são apenas duas: o arquivo cotendo a matriz de expressão, normalizada em TPM (Transcrito por milhão), e o arquivo que contém os clusters de coexpressão gerados pelo algoritmo MCL, cada linha deve conter o nome do cluster seguido de ":" e logo depois todos os genes pertencentes ao cluster separados por tabulação.

3.7. Carregamento (*loading*) dos dados

Foi usado o esquema de banco de dados Chado que é muito utilizado para armazernamento de dados biológicos de uma ampla variedade de organismos, especialmente conjuntos de informações que direta ou indiretamente podem ser associadas a sequências de genoma ou ao RNA primário e produtos proteicos codificados por um genoma (MUNGALL *et al.*, 2007).

Para realizar o carregamento dos dados e resultados de análises no Chado (MUNGALL *et al.*, 2007), foram utilizados *scripts* em *python* (*loaders)* disponibilizados pelo software livre Machado (DE ALVARENGA MUDADU; ZERLOTINI, 2020), que é um aplicativo *Django* que contém ferramentas para interagir com um banco de dados Chado, fornecendo aos usuários uma estrutura para carregar, armazenar, pesquisar e visualizar dados biológicos.

O sistema de gerenciamento de banco de dados (SGBD) utilizado foi o PostgreSQL v.10.14 (https://www.postgresql.org/) e a linguagem de programação foi o *python* v3.6 e um ambiente virtual (*virtualenv*). O passo a passo instalação e configuração do Machado juntamente com o Chado e carregamento dos dados está devidamente descrito na página do Machado no *github* https://machado.readthedocs.io/en/latest/dataload.html#. Vale ressaltar, que o Machado contém *loaders* para os principais formatos de dados de bioinformática: fasta, gff, obo, bibtex, blast, interproscan, orthomcl, dados de expressão gênica entre outtros. Ao final do carregamento dos dados, foi feita a indexação do banco de dados usando o *Elasticsearch* (https://www.elastic.co/pt/elasticsearch/) para que as consultas sejam realizadas de forma mais rápida para os usuários.

3.8. Instalação e configuração da interface web

Para visualização e navegação nos dados armazenados o Machado também disponibiliza uma interface web desenvolvida em Django e os procedimentos para instalação, indexação dos dados e configuração do Jbrowse estão disponíveis em https://machado.readthedocs.io/en/latest/visualization.html#. Além disso, Machado também possui APIs para consulta dos dados armazenados https://www.machado.cnptia.embrapa.br/plantannot2/api/. A interface web para navegação, buscas e consultas de todos os dados armazenados está disponível publicamente no endereço web https://www.machado.cnptia.embrapa.br/plantannot2/.

3.9. Protocolos de filtragem de PFDs

O site do PlantAnnot2 (https://www.machado.cnptia.embrapa.br/plantannot2/) fornece vários filtros e uma caixa de pesquisa de texto que permite pesquisar moléculas por seus recursos de anotação desejados de várias formas. Esses filtros são necessários para obter as PFDs de interesse e tentar relacioná-las a estresses abióticos usando dados de ortoologia e coexpressão. O *menu* Filtros está dividido em 10 campos, mostrados na figura 7, são eles:

1. ***Organism**:* para filtrar o organismo de interesse, existem 67 organismos diferentes.
2. ***Feature type**:* para filtrar o tipo de sequencia, pode ser gene, mRNA ou proteína (polipeptídeo).
3. ***Orthology**:* para filtar apenas proteínas que pertençam ou não a grupos de ortólogos.
4. ***Coexpression**:* para filtar apenas mRNAs que pertençam ou não a clusters de coexpressão.
5. ***Orthologous_coexpression***: para filtrar apenas grupos de ortólogos que tenham ou não pelo menos uma proteína que seu mRNA também pertença a um *cluster* de coexpressão.
6. ***Analyse**s*: para filtrar apenas proteínas que tenha correspondência (*match*) ou não no Diamond e/ou no Interproscan.
7. ***Biomaterial**:* para filtrar o tipo de tecido da planta utilizado na análise de RNA-seq.
8. ***Treatment**:* para filtrar o tipo de estresse abiótico.
9. ***Orthologous group**:* para filtrar o(s) grupo(s) de ortólogo(s).
10. ***Coexpression group**:* para filtrar o(s) *cluster(s)* de coexpressão(ões).

É importante ressaltar que no filtro "*Organism"* da figura 7, não estão sendo exibidos todos os organismos e também que, se nenhuma opção é selecionada para qualquer dos filtros, equivale a marcar todas as opções daquele determinado filtro, por exemplo, se não for selecionado nenhum organismo no filtro 1, é o mesmo de estar selecionando todos os organismos.

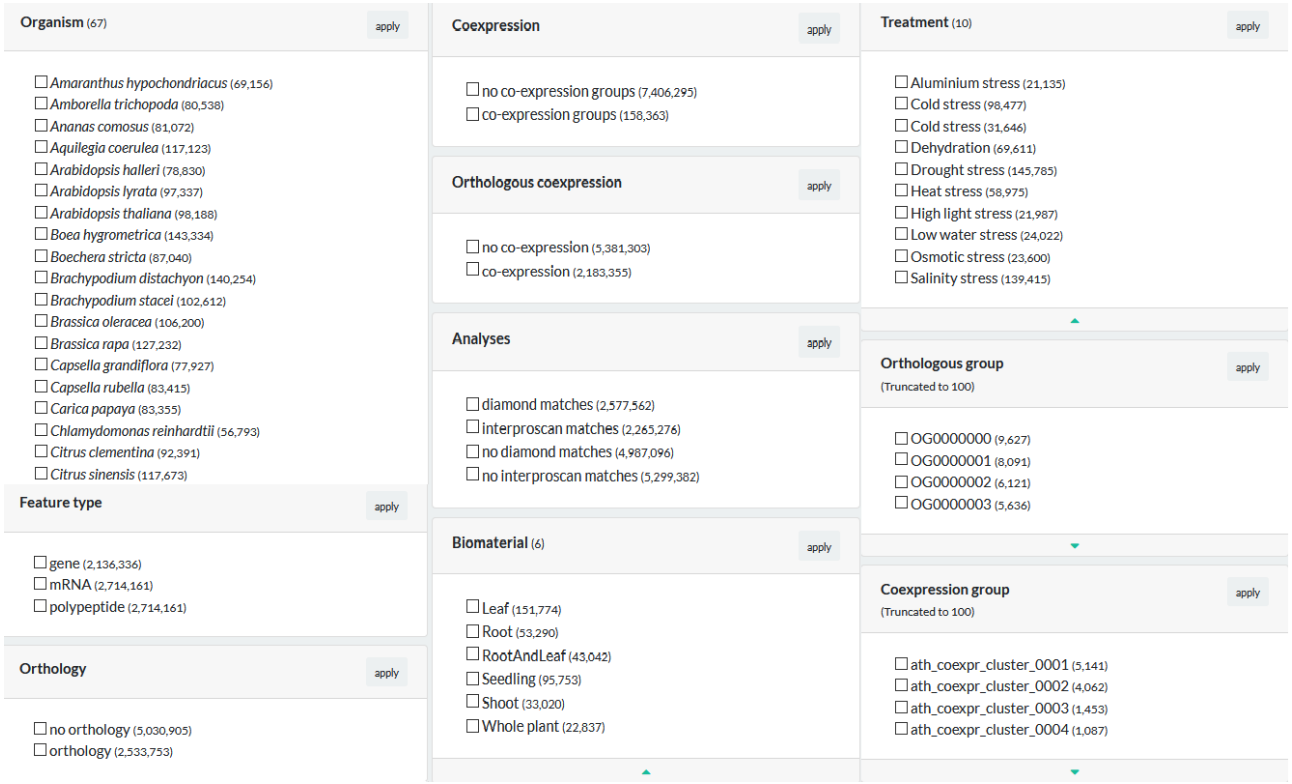

Figura 7 – Filtros do PlantAnnot2. Retirado de https://www.machado.cnptia.embrapa.br/plantannot2/ e adaptado.

Alguns protocolos básicos foram criados, para orientar os usuários, como exemplos de diferentes maneiras de selecionar PFDs. O filtro "*Feature type*" (Figura 7) possui três tipos de moléculas, das quais a caixa *polypeptide* é a única que estará sempre marcada e as demais em branco, nesses exemplos de protocolos. Usando outros 4 filtros (*Orthology*, *Coexpression*, *Orthologous coexpression* e *Analyses*), 6 protocolos foram criados (Tabela 2), e esses protocolos estão disponibilizados publicamente no site protocols.io (https://www.protocols.io/). Vale salientar que há uma pequena diferença entre os protocolos que usam busca textual no PlantAnnot e no PlantAnnot2, no primeiro usamos na busca textual "*Unknown function*" e no segundo apenas "*Unknown*", como será mostrado a seguir. O protocolo A (VIANA, Marcos; ZERLOTINI; MUDADU, [*s. d.*]) busca falta de correspondência de proteínas (*no-hits*) no NCBI-nr e de assinaturas de domínio de proteína nos bancos de dados do Interproscan. Protocolo B (VIANA, Marcos; ZERLOTINI; MUDADU, [*s. d.*]) busca por falta de correspondência no NCBI-nr e presença de assinaturas de domínio na tentativa de selecionar Domínios de Função Desconhecida (*Domain Unknown Function* - DUF) do PFAM e a pesquisa de texto "*Unknown*" (Figura 8). Protocolo C (VIANA, Marcos; ZERLOTINI; MUDADU, [*s. d.*]) busca correspondência de proteínas no NCBI-nr, mas que não há assinaturas de domínio de proteína e usando a pesquisa de texto "*Unknown*" . Protocolo D (VIANA, Marcos; ZERLOTINI;

MUDADU, [*s. d.*]), E (VIANA, Marcos; ZERLOTINI; MUDADU, [*s. d.*]) e F (VIANA, Marcos; ZERLOTINI; MUDADU, [*s. d.*]) são os mesmos protocolos de A-B-C respectivamente, mas usando grupos ortólogos para encontrar proteínas homólogas e que nesse grupo exista pelo menos uma proteína cujo mRNA pertence a algum *cluster* de coexpressão relacionado a estresse abiótico.

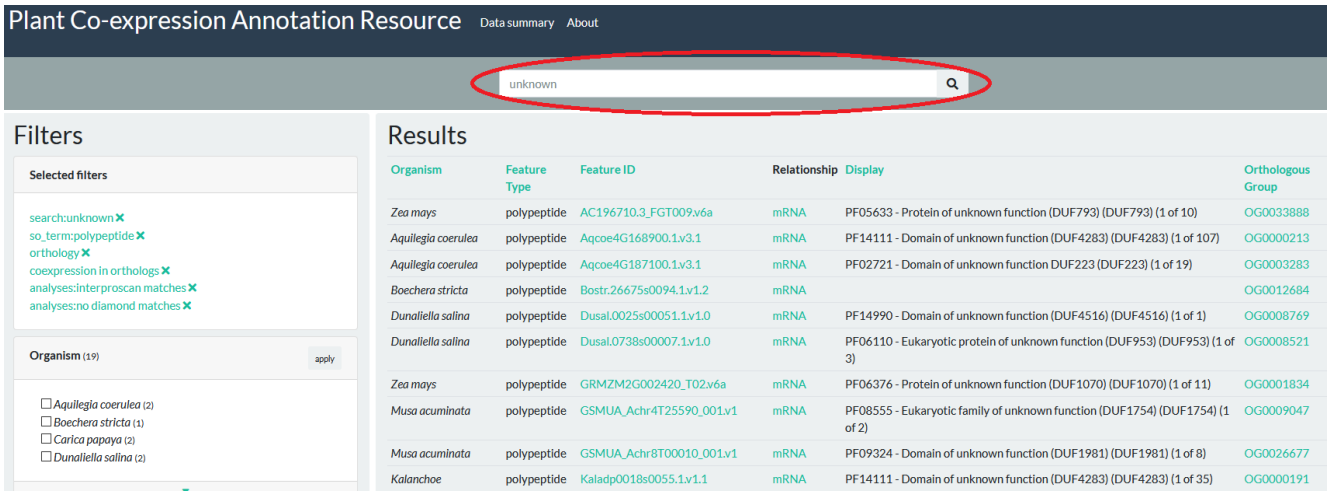

Figura 8 – Uso de busca textual no PlantAnnot2. Retirado de https://www.machado.cnptia.embrapa.br/plantannot2/.

Nome	**Objetivo**	**Filtros (Caixas que devem ser marcadas)**
Protocolo A	Recupera PFDs de organismos cujas proteínas ainda **não** estão no nr do NCBI e também **não** foi encontrado domínio pelo Interproscan.	*Analyses: **no** diamond matches* *Analyses: **no** interproscan matches*
Protocolo B	Recupera PFDs de organismos que já possuem a proteínas depositada no nr do NCBI e **não** foi encontrado domínio pelo Interproscan, usando busca textual.	*Analyses: **no** diamond matches* *Analyses: interproscan matches* *Text search: "Unknown"*
Protocolo C	Recupera PFDs de organismos cujas proteínas ainda **não** estão no nr do NCBI e possuem DUF no PFAM, usando busca textual.	*Analyses: diamond matches* *Analyses: **no** interproscan matches* *Text search: "Unknown"*

Protocolo D	Mesmo do protocolo A e usando grupos ortólogos e de coexpressão para relacionar essas PFDs a estresses abióticos.	*Analyses: **no** diamond matches* *Analyses: **no** interproscan matches* *Orthology: orthology* *Orthologs_coexpression: co-expression*
Protocolo E	Mesmo do protocolo B e usando grupos ortólogos e de coexpressão para relacionar essas PFDs a estresses abióticos.	*Analyses: **no** diamond matches* *Analyses: interproscan matches* *Text search: "Unknown"* *Orthology: orthology* *Orthologs_coexpression: co-expression*
Protocolo F	Mesmo do protocolo C e usando grupos ortólogos e de coexpressão para relacionar essas PFDs a estresses abióticos.	*Analyses: diamond matches* *Analyses: **no** interproscan matches* *Text search: "Unknown"* *Orthology: orthology* *Orthologs_coexpression: co-expression*

Tabela 2 – Protocolos de busca de PFDs

3.10. Enriquecimento de vias

Foi usado o software ShinyGO v0.61: Gene Ontology Enrichment Analysis + more (GE et al., 2020) para fazer o enriquecimento de vias dos organismos *Arabidopsis thaliana* e *Oryza sativa*. Usando os filtros do plantannnot (*Organism + feature + Coexpression*) (Figura 7) foram selecionados todos os mRNAs que pertencem a *clusters* de coexpressão desses dois organismos. Pela interface web da própria ferramenta PlantAnnot2 conseguimos baixar a relação de desses mRNAs em um arquivo no formato .tsv. Na figura 9 temos um exemplo da busca feita para *Arabidopsis thaliana*, depois de marcar os filtros (em destaque no lado esquerdo superior da figura 9) clique no ícone TSV canto superior direito da tela, também em destaque na figura 9.

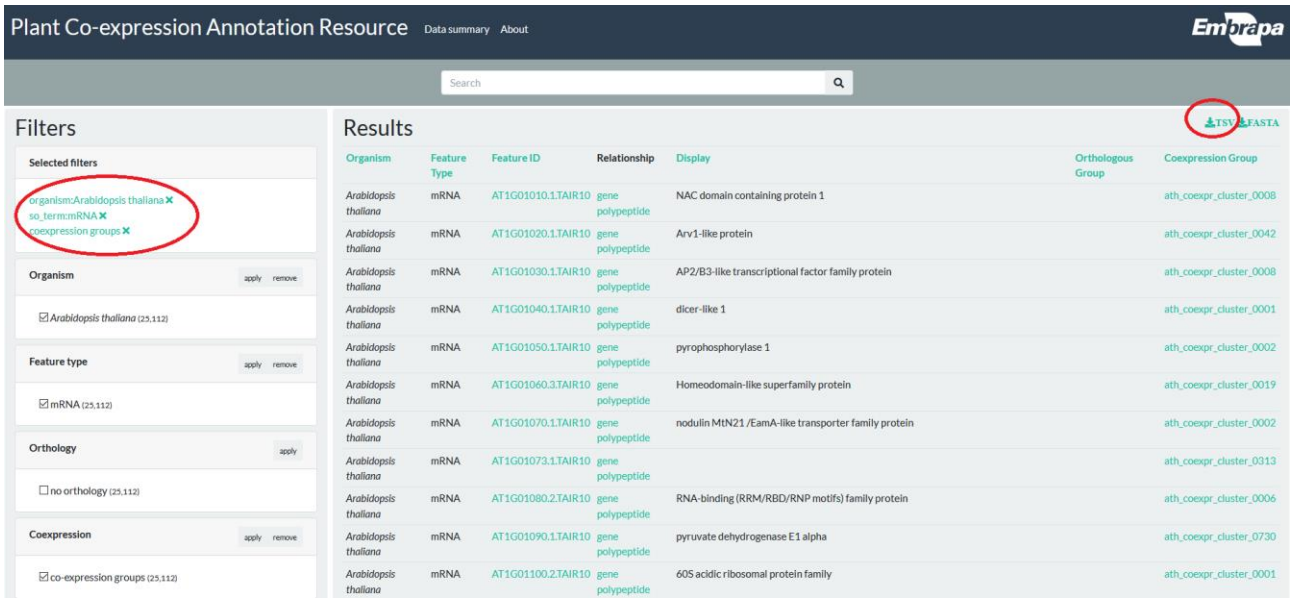

Figura 9 – Baixar resultado de uma busca em arquivo formato .tsv

CAPÍTULO 4: RESULTADOS

4.1. Dados armazenados

Depois de armazenados todos os dados que compõem o PlantAnnot2, como os de genomas, expressão gênica e resultados das análises, o banco de dados ficou com mais de 426 milhões de registros. Na tabela 3 estão alguns números sobre dados armazenados no PlantAnnot2 e também do PlantAnnot, para posterior comparação.

	PlantAnnot2	PlantAnnot
Quantidade de genomas de organismos	67	53
Quantidade de amostras (*samples*) de RNA-seq	154	53
Quantidade de genes	2.136.336	1.862.010
Quantidade de mRNAs	2.714.161	2.332.974
Quantidade de proteínas	2.714.161	2.332.974
Tamanho do Banco de dados (GB)	53	49

Tabela 3 – Dados armazenados no PlantAnnot2 x PlantAnnot

4.2. Homologia

As 2.714.161 proteínas foram usadas como entrada para os softwares OrhtoMCL e OrthoFinder construir grupos de ortólogos (*orthogroups*) e dessa forma obteve-se os resultados exibidos na tabela 4.

	OrthoMCL	OrthoFinder	OrthoFinder (Com CD-HIT)
Número de *orthogroups*	183.904	91.174	69.721
Média de proteínas por *orthogroup*	12	27	26
Tamanho do menor *orthogroup*	2	2	2
Desvio Padrão	34	112	106
Q1 – quantidade de proteínas por *orthogroup*	2	2	2
Mediana (Q2) – quantidade de proteínas por *orthogroup*	3	3	3
Q3 – quantidade de proteínas por *orthogroup*	6	8	9
Quantidade de proteínas do maior *cluster*	4.701	9.627	10.258
Quantidade total de proteínas	2.714.124	2.714.124	2.018.191
Quantidade de proteínas sem o*rthogroups*	450.138	180.371	181.019
Quantidade total de proteínas em o*rthogroups*	2.263.986	2.533.753	1.837.172
Porcentagem de proteínas em o*rthogroups*	83%	93%	91%
Quantidade de o*rthogroups* com todas as espécies presentes	-	1.360	1.362
Quantidade de orthogroups com *single-copy*	-	0	0
Quantidade de o*rthogroups* de espécie específicos	-	40.715	24.064
Quantidade de genes espécie específicos em *orthogroups*	-	180.696	115.960

Tabela 4 – Comparativos de resultados de homologia entre OrthoMCL, OrthoFinder e OrthoFinder + cd-hit do PlantAnnot2.

O OrthoFinder fornece mais saídas que o OrthoMCL, entre elas, uma análise estatística abrangente e alguns desses dados foram adicionados na tabela 4, apenas para as análises realizadas com esse software, como: Quantidade de o*rthogroups* com todas as espécies presentes, Quantidade de orthogroups com *single-copy*, que são grupos de ortólogos que possuem apenas 1 proteína para cada organismo envolvido na análise, Quantidade de o*rthogroups* espécie específicos, Quantidade de genes espécie específicos em *orthogroups* entre outras. Dado interessante, e que não consta na tabela, é que 50% de todas as proteínas da análise realizada apenas com o OrthoFinder estão em grupos com 176 proteínas ou mais.

Como a análise de homologia da primeira versão do PlantAnnot foi realizada com o OrthoMCL, os dados da tabela 4 referentes a esse software também servem para comparação entre as duas versões do sistema. Para complementar, no PlantAnnot foi conseguido a construção de

164.267 grupos de ortólogos com o agrupamento de 81% de todas as 2.332.974 proteínas, além disso a médias de proteínas por grupo foi a mesma, assim como a mediana(Q2), Q1 e Q3, a quantidade de proteínas no maior cluster ficou em 4.587.

4.3. Redes de coexpressão

Para construir redes de coexpressão, foram utilizadas 154 amostras baixadas do GEO/NCBI, de ensaios de diferentes estresses abióticos, como: seca, estresse por alumínio, frio, estresse osmótico, salínico entre outros, além disso, foram utilizados diferentes tipos de tecidos e estágios de desenvolvimento da planta, como: folha, raiz, muda, planta adulta, entre outros (veja Tabela suplementar 1) foram utilizadas para extrair os dados de expressão gênica, e então quatro redes de coexpressão foram construídas, uma para cada um dos organismos *Arabidopsis thaliana* (ATH), *Oryza sativa* (OSA), *Glycine max* (GMA) e *Zea mays* (ZMA) e posteriormente foram criados os c*lusters* de coexpressão usando o software MCL que agrupa os transcritos com expressão mais correlacionada. Na tabela 5 temos alguns dados estatísticos sobre os clusters de coexpressão gerados.

	ATH	OSA	ZMA	GMA	TOTAL
Quantidade de amostras	87	37	12	18	154
Quantidade de clusters	907	680	78	310	1.975
Média de genes por cluster	28	52	601	165	-
Desvio Padrão	230	321	1.148	853	-
Q1 - Quantidade de genes por *cluster*	3	3	4	4	-
Mediana (Q2) - Quantidade de genes por *cluster*	4	5	10	6	-
Q3 - Quantidade de genes por *cluster*	9	10	604	12	-
Tamanho do menor *cluster*	2	2	2	2	-
Tamanho do maior *cluster*	5.141	6.338	5.413	9.178	-
Quantidade total de genes em *clusters* de coexpressão gênica	25.112	35.377	46.862	51.012	158.363
Porcentagem de genes em *clusters* de coexpressão gênica	71%	67%	53%	57%	-
Quantidade de genes em nenhum *cluster* de coexpressão gênica	10.274	17.047	41.898	37.635	106.854

Tabela 5 – *Clusters* de coexpressão gênica do PlantAnnot2.

A primeira versão do PlantAnnot teve 875 clusters de coexpressão no total, sendo 427 para Arabidopsis thaliana, 137 para Oryza sativa, 65 para Zea mays e 246 para Glycine max. Foram

146.401 genes agrupados em *clusters* de coexpressão. A quantidade média de genes por cluster diminuiu em todos os organismos no PlantAnnot2 em relação ao PlantAnnot.

4.4. Caracterização de PFDs

Depois de analisar todas as 2.714.161 proteínas com Diamond e InterproScan, foram caracterizadas 78.416 PFDs (Tabela 6 - Protocolo A) com sequências que não tiveram correspondências NCBI-nr e em nenhum banco de dados do InterproScan. Outra maneira menos estringente de encontrar PFDs é filtrar por correspondências no InterproScan (por exemplo: tentando selecionar domínios DUF do PFAM), porém com anotações não informativas usando a pesquisa textual e buscando, por exemplo, pela palavra "*Unknown*" (desconhecida) (Tabela 6 - Protocolo B) ou buscar proteínas com correspondências encontradas pelo Diamond, e também utilizando a pesquisa textual buscando por "*Unknown*" (Tabela 6 - Protocolo C), dessa forma, foram recuperadas 2.195 e 5.514 PFDs respectivamente.

Organismo		Protocolo					
	Versão Genoma	A	B	C	D	E	F
Amaranthus hypochondriacus	v1.0	899	3	3	53	0	3
Amborella trichopoda	v1.0	55	0	2	0	0	1
Ananas comosus	v3	1516	0	3	89	0	3
Aquilegia coerulea	v3.1	2240	8	53	59	2	48
Arabidopsis halleri	v1.1	339	0	18	36	0	17
Arabidopsis lyrata	v2.1	610	0	7	19	0	7
Arabidopsis thaliana	TAIR10	324	0	131	27	0	130
Boea hygrométrica	GCA_001598015.1	39	0	2	1	0	0
Boechera stricta	v1.2	559	5	10	55	1	9
Brachypodium distachyon	v3.1	17	0	76	1	0	67
Brachypodium stacei	v1.1	1068	0	38	36	0	37
Brassica oleracea	v1.0	196	0	12	11	0	1
Brassica rapa	FPsc v1.3	18	0	22	3	0	22
Capsella grandiflora	v1.1	182	0	15	20	0	15
Capsella rubella	v1.0	1	0	10	0	0	10
Carica papaya	ASGPBv0.4	3283	2	1	192	2	0
Chlamydomonas reinhardtii	v5.5	20	0	2	0	0	0
Citrus clementenina	v1.0	8	0	25	1	0	25
Citrus sinensis	v1.1	7	0	29	0	0	28
Coccomyxa subellipsoidea	*C-169v2.0*	39	1	6	0	0	0
Cucumis sativus	v1.0	976	0	14	5	0	14
Daucus carota	v2.0	6	0	0	0	0	0

Dunaliella salina	v1.0	2827	13	2	36	2	1
Eucalyptus grandis	v2.0	61	0	20	1	0	19
Eutrema salsugineum	v1.0	3	0	11	0	0	11
Fragaria vesca	v1.1	3178	9	1	181	1	0
Glycine max	Wm82.a2.v1	12	0	95	1	0	95
Gossypium raimondii	v2.1	16	0	67	1	0	66
Kalanchoe fedtschenkoi	v1.1	1947	13	37	110	5	36
Kalanchoe laxiflora	v1.1	1587	10	104	65	1	103
Linum usitatissimum	v1.0	1614	28	4	118	10	4
Malus domestica	v1.0	3477	1	41	132	1	38
Manihot esculenta	v6.1	19	0	27	1	0	27
Marchantia polymorpha	V3.1	13	0	15	0	0	11
Medicago truncatula	Mt4.0v1	227	0	42	7	0	39
Micromonas pusilla	*CCMP1545v3.0*	50	2	6	1	0	1
Micromonas sp.	*RCC299v3.0*	10	0	0	0	0	0
Mimulus guttatus	v2.0	721	2	34	26	1	34
Musa acuminata	v1	3019	3	0	46	2	0
Oropetium thomaeum	v1.0	2495	7	3	117	5	3
Oryza sativa	v7_JGI	733	0	17	141	0	17
Ostreococcus lucimarinus	v2.0	5	0	10	0	0	3
Panicum hallii	v2.0	1746	1	54	31	0	51
Panicum virgatum	v1.1	10695	5	102	192	2	92
Pearl millet	V1.1	2005	0	0	75	0	0
Phaseolus vulgaris	v2.1	114	0	28	10	0	28
Physcomitrella patens	v3.3	978	0	250	10	0	178
Populus Simonii	v2.0	3775	0	0	104	0	0
Populus trichocarpa	v3.0	98	0	102	3	0	99
Prunus pérsica	v2.1	13	0	40	1	0	39
Ricinus communis	v0.1	14	0	0	0	0	0
Salix purpúrea	v1.0	1256	0	0	20	0	0
Selaginella moellendorffii	v1.0	8	0	2	0	0	2
Setaria itálica	v2.2	12	1	49	1	1	45
Setaria viridis	v1.1	50	1	55	1	1	52
Solanum lycopersicum	iTAG2.4	2209	0	1	26	0	1
Solanum tuberosum	v4.03	3350	2027	2289	57	15	752
Sorghum bicolor	v3.1.1	13	0	19	1	0	16
Sphagnum fallax	v0.5	3081	26	33	93	0	21
Spirodela polyrhiza	v2	1151	12	1	105	5	1
Theobroma cacao	v1.1	151	4	1400	0	0	57
Trifolium pratense	v2	1251	3	13	64	0	12
Vitis vinífera	Genoscope.12X	119	1	0	0	0	0
Volvox carteri	V2.1	993	0	5	1	0	2
Zea mays	284_AGPv3	9783	3	46	2068	2	44
Zea maysPH207	v1.1	1094	3	1	217	1	1

Zostera marina	v2.2	41	1	9	1	0	9
TOTAL		78416	2195	5514	4.673	60	2447

Tabela 6 – Proteínas de Função Desconhecida recuperadas no PlantAnnot2 por protocolo de busca.

Foi feita uma análise de buscas de PFDs usando os protocolos D, E e F para os 2 tipos de homologia que testamos, OrthoMCL e OrthoFinder, e também para a análise do Orthofinder com cd-hit, na tabela 7 são apresentados os resultados. Vale salientar que os protocolos D, E e F da tabela 6 são resultados obtidos com a análise de homologia do Orthofinder.

	OrthoMCL	OrthoFinder	OrthoFinder (Com CD-HIT)
Protocolo D	2.272	4.673	3.942
Protocolo E	39	60	60
Protocolo F	1.854	2.447	1.518

Tabela 7 – Comparação de quantidade de PFDs recuperadas por análise de homologia no PlantAnnot2.

Aqui também é possível usar os resultados do OrthoMCL como comparativo em relação as duas versões do PlantAnnot, dessa forma a comparação passa a ser mais justa, visto que o PlantAnnot2 usa mais dados de genômica e RNA-seq. Mesmo assim, para efeito de comparação, as quantidades de PFDs recuperadas com o PlantAnnot utilizando o protocolo A foi de 72.266, protocolo B foi 2.409 e protocolo C foi 6.569. E ao relacionar as PFDs com dados de ortologia e coexpressão gênica, foram obtidas 1.364 PFDs com protocolo D, 13 PFDs como protocolo E e 2.280 PFDs com protocolo F.

4.5. Anotação de PFDs

Foram encontradas 30.276 PFDs como membros de grupos ortólogos que podem ser fonte de informação funcional e anotação (Protocolo A, acrescido do filtro "*Ortology: ortology*") e recuperadas 4.673 PFDs que pertenciam a algum grupo ortólogo e foram relacionados a grupos de coexpressão usando filtros que foram criados para automatizar essa seleção (Tabela 6, Protocolo D). Ao modificar o protocolo D e a pesquisa de texto por "*Unknown*", além da filtragem para apenas correspondências de Interproscan ou apenas correspondências do Diamond, poderíamos anotar 60 e 2447 PFDs respectivamente (Tabela 6, protocolos E e F respectivamente).

Realizamos curadoria manual na literatura em busca de estudos relacionados as anotações econtradas em proteínas com função conhecida que pertencem ao mesmo grupo de ortólogos de todas as PFDs recuperadas usando protocolo D para os organismos *Pearl millet, Oropetium thomaeum* e

Populus simonii, e usando Protocolo E com uma pequena variação para *Boea hygrometrica*, alterando a palavra da busca textual de "*Unknown*" para "*hypothetica*l". Dessa forma foram recuperadas 517 PFD, para essa amostra com 4 organismos, e encontramos estudos relacionados a estresses abióticos, em média, para 67,5% das anotações que podem ser preditas para essas PFDs.

Com a versão 1 do PlantAnnot usando o protocolo A foram encontradas 21.895 PFDs pertencentes a grupos de ortólogos e 1.364 relacionadas com grupos ortólogos e *clusters* de coexpressão (Tabela 2, protocolo D), além disso foram recuperadas 13 e 2280 PFDs usando os protocolos E e F respectivamente. Foram encontrados estudos relacionados a estresse abiótico para as anotações do homólogos das PFDs para 11,6% delas, excluindo o organismo Zea mays que também ficou *outlier* nesse estudo o percentual sobe para 35%.

4.5.1. Estudos de caso

Foram encontradas muitas proteínas com função conhecida e homólogas de PFDs, que segundo estudos encontrados na literatura acadêmica e científica, executam papéis relevantes com relação a respostas a estresses abióticos. São apresentados aqui 4 estudos de casos em que algumas PFDs são ortólogas de proteínas já anotadas e que fazem parte do núcleo de sinalização do ABA (Seção 1.1).

No primeiro estudo, foram encontradas 2 PFDs de *Boea*, com os identificadores KZV23183.1 e KZV23184.1, que pertencem ao grupo ortólogo OG0010070 (https://www.machado.cnptia.embrapa.br/plantannot2/find/?selected_facets=orthologous_group:OG0010070) composto por 53 proteínas de 16 organismos diferentes, entre eles, outro organismo como características de resistência a seca, que é a *Populus simonii*, possui 6 proteínas nesse mesmo grupo. Foi encontrada a anotação "*K14496 - abscisic acid receptor PYR/PYL family (PYL)*". Vale ressaltar ainda que nesse grupo de ortólogos existe também a proteína Glyma.06G118700.1.Wm82.a2.v1 de *Glycine max* cujo seu mRNA pertence ao grupo de coexpressão *gma_coexpr_cluster_0003* (https://www.machado.cnptia.embrapa.br/plantannot2/find/?selected_facets=coexpression_group:gma_coexpr_cluster_0003) e teve sua expressão aumentada em 15 ensaios diferentes de tratamentos de seca, desidratação e estresse salínico. A anotação encontrada é referente a família de proteínas do receptor ABA PYR/PYL/RCAR (PYL) (seção 1.1).

O segundo estudo de caso também é sobre uma PFD da *Boea hygrometrica* que tem o identificador KZV51863.1 e pertencente ao grupo ortólogo OG0000607 (https://www.machado.cnptia.embrapa.br/plantannot2/find/?selected_facets=orthologous_group:OG0000607) composto por 527 proteínas e pelo menos 276 delas com alguma anotação relacionada a

"*Protein phosphatase 2C family protein (PP2C)"* cuja a função já explicamos na introdução (seção 1.1).

O terceiro estudo de caso é referente a PFD de *Pearl millet* identificada pelo *ID* Pgl_GLEAN_10014316 e pertence ao grupo de ortólogos OG0002090 (https://www.machado.cnptia.embrapa.br/plantannot2/find/?selected_facets=orthologous_group%3AOG0002090) composto por 241 proteínas, sendo que 75 delas possuem alguma anotação referente a "serine/threonine protein kinase" encontrada em 35 organismos diferentes, entre eles, *Boea hygrométrica* com 2 proteínas, *Oropetium thomaeum* com 2 proteínas, *Populus simonii* com 3 proteínas e *Sorghum bicolor* com 7. Nesse grupo de ortólogos temos ainda 17 proteínas com mRNAs participando de cluster de coexpressão gênica.

O quarto estudo de caso é referente as PFDs identificadas pelos IDs Simonii00031906-RA de *Populus simonii* e Oropetium_20150105_06861A.v1.0 de *Oropetium thomaeum* que pertencem ao grupo de ortólogos OG0000165 (https://www.machado.cnptia.embrapa.br/plantannot2/find/?selected_facets=orthologous_group:OG0000165) composto por 1068 proteínas e 432 delas com anotações referentes a "*BASIC-LEUCINE ZIPPER (BZIP) TRANSCRIPTION FACTOR FAMILY PROTEIN"*(seção 1.1). Vale salientar que nesse grupo de ortólogos temos também existem 72 proteínas com seus mRNAs pertencentes a *clusters* de coexpressão de ensaios de estresses abióticos diversos.

4.6. Enriquecimentos de vias.

O resultado das análises de enriquecimento de vias para *Arabidopsis thaliana* e *Oryza sativa* seguem abaixo nas tabelas 8 e 9 respectivamente.

PLANTANNOT2				**PLANTANNOT**			
Enrichment FDR	Genes in list	Total genes	Functional Category	Enrichment FDR	Genes in list	Total genes	Functional Category
2.68350744570545e-63	3109	3194	Response to chemical	5.91710576613338e-95	3081	3194	Response to chemical
6.5634846911007e-47	4019	4203	Cellular protein metabolic process	1.46475797179233e-83	2102	2150	Response to abiotic stimulus
3.28757193550615e-46	4425	4644	Protein metabolic process	1.93563822804856e-66	2121	2194	Response to organic substance

7.24206620245944e-45	3472	3619	Cellular component organization or biogenesis	4.82256427324308e-66	3116	3286	Developmental process
3.60896928585241e-44	2097	2150	Response to abiotic stimulus	4.82256427324308e-66	3417	3619	Cellular component organization or biogenesis
1.26475212488933e-42	2136	2194	Response to organic substance	8.97766242683215e-65	2964	3121	Anatomical structure development
3.21200433447951e-40	1741	1779	Response to oxygen-containing compound	2.35342439841026e-63	1733	1779	Response to oxygen-containing compound
4.50436717016634e-39	3156	3293	Cellular component organization	4.91931964633713e-59	3108	3293	Cellular component organization
1.78699204072062e-37	3374	3532	Macromolecule modification	3.73522803338873e-55	2611	2752	Multicellular organism development
3.79605951324249e-37	3145	3286	Developmental process	1.27276872228562e-54	4025	4327	Gene expression
5.87737204071954e-37	3480	3648	Response to stress	6.78425019292935e-53	1787	1853	Response to hormone
5.64589336956213e-36	2989	3121	Anatomical structure development	7.93089983411663e-53	1819	1888	Response to endogenous stimulus
2.93359544801885e-33	2640	2752	Multicellular organism development	4.73940952526373e-52	2862	3037	Multicellular organismal process
3.00100534638918e-33	3041	3184	Cellular response to stimulus	8.02129516950655e-51	1858	1934	Organelle organization
8.40774515369598e-33	1832	1888	Response to endogenous stimulus	8.86572900871184e-51	3304	3532	Macromolecule modification
1.37800692397954e-32	1810	1865	Organonitrogen compound biosynthetic process	4.40281939425123e-50	3902	4203	Cellular protein metabolic process
3.54925955734536e-32	1798	1853	Response to hormone	2.14244282938114e-49	4542	4925	Nucleobase-containing compound metabolic process

4.25580788117434e-32	1901	1963	Small molecule metabolic process	7.73074928515591e-46	1298	1334	Response to acid chemical
2.1598705109779e-31	4091	4327	Gene expression	1.73816287127611e-45	4281	4644	Protein metabolic process
5.15760698097071e-31	2899	3037	Multicellular organismal process	6.56657789126971e-43	2609	2779	Establishment of localization

Tabela 8 – Enriquecimento de vias *Arabidopsis thaliana*, comparação PlantaAnnot x PlnatAnnot2

PLANTANNOT2				**PLANTANNOT**			
Enrichment FDR	Genes in list	Total genes	Functional Category	Enrichment FDR	Genes in list	Total genes	Functional Category
4.92886533173971e-39	665	956	Response to chemical	6.7241011954901e-43	655	956	Response to chemical
1.45700275498235e-32	354	465	Response to organic substance	3.72812695056049e-36	874	1385	Cellular component organization or biogenesis
1.69275864569368e-32	1915	3282	Biological regulation	1.07296898532145e-34	1226	2057	Response to stimulus
1.56682952880123e-31	320	415	Response to endogenous stimulus	5.5936309258478e-34	347	465	Response to organic substance
1.56682952880123e-31	320	415	Response to hormone	9.3994818267206e-33	1846	3282	Biological regulation
1.77579803350366e-29	879	1385	Cellular component organization or biogenesis	1.09145159880948e-31	312	415	Response to endogenous stimulus
3.49422876057482e-29	1245	2057	Response to stimulus	1.09145159880948e-31	312	415	Response to hormone
5.81872408279404e-27	1713	2957	Regulation of biological process	5.3647046353627e-31	784	1253	Cellular component organization
7.48343309466506e-26	511	755	Developmental process	7.12041517445805e-31	2610	4836	Organonitrogen compound metabolic process
5.60935702944753e-25	789	1253	Cellular component organization	8.52420817575228e-29	1791	3217	Cellular protein metabolic process
1.19606101698526e-24	482	711	Anatomical structure development	1.52651297521117e-27	1653	2957	Regulation of biological process
2.74815721236103e-24	2668	4836	Organonitrogen compound metabolic process	7.01612469412796e-27	498	755	Developmental process
6.29259514891767e-23	1825	3217	Cellular protein metabolic process	1.24132911651916e-26	2105	3871	Protein metabolic process

1.96498370247083e-22	2160	3871	Protein metabolic process	2.50955920774673e-25	469	711	Anatomical structure development
1.04529103671901e-21	259	348	Response to oxygen-containing compound	5.95436681155732e-25	2119	3921	Organic substance biosynthetic process
5.45314968727441e-21	1542	2697	Regulation of cellular process	9.23592458657387e-25	258	348	Response to oxygen-containing compound
6.16552889334113e-21	429	638	Multicellular organism development	3.20681162987299e-24	2198	4091	Biosynthetic process
1.75651231704631e-20	521	802	Multicellular organismal process	5.97049892076965e-24	2075	3846	Cellular biosynthetic process
3.75641546303754e-20	244	329	Cell wall organization or biogenesis	3.74094904588995e-21	1638	3002	Gene expression
3.89073782717296e-20	2169	3921	Organic substance biosynthetic process	3.74094904588995e-21	320	467	Response to abiotic stimulus
9.9024655915671e-20	2253	4091	Biosynthetic process	4.57625985770183e-21	416	638	Multicellular organism development
1.3967888380235e-19	349	507	Cellular response to chemical stimulus	6.11172269526891e-21	1484	2697	Regulation of cellular process
1.5472036150175e-19	1691	3002	Gene expression	1.72364158419654e-20	504	802	Multicellular organismal process
2.74478292161871e-19	2124	3846	Cellular biosynthetic process	1.72364158419654e-20	238	329	Cell wall organization or biogenesis
6.10556788211723e-18	1217	2112	Regulation of metabolic process	2.08190586116114e-20	634	1045	Response to stress
1.68739510062176e-17	320	467	Response to abiotic stimulus	3.6037521270796e-20	340	507	Cellular response to chemical stimulus
2.43122230630209e-17	1591	2837	Macromolecule biosynthetic process	2.45555157335829e-19	1184	2120	Oxidation-reduction process
4.73006888252998e-17	1145	1985	Regulation of cellular metabolic process	4.97674180995265e-19	1394	2541	Macromolecule modification
4.99142474963292e-17	548	873	Catabolic process	6.39710200684026e-19	1541	2837	Macromolecule biosynthetic process
6.26568459264943e-17	642	1045	Response to stress	7.41886683515559e-19	1312	2380	Protein modification process

Tabela 9 – Enriquecimento de vias *Oryza sativa*, comparação PlantaAnnot x PlnatAnnot2

As redes abaixo de *Arabidopsis thaliana* e *Oryza sativa*, figuras 10 e 11 respectivamente, mostra a relação entre caminhos enriquecidos. Duas vias (nós) são conectadas se compartilharem 20% (padrão) ou mais genes. Os nós mais escuros são conjuntos de genes mais significativamente enriquecidos. Nós maiores representam conjuntos de genes maiores. Bordas mais espessas representam mais genes sobrepostos (GE *et al.*, 2020).

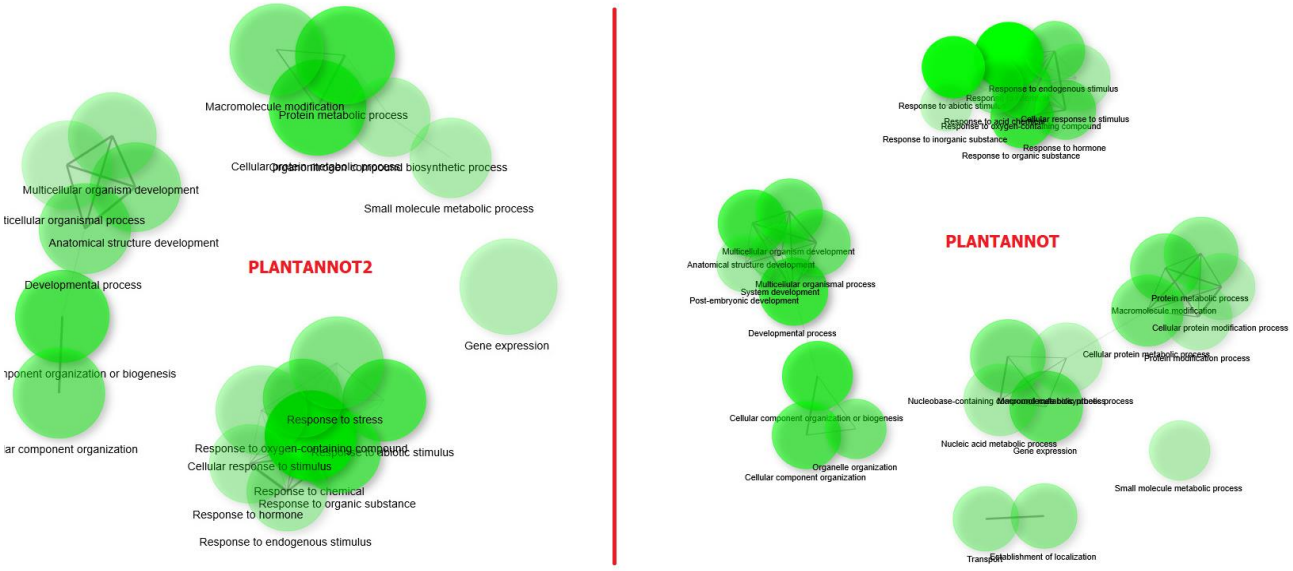

Figura 10 – Redes de enriquecimento para *Arabidopsis thaliana* no Plantannot2 x PlantAnnot

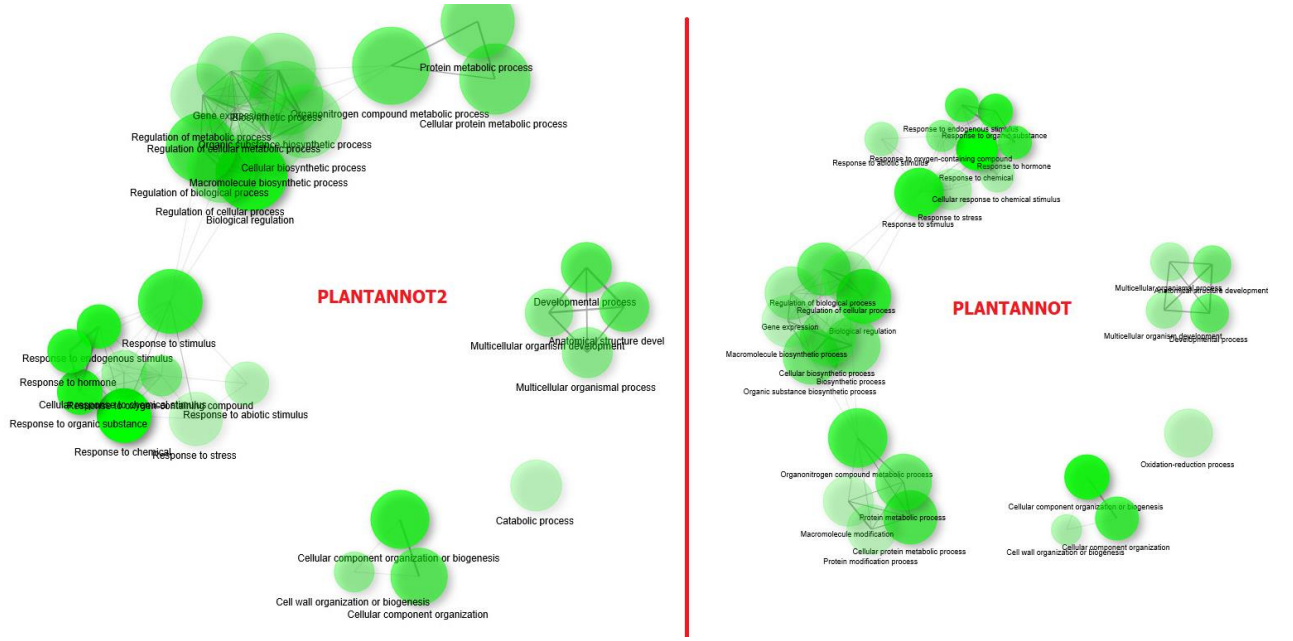

Figura 11 – Redes de enriquecimento para *Oryza sativa* no Plantannot2 x PlantAnnot

CAPÍTULO 5: DISCUSSÃO

5.1. Dados armazenados

Houve um grande trabalho de normalização dos dados genômicos utilizados no PlantAnnot2, não existe um consenso na forma de identificação dos genes, mRNA e proteínas nos genomas, nem dentro do mesmo repositório. Praticamente 96% dos dados genômicos usados foram baixados do *Phytozome*, mesmo assim, ainda foi necessário normalizar a grande maioria desses dados. Um problema comum, é que os identificadores usados no arquivo *GFF* para os mRNAs não serem iguais ao identificador das sequencias no arquivo fasta. Essa tarefa, em uma imensa base de dados, com 67 genomas de plantas, passa a ser um trabalho hercúleo e desafiador. Com scripts em *Python* e *shell script* foi feita a normalização desses dados. A normalização de dados biológicos ainda é uma das principais atividades da área de bioinformática (AMBROSINO *et al.*, 2020).

Em relação a primeira versão do PlantAnnot foram adicionados mais 14 novos genomas e 101 amostras (samples) de RNA-seq, consequentemente houve aumento no número de genes, mRNAs e polipeptídios armazenados (Tabela 3), assim como a quantidade de análises, grupos de ortólogos e clusters de coexpressão que serão melhores detalhados nas próximas seções.

5.2. Homologia

Nas análises de homologia, foram testados dois softwares para montagem de grupos ortólogos, o OrthoMCL, que foi usado na versão 1 do PlantAnnot e é o mais usados para essa finalidade (LI, Li; STOECKERT; ROOS, 2003), e o OrthoFinder que vem se destacando e ganhando muitos adeptos, por sua simplicidade de execução, escalabilidade, rapidez na execução, maior acurácia na montagem dos grupos de ortólogos, maior quantidade de saídas disponibilizadas (inclusive com uma análise estatística abrangente) e por envolver também análises filogenéticas para criar os grupos de ortólogos.

Um dos grandes diferenciais do OrthoFinder é que ele não utiliza *e-values* para a avaliação da similaridade entre sequências, pois podem ocorrer falhas no uso dessa medida sem uma devida normalização. O *e-value* mínimo que pode ser obtido para uma determinada sequência de consulta (*query*) diminui com o aumento do tamanho da sequência até que, em um determinado comprimento, o limite inferior dos *e-values* seja atingido e o *BLAST* retorna um valor eletrônico igual a 0. Isso cria um problema, pois sequências longas frequentemente terão *hits* de baixa qualidade com melhores *e-values* do que os melhores *hits* possíveis de sequências curtas. Esse viés de comprimento é removido

nas análises com OrthoFinder, pois é usado uma medida de similaridade entre sequências com base no escore de *bits* normalizado, que leva em consideração os comprimentos das sequências de consulta e ocorrência e a distância filogenética entre as espécies. (EMMS; KELLY, 2015)

Na metodologia do OrthoFinder, a ortologia é definida pela filogenia, e não apenas por medidas de similaridade de sequência, métodos que usam tais pontuações para definir ortólogos na ausência de filogenia podem apenas fornecer suposições. A única maneira de ter certeza de que a atribuição da ortologia está correta é conduzindo uma reconstrução filogenética de todos os genes descendentes de um único gene, este conjunto de genes é um grupo de ortólogos (*orthogroup)*. (EMMS; KELLY, 2019)(LI, Li; STOECKERT; ROOS, 2003). O OrthoFinder superou o OrthoMCL nos testes realizados com o OrthoBench (TRACHANA *et al.*, 2011), medidos pelo *F-score*, ficando até 25% melhor, é importante ressaltar que a acurácia e a recuperação (*recall)* foram balanceados, demonstrando que o método não é tendencioso para o agrupamento excessivo ou insuficiente de proteínas.

A principal diferença entre os *orthogroups* inferidos pelas duas ferramentas é que aqueles produzidos por OrthoFinder abrangem um número maior de espécies do que aqueles recuperados pelo OrthoMCL, portanto, *orthogroups* produzidos pelo primeiro abrangem maiores distâncias filogenéticas (EMMS; KELLY, 2015). Corroborando isso, pode-se verificar na tabela 4 que os grupos construidos com OrthoMCL tendem a ser menores em média, por volta da metade, que os gerados pelo OrthoFinder, e nesse mesmo sentido, o OrthoMCL gerou mais que o dobro de grupos.

Esse resultado, a principio, nos deixou com receio de perder anotações de PFDs no caso de escolher o OrthoFinder, no entanto, podese perceber pelos resultados da tabela 7 que foi conseguido anotar com ele, usando protocolo D, mais que o dobro de PFDs do que usando o OrthoMCL. Outro ponto é que a análise com orthofinder conseguiu agrupar 93% de todas as proteínas, enquanto que com o OrthoMCL foram 83%, isto é, embora existam menos grupos com o orthofinder, esses grupos envolvem 10% a mais de proteínas que a análise com OrthoMCL, aumentando assim o espaço de busca por PFDs agrupadas em *orthogroups*.

Na comparação entre os resultados, os do OrthoFinder foram melhores para o objetivo de agrupar PFDs para prospecção de função. Então por esta, e todas as outras vantagens aqui citadas, o OrthoFinder foi escolhido para ser o software de montagem de grupos de ortólogos no PlantAnnot2. Considera-se que comparar os resultados de homologia do Orthofinder e do OrthoMCL no contexto do sistema é o mesmo que comparar as versões PlantAnnot2 e o PlantAnnot no que se refere a homologia.

Também foi realizado testes usando resultados de homologia fazendo uma combinação entre OrtoFinder e o software CD-HIT. Foi decidido não manter esse resultado pois há poissibilitdade do cd-hit remover duplicações, que são muito presentes em genomas de plantas (FLAGEL; WENDEL, 2009; PANCHY; LEHTI-SHIU; SHIU, 2016), e elas, assim como eventos de splicing podem ter papéis importantes em respostas a estresses abióticos (HAAK *et al.*, 2017) (VANBUREN *et al.*, 2015). Além disso, foram perdidas 15% de anotações de PFDs como pode ser verificado na na tabela 7.

5.3. Redes de coexpressão

Foram usados no PlantAnnot2 apenas dados de RNA-seq baixados do GEO/NCBI por ser um banco de dados que armazena conjunto de dados de expressão gênica com curadoria (EDGAR; DOMRACHEV; LASH, 2002), portanto dados mais confiáveis e com mais informações sobre as amostras (metadados), o que os torna mais rastreáveis.

Mesmo com essa curadoria, ainda foram removidas 10 amostras baixadas do GEO/NCBI das análises, pois não atingiram o padrão de qualidade mínimo exigida pelo software HTSEQ-count. Por recomendação do LSTrAP, deveria existir mais de 20 amostras de cada organismo para montagem das redes de coexpressão, no entando, para Glycine max houve 18 amostras e para Zea mays 12 amostras, pois não foram encontradas mais análises que atendessem ao padrão de qualidade para esses organismos no GEO/NCBI.

Ainda a respeito das amostras, importante falar que na versão 2 do PlantAnnot, diferente da versão1 que usou só amostras de folhas de plantas, além de ter 101 amostras a mais, foram usadas análises com tecidos diferentes e em fases de desenvolvimento diversos da planta. Essa é uma boa prática para montagem das redes de coexpressão, pois verificar os padrões de expressão em vários tecidos e em estágios de desenvolvimento diferentes podem lançar luz sobre quando e onde um gene é necessário, e fornecer pistas sobre a função do gene (PROOST; KRAWCZYK; MUTWIL, 2017).

Como foram usadas quase 3 vezes mais amostras nessa versão do PlantAnnot para construção dos clusters de coexpressão, consequentemente foram obtidos 2,25x mais clusters de coexpressão gênica que na primeira versão do sistema. Pode-se perceber também na tabela 5 que quanto mais amostras são utilizadas para um determinado organismo, mais genes daquele organismo são agrupados em *clusters*. Com o aumento do número de clusters e de genes agrupados, o espaço de consulta por PFDs possivelmente relacionadas a estresses abióticos foi expandido na nova versão do PlantAnnot.

5.4. Caracterização de PFDs

Na busca por PFDs usando o Protocolo A (Tabela 2), a análise usando o Diamond é bastante estringente pois o *e-value* padrão é 10^{-3}, valor que considerado alto, como explicado a seguir. Fazendo uma análise rápida, como o banco de dados NCBI-nr usado tinha 219.524.215 sequencias de proteínas, então esse *e-value* indica que o número esperado de sequencias do nr que poderiam alinhar significativamente ao acaso com o a sequência de consulta (*query*) seria de 219.524. Quanto maior o *e-value*, menor será a restrição para os *matches* de *queries* no NCBI-nr.

É provável que foram perdidas PFDs usando esse *e-value* padrão (lembrando que são recuperadas no protocolo D os ***no match***), mas mesmo sendo rigorosos com relação ao *no match* no Diamond, ainda foi possível recuperar 136.599 sequencias de proteínas. Por outro lado, se fosse configurado um *e-value* menor, a busca por matches seria mais restritiva, e provavelmente seria recuperado um número maior de possíveis PFDs, desde que também não tivessem *match* no Interproscan, no entanto, com um número muito grande de PFD recuperadas, além da possibilidade de existir muitos falsos positivos, a posterior análise manual delas poderia se tornar inviável pelo tempo que teria de ser dedicado a isso. Uma forma de tentar recuperar PFD perdidas devido ao alto *e-value*, é utilizando as buscas textuais no PlantAnnot 2 (Tabela 2 – Protocolo B, C, E e F).

O Diamond foi utilizado apenas como filtro para recuperar PFDs, no entanto ele poderia ter sido usado também para tentar anotar as PFDs, se tivesse sido recuperado mais *matches* em sua análise (atualmente só é recuperado o melhor, que muitas vezes é a própria proteína depositada no NCBI-nr) e fossem analisadas suas anotações. Essa análise do Diamond também pode ter grande utilidade para tentar anotar PFDs de um genoma novo ainda não publicado. Por fim, considera-se que proteínas com *no match* no Diamond e no Interproscan, que faz buscas em 15 bancos de dados diferentes, estão bem caracterizadas como proteínas de função desconhecida (PFD) (GOLLERY *et al.*, 2006).

5.5. Anotação de PFDs

O Protocolo D (Tabela 2) para busca de PFD relacionadas com ortologia e coexpressão foi considerado muito restritivo, uma vez que só recupera proteínas que não possuem nenhuma anotação (*no match* no Diamond e *no match* no Interproscan). Então, para tentar recuperar PFDs possivelmente perdidas usando esse método, foram criados outros protocolos de buscas, são os protocolos E e F respectivamente. Neles podem ser realizadas buscas textuais por palavras que possam vir a caracterizar proteínas ou domínio de função desconhecidas, pois muitas PFDs possuem anotações

não informativas, tais como*: "Unknown protein", "putative protein", "expressed protein", "hypothetical" ou "uncharacterized"*. Infelizmente não existe uma nomenclatura padrão para nomeá-las e isso vai variar muito de organismo para organismo.

Por exemplo, se for feita uma pequena mudança no protocolo E, e usado na busca textual a palavra *"uncharacterized"*, foram recuperadas 2.262 possíveis PFDs para todos os organismos, dentre elas existem 1.101 para o *Sorghum bicolor*, planta que também possui importantes características de tolerância a seca, enquanto que utilizando o protocolo D tinha sido possível recuperar apenas 1 PFD para o Sorgo. Existem muitas outras possibilidades de pesquisas que podem ser usadas no PlantAnnot2 e não apenas os 6 protocolos de buscas, que foram usados aqui apenas como exemplos de busca. Vale salientar que essa forma de busca de PFDs entre a versão mais nova e a mais antiga do sistema continua a mesma.

Outra forma de busca pode ser feita, é através de uma pesquisa reversa, onde ao invés de filtrar PFDs, primeiro associando elas a ortologia e coexpressão para encontrar possíveis anotações. É possível fazer diretamente uma pesquisa pela anotação que deseja-se encontrar, como por exemplo "Ovate", dessa forma, foram encontradas muitas proteínas com anotação "Ovate Family Protein" e essa proteína possui vários estudos sobre seu importante papel em respostas a estresses abióticos (LI, Huifeng *et al.*, 2019; MA *et al.*, 2017; SUN *et al.*, 2020), além disso, está presente em 52 organismos dentro de um mesmo grupo ortólogo (OG0000143), entre eles *Boea hygrométrica* e da *Populus simonii*, portanto é uma proteínas bem conservada.

Na tabela 6 pode-se verificar que a quantidade de PFDs recuperadas usando o protocolo D para *Zea mays* está fora do padrão (*outlier) com* 2.068. O genoma do milho passou por várias rodadas de duplicação do genoma, estimasse que 25% dele seja composto de duplicações (SCHNABLE *et al.*, 2009). Na análise do OrthoFinder foram encontradas 26.465 duplicações, 1.569 PFDs recuperadas com protocolo D pertencem a grupos de ortólogos formado exclusivamente com proteínas de *Zea mays*, isso representa 75% de todas as PFDs desse organismo. Além disso, apenas 5% das PFDs estão em grupos que têm 67 ou mais proteínas, 617 dos grupos exclusivos de *Zea mays* são compostos por apenas com 2 proteínas, 912 com 3 ou menos, 1357 tem 10 ou menos proteínas. A mediana de quantidade de proteínas em grupos ortólogos de Zea mays é 3, o terceiro quartil (Q3) é 7.25. Podemos perceber que *Zea mays* possui muitos grupos pequenos. Na primeira versão do PlantAnnot o organismo *Zea mays* teve esse mesmo comportamento.

Além disso, Zea *mays* é o segundo organismo com mais PFDs (Protocolo A – no match Diamond e no match Interproscan), ele possui 9.783 (tabela 6), ficando atrás apenas do organismo *Panicum virgatum* que possui 10.695. Vale salientar que o número de PFDs desses dois organismos estão bem superiores que dos outros, o terceiro colocado em número de PFDs possui 3.775(*Populus*

simonii). Quando recuperamos apenas PFDs que pertençam a algum grupo ortólogo, *Zea mays* tem 3.515 e *Panicum virgatum*, 6.329, mas quando é filtrada apenas PFDs que pertençam a grupos de ortólogos que tenha alguma proteína com seu mRNA em *clusters* de coexpressão, aí o *Zea mays* se destaca, pois ele possui dados de expressão e *Panicum virgatum* não. Nesse cenário, *Zea mays* ficou com 2.068 PFDs, enquanto *Panicum virgatum* ficou com apenas 192. Vale ainda salientar o fato de que os outros 3 organismos, *Glycine max, Arabidopsis thaliana* e *Oryza sativa*, que também foram usados para montagem dos *clusters* de coexpressão, não ficaram *outliers*, eles têm respectivamente 1, 27 e 141 PFDs recuperados com o protocolo D. Porém, eles possuem bem menos PFDs (protocolo A) que *Zea mays*: *Arabidopsis thaliana* tem 46, *Glycine max* tem 1 e *Oryza sativa* tem 161.

Na tabela 7 é possível verificar que os testes realizados com o OrthoMCL e com o Orthofinder + cd-hit diminuíram a recuperação de PFDs usando os protocolos D, E e F em relação a análise realizada apenas com o Orthofinder. Usando essa última análise, foi possível mais que dobrar a quantidade de PFD recuperadas usando o protocolo D, que é considerado o padrão ouro de busca por PFD. E em relação ao PlantAnnot, o Plantannot2 recuperou, usando o protocolo D, mais de 3x a quantidade de PFDs da primeira versão do sistema. Além disso, 78% de todas as PFDs recuperadas com protocolo D na primeira versão do PlantAnnot também foram encontradas nessa versão mais nova. Lembrando que o protocolo D caracteriza como PFDs aquelas com *no match* no Diamond e no Interproscan. A diferença de tempo entre as análises de Diamond e Interproscan foi de quase 24 meses e os bancos de dados de ambas as ferramentas são atualizados com frequência, é possível que por isso, nem todas as PFDs do PlantAnnot se enquadram mais na categoria de PFD nas análises realizadas pelo PlantaAnnot2.

Sobre o percentual médio de estudos relacionados a estresses abióticos encontrados para as anotações, 67,5% (variando entre 62% a 74%), vale ressaltar que a pesquisa manual foi apenas sobre a anotação exibida no campo *display* da *feature* proteína, é possível que se fosse aprofundada mais a pesquisa, buscando as anotações encontradas nos bancos de dados do Interproscan, fossem achados mais estudos, no entanto, o grande volume de dados envolvido na análise dificulta essa busca manual. Só foram analisados organismos com característica de tolerância a estresses abióticos, não sabemos esse percentual para outros organismos comuns. Outra característica que foi encontrada referente as anotações, é que esse percentual aumenta bastante para grupos com mais de 68 proteínas, passa a 79% (variando entre 75% e 85%), então é mais fácil encontrar estudos para as anotações recuperadas em grupos maiores (vide seção 5.2). Com isso, a versão 2 do PlantAnnot quase dobrou o percentual de estudos relacionados a estresses abióticos encontrados nas anotações dos homólogos de PFDs, saiu de 35% para 67,5%.

5.6. Estudos de caso

No primeiro estudo de caso, a família de proteínas do receptor ABA PYR/PYL/RCAR (PYL) (seção 1.1) é regulador negativo da proteína fosfatase tipo 2C (PP2C) e regulador positivo da classe III da proteína quinase relacionada ao SNF-1 (SnRK2). O ABA se liga a uma proteína PYL, resultando na inibição de PP2Cs através do complexo ABA-PYL-PP2C (Figura 1B). Este complexo leva ao acúmulo de SnRK2s fosforilados, o que leva à fosforilação de fatores de ligação do elemento responsivo a ABA (ABFs) e subsequente expressão do gene ABA para respostas celulares apropriadas e estresses abióticos em plantas (DUARTE *et al.*, 2019)(CUTLER *et al.*, 2010).

No segundo estudo de caso, vale salientar que a anotação "*Protein phosphatase 2C family protein (PP2C)"* apareceu em 57 organismos diferentes, então ela parece ser bem conservada, e entre esses organismos temos a *Populus simonii* com 5 proteínas e o *Oropetium thomaeum* (VANBUREN *et al.*, 2015) com 4 proteínas, que são espécies com importantes características de tolerância a seca. Nesse grupo temos também 21 proteínas cujos mRNAs pertencem a clusters de coexpressão de ensaios de estresses abióticos diversos.

No terceiro estudo de caso, a anotação "serine/threonine protein kinase" refere-se a membros da família SnRK2 que são serina/treonina quinases específicas de plantas envolvidas na resposta das plantas a estresses abióticos e ao desenvolvimento de plantas dependentes de ácido abscísico (ABA), elas são considerados um dos principais reguladores da resposta da planta ao ABA e sua função já foi descrita anteriormente(seção 1.1.1) quando explicamos os fatores de sinalização do ABA na figura 1 (KULIK *et al.*, 2011)

No estudo de caso 4, Esses genes bZIP são nomeados como proteínas de ligação a elementos responsivos a ABA (AREB) ou fatores de ligação ABRE (ABF), e foram funcionalmente caracterizados como importantes fatores na sinalização de ABA quando a planta é submetida a estresse (YOSHIDA *et al.*, 2010)(FUJITA; YOSHIDA; YAMAGUCHI-SHINOZAKI, 2013). Os ABFs também são induzíveis por vários tratamentos de estresse, e cada ABF exibe um padrão de indução exclusivo, sugerindo que eles provavelmente estão envolvidos em diferentes vias de sinalização de estresse mediadas por ABA (CHOI *et al.*, 2000). Além disso, a anotação mostrou-se bem conservada, pois existe em 57 organismos diferentes do mesmo grupo.

No caso da rota de transcrição de sinal do ABA (seção 1.1.1) é possível perceber que existem genes *flat* (plano), que podem permanecer com a expressão estável (não são *upregulated* ou *downregulaed*), e consequentemente não participam de nenhum *cluster* de coexpressão gênica, dificultando assim a recuperação deles através do algoritmo padrão de busca por PFD apresentado nesse trabalho (protocolos D, E e F da tabela 2), que tenta relacioná-las a *clusters* de coexpressão. No

entanto, alguns desses genes podem executar papeis chaves em relação a resposta a estresses abióticos. As plantas utilizam muitos receptores quinases e quinases de transdução de sinal para realizar as respostas fisiológicas de células-alvo a hormônios. A percepção do hormônio, como o ABA, por um receptor, como o PYR/PYL/RCAR, causará eventos transcricionais ou pós-transcricionais, como a fosforilação, que, resultam em respostas fisiológicas ou de desenvolvimento na planta (TAIZ *et al.*, 2017).

O PlantAnnot2 possibilita recuperar esses genes e tentar inferir função para eles através de homologia. Utilizando os filtros da ferramenta, é possível filtrar todas as proteínas que pertencem a grupos de ortólogos e que nesse grupo não exista qualquer proteína com gene presente em *cluster* de coexpressão. No entanto, pode existir um viés nessa forma de tentar recuperar os genes com expressão *flat*, pois podem ser recuperados também genes que foram poucos coexpressos, isto é, com coeficiente de *pearson* (PCC) menor que 0,7, que foi o *cut off* utilizado para montagem dos *clusters*. Para filtrar mais ainda os resultados pode-se utilizar a busca textual para recuperar só proteínas de interesse, como as quinases, e analisar os grupos de ortólogos delas em busca de PFDs.

Corroborando a discussão levantada na seção 5.5 em relação a diferença de tempo entre as datas em que foram realizadas as análises do Diamond e do Interproscan. No estudo de caso utilizado no artigo da versão 1 do Plantannot (VIANA, Marcos José Andrade; ZERLOTINI; DE ALVARENGA MUDADU, 2021), em que a proteína ID "Oropetium_20150105_06293A.v1.0" da planta *Oropetium thomaeum* foi caracterizada como PFD por não ter encontrado, à época, correspondência (*match*) no nr-NCBI nem nos bancos de dados do Interproscan. Na versão atual a mesma proteína não foi caracterizada como PFD por já existir correspondência para ela no nr-NCBI e ter sido encontrada anotação no Interproscan, como pode ser conferido no link https://www.machado.cnptia.embrapa.br/plantannot2/feature/?feature_id=23700890. O mesmo ocorreu para os outros 3 estudos de casos do PlantAnnot referentes aos IDs "KZV45975.1", "KZV43328.1" e "KZV34923.1" da planta *Boea hygrométrica*, todas as 3 proteínas, no PlantAnnot2, encontraram correspondências nas análises do Diamond e do Interproscan e, portanto, não foram caracterizada como PFDs.

5.7. Enriquecimentos de vias.

Após analisar todos os genes agrupados em *clusters* de coexpressão das plantas *Arabidopsis thaliana* e *Oriza sativa*, verificou-se que eles estão enriquecidos para a via de "Resposta a estímulus abióticos" (tabelas 8 e 9), assim como para outras vias importantes que estão conectadas a essa por compartilharem 20% ou mais genes (Figuras 10 e 11), como por exemplo as vias de "Resposta a

quimicos", "Resposta a substâncias orgânicas", "Resposta a compostos oxigênios", "Resposta a estresses", "Resposta celular a estimulus" e "Resposta a hormônios" entre outras.

Não foi percebido muita diferença entre os enriquecimentos feitos para genes do PlantAnnot e o PlantAnnot2 para as análises de enriquecimento feitas para Oriza sativa. Já para para Arabidopsis thaliana, houve uma certa melhora no enriquecimento, principalmente no que se refere ao enriquecimentos dos genes em outras vias que são interligadas a via de "Respostas a estímulus abióticos. Isso pode significar, que quanto maior o número de amostras envolvidas na montagem dos clusters, melhor pode ficar a montagem deles.

Outro ponto que pode-se destacar, é que o número de genes na via de "Resposta a estímulos abióticos", e em outras interligadas a ela, aumentou, por exemplo, de 12.643 para *Arabidopsis thaliana* no PlantAnnot para 19.234 no PlantAnnot2. Então, houve um crescimento de 52% na quantidade de genes em vias importantes para respostas a estresses abióticos em plantas.

5.8. Ferramentas semelhantes

Existem algumas ferramentas web com objetivos semelhantes, atendendo parcialmente ou superando os objetivos do *PlantAnnot2*. No entanto essas ferramentas ou só funcionam online e não possuem código aberto, como é o caso do *PLAZA 3.0* (CO-EXPRESSION; BEL; COPPENS, 2017) e da *GeNET* (DESAI *et al.*, 2017), ou exigem dados de entrada como uma lista de genes e uma matriz de expressão gênica pré-processada, como exemplo temos o CoExpNetViz (TZFADIA *et al.*, 2016), ou são mais genéricas e buscam qualquer tipo de anotação em qualquer tipo de organismo, como o *conekt* (PROOST; MUTWIL, 2018) ou usam dados de microarray, que diferente do RNA-seq, não conseguem recuperar todos os dados de expressão gênica, é o caso do *genevestigator* (HRUZ *et al.*, 2008). O PlantAnnot2 tem um papel bem específico de buscar genes possivelmente relacionados em repostas a stresses abióticos em plantas. Um de seus diferenciais é a grande quantidade de organismos, são 67 genomas exclusivamente de plantas envolvidos nesse estudo, outras ferramentas como *Planet* (MUTWIL *et al.*, 2011) , o *PODC (Plant Omics Data Center)* (OHYANAGI *et al.*, 2015) *e o NetMiner-an* (YU *et al.*, 2018), quando não usam apenas um organismos, usam pouca quantidade e muitas vezes apenas de sistemas modelos como *Arabidopsis thaliana*. O algoritmo utilizado no PlantAnnot2 para caractrerizar e anotar as PFDs inclui meta análises de bioinformática e cruzamentos de dados que envolve buscas de similaridades de sequências, ortologia e redes de coexpressão gênica. Outras ferramentas tentam inferir função usando só dados de expressão gênica, como por exemplo *WebCEMiTool* (CARDOZO *et al.*, 2019), ou apenas dados de ortologia, como a *OrthoVenn* (WANG

et al., 2015) e *OrthoVenn2* (XU *et al.*, 2019), sem tentar integrá-los. A maioria das ferramentas citadas, que usam dados de ortologia para inferências de função, utilizam o OrthoMCL, que como já foi discutido aqui, possui viés com relação a montagem dos grupos ortólogos.

A escolha dos dados foi criteriosa, retirando a maioria de um banco de dados confiável que é o Phytozome v12.1. Foram envolvidos nas análises espécies de plantas com importantes características de tolerância a estresses abióticos, como a: *Boea hygrometrica*, *Orophetium thomaeum* e *Sorghum bicolor*, e os novíssimos genomas da *Pearl millet e Populus simonii*. Os dados de expressão gênica foram baixados apenas do GEO/NCBI que possui dados mais curados que os do SRA/NCBI, onde poderíamos encontrar muitos outros dados de expressão gênica, no entanto, tão importante quanto os dados que usamos como entrada para as análises, são os dados que deixamos de usar, pois esses últimos podem gerar ruídos indesejáveis. As análises realizadas no PlantAnnot2 usam ferramentas que têm alto níveis de acurácia e que prezam pela qualidade dos dados analisados, todos os dados reprovados pelos testes de qualidades foram removidos das análises.

O esquema de banco de dados utilizado, *Generic Model Organism Database* (GMOD) Chado, é um clássico bastante conhecido e testado para armazenamento de dados biológicos, além de ser totalmente compatível com mais de 40 ferramentas que compõem o projeto GMOD (http://gmod.org/wiki/GMOD_Components). Um dos grandes obstálculos para utilização do *Chado* era o carregamento dos dados, e o software *Machado*, usado aqui, fornece scripts em python para carregamento dos principais tipos de dados biológicos, além de uma intereface web para navegação pelos dados armazenados, além de ser um software muito bem documentado, de código aberto e disponibilizado de graça via *github* (https://github.com/lmb-embrapa). Ainda são disponibilizados APIs (https://www.machado.cnptia.embrapa.br/plantannot2/api/) para integração, comunicação, buscas e recuperação de dados por meios de outros sistemas e/ou plataformas, como *mobiles*.

CAPÍTULO 6: CONCLUSÃO

Plantas geneticamente modificadas tolerantes as grandes mudanças climáticas pelas quais o mundo vem passando, serão fundamentais para que o Brasil e o mundo atendam a crescente demanda por alimentos prevista até 2050 pela Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) de 70%, e dessa forma, contribuir com a segurança alimentar mundial. Além disso, a criação dessas plantas no Brasil pode trazer outros benefícios, tais como a expansão de fronteiras agrícolas. Plantas de alto valor agronômico que atualmente não são cultivadas em certas

regiões por falta de condições climáticas ou de solo adequados, poderiam vir a ser, incrementando ainda mais a produção agrícola nacional e desenvolvendo outras áreas de nosso imenso país.

O surgimento de abordagens ômicas têm papel fundamental em promover pesquisas eficazes no campo, facilitando as investigações de modelos de referência para um número crescente de espécies. Abordagens multiníveis integradas, baseadas em investigações moleculares em níveis de genômica, transcriptômica, proteômica e metabolômica, agora são viáveis, expandindo as oportunidades de esclarecer os principais aspectos moleculares envolvidos nas respostas a estresses abióticos. (AMBROSINO *et al.*, 2020).

O PlantAnnot2 mostrou que pode ser valiosa na busca de alvos interessantes para serem usadas como prova de conceito em pipelines de produção de plantas geneticamente modificadas tolerantes a estresses abióticos. Além disso, acredita-se que as melhorias realizadas nessa versão do sistema fizeram com que fosse possível recuperar mais de 3 vezes a quantidade de PFDs encontradas na primeira versão do sistema (usando protocolo D) possivelmente associadas em respostas a estresses abióticos em plantas e também que saltasse de 35% para 67,5% o percentual de sucesso em buscas de estudos relacionados a estresses abióticos para as possíveis anotações dessas PFDs. Por último, é importante salientar que o pipeline e o sistema da web proposto têm uma cobertura muito mais ampla do que a descrita, pois eles podem suportar genomas de qualquer tipo de organismo e seus transcritos em qualquer situação. Plantas e estresses abióticos foram escolhidos como um "estudo de caso".

CAPÍTULO 7: PERSPECTIVAS

- Usar técnicas de aprendizagem de máquina para calcular atributos que consigamos calcular só a partir da sequência primária, só a partir da sequência de DNA ou Proteína. Fazer um grande trabalho de curadoria semi manual, encontrar os genes que já sabemos que tem um papel importante na resposta a stress abióticos e os genes que não tem papel nenhum. Aí usamos esses grupos para treinar nosso modelo e aplicamos em outros genes no hit... para verificar se conseguimos recuperar alguns genes também dessa forma.
- Inserir nos sistemas outros tipos de dados, como metabolômica.
- Depois de identificados fatores de transcrição (FT) tentar criar os subclusters com eles e verificar se existem PFDs, pois os FT além de serem “genes hubs” são proteínas envolvidas nas etapas iniciais de expressão e regulação gênica e na transdução de sinais, em resposta aos estresses.
- Analisar também RNAs não codificantes (Long Non Coding, MicroRNAs).

- Incluir plugin Cytoscape para ver as redes de coexpressão no *browser*.
- Criar subcluster de coexpressão menores para serem melhor visualizados.
- Automatizar a busca por estudos de estresses abióticos referentes as anotações encontradas em grupos ortólogo de PFDs.
- Possibilitar upload de dados por parte dos usuários.
- Buscar estudos de casos para tentar tampar *gaps* em vias biológicas, e nesses casos o psi-blast pode ser utilizado para validar o resultado.

Bibliografia

ALEXEYENKO, A.; TAMAS, I.; LIU, G.; SONNHAMMER, E. L. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. **Bioinformatics**, vol. 22, no. 14, p. e9–e15, 2006. DOI 10.1093/bioinformatics/btl213. Available at: https://doi.org/10.1093/bioinformatics/btl213.

ALLEN, J. D.; XIE, Y.; CHEN, M.; GIRARD, L.; XIAO, G. Comparing statistical methods for constructing large scale gene networks. **PLoS ONE**, vol. 7, no. 1, p. 17–19, 2012. https://doi.org/10.1371/journal.pone.0029348.

ALTENHOFF, A. M.; SCHNEIDER, A.; GONNET, G. H.; DESSIMOZ, C. OMA 2011: Orthology inference among 1000 complete genomes. **Nucleic Acids Research**, vol. 39, no. SUPPL. 1, p. 289–294, 2011. https://doi.org/10.1093/nar/gkq1238.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, vol. 25, no. 17, p. 3389–3402, 1 Sep. 1997. DOI 10.1093/nar/25.17.3389. Available at: https://doi.org/10.1093/nar/25.17.3389.

AMBROSINO, L.; COLANTUONO, C.; DIRETTO, G.; FIORE, A.; CHIUSANO, M. L. Bioinformatics Resources for Plant Abiotic Stress Responses: State of the Art and Opportunities in the Fast Evolving -Omics Era Luca. **Plants**, vol. 9, no. 5, 2020. https://doi.org/10.3390/plants9050591.

ANDERS, S.; PYL, P. T.; HUBER, W. HTSeq—a Python framework to work with high-throughput sequencing data. **Bioinformatics**, vol. 31, no. 2, p. 166–169, 15 Jan. 2015. DOI 10.1093/bioinformatics/btu638. Available at: https://doi.org/10.1093/bioinformatics/btu638.

ATTWOOD, T. K.; COLETTA, A.; MUIRHEAD, G.; PAVLOPOULOU, A.; PHILIPPOU, P. B.; POPOV, I.; ROMÁ-MATEO, C.; THEODOSIOU, A.; MITCHELL, A. L. The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. **Database**, vol. 2012, 1 Jan. 2012. DOI 10.1093/database/bas019. Available at: https://doi.org/10.1093/database/bas019.

BAILEY-SERRES, J.; PARKER, J. E.; AINSWORTH, E. A.; OLDROYD, G. E. D.; SCHROEDER, J. I. Genetic strategies for improving crop yields. **Nature**, vol. 575, no. 7781, p. 109–118, 6 Nov. 2019. DOI 10.1038/s41586-019-1679-0. Available at: http://www.nature.com/articles/s41586-019-1679-0.

BALLOUZ, S.; VERLEYEN, W.; GILLIS, J. Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers. **Bioinformatics**, vol. 31, no. 13, p. 2123–2130, 2015.

https://doi.org/10.1093/bioinformatics/btv118.

BARTELS, D; MATTAR, M. Z. M. Oropetium thomaeum. A resurrection grass with a diploid genome. **Maydica (Italy)**, 2002. .

BARTELS, Dorothea; SALAMINI, F. Desiccation Tolerance in the Resurrection Plant. **Plant Physiol.**, vol. 127, no. December, p. 1346–1353, 2001. DOI https://doi.org/10.1104/pp.010765. Available at: https://doi.org/10.1104/pp.010765%0A.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, vol. 30, no. 15, p. 2114–2120, 1 Aug. 2014. DOI 10.1093/bioinformatics/btu170. Available at: https://doi.org/10.1093/bioinformatics/btu170.

BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, vol. 12, no. 1, p. 59–60, 2014. https://doi.org/10.1038/nmeth.3176.

CARDOZO, L. E.; RUSSO, P. S. T.; GOMES-CORREIA, B.; ARAUJO-PEREIRA, M.; SEPÚLVEDA-HERMOSILLA, G.; MARACAJA-COUTINHO, V.; NAKAYA, H. I. WebCEMiTool: Co-expression modular analysis made easy. **Frontiers in Genetics**, vol. 10, no. MAR, p. 1–5, 2019. https://doi.org/10.3389/fgene.2019.00146.

CHALLINOR, A. J.; SIMELTON, E. S.; FRASER, E. D. G.; HEMMING, D.; COLLINS, M. Increased crop failure due to climate change: Assessing adaptation options using models and socio-economic data for wheat in China. **Environmental Research Letters**, vol. 5, no. 3, 2010. https://doi.org/10.1088/1748-9326/5/3/034012.

CHEN, Kong; LI, G. J.; BRESSAN, R. A.; SONG, C. P.; ZHU, J. K.; ZHAO, Y. Abscisic acid dynamics, signaling, and functions in plants. **Journal of Integrative Plant Biology**, vol. 62, no. 1, p. 25–54, 2020. https://doi.org/10.1111/jipb.12899.

CHEN, Kunling; WANG, Y.; ZHANG, R.; ZHANG, H.; GAO, C. CRISPR / Cas Genome Editing and Precision Plant Breeding in Agriculture. , p. 1–31, 2019. .

CHOI, H. I.; HONG, J. H.; HA, J. O.; KANG, J. Y.; KIM, S. Y. ABFs, a family of ABA-responsive element binding factors. **Journal of Biological Chemistry**, vol. 275, no. 3, p. 1723–1730, 2000. https://doi.org/10.1074/jbc.275.3.1723.

CINTRA, L. C. **Instruções para uso do Open Grid Engine no Laboratório Multiusuário de Bioinformática**. Campinas-SP: [*s. n.*], 2016. Available at: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1066147/1/Doc154.pdf.

CO-EXPRESSION, E. P.; BEL, M. Van; COPPENS, F. A guide to the PLAZA 3.0 plant comparative genomic database. [*S. l.*: *s. n.*], 2017. vol. 1533, p. 201–212. DOI 10.1007/978-1-4939-6658-5. Available at: http://link.springer.com/10.1007/978-1-4939-6658-5.

CUTLER, S. R.; RODRIGUEZ, P. L.; FINKELSTEIN, R. R.; ABRAMS, S. R. **Abscisic acid:**

Emergence of a core signaling network. [*S. l.*: *s. n.*], 2010. vol. 61, . https://doi.org/10.1146/annurev-arplant-042809-112122.

DALQUEN, D. A.; DESSIMOZ, C. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. **Genome Biology and Evolution**, vol. 5, no. 10, p. 1800–1806, 2013. https://doi.org/10.1093/gbe/evt132.

DE ALVARENGA MUDADU, M.; ZERLOTINI, A. Machado: Open source genomics data integration framework. **GigaScience**, vol. 9, no. 9, p. 1–6, 2020. https://doi.org/10.1093/gigascience/giaa097.

DESAI, A. P.; RAZEGHIN, M.; MERUVIA-PASTOR, O.; PEÑA-CASTILLO, L. GeNET: A web application to explore and share Gene Co-expression Network Analysis data. **PeerJ**, vol. 2017, no. 8, p. 3678, 2017. https://doi.org/10.7717/peerj.3678.

DESSIMOZ, C.; GABALDÓN, T.; ROOS, D. S.; SONNHAMMER, E. L. L.; HERRERO, J.; ALTENHOFF, A.; APWEILER, R.; ASHBURNER, M.; BLAKE, J.; BOECKMANN, B.; BRIDGE, A.; BRUFORD, E.; CHERRY, M.; CONTE, M.; DANNIE, D.; DATTA, R.; DOMELEVO ENTFELLNER, J. B.; EBERSBERGER, I.; GALPERIN, M.; … ZDOBNOV, E. Toward community standards in the quest for orthologs. **Bioinformatics**, vol. 28, no. 6, p. 900–904, 2012. https://doi.org/10.1093/bioinformatics/bts050.

DOLFERUS, R. To grow or not to grow: A stressful decision for plants. **Plant Science**, 2014. DOI 10.1016/j.plantsci.2014.10.002. Available at: http://dx.doi.org/10.1016/j.plantsci.2014.10.002.

DUARTE, K. E.; DE SOUZA, W. R.; SANTIAGO, T. R.; SAMPAIO, B. L.; RIBEIRO, A. P.; COTTA, M. G.; DA CUNHA, B. A. D. B.; MARRACCINI, P. R. R.; KOBAYASHI, A. K.; MOLINARI, H. B. C. Identification and characterization of core abscisic acid (ABA) signaling components and their gene expression profile in response to abiotic stresses in Setaria viridis. **Scientific Reports**, vol. 9, no. 1, p. 1–16, 2019. https://doi.org/10.1038/s41598-019-40623-5.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. **Nucleic Acids Research**, vol. 30, no. 1, p. 207–210, 2002. https://doi.org/10.1093/nar/30.1.207.

EMMS, D. M.; KELLY, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. **Genome Biology**, vol. 20, no. 1, p. 1–14, 2019. https://doi.org/10.1186/s13059-019-1832-y.

EMMS, D. M.; KELLY, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biology**, vol. 16, no. 1, p. 1–14, 2015. DOI 10.1186/s13059-015-0721-2. Available at: http://dx.doi.org/10.1186/s13059-015-0721-2.

ENRIGHT, A. J.; VAN DONGEN, S.; OUZOUNIS, C. A. An efficient algorithm for large-scale

detection of protein families. **Nucleic Acids Research**, vol. 30, no. 7, p. 1575–1584, 1 Apr. 2002. DOI 10.1093/nar/30.7.1575. Available at: https://doi.org/10.1093/nar/30.7.1575.

FAO. **Food Security and Nutrition in the World**. [*S. l.*: *s. n.*], 2020.

FEDOROFF, N. V.; BATTISTI, D. S.; BEACHY, R. N.; COOPER, P. J. M.; FISCHHOFF, D. A.; HODGES, C. N.; KNAUF, V. C.; LOBELL, D.; MAZUR, B. J.; MOLDEN, D.; REYNOLDS, M. P.; RONALD, P. C.; ROSEGRANT, M. W.; SANCHEZ, P. A.; VONSHAK, A.; ZHU, J. K. Radically rethinking agriculture for the 21st century. **Science**, vol. 327, no. 5967, p. 833–834, 2010. https://doi.org/10.1126/science.1186834.

FINN, R. D.; ATTWOOD, T. K.; BABBITT, P. C.; BATEMAN, A.; BORK, P.; BRIDGE, A. J.; CHANG, H. Y.; DOSZTANYI, Z.; EL-GEBALI, S.; FRASER, M.; GOUGH, J.; HAFT, D.; HOLLIDAY, G. L.; HUANG, H.; HUANG, X.; LETUNIC, I.; LOPEZ, R.; LU, S.; MARCHLER-BAUER, A.; … MITCHELL, A. L. InterPro in 2017-beyond protein family and domain annotations. **Nucleic Acids Research**, vol. 45, no. D1, p. D190–D199, 2017. https://doi.org/10.1093/nar/gkw1107.

FINN, R. D.; COGGILL, P.; EBERHARDT, R. Y.; EDDY, S. R.; MISTRY, J.; MITCHELL, A. L.; POTTER, S. C.; PUNTA, M.; QURESHI, M.; SANGRADOR-VEGAS, A.; SALAZAR, G. A.; TATE, J.; BATEMAN, A. The Pfam protein families database: towards a more sustainable future. **Nucleic Acids Research**, vol. 44, no. D1, p. D279–D285, 4 Jan. 2016. DOI 10.1093/nar/gkv1344. Available at: https://doi.org/10.1093/nar/gkv1344.

FITCH, W. M. Distinguishing homologous from analogous proteins. **Systematic Zoology**, vol. 19, no. 2, p. 99–113, 1970. https://doi.org/10.2307/2412448.

FLAGEL, L. E.; WENDEL, J. F. Gene duplication and evolutionary novelty in plants. **New Phytologist**, vol. 183, no. 3, p. 557–564, 2009. https://doi.org/10.1111/j.1469-8137.2009.02923.x.

FUJITA, Y.; YOSHIDA, T.; YAMAGUCHI-SHINOZAKI, K. Pivotal role of the AREB/ABF-SnRK2 pathway in ABRE-mediated transcription in response to osmotic stress in plants. **Physiologia Plantarum**, vol. 147, no. 1, p. 15–27, 2013. https://doi.org/10.1111/j.1399-3054.2012.01635.x.

GABALDÓN, T.; KOONIN, E. V. Functional and evolutionary implications of gene orthology. **Nature Reviews Genetics**, vol. 14, no. 5, p. 360–366, 2013. DOI 10.1038/nrg3456. Available at: http://dx.doi.org/10.1038/nrg3456.

GAO, H.; GADLAGE, M. J.; LAFITTE, H. R.; LENDERTS, B.; YANG, M.; SCHRODER, M.; FARRELL, J.; SNOPEK, K.; PETERSON, D.; FEIGENBUTZ, L.; JONES, S.; ST CLAIR, G.; RAHE, M.; SANYOUR-DOYEL, N.; PENG, C.; WANG, L.; YOUNG, J. K.; BEATTY, M.; DAHLKE, B.; … MEELEY, R. B. Superior field performance of waxy corn engineered using CRISPR–Cas9. **Nature Biotechnology**, vol. 38, no. 5, p. 579–581, 2020. DOI 10.1038/s41587-020-

0444-0. Available at: https://doi.org/10.1038/s41587-020-0444-0.

GE, S. X.; JUNG, D.; JUNG, D.; YAO, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. **Bioinformatics**, vol. 36, no. 8, p. 2628–2629, 2020. https://doi.org/10.1093/bioinformatics/btz931.

GENTZSCH, W. Sun Grid Engine: towards creating a compute power grid. 2001. **Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid** [...]. [*S. l.*: *s. n.*], 2001. p. 35–36. https://doi.org/10.1109/CCGRID.2001.923173.

GOLLERY, M.; HARPER, J.; CUSHMAN, J.; MITTLER, T.; GIRKE, T.; ZHU, J.-K.; BAILEY-SERRES, J.; MITTLER, R. What makes species unique? The contribution of proteins with obscure features. **Genome biology**, vol. 7, no. 7, p. R57, 2006. DOI 10.1186/gb-2006-7-7-r57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16859532.

GOODSTEIN, D. M.; SHU, S.; HOWSON, R.; NEUPANE, R.; HAYES, R. D.; FAZO, J.; MITROS, T.; DIRKS, W.; HELLSTEN, U.; PUTNAM, N.; ROKHSAR, D. S. Phytozome: A comparative platform for green plant genomics. **Nucleic Acids Research**, vol. 40, no. D1, p. 1178–1186, 2012. https://doi.org/10.1093/nar/gkr944.

GSA. Genetics Society of America (GSA) Journal. [*s. d.*]. Available at: https://genetics-gsa.org/.

HAAK, D. C.; FUKAO, T.; GRENE, R.; HUA, Z.; IVANOV, R.; PERRELLA, G.; LI, S. Multilevel regulation of abiotic stress responses in plants. **Frontiers in Plant Science**, vol. 8, no. September, p. 1–24, 2017. https://doi.org/10.3389/fpls.2017.01564.

HAFT, D. H.; SELENGUT, J. D.; RICHTER, R. A.; HARKINS, D.; BASU, M. K.; BECK, E. TIGRFAMs and Genome Properties in 2013. **Nucleic Acids Research**, vol. 41, no. D1, p. D387–D395, 1 Jan. 2013. DOI 10.1093/nar/gks1234. Available at: https://doi.org/10.1093/nar/gks1234.

HIRAYAMA, T.; SHINOZAKI, K. Research on plant abiotic stress responses in the post-genome era: Past, present and future. **Plant Journal**, vol. 61, no. 6, p. 1041–1052, 2010. https://doi.org/10.1111/j.1365-313X.2010.04124.x.

HRUZ, T.; LAULE, O.; SZABO, G.; WESSENDORP, F.; BLEULER, S.; OERTLE, L.; WIDMAYER, P.; GRUISSEM, W.; ZIMMERMANN, P. Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. **Advances in BioinformaticHRUZ, T. et al. Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. Advances in Bioinformatics, v. 2008, p. 1–5, 2008. s**, vol. 2008, p. 1–5, 2008. https://doi.org/10.1155/2008/420747.

HULO, N.; SIGRIST, C. J. A.; LE SAUX, V.; LANGENDIJK-GENEVAUX, P. S.; BORDOLI, L.; GATTIKER, A.; DE CASTRO, E.; BUCHER, P.; BAIROCH, A. Recent improvements to the PROSITE database. **Nucleic Acids Research**, vol. 32, no. suppl_1, p. D134–D137, 1 Jan. 2004. DOI

10.1093/nar/gkh044. Available at: https://doi.org/10.1093/nar/gkh044.

ICRISAT. International Pearl Millet do Genome Sequencing Consortium (IPMGSC). 2020. Available at: https://cegresources.icrisat.org/data_public/PearlMillet_Genome/v1.1/.

JONES, P.; BINNS, D.; CHANG, H. Y.; FRASER, M.; LI, W.; MCANULLA, C.; MCWILLIAM, H.; MASLEN, J.; MITCHELL, A.; NUKA, G.; PESSEAT, S.; QUINN, A. F.; SANGRADOR-VEGAS, A.; SCHEREMETJEW, M.; YONG, S. Y.; LOPEZ, R.; HUNTER, S. InterProScan 5: Genome-scale protein function classification. **Bioinformatics**, vol. 30, no. 9, p. 1236–1240, 2014. https://doi.org/10.1093/bioinformatics/btu031.

KASUGA, M.; LIU, Q.; MIURA, S.; YAMAGUCHI-SHINOZAKI, K.; SHINOZAKI, K. Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. **Nature Biotechnology**, vol. 17, no. 3, p. 287–291, 1999. DOI 10.1038/7036. Available at: https://doi.org/10.1038/7036.

KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature Methods**, vol. 12, no. 4, p. 357–360, 2015. DOI 10.1038/nmeth.3317. Available at: https://doi.org/10.1038/nmeth.3317.

KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. **Annual Review of Genetics**, vol. 39, p. 309–338, 2005. https://doi.org/10.1146/annurev.genet.39.073003.114725.

KULIK, A.; WAWER, I.; KRZYWIŃSKA, E.; BUCHOLC, M.; DOBROWOLSKA, G. SnRK2 protein Kinases - Key regulators of plant response to abiotic stresses. **OMICS A Journal of Integrative Biology**, vol. 15, no. 12, p. 859–872, 2011. https://doi.org/10.1089/omi.2011.0091.

LAFOND, M.; MEGHDARI MIARDAN, M.; SANKOFF, D. Accurate prediction of orthologs in the presence of divergence after duplication. **Bioinformatics**, vol. 34, no. 13, p. i366–i375, 2018. https://doi.org/10.1093/bioinformatics/bty242.

LEES, J.; YEATS, C.; PERKINS, J.; SILLITOE, I.; RENTZSCH, R.; DESSAILLY, B. H.; ORENGO, C. Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. **Nucleic Acids Research**, vol. 40, no. D1, p. D465–D471, 1 Jan. 2012. DOI 10.1093/nar/gkr1181. Available at: https://doi.org/10.1093/nar/gkr1181.

LENTON, T. M.; CROUCH, M.; JOHNSON, M.; PIRES, N.; DOLAN, L. First plants cooled the Ordovician. **Nature Geoscience**, vol. 5, no. 2, p. 86–89, 2012. DOI 10.1038/ngeo1390. Available at: http://dx.doi.org/10.1038/ngeo1390.

LETUNIC, I.; DOERKS, T.; BORK, P. SMART 7: recent updates to the protein domain annotation resource. **Nucleic Acids Research**, vol. 40, no. D1, p. D302–D305, 1 Jan. 2012. DOI 10.1093/nar/gkr931. Available at: https://doi.org/10.1093/nar/gkr931.

LI, H.; DONG, Q.; ZHAO, Q.; RAN, K. Genome-wide identification, expression profiling, and

protein-protein interaction properties of ovate family proteins in apple. **Tree Genetics & Genomes**, vol. 15, no. 3, p. 45, 2019. DOI 10.1007/s11295-019-1354-5. Available at: https://doi.org/10.1007/s11295-019-1354-5.

LI, L.; STOECKERT, C. J. J.; ROOS, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. **Genome Research**, vol. 13, no. 9, p. 2178–2189, 2003. DOI 10.1101/gr.1224503.candidates. Available at: http://genome.cshlp.org/cgi/content/full/13/9/2178.

LI, W.; GODZIK, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, vol. 22, no. 13, p. 1658–1659, 2006. https://doi.org/10.1093/bioinformatics/btl158.

LUHUA, S.; HEGIE, A.; SUZUKI, N.; SHULAEV, E.; LUO, X.; CENARIU, D.; MA, V.; KAO, S.; LIM, J.; GUNAY, M. B.; OOSUMI, T.; LEE, S. C.; HARPER, J.; CUSHMAN, J.; GOLLERY, M.; GIRKE, T.; BAILEY-SERRES, J.; STEVENSON, R. A.; ZHU, J.-K.; MITTLER, R. Linking genes of unknown function with abiotic stress responses by high-throughput phenotype screening. **Physiologia Plantarum**, vol. 148, no. 3, p. 322–333, Jul. 2013. DOI 10.1111/ppl.12013. Available at: http://doi.wiley.com/10.1111/ppl.12013.

MA, Y.; YANG, C.; HE, Y.; TIAN, Z.; LI, J. Rice OVATE family protein 6 regulates plant development and confers resistance to drought and cold stresses. **Journal of Experimental Botany**, vol. 68, no. 17, p. 4885–4898, 13 Oct. 2017. DOI 10.1093/jxb/erx309. Available at: https://doi.org/10.1093/jxb/erx309.

MAIA, G.; MORAES, A. De. Fatores de estresse no milho são diversos e exigem monitoramento constante. **Visão Agrícola**, vol. jul-dez, no. 13, p. 30–34, 2015. .

MANTRI, N.; PATADE, V.; PENNA, S.; FORD, R.; PANG, E. Abiotic Stress Responses in Plants. **Journal of Chemical Information and Modeling**. [*S. l.*: *s. n.*], 2013. vol. 53, p. 1689–1699. https://doi.org/10.1017/CBO9781107415324.004.

MARCHLER-BAUER, A.; DERBYSHIRE, M. K.; GONZALES, N. R.; LU, S.; CHITSAZ, F.; GEER, L. Y.; GEER, R. C.; HE, J.; GWADZ, M.; HURWITZ, D. I.; LANCZYCKI, C. J.; LU, F.; MARCHLER, G. H.; SONG, J. S.; THANKI, N.; WANG, Z.; YAMASHITA, R. A.; ZHANG, D.; ZHENG, C.; BRYANT, S. H. CDD: NCBI's conserved domain database. **Nucleic Acids Research**, vol. 43, no. D1, p. D222–D226, 28 Jan. 2015. DOI 10.1093/nar/gku1221. Available at: https://doi.org/10.1093/nar/gku1221.

MCCORMICK, R. F.; TRUONG, S. K.; SREEDASYAM, A.; JENKINS, J.; SHU, S.; SIMS, D.; KENNEDY, M.; AMIREBRAHIMI, M.; WEERS, B. D.; MCKINLEY, B.; MATTISON, A.; MORISHIGE, D. T.; GRIMWOOD, J.; SCHMUTZ, J.; MULLET, J. E. The Sorghum bicolor

reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. **Plant Journal**, vol. 93, no. 2, p. 338–354, 2018. https://doi.org/10.1111/tpj.13781.

MI, H.; POUDEL, S.; MURUGANUJAN, A.; CASAGRANDE, J. T.; THOMAS, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. **Nucleic Acids Research**, vol. 44, no. D1, p. D336–D342, 4 Jan. 2016. DOI 10.1093/nar/gkv1194. Available at: https://doi.org/10.1093/nar/gkv1194.

MOLINARI, H. B. C.; VIEIRA, L. R.; SILVA, N. V. e; PRADO, G. S.; FILHO, J. H. L. **Tecnologia CRISPR na edição genômica de plantas: biotecnologia aplicada à agricultura**. EMBRAPA. Brasília,DF: [*s. n.*], 2020. Available at: https://www.alice.cnptia.embrapa.br/alice/handle/doc/1126157.

MORTAZAVI, A.; WILLIAMS, B. A.; MCCUE, K.; SCHAEFFER, L.; WOLD, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature Methods**, vol. 5, no. 7, p. 621–628, 2008. DOI 10.1038/nmeth.1226. Available at: https://doi.org/10.1038/nmeth.1226.

MUNGALL, C. J.; EMMERT, D. B.; GELBART, W. M.; DE GREY, A.; LETOVSKY, S.; LEWIS, S. E.; RUBIN, G. M.; SHU, S. Q.; WIEL, C.; ZHANG, P.; ZHOU, P. A Chado case study: An ontology-based modular schema for representing genome-associated biological information. **Bioinformatics**, vol. 23, no. 13, p. 337–346, 2007. https://doi.org/10.1093/bioinformatics/btm189.

MUTWIL, M.; KLIE, S.; TOHGE, T.; GIORGI, F. M.; WILKINS, O.; CAMPBELL, M. M.; FERNIE, A. R.; USADEL, B.; NIKOLOSKI, Z.; PERSSON, S. PlaNet: Combined sequence and expression comparisons across plant networks derived from seven species. **Plant Cell**, vol. 23, no. 3, p. 895–910, 2011. https://doi.org/10.1105/tpc.111.083667.

NEHRT, N. L.; CLARK, W. T.; RADIVOJAC, P.; HAHN, M. W. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. **PLOS Computational Biology**, vol. 7, no. 6, p. e1002073, 9 Jun. 2011. Available at: https://doi.org/10.1371/journal.pcbi.1002073.

NEPOMUCENO, A. L.; NEUMAIER, N.; FARIAS, J. R. B.; OYA, T. Tolerância à seca em plantas: mecanismos fisiológicos e moleculares. **Biotecnologia Ciência & Desenvolvimento**, vol. 23, no. 1, p. 12–18, 2001. .

NOGUÉ, F.; MARA, K.; COLLONNIER, C.; CASACUBERTA, J. M. Genome engineering and plant breeding: impact on trait discovery and development. **Plant Cell Reports**, vol. 35, no. 7, p. 1475–1486, 2016. https://doi.org/10.1007/s00299-016-1993-z.

NUCCIO, M. L.; PAUL, M.; BATE, N. J.; COHN, J.; CUTLER, S. R. Where are the drought tolerant crops? An assessment of more than two decades of plant biotechnology effort in crop improvement. **Plant Science**, vol. 273, p. 110–119, Aug. 2018. DOI 10.1016/j.plantsci.2018.01.020. Available at:

https://linkinghub.elsevier.com/retrieve/pii/S016894521731213X.

NUTAN, K. K.; RATHORE, R. S.; TRIPATHI, A. K.; MISHRA, M.; PAREEK, A.; SINGLA-PAREEK, S. L. Integrating the dynamics of yield traits in rice in response to environmental changes. **Journal of Experimental Botany**, vol. 71, no. 2, p. 490–506, 7 Jan. 2020. DOI 10.1093/jxb/erz364. Available at: https://academic.oup.com/jxb/article/71/2/490/5549718.

OATES, M. E.; STAHLHACKE, J.; VAVOULIS, D. V; SMITHERS, B.; RACKHAM, O. J. L.; SARDAR, A. J.; ZAUCHA, J.; THURLBY, N.; FANG, H.; GOUGH, J. The SUPERFAMILY 1.75 database in 2014: a doubling of data. **Nucleic Acids Research**, vol. 43, no. D1, p. D227–D233, 28 Jan. 2015. DOI 10.1093/nar/gku1041. Available at: https://doi.org/10.1093/nar/gku1041.

OHYANAGI, H.; TAKANO, T.; TERASHIMA, S.; KOBAYASHI, M.; KANNO, M.; MORIMOTO, K.; KANEGAE, H.; SASAKI, Y.; SAITO, M.; ASANO, S.; OZAKI, S.; KUDO, T.; YOKOYAMA, K.; AYA, K.; SUWABE, K.; SUZUKI, G.; AOKI, K.; KUBO, Y.; WATANABE, M.; … YANO, K. Plant Omics Data Center: An Integrated Web Repository for Interspecies Gene Expression Networks with NLP-Based Curation. **Plant and Cell Physiology**, vol. 56, no. 1, p. e9–e9, 1 Jan. 2015. DOI 10.1093/pcp/pcu188. Available at: https://doi.org/10.1093/pcp/pcu188.

OLIVER, M. J.; TUBA, Z.; MISHLER, B. D. The evolution of vegetative desiccation tolerance in land plants. **Plant Ecology**, vol. 151, no. 1, p. 85–100, 2000. https://doi.org/10.1023/A:1026550808557.

ONU. **World population prospects 2019**. [*S. l.*: *s. n.*], 2019. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12283219.

PANCHY, N.; LEHTI-SHIU, M.; SHIU, S. H. Evolution of gene duplication in plants. **Plant Physiology**, vol. 171, no. 4, p. 2294–2316, 2016. https://doi.org/10.1104/pp.16.00523.

PEDRUZZI, I.; RIVOIRE, C.; AUCHINCLOSS, A. H.; COUDERT, E.; KELLER, G.; DE CASTRO, E.; BARATIN, D.; CUCHE, B. A.; BOUGUELERET, L.; POUX, S.; REDASCHI, N.; XENARIOS, I.; BRIDGE, A. HAMAP in 2015: updates to the protein family classification and annotation system. **Nucleic Acids Research**, vol. 43, no. D1, p. D1064–D1070, 28 Jan. 2015. DOI 10.1093/nar/gku1002. Available at: https://doi.org/10.1093/nar/gku1002.

POTENZA, E.; DOMENICO, T. Di; WALSH, I.; TOSATTO, S. C. E. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. **Nucleic Acids Research**, vol. 43, no. D1, p. D315–D320, 28 Jan. 2015. DOI 10.1093/nar/gku982. Available at: https://doi.org/10.1093/nar/gku982.

PRADO, J. R.; SEGERS, G.; VOELKER, T.; CARSON, D.; DOBERT, R.; PHILLIPS, J.; COOK, K.; CORNEJO, C.; MONKEN, J.; GRAPES, L.; REYNOLDS, T.; MARTINO-CATT, S. Genetically Engineered Crops: From Idea to Product. **Annual Review of Plant Biology**, vol. 65, no. 1, p. 769–

790, 2014. https://doi.org/10.1146/annurev-arplant-050213-040039.

PROOST, S.; KRAWCZYK, A.; MUTWIL, M. LSTrAP: Efficiently combining RNA sequencing data into co-expression networks. **BMC Bioinformatics**, vol. 18, no. 1, p. 1–9, 2017. https://doi.org/10.1186/s12859-017-1861-z.

PROOST, S.; MUTWIL, M. CoNekT: An open-source framework for comparative genomic and transcriptomic network analyses. **Nucleic Acids Research**, vol. 46, no. W1, p. W133–W140, 2018. https://doi.org/10.1093/nar/gky336.

QASEEM, M. F.; QURESHI, R.; SHAHEEN, H. Effects of Pre-Anthesis Drought, Heat and Their Combination on the Growth, Yield and Physiology of diverse Wheat (Triticum aestivum L.) Genotypes Varying in Sensitivity to Heat and drought stress. **Scientific Reports**, vol. 9, no. 1, p. 1–12, 2019. DOI 10.1038/s41598-019-43477-z. Available at: http://dx.doi.org/10.1038/s41598-019-43477-z.

QUEVILLON, E.; SILVENTOINEN, V.; PILLAI, S.; HARTE, N.; MULDER, N.; APWEILER, R.; LOPEZ, R. InterProScan: protein domains identifier. **Nucleic Acids Research**, vol. 33, no. Web Server, p. W116–W120, 1 Jul. 2005. DOI 10.1093/nar/gki442. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki442.

RAO, X.; DIXON, R. A. Co-expression networks for plant biology: Why and how. **Acta Biochimica et Biophysica Sinica**, vol. 51, no. 10, p. 981–988, 2019. https://doi.org/10.1093/abbs/gmz080.

RUPRECHT, C.; VAID, N.; PROOST, S.; PERSSON, S.; MUTWIL, M. Beyond Genomics: Studying Evolution with Gene Coexpression Networks. **Trends in Plant Science**, vol. 22, no. 4, p. 298–307, 2017. DOI 10.1016/j.tplants.2016.12.011. Available at: http://dx.doi.org/10.1016/j.tplants.2016.12.011.

SAH, S. K.; REDDY, K. R.; LI, J. Abscisic acid and abiotic stress tolerance in crop plants. **Frontiers in Plant Science**, vol. 7, no. MAY2016, p. 1–26, 2016. https://doi.org/10.3389/fpls.2016.00571.

SCHEBEN, A.; EDWARDS, D. Bottlenecks for genome-edited crops on the road from lab to farm. **Genome Biology**, vol. 19, no. 1, p. 178, 26 Dec. 2018. DOI 10.1186/s13059-018-1555-5. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1555-5.

SCHNABLE, P. S.; WARE, D.; FULTON, R. S.; STEIN, J. C.; WEI, F.; PASTERNAK, S.; LIANG, C.; ZHANG, J.; FULTON, L.; GRAVES, T. A.; MINX, P.; REILY, A. D.; COURTNEY, L.; KRUCHOWSKI, S. S.; TOMLINSON, C.; STRONG, C.; DELEHAUNTY, K.; FRONICK, C.; COURTNEY, B.; … WILSON, R. K. The B73 maize genome: Complexity, diversity, and dynamics. **Science**, vol. 326, no. 5956, p. 1112–1115, 2009. https://doi.org/10.1126/science.1178534.

SERIN, E. A. R.; NIJVEEN, H.; HILHORST, H. W. M.; LIGTERINK, W. Learning from Co-expression Networks: Possibilities and Challenges. **Frontiers in Plant Science**, vol. 7, 8 Apr. 2016.

DOI 10.3389/fpls.2016.00444. Available at: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00444/abstract.

SHAMEER, K.; NAIKA, M. B. N.; SHA, K. M.; SOWDHAMINI, R. Decoding systems biology of plant stress for sustainable agriculture development and optimized food production. vol. 145, 2019. https://doi.org/10.1016/j.pbiomolbio.2018.12.002.

SMOOT, M. E.; ONO, K.; RUSCHEINSKI, J.; WANG, P.-L.; IDEKER, T. Cytoscape 2.8: new features for data integration and network visualization. **Bioinformatics**, vol. 27, no. 3, p. 431–432, 1 Feb. 2011. DOI 10.1093/bioinformatics/btq675. Available at: https://doi.org/10.1093/bioinformatics/btq675.

SONG, L.; LANGFELDER, P.; HORVATH, S. Comparison of co-expression measures: Mutual information, correlation, and model based indices. **BMC Bioinformatics**, vol. 13, no. 1, 2012. https://doi.org/10.1186/1471-2105-13-328.

STAMBOULIAN, M.; GUERRERO, R. F.; HAHN, M. W.; RADIVOJAC, P. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. **Bioinformatics (Oxford, England)**, vol. 36, no. 1, p. i219–i226, 2020. https://doi.org/10.1093/bioinformatics/btaa468.

STEUER, R.; KURTHS, J.; DAUB, C. O.; WEISE, J.; SELBIG, J. The mutual information: Detecting and evaluating dependencies between variables. **Bioinformatics**, vol. 18, no. SUPPL. 2, p. 231–240, 2002. https://doi.org/10.1093/bioinformatics/18.suppl_2.S231.

SUN, X.; MA, Y.; YANG, C.; LI, J. Rice OVATE family protein 6 regulates leaf angle by modulating secondary cell wall biosynthesis. **Plant Molecular Biology**, vol. 104, no. 3, p. 249–261, 2020. DOI 10.1007/s11103-020-01039-2. Available at: https://doi.org/10.1007/s11103-020-01039-2.

TAIZ, L.; ZEIGER, E.; MOLLER, I. max; MURPHY, A. **Fisiologia e desenvolvimento vegetal Diversidade vegetal**. [*S. l.*: *s. n.*], 2017. vol. 6 ed., .

TRACHANA, K.; LARSSON, T. A.; POWELL, S.; CHEN, W. H.; DOERKS, T.; MULLER, J.; BORK, P. Orthology prediction methods: A quality assessment using curated protein families. **BioEssays**, vol. 33, no. 10, p. 769–780, 2011. https://doi.org/10.1002/bies.201100062.

TZFADIA, O.; DIELS, T.; DE MEYER, S.; VANDEPOELE, K.; AHARONI, A.; VAN DE PEER, Y. CoExpNetViz: Comparative co-expression networks construction and visualization tool. **Frontiers in Plant Science**, vol. 6, no. JAN2016, p. 1–7, 2016. https://doi.org/10.3389/fpls.2015.01194.

USADEL, B.; OBAYASHI, T.; MUTWIL, M.; GIORGI, F. M.; BASSEL, G. W.; TANIMOTO, M.; CHOW, A.; STEINHAUSER, D.; PERSSON, S.; PROVART, N. J. Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats. **Plant, Cell and Environment**, vol.

32, no. 12, p. 1633–1651, 2009. https://doi.org/10.1111/j.1365-3040.2009.02040.x.

VADEZ, V.; HASH, T.; BIDINGER, F. R.; KHOLOVA, J. II.1.5 Phenotyping pearl millet for adaptation to drought. **Frontiers in Physiology**, vol. 3 OCT, no. October, p. 1–12, 2012. https://doi.org/10.3389/fphys.2012.00386.

VAN DONGEN, S. **Graph stimulation by flow clustering**. 2000. University of Utrecht f. 2000. Available at: https://micans.org/mcl/%0Ahttp://www.mendeley.com/research/mcl-a-cluster-algorithm-for-graphs/.

VANBUREN, R.; BRYANT, D.; EDGER, P. P.; TANG, H.; BURGESS, D.; CHALLABATHULA, D.; SPITTLE, K.; HALL, R.; GU, J.; LYONS, E.; FREELING, M.; BARTELS, D.; TEN HALLERS, B.; HASTIE, A.; MICHAEL, T. P.; MOCKLER, T. C. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. **Nature**, vol. 527, no. 7579, p. 508–511, 2015. DOI 10.1038/nature15714. Available at: http://dx.doi.org/10.1038/nature15714.

VANBUREN, R.; WAI, C. M.; KEILWAGEN, J.; PARDO, J. A chromosome-scale assembly of the model desiccation tolerant grass Oropetium thomaeum. **Plant Direct**, vol. 2, no. 11, p. 1–9, 2018. https://doi.org/10.1002/pld3.96.

VANBUREN, R.; WAI, C. M.; ZHANG, Q.; SONG, X.; EDGER, P. P.; BRYANT, D.; MICHAEL, T. P.; MOCKLER, T. C.; BARTELS, D. Seed desiccation mechanisms co-opted for vegetative desiccation in the resurrection grass Oropetium thomaeum. **Plant Cell and Environment**, vol. 40, no. 10, p. 2292–2306, 2017. https://doi.org/10.1111/pce.13027.

VARSHNEY, R. K.; SHI, C.; THUDI, M.; MARIAC, C.; WALLACE, J.; QI, P.; ZHANG, H.; ZHAO, Y.; WANG, X.; RATHORE, A.; SRIVASTAVA, R. K.; CHITIKINENI, A.; FAN, G.; BAJAJ, P.; PUNNURI, S.; GUPTA, S. K.; WANG, H.; JIANG, Y.; COUDERC, M.; … XU, X. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. **Nature Biotechnology**, vol. 35, no. 10, p. 969–976, 2017. DOI 10.1038/nbt.3943. Available at: http://dx.doi.org/10.1038/nbt.3943.

VIANA, M. J. A.; ZERLOTINI, A.; DE ALVARENGA MUDADU, M. Plant Co-expression Annotation Resource: a webserver for identifying targets for genetically modified crop breeding pipelines. 2020. https://doi.org/10.1101/2020.05.22.110510.

VIANA, M. J. A.; ZERLOTINI, A.; DE ALVARENGA MUDADU, M. Plant Co-expression Annotation Resource: a webserver for identifying targets for genetically modified crop breeding pipelines. **BMC Bioinformatics**, , p. 1–14, 2021. DOI 10.1101/2020.05.22.110510. Available at: https://doi.org/10.1186/s12859-020-03792-z.

VIANA, M.; ZERLOTINI, A.; MUDADU, M. Protocol A - Plantannot. [*s. d.*]. DOI 10.17504/protocols.io.bgcvjsw6. Available at: https://dx.doi.org/10.17504/protocols.io.bgcvjsw6.

VIANA, M.; ZERLOTINI, A.; MUDADU, M. Protocol B - Plantannot. [*s. d.*]. DOI 10.17504/protocols.io.bgdgjs3w. Available at: https://dx.doi.org/10.17504/protocols.io.bgdgjs3w.

VIANA, M.; ZERLOTINI, A.; MUDADU, M. Protocol C - Plantannot. [*s. d.*]. DOI 10.17504/protocols.io.bgdijs4e. Available at: https://dx.doi.org/10.17504/protocols.io.bgdijs4e.

VIANA, M.; ZERLOTINI, A.; MUDADU, M. Protocol D - Plantannot. [*s. d.*]. DOI 10.17504/protocols.io.bgd6js9e. Available at: https://dx.doi.org/10.17504/protocols.io.bgd6js9e.

VIANA, M.; ZERLOTINI, A.; MUDADU, M. Protocol E - Plantannot. [*s. d.*]. DOI 10.17504/protocols.io.bgdjjs4n. Available at: https://dx.doi.org/10.17504/protocols.io.bgdjjs4n.

VIANA, M.; ZERLOTINI, A.; MUDADU, M. Protocol F - Plantannot. [*s. d.*]. DOI 10.17504/protocols.io.bgdkjs4w. Available at: https://dx.doi.org/10.17504/protocols.io.bgdkjs4w.

WALT, S. van der; COLBERT, S. C.; VAROQUAUX, G. The NumPy Array: A Structure for Efficient Numerical Computation. **Computing in Science & Engineering**, vol. 13, no. 2, p. 22–30, 2011. https://doi.org/10.1109/MCSE.2011.37.

WANG, Y.; COLEMAN-DERR, D.; CHEN, G.; GU, Y. Q. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. **Nucleic Acids Research**, vol. 43, no. W1, p. W78–W84, 2015. https://doi.org/10.1093/nar/gkv487.

WOŹNIAK, E.; WASZKOWSKA, E.; ZIMNY, T.; SOWA, S.; TWARDOWSKI, T. The Rapeseed Potential in Poland and Germany in the Context of Production, Legislation, and Intellectual Property Rights. **Frontiers in Plant Science**, vol. 10, 5 Nov. 2019. DOI 10.3389/fpls.2019.01423. Available at: https://www.frontiersin.org/article/10.3389/fpls.2019.01423/full.

WU, C. H.; NIKOLSKAYA, A.; HUANG, H.; YEH, L. L.; NATALE, D. A.; VINAYAKA, C. R.; HU, Z.; MAZUMDER, R.; KUMAR, S.; KOURTESIS, P.; LEDLEY, R. S.; SUZEK, B. E.; ARMINSKI, L.; CHEN, Y.; ZHANG, J.; CARDENAS, J. L.; CHUNG, S.; CASTRO-ALVEAR, J.; DINKOV, G.; BARKER, W. C. PIRSF: family classification system at the Protein Information Resource. **Nucleic Acids Research**, vol. 32, no. suppl_1, p. D112–D114, 1 Jan. 2004. DOI 10.1093/nar/gkh097. Available at: https://doi.org/10.1093/nar/gkh097.

WU, H.; YAO, D.; CHEN, Y.; YANG, W.; ZHAO, W.; GAO, H.; TONG, C. De novo genome assembly of populus simonii further supports that populus simonii and populus trichocarpa belong to different sections. **G3: Genes, Genomes, Genetics**, vol. 10, no. 2, p. 455–466, 2020. https://doi.org/10.1534/g3.119.400913.

XIAO, L.; YANG, G.; ZHANG, L.; YANG, X.; ZHAO, S.; JI, Z.; ZHOU, Q.; HU, M.; WANG, Y.; CHEN, M.; XU, Y.; JIN, H.; XIAO, X.; HU, G.; BAO, F.; HU, Y.; WAN, P.; LI, L.; DENG, X.; … HE, Y. The resurrection genome of Boea hygrometrica: A blueprint for survival of dehydration. **Proceedings of the National Academy of Sciences of the United States of America**, vol. 112, no.

18, p. 5833–5837, 2015. https://doi.org/10.1073/pnas.1505811112.

XU, L.; DONG, Z.; FANG, L.; LUO, Y.; WEI, Z.; GUO, H.; ZHANG, G.; GU, Y. Q.; COLEMAN-DERR, D.; XIA, Q.; WANG, Y. OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. **Nucleic Acids Research**, vol. 47, no. W1, p. W52–W58, 2019. https://doi.org/10.1093/nar/gkz333.

YAMAGUCHI-SHINOZAKI, K.; SHINOZAKI, K. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. **Annual Review of Plant Biology**, vol. 57, p. 781–803, 2006. https://doi.org/10.1146/annurev.arplant.57.032905.105444.

YOSHIDA, T.; FUJITA, Y.; SAYAMA, H.; KIDOKORO, S.; MARUYAMA, K.; MIZOI, J.; SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. **Plant Journal**, vol. 61, no. 4, p. 672–685, 2010. https://doi.org/10.1111/j.1365-313X.2009.04092.x.

YOU, J.; CHAN, Z. Ros regulation during abiotic stress responses in crop plants. **Frontiers in Plant Science**, vol. 6, no. DEC, p. 1–15, 2015. https://doi.org/10.3389/fpls.2015.01092.

YU, H.; JIAO, B.; LU, L.; WANG, P.; CHEN, S.; LIANG, C.; LIU, W. NetMiner-an ensemble pipeline for building genome-wide and high-quality gene coexpression network using massive-scale RNA-seq samples. **PLoS ONE**, vol. 13, no. 2, p. 1–23, 2018. https://doi.org/10.1371/journal.pone.0192613.

ZHU, J. K. Abiotic Stress Signaling and Responses in Plants. **Cell**, vol. 167, no. 2, p. 313–324, 2016. DOI 10.1016/j.cell.2016.08.029. Available at: http://dx.doi.org/10.1016/j.cell.2016.08.029.

ZHU, J.; TIAN, J.; WANG, J.; NIE, S. Variation of traits on seeds and germination derived from the hybridization between the sections Tacamahaca and Aigeiros of the genus populus. **Forests**, vol. 9, no. 9, 2018. https://doi.org/10.3390/f9090516.

APENDICE A – TABELAS SUPLEMENTARES

Tabela suplementar 1: Dados de RNA-seq utilizados para montagem dos clusters de coexpressão gênica.

Organismo	GEO series	GEO samples	SRA	Condição	Tecido	Data
Arabidopsis thaliana	GSE134391	GSM3945669	SRR9696660	Heat stress rep1	Leaf	Nov-21-2019
Arabidopsis thaliana	GSE134391	GSM3945673	SRR9696664	Heat stress rep2	Leaf	Nov-21-2019
Arabidopsis thaliana	GSE134391	GSM3945677	SRR9696668	Heat stress rep3	Leaf	Nov-21-2019
Arabidopsis thaliana	GSE134391	GSM3945668	SRR9696659	Highlight stress rep1	Leaf	Nov-21-2019
Arabidopsis thaliana	GSE134391	GSM3945672	SRR9696663	Highlight stress rep2	Leaf	Nov-21-2019
Arabidopsis thaliana	GSE134391	GSM3945676	SRR9696667	High light stress rep1	Leaf	Nov-21-2019
Arabidopsis thaliana	GSE127805	GSM3639178	SRR8666075	low water WT stress replicate 1	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639178	SRR8666076	low water WT stress replicate 1	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639178	SRR8666077	low water WT stress replicate 1	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639182	SRR8666087	low water WT stress replicate 2	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639182	SRR8666088	low water WT stress replicate 2	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639182	SRR8666089	low water WT stress replicate 2	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639186	SRR8666099	low water WT stress replicate 3	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639186	SRR8666100	low water WT stress replicate 3	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE127805	GSM3639186	SRR8666101	low water WT stress replicate 3	Seedling	Oct-01-2019
Arabidopsis thaliana	GSE85292	GSM2264185	SRR4011032	Aluminium stress Col_4h	Root	Aug-08-2019
Arabidopsis thaliana	GSE116069	GSM3208265	SRR7405096	200 mM NaCl for 16 hrs Col_0_NaCl1	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208265	SRR7405097	200 mM NaCl for 16 hrs Col_0_NaCl1	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208265	SRR7405098	200 mM NaCl for 16 hrs Col_0_NaCl1	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208265	SRR7405099	200 mM NaCl for 16 hrs Col_0_NaCl1	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208266	SRR7405100	200 mM NaCl for 16 hrs Col_0_NaCl2	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208266	SRR7405101	200 mM NaCl for 16 hrs Col_0_NaCl2	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208266	SRR7405102	200 mM NaCl for 16 hrs Col_0_NaCl2	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208266	SRR7405103	200 mM NaCl for 16 hrs Col_0_NaCl2	Seedling	Jun-16-2019

Arabidopsis thaliana	GSE116069	GSM3208267	SRR7405104	200 mM NaCl for 16 hrs Col_0_NaCl3	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208267	SRR7405105	200 mM NaCl for 16 hrs Col_0_NaCl3	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208267	SRR7405106	200 mM NaCl for 16 hrs Col_0_NaCl3	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE116069	GSM3208267	SRR7405107	200 mM NaCl for 16 hrs Col_0_NaCl3	Seedling	Jun-16-2019
Arabidopsis thaliana	GSE130729	GSM3752562	SRR9016746	cold stress 4oC 0h rep 1	Seedling	May-07-2019
Arabidopsis thaliana	GSE130729	GSM3752563	SRR9016747	cold stress 4oC 0h rep 2	Seedling	May-07-2019
Arabidopsis thaliana	GSE130729	GSM3752564	SRR9016748	cold stress 4oC 3h rep 1	Seedling	May-07-2019
Arabidopsis thaliana	GSE130729	GSM3752566	SRR9016750	cold stress 4oC 24h rep 1	Seedling	May-07-2019
Arabidopsis thaliana	GSE130729	GSM3752567	SRR9016751	cold stress 4oC 24h rep 2	Seedling	May-07-2019
Arabidopsis thaliana	GSE116004	GSM3204821	SRR7367104	cold stress 10oC 1h ZT1 rep2	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204822	SRR7367105	cold stress 10oC 1h ZT1 rep3	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204823	SRR7367106	cold stress 10oC 1h ZT1 rep4	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204836	SRR7367119	heat stress 37oC 1h ZT1 rep1	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204837	SRR7367120	heat stress 37oC 1h ZT1 rep2	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204838	SRR7367121	heat stress 37oC 1h ZT1 rep3	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204839	SRR7367122	heat stress 37oC 1h ZT1 rep4	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204844	SRR7367127	cold stress 10oC 1h ZT6 rep1	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204846	SRR7367129	cold stress 10oC 1h ZT6 rep2	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204850	SRR7367133	heat stress 37oC 1h ZT6 rep1	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE116004	GSM3204853	SRR7367136	heat stress 37oC 1h ZT6 rep2	Seedling	Mar-18-2019
Arabidopsis thaliana	GSE127819	GSM3639336	SRR8667484	cold stress 4oC 0h rep1	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639337	SRR8667485	cold stress 4oC 0h rep2	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639338	SRR8667486	cold stress 4oC 0h rep3	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639339	SRR8667487	cold stress 4oC 3h rep1	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639340	SRR8667488	cold stress 4oC 3h rep2	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639341	SRR8667489	cold stress 4oC 3h rep3	Seedling	Mar-06-2019

Arabidopsis thaliana	GSE127819	GSM3639342	SRR8667490	cold stress 4oC 24h rep1	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639342	SRR8667491	cold stress 4oC 24h rep1	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639343	SRR8667492	cold stress 4oC 24h rep2	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE127819	GSM3639344	SRR8667493	cold stress 4oC 24h rep3	Seedling	Mar-06-2019
Arabidopsis thaliana	GSE119330	GSM3370069	SRR7774144	heat stress 30oC 1h AM ZT 0.5 rep1	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370070	SRR7774145	heat stress 30oC 1h AM ZT 0.5 rep2	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370071	SRR7774146	heat stress 30oC 1h AM ZT 0.5 rep3	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370072	SRR7774147	heat stress 30oC 1h AM ZT 0.5 rep4	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370073	SRR7774148	heat stress 30oC 1h PM ZT 12.5 rep1	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370074	SRR7774149	heat stress 30oC 1h PM ZT 12.5 rep2	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370075	SRR7774150	heat stress 30oC 1h PM ZT 12.5 rep3	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370076	SRR7774151	heat stress 30oC 1h PM ZT 12.5 rep4	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370079	SRR7774154	heat stress 30oC 1h AM ZT 0.5 and constant light	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE119330	GSM3370080	SRR7774155	heat stress 30oC 1h PM ZT 12.5 and constant light	Seedling	Feb-01-2019
Arabidopsis thaliana	GSE116964	GSM3265402	SRR7511617	cold stress 4oC 24h	Leaf	Jan-31-2019
Arabidopsis thaliana	GSE108610	GSM2906494	SRR6427370	drought stress without water for 5 days rep1	Whole plant	Dec-01-2018
Arabidopsis thaliana	GSE108610	GSM2906495	SRR6427371	drought stress without water for 5 days rep2	Whole plant	Dec-01-2018
Arabidopsis thaliana	GSE119382	GSM3373992	SRR7779219	drought stress without water for 5 days rep1	Root	Sep-04-2018
Arabidopsis thaliana	GSE119382	GSM3373993	SRR7779220	drought stress without water for 5 days rep2	Root	Sep-04-2018
Arabidopsis thaliana	GSE119382	GSM3373994	SRR7779221	drought stress without water for 5 days rep3	Root	Sep-04-2018
Arabidopsis thaliana	GSE118298	GSM3324327	SRR7659148	heat stress	Leaf	Aug-08-2018
Arabidopsis thaliana	GSE118298	GSM3324328	SRR7659149	heat stress	Leaf	Aug-08-2018
Arabidopsis thaliana	GSE118298	GSM3324329	SRR7659150	heat stress	Leaf	Aug-08-2018
Arabidopsis thaliana	GSE112368	GSM3068810	SRR6902930	low water WT stress replicate 1	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068810	SRR6902931	low water WT stress replicate 1	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068810	SRR6902932	low water WT stress replicate 1	Seedling	Jul-20-2018

Arabidopsis thaliana	GSE112368	GSM3068814	SRR6902944	low water WT stress replicate 2	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068814	SRR6902945	low water WT stress replicate 2	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068814	SRR6902946	low water WT stress replicate 2	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068818	SRR6902956	low water WT stress replicate 3	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068818	SRR6902957	low water WT stress replicate 3	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE112368	GSM3068818	SRR6902958	low water WT stress replicate 3	Seedling	Jul-20-2018
Arabidopsis thaliana	GSE85653	GSM2280286	SRR4033018	Heat stress rep1	Leaf	May-30-2018
Arabidopsis thaliana	GSE85653	GSM2280287	SRR4033019	Heat stress rep2	Leaf	May-30-2018
Arabidopsis thaliana	GSE85653	GSM2280288	SRR4033020	Heat stress rep3	Leaf	May-30-2018
Arabidopsis thaliana	GSE114379	GSM3140728	SRR7160928	osmotic stress r1	Seedling	May-14-2018
Arabidopsis thaliana	GSE114379	GSM3140729	SRR7160929	osmotic stress r2	Seedling	May-14-2018
Arabidopsis thaliana	GSE114379	GSM3140730	SRR7160930	osmotic stress r3	Seedling	May-14-2018
Arabidopsis thaliana	GSE103964	GSM2786977	SRR6048928	cold stress 4oC	Whole plant	Apr-18-2018
Arabidopsis thaliana	GSE103964	GSM2786977	SRR6048929	cold stress 4oC	Whole plant	Apr-18-2018
Arabidopsis thaliana	GSE107820	GSM2881006	SRR6359065	heat stress 43oC	Root	Mar-21-2018
Arabidopsis thaliana	GSE107820	GSM2881007	SRR6359066	heat stress 43oC	Root	Mar-21-2018
Arabidopsis thaliana	GSE103041	GSM2752982	SRR5968352	WT heat stress 44oC	Seedling	Feb-13-2018
Arabidopsis thaliana	GSE93979	GSM2466002	SRR5196729	WT drought rep1	Leaf	Jun-13-2017
Arabidopsis thaliana	GSE93979	GSM2466003	SRR5196730	WT drought rep1	Leaf	Jun-13-2017
Arabidopsis thaliana	GSE93420	GSM2453038	SRR5167847	WT dehydration1	Leaf	Apr-11-2017
Arabidopsis thaliana	GSE93420	GSM2453039	SRR5167848	WT dehydration2	Leaf	Apr-11-2017
Arabidopsis thaliana	GSE93420	GSM2453040	SRR5167849	WT dehydration3	Leaf	Apr-11-2017
Arabidopsis thaliana	GSE74864	GSM1936771	SRR2932377	dehydration 3h rep1	Seedling	Apr-05-2017
Arabidopsis thaliana	GSE74864	GSM1936772	SRR2932378	dehydration 3h rep1	Seedling	Apr-05-2017
Arabidopsis thaliana	GSE74864	GSM1936773	SRR2932379	dehydration 3h rep3	Seedling	Apr-05-2017
Arabidopsis thaliana	GSE94015	GSM2467113	SRR5197907	WT RL3h rep1 heat stress 37C 3h	Leaf	Mar-15-2017

Arabidopsis thaliana	GSE94015	GSM2467114	SRR5197908	WT RL3h rep2 heat stress 37C 3h	Leaf	Mar-15-2017
Arabidopsis thaliana	GSE94015	GSM2467115	SRR5197909	WT RL3h rep3 heat stress 37C 3h	Leaf	Mar-15-2017
Arabidopsis thaliana	GSE72806	GSM1872392	SRR2302914	Col h1R heat stress 44oC 1h	Leaf	Oct-24-2016
Arabidopsis thaliana	GSE72806	GSM1872393	SRR2302915	Col h2R heat stress 44oC 1h	Leaf	Oct-24-2016
Arabidopsis thaliana	GSE72806	GSM1872394	SRR2302916	Col h3R heat stress 44oC 1h	Leaf	Oct-24-2016
Arabidopsis thaliana	GSE72806	GSM1872389	SRR2302911	Col s1R salinity stress	Leaf	Oct-24-2016
Arabidopsis thaliana	GSE72806	GSM1872390	SRR2302912	Col s2R salinity stress	Leaf	Oct-24-2016
Arabidopsis thaliana	GSE72806	GSM1872391	SRR2302913	Col s3R salinity stress	Leaf	Oct-24-2016
Oryza sativa	GSE132183	GSM3852644	SRR9201531	Salt stress 100 mN NaCl rep1	Root	Jun-05-2019
Oryza sativa	GSE132183	GSM3852645	SRR9201532	Salt stress 100 mN NaCl rep2	Root	Jun-05-2019
Oryza sativa	GSE132183	GSM3852646	SRR9201533	Salt stress 100 mN NaCl rep3	Root	Jun-05-2019
Oryza sativa	GSE132183	GSM3852656	SRR9201543	Salt stress 100 mN NaCl rep1	Shoot	Jun-05-2019
Oryza sativa	GSE132183	GSM3852657	SRR9201544	Salt stress 100 mN NaCl rep2	Shoot	Jun-05-2019
Oryza sativa	GSE132183	GSM3852658	SRR9201545	Salt stress 100 mN NaCl rep3	Shoot	Jun-05-2019
Oryza sativa	GSE100713	GSM2691964	SRR5796952	Heat stress 42oC 10min then 30oC 10min rep1	Shoot	Nov-12-2018
Oryza sativa	GSE100713	GSM2691966	SRR5796954	Heat stress 42oC 10min then 30oC 50min rep1	Shoot	Nov-12-2018
Oryza sativa	GSE100713	GSM2691967	SRR5796955	Heat stress 42oC 10min then 30oC 50min rep2	Shoot	Nov-12-2018
Oryza sativa	GSE100713	GSM2691968	SRR5796956	Heat stress 42oC 10min then 30oC 1h50min rep1	Shoot	Nov-12-2018
Oryza sativa	GSE100713	GSM2691969	SRR5796957	Heat stress 42oC 10min then 30oC 1h50min rep2	Shoot	Nov-12-2018
Oryza sativa	GSE100713	GSM2691970	SRR5796958	Heat stress 42oC 10min then 30oC 9h50min rep1	Shoot	Nov-12-2018
Oryza sativa	GSE100713	GSM2691971	SRR5796959	Heat stress 42oC 10min then 30oC 9h50min rep2	Shoot	Nov-12-2018
Oryza sativa	GSE86215	GSM2298194	SRR4098183	Drought stress air dried	Seedling	Jun-13-2018
Oryza sativa	GSE86215	GSM2298197	SRR4098186	Heat stress 50oC	Seedling	Jun-13-2018
Oryza sativa	GSE107425	GSM2866848	SRR6326702	Drought stress	Shoot	Feb-28-2018
Oryza sativa	GSE107425	GSM2866849	SRR6326703	Drought stress	Shoot	Feb-28-2018
Oryza sativa	GSE101734	GSM2714235	SRR5856930	Salt stress	Leaf	Jul-22-2017

Oryza sativa	GSE101734	GSM2714236	SRR5856931	Salt stress	Leaf	Jul-22-2017
Oryza sativa	GSE101734	GSM2714237	SRR5856932	Salt stress	Leaf	Jul-22-2017
Oryza sativa	GSE77510	GSM2053502	SRR3140959	Heat stress 45oC 12h	Leaf	Dec-21-2017
Oryza sativa	GSE77510	GSM2053503	SRR3140960	Heat stress 45oC 24h	Leaf	Dec-21-2017
Oryza sativa	GSE92989	GSM2441712	SRR5134065	Drougth stress 2 days	Root	Jun-12-2017
Oryza sativa	GSE92989	GSM2441713	SRR5134066	Drougth stress 2 days	Root	Jun-12-2017
Oryza sativa	GSE92989	GSM2441714	SRR5134067	Drougth stress 3 days	Root	Jun-12-2017
Oryza sativa	GSE92989	GSM2441715	SRR5134068	Drougth stress 3 days	Root	Jun-12-2017
Oryza sativa	GSE78972	GSM2082859	SRR3209771	Long Day Drought stress S3	Leaf	Mar-01-2017
Oryza sativa	GSE78972	GSM2082860	SRR3209772	Long Day Drought stress S4	Leaf	Mar-01-2017
Oryza sativa	GSE78972	GSM2082863	SRR3209775	Short Day Drought stress S7	Leaf	Mar-01-2017
Oryza sativa	GSE78972	GSM2082864	SRR3209776	Short Day Drought stress S8	Leaf	Mar-01-2017
Oryza sativa	GSE78972	GSM2082866	SRR3209778	Long Day Drought stress S10	Leaf	Mar-01-2017
Oryza sativa	GSE78972	GSM2082868	SRR3209780	Short Day Drought stress S12	Leaf	Mar-01-2017
Oryza sativa	GSE80811	GSM2137964	SRR3466960	drought stress 1d	Leaf	Feb-14-2017
Oryza sativa	GSE80811	GSM2137964	SRR3466961	drought stress 1d	Leaf	Feb-14-2017
Oryza sativa	GSE80811	GSM2137965	SRR3466962	drought stress 2d	Leaf	Feb-14-2017
Oryza sativa	GSE80811	GSM2137965	SRR3466963	drought stress 2d	Leaf	Feb-14-2017
Oryza sativa	GSE80811	GSM2137966	SRR3466964	drought stress 3d	Leaf	Feb-14-2017
Oryza sativa	GSE80811	GSM2137966	SRR3466965	drought stress 3d	Leaf	Feb-14-2017
Oryza sativa	GSE95668	GSM2520922	SRR5311340	heat stress 35oC 6h	Leaf	Nov-07-2017
Oryza sativa	GSE95668	GSM2520923	SRR5311341	heat stress 35oC 6h	Leaf	Nov-07-2017
Zea mays	GSE137780	GSM4087772	SRR10153120	Drought stress rep1	Leaf	Sep-21-2019
Zea mays	GSE137780	GSM4087773	SRR10153121	Drought stress rep2	Leaf	Sep-21-2019
Zea mays	GSE71723	GSM1843772	SRR2144414	drought stress	Leaf	Feb-04-2016
Zea mays	GSE71723	GSM1843780	SRR2144422	drought stress	Leaf	Feb-04-2016

Zea mays	GSE71723	GSM1843788	SRR2144430	drought stress	Leaf	Feb-04-2016
Zea mays	GSE71723	GSM1843796	SRR2144438	drought stress	Leaf	Feb-04-2016
Zea mays	GSE71377	GSM1833214	SRR2129983	drought stress	Leaf	Jan-22-2016
Zea mays	GSE71046	GSM1826061	SRR2106186	wt Salt T7 Rep1	Leaf	Jan-14-2016
Zea mays	GSE71046	GSM1826073	SRR2106198	wt Salt T0 Rep2Rep3	Leaf	Jan-14-2016
Zea mays	GSE71046	GSM1826077	SRR2106202	wt Salt T7 Rep2Rep3	Leaf	Jan-14-2016
Zea mays	GSE76939	GSM2041248	SRR3105596	cold stress 4oC	Seedling	Jan-19-2016
Zea mays	GSE76939	GSM2041249	SRR3105597	cold stress 4oC	Seedling	Jan-19-2016
Glycine max	GSE93322	GSM2451231	SRR5163165	salt stress 150 mM NaCl 6h rep1	RootAn-dLeaf	Aug-01-2019
Glycine max	GSE93322	GSM2451232	SRR5163166	salt stress 150 mM NaCl 6h rep2	RootAn-dLeaf	Aug-01-2019
Glycine max	GSE117686	GSM3307258	SRR7601341	cold stress 4oC 24h	Leaf	Jul-27-2018
Glycine max	GSE117686	GSM3307259	SRR7601342	cold stress 4oC 24h	Leaf	Jul-27-2018
Glycine max	GSE117686	GSM3307260	SRR7601343	cold stress 4oC 24h	Leaf	Jul-27-2018
Glycine max	GSE98958	GSM2628302	SRR5569810	dehydrated	Leaf	May-31-2018
Glycine max	GSE98958	GSM2628302	SRR5569811	dehydrated	Leaf	May-31-2018
Glycine max	GSE98958	GSM2628303	SRR5569812	dehydrated	Leaf	May-31-2018
Glycine max	GSE98958	GSM2628303	SRR5569813	dehydrated	Leaf	May-31-2018
Glycine max	GSE69571	GSM1704043	SRR2051086	salt stress	Leaf	Jul-11-2017
Glycine max	GSE69571	GSM1704044	SRR2051087	salt stress	Leaf	Jul-11-2017
Glycine max	GSE69571	GSM1704045	SRR2051088	salt stress	Leaf	Jul-11-2017
Glycine max	GSE69571	GSM1704046	SRR2051089	salt stress	Leaf	Jul-11-2017
Glycine max	GSE70310	GSM1723542	SRR2079645	drought 15 days	Leaf	Aug-31-2015
Glycine max	GSE70310	GSM1723542	SRR2079646	drought 15 days	Leaf	Aug-31-2015
Glycine max	GSE70310	GSM1723542	SRR2079647	drought 15 days	Leaf	Aug-31-2015
Glycine max	GSE69469	GSM1701586	SRR2048167	drought 3days ZT0 8h R1	Leaf	Jul-07-2015
Glycine max	GSE69469	GSM1701592	SRR2048173	drought 3days ZT4 12h R1	Leaf	Jul-07-2015

Glycine max	GSE69469	GSM1701598	SRR2048179	drought 3days ZT8 16h R1	Leaf	Jul-07-2015
Glycine max	GSE69469	GSM1701604	SRR2048185	drought 3days ZT12 20h R1	Leaf	Jul-07-2015
Glycine max	GSE69469	GSM1701610	SRR2048191	drought 3days ZT16 24h R1	Leaf	Jul-07-2015
Glycine max	GSE69469	GSM1701616	SRR2048197	drought 3days ZT20 4h R1	Leaf	Jul-07-2015