UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA

FELIPE BITENCOURT MARTINS

IDENTIFICAÇÃO DE CONTAMINANTES E ANÁLISES MULTI ÔMICAS APLICADAS NO MELHORAMENTO MOLECULAR DE GRAMÍNEAS FORRAGEIRAS TROPICAIS POLIPLOIDES

CONTAMINANT IDENTIFICATION AND MULTI-OMIC ANALYSIS APPLIED IN POLYPLOID TROPICAL FORAGE GRASSES MOLECULAR BREEDING

CAMPINAS
2022

FELIPE BITENCOURT MARTINS

IDENTIFICAÇÃO DE CONTAMINANTES E ANÁLISES MULTI ÔMICAS APLICADAS NO MELHORAMENTO MOLECULAR DE FORRAGEIRAS TROPICAIS POLIPLOIDES

CONTAMINANT IDENTIFICATION AND MULTI-OMIC ANALYSIS APPLIED IN POLIPLOID TROPICAL FORAGES MOLECULAR BREEDING

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do Título de Doutor em Genética e Biologia Molecular, na área de Bioinformática

Thesis presented to the Biology Institute of the University of Campinas in a partial fulfillment of the requirements for the degree of Doctor in Genetics and Molecular Biology, in the area of Bioinformatics

Orientadora: Prof.ª Anete Pereira de Souza

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO FELIPE BITENCOURT MARTINS E ORIENTADA PELA PROFESSORA ANETE PEREIRA DE SOUZA.

CAMPINAS

2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

M366i

Martins, Felipe Bitencourt, 1993-
Identificação de contaminantes e análises multiômicas aplicadas no melhoramento molecular de gramíneas forrageiras tropicais poliploides / Felipe Bitencourt Martins. – Campinas, SP : [s.n.], 2022.

Orientador: Anete Pereira de Souza.
Coorientador: Rosangela Maria Simeão.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Plantas - Melhoramento genético. 2. Gramínea. 3. Bioinformática. 4. Aprendizado de máquina. I. Souza, Anete Pereira de, 1962-. II. Simeão, Rosangela Maria. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações Complementares

Título em outro idioma: Contaminant identification and multiomic analysis applied in polyploid tropical forage grasses molecular breeding
Palavras-chave em inglês:
Plant breeding
Grasses
Bioinformatics
Machine learning
Área de concentração: Bioinformática
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Anete Pereira de Souza [Orientador]
Renato Vicentini dos Santos
Rafael Vasconcelos Ribeiro
Letícia Aparecida de Castro Lara
Lilian Padilha
Data de defesa: 06-12-2022
Programa de Pós-Graduação: Genética e Biologia Molecular

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-2509-2901
- Currículo Lattes do autor: http://lattes.cnpq.br/7868956459473429

Campinas, 6 de dezembro de 2022.

COMISSÃO EXAMINADORA

Prof.ª Anete Pereira de Souza

Prof. Renato Vicentini dos Santos

Prof. Rafael Vasconcelos Ribeiro

Dr.ª Lilian Padilha

Dr.ª Letícia Aparecida de Castro Lara

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia.

AGRADECIMENTOS

Agradeço primeiramente a orientadora professora Anete por ter me dado a oportunidade de ser seu aluno de iniciação científica e posteriormente confiado a mim o desenvolvimento de um projeto de pesquisa de doutorado direto em seu laboratório. Agradeço pelo apoio, pelo otimismo, pelo exemplo, pela preocupação e pela compreensão que fizeram com que esse longo período de pós-graduação fosse mais agradável e proveitoso.

Agradeço a Dra. Rosangela, minha coorientadora, que me auxiliou no entendimento e execução do projeto. Agradeço aos funcionários e técnicos do LAGM que sempre garantiram uma boa estrutura para a realização da pesquisa. Agradeço a todos os colegas de laboratório, que sempre foram muito solícitos a dividir seus conhecimentos e tornaram o ambiente de trabalho mais prazeroso. Agradeço a Aline, Alexandre e Rebecca que participaram ativamente do desenvolvimento desta tese, trabalhando juntos em diversos momentos. Especialmente a Aline que sempre esteve muito atenta a minha vida acadêmica me auxiliando em diversos aspectos e ao Alexandre que com seu conhecimento, empenho e dedicação me inspirou e auxiliou a desenvolver um trabalho que jamais faria sozinho. Agradeço ao professor Augusto da ESALQ, que se tornou uma grande inspiração para mim tanto como professor quanto como pesquisador, me mostrando a importância da estatística para a ciência.

Agradeço a minha família pelo auxílio financeiro que tornou possível a realização desta pós-graduação sem grandes preocupações. Agradeço aos meus amigos, tanto os que fiz antes da graduação e estão presentes em minha vida até hoje, quanto os que conheci na universidade e compartilharam a vida acadêmica comigo. Ter o apoio e companhia deles para roles, bares, festas e viagens foi essencial para a preservação da minha saúde mental.

Por fim, por acreditar em um mundo onde o conhecimento é democratizado, agradeço a todos que produzem e disponibilizam de forma gratuita materiais didáticos na internet, e que participam de fóruns de discussão compartilhando seus conhecimentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

Gramíneas forrageiras tropicais, especialmente espécies do gênero *Urochloa*, desempenham um importante papel na pecuária, sendo a principal fonte de alimento para os animais nas regiões tropicais e subtropicais. Apesar do melhoramento convencional de forrageiras ter liberado cultivares superiores nas décadas recentes, os ganhos genéticos foram baixos e o melhoramento molecular ainda não é usado devido aos escassos recursos genômicos e a alta complexidade genômica das espécies. Nesse cenário, essa tese tem como objetivo investigar métodos e técnicas de melhoramento molecular em forrageiras tropicais poliploides para produzir conhecimentos que auxiliem suas aplicações nos programas de melhoramento. Hibridização artificial e cruzamentos controlados em forrageiras são desafiadores devido a apomixia e a autogamia em certas espécies, e devido a outras situações que podem causar a inserção de contaminantes na população de interesse. Dessa forma, esta tese propõe uma metodologia multivariada semi-automatizada para a detecção e classificação de contaminantes em progênies poliploides biparentais, incluindo clones apomíticos, indivíduos de autofecundação, meios irmãos e contaminantes completos. A metodologia foi estabelecida utilizando dados de genotipagem por sequenciamento (GBS) processados para obtenção de marcadores SNP com informação de dosagem alélica, integrando análises de componentes principais (PCA), métricas genotípicas baseadas em segregação mendeliana e análises de clusterização. O método implementado no aplicativo polyCID permitiu, de forma rápida e simplificada, a identificação de todos os contaminantes em todas as progênies simuladas, e a detecção de contaminantes putativos em três progênies reais. Além disso, como a maioria das espécies forrageiras tropicais são apomíticas e tetraploides, *U. ruziziensis* tem uma importância ímpar por ser uma espécie diploide e sexual que pode ser tetraploidizada para ser usada em cruzamentos interespecíficos com espécies apomíticas. Assim, investigamos também a aplicabilidade de predição genômica de famílias (GWFP) em famílias de meios irmãos tetraploides de *U. ruziziensis* para predizer crescimento e produção de biomassa. Aprendizado de máquina e seleção de features foram utilizados para reduzir a em média 11 o número de marcadores necessários para a predição, e elevar a habilidade preditiva a aproximadamente 0.9 em todos os fenótipos. Ademais, para investigar a regulação das características agronômicas, a posição dos marcadores de maior importância para a predição foram considerados como QTL putativos, e em uma abordagem multi ômica, genes obtidos no transcriptoma da espécie fisicamente ligados aos marcadores foram mapeados. Uma rede de co-expressão foi modelada permitindo não apenas a investigação dos genes mapeados, mas também seus módulos de co-expressão. A anotação funcional mostrou que os

genes estão majoritariamente associados ao transporte de auxina e biossíntese de lignina, flavonols e ácido fólico, enquanto os genes co-expressos estão associados ao metabolismo de DNA, respostas a estresses e ao ritmo circadiano. A metodologia estabelecida fornece uma abordagem viável de melhoramento assistido por marcadores em forrageiras tropicais poliploides, e identificou regiões específicas para estudos moleculares futuros das características agronômicas investigadas. Esta tese mostra que os métodos e ferramentas do melhoramento molecular, desenvolvidos e amplamente aplicados em plantas modelo, podem ser adaptados para o melhoramento de espécies com genomas de alta complexidade, tais como as forrageiras tropicais, contribuindo para a formação de um arcabouço de conhecimento que auxiliará sua implementação nos programas de melhoramento dessas espécies.

Palavras-chave: forrageiras tropicais poliploides, melhoramento molecular, contaminantes, predição genômica, aprendizado de máquina, redes de co-expressão.

ABSTRACT

Tropical forage grasses, especially species of the genus *Urochloa*, play an important role in cattle production being the main food source for the animals in the tropical/subtropical regions. Although conventional breeding of tropical forage grasses has resulted in the release of superior cultivars in recent decades, the genetic gains are low and molecular breeding still not being used due to the genus' scarce genomic resources and genomic complexity. In this scenario, this thesis aims to investigate the application of molecular breeding methods and techniques in polyploid tropical forage grasses and produce useful knowledge to auxiliate breeding programs. Artificial hybridization and controlled crosses in tropical forages may be challenging especially due to apomixis and autogamy in certain species, and other situations that can cause the insertion of contaminants in the population of interest. This way, we proposed a semi-automated multivariate methodology for the detection and classification of putative contaminants, including apomictic clones, self-fertilized individuals, half-siblings, and full contaminants, in biparental polyploid progenies. We established a pipeline using genotyping-by-sequencing (GBS) data encoded as allele dosages of SNP markers by integrating principal component analysis, genotypic analysis measures based on Mendelian segregation, and clustering analysis. The method implemented in the fast and user-friendly polyCID app allowed the correct identification of all contaminants in all simulated progenies and the detection of putative contaminants in three real progenies. Furthermore, as most of the species are apomictic and tetraploid, *U. ruziziensis*, a sexual diploid species that can be tetraploidized to be used in interspecific crosses with apomictic species takes special importance. This way, we investigated the applicability of genome wide family prediction (GWFP) in tetraploidized *U. ruziziensis* half-sibling families to predict growth and biomass production. Machine learning and feature selection algorithms were used to reduce the number of markers necessary for prediction to approximately 11 markers and to enhance the predictive ability to approximately 0.9 across the phenotypes. Beyond that, to investigate the regulation of the agronomic traits, the position of the markers with more importance for the prediction were considered as putative QTLs, and in a multi-omic approach, genes obtained in the species transcriptome were mapped linked to those markers. Furthermore, a gene co-expression network was modeled enabling not only the investigation of the mapped genes but their co-expressed genes too. The functional annotation showed that the mapped genes are mainly associated with auxin transport and biosynthesis of lignin, flavonol and folic acid, while the co-expressed genes are associated with DNA metabolism, stress response and circadian rhythm. The methodology provided a viable marker-assisted breeding approach for

tropical forages and identified target regions for future molecular studies on these agronomic traits. This thesis shows that molecular breeding methods and tools, developed and broadly applied in model plants, with the necessary adjustments, can be used in the breeding of orphan species with high complexity genome like the tropical forages, contributing to the formation of a knowledge that will auxiliate its implementation in those species breeding programs.

Keywords: polyploid tropical forages, molecular breeding, contaminants, genomic prediction, machine learning, co-expression networks.

SUMÁRIO

ORGANIZAÇÃO DA TESE

Esta tese está organizada em oito partes que se iniciam a partir de uma **Introdução Geral**, onde são abordados os principais aspectos do cultivo de forrageiras tropicais no Brasil, indicando a sua importância e as dificuldades encontradas pelos programas de melhoramento. Em seguida, é apresentada uma breve **Revisão Bibliográfica** acerca dos principais assuntos abordados nos capítulos, e então são apresentados os **Objetivos Gerais e Específicos** que foram investigados e alcançados nos capítulos.

No **Capítulo I** desenvolvemos uma ferramenta para a identificação semi automática de contaminantes em populações biparentais tetraplóides e hexaplóides utilizando marcadores SNPs. Implementada em um R Shiny web app chamado polyCID, com interface visual para facilitar seu uso, nossa metodologia integra análises de componentes principais (PCA), análises de genótipo e algoritmos de agrupamento para identificar contaminantes. Em nossas simulações, obtivemos 100% de identificação correta em populações compostas por 200 indivíduos genotipados com mais de 689 marcadores. A metodologia foi aplicada em três populações reais de forrageiras tropicais, duas tetraploides (*U. decumbens e M. maximus*) e uma hexaploide (*U. humidicola*), e foram identificados contaminantes do tipo clone apomíticos em todas as populações.

No **Capítulo II** investigamos a aplicabilidade de predição genômica em bulks de famílias de meios irmãos em *U. ruziziensis*, e os processos metabólicos envolvidos na expressão das características preditas. A população foi genotipada por GBS (do inglês Genotyping-By-Sequencing) e fenotipada em nove cortes para características agronômicas relacionadas a produção de biomassa e crescimento. Modelos convencionais e de aprendizado de máquina foram comparados, assim como subconjuntos de marcadores obtidos através de métodos de seleção de *features*. Obtivemos uma média de aproximadamente 0.9 de habilidade preditiva utilizando conjuntos de em média 11 marcadores por característica. Com um transcriptoma da espécie, construímos uma rede de co-expressão, e identificamos os genes que estão fisicamente ligados aos marcadores de maior importância para a predição. A partir da anotação funcional destes genes e daqueles co-expressos, obtivemos um panorama restrito e um amplo dos processos biológicos que atuam na regulação das características agronômicas investigadas.

Após os capítulos, um breve **Resumo dos Resultados** está descrito, seguido pelas **Conclusões Gerais** relacionadas aos capítulos e por fim as **Perspectivas** referentes a pesquisa desenvolvida na tese.

INTRODUÇÃO

As gramíneas forrageiras tropicais são de extrema importância para a pecuária brasileira, compondo a base da alimentação bovina do país e influenciando no custo da produção da carne e do leite, e na segurança alimentar (Dias-Filho, 2016). As pastagens formadas por essas gramíneas forrageiras ocuparam cerca de 163 milhões de hectares e permitiram alimentar um rebanho total de 196 milhões de cabeças no ano de 2021 (ABIEC, 2022). Esses valores posicionaram o Brasil como portador do maior rebanho bovino e principal exportador do planeta, fazendo com que a pecuária total representasse aproximadamente 10% do PIB nacional (ABIEC, 2022). Paralelamente ao mercado da carne, o comércio brasileiro de sementes de forrageiras tropicais também possui grande importância econômica. Devido ao desenvolvimento de pastagens melhoradas, o Brasil passou a ser o maior produtor, consumidor e exportador dessas sementes, movimentando cerca de 440 milhões de dólares por ano, o que representa 11% do mercado de sementes no Brasil (Jank et al., 2014; Campante, 2019).

Estima-se que mais da metade das pastagens cultivadas no Brasil sejam formadas por gramíneas do gênero *Urochloa* (Macedo, 2006), e as espécies de maior importância neste gênero são: *U. brizantha*, *U. decumbens*, *U. ruziziensis* e *U. humidicola*, todas de origem africana e com níveis de ploidia variados (Renvoize et al, 1996). A introdução comercial de espécies do gênero *Urochloa* nas pastagens brasileiras é relativamente recente, sendo a primeira em 1952 com um acesso de *U. decumbens* (cv. Ipean) (Valle et al., 2008). Ao longo do tempo, devido a problemas especialmente relacionados a cigarrinhas-das-pastagens, fotossensibilização e baixa produtividade, a pecuária brasileira sofreu grandes prejuízos, o que incentivou a busca por novas cultivares. A fim de mitigar problemas como esses através do melhoramento de espécies e diversificação das pastagens, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) trouxe para o Brasil parte do germoplasma coletado entre 1984 e 1985 no leste africano pelo Centro Internacional de Agricultura Tropical (CIAT) (Valle, 1990; Valle et al., 2008).

Uma importante característica das forrageiras tropicais é a capacidade de se reproduzir por apomixia, que é uma forma assexual de reprodução na qual são geradas sementes geneticamente idênticas à planta mãe, fazendo com que a progênie seja composta por clones (Bicknell, 2004). Apesar de ser muito interessante devido a possibilidade da fácil fixação de genótipos superiores, os programas de melhoramento acabam tendo dificuldades para realizar cruzamentos devido a ausência de genótipos sexuais. Nesse contexto, o que tem sido feito é a poliploidização de genótipos diplóides sexuais para que sejam utilizados em

cruzamentos com genótipos poliploides apomíticos (Simioni e Valle, 2009, Euclides et al., 2010; Jank et al., 2014). Ainda que haja formas de identificar se a planta é apomítica ou não, como por exemplo através da análise de sacos embrionários (Valle, 1990), há a possibilidade da planta diplóide sexual ao ser poliploidizada se tornar apomítica facultativa e expressar essa característica apenas em situação de campo, contaminando progênies híbridas com clones apomíticos.

A *Urochloa ruziziensis* é a principal espécie utilizada para poliploidização e aplicação como genitor feminino em cruzamentos interespecíficos no gênero *Urochloa*. Além disso, é uma espécie com alta qualidade nutricional (Simioni e Valle, 2009; Timbó et al., 2014) e de importância para a sustentabilidade do agronegócio brasileiro, sendo muito utilizada em sistemas integrados que unem lavoura, pecuária e floresta (Balbino et al., 2011). Apesar de sua importância, e tendo sido utilizada para o lançamento de diversas cultivares híbridas, o melhoramento específico de *U. ruziziensis* ainda é muito recente e teve sua primeira cultivar lançada pela Embrapa em 2022, a BRS Integra.

Nesse contexto, técnicas e métodos de melhoramento molecular, como a genotipagem, construção de mapas genéticos, mapeamento de QTL, seleção/predição genômica e análises de transcriptoma ainda são pouquíssimos utilizados e investigados na espécie. Dessa forma, este trabalho teve dois objetivos principais para expandir o conhecimento relacionado ao melhoramento molecular de gramíneas forrageiras tropicais: (1) desenvolvimento de uma ferramenta capaz de fazer a identificação de contaminantes em populações biparentais tetra e hexaploides genotipadas com marcadores SNPs; (2) avaliar a aplicação de GWFP (predição genômica ampla em famílias) em *U. ruziziensis* e através de uma abordagem multi ômica investigar vias metabólicas que atuam na regulação dos fenótipos avaliados.

REVISÃO BIBLIOGRÁFICA

Gênero *Urochloa*: importância e melhoramento genético

O gênero *Urochloa* P. Beauv. (sinonímia *Brachiaria* (Trin.) Griseb.) é pertencente à família Poaceae, à subfamília Panicoideae e à tribo Paniceae, e possui diversas espécies distribuídas em regiões tropicais e subtropicais, em habitats que vão desde várzeas inundáveis até savanas (Valle et al., 2009). No Brasil, estima-se que mais de 50% da área de pastagem plantada são cultivadas com espécies do gênero *Urochloa* (Figueiredo et al. 2012), sendo que as mais utilizadas são *U. brizantha*, *U. decumbens*, *U. ruziziensis* e *U. humidicola*, todas de origem africana e com níveis de ploidia variados (Renvoize et al, 1996). A importância do gênero para o desenvolvimento da pecuária no Brasil reside no fato de que a alimentação de ruminantes é realizada em pastos, de onde estes animais retiram aproximadamente 90% dos nutrientes que consomem (Euclides et al. 2010). Por possuírem características essenciais como a alta produção de matéria seca, elevada adaptação a diversos tipos de solos e variabilidade para resistência a pragas e doenças, algumas espécies do gênero são ótimas opções para serem usadas como pastagens (Valle et al., 2008; Figueiredo, 2011; Mendonça, 2012).

A introdução comercial de espécies do gênero *Urochloa* nas pastagens brasileiras é relativamente recente, sendo a primeira em 1952 com um acesso de *U. decumbens* (cv. Ipean), que apresentou baixa produção de sementes. Na década seguinte, outra cultivar de *U. decumbens* foi introduzida, conhecida como cv. Basilisk, a qual apresentou excelente adaptação às condições locais e rapidamente tornou-se a principal espécie de forrageira do país. No mesmo período, houve a introdução de cultivares de *U. ruziziensis*, *U. humidicola* e *U. arrecta* (Valle et al., 2008). No início da década de 1980, a monocultura de *U. decumbens* passou a apresentar sérios problemas, tais como a suscetibilidade à cigarrinha-das-pastagens e a fotossensibilização de animais, acarretando em grandes prejuízos para a pecuária brasileira. A busca por novas cultivares para solucionar esses problemas levou ao lançamento da primeira cultivar de *U. brizantha* em 1984, chamada de cv Marandu, a qual passou a substituir gradualmente as pastagens de *U. decumbens* por possuir diversos atributos agronômicos relevantes, dentre eles, a resistência às cigarrinhas-das-pastagens (Nunes et al., 1984).

Nesse contexto, devido à necessidade de diversificação das pastagens e aumento da produtividade, foi imprescindível a obtenção de um banco de germoplasma de *Urochloa* para dar início aos programas de melhoramento. Sendo assim, viagens de coleta foram realizadas durante 1984 e 1985 no leste africano, região com grande diversidade e dispersão do gênero

(Valle et al., 2008) pelo Centro Internacional de Agricultura Tropical (CIAT) (Keller-Grein et al., 1996; Miles e Valle, 1996), e boa parte do material coletado foi duplicado e trazido para o Brasil para constituir a coleção da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), que é utilizada até hoje pelos programas de melhoramento. Além dos germoplasmas já citados pertencentes ao CIAT e a Embrapa, são mantidos mais cinco germoplasmas de *Urochloa* distribuídos pelo mundo. Sendo os do Centro Internacional de Pecuária da África - ILCA, do Centro de Recursos Genéticos de Forrageiras Tropicais Australiano - ATFGRC/CSIRO, do Departamento de Agricultura dos Estados Unidos - USDA, do Germoplasma Nacional do Quênia - GBK e do Instituto Pasteur Roodeplaat/Conselho Africano de Pesquisa - RGI/ARC (Keller-Grein et al., 1996).

Apesar da obtenção do germoplasma, os programas de melhoramento ainda tinham um grande desafio a ser superado. As principais espécies do gênero (*U. brizantha, U. decumbens* e *U. humidicola*) são predominantemente apomíticas, isso é, se reproduzem assexuadamente através de sementes (Valle et al., 2008, Valle et al., 2009). Essa característica limita o melhoramento das espécies, já que na ausência de parentais sexuais, não é possível realizar cruzamentos, deixando a princípio apenas a seleção direta no germoplasma como forma de lançar cultivares (Jank et al., 2005; Miles, 2007). Ainda que a seleção direta seja uma forma rápida e simples de melhoramento, raramente uma planta apomítica possui todos os caracteres de interesse. Para superar esse obstáculo, a espécie *U. ruziziensis*, originalmente diploide e sexual, teve sete genótipos poliploidizados artificialmente (Swenne et al. 1981). Dessa forma, em conjunto com as espécies tetraploides *U. brizantha* e *U. decumbens*, esses novos genótipos formaram um complexo agâmico, possibilitando a realização de cruzamentos inter-específicos e a obtenção/seleção de híbridos superiores (Valle et al., 2008, Valle et al., 2009). Embora a apomixia represente uma limitação para a realização de cruzamentos, é uma característica de muita importância para a comercialização de cultivares de gramíneas forrageiras. Genótipos superiores que se reproduzem por apomixia podem ter seus clones comercializados através de suas sementes, sendo fácil a obtenção de grandes quantidades de sementes e garantindo uma maior uniformidade no cultivo.

O estudo e a avaliação dos germoplasmas pelos programas de melhoramento permitiu que diversos cultivares fossem lançados. Atualmente existem 20 cultivares comercialmente disponíveis no mundo, cultivares das espécies *U. ruziziensis* (Kennedy e BRS Integra), *U. brizantha* (Marandu, MG-4, Xaraés, BRS Piatã, BRS Paiaguás, Braúna MG 13 e BRS Ybaté), *U. decumbens* (Basilisk) e *U. humidicola* (BRS Tupi), além de diversas cultivares híbridas de *U. ruziziensis* com as espécies tetraploides, sendo a Embrapa e a Barenbrug do

Brasil Sementes LTDA as instituições responsáveis por boa parte dos cultivares (Ferreira et al., 2021).

Comparado a outras gramíneas de importância, como trigo e arroz, o conhecimento genético/genômico do gênero *Urochloa* é muito restrito, fazendo com que a aplicação de métodos de melhoramento molecular, como por exemplo através seleção assistida por marcadores, seja muito limitado. Apesar de características intrínsecas como o modo de reprodução e a poliploidia, e também baixos investimentos dificultarem os estudos no gênero, os avanços das tecnologias genômicas e a redução de seus custos, aliados ao surgimento de métodos estatísticos apropriados, permitiram considerável progresso no conhecimento genético/genômico dessas espécies nos últimos anos.

Dentre esses avanços está principalmente o desenvolvimento de marcadores moleculares, que a princípio foram utilizados para estudos de diversidade genética. Com marcadores do tipo microssatélite (SSR) foram feitos trabalhos em painéis de genótipos com mais de uma espécie do gênero (Chiari et al., 2008; Almeida et al., 2011; Garcia et al., 2013; Kuwi et al., 2018; Namazzi et al., 2020), em acessos de *U. humidicola* (Jungman et al., 2009a; Jungmann et al., 2010; Santos et al., 2015; Vigna et al., 2016a), *U. brizantha* (Jungmann et al., 2009b; Vigna et al., 2011; Braga et al., 2017; Tegegn et al., 2019), *U. decumbens* (Ferreira et al., 2016; Souza et al., 2018) e *U. ruziziensis* (Azevedo et al., 2011; Silva et al., 2013; Pessoa-Filho et al., 2015). Além da avaliação de diversidade, marcadores moleculares também foram utilizados para construir mapas genéticos. Até o momento existem mapas para populações híbridas (Worthington et al., 2016; Thaikua et al., 2016; Worthington et al., 2021), para *U. humidicola* (Vigna et al., 2016a, Worthington et al., 2019) e para *U. decumbens* (Ferreira et al., 2016), que foram construídos com marcadores SSR e SNP. Já em relação a associação genômica ampla (GWAS) e seleção genômica, os estudos são ainda mais escassos. Ambas metodologias foram investigadas em uma população híbrida do cruzamento entre *U. brizantha* e *U. ruziziensis* (Matias et al., 2019a; Matias et al., 2019b) e seleção genômica foi avaliada em uma população biparental de *U. decumbens* (Aono et al., 2022), os três trabalhos foram realizados utilizando marcadores SNPs.

Além de trabalhos com genotipagem, também estão disponíveis estudos de transcriptoma com espécies do gênero. A primeira investigação a nível de transcritos foi realizada utilizando material foliar de dois genótipos contrastantes de *U. humidicola* coletados em condições de campo (Vigna et al., 2016b). Com o objetivo de investigar as bases moleculares da resistência a estresses abióticos, foram realizadas análises de transcriptoma em *U. decumbens* (Basilisk) por Salgado et al. (2017) e por Worthington et al.

(2021), que também utilizou no estudo o genótipo de *U. ruziziensis* (BRX 44-02). Já o trabalho de Jones et al. (2021) foi feito utilizando híbridos interespecíficos para estudar as respostas ao estresse causado pela falta de água. Por fim, o trabalho mais recente envolvendo RNA-seq, foi feito com um painel diverso de espécies do gênero *Urochloa* com o objetivo de identificar alelos LOF (*loss-of-function*) relacionados a digestibilidade e conteúdo lipídico da biomassa produzida pelas plantas (Hanley et al., 2021).

Espécie *U. ruziziensis*: importância e melhoramento genético

A espécie *Urochloa ruziziensis* (R. Germ. and C.M. Evrard), também conhecida como: “Congo signal grass”, “Congo grass” e “Ruzi grass”, é originária da África, sendo encontrada principalmente no vale do rio Ruzizi, que fica na divisa entre a República Democrática do Congo, Burundi e Ruanda, em áreas de solos frescos e não inundáveis (Keller-Grein et al. 1996). Quanto às suas características morfológicas, *U. ruziziensis* é uma espécie perene com 1-1.5 m de altura, rizomas curtos, robustos e globosos; nós comprimidos, de cor escura e glabros; lâmina linear-lanceolada, de 10 a 30 cm de comprimento, 10 a 15mm de largura, pubescentes, verde amarelada, esparsamente pilosa; possui rizomas fortes, em forma de tubérculos arredondados e com até 15 mm de diâmetro; inflorescência em panícula com racemos bilaterais, terminais, de 15 a 25 cm de comprimento com 3 a 7 às vezes até 9 ramificações primárias e alternas; a inflorescência está formada por 3-6 racemos de 4-10 mm de comprimento; ráquis largamente alada, com 4 mm de largura, geralmente de cor arroxeada; espiguetas de 8 a 5 mm de comprimento, pilosas na parte apical, bisseriadas ao longo da ráquis; a gluma inferior tem 3 mm de comprimento e surge de 0.5 a 1 mm abaixo do resto da espigueta e o flósculo fértil apresenta 4 mm de comprimento (Seiffert, 1980; Valle et al., 2010).

Em relação a suas características agronômicas, a espécie possui um estabelecimento e crescimento rápido, que ocorre geralmente no início do período das chuvas, e seu florescimento é concentrado com alta produção de sementes (Valle et al. 2010). A biomassa produzida por essa pastagem possui elevado valor nutricional, com altas taxas de degradabilidade da matéria seca e da proteína bruta, baixo teor de fibra em detergente neutro e alta aceitação pelo gado, devido à boa palatabilidade (Simioni and Valle 2009; Lopes et al. 2010; Souza Sobrinho et al. 2010; Timbó et al. 2014). Apesar dessas boas características, a espécie apresenta baixa adaptação a solos mal drenados e de baixa fertilidade, alta suscetibilidade às cigarrinhas das pastagens, baixa competição com invasoras e baixa tolerância à seca (Valle et al. 2010). Uma interessante característica da espécie é sua

importância para a sustentabilidade do agronegócio brasileiro, sendo muito utilizada em sistemas integrados, tais como ILP e ILPF, que unem lavoura, pecuária e floresta (Balbino et al., 2011). Esse sistemas reúnem atributos raros em outros sistemas de produção de alimentos, são mais eficientes no uso dos recursos naturais (Wrigth et al., 2012), promovem ciclagem de nutrientes e melhoria do solo (Salton et al., 2014), reduzem os custos de produção mantendo níveis de produtividade elevados (Ryschawy et al., 2012; Balbinot et al., 2009) e produzem inúmeros serviços ecossistêmicos (Sanderson et al., 2013).

Pelo fato de ser possível obter artificialmente tetraplóides sexuais de *U. ruziziensis*, a representatividade da espécie dentre as cultivares de *Urochloa* disponíveis no mundo se dá principalmente em cultivares híbridas. Atualmente estão disponíveis apenas as cultivares Kennedy (1966) e BRS Integra (2020) especificamente de *U. ruziziensis*, enquanto cultivares híbridas estão disponíveis: Mulato I e II (2000, 2005), Mixe Drwn 12 e LN 45 (ambas de 2013), BRS RB331 Ipyporã (2017), Cayana (2020), Convert 330 (2021) e BARG156 780J (2021), que são resultado de hibridações com *U. brizantha* e *U. decumbens* (Ferreira et al., 2021).

Quanto a genética e aos recursos genético/genômicos da espécie, avaliações citogenéticas mostraram que as plantas diplóides apresentam 18 cromossomos (2n = 2x = 18) (Bernini and Marin-Morales, 2001) e as artificialmente poliploidizadas apresentam 36 cromossomos (2n = 4x = 36) (Timbó et al., 2014), e investigaram o comportamento dos cromossomos em todas as fases da meiose (Risso-Pascotto et al., 2003; Risso-Pascotto et al., 2005). Dentro do gênero, *U. ruziziensis* é a única espécie que possui o genoma sequenciado, dos dois sequenciamentos disponíveis ambos são de indivíduos diplóides, um deles está a nível de cromossomo, obtido do genótipo “C6” através do sequenciamento PacBio baseado na tecnologia SMRT *(Single Molecule Real Time)* (Pessoa-Filho et al., 2019), e o outro, do genótipo CIAT 26162, a nível de *scaffolds*, que foi obtido por WGA *(Whole Genome Assembly)* através da montagem de *reads* curtos (Worthington et al., 2021).

Além do genoma sequenciado, a espécie possui marcadores moleculares do tipo microssatélite (SSR) identificados, esses marcadores foram desenvolvidos principalmente para avaliações de diversidade, como nos trabalhos de Azevedo et al. (2011) e Pessoa-Filho et al. (2015), que reportaram alta diversidade e alta heterozigosidade nos painéis de genótipos. Além desses dois trabalhos, Silva et al. (2013) validou um conjunto de marcadores SSR identificados através de uma montagem do genoma parcial *de novo* utilizando *reads* Illumina *single-end*, que podem ser utilizados em análises genéticas e em seleção assistida por marcadores. Apesar de não especificamente com populações de *U.ruziziensis*, mapas

genéticos foram construídos para populações híbridas obtidas de cruzamentos com outras espécies do gênero. O primeiro foi construído utilizando marcadores AFLP para uma população oriunda do cruzamento entre a cultivar apomítica "Mulato" e a cultivar de *U. ruziziensis* "Miyaokikoku" (Thaikua et al., 2016). Depois, já com marcadores do tipo SNP, mais dois mapas foram construídos utilizando cruzamentos envolvendo os acessos sexuais tetraploidizados BRX 44-0 e BRX 44-2 com o acesso CIAT 606 da cultivar Basilisk de *U. decumbens* (Worthington et al. 2016; Worthington et al. 2021). Ademais, para investigar as bases moleculares da tolerância a solos com alumínio, foi feito um transcriptoma dos acessos BRX 44-2 e CIAT 606 comparando os perfis de expressão sob situação de alta concentração ou ausência de alumínio no solo (Worthington et al. 2021). Já o trabalho de Hanley et al. (2021) que investigou um painel diverso de *Urochloa*, é o mais recente que produziu dados de RNA-seq de plantas da espécie *U. ruziziensis*, estando incluso no painel 11 genótipos da espécie.

Identificação de contaminantes

Um dos problemas relacionados ao método de artificialmente poliploidizar genótipos sexuais de gramíneas forrageiras tropicais é a possibilidade da planta passar a ser apomítica facultativa, sendo capaz de produzir simultaneamente híbridos oriundos do cruzamento de interesse e clones apomíticos (Smith, 1972). Ainda que seja possível avaliar a apomixia nos genótipos através da análise de sacos embrionários (Valle, 1990), plantas que apresentam apomixia facultativa podem não expressar a característica durante a avaliação. Além disso, devido ao sistema reprodutivo dessas plantas e a anemofilia (polinização das flores por ação do vento), existe a possibilidade de indivíduos serem gerados por autofecundação ou por fecundação através de pólen vindo de outro campo (Bateman, 1947; Simeão et al., 2016). Somando as situações descritas a ocorrência de erros durante a coleta e manejo de sementes, mesmo em cruzamentos controlados, contaminantes como clones apomíticos, progênies de autofecundação e híbridos de cruzamentos inesperados podem ser inseridos nas populações de interesse. O comprometimento de progênies puramente híbridas, a depender do nível de contaminação, pode trazer grandes vieses para análises genéticas e genômicas, como testes de segregação e de desequilíbrio de ligação, construção de mapas genéticos e mapeamento de QTL, que são fundamentais para o entendimento das relações entre genótipo e fenótipo (Kemble et al., 2019).

Tradicionalmente, a identificação de híbridos tem sido feita através de caracteres morfológicos e marcadores microssatélite (SSR) (Santos et al., 2014; Jha et al., 2016; Zhao et

al., 2017; Patella et al., 2019), que são métodos com diversas desvantagens. Análises de características morfológicas demandam muito tempo, apresentam baixo rendimento e possuem acurácias que são influenciadas por fatores ambientais (Zhao et al., 2017). Já o desenvolvimento de marcadores microssatélites é um processo demorado e trabalhoso, que depende de informações genômicas obtidas previamente, desenvolvimento de *primers* específicos e otimização da técnica de PCR (*Polymerase Chain Reaction*) (Vieira et al., 2016), além de apresentar frequentemente erros de genotipagem em poliploides (Guichoux et al., 2011; Hodel et al., 2016). Nesse sentido, os marcadores SNPs, que podem ser obtidos através dos métodos de genotipagem baseados em sequenciamento de nova geração, como o GBS (*genotyping-by-sequencing)* (Elshire et al. 2011; Poland et al., 2012), podem ser obtidos rapidamente, por preços acessíveis e de forma confiável em espécies com pouca informação genômica disponível e com grandes genomas, como as espécies poliploides (Elshire et al., 2011; Poland et al., 2012; Ferreira et al., 2019; Deo et al., 2020; Mollinari et al., 2020).

Atualmente, existem alguns *software* capazes de atribuir paternidade ou nível de parentesco entre amostras diplóides, fazendo uso de marcadores microssatélites ou SNPs através de diferentes métodos estatísticos (Kalinowski et al., 2007; Jones and Wang, 2010; Hayes, 2011; Anderson, 2012; Heaton et al., 2014; Huisman, 2017; Grashei et al., 2018; Whalen et al., 2019). Para poliploides, os programas estão limitados a marcadores microssatélites (Spielmann et al., 2015; Zwart et al., 2016), ou métodos para estimar índices de parentesco porém incapazes de classificar amostras (Huang et al., 2015; Amadeu et al., 2020). Ainda que haja opções para tentar identificar contaminantes em populações biparentais poliploides, as ferramentas não foram desenvolvidas especificamente para essa tarefa, o que acaba demandando alto conhecimento da ferramenta e do método pelo usuário para que a análise seja adaptada a esse objetivo.

Predição e seleção genômica

Seleção genômica é um método de melhoramento assistido por marcadores moleculares proposto por Bernardo (1994) e amplamente difundido pelo trabalho de Meuwissen et al. (2001). O método se baseia no uso de dados de genotipagem e fenotipagem para desenvolver modelos de predição capazes de estimar valores genéticos de amostras que foram apenas genotipadas. Isso é possível porque com a obtenção de muitos marcadores que cobrem todo o genoma, parte desses marcadores estarão em desequilíbrio de ligação com os QTL que influenciam a característica (Meuwissen et al. 2001). As principais vantagens da seleção genômica estão relacionadas a redução dos gastos e do tempo necessário para o

processo de melhoramento. Além de não ser necessário fazer a fenotipagem, que é um processo caro, as plantas podem ser avaliadas logo que brotam, não sendo necessária a manutenção da população no campo por muito tempo (Simeão-Resende et al., 2014, Crossa et al., 2017).

As populações necessárias para a realização de um programa de seleção genômica geralmente são três, essas populações são conjuntos de genótipos nos quais os modelos preditivos são treinados, validados e aplicados. A primeira população é chamada de população de treino, os indivíduos dessa população são genotipados e fenotipados para as características de interesse e, a partir desses dados, o modelo estatístico é treinado, criando a relação entre os valores genéticos e os marcadores. A segunda população é chamada de população de teste, a qual geralmente é uma população menor que a de treino, mas também possui seus indivíduos genotipados e fenotipados. A função dessa população é avaliar a performance do modelo de predição genômica, comparando os valores estimados com os valores reais observados. Por fim, a terceira população é a de melhoramento, nessa população os indivíduos são apenas genotipados e seus valores genéticos são estimados pelo modelo previamente treinado, podendo haver assim a seleção das melhores amostras sem a necessidade da fenotipagem (Goddard e Hayes, 2007; Nakaya e Isobe, 2012; Desta e Ortiz, 2014).

Muitas vezes a população de treino e teste são a mesma população e a modelagem é avaliada através do método de validação cruzada, nesse método a população genotipada e fenotipada é dividida em *"k"* grupos, dos quais *"k-1"* são utilizados como população treino e o último grupo como população de teste, isso é feito *"k"* vezes de forma que todos os grupos sejam utilizados tanto como treino quanto como teste. Em cada rodada da validação cruzada, as métricas de avaliação são computadas e por fim uma média é obtida. Das métricas utilizadas para avaliar modelos de regressão, as mais utilizadas em seleção genômica são a habilidade preditiva (coeficiente de correlação de Pearson) e o erro quadrático médio (Schulz-Streeck et al., 2012; Ould Estaghvirou et al., 2013).

Desde que a seleção genômica foi proposta, diferentes modelos estatísticos foram investigados. Dentre os mais aplicados, estão os modelos paramétricos lineares mistos, como o RR-BLUP, GBLUP e Bayes(A,B,C), modelos semi-paramétricos como o RKHS e modelos não paramétricos, como os algoritmos de aprendizado de máquina: SVM, RF e redes neurais (VanRaden, 2008; Endelman, 2011; Pérez et al., 2014; Wang et al., 2018; Montesinos-López et al., 2021). Quando comparados, geralmente o melhor modelo varia a depender da espécie e do fenótipo avaliado, não havendo assim um modelo que seja superior aos outros em todas as

situações. Além do modelo, a acurácia da seleção genômica é afetada por vários fatores, como o tamanho amostral, distância genética, densidade de marcadores, herdabilidade e desequilíbrio de ligação entre marcadores e QTL (Wang et al., 2018).

No caso de *U. ruziziensis* e outras espécies forrageiras, como alfafa e azevém, que podem ser melhoradas através de seleção de famílias completas ou de meios irmãos (Simeão et al., 2012; Simeão et al., 2016a,b; Biazzi et al., 2017; Cericola et al., 2018; Jia et al, 2018; Andrade et al., 2022), a aplicação de GWFP (predição genômica ampla em famílias) é uma possibilidade. Nessa situação, tanto a fenotipagem quanto a genotipagem são feitas a nível de família e não de indivíduo, sendo assim uma forma de redução de gastos, uma vez que grandes populações podem ser reduzidas a amostras de famílias (Rios et al., 2021).

Associações genótipo-fenótipo e análises multi ômicas

A identificação de associações genótipo-fenótipo é de suma importância tanto para programas de melhoramento quanto para investigações metabólicas/fisiológicas. Marcadores moleculares em desequilíbrio de ligação com QTL podem ser utilizados auxiliando programas de melhoramento a selecionar amostras de maior valor genético e na identificação de regiões genômicas que atuam na regulação da característica avaliada. Atualmente existem diferentes métodos, com vantagens e desvantagens, que são utilizados para identificar essas associações.

Um método tradicional é através do mapeamento de QTL, que apesar da sua importância, possui desvantagens que impedem sua aplicação constante em programas de melhoramento (Bernardo, 2008). Além da necessidade de populações biparentais segregantes (RILs, NILs, F2, etc), o alto desequilíbrio de ligação induzido nessas populações experimentais restringe a relevância dos resultados apenas à população estudada e a um número reduzido de marcadores ancorados a grupos de ligação. Ademais, o mapeamento possui melhor eficiência para características controladas por "*major-genes*", que são incomuns para características de importância agronômica (Mohan et al., 1997; Goddard e Hayes, 2007; Dhingani et al., 2015; Crossa et al., 2017). Outra forma de detecção de QTL é através do método de GWAS (Estudos de Associação Genômica Ampla), que apesar de assim como no mapeamento de QTL fazer uso de modelos lineares mistos para identificar as associações, não necessita de populações biparentais, podendo utilizar populações de alta diversidade para identificar as associações e os resultados não estão restritos a população estudada (Myles et al. 2009, Bernardo, 2016; Crossa et al., 2017). Porém, como nessas populações o desequilíbrio de ligação é muito menor, é necessário que as amostras sejam

genotipadas com grandes quantidades de marcadores distribuídos por todo o genoma (Zargar et al., 2015).

Apesar de não ser um método tão estabelecido como o mapeamento de QTL e o GWAS, a popularização do uso de algoritmos de aprendizado de máquina para predição genômica, fez com que os métodos de seleção de *features* passassem a ser utilizados como forma de identificar marcadores associados a fenótipos. Seleção de features é uma estratégia usada em ciência de dados para remover dados redundantes, incorretos ou irrelevantes em uma modelagem, e em alguns casos aumentar a performance do modelo (Miao & Niu, 2016; Cai et al., 2018). No contexto da predição genômica, ao selecionar os melhores marcadores para a predição, a seleção de *features* acaba por identificar um conjunto de marcadores que possivelmente estão em desequilíbrio de ligação com os QTL que influenciam o fenótipo. Essa estratégia já foi aplicada em diversas espécies, sendo capaz de aumentar a performance dos modelos, ou pelo menos manter a performance utilizando conjuntos muito menores de marcadores (Li et al. 2018; Azodi et al., 2019 ; Luo et al., 2021; Piles et al., 2021). Além disso, outros trabalhos vêm mostrando que a posição desses marcadores pode ser utilizado em investigações de processos moleculares relacionados ao fenótipo avaliado (Steinfath et al., 2010; Heer et al., 2018; Zhou et al., 2019; Aono et al., 2020; Pimenta et al., 2021; Aono et al., 2022; Pimenta et al., 2022).

Apesar da existência de diferentes métodos para estabelecer relações estatísticas entre genótipos e fenótipos, os genes, moléculas e interações moleculares envolvidas nessas relações não são explícitas e demandam diversos outros tipos de análises para serem elucidadas. Nesse sentido, a integração de dados biológicos, chamada de análises multi ômicas, tem como objetivo fazer uso de informações biológicas, especialmente moleculares, de diferentes níveis metabólicos para compreender de forma sistêmica e holística processos biológicos complexos que controlam as características investigadas (Richardson, Tseng e Sun, 2016; Yan et al., 2018). Os métodos multi ômicos abrangem especialmente os dados genômicos, proteômicos, transcriptômicos, metabolômicos e epigenômicos, podendo ainda ser estendido para informações biológicas mais específicas, como dados lipidômicos, fosfoproteômicos ou glicoproteômicos (Subramanian et al., 2020). A disponibilidade de dados multi ômicos e métodos estatísticos apropriados para analisar e integrar esses dados tem permitido seu uso em pesquisas de diversas áreas, como em humanos, especialmente estudos de doenças (Yang et al, 2014, Vasaikar et al., 2018; Lloyd-Price et al., 2019; Hill et al., 2022), microrganismos (Borin et al, 2018; Rosolen et al, 2022), animais (Gaddis et al, 2016; Mateescu et al, 2017) e plantas, incluindo trabalhos direcionados ao melhoramento

molecular (Scossa, Alseekh e Fernie, 2021; Francisco et al., 2021; Knoch et al., 2021; Cao et al., 2022; Cardoso-Silva et al, 2022)

Considerando que nos programas de melhoramento muitas vezes são feitos trabalhos de genômica e transcriptômica, uma forma interessante de integrar esses dados é através das redes de co-expressão gênica. A capacidade das redes de simular sistemas biológicos complexos e inferir novas associações biológicas tem revolucionado a pesquisa na área (D'haeseleer et al., 2000; Liu et al., 2020). Baseado no princípio "*guilt by association*", que no caso das redes construídas com dados transcriptômicos, constata que genes com funções biológicas correlatas tendem a interagir nas redes (Oliver, 2000; Wolfe et al., 2005; Childs et al., 2011), é possível estudar relações regulatórias, inferir vias metabólicas e fazer transferência de anotação funcional (Rao & Dixon, 2019). Em resumo, as redes são construídas considerando a correlação do nível de expressão entre dois genes em diferentes amostras, e genes com correlação acima de um limiar definido são considerados co-expressos (Rao & Dixon, 2019). Dessa forma, utilizando resultados de associações genótipo-fenótipo (QLT) obtidos através dos dados genômicos e identificando genes fisicamente ligados a esses loci, é possível isolá-los na rede junto de seus módulos de co-expressão e expandir a investigação para conjuntos gênicos que interagem e atuam na características de interesse (Francisco et al., 2021; Pimenta et al., 2022).

OBJETIVOS

Objetivo Geral

Desenvolver uma ferramenta de fácil uso para a identificação de contaminantes em populações biparentais tetra e hexaplóides e investigar vias metabólicas relacionadas a produção de biomassa e crescimento em *U. ruziziensis* através de uma abordagem multi ômica integrando predição genômica e rede de co-expressão.

Objetivos Específicos

- Desenvolver uma metodologia para a identificação de contaminantes em populações biparentais tetra e hexaplóides;
- Identificar contaminantes em populações reais de forrageiras;
- Implementar a metodologia em um aplicativo com interface gráfica;
- Avaliar a aplicabilidade de predição genômica em bulks de famílias de meios irmãos;
- Comparar modelos convencionais e de aprendizado de máquina para predição genômica;
- Avaliar o uso de métodos de seleção de features para reduzir o conjunto de marcadores necessários para a predição;
- Identificar marcadores associados a características agronômicas através de sua importância para a predição;
- Montar um transcriptoma e fazer a anotação funcional dos genes;
- Identificar genes fisicamente ligados aos marcadores selecionados;
- Investigar a função dos genes identificados;
- Construir uma rede de co-expressão utilizando o transcriptoma;
- Investigar os processos biológicos relacionados aos genes previamente identificados em conjunto com genes co-expressos na rede;
- Diferenciar os processos biológicos relativos às temporadas de chuva e seca.

CAPÍTULO I

A semi-automated SNP-based approach for contaminant identification in biparental polyploid populations of tropical forage grasses.

Autores: Felipe Bitencourt Martins, Aline Costa Lima Moraes, Alexandre Hild Aono, Rebecca Caroline Ulbricht Ferreira, Lucimara Chiari, Rosangela Maria Simeão, Sanzio Carvalho Lima Barrios, Mateus Figueiredo Santos, Liana Jank, Cacilda Borges do Valle, Bianca Baccili Zanotto Vigna e Anete Pereira de Souza.

Publicado no Periódico Frontiers in Plant Science
https://www.frontiersin.org/articles/10.3389/fpls.2021.737919/full
Volume 12:737919. doi: 10.3389/fpls.2021.737919, outubro de 2021.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fpls.2021.737919

A Semi-Automated SNP-Based Approach for Contaminant Identification in Biparental Polyploid Populations of Tropical Forage Grasses

Felipe Bitencourt Martins[1†], *Aline Costa Lima Moraes*[1†], *Alexandre Hild Aono*[1], *Rebecca Caroline Ulbricht Ferreira*[1], *Lucimara Chiari*[2], *Rosangela Maria Simeão*[2], *Sanzio Carvalho Lima Barrios*[2], *Mateus Figueiredo Santos*[2], *Liana Jank*[2], *Cacilda Borges do Valle*[2], *Bianca Baccili Zanotto Vigna*[3] *and Anete Pereira de Souza*[1,4*]

[1] Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), São Paulo, Brazil, [2] Embrapa Gado de Corte, Brazilian Agricultural Research Corporation, Campo Grande, Brazil, [3] Embrapa Pecuária Sudeste, Brazilian Agricultural Research Corporation, São Paulo, Brazil, [4] Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP), São Paulo, Brazil

OPEN ACCESS

Edited by:
Kun Lu,
Southwest University, China

Reviewed by:
Cheng-Ruei Lee,
National Taiwan University, Taiwan
Sukhjiwan Kaur,
Agriculture Victoria, Australia

***Correspondence:**
Anete Pereira de Souza
anete@unicamp.br

[†] *These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *07 July 2021*
Accepted: *20 September 2021*
Published: *22 October 2021*

Citation:
Martins FB, Moraes ACL, Aono AH, Ferreira RCU, Chiari L, Simeão RM, Barrios SCL, Santos MF, Jank L, do Valle CB, Vigna BBZ and de Souza AP (2021) A Semi-Automated SNP-Based Approach for Contaminant Identification in Biparental Polyploid Populations of Tropical Forage Grasses. Front. Plant Sci. 12:737919. doi: 10.3389/fpls.2021.737919

Artificial hybridization plays a fundamental role in plant breeding programs since it generates new genotypic combinations that can result in desirable phenotypes. Depending on the species and mode of reproduction, controlled crosses may be challenging, and contaminating individuals can be introduced accidentally. In this context, the identification of such contaminants is important to avoid compromising further selection cycles, as well as genetic and genomic studies. The main objective of this work was to propose an automated multivariate methodology for the detection and classification of putative contaminants, including apomictic clones (ACs), self-fertilized individuals, half-siblings (HSs), and full contaminants (FCs), in biparental polyploid progenies of tropical forage grasses. We established a pipeline to identify contaminants in genotyping-by-sequencing (GBS) data encoded as allele dosages of single nucleotide polymorphism (SNP) markers by integrating principal component analysis (PCA), genotypic analysis (GA) measures based on Mendelian segregation, and clustering analysis (CA). The combination of these methods allowed for the correct identification of all contaminants in all simulated progenies and the detection of putative contaminants in three real progenies of tropical forage grasses, providing an easy and promising methodology for the identification of contaminants in biparental progenies of tetraploid and hexaploid species. The proposed pipeline was made available through the polyCID Shiny app and can be easily coupled with traditional genetic approaches, such as linkage map construction, thereby increasing the efficiency of breeding programs.

Keywords: GBS, apomictic clones, self-fertilization, half-sibling, allele dosage, principal component analysis, clustering analysis, shiny

INTRODUCTION

The concept of artificial crossings to generate experimental plant populations was introduced scientifically in the historical work of Mendel (1866) and became a fundamental tool for genetics studies and breeding programs, maximizing genetic gains by the selection of superior genotypes (Bourke et al., 2018). Although this concept is well-known and applied in important crops (Goulet et al., 2017), there are few commercial cultivars of tropical forage grasses originating from artificial hybridization (Azevedo et al., 2019). Perennial tropical forage grasses are recognized worldwide for their economic importance as food for beef and dairy cattle in the tropical and subtropical regions (Pereira et al., 2018a; ABIEC, 2020). In addition to the recently initiated breeding programs and long selection cycles, some intrinsic biological characteristics, including different reproductive modes (sexual and facultative apomixis), levels of ploidy, and self-incompatibility (SI) within and between the species, are challenges faced by breeders when performing controlled crosses using these plants (Lutts et al., 1991; Jank et al., 2011; Pereira et al., 2018a; Worthington et al., 2019).

Apomixis is a type of asexual reproduction through seeds that produces a progeny which is genetically identical to the maternal plant (Bicknell, 2004; Hand and Koltunow, 2014). Thus, to explore the genetic diversity of polyploid apomictic forage grasses, controlled crosses are performed between sexual and apomictic (pollen donor) parents with contrasting traits and the same ploidy level. In most species, the ploidy of the sexual plants does not match with the ploidy of the apomictic plants; this way, it is necessary the artificial polyploidization (usually chromosome duplication) of the sexual ones to perform the crosses at the same ploidy level (Pinheiro et al., 2000; Simioni and Valle, 2009; Acuña et al., 2019). However, because of the reproductive system of these plants, during the crosses, some individuals are also generated by foreign pollen or by self-fertilization of female parents. Some of these scenarios can also occur in other species, such as sugarcane, eucalyptus, sainfoin, and lettuce (Santos et al., 2014; Subashini et al., 2014; Kempf et al., 2015; Patella et al., 2019). Also, if facultative apomictic plants (i.e., apomictic plants in which sexual reproduction events are also observed) are used as females, they simultaneously generate hybrid by crossings and clones by apomixis (Smith, 1972). Contamination by physical admixture during seed harvesting and handling is also possible, especially when crosses are performed in the field, as these species are mostly anemophilous (i. e., the pollination of these species occurs by the wind) (Bateman, 1947; Simeão et al., 2016a). In this context, it is evident that controlled crosses may not avoid contamination, compromising the attainment of pure hybrid progeny and, consecutively, unbiased genetic and genomic methods, such as segregation tests, linkage map construction, quantitative trait locus (QTL) mapping and linkage disequilibrium analysis, which are fundamental for understanding the genotype and its relationship to the phenotype (Kemble et al., 2019).

Traditionally, hybrid identification has been performed on the basis of morphological traits and microsatellite markers (Santos et al., 2014; Jha et al., 2016; Zhao et al., 2017; Patella et al., 2019). However, both methodologies have disadvantages. Morphological traits are time-consuming and have low throughput, with accuracies influenced by environmental factors (Zhao et al., 2017), while developing microsatellite markers is an expensive and time-consuming process that requires previously obtained genomic sequence information and investment in terms of designing locus-specific primers and optimizing PCR conditions (Vieira et al., 2016). Moreover, size estimates across alleles at each locus are imprecise, especially in polyploids, such as tropical forage grasses, leading to frequent genotyping errors (Guichoux et al., 2011; Hodel et al., 2016). Therefore, there is a need to develop alternative methodologies using molecular markers to quickly and efficiently distinguish true hybrids resulting from the breeding program crosses from those resulting from accidental selfing or contamination in biparental populations.

Single nucleotide polymorphism (SNP) markers have been shown to be an excellent tool for genomic studies in function of their high-throughput nature, low error rates, and abundance in eukaryote genomes (Helyar et al., 2011). Additionally, genotyping methodologies based on next-generation sequencing (NGS), such as genotyping-by-sequencing (GBS) proposed by Elshire et al. (2011) and Poland et al. (2012), have been demonstrated to be quick, affordable, and highly robust for discovering and profiling a large number of SNP loci, even in species with no genomic information available and large genomes, such as polyploids (Elshire et al., 2011; Poland et al., 2012; Ferreira et al., 2019; Deo et al., 2020; Mollinari et al., 2020). In the last few years, many studies using SNP markers in tropical forage grasses, mainly coupled with principal component analysis (PCA) to investigate the structure of the progenies and remove putative contaminants, have been published (Lara et al., 2019; Deo et al., 2020; Zhang et al., 2020). Even though PCA can be used to retain and explore most of the variations in large SNP datasets through the first principal components (PCs) (Jolliffe and Cadima, 2016), such a multivariate technique is not appropriate for contaminant identification, which requires more specific approaches, such as pedigree reconstruction, sibship and parentage assignment.

The different methods for identifying the parents of a progeny are based on exclusion (Zwart et al., 2016; McClure et al., 2018), likelihood-based (Spielmann et al., 2015), and Bayesian (Christie et al., 2013) techniques, using Mendel's laws to infer relationships between samples through genotyped loci (Thompson, 1975; Thompson and Meagher, 1987). This evaluation is generally

Abbreviations: AC, Apomictic clone; CA, Clustering analysis; FC, Full contaminant; GA, Genotype analysis; GBS, Genotyping-by-sequencing; HP, Hybrid progeny; HS, Half-sibling; IBD, Identity-by-descent; MRAC, mean rate of ACs correctly identified; MRC, Mean rate of contaminants (correctly identified); MRCC, Mean rate of cross-contaminants (HSs/FCs) correctly identified; MRH, Mean rate of hybrids correctly identified; MRSP, Mean rate of SPs correctly identified; NIPALS, Non-linear iterative partial least squares; NGS, Next-generation-sequencing; P1/Parent 1, Female parent for simulated or real population; P2/Parent 2, Male parent for simulated or real population; PC, Principal component; PCA, Principal component analysis; PCR, Polymerase chain reaction; QTL, Quantitative trait loci; RAPD, Random amplified polymorphic DNA; SNP, Single nucleotide polymorphism; SP, Self-fertilization progeny of one of the parents; SSR, Simple sequence repeats.

based on pairwise Mendelian segregation tests, comparing individuals and generating different measures that account for the similarity between a sample and one of the parents or for a rate of unexpected genotypes in each sample considering the genotypes of both parents. Therefore, such genotype analyses (GAs) can be used to define what is not genotypically similar and consecutively represents an experimental contaminant. In this work, we propose to use GA measures for performing clustering analyses (CAs) and automatically identifying contaminants in forage grass biparental populations, grouping individuals based on GA similarity measures instead of their raw SNP data. Although CA of large SNP datasets has been extensively used to discover patterns in population relatedness and structure (Gori et al., 2016; Muniz et al., 2019; Yousefi-Mashouf et al., 2021), its use for parentage assignment is not common because of the non-specificity and constancy of the clusters, but has already been combined with previously described techniques for parentage and sibship inference in diploids (Ellis et al., 2018).

Instead of relying strictly on PCA for population analyses and *ad hoc* decisions (Deo et al., 2020; Zhang et al., 2020), we created an semi automated pipeline, combining GA and CA that allow us not only to precisely identify but also to list the types of contaminants in a biparental cross. For this purpose, we simulated several biparental progenies with contaminants to (1) identify dispersion patterns in a PCA biplot that can suggest the presence of contaminants, (2) create appropriate GA measures for contaminant identification in polyploid forage grass samples, generating scores for all individuals, and (3) integrate such scores in an automatic CA to separate the real hybrids from the contaminants. These steps led to the formulation of a unified methodology, which we applied to biparental progenies of three different species of tropical forage grasses: *Megathyrsus maximus* (Jacq.), syn. *Panicum maximum* Jacq. (B. K. Simon & S. W. L. Jacobs), *Urochloa decumbens* (Stapf), syn. *Brachiaria decumbens* Stapf (R. D. Webster) and *Urochloa humidicola* (Rendle), syn. *Brachiaria humidicola* (Rendle, Schweick) (Morrone and Zuloaga, 1992; Torres-González and Morton, 2005). The implemented pipeline was made available through a Shiny app and has a high potential to be employed in pre-breeding stages, as well as in genomic studies involving polyploid biparental progenies in general.

MATERIALS AND METHODS

The following sections describe the steps involved in the generation of real and simulated data and their use to propose a methodology for contaminant identification in biparental crosses. First, the genotyping and allele dosage estimation for biparental F_1 populations of three tropical forage species are presented (2.1, 2.2, and 2.3). Then, different biparental crossings are simulated (2.4). Finally, contaminant identification methodologies are applied to the simulated and real data (2.5, 2.6, 2.7, and 2.8).

Plant Material

Genotypic data were obtained from biparental F_1 progenies of *Urochloa humidicola* (a segmental allopolyploid, with 2n = 6x = 36), *Urochloa decumbens* (a segmental allopolyploid, with 2n = 4x = 36), and *Megathyrsus maximus* (an autopolyploid, with 2n = 4x = 32), three important species of tropical forage grasses used in the pastures of tropical and subtropical areas. All these intraspecific crossings were performed by the Brazilian Agricultural Research Corporation (Embrapa) Gado de Corte, located in Campo Grande, Mato Grosso do Sul, Brazil (20°27′S, 54°37′W, 530 m), and are part of the breeding programs of this research institution. Details about the crossing were described by Deo et al. (2020) for *M. maximus* and by Barrios et al. (2013) for *U. decumbens*. For *U. humidicola*, the crossings were manually performed in controlled crosses in greenhouses at Embrapa. Plants from the male genitor were cultivated in the field and pollen grains were collected in the day of the crossings or in the day before and stored overnight in a Petri plate in a refrigerator. Plants from the female genitor were cultivated in pots in the greenhouse and the inflorescences had the spikelets at anthesis removed with a tweezer, only those remaining spikelets were going to be opened in the next day. At the crossing day, the spikelets at anthesis were pollinated with the collected pollen grains and the inflorescences were covered with a paper bag and identified. After dehiscence, the F_1 seeds were collected and processed until germination in trays and then planted in the field in single plots.

The *U. humidicola* progeny consisted of 279 hybrids obtained from a cross between the sexual accession H031 (CIAT 26146) and the apomictic cultivar *U. humidicola* cv. BRS Tupi, as described by Vigna et al. (2016). The cross between *U. decumbens* D24/27 (sexual diploid accession tetraploidized by colchicine) and the apomict *U. decumbens* cv. Basilisk generated a progeny with 239 hybrids (Ferreira et al., 2019). Finally, the progeny of *M. maximus* included 136 hybrids originating from a cross between the sexual genotype S10 and *M. maximus* cv. Mombaça (apomictic parent) (Deo et al., 2020). The apomixis in the cultivars BRS Tupi, Basilisk, and Mombaça is of the pseudogamic aposporic types.

Genotyping-By-Sequencing Library Preparation

Genotyping-by-sequencing (GBS) libraries of the *U. decumbens* and *M. maximus* progenies were built and sequenced as described by Ferreira et al. (2019) and Deo et al. (2020), respectively. For the progeny of *U. humidicola*, DNA was extracted following Vigna et al. (2016), and the GBS libraries were built according to Poland et al. (2012), containing five replicates for each parent and one for each hybrid. Genomic DNA (210 ng of DNA per individual) was digested using a combination of a rarely cutting enzyme (PstI) and a frequently cutting enzyme (MspI). Subsequently, the libraries were sequenced as 150-bp single-end reads using the High Output v2 Kit (Illumina, San Diego, CA, USA) in the NextSeq 500 platform (Illumina, San Diego, CA, USA). The quality of the resulting sequence data was evaluated using the FastQC toolkit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

GBS-SNP Discovery and Allele Dosage

We analyzed the raw data of the three biparental progenies using the Tassel-GBS pipeline (Glaubitz et al., 2014) modified for polyploids (Pereira et al., 2018b), which considers the original read depths for each SNP allele. The Bowtie2 algorithm version 2.1 (Langmead and Salzberg, 2012) was used to align the reads of the *Urochloa* spp. and *M. maximus* against the reference genomes of *Setaria viridis* v1.0 and *Panicum virgatum* v1.0, respectively, since the reference genomes are not available for the species under study. In this stage, a limit of 20 dynamic programming problems (D), a maximum of four times to align a read (R), and a very-sensitive-local argument were considered. Both genomes used as references were retrieved from the Phytozome database (Goodstein et al., 2012).

For quality purposes, the SNPs were submitted to a filtering procedure using VCFtools (Danecek et al., 2011), with the following parameters: maximum number of alleles of two (to include only bi-allelic loci), maximum missing data per marker of 25%, and minimum read depth per individual of 20 reads for *M. maximus* and *U. decumbens,* and 40 reads for *U. humidicola*. Due to the polyploid nature of the species, a high sequence depth is required to identify the genotypic class accurately (Cappai et al., 2020; Ferrão et al., 2020; Mollinari et al., 2020), and even higher values were used for *U. humidicola* because it is a hexaploid. Finally, the Updog R package (Gerard et al., 2018) was used to estimate the allele dosage of each SNP locus, with a fixed ploidy parameter of four for *M. maximus* and *U. decumbens*, and six for *U. humidicola*. The flexdog function was used with the "f1" population model for the three populations. The posterior proportion of mis-genotyped individuals (prop_mis) was set at six different values (0.05, 0.1, 0.15, 0.20, 0.25, and 0.3) for *M. maximus* and *U. decumbens*, aiming to compare the rates of the tetraploid dosages in the parents and assess the influence of the number and quality of the markers in further analysis. For the hexaploid population of *U. humidicola*, prop_mis was set at 0.2.

The genotyping data were organized into marker matrices $\mathbf{M}_{(n \times m)}$, where n denotes the samples, m denotes the markers, and the allele dosage genotypes are encoded as 0, 1, 2, 3, 4, 5, or 6 for nulliplex, simplex, duplex, triplex, quadruplex, quintuplex, and hexaplex data, respectively.

Simulated Data

Biparental F_1 populations were simulated using the PedigreeSim R package (Voorrips and Maliepaard, 2012), a software package that simulates meiosis and uses this information to create cross populations in tetraploid species. To create the linkage map required by PedigreeSim, the previously published map for *M. maximus* (Deo et al., 2020) was used as a model to estimate the main parameters, such as the number and size of chromosomes, density, gap regions, and centromere position. Eight chromosomes with sizes between 90 and 120 centimorgans (cM) and 600–900 SNP markers per chromosome, both randomly sampled, were created. In addition, the markers were considered to be distributed along the chromosomes with a minimum distance between adjacent markers of 0.1 cM. The centromere position was sampled between 10 and 50 cM, preferential pairing was set to zero, and the quadrivalent fraction was set for natural pairing. In this case, the fraction of quadrivalents arises automatically from the pairing process at the telomeres. Other options of the software were kept as default. All these files were created using R software (R Core Team, 2020).

To perform the crosses, four parents (P1, P2, P3, and P4) were created, and the genotypes of these parents were simulated based on the rate of allele dosages of parents genotyped in real biparental progenies: P1 and P2 from *M. maximus* (Deo et al., 2020) and P3 and P4 from *U. decumbens* (Ferreira et al., 2019). Considering these rates, the haplotypes of each of the four homologous chromosomes were randomly created for each parent. The simulated crosses between these parents were based on the following combinations: P1 × P2, P1 × P1 (self-fertilization), P1 × P3, P1 × P4, and P3 × P4, with a progeny size of 200.

The results of the simulated crosses were converted into marker matrices (**M**), and all subsequent manipulations were performed using R software (R Core Team, 2020). To insert genotyping errors, 5% of the genotypes were randomly replaced by other genotype values with equal probability, and between 1 and 5% of the genotypes of each marker were removed to simulate the missing data (NAs). Clonal individuals were simulated by duplicating the genotype of a parent, and errors and NAs were inserted as described above.

Using the tetraploid populations created in the PedigreeSim software, four scenarios were established to analyze the different types of contaminants that could occur in biparental populations of tropical forage grasses. The first two scenarios were represented by contaminants resulting from the reproductive mode of parents, which can reproduce by (1) apomictic clones (ACs), or (2) self-fertilization progeny of one of the parents (SPs), resulting in segregating individuals. The last two scenarios represent (3) cross-contamination, that is, when fertilization occurs by foreign pollen, resulting in half-siblings (HSs), or (4) when physical mixtures occur during seed handling, resulting in full contaminants (FCs). In each of the four possible scenarios, the size of the base population was 200 hybrids (HPs), and the HPs were progressively replaced by contaminants until 25% of the samples were contaminants. In addition, to investigate a joint scenario with four parents (P1, P2, P3, and P4) with AC and SP contamination, a population of 1,200 individuals (200 P1-ACs, 200 P1-SPs, 200 HPs from P1 × P2, 200 HPs from P1 × P3, 200 HPs from P1 × P4, and 200 HPs from P3 × P4) was created. These described populations were constructed to investigate how contaminants influence principal component analysis (PCA) scatter plot dispersion patterns.

For the evaluation of the proposed contaminant identification method, 6,000 populations were simulated. Each one was composed of 200 individuals with a random number of contaminants, ranging between 1 and 50 and distributed per contaminant type considering random probabilities between 0.1 and 0.8. The populations were divided into six equal size groups according to the number of genotyped markers. Considering n as the total simulated markers, the groups were composed of: n/2, n/4, n/8, n/16, n/32, and n/64 markers. For each population, the subset of markers used was randomly sampled from the total simulated markers. Furthermore, a biparental population with

200 individuals [150 HP (P1 × P2), 10 AC (P1), 10 SP (P1), 10 HS1 (P1 × P3), 10 HS2 (P1 × P4), and 10 FC (P3 × P4)] was also simulated to exemplify the use of GA and CA in the contaminant identification.

Principal Component Analysis

Principal component analyses were performed by the R package pcaMethods (Stacklies et al., 2007) using the non-linear estimation by iterative partial least squares (NIPALS) algorithm (Wold and Krishnaiah, 1966) to calculate the eigenvalues with missing data imputation. Given a matrix $X_{m,n}$ representing the n random variables (herein SNPs) across m individuals, this analysis transforms X by multiplying it by the orthogonal eigenvectors, generating a matrix $X_{m,p}$ of new p variables [the principal components (PCs)] with specific mathematical properties (Maćkiewicz and Ratajczak, 1993). The ggplot2 R package (Wickham and Chang, 2016) was used to construct scatter plots of the first two PCs. These graphical visualizations were used to identify clustering patterns that may be associated with contaminants in the progeny.

Genotypic Analysis

The term genotypic analysis (GA) is employed here to refer to an analysis that evaluates all the samples of a progeny considering what is genotypically expected for a contaminant. Three different measures were created for evaluating the samples: GA-I for AC identification and GA-II for SP identification, both accounting for a similarity rate between the sample and one of the parents (computed separately for each), and GA-III, accounting for a rate of unexpected genotypes in each sample considering the genotypes of both the parents, enabling the identification of half-siblings (HSs) and full contaminants (FCs) in the progeny.

To investigate whether an individual x is an AC of a parent p, the GA-I scores were calculated using the marker matrix **M** with n rows (individuals) and m columns (markers). Then, the similarity between x and p was the proportion of allele dosages in $M_{x,i}$ that satisfied the condition, $M_{x,i} = M_{p,i}$ with $1 \leq i \geq m$. This measure is based on the presumption that, given Mendel's law, each individual inherits genetic material from its parents (Mendel, 1866; Miko, 2008). However, if one of the parents reproduces through apomixis, a genetically identical progeny is produced (Hand and Koltunow, 2014). Therefore, in a suite of Mendelian loci, if a putative individual shows a high similarity (GA-I close to 1.00) with one of the parents, it can be considered a clone of this parent.

In the case of SP samples, the GA-II scores were calculated by computing the similarity between the progeny samples and the parents considering only nulliplex allele dosages; i.e., for a parent p and an individual x, GA-II was the proportion of allele dosages in $M_{x,i}$ (with $1 \leq i \leq m$) that satisfied $M_{x,i} = M_{p,i} = 0$. If a parent reproduces through self-fertilization, Mendelian segregation should be observed. Using a tetraploid species as an example, a parent with the genotype AABB at a specific locus, after self-fertilization, would generate a progeny with genotypes in all possible doses (AAAA, AAAB, AABB, ABBB, and BBBB) (Hackett et al., 2013). However, if we focus only on the markers for which the parent had a nulliplex genotype (AAAA), the progeny produced would be genetically identical to the parent at those loci. Thus, GA-II computes a similarity rate between the sample and the parent considering only those markers; in this situation, it was expected that SP contaminants would present GA-II scores close to 1.00.

For the HSs and FCs, the GA-III term calculates the rate of unexpected allele dosages for the progeny individuals across all the markers. Considering the combination of gametes for parent *p1* and *p2* at a SNP i with $1 \leq i \geq m$, the GA-III of an individual x is the proportion of unexpected allele dosages for its set of markers. Considering the allele dosage of each parent at each marker, it is possible to define which dosage is not expected in their progeny. For example, if one parent is nulliplex (AAAA) for a marker and the other is simplex (AAAB), the gametes produced by the nulliplex are all AA, and for the simplex, they can be AA or AB (Hackett et al., 2013). Their combination can produce a progeny with only nulliplex or simplex for this marker, and the presence of other dosage types is an evidence for the fact that this individual may not belong to the cross. In this way, for all markers, GA-III tested whether the genotype of this sample was expected considering both parental genotypes, computing a rate of unexpected genotypes for each sample (**Supplementary Table 1**). In this analysis, it was expected that HSs and FCs would show higher GA-III scores than HPs, enabling their identification.

Clustering Analysis

The contaminant identification process is based on a clustering analysis (CA) performed using an average linkage hierarchical clustering approach with R software (R Core Team, 2020). Considering the GAs calculated, pairwise Euclidean distances between these values were calculated across the progeny and were used to obtain 27 different clustering indexes (**Supplementary Table 2**) with numbers of clusters varying from 2 to 15, implemented in the R package NbClust (Charrad et al., 2014). The package automatically calculates the indexes, defines the best clustering scheme based on majority rule (i.e., most indicated number of clusters), and classifies the samples into clusters.

Contaminant identification was then performed with the best clustering configuration scheme. Individuals in groups separated from most of the population were considered contaminants and classified according to the following rules applied to the GA measures within these clusters: (1) individuals within a cluster having the greatest GA-I values for one parent were considered ACs when the median of these measures was >0.75; (2) individuals within a cluster with the median GA-II values for one parent >0.75 and not belonging to Group (1) were considered SPs; and (3) individuals not belonging to Groups (1) and (2) and with a within-group minimum GA-III value greater than the maximum measure of the group with the minimum GA-III median were considered HSs/FCs. Therefore, in a simplified and automated process with only the threshold of GA measures as an *ad hoc* decision, we obtained the final data set with parents and their corresponding true hybrids.

The method was evaluated on a set of 1,000 simulated populations, computing the following metrics considering the most indicated clustering scheme: mean rate of hybrids correctly identified (MRH), mean rate of contaminants correctly identified (MRC), mean rate of apomictic clones correctly identified (MRAC), mean rate of SPs correctly identified (MRSP), and mean rate of cross-contaminants (HSs/FCs) correctly identified (MRCC). Furthermore, the same metrics were computed considering the three most indicated clustering schemes; in this situation, the highest rate among the three schemes for each simulated population was used to calculate the mean.

Contaminant Identification in Real Data

Combining GA, CA, and PCA, we established a four-step contaminant identification methodology as follows, and applied it to the real populations (**Figure 1,** Part III Contaminant Identification):

1. Construction of a scatter plot with the first PCs from a PCA performed with the SNP data organized according to allele dosage, looking for evidence of contaminants in the population. When no contaminants are detected, simulated clones (25% of the population) from one of the parents can be artificially added to the population, changing the dispersion pattern of individuals and inducing contaminant separation;
2. Calculation of five different GA measures for each individual (GA-I and GA-II, considering Parents 1 and 2, respectively, and GA-III). GAI and GAII were calculated in the same way for all ploidy, but for hexaploid progeny, GAIII was adapted considering its respective segregation;
3. Performance of CA using the GA data to identify clusters in the population;
4. Visual inspection of the histograms, to classify the clusters according to the GA value differences described in section Clustering analysis. This step is done in a sequential process, in which the first ACs are identified and removed, then SPs, and lastly, HSs/FCs are identified and removed.
5. Recalculation of PCA to confirm in the biplot the expected dispersion pattern of a population with no contaminants.

All these procedures were unified in polyCID Shiny app, created using R software together with the libraries shiny (Chang et al., 2021), shinydashboard (https://cran.r-project.org/web/packages/shinydashboard/index.html), and DT (https://cran.r-project.org/web/packages/DT/index.html). polyCID is an R-Shiny Web graphical user interface (GUI) that combines all the described analyses in a simple way and provides a user-friendly tool, fully available and documented at https://github.com/lagmunicamp/polycid.

RESULTS

The results are organized as follows. First, the genotyping and allele dosage information for the three biparental progenies of the tropical forage species is presented (3.1). Next, the application of principal component analysis (PCA) to the simulated data is shown (3.2). Then, the use of GA and CA in contaminant identification in the simulated data is described (3.3), and finally, the results obtained from the contaminant identification in real data are presented (3.4). Furthermore, for simulated and real populations, P1/Parent 1 is the female parent and P2/Parent 2 is the male parent.

GBS-SNP Discovery and Allele Dosage Estimation

After SNP calling using the Tassel- genotyping-by-sequencing (GBS) pipeline (Glaubitz et al., 2014) modified for polyploids (Pereira et al., 2018b), filtering markers for missing data (NAs) and read depth with VCFtools (Danecek et al., 2011), we obtained 15,279 SNP markers for *Urochloa humidicola*, 8,036 for *Urochloa decumbens*, and 6,337 for *Megathyrsus maximus*. Three individuals ("Bh181," "Bh226," and "Bh245") of the *U. humidicola* progeny were removed because of the high content of missing data (>44%).

The Updog R package (Gerard et al., 2018) was used to estimate the allele dosage for the SNP loci identified in each progeny. For the six values of prop_mis used (0.05, 0.10, 0.15, 0.20, 0.25, and 0.30), 4,003, 5,179, 5,863, 6,068, 6,161, and 6,215 markers were obtained for *M. maximus* and 1,195, 1,745, 2,303, 2,862, 3,165, and 3,243 markers were obtained for *U. decumbens*, respectively, while 7,253 markers were obtained for *U. humidicola* using a prop_mis value of 0.20.

Principal Component Analysis

Marker matrices of each simulated scenario were used to perform a PCA, looking for patterns spanned by the first two PCs that can aid in the identification of contaminant samples. Details of these simulated scenarios, such as the size of chromosomes, position of centromeres, and the number of markers can be found in **Supplementary Tables 3-5**. The PCA scatter plot of the simulated population without contaminants had hybrids and parents distributed with no apparent clustering patterns among the individuals, with 4% of variance explained by the first two principal components (PCs) (**Supplementary Figure 1**).

The same biplot distribution was observed when only one contaminant was added to the biparental population, i.e., an apomictic clone (AC) (**Figure 2A**), self-fertilization progeny (SP) (**Supplementary Figure 2**), half-sibling (HS) (**Supplementary Figure 3**), or full contaminant (FC) (**Supplementary Figure 4**). In these situations, the genetic variation related to contamination could not be detected by the first components and therefore assessed by visual inspection. When the number of contaminants was progressively increased in the scenarios, the dispersion pattern of the scatter plots began to reveal the separation of the contaminants from the hybrids. For the scenarios, five (**Figure 2B**), four (**Supplementary Figure 5**), six (**Supplementary Figure 6**), and three (**Supplementary Figure 7**) contaminants were necessary to clearly visualize the separation. Adding these contaminants changed the source of variation in the first PCs, which changed little (<0.2%). As the number of contaminants increased to 25% of the population, it was possible to observe in the PCA biplot that the hybrids were projected between the parents, the ACs/SPs were closer to the parent of origin (**Figure 2C** and **Supplementary Figure 8**), the HSs/FCs formed separated groups

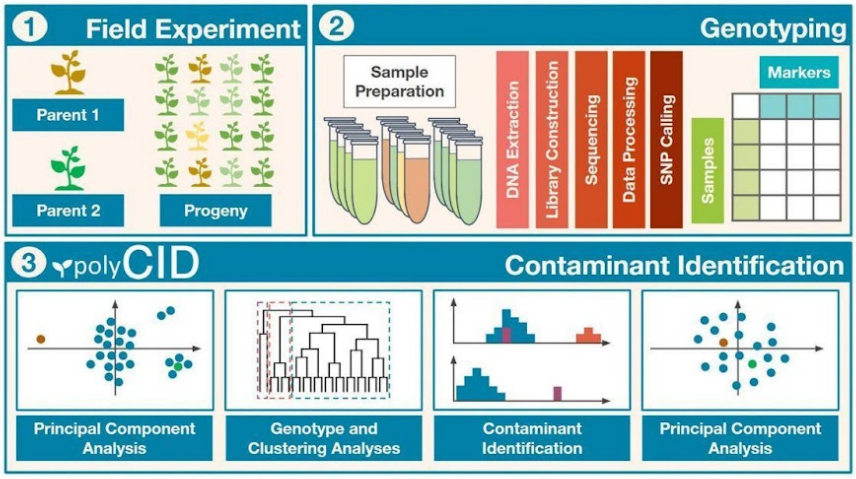


FIGURE 1 | Workflow of the contaminant identification process. (I) Field experiment to obtain the biparental progeny. (II) Population genotyping and bioinformatic analyses. (III) Methodology with principal component analysis (PCA), genotype analysis (GA), and clustering analysis (CA) to identify and remove contaminants.

(**Supplementary Figures 9, 10**), and the sums of variance in the first two PCs changed to values between 10.8 and 17%.

Considering that the analysis of the first two PCs through a PCA scatter plot could not reveal contaminants at low frequencies, biparental populations with 199 HPs and one contaminant were simulated, and 50 ACs (25%) of one of the parents were included. This unique contaminant may be an SP (Scenario 2), AC (Scenario 3), or FC (Scenario 4). As a result, we observed that the inclusion of these simulated clones, which occurs in real populations, changed the sums of variance in the first two PCs to a value of ∼10.3% and increased the dispersion pattern in the PCA scatter plot, leading to the formation of different subgroups that allowed for the visualization of SP or HS contaminants (**Supplementary Figures 11**, **12**). On the other hand, FCs and HPs were grouped together and could not be identified visually in the scatter plot (**Supplementary Figure 13**).

Finally, when simulating an open pollination population with four different possible parents, the biplot of the PCs was able to provide visual separation of the different progenies. It was possible to identify each cross since HPs formed a subgroup between their respective parents. In addition, the AC and SP contaminants grouped together with their parents (**Figure 3**).

Semi Automatic Contaminant Identification

To look for the patterns in contaminant genotype analyses (GA) measures, the three described GAs were calculated in a simulated population of 200 samples composed of 150 hybrids (HPs) and 50 contaminants (10 ACs, 10 SPs, 10 HS1s, 10 HS2s, and 10 FCs); thus, five different values for each putative hybrid were generated. We analyzed how GA histograms behave for each type of contamination. In **Figure 4A**, AC individuals formed a group with the greatest GA-I scores for Parent 1 (red circle) and were removed to analyze the other histograms. In the same way, the GA-II histogram (**Figure 4B**) showed that the SP samples had the highest scores for Parent 1 (red circle), and these individuals were also removed. We believe that mutations, missing data, and sequencing/genotyping errors are events responsible for the differences between the expected scores (pretty close to 1) and the observed (about 0.8 to 0.9). Finally, in **Figure 4C**, the GA-III histogram showed that the HP samples had lower scores than the HS and FC contaminants. For a correct hybrid definition, these contaminants were also removed to generate a proper hybrid data set.

By using the idea underlying these visual histogram inspections, we implemented on GA measures a clustering-based approach for automatic contaminant identification. The established methodology employs a single hierarchical clustering algorithm on a different range of cluster numbers, defining the best clustering scheme with 27 clustering indexes (**Supplementary Table 2**). Employing this approach on the simulated population previously described, we observed that the defined CA separated the samples into six different groups: one for the HP and five for each type of contaminant, exactly corresponding to the simulated categories (**Figure 4**). Therefore, we evaluated its accuracy on additional 6,000 simulated populations and checked its appropriateness using six set sizes of markers in a broad range of possible contamination scenarios. The sets were of the following sizes: 2,758, 1,379, 689,

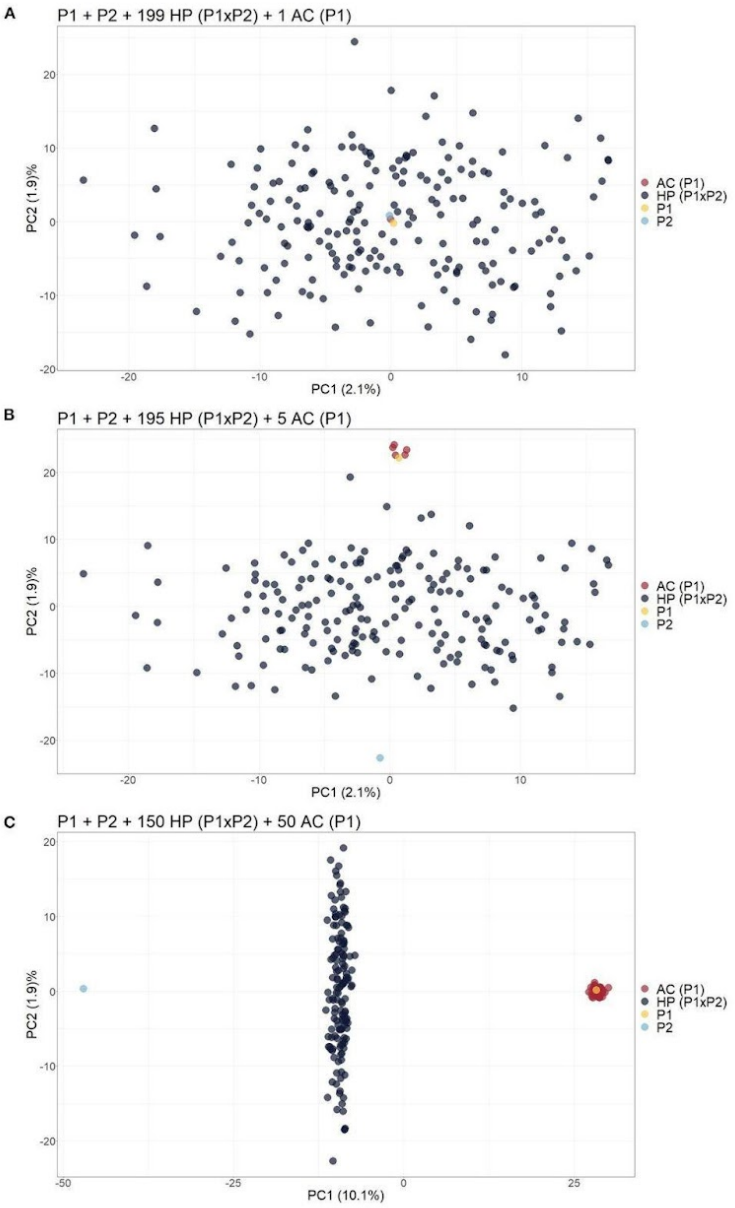


FIGURE 2 | Principal component analysis (PCA)-based scatter plots showing the change in dispersion pattern as the apomictic clone (AC) of P1 increases in frequency in the simulated population. **(A)** Progeny with 199 hybrids (HPs) and 1 AC; **(B)** Progeny with 195 HPs and 5 ACs; **(C)** Progeny with 150 HPs and 50 ACs. The axis represents the first and second principal components, explaining 2.1% and 1.9% of the variance, respectively, for **(A)**, 2.1% and 1.9% for **(B)** and 10.1% and 1.9% for **(C)**.

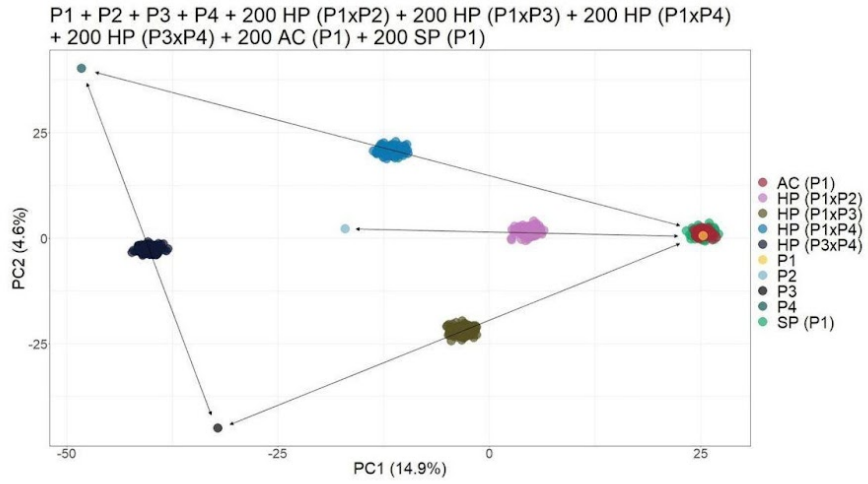


FIGURE 3 | Principal component analysis (PCA)-based scatter plot showing the simulated population composed of four parents (P1, P2, P3, and P4), 200 apomictic clones (ACs) (P1), 200 self-fertilization progenies -(SPs) (P1), 200 half-siblings (HSs) (P1 × P2), and 200 hybrids (HPs) (P1 × P3, P1 × P4, and P3 × 43). The axis represents the first and second principal components, with 14.9% and 4.6% of the variance explained, respectively.

344, 172, and 86 markers. For each marker's set size of the, 1,000 populations were simulated sampling markers from a total of 5,516.

The mean rate of hybrids (MRHs) correctly identified was 100% for all sets of markers, except for the smallest one (86 markers), which had a slight reduction. On the other hand, the mean rate of contaminants (MRC) was around 90% for the three largest sets (2,758, 1,379, and 689 markers), which started decreasing, reaching the value of 48% in the smallest one (**Figure 5A**). It was possible to observe that the methodology failed only for the smallest set (86 markers), in which a true hybrid was considered a contaminant, but it rarely discarded reliable data. Regarding the contaminant classification and considering the largest sets of markers, 69, 72, and 84% were observed for the mean rate of apomictic clone (MRAC), mean rate of self-fertilization progeny (MRSP), and mean rate of cross-contaminants correctly identified (MRCC), respectively. Then, we observed a slight reduction in the 689 markers set, which showed values of 63% (MRAC), 66% (MRSP), and 76% (MRCC), followed by 49% (MRAC), 50% (MRSP), and 15% (MRCC) in the smallest set (**Figure 5A**).

In the function of these modest values, we also evaluated the method efficiency on the second and third best clustering configurations identified by the calculated indexes. Considering the best group separation within these three possible configurations also noticed in GA histograms, we achieved a performance improvement in all set markers, reaching an approximate accuracy of 100% in the three largest sets for all types of samples. Next, we observed a slight reduction (to values higher than 90%) in the set of 344 markers, and more prominent reductions in the two smallest sets, reaching the values of 49% (MRAC), 85% (MRSP), and 40% (MRCC) (**Figure 5B**). These findings suggest that, in real applications, such evaluations in these three cluster configurations may represent an additional step for increasing the method's reliability.

Contaminant Identification in Real Populations

After investigating with simulated populations, the proposed methodology was applied to real genotyping data from three biparental F_1 progenies of tropical forage grasses. For the progeny of *M. maximus*, the PCA plots with different values of prop_mis showed similar sample dispersion patterns and a reduction in variance explained by the first two PCs from 10.1 to 7.5% as the number of markers increased. Therefore, the dataset obtained with the default value of prop_mis = 0.20 was used in the further analysis. Even though there was no clear group formation in the PCA scatter plot, the pattern of parents on the opposite sides and HPs grouped between them provided evidence for the presence of contaminants (**Supplementary Figure 14A**). Similarly, in the PCA with simulated ACs, these two individuals remained close to Parent 2 (*M. maximus* cv. Mombaça) (**Supplementary Figure 14B**).

The clustering analysis (CA) with GA revealed two clusters in the *M. maximus* progeny, with 134 and two samples. The GA-I histogram for Parent 2 (*M. maximus* cv. Mombaça) showed that the cluster with two samples had high scores and must be considered putative ACs of Parent 2 (*M. maximus* cv. Mombaça) (**Supplementary Figure 15A**). On the other hand, GA II and

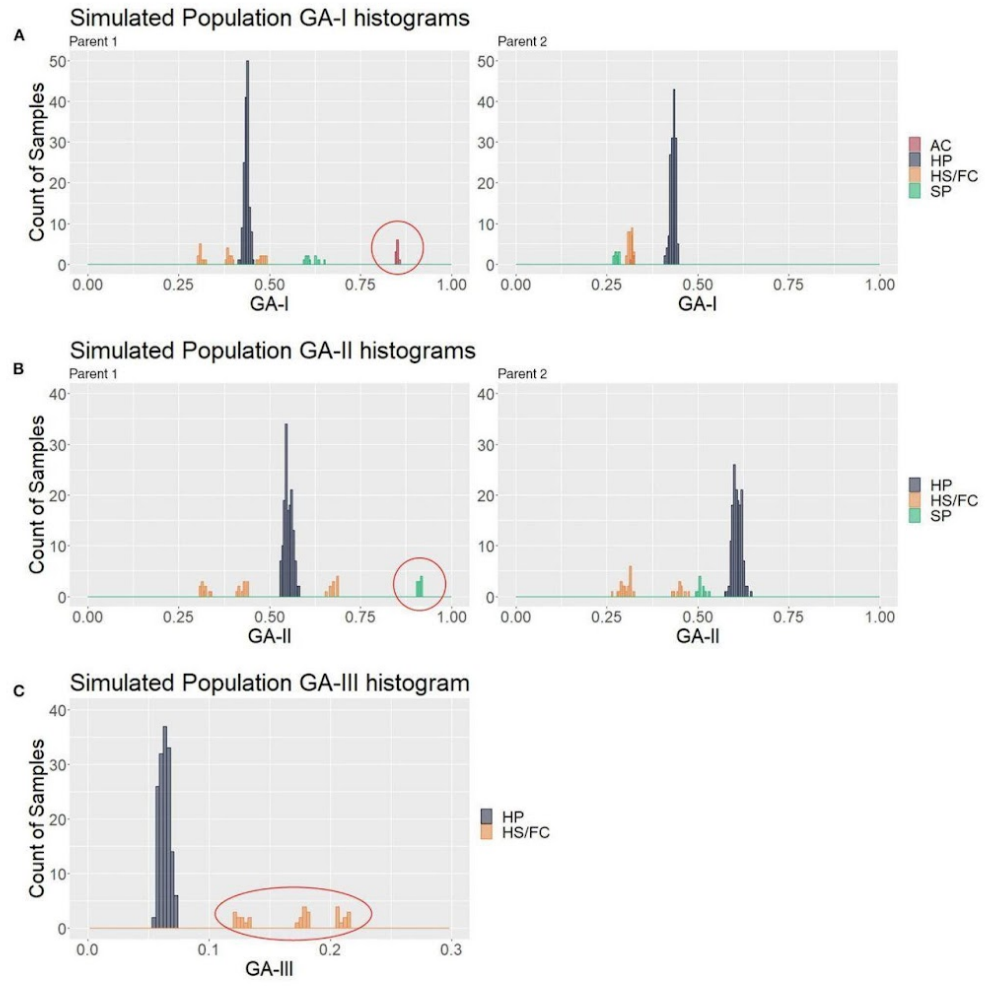


FIGURE 4 | Genotype analysis (GA)-based histograms of the simulated population with samples classified in the clustering analyses (CAs). **(A–C)** show the results from GA-I, GA-II, and GA-III, respectively. Red circles highlight the identified contaminants, i.e., apomictic clones (AC), self-progeny (SP), and half siblings/full contaminants (HSs/FCs), in contrast to the hybrid progeny (HP).

III showed no evidence for other types of contaminants in the *M. maximus* progeny (**Supplementary Figures 15B,C**). The exclusion of these two individuals resulted in a PCA scatter plot with the expected pattern (**Supplementary Figure 14C**).

For the progeny of *U. decumbens*, the PCA biplots for the different values of prop_mis showed different sample dispersion patterns (data not shown). As this is a very intuitive measure for the quality of SNPs when estimating allele dosage (Gerard et al., 2018), we chose to be conservative and used the most restrictive filter, 0.05, ensuring the selection of markers with high quality. The first PCA scatter plot showed strong evidence of contaminants in the population (**Figure 6A**). The algorithm found three clusters with 184, 49, and 3 samples. In the GA histograms, both Clusters 2 and 3 had high GA-I scores for Parent 2 (*U. decumbens* cv. Basilisk), providing evidence that those samples were putative ACs of this parent (**Figure 7A**).

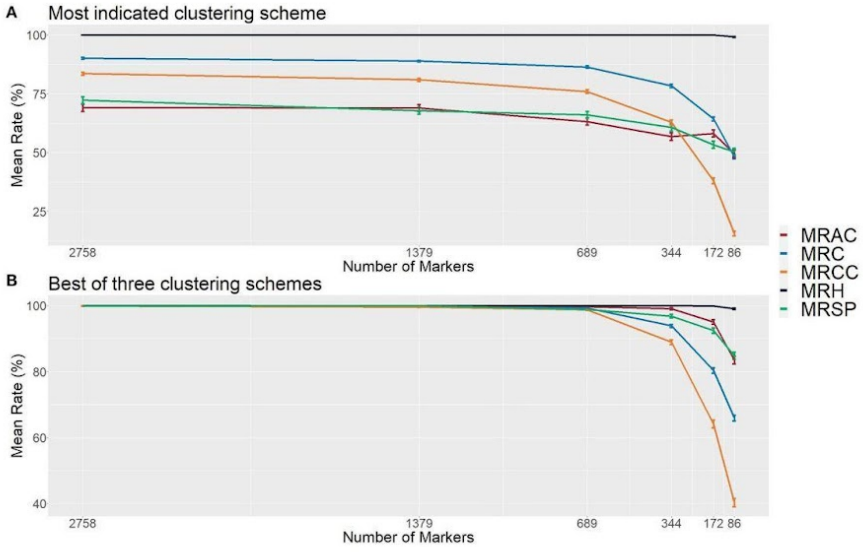


FIGURE 5 | Evaluation metrics of the proposed methodology considering six different set sizes of makers in simulated populations (1,000 populations for each set). **(A)** The most indicated clustering scheme, and **(B)** the best out of the first three most indicated schemes found with the 27 employed clustering indexes. The indicated rates are the mean rate of hybrids correctly identified (MRH), the mean rate of contaminants correctly identified (MRC), the mean rate of apomictic clones correctly identified (MRAC), the mean rate of self-fertilized progeny correctly identified (MRSP), and the mean rate of cross-contaminants (HSs/FCs) correctly identified (MRCC).

The other GA histograms showed no clear evidence of other contaminants (**Figures 7B,C**), except two individuals that could be considered suspicious in GA-II. In this case, we followed the clustering results and did not consider these individuals as contaminants. But this is an *ad hoc* decision, so the user can choose to be conservative and remove outliers. Once again, after the elimination of these ACs, the PCA scatter plot showed the expected pattern for progeny without contaminants (**Figure 6B**).

For the hexaploid biparental population of *U. humidicola*, the scatter plot of the first PCs showed strong evidence for the presence of AC and/or SP contaminants (**Supplementary Figure 16A**). The clustering analysis of the GA scores separated the progeny into two clusters with 211 and 65 samples. The histogram of GA-I for Parent 1 (*U. humidicola* H031) showed that the cluster with 65 samples had scores close to 1.0 (**Supplementary Figure 17A**), representing putative ACs of the respective parent. GA-II and GA-III showed no evidence of contaminants (**Supplementary Figures 17B,C**). Finally, the PCA without the previously identified contaminants also showed the expected pattern for progeny without contaminants (**Supplementary Figure 16B**).

The PolyCID Shiny App

Finally, we implemented the polyCID Shiny app, a Web graphical user interface (GUI) that provides all previously described analyses in a user-friendly tool that allows users to identify contaminants in biparental progeny in a simple way. The polyCID is completely R based, easy to install and presents a graphical interface designed for non-expert users, with several functions for interactive visualization of the results. The package accepts SNP data in the form of marker matrices with allele dosage information, loads this information, and performs the four-step contaminant identification methodology, as described in section Contaminant identification in real data. The Shiny-based GUI is included in the package as a standalone application, available at https://github.com/lagmunicamp/polycid.

DISCUSSION

Experimental populations used in the breeding programs are usually derived from a controlled cross between two or more parents, but depending on the field experiment, the species analyzed, and its reproductive biology, individuals may be generated from the mixtures of seeds, foreign pollination during open pollinated crosses, self-fertilization, or apomixis by one of the parents during the crosses. The non-identification of these contaminant individuals can interfere not only in the selection cycles of breeding programs (Telfer et al., 2015) but also in the studies of genetic diversity (Ji et al., 2013), population structure (Alam et al., 2018), linkage mapping (Deo et al., 2020), and association mapping (Laucou et al., 2018), since it can generate biased results.

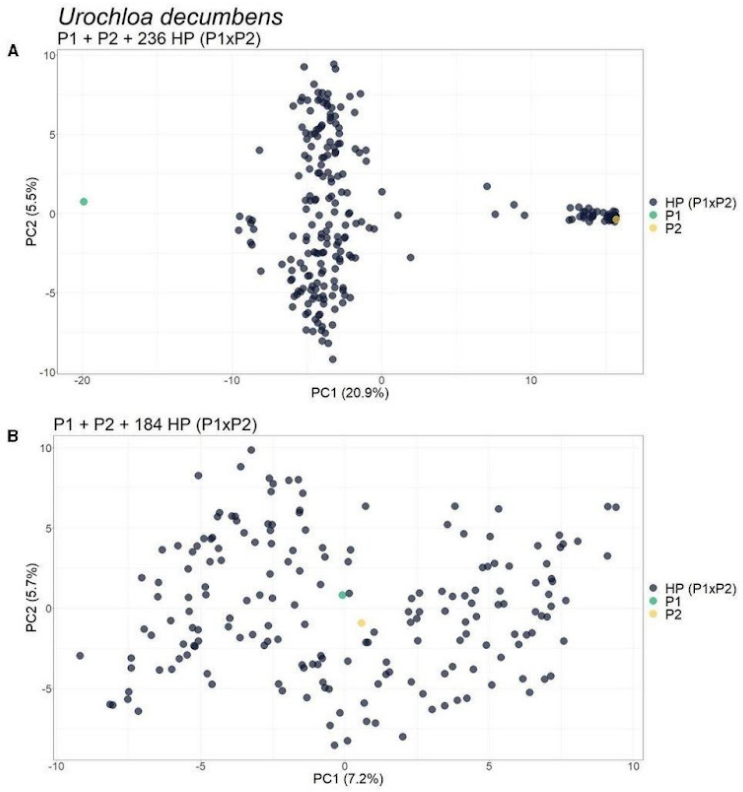


FIGURE 6 | Principal component analysis (PCA)-scatter plots showing the *Urochloa decumbens* progeny for the set of single nucleotide polymorphism (SNP) markers filtered by 0.05 for *prop_mis*. **(A)** Original population composed of two parents (P1 and P2) and their progeny of 236 hybrids [HP (P1 × P2)]; **(B)** Population without the 52 apomictic clones (ACs) identified. The axis represents the first and second principal components, explaining 20.9 and 5.5%, respectively, of the variance for **(A)** and 7.2 and 5.7% for **(B)**.

In most available studies, the identification of contaminants involved the use of microsatellites and morphological markers, but this strategy can be costly and time-consuming (Santos et al., 2014; Jha et al., 2016; Zhao et al., 2017; Patella et al., 2019), especially for polyploid species. In these cases, progeny evaluation is often performed using few microsatellite markers in polyacrylamide gels, and frequently, other analyses are needed, such as genetic distance analysis (Santos et al., 2014). The low number of microsatellite markers, usually in the tens or hundreds, may prevent the identification of contaminants. Considering this scenario, we used genotyping-by-sequencing (GBS) (Poland et al., 2012) to identify thousands of single nucleotide polymorphism (SNP) markers with allele dosage information and to propose a methodology that facilitates the identification of contaminants in biparental crossbreeding of polyploid species. Despite the emergence of several pipelines for the analysis of GBS data in polyploids, the application of these markers in parentage analysis is still little explored.

Currently, several software packages can deal with genetic data to assign paternity or parentage to individuals at the diploid level using microsatellite or SNP markers and the likelihood-based or Bayesian methods (Kalinowski et al., 2007; Jones and Wang, 2010; Anderson, 2012; Huisman, 2017), in addition to other approaches (Hayes, 2011; Heaton et al., 2014; Grashei et al., 2018; Whalen et al., 2019). For polyploids, the few resources available are limited to microsatellite data (Spielmann et al., 2015; Zwart et al., 2016). Another common approach in polyploids is to estimate pairwise relatedness (r) (Huang et al., 2015; Amadeu et al., 2020), for example, to assess the relationships between parents, offspring, full-sibs and half-sibs in progenies.

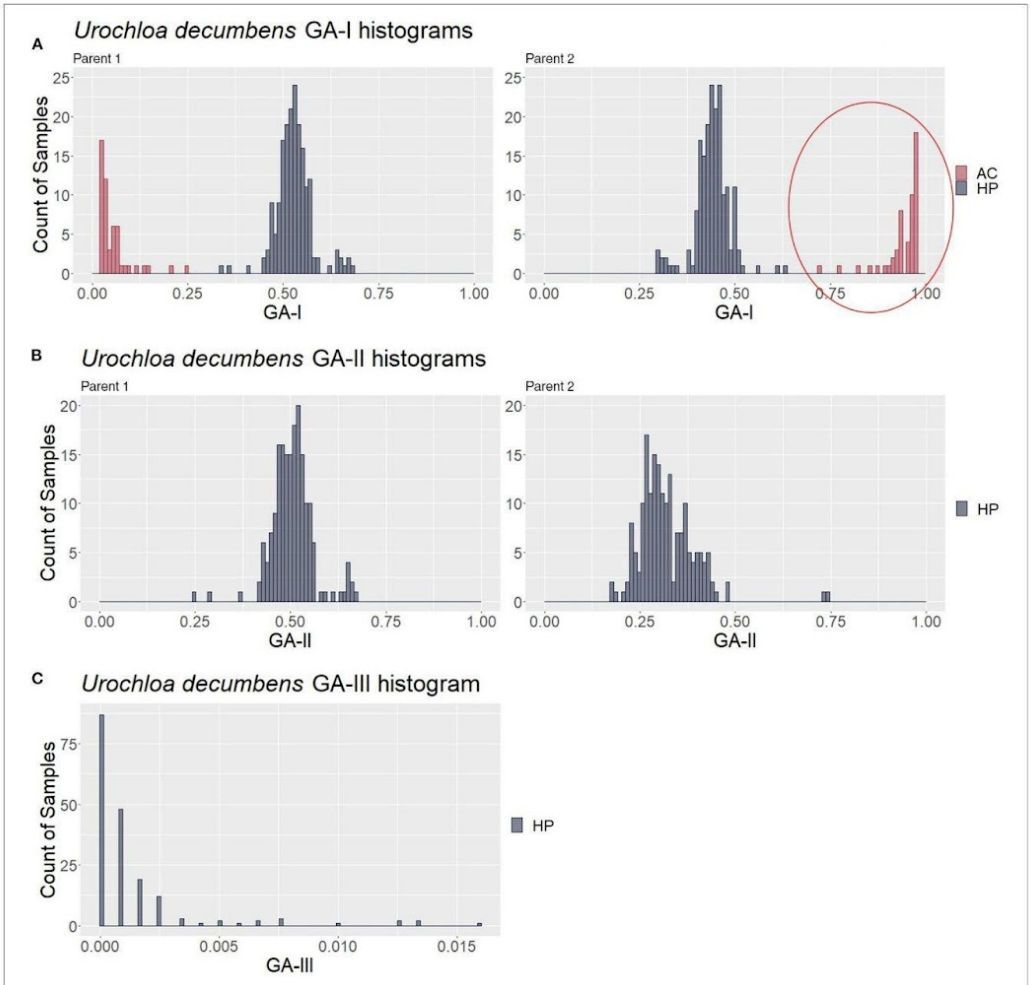


FIGURE 7 | Genotype analysis (GA)-based histograms of *Urochloa decumbens* with samples classified in the clustering analysis (CA). **(A–C)** show the results from GA-I, GA-II, and GA-III, respectively. The red circle highlights the identified apomictic clone (AC) contaminants in contrast to the hybrid progeny (HP).

In addition, identity-by-descent (IBD) has been used to assess the probabilities of inheritance of particular combinations of parental haplotypes (Zheng et al., 2016), which are also quite difficult to evaluate in polyploid progenies. For both approaches, the parameters are estimated in a pairwise manner, and the results are evaluated for each pair, making the analysis even more laborious.

For breeding programs that make use of biparental crosses, the major challenge is to precisely identify whether there are contaminating individuals to be excluded from the progeny (Martuscello et al., 2009; Ma and Amos, 2012; Santos et al., 2014; Subashini et al., 2014; Simeão et al., 2016b; Matias et al., 2019; Deo et al., 2020). In this context, no studies have proposed a unified pipeline focused on identifying the most common contaminants in biparental crossings, especially in polyploid species, and supplying such a pipeline is the main objective of this work. Therefore, we propose an unprecedented semi-automatized pipeline that is based on principal component

analysis (PCA), genotypic analysis (GA), and clustering analysis (CA) to identify and classify all types of contaminants in a biparental progeny. The proposed methodology was developed and tested in F_1 biparental crosses of tropical forage grasses, but it can be applied to any tetraploid or hexaploid species since the parents of the F_1 biparental cross are known.

Contaminant Identification in Simulated Data Based on PCA, GA, and CA

Principal component analysis (PCA) is a multivariate data technique used to represent a dataset as orthogonal variables named principal components (PCs). Aiming at reducing the dimensionality of a set of variables through linear combinations, repeated information can be removed while the maximum variance–covariance structure of these variables is maintained (Jolliffe and Cadima, 2016). As the first two components explain the most variance in the SNP data, a scatter plot of the samples in a Cartesian plane with these PCs is a way to visually identify similarities and differences, and determine whether samples can be grouped (Ringnér, 2008). Our results showed that in a simulated biparental F_1 progeny without the presence of contaminants, the first components showed a two-dimensional pattern in which the population was distributed between the two parents (**Supplementary Figure 1**), which was expected since these individuals were closely related. As the first PCs generally reflect the variance related to the population structure in the sample, individuals from the same population form a unique cluster in a subspace spanned by the first two eigenvectors (Ma and Amos, 2012).

Considering the four simulated scenarios described above, a contaminant frequency of ∼3% in a progeny is needed to observe a different pattern of PCs that allows the identification of contaminants (**Figure 2B** and **Supplementary Figures 5–7**), which shows the inefficiency of employing a PCA biplot for such an approach. In cases with a lower percentage of apomictic clones (ACs), self-fertilization progenies (SPs), or half-siblings (HSs), duplicating the genotype of one of the parents to generate artificial clones proved to be an alternative way to change the dispersion pattern of individuals, inducing the projection of contaminants as separated from the real hybrids (**Supplementary Figures 11**, **12**). This occurred because the values for the linear combination increased for the PC1 vector, and the source of variation changed to be based on the presence of inserted clones. On the other hand, we found that full contaminants (FCs) could be detected with fewer contaminating individuals (1.5% contaminants in relation to the total population) due to the different genetic backgrounds in relation to the progeny. This high genetic variability modifies the first components and thereby facilitates the identification of FCs in the PCA.

In general, PCA has a better-defined pattern that allows for more inferences about population relationships, not at the individual level (Patterson et al., 2006). It has been widely performed using microsatellite and SNP markers for diploid and polyploid species to evaluate population structure (Larsen et al., 2018; Lara et al., 2019), to infer genetic ancestry (Byun et al., 2017), to predict genomic breeding values (Macciotta et al., 2010), and for other applications. However, for contaminant identification, the use of the first components from PCA, even those successfully employed in forage grass polyploids (Lara et al., 2019; Deo et al., 2020), proved to be insufficient in most scenarios; therefore, other approaches are required.

In the pipeline described here, we propose the use of PCA to visualize the data and produce information that suggest the presence of possible contaminants in biparental crosses. The main limitation of PCA lies in cases with few contaminants, i.e., <3% of the progeny, which has already been reported in tropical forage grasses (Deo et al., 2020). Artificially inserting simulated clones from one of the parents changed the dispersion pattern in most cases; however, when the contaminants were HSs or FCs, the variance was still not captured by the first components. Therefore, PCA itself could not effectively identify and classify the contaminants and, for this reason, was combined with other techniques. We suggested the use of specific GA measures as inputs for CAs as a methodological workflow capable of identifying contaminants regardless of the type or quantity, overcoming the limitation of PCA in identifying contaminants in proportions below 3% of the total population.

The fundamental idea underlying GA-I, GA-II, and GA-III was to identify incompatibilities between putative hybrids and their parents as a strategy to conclusively demonstrate their parentage. For such analyses, it is expected that the GA scores from different populations (here, hybrids and contaminants) form different distributions with specific parameters. Although there are other approaches for parentage estimation already discussed in the literature, such as Identity by Descent (IBD) or pairwise relatedness (r) (Huang et al., 2015; Zheng et al., 2016; Amadeu et al., 2020), these measures indicate how close an individual is to another in a given population, regardless of the degree of relationship. GA measures, on the other hand, differ from these in terms of their focus on the genetic relationships in biparental populations for which both parents are known. Here, the main objective is to compute scores that are related to the type of contaminants expected in such populations, enabling not only identification but also classification.

In all simulated populations with 689 or more genotyped markers, the proposed methodology could correctly identify and classify almost 100% of the samples, ratifying the appropriateness of the proposed pipeline. The size of the markers set employed in different scenarios has been demonstrated to have a large effect on the accuracy of the methodology, as we observed a positive correlation between the two variables. Nevertheless, considering the most indicated clustering scheme, sets with more than 689 markers did not cause an expressive accuracy increase (**Figure 5**). Previous studies have evaluated accuracies in the function of number of markers in different genomic approaches, such as parentage assignment and genomic selection, and found similar results (Wang, 2012; Arruda et al., 2015; Lenz et al., 2017; Whalen et al., 2019). However, finding and generalizing the optimal number of markers for this methodology is complicated because it may be influenced by various factors, including the species, population size, contaminants quantity/type, and sequencing/genotyping techniques.

Even though the CA identifies different groups of individuals with similar GA measures, the association of each group with a contaminant type requires an additional step, which is important because identifying the type of contamination (in the case of AC or SP contamination) can assist the breeder to better understand the reproductive biology of the species or genotype. On the other hand, identifying HS or FC contaminants highlights the need for greater control during the field experiment, avoiding foreign pollen or seed mixtures. Interestingly, we noticed that each cluster captured a distinct pattern in the GA measures, a phenomenon that can be leveraged to decipher the contaminant origin of the individuals. Importantly, by using the proposed approach, we did not find any configuration in which true hybrids were discarded, which is of great value for real applications.

In summary, our proposal is a unique methodology that brings together all types of contamination in a single identification pipeline, representing an important resource for breeders, who need specific tools to deal with such contamination. Instead of relying solely on the putative population structure revealed by PCA methodologies, genotypic analysis (GA) indexes are calculated, taking into account the genetics behind the origin of the contaminants. Compared to the exclusive use of PCA, this pipeline identifies one or a few contaminating individuals with more confidence. This increased confidence makes this methodology ideal for situations in the field that lead to mixtures of seeds or foreign pollen during fertilization, which usually occurs at low rates.

Contamination Identification in Real Data

Principal component analysis, GA, and CA using genotypic data from the *Megathyrsus maximus*, *Urochloa decumbens*, and *Urochloa humidicola* F_1 progenies led to the conclusion that these real progenies had AC contaminants (**Supplementary Figures 15, 17** and **Figure 7**). For *M. maximus*, the two detected clones (1.4% of the population) corroborated the findings of Deo et al. (2020), while for *U. decumbens*, 52 individuals (21.7% of the population) were identified as clones of the male parent. It is possible that these clones were inserted into these two progenies during seed collection. Additionally, the male parent was used as a control in the field experiments, and the plants may have produced seeds and/or seedlings that became mixed with the real progeny. As the female parent of these populations was entirely sexual, the absence of SPs suggests the predominance of allogamy in these plants and self-incompatibility as the main mechanism to guarantee this mode of reproduction.

We extended this methodology for the identification of contaminants in hexaploid species, represented in this study by *U. humidicola* ($2n = 6x = 36$). GA-I and GA-II were performed in the same way as for tetraploid species, but GA-III was adapted considering the segregation and possible combination of gametes in hexaploid species. For the progeny of *U. humidicola*, the combined PC and GA-I histogram analysis suggested the presence of 61 clones of the female parent (21.8% of the population). This result suggests that the genotype H031 (CIAT 26146) also reproduces through facultative apomixis, even though it has been widely cited in the literature as a unique obligate sexual genotype of *U. humidicola* (Jungmann et al., 2010; Vigna et al., 2016). It is known that the expression of apomixis in the same genotype may vary with the flowering season in other grasses (Rios et al., 2013). It might be that the mode of reproduction of H031 was evaluated at the end of flowering or under a specific environmental condition when the proportion of sexuality was greater than apomixis; therefore, this genotype might be a facultative apomict with high rates of sexuality (Karunarathne et al., 2020). In addition, the sexual genotypes of the *Urochloa* spp. can present a certain degree of self-incompatibility (SI) (Keller-Grein et al., 1996; Dusi et al., 2010), and Worthington et al. (2019) reported the detection of 12 individuals derived from accidental self-pollination of *U. humidicola* H031. Therefore, there is a need to enrich the current understanding of *U. humidicola* biology and reproduction mode, which are important for developing suitable breeding and selection methods (Barcaccia and Albertini, 2013).

All three forage progenies used in this work have already been used in the studies previously developed for the construction of genetic maps. Deo et al. (2020) identified and removed two contaminants in *M. maximus* progeny by PCA, which were also identified as contaminants by our methodology. However, for the progeny of *U. decumbens* (Ferreira et al., 2019) and *U. humidicola* (Vigna et al., 2016), only an analysis of the bands of the hybrids identified by genotyping with dozens of microsatellites or single sequence repeats (SSRs) and random amplified polymorphic DNA (RAPD) markers (Bitencourt et al., 2008), respectively, was performed, and no clones could be identified through this approach. Therefore, the absence of an adequate methodology and/or a sufficient number of markers for the prior identification of contaminants has resulted in genetic maps constructed with genetic information including some false hybrids, and consequently, these maps may contain bias that should be considered by researchers.

Our methodology proved to be useful in practical situations of breeding programs of tropical forage grasses, including the identification of different progenies from multiparent crosses, which may be extended to other polyploid crops. The identification of contaminants in the early stages of breeding cycles can greatly increase the efficiency of programs, preventing costs with false hybrids that might otherwise only be discarded in the later phases of selection. Conversely, it allows for the size of the useful population to increase, optimizing the breeding populations. Although the use of molecular markers is not yet a reality in many breeding programs, it is important to assess potential expenses brought by false hybrids, which might surpass the cost of large-scale genotyping technologies (such as GBS), which have been experiencing considerable cheapening in the recent years. PCA, GA, and CA were combined in a simple and semi-automated pipeline, and the coupling of a low-cost genotyping with such pipeline thus allows for a more precise and efficient detection of incompatibilities between a group of putative hybrids and the

identification of contaminants in biparental crosses of tetraploid and hexaploid species.

The implementation of this simple approach in the identification of contaminants in biparental progenies of polyploid species can increase the efficiency of breeding programs. In this context, the polyCID Shiny app was designed to enhance the ability of breeders to use our methodology, even with no bioinformatics expertise. Great advances in sequencing technologies and genotyping tools have enabled us to explore vast amounts of genetic data in a more cost-effective and faster way; however, the ability to handle and apply this genome information to breeding remains a significant barrier for most breeders and experimental researchers. Therefore, we designed the polyCID Shiny app as an interactive and user-friendly application that is completely R based and easy to install, incorporating the analysis in a single environment and enabling the users to extract information on contaminant individuals without requiring knowledge of a programming language.

Finally, although our analyses were performed with real and simulated progenies of tropical forage grasses, this methodology can be extended to any biparental progeny of tetraploid or hexaploid species. It can be applied in the early stages of genomic studies with GBS in biparental polyploid progenies, such as genetic linkage map construction and genomic prediction, to identify possible contaminants. However, as the price of SNP genotyping is constantly decreasing and other polyploid genotyping tools are emerging, the application of our methodology even in experiments that do not involve SNPs may be possible, mainly in the intermediate and final stages of the breeding program to confirm the absence of contamination in the final stages and cultivar release. In the case of genotyping with a lower number of molecular markers, it is suggested that simulation studies be carried out *a priori*, taking into account how the number and quality of the markers affect the final results.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA703438, https://www.ncbi.nlm.nih.gov/, SRP148665, https://www.ncbi.nlm.nih.gov/, PRJNA563938.

AUTHOR CONTRIBUTIONS

FM, AM, AA, BV, and AS conceived the study. LC, RS, SB, MS, LJ, and CV conducted the field experiments. AM, RF, and BV performed the laboratory experiments. FM, AM, and AA analyzed the data. FM, AM, AA, RF, and BV wrote the manuscript. AA and FM implemented the Shiny web app. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Fundação de Amparo à Pesquisa de do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Computational Biology Programme and Financial Code 001), Embrapa, and the Associação para o Fomento à Pesquisa de Melhoramento de Forrageiras (UNIPASTO). FM received a Ph.D. fellowship from CAPES (88882.329502/2019-01); AA received a Ph.D. fellowship from FAPESP (2019/03232-6); RF received a PD fellowship from FAPESP (2018/19219-6); SB, LJ, and AS received research fellowships from CNPq (315271/2018-3, 315456/2018-3, and 312777/2018–3, respectively).

ACKNOWLEDGMENTS

We would like to acknowledge the Fundação de Amparo à Pesquisa de do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). We also acknowledge the Brazilian Agricultural Research Corporation (Embrapa Gado de Corte) for providing the populations used in this study. This manuscript was previously posted to bioRxiv at https://www.biorxiv.org/content/10.1101/2021.07.01.450796v1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.737919/full#supplementary-material

REFERENCES

ABIEC. (2020). *Beef Report Perfil da Pecuária No Brasil*. Available online at: http://www.abiec.com.br/controle/uploads/arquivos/sumario2019portugues.pdf (accessed July 22, 2020).

Acuña, C. A., Martínez, E. J., Zilli, A. L., Brugnoli, E. A., Espinoza, F., Marcón, F., et al. (2019). Reproductive systems in paspalum: relevance for germplasm collection and conservation, breeding techniques, and adoption of released cultivars. *Front. Plant Sci.* 10:1377. doi: 10.3389/fpls.2019.01377

Alam, M., Neal, J., O'Connor, K., Kilian, A., and Topp, B. (2018). Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. *PLoS ONE* 13:e0203465. doi: 10.1371/journal.pone.0203465

Amadeu, R. R., Lara, L. A. C., Munoz, P., and Garcia, A. A. F. (2020). Estimation of molecular pairwise relatedness in autopolyploid crops. *G3 Genes Genomes Genet.* 10, 4579–4589. doi: 10.1534/g3.120.401669

Anderson, E. C. (2012). Large-scale parentage inference with SNPs: an efficient algorithm for statistical confidence of parent pair allocations. *Stat. Appl. Genet. Mol. Biol.* 11:296–302. doi: 10.1515/1544-6115.1833

Arruda, M. P., Brown, P. J., Lipka, A. E., Krill, A. M., Thurber, C., and Kolb, F. L. (2015). Genomic selection for predicting fusarium head blight resistance in a wheat breeding program. *Plant Genome* 8:1–12. doi: 10.3835/plantgenome2015.01.0003

Azevedo, A. L. S., Pereira, J. F., and Machado, J. C. (2019). *Melhoramento de Forrageiras na Era Genômica*. Brasília: Embrapa.

Barcaccia, G., and Albertini, E. (2013). Apomixis in plant reproduction: a novel perspective on an old dilemma. *Plant Reprod.* 26, 159–179. doi: 10.1007/s00497-013-0222-y

Barrios, S. C. L., Do Valle, C. B., Alves, G. F., Simeão, R. M., and Jank, L. (2013). Reciprocal recurrent selection in the breeding of Brachiaria decumbens. *Trop. Grasslands-Forrajes Trop.* 1, 52–54. doi: 10.17138/TGFT(1)52-54

Bateman, A. J. (1947). Contamination of seed crops. *J. Genet.* 48, 257–275. doi: 10.1007/BF02989385

Bicknell, R. A. (2004). Understanding apomixis: recent advances and remaining conundrums. *Plant Cell Online* 16, S228–S245. doi: 10.1105/tpc.017921

Bitencourt, G. A., Chiari, L., Valle, C. B., Salgado, L. R., and Leguizamon, G. O. C. (2008). "Uso de marcadores RAPD na identificação de híbridos de brachiaria humidicola," in *Boletim Pesquisa*, Vol. 23. Campo Grande: Embrapa Gado de Corte, 19.

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Front. Plant Sci.* 9:513. doi: 10.3389/fpls.2018.00513

Byun, J., Han, Y., Gorlov, I. P., Busam, J. A., Seldin, M. F., and Amos, C. I. (2017). Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. *BMC Genom.* 18:789. doi: 10.1186/s12864-017-4166-8

Cappai, F., Amadeu, R. R., Benevenuto, J., Cullen, R., Garcia, A., Grossman, A., et al. (2020). High-resolution linkage map and QTL analyses of fruit firmness in autotetraploid blueberry. *Front. Plant Sci.* 11:562171. doi: 10.3389/fpls.2020.562171

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2021). *Shiny: Web Application Framework for R. R Package Version 1.6.0.* Available online at: https://CRAN.R-project.org/package=shiny. (accessed July 1, 2021).

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* 61, 1–36. doi: 10.18637/jss.v061.i06

Christie, M. R., Tennessen, J. A., and Blouin, M. S. (2013). Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching. *Bioinformatics* 29, 725–732. doi: 10.1093/bioinformatics/btt039

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., De Pristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Deo, T. G., Ferreira, R., Lara, L., Moraes, A., Alves-Pereira, A., de Oliveira, F. A., Garcia, A., Santos, M. F., Jank, L., and de Souza, A. P. (2020). High-resolution linkage map with allele dosage allows the identification of regions governing complex traits and apospory in Guinea grass (*Megathyrsus maximus*). *Front. Plant Sci.* 11, 15. doi: 10.3389/fpls.2020.00015

Dusi, D. M. A., Alves, E. R., Willemse, M. T. M., Falcão, R., do Valle, C. B., and Carneiro, V. T. C. (2010). Toward *in vitro* fertilization in *Brachiaria* spp. *Sex. Plant Reprod.* 23, 187–197. doi: 10.1007/s00497-010-0134-z

Ellis, T. J., Field, D. L., and Barton, N. H. (2018). Efficient inference of paternity and sibship inference given known maternity via hierarchical clustering. *Mol. Ecol. Resour.* 18, 988–999. doi: 10.1111/1755-0998.12782

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Ferrão, L. F. V., Johnson, T. S., Benevenuto, J., Edger, P. P., Colquhoun, T. A., and Munoz, P. R. (2020). Genome-wide association of volatiles reveals candidate loci for blueberry flavor. *New Phytol.* 226, 1725–1737. doi: 10.1111/nph.16459

Ferreira, R. C. U., Lara, L. A. D. C., Chiari, L., Barrios, S. C. L., do Valle, C. B., Valério, J. R., et al. (2019). Corrigendum: genetic mapping with allele dosage information in tetraploid *Urochloa decumbens* (Stapf) R. D. Webster reveals insights into spittlebug (Notozulia entreriana Berg) resistance. *Front. Plant Sci.* 10:92. doi: 10.3389/fpls.2019.00855

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., and Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics* 210, 789–807. doi: 10.1534/genetics.118.301468

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9:e90346. doi: 10.1371/journal.pone.0090346

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Gori, K., Suchan, T., Alvarez, N., Goldman, N., and Dessimoz, C. (2016). Clustering genes of common evolutionary history. *Mol. Biol. Evol.* 33, 1590–1605. doi: 10.1093/molbev/msw038

Goulet, B. E., Roda, F., and Hopkins, R. (2017). Hybridization in plants: old ideas, new techniques. *Plant Physiol.* 173, 65–78. doi: 10.1104/pp.16.01340

Grashei, K. E., Ødegård, J., and Meuwissen, T.H.E. (2018). Using genomic relationship likelihood for parentage assignment. *Genet. Sel. Evol.* 50:26. doi: 10.1186/s12711-018-0397-7

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., et al. (2011). Current trends in microsatellite genotyping. *Mol. Ecol. Resour.* 11, 591–611. doi: 10.1111/j.1755-0998.2011.03014.x

Hackett, C. A., McLean, K., and Bryan, G. J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS ONE* 8:e63939. doi: 10.1371/journal.pone.0063939

Hand, M. L., and Koltunow, A. M. G. (2014). The genetic control of apomixis: asexual seed formation. *Genetics* 197, 441–450. doi: 10.1534/genetics.114.163105

Hayes, B. J. (2011). Technical note: efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *J. Dairy Sci.* 94, 2114–2117. doi: 10.3168/jds.2010-3896

Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., et al. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS ONE* 9:e94851. doi: 10.1371/journal.pone.0094851

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., et al. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* 11, 123–136. doi: 10.1111/j.1755-0998.2010.02943.x

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., et al. (2016). The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. *Appl. Plant Sci.* 4:1600025. doi: 10.3732/apps.1600025

Huang, K., Guo, S. T., Shattuck, M. R., Chen, S. T., Qi, X. G., Zhang, P., et al. (2015). A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity* 114, 133–142. doi: 10.1038/hdy.2014.88

Huisman, J. (2017). Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. *Mol. Ecol. Resour.* 17, 1009–1024. doi: 10.1111/1755-0998.12665

Jank, L., Valle, C. B., and Resende, R. M. S. (2011). Breeding tropical forages. *Crop Breed. Appl. Biotechnol.* 11, 27–34. doi: 10.1590/S1984-70332011000500005

Jha, N. K., Jacob, S. R., Nepolean, T., Jain, S. K., and Kumar, M. B. A. (2016). SSR markers based DNA fingerprinting and it's utility in testing purity of eggplant hybrid seeds. *Qual. Assur. Saf. Crops Foods* 8, 333–338. doi: 10.3920/QAS2015.0689

Ji, K., Zhang, D., Motilal, L. A., Boccara, M., Lachenaud, P., and Meinhardt, L. W. (2013). Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. *Genet. Resour. Crop Evol.* 60, 441–453. doi: 10.1007/s10722-012-9847-1

Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374:20150202. doi: 10.1098/rsta.2015.0202

Jones, O. R., and Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Mol. Ecol. Resour.* 10, 551–555. doi: 10.1111/j.1755-0998.2009.02787.x

Jungmann, L., Vigna, B. B. Z., Boldrini, K. R., Sousa, A. C. B., do Valle, C. B., Resende, M. S., et al. (2010). Genetic diversity and population structure analysis of the tropical pasture grass *Brachiaria humidicola* based on microsatellites, cytogenetics, morphological traits, and geographical origin. *Genome.* 53, 698–709. doi: 10.1139/G10-055

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* 16, 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Karunarathne, P., Reutemann, A. V., Schedler, M., Glücksberg, A., Martínez, E. J., Honfi, A. I., et al. (2020). Sexual modulation in a polyploid grass: a reproductive

contest between environmentally inducible sexual and genetically dominant apomictic pathways. *Sci. Rep.* 10:8319. doi: 10.1038/s41598-020-64982-6

Keller-Grein, G., Maass, B. L., and Hanson, J. (1996). "Natural variation in *Brachiaria* and existing germplasm collections," in *Brachiaria: Biology, Agronomy and Improvement*, eds J. W. Miles, B. L. Maass, and C. B. Valle (Brasília: Embrapa, CIAT, Cali), 16–42.

Kemble, H., Nghe, P., and Tenaillon, O. (2019). Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. *Evol. Appl.* 12, 1721–1742. doi: 10.1111/eva.12846

Kempf, K., Grieder, C., Walter, A., Widmer, F., Reinhard, S., and Kölliker, R. (2015). Evidence and consequences of self-fertilisation in the predominantly outbreeding forage legume *Onobrychis viciifolia*. *BMC Genet.* 16:117. doi: 10.1186/s12863-015-0275-z

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lara, L. A. C., Santos, M. F., Jank, L., Chiari, L., Vilela, M. D. M., Amadeu, R. R., et al. (2019). Genomic selection with allele dosage in *Panicum maximum* Jacq. *G3 Genes Genomes Genet.* 9, 2463–2475. doi: 10.1534/g3.118.200986

Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., and Migicovsky, Z., Myles, S., et al. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PLoS ONE* 13:e0201889. doi: 10.1371/journal.pone.0201889

Laucou, V., Launay, A., Bacilieri, R., Lacombe, T., Adam-Blondon, A.-F., Bérard, A., et al. (2018). Extended diversity analysis of cultivated grapevine *Vitis vinifera* with 10K genome-wide SNPs. *PLoS ONE* 13:e0192540. doi: 10.1371/journal.pone.0192540

Lenz, P. R. N., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., and Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). *BMC Genomics* 18:335. doi: 10.1186/s12864-017-3715-5

Lutts, S., Ndikumana, J., and Louant, B. P. (1991). Fertility of *Brachiaria ruziziensis* in interspecific crosses with *Brachiaria decumbens* and *Brachiaria brizantha*: meiotic behaviour, pollen viability and seed set. *Euphytica* 57, 267–274. doi: 10.1007/BF00039673

Ma, J., and Amos, C. I. (2012). Principal components analysis of population admixture. *PLoS ONE* 7:e40115. doi: 10.1371/journal.pone.0040115

Macciotta, N. P. P., Gaspa, G., Steri, R., Nicolazzi, E. L., Dimauro, C., Pieramati, C., et al. (2010). Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *J. Dairy Sci.* 93, 2765–2774. doi: 10.3168/jds.2009-3029

Maćkiewicz, A., and Ratajczak, W. (1993). Principal components analysis (PCA). *Comput. Geosci.* 19, 303–342. doi: 10.1016/0098-3004(93)90090-R

Martuscello, J. A., Jank, L., Fonseca, D. M. D., Cruz, C. D., and Cunha, D. N. F. V. (2009). Among and within family selection and combined half-sib family selection in *Panicum maximum* Jacq. *Rev. Bras. Zootec.* 38, 1870–1877. doi: 10.1590/S1516-35982009001000003

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Endelman, J. B., et al. (2019). On the accuracy of genomic prediction models considering multi-trait and allele dosage in *Urochloa* spp. interspecific tetraploid hybrids. *Mol. Breed.* 39:100. doi: 10.1007/s11032-019-1002-7

McClure, M. C., McCarthy, J., Flynn, P., McClure, J. C., Dair, E., O'Connell, D. K., et al. (2018). SNP data quality control in a National beef and dairy cattle system and highly accurate SNP based parentage verification and identification. *Front. Genet.* 9:84. doi: 10.3389/fgene.2018.00084

Mendel, G. (1866). Versuche über pflanzen-hybriden. *Verh. Naturforschenden Ver. Brünn* 4, 3–47. doi: 10.5962/bhl.title.61004

Miko, I. (2008). Gregor Mendel and the principles of inheritance. *Nat. Educ.* 1:134. Available online at: https://www.nature.com/scitable/topicpage/gregor-mendel-and-the-principles-of-inheritance-593/

Mollinari, M., Olukolu, B. A., Pereira, G. D. S., Khan, A., Gemenet, D., Yencho, G. C., et al. (2020). Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. *G3 Genes Genomes Genet.* 10, 281–292. doi: 10.1534/g3.119.400620

Morrone, O., and Zuloaga, F. O. (1992). Revision de las especies sudamericanas nativas e introducidas de los generos *Brachiaria* y *Urochloa* (*Poaceae*: *Panicoideae*: *Paniceae*). *Darwiniana* 31, 43–109.

Muniz, A. C., Lemos-Filho, J. P., Buzatti, R. S. O., Ribeiro, P. C. C., Fernandes, F. M., and Lovato, M. B. (2019). Genetic data improve the assessment of the conservation status based only on herbarium records of a neotropical tree. *Sci. Rep.* 9:5693. doi: 10.1038/s41598-019-41454-0

Patella, A., Palumbo, F., Galla, G., and Barcaccia, G. (2019). The molecular determination of hybridity and homozygosity estimates in breeding populations of lettuce (*Lactuca sativa* L.). *Genes* 10:916. doi: 10.3390/genes10110916

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Pereira, G. S., Garcia, A. A. F., and Margarido, G. R. A. (2018b). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinform.* 19:398. doi: 10.1186/s12859-018-2433-6

Pereira, J. F., Azevedo, A. L. S., Pessoa-Filho, M., Romanel, E. A. C., Pereira, A. V., Vigna, B. B. Z., et al. (2018a). Research priorities for next-generation breeding of tropical forages in Brazil. *Crop Breed. Appl. Biotechnol.* 18, 314–319. doi: 10.1590/1984-70332018v18n3n46

Pinheiro, A. A., Pozzobon, M. T., do Valle, C. B., Penteado, M. I. O., and Carneiro, V. T. C. (2000). Duplication of the chromosome number of diploid *Brachiaria brizantha* plants using colchicine. *Plant Cell Rep.* 19, 274–278. doi: 10.1007/s002990050011

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7:e32253. doi: 10.1371/journal.pone,0.0032253

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ringnér, M. (2008). What is principal component analysis? *Nat. Biotechnol.* 26, 303–304. doi: 10.1038/nbt0308-303

Rios, E., Blount, A., Kenworthy, K., Acuña, C., and Quesenberry, K. (2013). Seasonal expression of apospory in *Bahiagrass*. *Trop. Grassl.* 1, 116–118. doi: 10.17138/TGFT(1)116-118

Santos, J. M. D., Barbosa, G. V. D. S., Neto, C. E. R., and Almeida, C. (2014). Efficiency of biparental crossing in sugarcane analyzed by SSR markers. *Crop Breed. Appl. Biotechnol.* 14, 102–107. doi: 10.1590/1984-70332014v14n2a18

Simeão, R., Silva, A., Valle, C., Resende, M. D., and Medeiros, S. (2016a). Genetic evaluation and selection index in tetraploid *Brachiaria ruziziensis*. *Plant Breed.* 135, 246–253. doi: 10.1111/pbr.12353

Simeão, R. M., Valle, C. B., and Resende, M. D. V. (2016b). Unravelling the inheritance, QSTand reproductive phenology attributes of the tetraploid tropical grass *Brachiaria ruziziensis* (Germain et Evrard). *Plant Breed.* 136, 101–110. doi: 10.1111/pbr.12429

Simioni, C., and Valle, C. B. (2009). Chromosome duplication in *Brachiaria* (A. Rich.) Stapf allows intraspecific crosses. Crop Breed. *Appl. Biotechnol.* 9, 328–334. doi: 10.12702/1984-7033.v09n04a07

Smith, R. L. (1972). Sexual reproduction in *Panicum maximum* Jacq. *Crop Sci.* 12, 624–627. doi: 10.2135/cropsci1972.0011183X001200050021x

Spielmann, A., Harris, S. A., Boshier, D. H., and Vinson, C. C. (2015). Orchard: paternity program for autotetraploid species. *Mol. Ecol. Resour.* 15, 915–920. doi: 10.1111/1755-0998.12370

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23, 1164–1167. doi: 10.1093/bioinformatics/btm069

Subashini, V., Shanmugapriya, A., and Yasodha, R. (2014). Hybrid purity assessment in *Eucalyptus* F1 hybrids using microsatellite markers. *3 Biotech* 4, 367–373. doi: 10.1007/s13205-013-0161-1

Telfer, E. J., Stovold, G. T., Li, Y., Silva-Junior, O. B., Grattapaglia, D. G., and Dungey, H. S. (2015). Parentage reconstruction in *Eucalyptus nitens* using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness. *PLoS ONE* 10:e0130601. doi: 10.1371/journal.pone.0130601

Thompson, E. A. (1975). The estimation of pairwise relationships. *Ann Human Genet* 39, 173–188. doi: 10.1111/j.1469-1809.1975.tb00120.x

Thompson, E. A., and Meagher, T. R. (1987). Parental and sib likelihoods in genealogy reconstruction. *Biometrics* 43, 585. doi: 10.2307/2531997

Torres-González, A. M., and Morton, C. M. (2005). Molecular and morphological phylogenetic analysis of *Brachiaria* and *Urochloa* (*Poaceae*). *Mol. Phylogenetics Evol.* 37, 36–44. doi: 10.1016/j.ympev.2005.06.003

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39, 312–328. doi: 10.1590/1678-4685-GMB-2016-0027

Vigna, B. B. Z., Santos, J. C. S., Jungmann, L., do Valle, C. B., Mollinari, M., Pastina, M. M., et al. (2016). Evidence of allopolyploidy in *Urochloa humidicola* based on cytological analysis and genetic linkage mapping. *PLoS ONE* 11:e0153764. doi: 10.1371/journal.pone.0153764

Voorrips, R. E., and Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinform.* 13:248. doi: 10.1186/1471-2105-13-248

Wang, J. (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics* 191, 183–194. doi: 10.1534/genetics.111.138149

Whalen, A., Gorjanc, G., and Hickey, J. M. (2019). Parentage assignment with genotyping-by-sequencing data. *J. Anim. Breed. Genet = Zeits. Tierz. Zucht.* 136, 102–112. doi: 10.1111/jbg.12370

Wickham, H., and Chang, W. (2016). Package 'ggplot2'. Vienna: R Foundation for Statistical Computing. doi: 10.1007/978-3-319-24277-4

Wold, H., and Krishnaiah, P. R. (1966). "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, ed P. R. Krishnaiah (New York, NY: Academic Press), 391–420.

Worthington, M., Ebina, M., Yamanaka, N., Heffelfinger, C., Quintero, C., Zapata, Y. P., et al. (2019). Translocation of a parthenogenesis gene candidate to an alternate carrier chromosome in apomictic *Brachiaria humidicola*. *BMC Genom.* 20:41. doi: 10.1186/s12864-018-5392-4

Yousefi-Mashouf, N., Mehrabani-Yeganeh, H., Nejati-Javaremi, A., Bailey, E., and Petersen, J. L. (2021). Genomic comparisons of Persian Kurdish, Persian Arabian and American thoroughbred horse populations. *PLoS ONE* 16:e0247123. doi: 10.1371/journal.pone.0247123

Zhang, J., Yang, J., Zhang, L., Luo, J., Zhao, H., Zhang, J., et al. (2020). A new SNP genotyping technology target SNP-seq and its application in genetic analysis of cucumber varieties. *Sci. Rep.* 10:5623. doi: 10.1038/s41598-020-62518-6

Zhao, X., Zhang, J., Zhang, Z., Wang, Y., and Xie, W. (2017). Hybrid identification and genetic variation of *Elymus sibiricus* hybrid populations using EST-SSR markers. *Hereditas* 154:15. doi: 10.1186/s41065-017-0053-1

Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., and Bink, M. C. A. M. (2016). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics* 203, 119–131. doi: 10.1534/genetics.115.185579

Zwart, A. B., Elliott, C., Hopley, T., Lovell, D., and Young, A. (2016). polypatex: anrpackage for paternity exclusion in autopolyploids. *Mol. Ecol. Resour.* 16, 694–700. doi: 10.1111/1755-0998.12496

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary Material

1 Supplementary Methods

1.1 Genotype Analysis of simulated data

Supplementary Table 1: Expected and not expected genotypes based on all possible parental allele dosage for a tetraploid.

Progenitors (P1xP2)	**Expected**	**Not expected**
0 x 0	0	1, 2, 3, 4
0 x 1; 1 x 0	0, 1	2, 3, 4
0 x 2; 2 x 0	0, 1, 2	3, 4
0 x 3; 3 x 0	1, 2	0, 3, 4
0 x 4, 4 x 0	2	0, 1, 3, 4
1 x 1	0, 1, 2	3, 4
1 x 2; 2 x 1	0, 1, 2, 3	4
1 x 3; 3 x 1	1, 2, 3	0, 4
1 x 4; 4 x 1	2, 3	0, 1, 4
2 x 2	0, 1, 2, 3, 4	NA
2 x 3; 3 x 2	1, 2, 3, 4	0
2 x 4; 4 x 2	2, 3, 4	0, 1
3 x 3	2, 3, 4	0, 1
3 x 4; 4 x 3	3, 4	0, 1, 2
4 x 4	4	0, 1, 2, 3

1.2 Clustering Indexes

Supplementary Table 2. Overview of the indexes implemented in the NbClust package.

Name of the index in NbClust	**Name of the index in the literature**	**References**
"kl"	KL	(Krzanowski and Lai, 1988)
"ch"	Calinski and Harabasz	(Calinski and Harabasz, 1974)
"hartigan"	Hartigan	(Hartigan, 1975)
"ccc"	Cubic clustering criterion (CCC)	(Sarle, 1983)
"scott"	n log (|T|/|W|)	(Scott, 1971)
"marriot"	k^2 |W|	(Marriot, 1971)
"trcovw"	Trace Cov W	(Milligan and Cooper, 1985)
"tracew"	Trace W	(Edwards and Cavalli-Sforza, 1965; Friedman and Rubin, 1967)
"friedman"	Trace $W^{-1}B$	(Friedman and Rubin, 1967)
"rubin"	|T|/|W|	(Friedman and Rubin, 1967)
"cindex"	C-index	(Hubert and Levin, 1976)
"db"	Davies and Bouldin	(Davies and Bouldin, 1979)
"silhouette"	Silhouette	(Rousseeuw, 1987)
"duda"	Je(2)/Je(1)	(Duda and Hart, 1973)
"pseudot2"	$Pseudot^2$	(Duda and Hart, 1973)
"beale"	Beale	(Beale, 1969)
"ratkowsky"	$c/k^{.5}$	(Ratkowsky and Lance, 1978)
"ball"	Ball and Hall	(Ball and Hall, 1965)
"ptbiserial"	Point-Biserial	(Kraemer, 1982)
"gap"	Gap	(Tibshirani et al., 2001)
"mcclain"	McClain and Rao	(Mcclain et al., 1975)
"gamma"	Gamma	(Baker and Hubert, 1975)
"gplus"	G(+)	(Rohlf, 1974)

"tau"	Tau	(Rohlf, 1974)
"dunn"	Dunn	(Dunn, 1974)
"sdindex"	SD	(Halkidi et al., 2000)
"sdbw"	SDbw	(Halkidi et al., 2001)

2 Supplementary Results

2.1 Simulation Details

The constructed linkage map used in PedigreeSim software consists of 8 chromosomes with 5,516 randomly distributed SNPs; the length, centromere position and number of markers of each chromosome can be found in Supplementary Table 1. Furthermore, as the software requires the parental genotypes for the markers to simulate the crosses, the *M. maximus* and *U. decumbens* parental allele dosage rates were used to genotype the simulated parents (P1, P2, P3 and P4), and the actual number of markers for each dosage in the set of 5,516 markers is shown in Supplementary Tables 2 and 3.

Supplementary Table 3. Information on the linkage map created for the simulation analysis.

Chromosome	**Size (cM)**	**Position of the centromere (cM)**	**Number of markers**
1	108	44	665
2	98	50	615
3	101	17	643
4	99	43	644
5	115	27	685
6	110	45	766
7	112	39	611
8	95	29	887

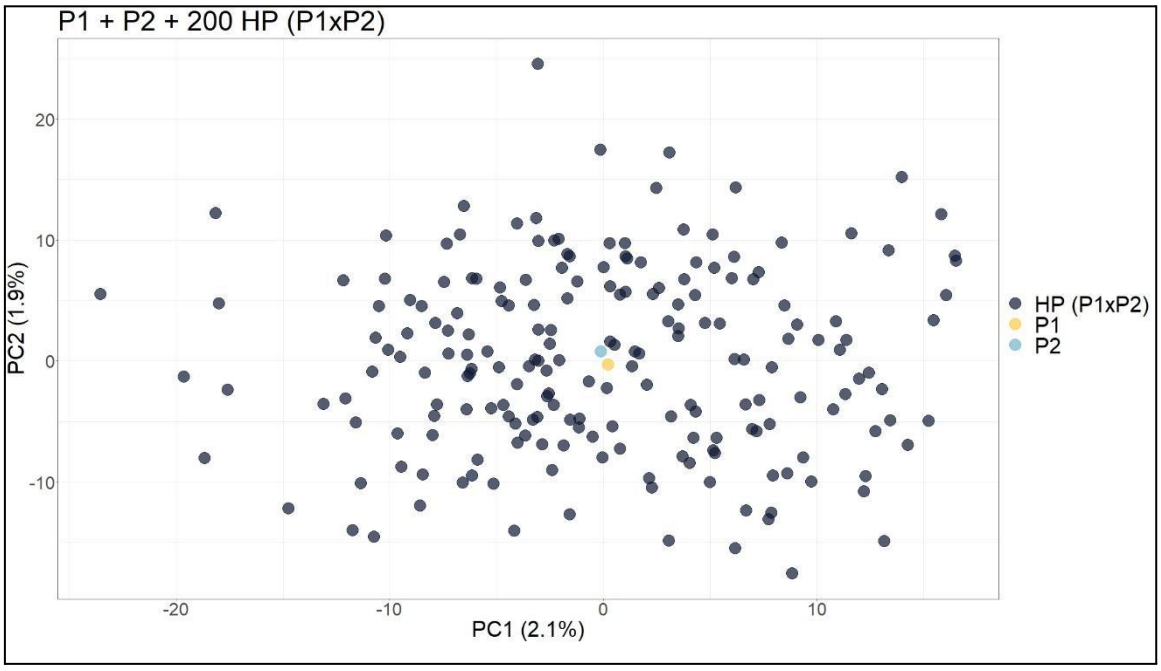
Supplementary Table 4. Rates of each allele dosage used to genotype the simulated parents. P1 and P2 were based on the parents of the *M. maximus* progeny, and P3 and P4 were based on the parents of the *U. decumbens* progeny.

	Nulliplex (0)	**Simplex (1)**	**Duplex (2)**	**Triplex (3)**	**Quadruplex (4)**
P1	0.4253	0.3899	0.1694	0.0110	0.0044
P2	0.3061	0.5255	0.1440	0.0213	0.0031
P3	0.5430	0.3008	0.1429	0.0098	0.0035
P4	0.1559	0.4930	0.2802	0.0671	0.0038

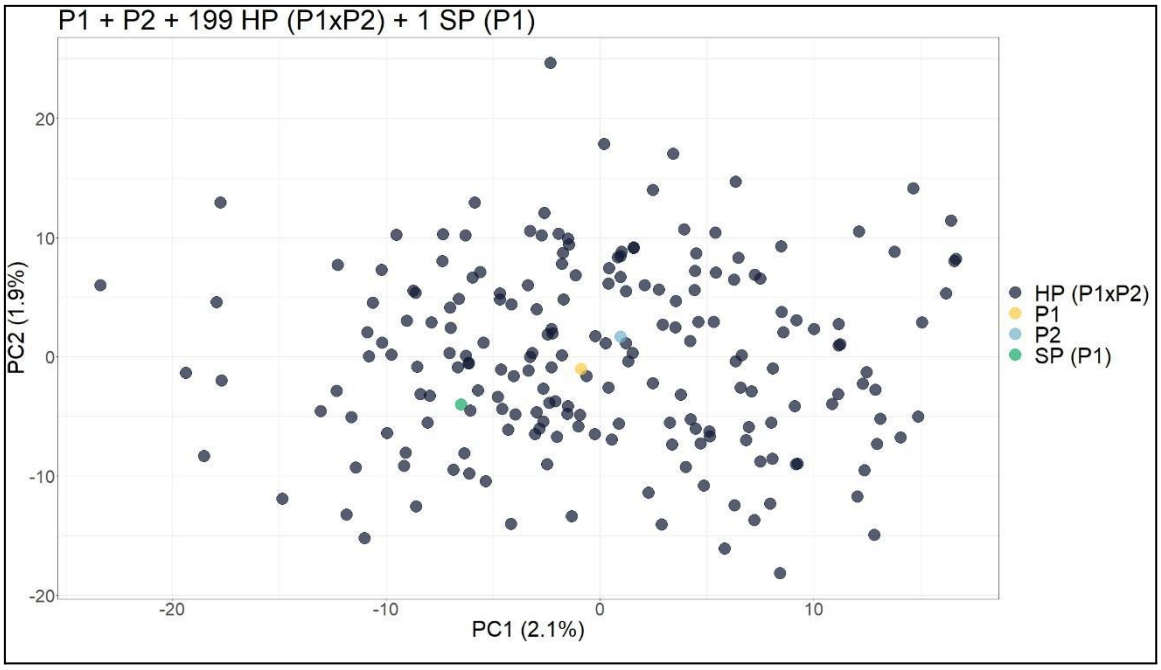
Supplementary Table 5. Number of SNP markers classified in each of the five possible dosage classes for the simulated parents, totaling 5,516 markers.

	Nulliplex (0)	**Simplex (1)**	**Duplex (2)**	**Triplex (3)**	**Quadruplex (4)**
P1	2,447	2,090	897	56	26
P2	1,652	2,876	833	138	17
P3	3,002	1,686	769	42	17
P4	848	2,705	1,552	388	23

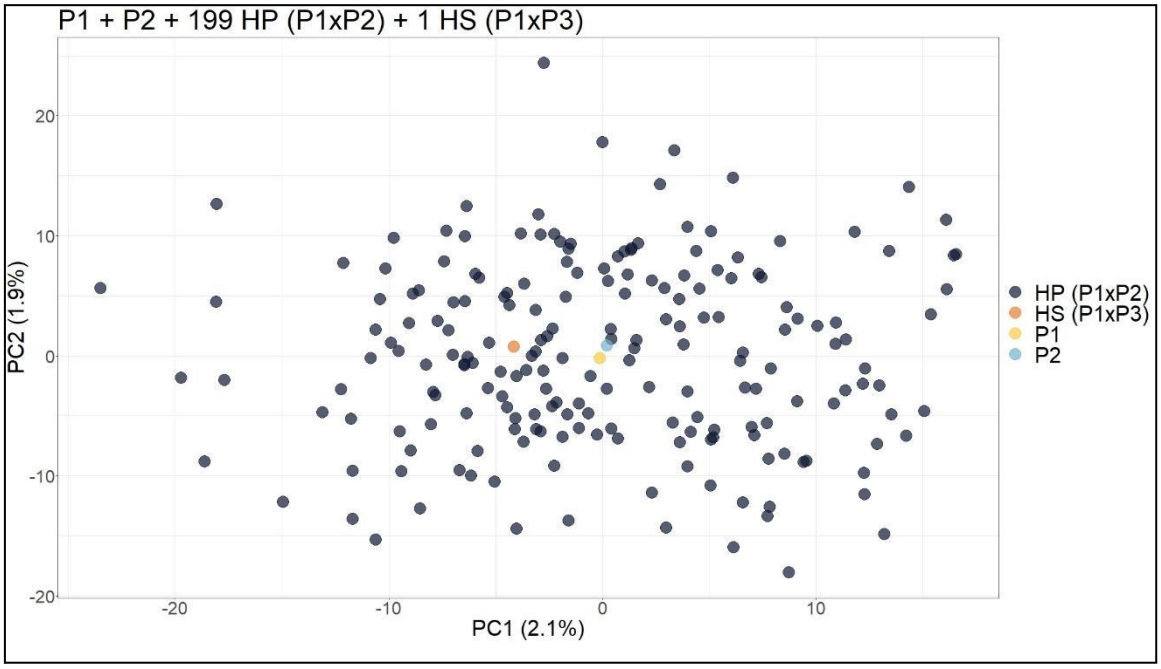
2.2 Principal Component Analysis of Simulated Populations



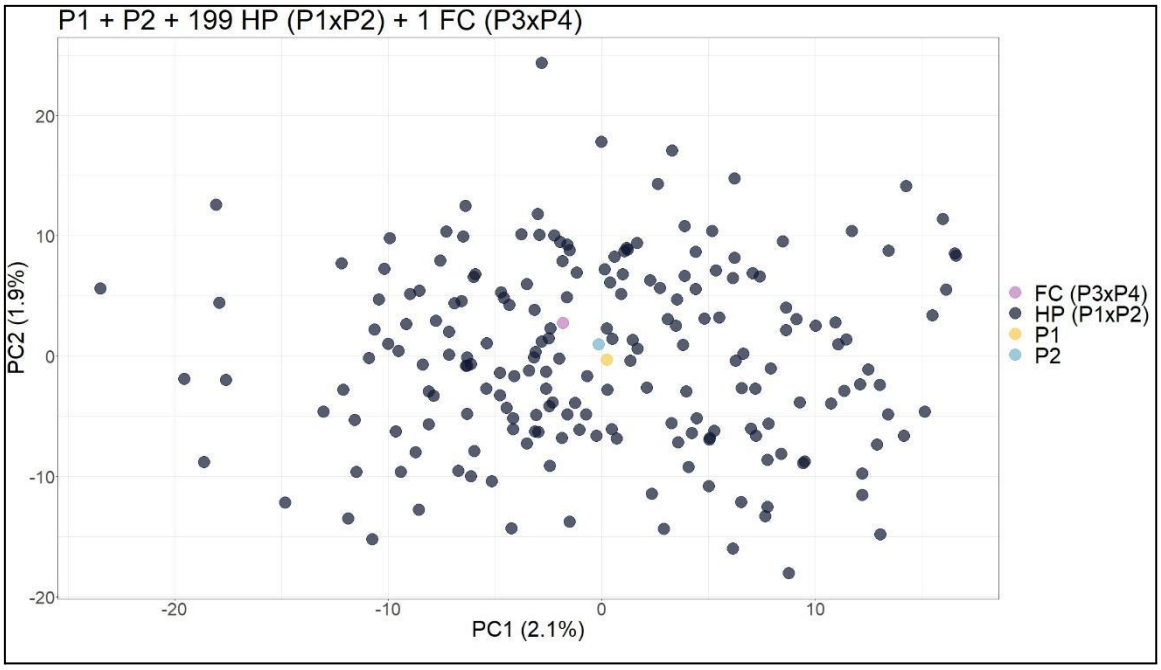
Supplementary Figure 1. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2) and 200 hybrids (HP (P1xP2)). The axes represent the first and second principal components, which explain 2.1% and 1.9% of the variance, respectively.



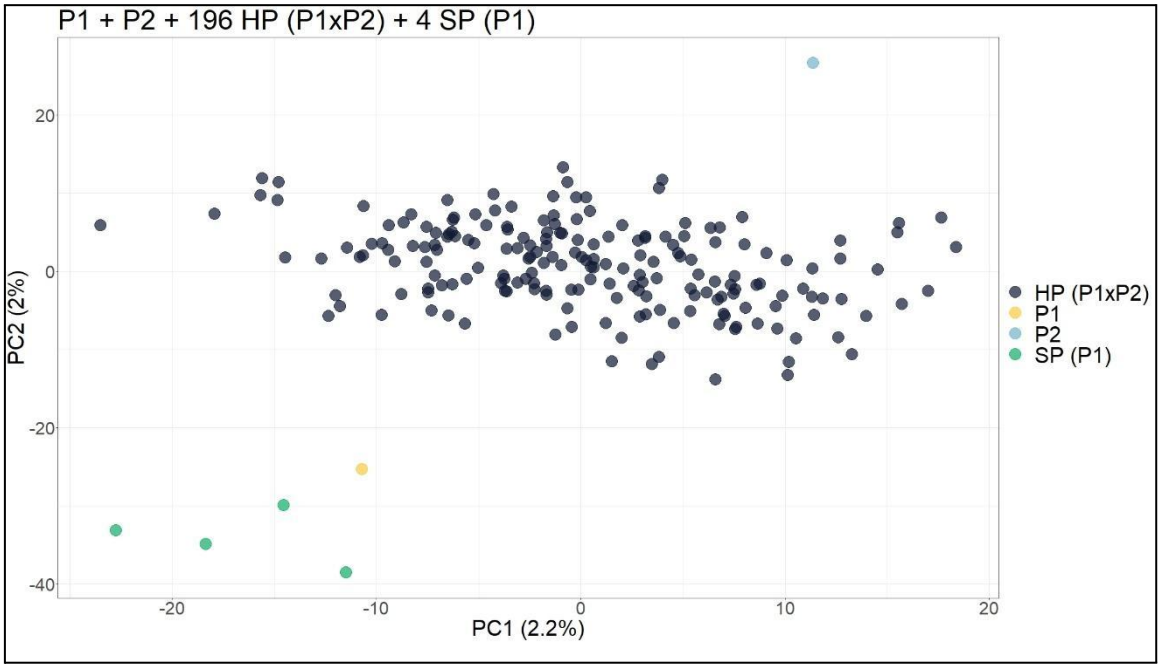
Supplementary Figure 2. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 199 hybrids (HP P1xP2) and one self-fertilization progeny (SP (P1xP1)). The axes represent the first and second principal components, which explain 2.1% and 1.9% of the variance, respectively.



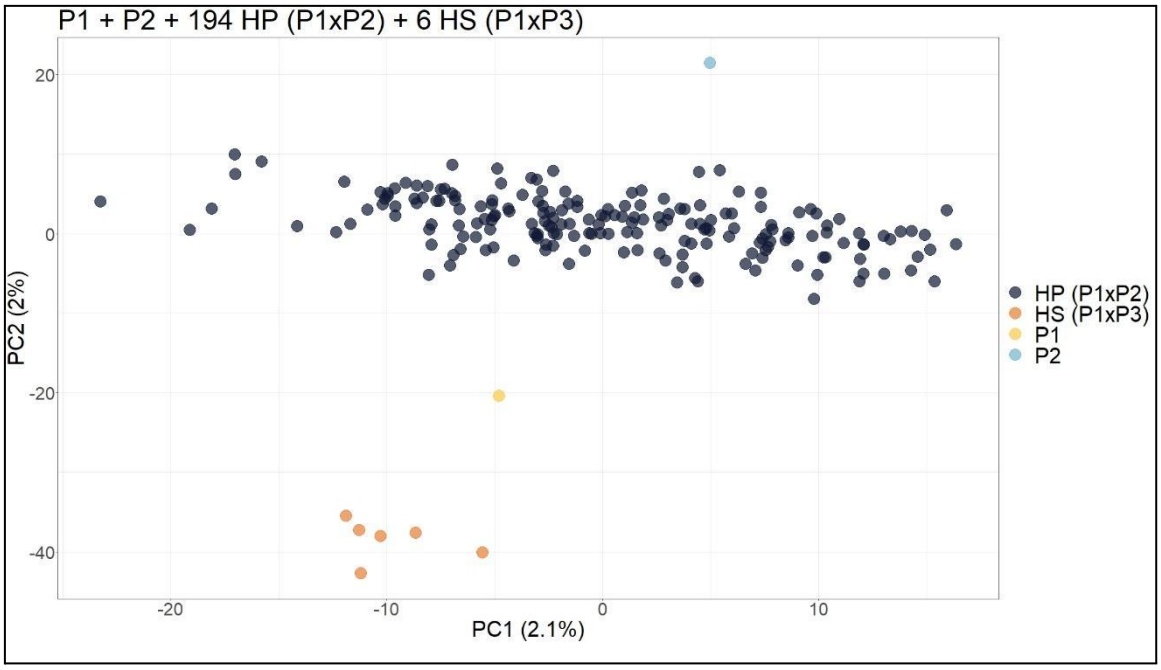
Supplementary Figure 3. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 199 hybrids (HP (P1 x P2)) and one half-sibling (HS (P1 x P3)). The axes represent the first and second principal components, which explain 2.1% and 1.9% of the variance, respectively.



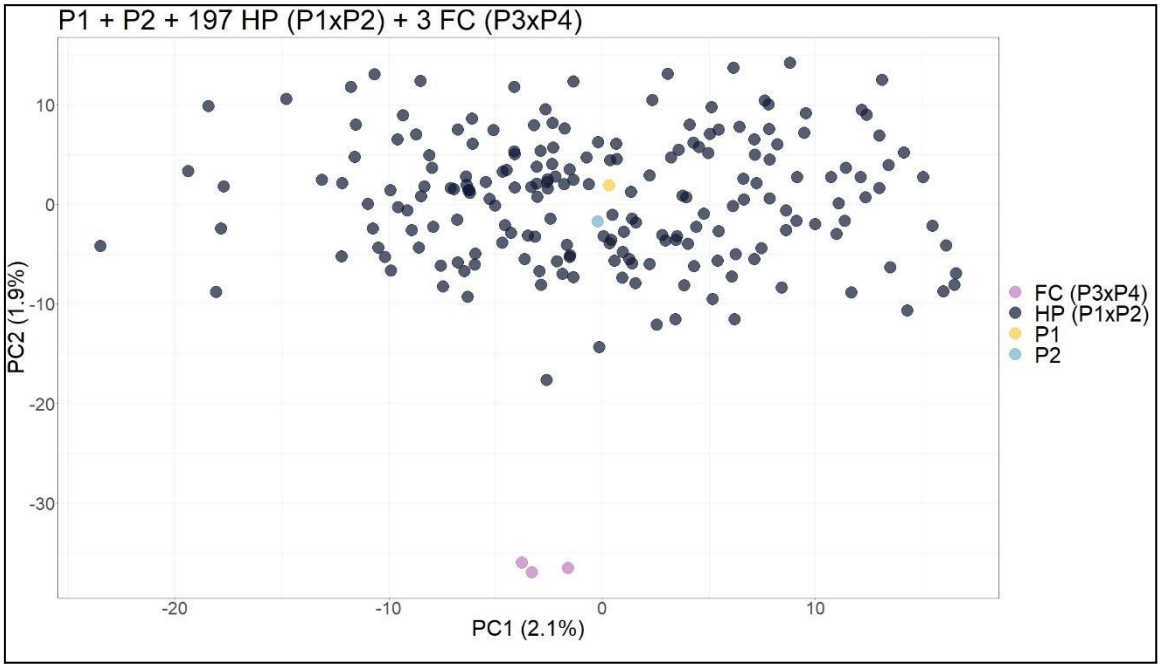
Supplementary Figure 4. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 199 hybrids (HP (P1 x P2)) and one full contaminant (FC (P3 x P4)). The axes represent the first and second principal components, which explain 2.1% and 1.9% of the variance, respectively.



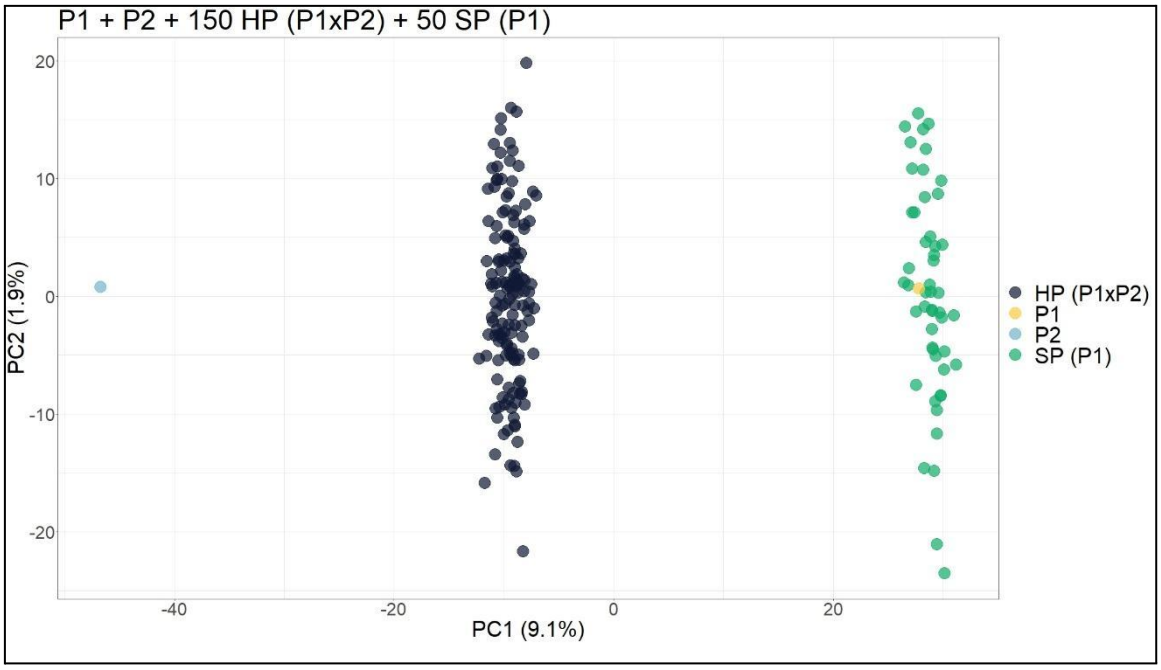
Supplementary Figure 5. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 196 hybrids (HP (P1 x P2)) and four self-fertilization progenies (SP (P1 x P1)). The axes represent the first and second principal components, which explain 2.2% and 2% of the variance, respectively.



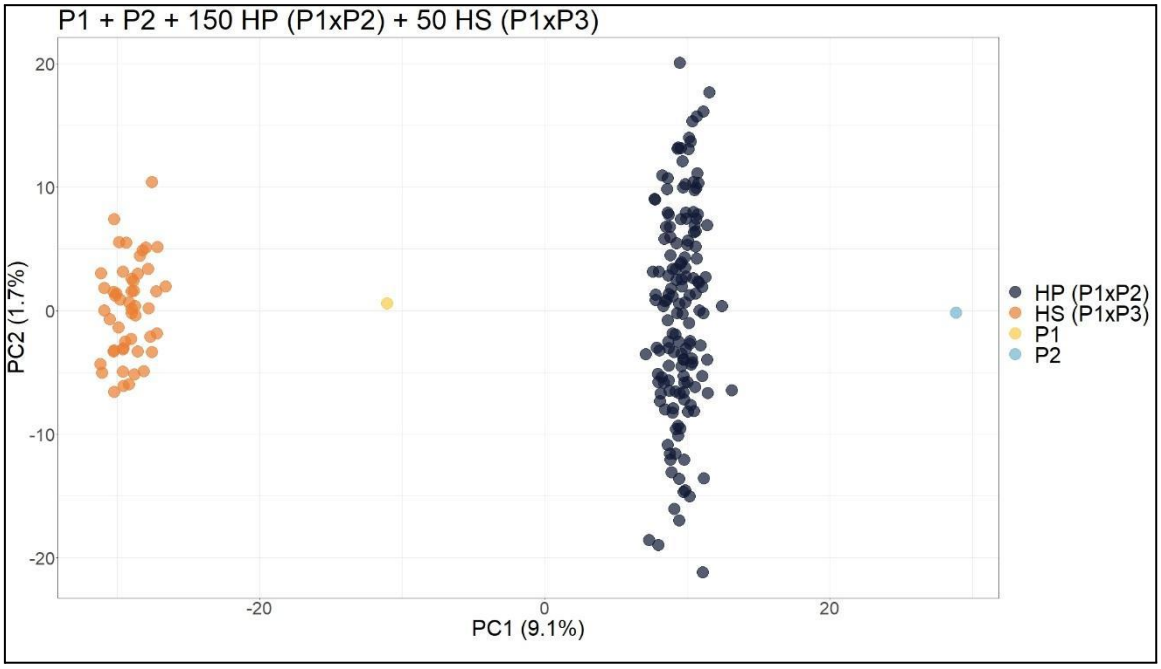
Supplementary Figure 6. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 194 hybrids (HP (P1 x P2)) and six half-siblings (HS (P1 x P3)). The axes represent the first and second principal components, which explain 2.1% and 2% of the variance, respectively.



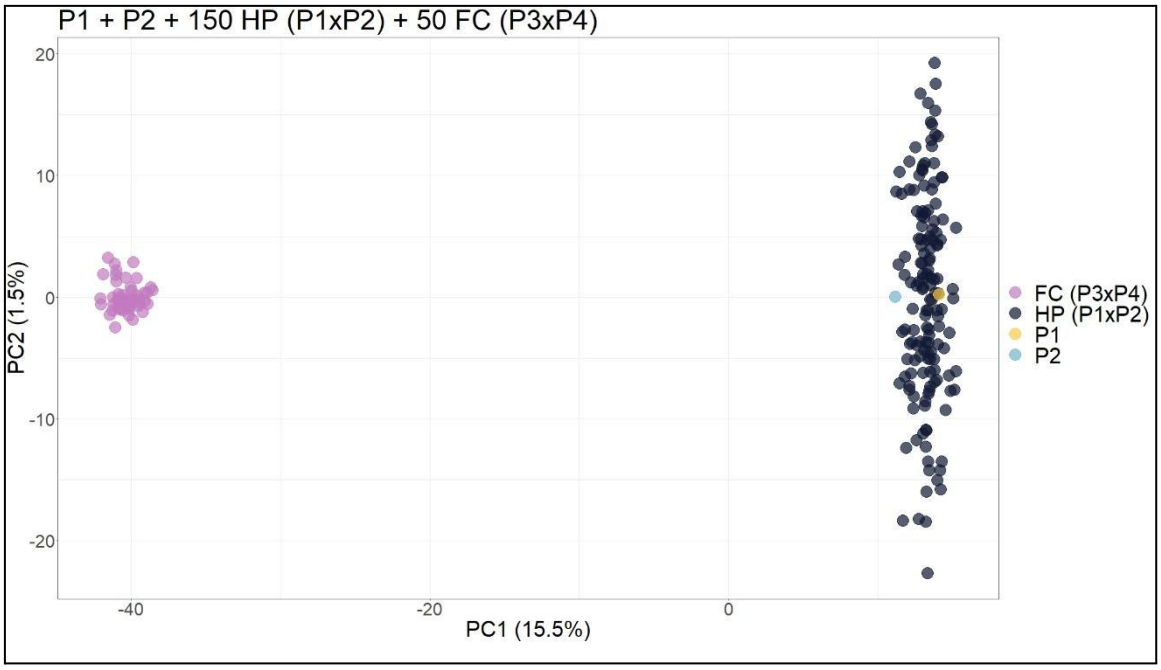
Supplementary Figure 7. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 197 hybrids (HP (P1 x P2)) and three full contaminants (FC (P3 x P4)). The axes represent the first and second principal components, which explain 2.1% and 1.9% of the variance, respectively.



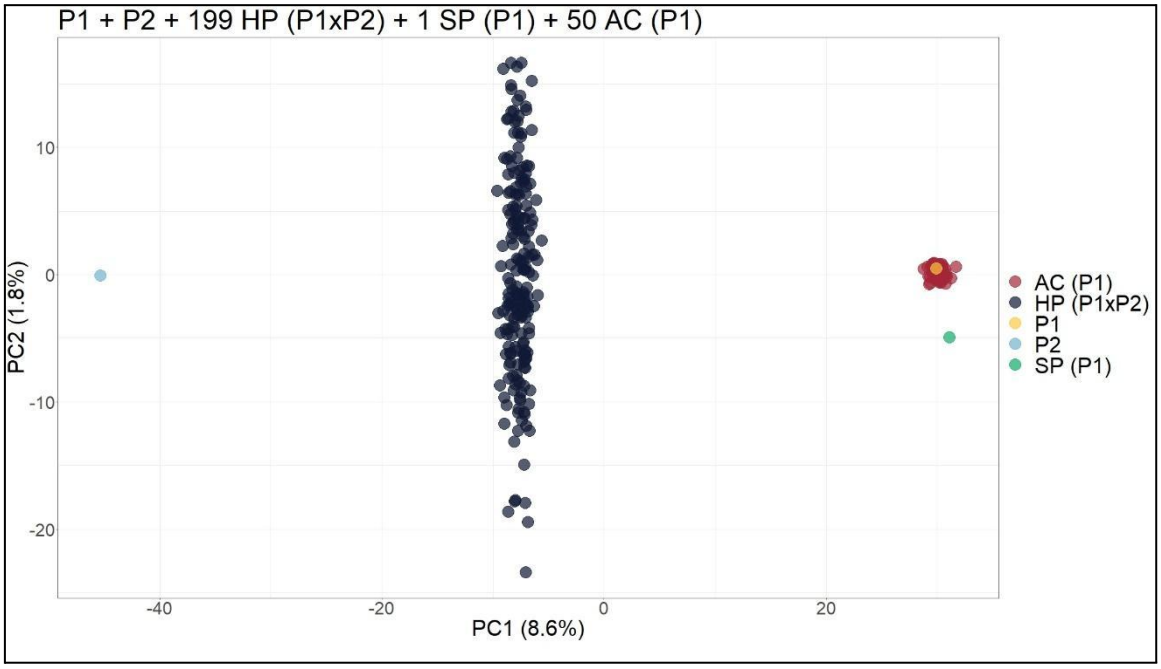
Supplementary Figure 8. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 150 hybrids (HP (P1 x P2)) and 50 self-fertilization progenies (SP (P1 x P1)). The axes represent the first and second principal components, which explain 9.1% and 1.9% of the variance, respectively.



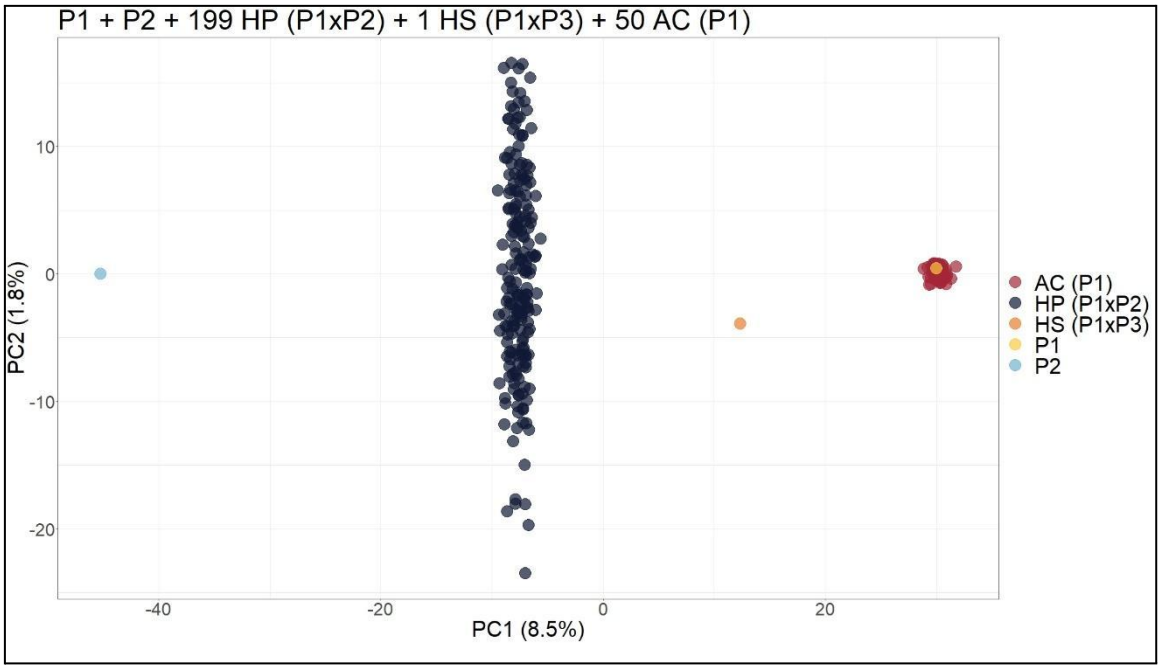
Supplementary Figure 9. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 150 hybrids (HP (P1 x P2)) and 50 half-siblings (HS (P1 x P3)). The axes represent the first and second principal components, which explain 9.1% and 1.7% of the variance, respectively.



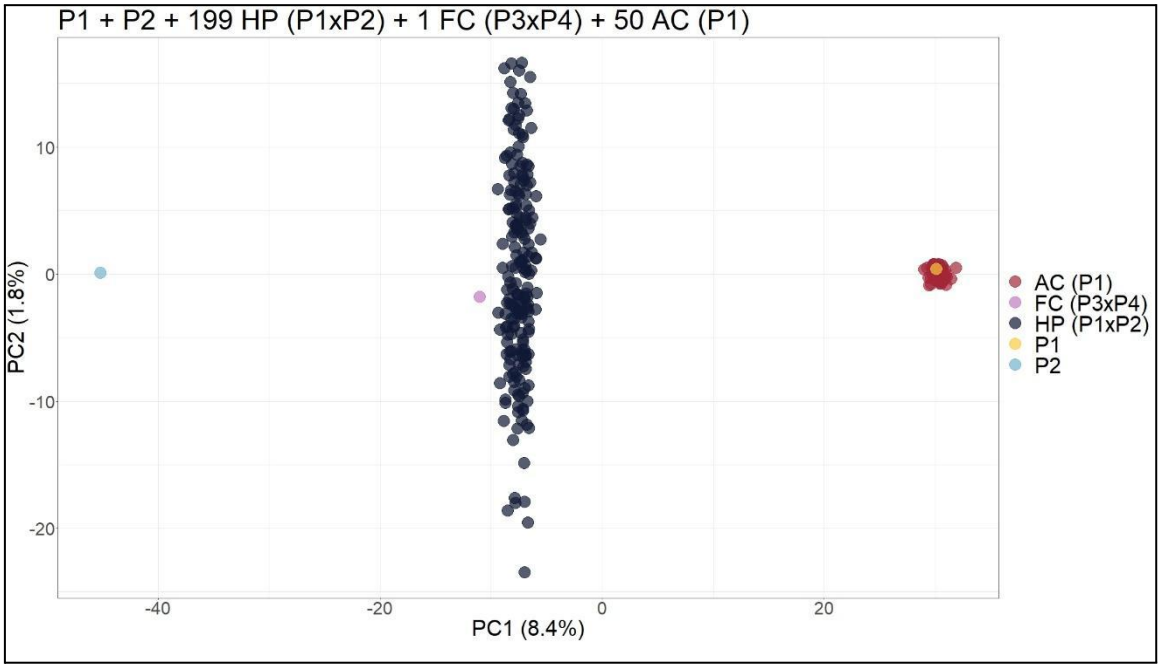
Supplementary Figure 10. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 150 hybrids (HP (P1 x P2)) and 50 full contaminants (FC (P3 x P4)). The axes represent the first and second principal components, which explain 15.5% and 1.5% of the variance, respectively.



Supplementary Figure 11. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 199 hybrids (HP (P1 x P2)), one self-fertilization (SP (P1 x P1)) and 50 apomictic clones (AC (P1)). The axes represent the first and second principal components, which explain 8.6% and 1.8% of the variance, respectively.

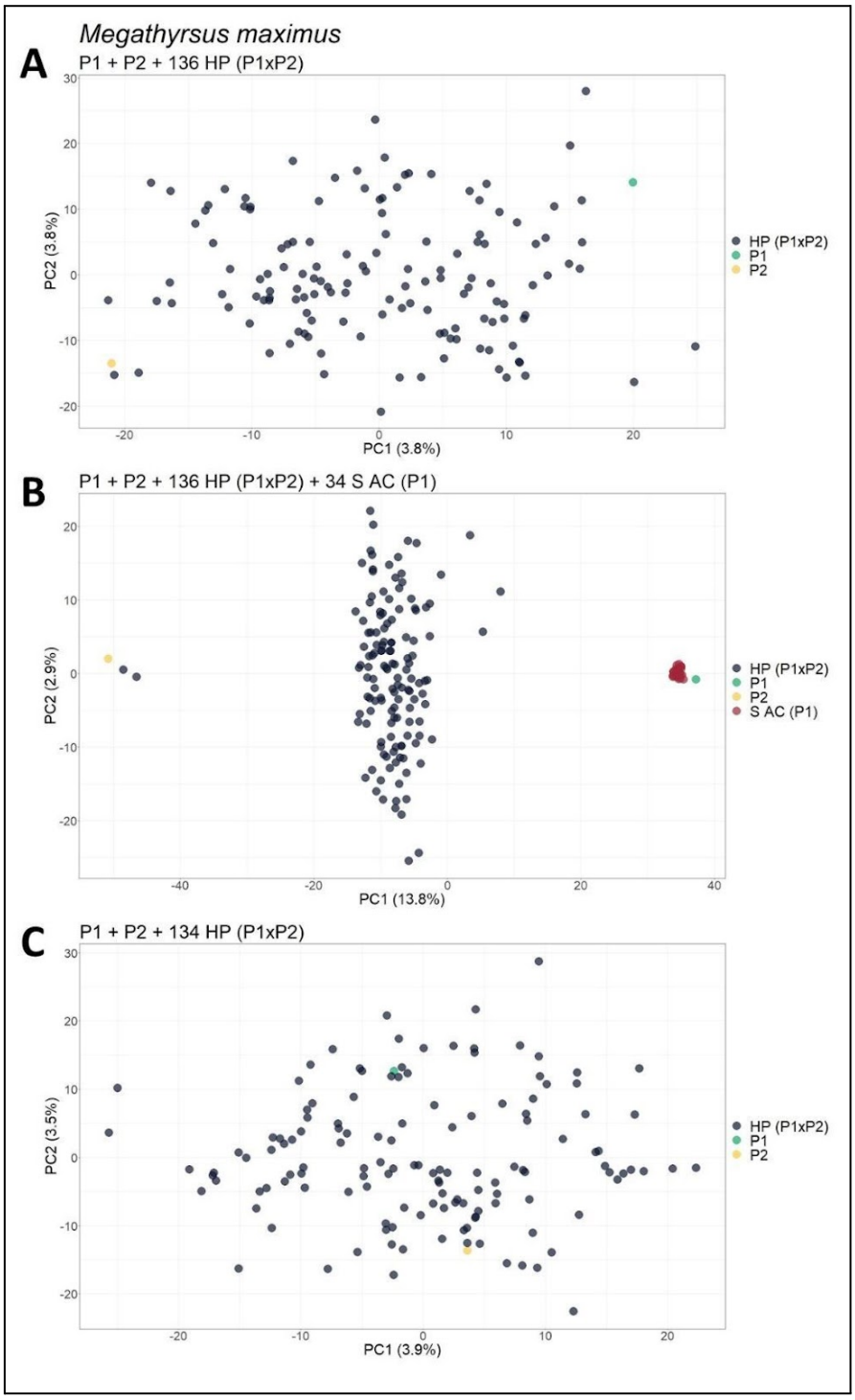


Supplementary Figure 12. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 199 HPs (P1 x P2), one half-sibling (HS (P1 x P3)) and 50 apomictic clones (AC (P1)). The axes represent the first and second principal components, which explain 8.5% and 1.8% of the variance, respectively.

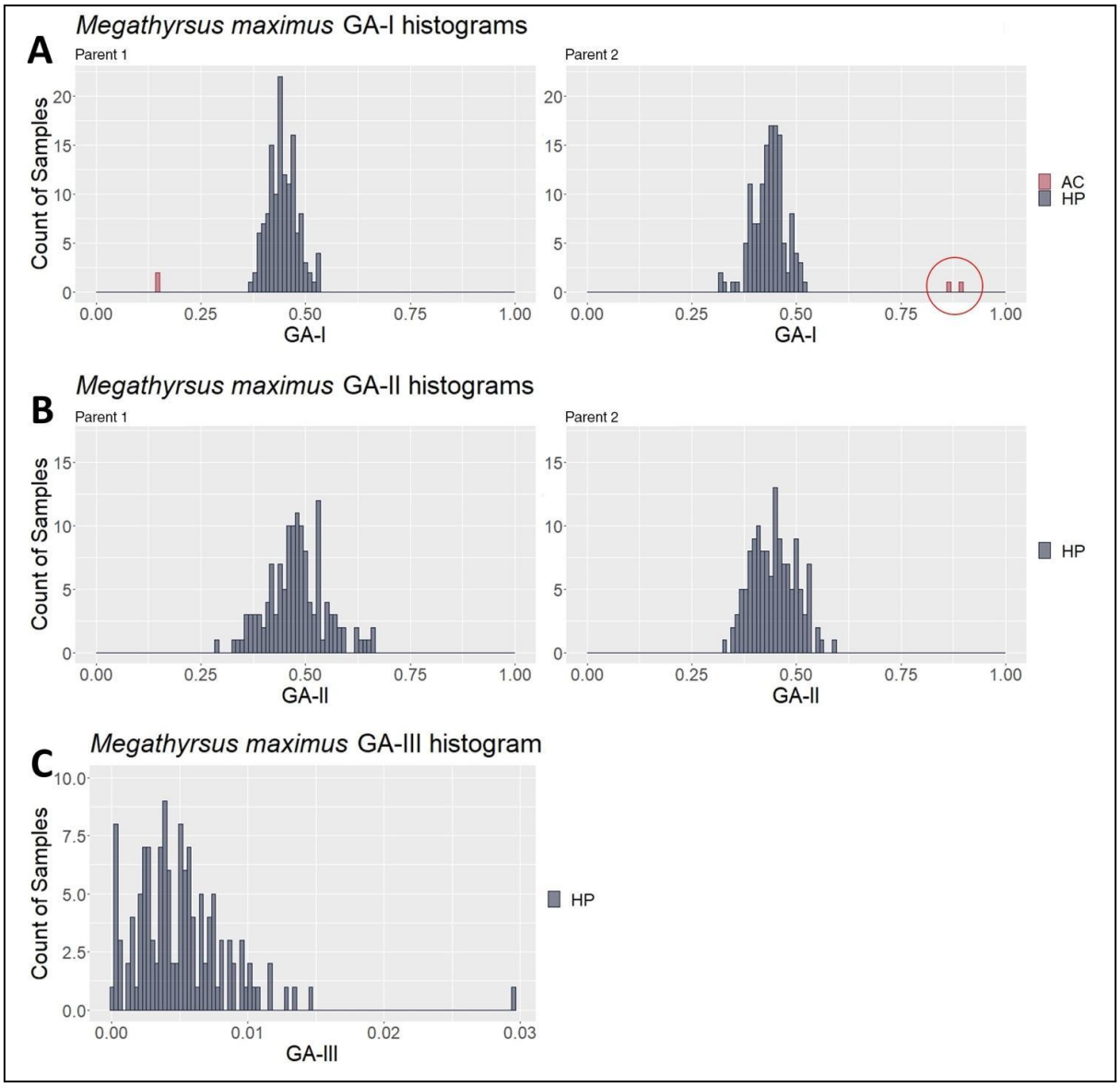


Supplementary Figure 13. Principal component analysis scatter plot showing the simulated population with two parents (P1 and P2), 199 HPs (P1 x P2), one full contaminant (FC (P3 x P4)) and 50 apomictic clones (AC (P1)). The axes represent the first and second principal components, which explain 8.4% and 1.8% of the variance, respectively.

2.3 Megathyrsus maximus

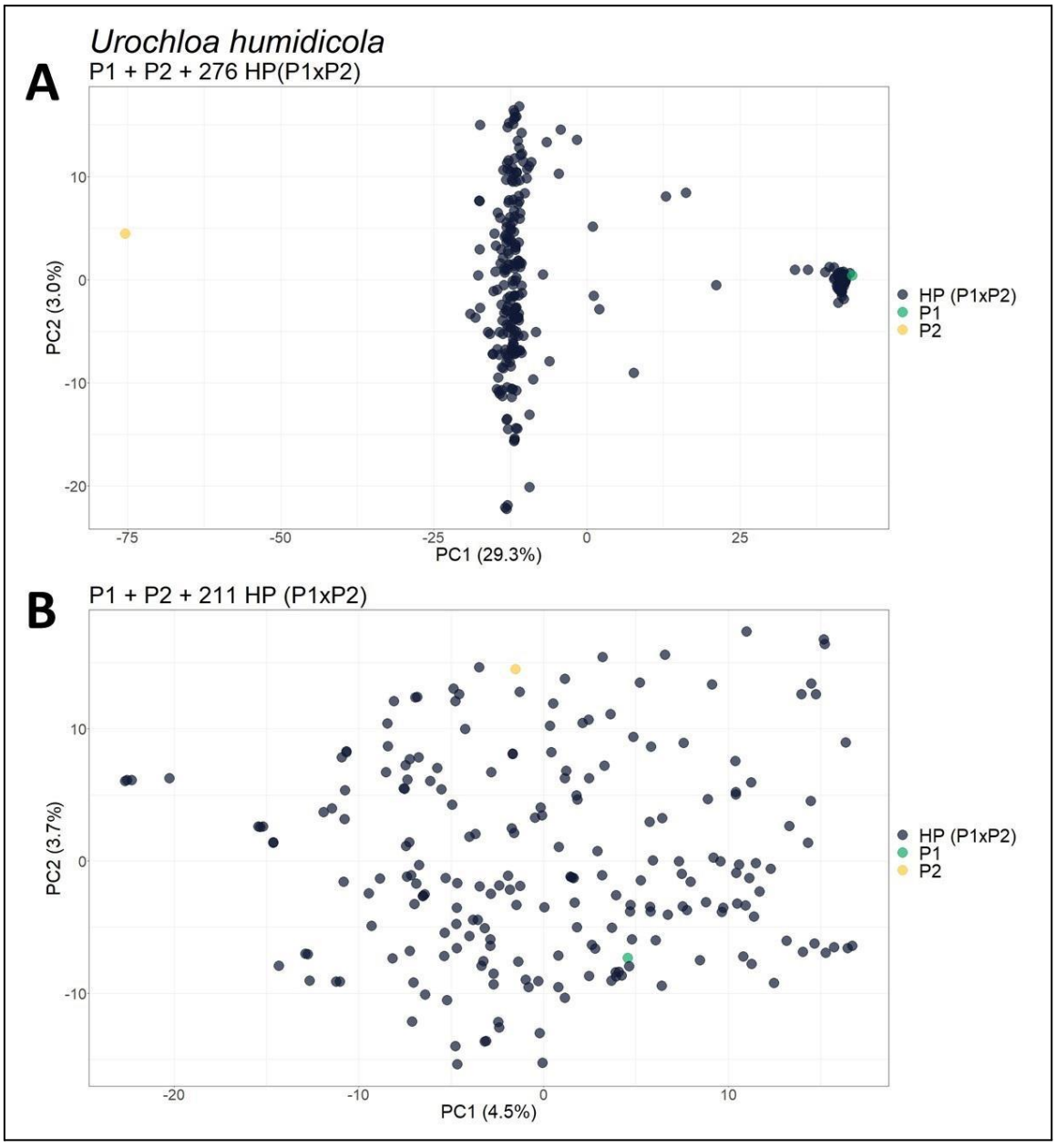


Supplementary Figure 14. Principal component analysis (PCA) scatter plots showing the *M. maximus* progeny for the set of SNP markers filtered by a *prop_mis* value of 0.20. (**A**) Original population composed of two parents (P1 and P2) and their progeny of 136 hybrids (HP (P1xP2)); (**B**) Population with 36 simulated apomictic clones (S AC (P1)); (**C**) Population without the 2 apomictic clones (AC) identified. The axes represent the first and second principal components, which explain 3.8% and 3.8% of the variance for (**A**), 13.8% and 2.9% for (**B**), and 3.9% and 3.5% for (**C**), respectively.

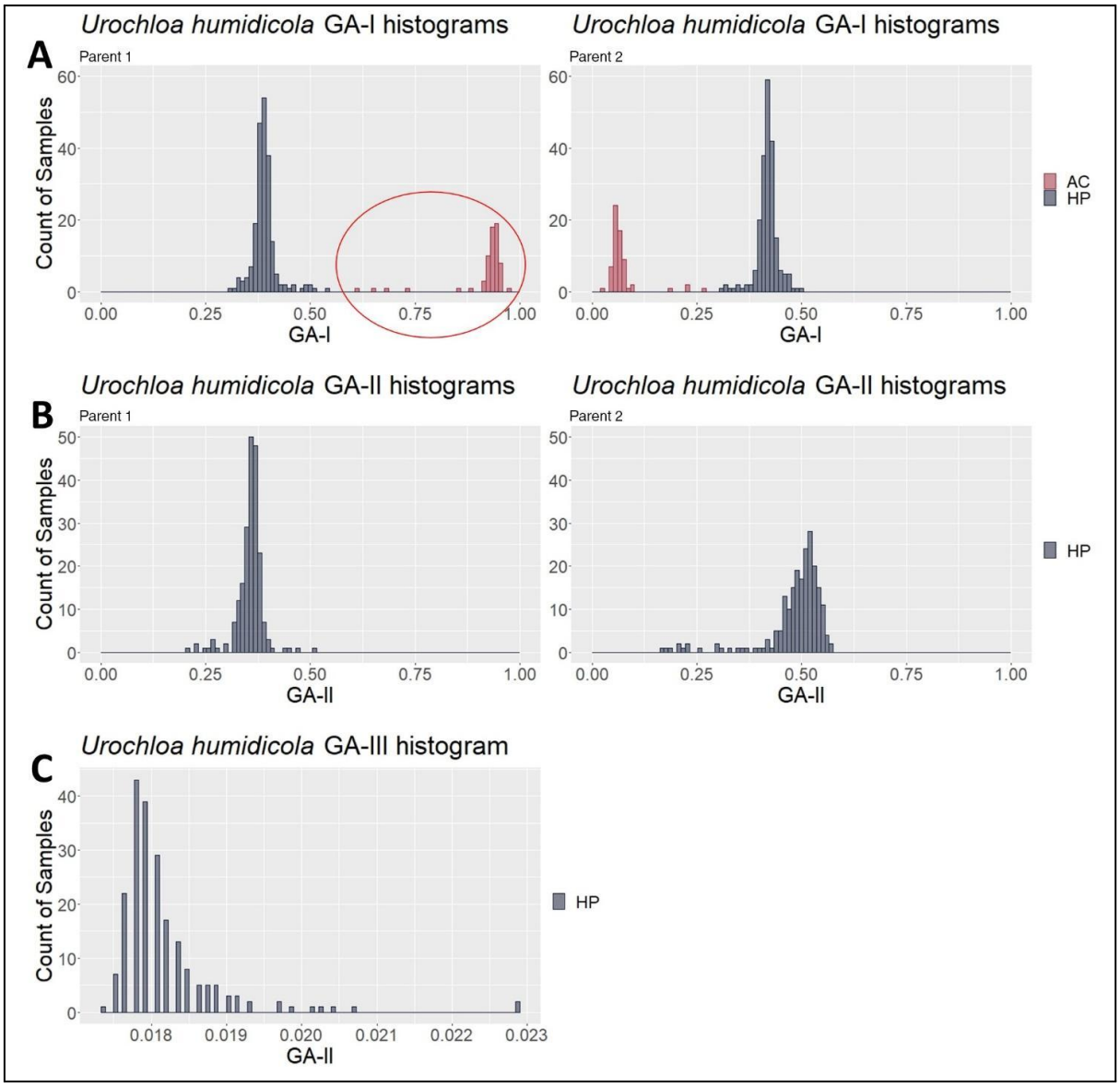


Supplementary Figure 15. *Megathyrsus maximus* GA histograms with samples classified by the two clusters obtained in the CA. (**A**), (**B**) and (**C**) show the results for GA-I, GA-II and GA-III, respectively. The red circle highlights the identified AC contaminants in contrast to the hybrid progeny (HP).

2.4 Urochloa humidicola



Supplementary Figure 16. Principal component analysis (PCA) scatter plots showing the *U. humidicola* progeny for the set of SNP markers filtered by a *prop_mis* value of 0.20. (**A**) Original population composed of two parents (P1 and P2) and their progeny of 276 hybrids (HP (P1xP2)); (**B**) Population without the 65 apomictic clones (AC) identified. The axes represent the first and second principal components, which explain 29.3% and 3.0% of the variance for (**A**) and 4.5% and 3.7% for (**B**), respectively.



Supplementary Figure 17. *Urochloa humidicola* GA histograms with samples classified by the two clusters obtained in the CA. (**A**), (**B**) and (**C**) show the results for GA-I, GA-II and GA-III, respectively. The red circle highlights the identified AC contaminants in contrast to the hybrid progeny (HP).

Supplementary Material References

Baker FB, Hubert LJ (1975). "Measuring the Power of Hierarchical Cluster Analysis." Journal of the American Statistical Association, 70(349), 31–38.

Ball GH, Hall DJ (1965). "ISODATA: A Novel Method of Data Analysis and Pattern Classification." Stanford Research Institute, Menlo Park. (NTIS No. AD 699616).

Beale EML (1969). Cluster Analysis. Scientific Control Systems, London.

Calinski T, Harabasz J (1974). "A Dendrite Method for Cluster Analysis." Communications in Statistics – Theory and Methods, 3(1), 1–27.

Davies DL, Bouldin DW (1979). “A Cluster Separation Measure.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

Duda RO, Hart PE (1973). Pattern Classification and Scene Analysis. John Wiley & Sons, New York.

Dunn J (1974). “Well Separated Clusters and Optimal Fuzzy Partitions.” Journal Cybernetics, 4(1), 95–104.

Edwards AWF, Cavalli-Sforza L (1965). “A Method for Cluster Analysis.” Biometrics, 21(2), 362–375.

Friedman HP, Rubin J (1967). “On Some Invariant Criteria for Grouping Data.” Journal of the American Statistical Association, 62(320), 1159–1178.

Halkidi M, Vazirgiannis M, Batistakis I (2000). “Quality Scheme Assessment in the Clustering Process.” In Principles of Data Mining and Knowledge Discovery, volume 1910 of Lecture Notes in Computer Science, pp. 265–276. Springer-Verlag, Berlin Heidelberg. Proceedings of the 4th European Conference, PKDD 2000, Lyon, France, September 13–16 2000.

Hartigan JA (1975). Clustering Algorithms. John Wiley & Sons, New York.

Hubert LJ, Levin JR (1976). “A General Statistical Framework for Assessing Categorical

Clustering in Free Recall.” Psychological Bulletin, 83(6), 1072–1080.

Kraemer HC (1982). Biserial Correlation. John Wiley & Sons. Reference taken from a SAS note about the BISERIAL macro on this Web Site: http://support.sas.com/kb/ 24/991.html.

Krzanowski WJ, Lai YT (1988). “A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering.” Biometrics, 44(1), 23–34.

Marriot FHC (1971). “Practical Problems in a Method of Cluster Analysis.” Biometrics, 27(3), 501–514.

McClain JO, Rao VR (1975). “CLUSTISZ: A Program to Test for The Quality of Clustering of a Set of Objects.” Journal of Marketing Research, 12(4), 456–460.

Milligan GW, Cooper MC (1985). “An Examination of Procedures for Determining the Number of Clusters in a Data Set.” Psychometrika, 50(2), 159–179.

Ratkowsky DA, Lance GN (1978). “A Criterion for Determining the Number of Groups in a Classification.” Australian Computer Journal, 10(3), 115–117.

Rohlf FJ (1974). “Methods of Comparing Classifications.” Annual Review of Ecology and

Systematics, 5, 101–113.

Rousseeuw P (1987). “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” Journal of Computational and Applied Mathematics, 20, 53–65.

Sarle WS (1983). “SAS Technical Report A-108, Cubic Clustering Criterion.” SAS Institute Inc. Cary, NC.

Scott AJ, Symons MJ (1971). “Clustering Methods Based on Likelihood Ratio Criteria.” Biometrics, 27(2), 387–397.

Tibshirani R, Walther G, Hastie T (2001). “Estimating the Number of Clusters in a Data Set Via the Gap Statistic.” Journal of the Royal Statistical Society B, 63(2), 411–423

CAPÍTULO II

Insights into the regulation of agronomic traits in *Urochloa ruziziensis* through multi-omic data integration.

Autores: Felipe Bitencourt Martins, Alexandre Hild Aono, Aline Costa Lima Moraes, Rebecca Caroline Ulbricht Ferreira, Marco Aurélio Caldas de Pinho Pessoa Filho, Mariana Rodrigues Motta, Mariane de Mendonça Vilela, Rosangela Maria Simeão e Anete Pereira de Souza.

Artigo ainda não publicado.

Abstract

Tropical forage grasses, especially species of the genus *Urochloa,* play an important role in cattle production being the main food source for the animals in the tropical/subtropical regions. Most of the species are apomictic and tetraploid, this gives special importance to *U. ruziziensis*, a sexual diploid species that can be tetraploidized to be used in interspecific crosses with apomictic species. As a means to auxiliate breeding programs, this study investigates the applicability of genome wide family prediction (GWFP) in *U. ruziziensis* half-sibling families to predict growth and biomass production. Machine learning and feature selection algorithms were used to reduce the number of markers necessary for prediction to approximately 11 markers and to enhance the predictive ability to approximately 0.9 across the phenotypes. Beyond that, to investigate the regulation of the agronomic traits, the position of the markers with more importance for the prediction were considered as putative QTLs, and in a multi-omic approach, genes obtained in the species transcriptome were mapped linked to those markers. Furthermore, a gene co-expression network was modeled enabling not only the investigation of the mapped genes but their co-expressed genes too. The functional annotation showed that the mapped genes are mainly associated with auxin transport and biosynthesis of lignin, flavonol and folic acid, while the co-expressed genes are associated with DNA metabolism, stress response and circadian rhythm. The results provided a viable marker-assisted breeding approach for tropical forages and identified target regions for future molecular studies on these agronomic traits.

1 - Introduction

Pastures composed of tropical forage grasses, with emphasis on grasses of the *Urochloa* genus, are the main food source for animals in the tropical and subtropical world, standing out in the economic sector related to the beef/milk and seed markets (Jank et al., 2014). The genetic improvement of *Urochloa* species is recent, started about 40 years ago, and challenging due to different polyploid levels, high heterozygosity, and predominantly mode of reproduction by apomixis (Ferreira et al., 2021; Simeão et al., 2021; Worthington et al., 2021). Among the main goals of the breeding programs are the launch of cultivars tolerant to biotic stresses, adapted to future climate changes, increased productivity and with better nutritional value to improve animal performance (Pereira et al., 2018a; Simeão et al., 2021).

These goals can be achieved more quickly with the inclusion of genomic selection (GS) in breeding cycles, which uses statistical models to perform genomic predictions (GP)

of plant performance from genetic markers, mainly single nucleotide polymorphisms (SNPs) (Daetwyler et al, 2013). Although the estimation of GP models has already been shown as feasible in other important polyploid crops (de Bem Oliveira et al., 2020; Pincot et al., 2020; Ferrão et al., 2021; Haile et al., 2021; Juliana et al., 2022; Petrasch et al., 2022), such a methodology has only recently started to be tested in *Urochloa* spp. (Matias et al., 2019a; Aono et al., 2022), and efforts must be made to achieve high- quality panels of markers and large-scale phenotyping (Simeão et al., 2021). Fortunately, two *Urochloa* spp. genomes came available recently, both of *U. ruziziensis* (2n=2x=18) (Pessoa-Filho et al., 2019; Worthington et al., 2021), enabling the detection of large numbers of SNPs with potential to improve model accuracy in GP analyses on *Urochloa* spp.

Traditionally, GP models employ a dense dataset of molecular markers to calculate genomic estimated breeding values (GEBV) at the individual level (Meuwissen et al., 2001). However, for *U. ruziziensis* and some other forage species, such as alfalfa and ryegrass, it is a common practice to employ the family (full or half-sibs) as the basic unit for phenotyping and selection (Simeão et al., 2012; Simeão et al., 2016a; Simeão et al., 2016b; Biazzi et al., 2017; Cericola et al., 2018; Jia et al., 2018; Andrade et al., 2022), making the development of genome-wide family prediction (GWFP) approaches more advantageous. By considering family-pools as the unit of measurement, genotyping efforts are reduced and, together with combined individual phenotypic values, the costs of developing GP models are also reduced (Zou et al., 2016; Rios et al., 2021; Murad Leite Andrade et al., 2022). In addition, the application of GWFP can increase the predictive ability and, consequently, affect the rate of genetic gains for complex traits, as reported in loblolly pine and alfalfa (Rios et al., 2021; Andrade et al., 2022).

To achieve family-pool markers, sequencing approaches can be employed to generate a large number of SNP markers (Elshire et al., 2011; Poland et al., 2012). Genotyping-by-Sequencing (GBS) is a low-cost and high-throughput genotyping method that can be used to identify SNPs even when no reference genome is available, although a reasonable sequencing depth must be needed to avoid too many missing data points (Thakral et al., 2022). GBS has been applied in several family-pool genotyping studies (Bélanger et al., 2016; Cericola et al., 2018; Schneider et al., 2022) because it obtains allele counts from the sequencing reads (Byrne et al.. 2013). This way, for family-pools GP and association studies, the allele counts can be directly used, without the necessity of estimating allelic dosages (Ashraf et al., 2014; Guo et al., 2018).

The incorporation of machine learning (ML) methods in GP has been controversial, with some studies showing advantages (Ma et al., 2018; Waldman et al., 2020; Aono et al., 2022), and others not (Montesinos-López et al., 2019; Zingaretti et al., 2020; Crossa et al., 2019). However, several studies have shown ML-based strategies to reduce marker density, including feature selection (FS) techniques to select polymorphisms associated with the phenotype, increasing the predictive power of GP methods (Li et al., 2018; Aono et al., 2020; Pimenta et al., 2021; Aono et al., 2022).

In contrast to predictions from GP models, association studies such as quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) attempt to find individual markers associated with larger amounts of genetic variation, in order to understand the genetic architecture of complex traits (Zhang et al., 2014). Although modeling different aspects, both approaches have complementary advantages, providing robust information to detect the potential candidate genes for agronomic important traits. Some methodologies originally used for GP were applied in GWAS to detect loci associated with the trait of interest (Goddard et al., 2016; Wang et al., 2020; Wolc et al., 2022). On the other hand, association studies have already shown to be useful to improve GP (Zhang et al., 2014; Bian et al., 2017; Jeong et al., 2020). However, it is yet to be better investigated how multiple sources of multi-omics data can provide information about the genes and biological processes associated with a group of selected markers that improved GP for a complex quantitative trait.

Once these polymorphisms are identified, other omics approaches can be used to ensure a better understanding of the biological relationships between genes related to complex traits (Scossa et al., 2021). Traditionally, data generated by multiple levels of biological information (genomics, transcriptomics, proteomics) were analyzed separately. However, in recent times, data integration followed by appropriate statistical analysis have become a tool with great potential in elucidating biological meaning of the studied traits in human (Yang et al, 2014), microorganisms (Borin et al., 2018; Rosolen et al., 2022), animals (Parker Gaddis et al., 2016; Mateescu et al., 2017), and plants (Francisco et al., 2021; Cardoso-Silva et al., 2022). Despite its economic importance and availability of some datasets, no study incorporating multi-omics has been carried out on *U. ruziziensis* nor any species of *Urochloa* spp.

Here, we developed a genome-wide family prediction (GWFP) approach for *U. ruziziensis* using different models and strategies, including ML and FS algorithms, and combined multi-omic approaches for a deeper investigation of genes and biological processes

involved in target traits for forage grass breeding programs. We hope that these results will provide insights about the genotype and phenotype associations, expression patterns and functional implications not only to *U. ruziziensis*, but to all tropical forage grasses.

2 - Materials and methods

The following sections describe an integrative methodology for analyzing phenotypic, genomic and transcriptomic data to investigate genes associated with agronomic traits in a breeding population of *Urochloa ruziziensis* (2n = 4x = 36) (Fig. 1).

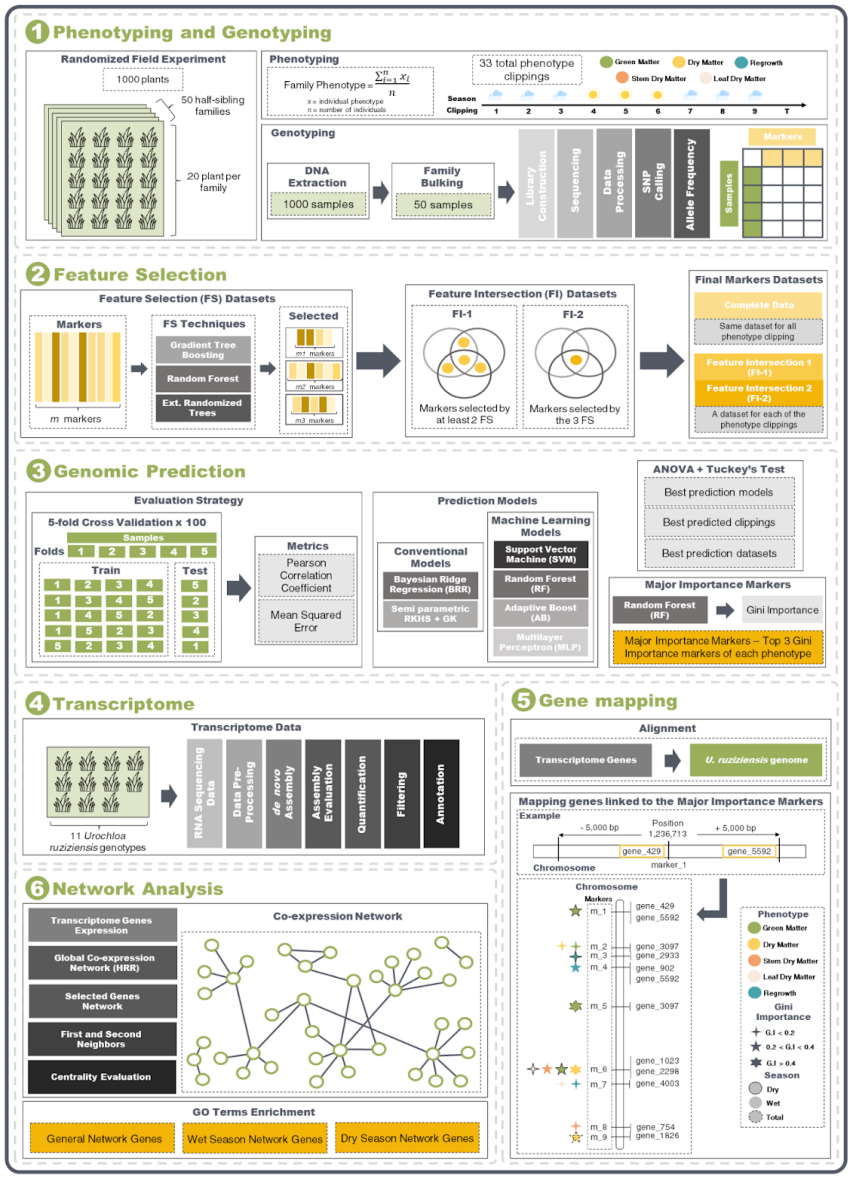


Figure 1. The approach established in this research can be divided into three main parts: (i) phenotyping and genotyping the population (1); (ii) identifying phenotypically associated markers through genomic prediction (2 and 3); and (iii) investigating the genes physically linked to the markers in a co-expression network (4, 5 and 6).

2.1 - *Urochloa ruziziensis* phenotyping

The progenies used in this study were generated as part of the *Urochloa* breeding program of the Brazilian Agricultural Research Corporation (Embrapa) Beef Cattle (EBC), Campo Grande, Mato Grosso do Sul State, Brazil (20∘27'S, 54∘37'W, 530m), as described by Simeão et al. (2012, 2016a,b). Briefly, seven sexual autotetraploid accessions (R30, R38, R41, R44, R46, R47 and R50) were replicated 20 times to compose an open pollination randomized field organized in 26 lines and 12 columns spaced by 2 meters. In 2012, 20 seeds from 59 plants with viable seed production and flowering synchrony were collected and used to compose the breeding progenies, totalizing 1180 half-sibling individuals. The experiment was planted in a randomized block design with one plant per plot, spaced by 1,5m (Simeão et al., 2016a,b). From the 59 half-sibling progenies, 50 were chosen based on the criteria of selecting the progenies with more plants that succeeded in the field.

The phenotypic evaluation was performed considering nine clippings at 15 cm height, being six of them in the wet season (1-3,7-9) and three in the dry season (4-6) where the climatological water balance was accessed through the AWC (Available Water Capacity) metric (Supplementary Fig. S1) (Simeao et al, 2016a,b). In addition to the nine clippings, we have a total (T) evaluation for each phenotype in the period. The agronomic traits evaluated in all clippings were: green matter yield (GM) and dry matter yield (DM), both in grams per family, and regrowth (RG), with scores varying from 0 to 6 as described by Figueiredo et al. (2012). In addition, in two clippings and T, about 200g of leaves and stems from each plant were used to estimate leaf dry matter yield (LDM) and stem dry matter yield (SDM). Considering that clipping 1 is discarded for the analysis, in total, we evaluated 33 different traits (clippings 2-9 and T for GM, DM and RG, clippings 2,5 and T for SDM and LDM).

Phenotype measures of each trait were evaluated according to the following mixed-effect model, implemented in the Selegen-REML/BLUP software (Resende, 2016):

$$y = \mu + Zm + e,$$

where y is the vector of the averaged phenotype by family, μ is the overall mean fitted as a fixed effect, $\boldsymbol{Z}$ is the incidence matrix of family effects, m is the vector of random family genetic effects and e is the vector of random error effects. The half-sib family narrow-sense heritability (h^2) was estimated based on:

$$h^2 = \frac{0.25Va}{0.25Va + \frac{\sigma}{r}},$$

in which *Va* is the additive variance, *r* is number of replications and:

$$\sigma = Ve+0.75Va,$$

where *Ve* is the error variance.

For creating GP models, we employed a family-based approach, where the mean trait value was calculated for each family; and to be able to compare the modeling results, all phenotype clippings were scaled between 0 and 1 using the Min-Max Scaler method. In order to perform a data descriptive analysis, we used boxplots to access the distribution and outliers, computed Pearson's correlation among all phenotype clippings and performed a Principal Component Analysis (PCA) to access population structure. The descriptive analysis was done in R (R Core Team, 2021), and all PCA and plots in the work were done using the package pcaMethods (Stacklies et al., 2007) and the package ggplot2 (Wickham & Chang, 2016), respectively.

2.2 - Genotyping

Genomic DNA of all individuals was extracted using the DNeasy Plant kit (QIAGEN), and pooled according to each family, totalizing 50 samples. Genotyping-by-sequencing (GBS) libraries were constructed following the method proposed by Poland et al. (2012) using a combination of a rarely cutting enzyme (EcoT22I) and a frequently cutting enzyme (MspI). Subsequently, libraries were sequenced as 150-bp single-end reads using the High Output v2 Kit (Illumina, San Diego, CA, USA) in a NextSeq 500 platform (Illumina, San Diego, CA, USA).

We performed the quality evaluation of GBS raw sequence reads using FastQC version 0.11.5 (Andrews, S. 2010) and SNP calling using the TASSEL-GBS pipeline (Glaubitz et al., 2014) modified for polyploids (Pereira et al., 2018b). The reads were aligned to the *U. ruziziensis* genome (Pessoa-Filho et al, 2019) using the BowTie 2.3.1 aligner (Langmead et al., 2009) and only reads aligned one time were kept. SNP markers were filtered using VCFtools v0.1.17 (Danecek et al., 2011) with the following criteria: a minimum sequencing depth of 20 reads, no more than 25% missing data per site, biallelic SNPs only, and removal of redundant (same genotypes in all samples) markers from the sets.

The allele frequency for each marker was estimated as the ratio between the number of reads for the alternative allele and the total number of reads. Missing data were replaced by the site mean of allele frequency. Furthermore, a principal component analysis (PCA) was performed in the complete genotype data to access population structure.

2.3 - Feature selection and genomic prediction

In order to create subsets of markers for each phenotype clipping, three FS techniques were applied to the SNP data using the Python 3 library scikit-learn with default parameters (Pedregosa et al., 2011): gradient tree boosting (FS-1) (Chen & Guestrin, 2016), extremely randomized trees (FS-2) (Geurts, Ernst & Wehenkel, 2008), and random forest (FS-3) (Breiman, 2001). Then, in order to perform the modeling, we created the Feature Intersection (FI) data sets by evaluating the intersection of the FS methods considering markers that were selected by at least two FS techniques (FI-1) and markers that were selected by all three FS techniques (FI-2), according to Aono et al. (2020) and Aono et al. (2022).

As GP strategies, we estimated different models considering six regression approaches across the 33 traits and the reduced (FI-1 and FI-2) and complete versions of the dataset. We used the following regression strategies for the creation of the models, two conventional genomic prediction models using 20,000 iterations with a thinning of 5 and 2,000 of burn-in: the semi-parametric reproducing kernel Hilbert space (RKHS) with a Gaussian kernel (GK) as the covariance function using the R package BGGE v0.6.5 (Granato et al., 2018); a Bayesian ridge regression (BRR) using the R package BGLR v1.0.9 (Perez & de los Campos, 2014)); and four ML algorithms using the Python 3 library scikit-learn v1.0.2 with default parameters (Pedregosa et al., 2011): support vector machine (SVM) (Cristianini & Shawe-Taylor, 2000), random forest (RF) (Breiman, 2001), adaptive boosting (AB) (Freund & Schapire, 1997), and multilayer perceptron (MLP) neural network (Popescu et al., 2009).

The evaluation of the previously described models for GP was performed using a k-fold (k=5) cross validation strategy, repeated 100 times and measuring two metrics: Predictive Ability (PA) as Pearson correlation coefficient and Mean Squared Error (MSE).

To compare the models, the phenotype clippings and the datasets, we used ANOVAs with multiple comparisons by Tukey's tests implemented in the agricolae R package (De Mendiburu & De Mendiburu, 2020). For PA we considered the best scenario the ones that in the Tukey's test had "a" or "a" combined with other letters, such, "ab" or "abc", which means the highest values. On the other hand, for MSE a scenario is better when its MSE value is

lower. So, we consider the best scenario as the ones with the higher letter, or combined with other letters (i.e., “f”, “ef” or “def”).

2.4 - Major Importance markers

Once the results showed the best data set of markers for each phenotype clipping, we used the Random Forest (Breiman, 2001) algorithm to model the data and obtain the impurity importance for each feature (markers). Also known as the Gini importance (GI), it is the normalized total reduction of the criterion brought by the feature, the sum of the features importance is equal to 1. To create a Major Importance set of markers, we selected the top 3 Gini importance (GI) markers in each phenotype clipping. If these three did not sum 0.5, the next ones were selected until the condition was satisfied. Furthermore, a Principal Component Analysis (PCA) was performed using the Major Importance markers dataset.

2.5 - Transcriptome assembly, quantification and annotation

A previous RNA-Seq data of 11 genotypes of *U. ruziziensis* was used to assess gene expression (Hanley et al., 2021; NCBI BioProject PRJNA513453). Raw data were quality-trimmed using Trimmomatic v0.39 (Bolger, Lohse & Usadel, 2014) with the following parameters: ILLUMINACLIP:IlluminaAdapters.fa:2:30:10, LEADING:3, TRAILING:3 SLIDINGWINDOW:4:20, MINLEN:75 and HEADCROP:12. Then, the good quality reads were *de novo* assembled by Trinity v2.5.1 (Grabherr et al., 2011) considering a minimum contig length (--min_contig_length) of 300 bases, and assembly integrity was evaluated using the Trinity.pl package utility.

SALMON 1.1.0 (Patro et al., 2017) was used to quantify the transcript expression, followed by the summarization of the counts at the gene level by tximport R package (Soneson et al., 2015). Only genes with more than one transcript per million (TPM) in at least three of the 11 samples were kept, generating the filtered assembly without low-level expression genes. The longest isoform for each gene was selected, and BUSCO v5.2.2 (Manni et al., 2021) was used to evaluate the annotation completeness against the Viridiplantae database. Finally, we aligned the filtered assembly to the UniProt database (Bateman et al., 2020) using Blastx and Blastn 2.10.0 (Altschul et al., 1990) with an e-value cutoff of 1e-10. Gene Ontology (GO) terms were retrieved using Trinotate software (Bryant et al., 2017), which performed the functional annotation.

2.6 - Genes linked with markers and GO enrichment

To identify genes physically linked to major importance markers (section 2.9), we aligned the genes obtained from the transcriptome assembly (section 2.10) against the *U. ruziziensis* genome (Pessoa-Filho et al., 2019). So, genes that aligned in a window of 5,000 bp up and downstream the marker position were noted as physically linked. The alignment was performed using Blastn 2.10.0 (Altschul et al., 1990) considering a coverage of 0.75 and e-value of 1e-6, and the linked genes were identified using R (R Core Team, 2021). To visualize the genes position in the genome, a physical map was constructed using MapChart v2.32 (Voorrips, 2002) with information of phenotype and season association, and Gini importance. In addition, a circular map was constructed using the R package circlize v0.4.14 (Gu et al., 2014) to show the associated genes that were duplicated.

Finally, in order to obtain a functional profile of the genes linked to the markers, a biological process GO terms enrichment was performed considering only terms common to *Arabidopsis thaliana*. This step was achieved with R package topGO (Alexa & Rahnenfuhrer, 2022), and GO terms with p-value < 0.01 were considered significantly enriched.

2.7 - Co-expression network

We modeled a gene co-expression network (GCN) using the transcript quantifications normalized in transcript per million (TPM) and the highest reciprocal rank (HRR) (Mutwil et al., 2010) approach, considering a limit of 30 edges. From the GCN, we selected the genes associated with the agronomic traits; also including highly correlated genes which were not considered in the network ranking (Pearson correlation coefficient ≥ 0.9 and a maximum p-value of 0.01 with Bonferroni correction). From this set established, we selected the first and second gene neighbors in the GCN.

The network visualization and evaluation were performed using the Cytoscape software v3.9.1 (Shannon et al., 2003). For each gene, we calculated the degree centrality measure as Barabási & Oltvai, 2004. Lastly, biological process GO term enrichments were performed for the selected genes, including first and second neighbors, to produce a general and seasonal functional profile of the metabolic pathways associated to the agronomics traits, same method of 2.6.

3 - Results

3.1 - Phenotypic and Genotypic Data Analyses

In our study, we evaluated five important phenotypes for forage grasses (GM, DM, RG, LMD, and STM) in different clippings, selected based on wet and dry seasons. Individual measures were averaged at subfamily level. The descriptive analysis of subfamily-based phenotypic data did not reveal any apparent patterns concerning the dispersion and skewness of the traits (Supplementary Fig. S2). No outliers were detected in 17 of the 33 traits. Even without any evident dispersion similarity between the phenotypes evaluated, the correlation analysis showed significant values for all the comparisons performed (Supplementary Fig. S3). We observed an average R Pearson correlation coefficient of 0.72 (Supplementary Fig. S3), with the highest correlations (~1) between the same clippings of GM and DM. Furthermore, early clippings (2 and 3) tended to be less correlated to all other measures. This scenario was more evident in GM, DM and SDM. In contrast, SDM-2 was the trait less correlated with the other phenotypes (Supplementary Fig. S3). The progeny means narrow-sense heritabilities for all phenotype clipings showed a mean value of 0.79, ranging from 0.44 (SDM-2) to 0.92 (LDM-5) (Supplementary Table S1).

GBS experiments generated ~720 million reads, which were converted into 1.3 million tags using the Tassel pipeline. A total of 77,413 SNP markers were identified in this step, of which 28,106 were retained after quality filters, allele frequency estimation, and missing genotype imputation. This final dataset of markers was denominated complete data (CD).

By using the phenotypic and genotypic data, we performed PCAs, plotting the dispersion of subfamilies using the scores of the two first principal components (PCs) (Supplementary Figs. S4-S5). Although arising from different sources of variation (the proportion of variance explained by the first two PCs was 85.2% and 57.2% for the phenotypic and genotypic data, respectively), similar patterns could be observed. For corroborating such a similarity, we colored the samples from the genotypic PCA scatter plot using the PC1 of the phenotypic data. Even without a pronounced presence of 3 groups, as the phenotypic PCA, the coloring in the genotypic PCA evidenced a clear association between both PCA results (Supplementary Fig. S5).

3.2 - Family Based Genomic Prediction

The predictive performances for the 33 traits at family level using the CD were evaluated by considering two conventional genomic prediction models (RKHS and BRR). Applying a 100 times 5-fold CV strategy, the RKHS model showed slight superior results in respect to BRR, with a mean PA of ~0.762 and mean MSE of ~0.025, contrasted to a mean PA of ~0.745 and a mean MSE of 0.026 in BRR. We observed a maximum PA of ~0.875 in the trait DM-8, and a minimum PA of ~0.490 in SDM-2. Aiming to obtain higher performances, four ML algorithms (SVM, RF, AB and MLP) were evaluated. Among the ML models, SVM had the best overall performance (PA mean of ~0.759 and MSE mean of ~0.026), however not superior to the RKHS approach. By considering the Tuckey's test results for MSE, SVM was highly overcome by the RKHS model, being considered the best approach in 13 traits while RKHS was considered in 30 (Supplementary Table S2-4). Our results indicate that when using CD for prediction, the ML algorithms were not able to outperform the conventional models.

In order to increase our predictive accuracies and evaluate putative trait-marker associations, specific subsets of SNPs were selected for each one of the 33 traits based on the intersections established between FS sets. Each one of the FS approaches generated a different quantity of markers: FS-1 selected sets with quantities ranging from 129 to 175 markers (mean of ~150, 0.53% of the CD); FS-2 from 484 to 1154 (mean of ~848, 3% of the CD); and FS-3 from 563 to 853 (mean of ~699, 2.5% of the CD). By considering the intersection approaches selected, we obtained FI-1 with SNP quantities ranging from 76 to 122 markers (mean of ~102, 0.36% of the CD) and FI-2 with quantities varying from 5 to 23 markers (mean ~11, 0.04% of the CD) (Supplementary Table S5). In addition to obtaining more restricted sets, these markers selected by FI have more evidence of trait associations, as they were selected by more than one algorithm. In this sense, model performances using the CD were contrasted with the use of models created from the data sets selected by FI-1 and FI-2.

The employment of the FI datasets increased the performance of all models for all traits. This increase was prominent in the models AB and RF, which presented the highest accuracies observed, overcoming RKHS in both FI sets. In respect to the six models evaluated, the FI-1 approach presented an improved overall performance when compared to FI-2, being considered by the Tuckey's test the best approach in 168 (FI-2 = 100) and 136 (FI-2 = 89) scenarios for PA and MSE respectively (Supplementary Table S6). However, individual results for the best models in each scenario were similar, as indicated by the best

model in FI-1 (AB with a mean PA of ~0.894 and a mean MSE of ~0.013) and FI-2 (RF with a mean PA of ~0.893 and a mean MSE of ~0.013) (Supplementary Table S2-4). Furthermore, when analyzing the clippings of a phenotype, we saw that the best performances for clippings in AB-FI-1 and RF-FI-2 varied in GM and DM, but for RG (clipping 3), SDM (clipping 5) and LDM (clipping 5) the results were equivalent (Supplementary Table S2-3 and 7).

In this sense, we observed that, for the prediction task, the combinations AB-FI-1 and RF-FI-2 can be employed with similar performances. However, for investigating trait-marker associations and catalog putative associated genes, the FI-2 represents a more restrictive approach. With sets (mean of ~11 markers) approximately ten times smaller than the sets of FI-1 (mean of ~102 markers), FI-2 markers provide a group of markers with a probable decreased number of false positive associations. Therefore, we considered the combination RF-FI-2 as the most promising approach to be employed in our datasets. In addition to a significant decrease in marker density through FI-2, the RF algorithm demonstrated high efficiency for prediction with an increase of 5.9% when compared to the RKHS PA using the FI-2 dataset or 17% when compared to the RKHS using the CD dataset. (Table 1).

Table 1. Comparison of RKHS and RF models predictive ability and mean squared error for all phenotype clippings using the FI-2 data sets.

Phenotype			Green Mater				Dry Mater				Leaf Dry Mater				Stem Dry Mater				Regrowth			
Dataset			RKHS	RF	Diff	%	RKHS	RF	Diff	%	RKHS	RF	Diff	%	RKHS	RF	Diff	%	RKHS	RF	Diff	%
Clipping	2	Predictive ability	0.857	0.850	-0.008	-0.9%	0.537	0.638	0.101	18.8%	0.835	0.875	0.040	4.8%	0.664	0.818	0.154	23.2%	0.868	0.896	0.029	3.3%
	3		0.870	0.869	-0.001	-0.1%	0.802	0.840	0.038	4.8%									0.940	0.956	0.017	1.8%
	4		0.894	0.912	0.018	2.0%	0.935	0.941	0.007	0.7%									0.859	0.911	0.051	6.0%
	5		0.866	0.917	0.051	5.9%	0.882	0.895	0.013	1.5%	0.897	0.924	0.027	3.0%	0.904	0.921	0.017	2.6%	0.826	0.869	0.043	5.2%
	6		0.863	0.903	0.040	4.6%	0.884	0.917	0.033	3.8%									0.833	0.884	0.051	6.1%
	7		0.844	0.923	0.079	9.3%	0.849	0.913	0.064	7.5%									0.862	0.896	0.034	4.0%
	8		0.881	0.913	0.032	3.6%	0.911	0.915	0.004	0.5%									0.860	0.872	0.012	1.4%
	9		0.897	0.944	0.047	5.2%	0.787	0.925	0.138	17.6%									0.813	0.843	0.030	3.7%
	T		0.867	0.937	0.070	8.1%	0.912	0.943	0.031	3.5%	0.900	0.952	0.052	5.8%	0.694	0.835	0.141	21.3%	0.886	0.932	0.046	5.2%
	2	Mean squared error	0.018	0.018	0.000	2.0%	0.044	0.035	-0.009	-21.1%	0.019	0.014	-0.005	-25.5%	0.033	0.021	-0.013	-37.5%	0.015	0.014	-0.001	-7.7%
	3		0.016	0.017	0.000	2.7%	0.014	0.013	0.000	-1.2%									0.011	0.008	-0.003	-27.9%
	4		0.015	0.013	-0.002	-10.5%	0.010	0.009	-0.001	-11.4%									0.017	0.012	-0.004	-26.1%
	5		0.016	0.012	-0.004	-25.6%	0.014	0.015	0.000	1.8%	0.012	0.011	-0.001	-8.4%	0.011	0.012	0.001	10.4%	0.016	0.013	-0.003	-19.8%
	6		0.018	0.012	-0.007	-36.0%	0.016	0.011	-0.005	-32.8%									0.024	0.016	-0.008	-32.1%
	7		0.019	0.010	-0.008	-44.4%	0.017	0.011	-0.006	-35.3%									0.015	0.011	-0.004	-28.0%
	8		0.020	0.015	-0.005	-25.6%	0.016	0.016	0.000	0.8%									0.014	0.013	-0.001	-8.7%
	9		0.012	0.008	-0.004	-36.2%	0.022	0.011	-0.011	-51.1%									0.017	0.014	-0.002	-14.5%
	T		0.018	0.009	-0.009	-48.0%	0.013	0.009	-0.004	-33.2%	0.011	0.007	-0.005	-41.4%	0.025	0.015	-0.009	-37.8%	0.013	0.008	-0.005	-35.9%

3.3 - Major Importance markers

Considering that the FS strategies employed in our study were based on ML algorithms estimated through a combination of decision trees, and also that the best model performances observed for FI-1 and FI-2 were AB and RF, respectively, we performed an additional approach to evaluate marker-trait associations based on decision tree structures: ranking the markers according to the RF scores estimated using the FI-2 selected markers. We selected the top three Gini importance markers for each trait, and if the top three did not present an importance sum of at least 0.5 (from a total of 1.0), we selected the next in the ranking until we achieved half of the total importance. In this way, we listed the markers with

the highest feature importances and avoided importance underrepresentation in traits. From all the FI-selected markers (283 markers) across the 33 traits, we could highlight a set of 69 markers with major prediction relevances, where only for SDM clipping 5 it was necessary to select four markers instead of three (Supplementary Table S8).

Furthermore, we performed a PCA to evaluate the subfamilies' dispersion considering this set of 69 major importance markers. The first two PCs explained 67.5% of the data variance, an intermediate value between the complete set of SNPs (57.2%) and the phenotypic data (85.2%) (Supplementary Fig. S6). Although the values of the first PCs seem to be inverted in such a PCA when compared to other ones performed, we observed a similar dispersion pattern (Supplementary Figs. S4 and S5). As we expected, the scatter plot displayed a group formation visually closer to the phenotypic PCA. Since the markers were selected through associations with the traits, there was a strong relation between the major importance data PC1 and the samples colored using the phenotypic PC1 values (Supplementary Fig. S7).

For evaluating the physical distribution of the FS selected markers, we created a *U. ruziziensis* physical map using the values obtained in the species genome. In addition to the set of 69 major importance selected markers, we also included in the map constructed all the FI-2 markers (Fig. 2). In regard to the markers distribution, we identified associations in all chromosomes with no clear pattern, except that there were extensive regions with little or no markers in the central region of chromosomes 1, 2, 5, 7 and 8, which we speculate as centromeric regions (Fig. 2). Chromosome 1 had more associations considering both FI and major importance sets, with little variation in representativeness, but it was the one with more SNPs identified too (Table 2). The smallest presence of associations was in Chromosomes 5 and 6 and the highest change in representativeness was in Chromosome 4 which had a reduction of 7% from FI to Major Importance (Table 2). Beyond that, especially in Chromosomes 1, 4 and 7, we observed regions with high density of minor importance markers around major importance markers, which may indicate that those are QTLs regions associated with the agronomic traits.

The Major Importance set was composed of various markers associated with more than one trait. As evidenced in the physical map, the marker associated with more traits clippings is in Chromosome 7, position 42,826,434. This marker was associated with four of the five phenotypes evaluated and was selected for nine clippings, of which in three it had Gini importance higher than 0.4 and in six GI between 0.2 and 0.4 (Fig. 2). Other markers were associated with various traits clippings, as a marker in Chromosome 1 position

69,834,400, that was associated with six traits clippings, and other three markers with four associations in Chromosome 1 and 3 (Fig. 2).

When considering the markers for each of the five traits without separating them by clippings, we accessed the sets intersections to quantify markers associated with more than one trait (Supplementary Fig. S7). Despite the difference in quantity between FI-2 and Major Importance sets, the logic relation among the traits sets was maintained: GM, DM and LDM were the phenotypes that shared more markers, while RG and SDM had proportionally more exclusive markers. Interestingly, SDM and RG were the traits with generally less correlation to the other traits too.

Table 2. Number and percentage of SNP markers identified/selected in each chromosome considering the Complete Data (CD), Feature Intersection (FI-2) and Top Gini Importance datasets.

Chromosome	Complete Data	Feature Intersection - 2	Top Gini Importance
1	4722 (16.8 %)	69 (23.4 %)	16 (23.2 %)
2	3565 (12.7 %)	36 (12.2 %)	10 (14.5 %)
3	3249 (11.6 %)	28 (9.5 %)	9 (13 %)
4	3552 (12.6 %)	42 (14.2 %)	5 (7.2 %)
5	1384 (4.9 %)	15 (5.1 %)	4 (5.8 %)
6	2508 (8.9 %)	23 (7.8 %)	2 (2.9 %)
7	3796 (13.5 %)	37 (12.5 %)	8 (11.6 %)
8	2127 (7.6 %)	18 (6.1 %)	8 (11.6 %)
9	2426 (8.6 %)	22 (7.5 %)	5 (7.2 %)
Scaffolds	777 (2.8 %)	5 (1,7 %)	2 (2.9 %)
Total	28106 (100 %)	295 (100 %)	69 (100%)

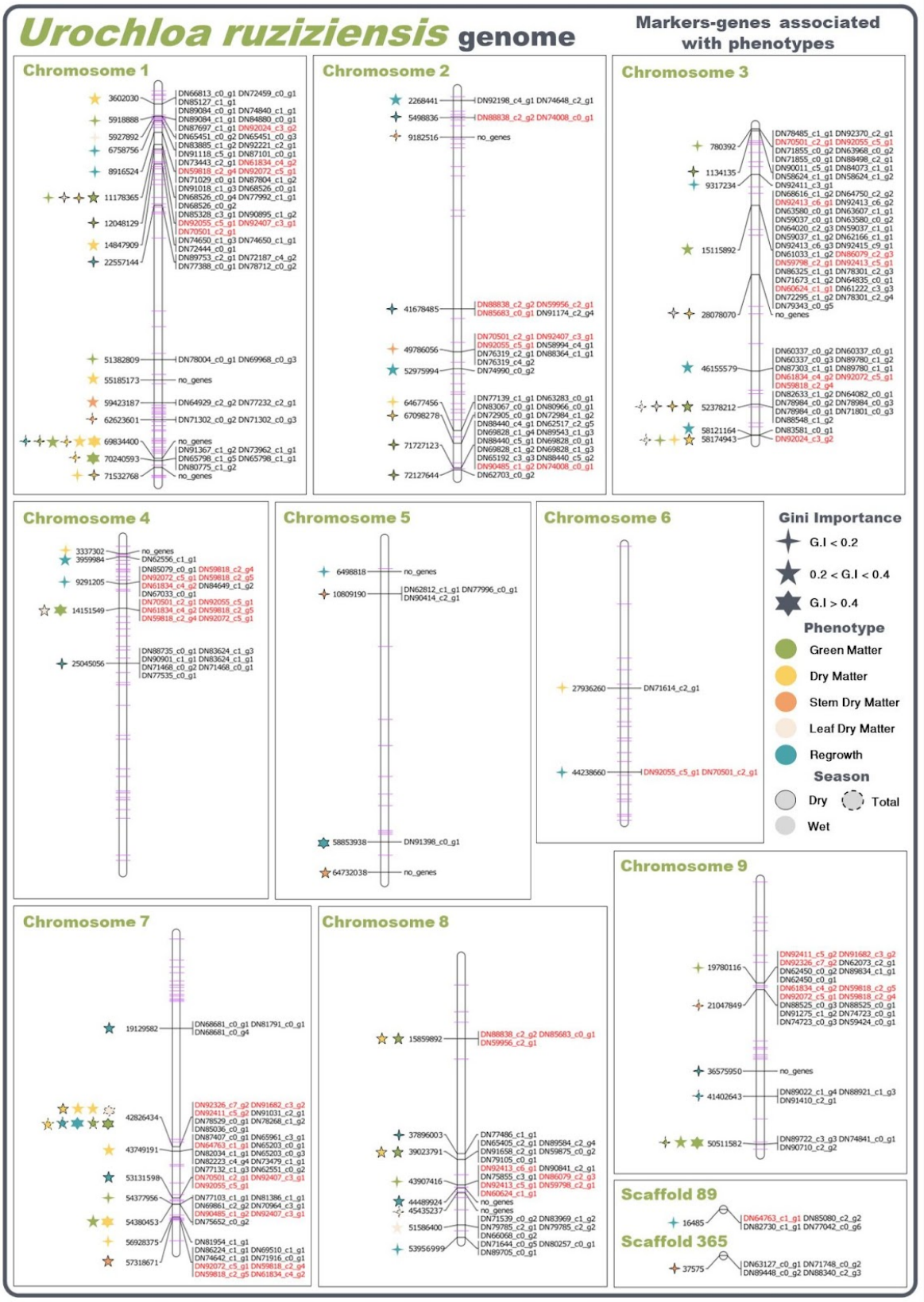


Figure 2. Physical map with the markers and genes associated with the phenotypes evaluated in the *Urochloa ruziziensis* population, with Gini Importance (GI) and Season indicated. Duplicated genes and minor importance markers (FI-2) mapped are represented in red and purple, respectively.

3.4 - Markers-genes associated with phenotypes.

To obtain insights into the agronomic traits regulation, we proceed to the identification of genes associated with the major importance markers. From the alignment of transcripts (Supplementary Results 2.1) with the reference genome of *U. ruziziensis* and, considering a window of 5,000 bp up/down stream the markers positions, we mapped a total of 217 genes (264 considering genes with multiple copies) physically close to 58 markers (Fig. 2 and Supplementary Table S8). We did not detect genes linked to all markers, as in Chromosome 1, where there were no genes in a marker region associated with six traits clippings, and in Chromosome 5, which from the four major importance markers, two had no linked genes (Fig. 2).

As previously stated, we found genes with multiple copies linked to more than one major importance marker region. There were 22 genes in this condition, which are highlighted in red in Fig. 2. For a better investigation of those genes, we mapped them in a circular map showing how the copies are positioned in the genome and united the copies information about trait/season associations and gini importance (Fig. 3). We identified five genes with six copies, three and two of them always appear together, and simultaneously, they are associated with seven traits clippings. Furthermore, we identified genes with 4, 3 and 2 copies associated to all evaluated traits and with different importances, demonstrating no clear pattern.

Regarding the functional annotation of the genes associated to the phenotypes, which is a subset of the complete transcriptome annotation obtained through the Trinotate and the UniProt database (Supplementary Results 2.1), we identified proteins/enzymes and GO terms for 100 of the 217 genes (Supplementary Table S8). Some interesting proteins were identified, as in the region associated with more traits, in Chromosome 7 (position 42,826,434). It had seven mapped genes, some were annotated as Cinnamoyl-CoA reductase 1 which participates in the lignin biosynthesis and DEAD-box ATP-dependent RNA helicase 25 which acts in response to abiotic stress. Furthermore, in Chromosome 1 (position 11,178,365), associated to four traits, there are genes annotated to the protein multidomain RHM2/MUM4, involved in UDP-D-glucose to UDP-L-rhamnose conversion which is required for the biosynthesis of cell wall (Supplementary Table S8).

Beyond specific protein annotation, to obtain a general functional profile of the proteins identified, we performed the enrichment of the biological process GO terms and obtained a profile with 18 significant terms (p-value < 0.01). The enrichment identified terms

associated with various phenotype clippings, such as “lignin biosynthetic process” , “auxin efflux” and “flavonol biosynthetic process” (Supplementary Table S10).

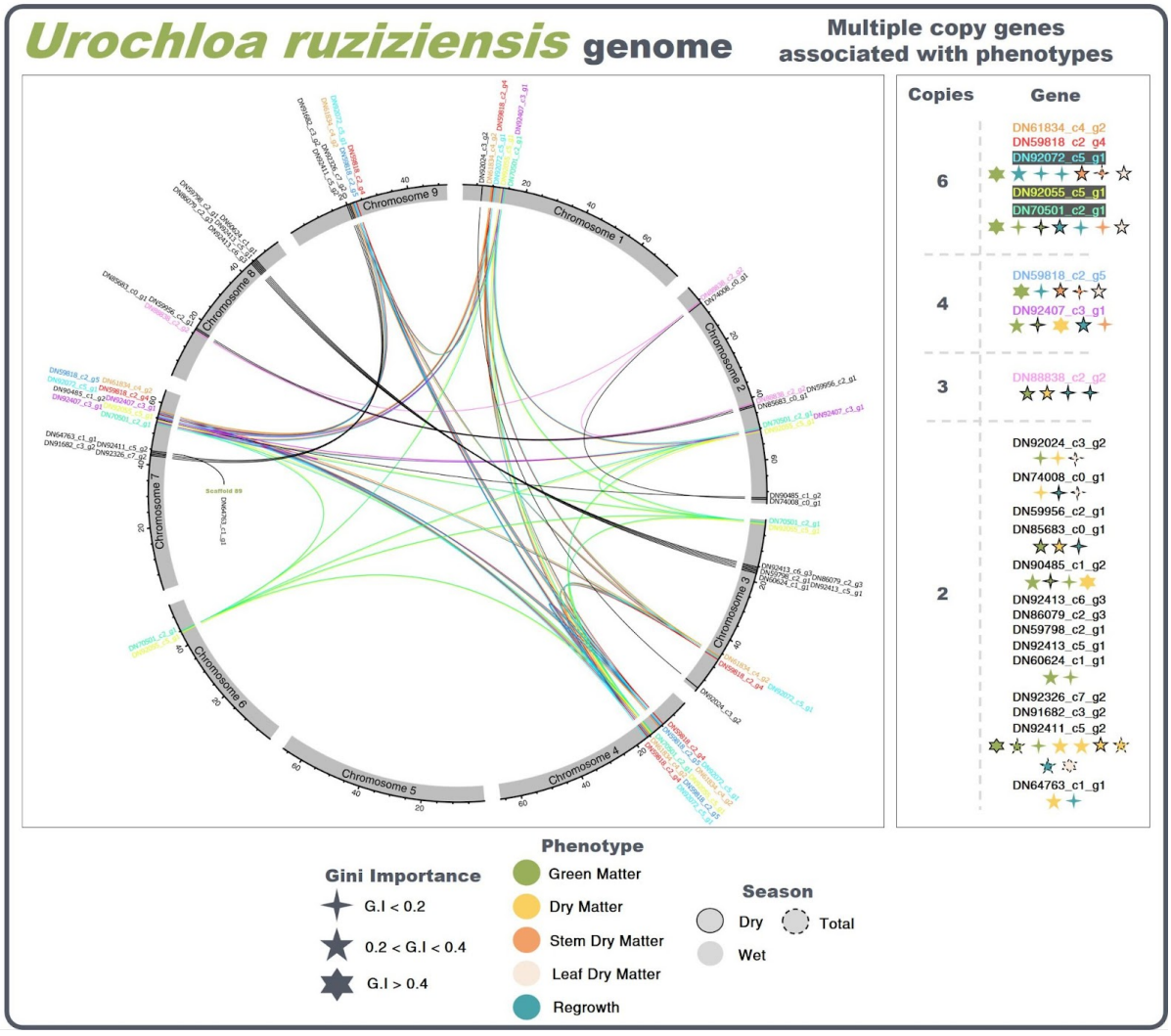


Figure 3. Circular map of the *U. ruziziensis* genome, indicating multiple copy genes identified as associated with the phenotypes evaluated, indicating Gini importance and Season. Same genes indicated with the same color, except genes with two copies.

3.5 - Co-expression network

In order to provide deeper insights into functional patterns of genes associated with the agronomic traits evaluated, we modeled a gene GCN using gene quantifications from *U. ruziziensis* accessions. From a total of 49,445 genes, we could estimate significant interactions between 14,141, represented as nodes in the network structure and associated through 17,812 edges (Supplementary Fig. S8). In the GCN, we found 54 genes from the 217 genes associated with the major importance markers. As we restricted the GCN created to the top 30 gene associations, we expanded the collection of 54 selected genes to more 109 by

considering correlations with a minimum Pearson coefficient of 0.9 and a Bonferroni corrected p-value of 0.01. This group of 153 genes was considered directly associated with the traits evaluated.

The GCN potential to elucidate the metabolic pathways lies in the possibility to identify genes that, even though were not selected by the prediction methodology, are co-expressed with them. In this sense, we expanded the set of 153 genes previously selected, to the GCN first (308 genes) and second gene neighbors (2233 genes), to compose an agronomic traits network with a total of 2704 genes (nodes) and 3453 edges (Fig 4-A).

The functional profile of the agronomic traits network, obtained through the enrichment of the biological process GO terms, identified 11 significant terms ($p < 0.01$) for the genes set without the second neighbors and 16 terms considering all genes in the network Supplementary Table S11). When we consider the restricted set without the second neighbors, we found enriched terms related to hormones as auxin efflux and abscisic acid transport; and biosynthetic processes of molecules such as the flavonoids. In the broader set, including second neighbors, we found terms related to DNA metabolism as: mismatch repair, DNA replication and DNA duplex unwinding. Other terms enriched were for biotic stress: response to chitin; and regulation of circadian rhythm.

Another interesting aspect of the GCN to investigate regulation of metabolic pathways, is that by identifying the genes with more connections in the network through the degree metric, we can find important genes that regulate the functioning of many others. Consequently, those genes with high degree values may have an important role in the expression of the phenotypes that we are studying. In our modeled agronomic traits network, we found some interesting genes with many connections, like the ones that translate into the ribosomal proteins 40S S6 and 60S L9, protein ELF4-LIKE 4 or 14-3-3 protein zeta (Supplementary Table S12).

Furthermore, to investigate the differences in respect to wet and dry seasons, we separated the genes of the general network related to each season and evidenced them in the network. The wet and dry parts of the network comprises 33 and 22 genes associated with the major importance markers, 58 and 54 highly correlated, 102 and 231 first neighbors, 1322 and 1359 second neighbors, and totalized 1515 and 1666 genes (nodes) with 1801 and 2205 edges, respectively (Fig 4-B,C). Comparing the seasons functional profile, we found shared terms as: flavonol biosynthetic process, auxin efflux and mitotic recombination-dependent replication fork processing. And season specific terms, like abscisic acid transport, isoleucine biosynthetic process, response to nematode and chaperone-mediated protein

folding for wet season. And for dry season, we found enriched terms as: pyridoxal phosphate biosynthetic process, response to water deprivation and response to chitin, which are terms related to stress response (Supplementary Table S13 and S14).

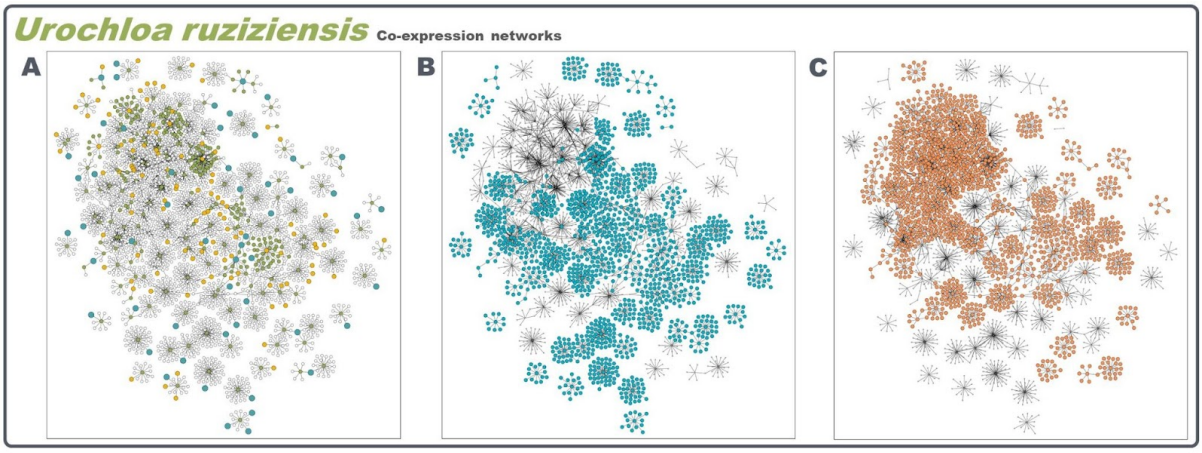


Figure 4. Selected and correlated genes co-expression network with first and second neighbors. A) Majo importance genes in blue, highly correlated genes in yellow, first neighbors in green and second neighbors in gray. B) Genes associated with the wet season trait clippings. C) Genes associated with the dry season trait clippings.

4 - Discussion

Tropical forage grasses play an important role in the global economy and in the supply of milk and meat. Thus, the increase in productivity and quality of these species, mainly from the *Urochloa* genus, can result in more sustainable livestock production and positively impact food and nutrition security. Although conventional breeding of tropical forage grasses has resulted in the release of superior cultivars in recent decades, the genetic gains are low and molecular breeding still not being used due to the genus scarce genomic resources and genomic complexity (Simeão et al., 2021, Ferreira et al., 2021). Besides the difficulties, the recent advances in omics approaches and computational methods for polyploid species has enabled the emergence of studies in *Urochloa* breeding important areas, such as genome assembly (Pessoa-Filho et al., 2018; Worthington et al., 2021), contaminant identification (Martins et al., 2021), transcriptomics (Vigna et al., 2016a, Worthington et al., 2021, Jones et al., 2021; Hanley et al., 2021), linkage and QTL mapping (Vigna et al., 2016b; Thaikua et al., 2016; Ferreira et al., 2016; Worthington et al., 2021), GWAS (Matias et al., 2019b) and genomic selection/prediction (Matias et al., 2019a, Aono et al., 2022). By integrating the few genetic resources available through bioinformatics and not explored techniques, such as machine learning feature selection and complex networks, we developed

a multi-omic approach capable of expanding the molecular breeding knowledge that can be used in the *Urochloa* genus.

Following this sense, we conducted our study considering 49 *Urochloa ruziziensis* subfamilies genotyped with 28,106 high quality markers and phenotyped in 33 characteristics, which are clippings for the agronomic phenotypes GM, DM, SDM, LDM and RG. In our first approach for genome wide family prediction (GWFP), we employed the complete markers dataset (CD) with conventional parametric and semiparametric GP models (BRR and RKHS) and obtained higher or at least equivalent PA in comparison to other GWFP studies. Even though not with *Urochloa* species, there are studies with forages that evaluated the prediction of dry matter. While we had a high mean PA of ~0.8 for the DM clippings, in alfalfa it had values smaller than 0.7 (Murad Leite Andrade et al., 2022) and in ryegrass it had a low PA of 0.34 (Guo et al., 2018). If we consider the GWFP of other phenotypes, like rust resistance and heading date in alfalfa (Murad Leite Andrade et al., 2022) or ligning , stiffness and diameter in loblolly pine (Rios et al., 2021), they were still smaller than our results, which had a mean PA across the 33 traits of ~0.762. The performance that we obtained seems even more promising when we compare to genomic prediction performed at individual level in tropical forages. The studies with *U.ruziziensis* interspecific hybrids (Matias et al., 2019a), *U. decumbens* (Aono et al., 2022) *M. maximum* (de C. Lara et al., 2019; Aono et al., 2022) and with *P. virgatum* (Lipka et al., 2014), obtained PA ranging from values close to zero to maximum values of 0.7 when evaluating several agronomic and morphological traits. The highest values were obtained for *U. decumbens* and *M. maximus* in agronomic traits after feature selection strategies with integration of ML methods that increased the PAs (Aono et al., 2022).

The application of ML algorithms in genomic prediction has been investigated in a variety of species and phenotypes (Grinberg et al., 2016; Lello et al., 2018; Zhao et al., 2020; Liang et al., 2020; Chung et al., 2021; Islam et al., 2021; Sandhu et al., 2021), and besides that there is no empirical evidence in support of the supremacy of ML methods over linear methods (Zingaretti et al., 2020; Varshney, 2021), it has performed better or at least equal to conventional models in various situations (Bellot et al., 2018; Abdollahi-Arpanahi et al., 2020; Liang et al., 2021; Wang et al. 2022). It is discussed that the ML methods finds its potential to outperform conventional models especially when dealing with complex phenotypes controlled by high dominance and epistatic effects (Wang et al., 2018; Tong & Nikoloski, 2021), and beyond that, there are no studies investigating its applicability in GWFP. This way, in addition to the parametric (BRR) and semiparametric (RKHS) models,

we evaluated four ML models (SVM, RF, AB and MLP) and none was able to outperform the RKHS model. Although SVM had competitive performances for PA, it was not as good for MSE; RF and AB had intermediate performances and MLP was the worst by far (Supplementary Tables S2-4). The bad performance of MLP may be related to the small sample size of our data set, since neural networks methods commonly require large data sets for good predictions (Bellot et al., 2018; Montesinos-López et al., 2021).

Although ML algorithms have been showing similar performances to the conventional models for genomic prediction, some of them offers methods to feature selection (FS), which is a common strategy in data science to remove redundant, incorrect or irrelevant data from the dataset and in some cases enhance the modeling performance (Miao & Niu, 2016; Cai et al., 2018). FS has been applied in genomic prediction and produced interesting results, enabling enhanced prediction or at least equivalent performances with smaller markers datasets (Li et al. 2018; Azodi et al., 2019 (A); Luo et al., 2021; Piles et al., 2021). On this wise, we applied three FS methods based on decision tree algorithms (gradient tree boosting, extremely randomized trees and random forest) in our markers dataset and following the intersection approach, we created two intersection (FI) datasets (FI-1: intersection of at least two FS; FI-2: intersection among all FS) for each of the 33 traits. This approach of using FI to compose smaller data sets with putatively more reliable markers for prediction, has been applied in grasses and produced improved results (Aono et al., 2020; Pimenta et al., 2021; Aono et al., 2022; Pimenta et al., 2022). For the traits evaluated in this study, the FI-1 and FI-2 dataset types had a mean size of ~102 and ~11 markers, respectively, and both produced improved prediction results in all models. Even though FI-1 had better performance across models, the best model for each FI (FI-1:AB; FI-2:RF) were equivalent, and represented a mean PA increase of 5.9% in comparison to the RKHS with CD (Table 1).

In regard to the analyzed agronomic traits, we observed a high correlation among evaluations and it was expected since the biomass characteristics evaluated are very similar and may be controlled mainly by the same metabolic processes. DM phenotyping is done by drying the GM material, and SDM/LDM by separating the DM into stem and leaf, furthermore, the biomass production depends on the plant growth capacity (RG). For that reason, the families narrow sense heritabilities for the traits were very similar too, and as stated in other studies, modeling performances are strongly influenced by heritabilities (Wang et al., 2018; Xu et al., 2018; Murad Leite Andrade et al., 2022), consequently, our prediction performances did not vary much and were correlated to the heritabilities (Supplementary Table S1, 2 and 3). Interestingly, the use of FS/FI had more impact increasing the PAs of the

clippings with less heritabilities (GM-2/3, DM-2/3, SDM-2, RG-8/9), making the prediction performances across the 33 traits less variable in comparison to CD.

We believe that such high PA in our predictions may be related mainly to the small population size and the low genetic diversity among samples. The diversity is low because the population was obtained from a open pollination cross of only seven sexual genotypes and as seen in the CD PCA scatterplot (Supplementary Fig. S5), the diversity is defined by the formation of 3 groups where most samples are concentrated in one group. This combination is reported to produce high predictive abilities as observed in wheat (Edwards et al., 2019). Furthermore, there are works demonstrating that increasing the population with genetically distant samples makes the prediction task more difficult and consequently reduces its accuracy (Lorenz & Smith, 2015; Berro et al., 2019).

Beyond the well known advantages of genomic selection related to the power to shorten the time of breeding process (Simeão-Resende et al., 2014) and to reduce the phenotyping costs (Crossa et al., 2017). The use of GWFP has other advantages, specially in forage breeding programs, which commonly uses families or plot level phenotyping for conventional breeding (Rios et al., 2021). Besides its apparent higher performance over individual level GP, when performed using ML models coupled with FS/FI strategies, it has the potential to strongly reduce the costs of genotyping. First because large populations can be reduced to family samples and second, because despite the necessity of genotyping the training population with a high number of markers in order to perform FS/FI, the samples that are going to be predicted can be genotyped only for the small FI set.

In addition, for its application in breeding programs or further studies, we would recommend experimental designs prioritizing more families instead of more individuals per family. Even though not with half-sibling families, there are works with tetraploid full siblings families that concluded that six individuals would be enough to represent the family variation both in genotyping and phenotyping (de Bem Oliveira et al. 2020; Rios et al., 2021), evidencing that bulking 20 individuals per family as we did may be unnecessary. Furthermore, to a certain extent, increasing the training population has the potential to improve the GWFP performance (Fé et al. 2015).

Beyond the applicability analysis of GWFP in *U. ruziziensis*, the present work aimed to investigate the agronomic traits metabolic regulation, and as a first step to achieve this objective, we identified marker-phenotype associations. To this extent, the high PA and the performance enhancing using FS/FI evidences that the selected sets of markers are putatively associated with QTLs, and can be used to study molecular processes involved in the trait

predicted (Steinfath et al., 2010; Heer et al., 2018; Zhou et al., 2019; Aono et al., 2020; Pimenta et al., 2021; Aono et al., 2022; Pimenta et al., 2022). In parallel to other approaches to identify genotype-phenotype associations, the FS techniques does not depend on specific bi-parental populations (RILs, NILs, F2, etc) as is necessary for QTL mapping (Mohan et al., 1997; Dhingani et al., 2015) and can identify non-linear and complex associations which is a limitation of linear models used in GWAS (Korte & Farlo, 2013).

Furthermore, the ML models based on decision trees offer a good prediction interpretability since it is possible to assess feature importance through Gini importance, and in the context of genomic prediction, it can rank markers based on the strength of the association with the modeled phenotype (Azodi et al., 2019b; Bayer et al., 2021; Medina et al., 2021). This way, considering that the best model for each FI dataset type was equivalent, we computed the RF Gini importance for the more restricted FI-2 datasets and selected the most significant features to compose an even smaller and more reliable set of agronomic traits associated markers (Supplementary Table S8). In this major importance set, we identified markers associated with more than one phenotype, the number of shared markers was representative only for GM, DM and LDM, which may indicate the similarity of the biological processes behind the phenotypes, while SDM and RG seems to be controlled by specific/exclusive metabolic pathways, which is corroborated by the smaller correlation of SDM and RG with the other phenotypes (Supplementary Fig S3 and S6).

By using half-sibling families bulks as a representation of the genetic variability available for breeding, genotyping similar agronomic traits in various clippings and selecting only the most influential markers in the predictions, we were able to minimize the limitations of the method due to small sample size and obtain a reliable set of markers to use as a guide to identify genes/regions involved in the regulation of the species' agronomic traits. In the absence of genome annotation, we employed a multi-omic approach integrating RNA-seq data in the analysis to identify and annotate expressed genes physically linked to the selected markers (Supplementary Table S8).

The enrichment of the terms related to the annotated genes evidences the methodology capacity of in fact identifying QTL regions influencing the evaluated agronomic traits (Supplementary Table S10). Associated with various phenotype clippings, we had terms related to lignin biosynthetic process, which has been previously reported to have strong influence in plant development (Yoon, Choi & An, 2015; Bahri et al., 2020), where mutants of lignin biosynthesis genes showed phenotypes of dwarfism/reduced plant growth (Schilmiller et al., 2009; Wagner et al., 2009; Li et al., 2009; Song & Wang, 2011),

altered morphology (Elkind et al., 1990; Piquemal et al., 1998; Jones et al., 2001; Franke et al., 2002) and browning tissues (Bout & Vermerris, 2003; Xu et al., 2011; Saballos et al. 2012). Terms related to auxin efflux were identified too and it is known its importance in regulation of growth. Since the auxin hormone effects are concentration dependent and it is mostly produced in the meristematic and other specific regions (Blakeslee, Peer & Murphy, 2005; Zhao, 2018), the transport and distribution of auxin within plant tissues constitutes an essential aspect of its function in plant organogenesis and morphogenesis (Woodward & Bartel, 2005). This transport is achieved through influx and efflux carrier proteins, providing essential directional and positional information for developmental processes, such as vascular differentiation, apical dominance, organ development and tropic growth (Benkova et al., 2003; Blancaflor & Masson, 2003; Friml et al., 2003; Blilou et al., 2005; Grieneisen et al., 2007).

Furthermore, flavonol biosynthetic process, another enriched term found in our results, is reported to regulate plant growth and development by regulating auxin transport, with effects mainly in root elongation, quantity and gravitropic response (Jacobs & Rubery, 1988; Brown et al., 2001; Santelia et al., 2008; Grunewald et al., 2012). The flavonols may have influence on auxin transportation by different mechanisms, like by affecting the transcriptions of genes encoding auxin transport proteins (Peer et al., 2004), by being kinase inhibitors modulating the phosphorylation status of auxin transport proteins (Agullo et al., 1997; Peer & Murphy, 2007), or by altering the redox state of the cell (Fernández-Marcos et al., 2013). In addition to that, flavonols have antioxidant functions, acting in stress response to UV radiation, wounding, drought, metal toxicity and nutrient deprivation, conditions that result in the accumulation of reactive oxygen species (ROS), which can damage cellular components and consequently affect plant development (Winkel-Shirley, 2001; Baskar, Venkatesh & Ramalingam, 2018; Agati et al., 2020). The list of terms associated with plant growth, development and stress response continues with folic acid biosynthetic process (Stakhove et al., 2000; Gorelova et al., 2017), galactolipid metabolic process (Kobayashi et al., 2007; Jouhet et al., 2007; Botté et al., 2011) and cellular response to cold.

In this sense, considering its ability to identify regions with known genes associated with the traits, another important aspect of the methodology is to put light on not annotated genes that should be investigated. In our results, more than half of the identified genes linked to the major importance markers did not have functional annotation, and more than that, some of them seems to be very important, having multiple copies and being associated with various traits (Fig. 3). These genes/regions are important targets to expand the knowledge on the

agronomic traits metabolic regulation and represents valuable information that can be applied in the species breeding.

Additionally, for going further in our multi-omic investigation and understanding of the metabolic pathways and regulation that controls the evaluated agronomic traits. We constructed a co-expression network and isolated the previously identified genes and their co-expressed (Figure 4-A). This integration has been employed in some species and produced interesting results (Calabrese et al., 2017; Schaefer et al., 2018; Yan et al., 2020; Francisco et al., 2021), the main difference is that instead of using GWAS to identify marker-phenotype associations, we employed genomic prediction with feature selection, similar to the method applied in the study of mosaic virus resistance in sugarcane (Pimenta et al., 2022). In this aspect, the capability of networks to simulate complex biological systems and infer novel biological associations has been revolutionizing molecular biology research (D'haeseleer et al., 2000; Liu et al., 2020), enabling the study of regulatory relationships, metabolic pathway inferences and annotation transference (Rao & Dixon, 2019). Based on the "guilt-by association" principle, which in the case of co-expression networks, states that genes with correlated biological functions tend to interact in the networks (Oliver, 2000, Wolfe et al., 2005, Childs et al., 2011). We could expand our set of identified genes to their co-expression modules and have a broader profile of the metabolic pathways acting in the phenotypes, and more than that, the annotated genes in the modules can be used to infer the biological functions of the not annotated genes.

Even though the investigation of the enriched terms in the agronomic traits network showed various terms with no clear relation to the phenotypes evaluated, such as seed maturation, meiosis processes or pollen sperm cell differentiation (Supplementary Table S11). The network expanded the genes related to the enriched terms already discussed, and enabled the identification of new genes involved in biological processes mainly related to DNA integrity, stability and metabolism, acting in mismatch repair, telomere capping and duplex unwinding, all reported to at some extend, affect the normal growth and development of plants (Karthika et al., 2020, Kim & Kim, 2018, Tuteja, 2003); related to the abscisic acid (ABA) transport, where the modulation of the hormone levels in tissues and cells is critical for balancing defense and growth processes when not in an optimal environments, with an important function controlling stomatal closure (Seo & Koshiba, 2011; Chen et al., 2020); related to the regulation of the circadian rhythm, which gives the plants the capacity to not only deal with the daily environmental changes but anticipate and prepare for them beforehand (Kim et al., 2017; Millar, 2016; Creux & Harmer, 2019), where the gene

ELF4-LIKE 4, a circadian rhythm key gene (Doyle et al., 2002), had its importance evidenced by being one of the genes with highest degree value in the network (Supplementary Table S12); and related to response to chitin, an important part of the plants immune system that is activated in the presence of pathogens such as fungi, arthropods and nematodes egg-shells (Kombrink et al., 2011; Sánchex-Vallet et al., 2015).

Furthermore, by isolating in the network the genes associated with each season, we could investigate metabolic processes affecting the plant development and production in the wet and dry periods. In our results, we identified in both seasons enriched terms for auxin efflux and flavonol biosynthetic process which as already discussed regulates auxin transport, indicating the importance of the hormone regardless of the season. For the wet season, beyond the abscisic acid transport already cited, we had terms reported to influence plant development such as isoleucine biosynthetic process (Yu et al., 2013) and response to nematodes, a pathogen capable of modifying plant physiology, development, metabolism and immunity (Eves-van den Akker, 2021). Even though being related to the phenotypes, we could not find in the literature studies specifically associating these processes with wet/rainy season. Differently, in the dry network we had response to water deprivation and protein transport terms enriched, which is evidence of the metabolic machinery necessary to deal with the abiotic stress.

Our work is innovative in different aspects and represents a great advance in the knowledge of molecular breeding techniques that can be applied in tropical forages. It is the first study investigating the applicability of GWFP in an *Urochloa* species and more than that, the first time that feature selection and machine learning algorithms are employed in GWFP to not only enhance the prediction metrics but drastically reduce the number of makers necessary for the prediction. Additionally, in a multi-omic approach, the markers selected were integrated with transcriptome data to model a co-expression network capable of offering insights into the regulation of plant growth and biomass production in the species. The results show that molecular breeding has a great potential to reduce breeding costs, accelerate the release of new varieties and provide means to metabolic investigations even in orphan species with high genomic complexity such as tropical forages.

5 - References

Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics, selection, evolution : GSE*, *52*(1), 12. https://doi.org/10.1186/s12711-020-00531-z

Agati, G.; Brunetti, C.; Fini, A.; Gori, A.; Guidi, L.; Landi, M.; Sebastiani, F.; Tattini, M. Are Flavonoids Effective Antioxidants in Plants? Twenty Years of Our Investigation. Antioxidants 2020, 9, 1098. https://doi.org/10.3390/antiox9111098

Agullo, G., Gamet-Payrastre, L., Manenti, S. et al. (1997) Relationship between flavonoid structure and inhibition of phosphatidylinositol 3-kinase: a comparison with tyrosine kinase and protein kinase C inhibition. Biochemical Pharmacology, 53, 1649–1657.

Alexa A., Rahnenfuhrer J. (2022). topGO: Enrichment Analysis for Gene Ontology. R package version 2.48.0.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. In Journal of Molecular Biology (Vol. 215, Issue 3, pp. 403–410). Elsevier BV. https://doi.org/10.1016/s0022-2836(05)80360-2

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aono, A.H., Costa, E.A., Rody, H.V.S. *et al.* Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. *Sci Rep* 10, 20057 (2020). https://doi.org/10.1038/s41598-020-77063-5

Aono, A. H., Ferreira, R., Moraes, A., Lara, L., Pimenta, R., Costa, E. A., Pinto, L. R., Landell, M., Santos, M. F., Jank, L., Barrios, S., do Valle, C. B., Chiari, L., Garcia, A., Kuroshu, R. M., Lorena, A. C., Gorjanc, G., & de Souza, A. P. (2022). A joint learning approach for genomic prediction in polyploid grasses. *Scientific reports*, *12*(1), 12499. https://doi.org/10.1038/s41598-022-16417-7

Ashraf, B. H., Jensen, J., Asp, T., & Janss, L. L. (2014). Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *127*(6), 1331–1341. https://doi.org/10.1007/s00122-014-2300-4

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., & Shiu, S.-H. (2019-A). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. In G3 Genes|Genomes|Genetics (Vol. 9, Issue 11, pp.

3691–3702). Oxford University Press (OUP). https://doi.org/10.1534/g3.119.400498

Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., & Shiu, S.-H. (2019-B). Transcriptome-Based Prediction of Complex Traits in Maize. In The Plant Cell (Vol. 32, Issue 1, pp. 139–151). Oxford University Press (OUP). https://doi.org/10.1105/tpc.19.00332 (B)

Bahri, B.A., Daverdin, G., Xu, X. et al. Natural Variation in Lignin and Pectin Biosynthesis-Related Genes in Switchgrass (Panicum virgatum L.) and Association of SNP Variants with Dry Matter Traits. Bioenerg. Res. 13, 79–99 (2020). https://doi.org/10.1007/s12155-020-10090-2

Baskar, V., Venkatesh, R., Ramalingam, S. (2018). Flavonoids (Antioxidants Systems) in Higher Plants and Their Response to Stresses. In: Gupta, D., Palma, J., Corpas, F. (eds) Antioxidants and Antioxidant Enzymes in Higher Plants. Springer, Cham. https://doi.org/10.1007/978-3-319-75088-0_12

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., & Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. In The Plant Genome (Vol. 14, Issue 3). Wiley. https://doi.org/10.1002/tpg2.20112

Barabási A. L., Oltvai Z. N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113. doi: 10.1038/nrg1272

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., … Teodoro, D. (2020). UniProt: the universal protein knowledgebase in 2021. In Nucleic Acids Research (Vol. 49, Issue D1, pp. D480–D489). Oxford University Press (OUP). https://doi.org/10.1093/nar/gkaa1100

Bélanger, S., Esteves, P., Clermont, I., Jean, M., & Belzile, F. (2016). Genotyping-by-Sequencing on Pooled Samples and its Use in Measuring Segregation Bias during the Course of Androgenesis in Barley. *The plant genome*, *9*(1), 10.3835/plantgenome2014.10.0073. https://doi.org/10.3835/plantgenome2014.10.0073

Bellot, P., de los Campos, G., & Pérez-Enciso, M. (2018). Can Deep Learning Improve Genomic Prediction of Complex Human Traits? In Genetics (Vol. 210, Issue 3, pp. 809–819). Oxford University Press (OUP). https://doi.org/10.1534/genetics.118.301298

Benková, E., Michniewicz, M., Sauer, M., Teichmann, T., Seifertová, D., Jürgens, G., & Friml, J. (2003). Local, Efflux-Dependent Auxin Gradients as a Common Module for Plant

Organ Formation. In Cell (Vol. 115, Issue 5, pp. 591–602). Elsevier BV. https://doi.org/10.1016/s0092-8674(03)00924-3

Berro, I., Lado, B., Nalin, R. S., Quincke, M. & Gutiérrez, L. Training population optimization for genomic selection. Plant Genome 12, 190028 (2019).

Bian, Y., & Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity*, *118*(6), 585–593. https://doi.org/10.1038/hdy.2017.4

Biazzi E, Nazzicari N, Pecetti L, Brummer EC, Palmonari A, et al. 2017. Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. *PLoS One*. 12:e0169234.

Blancaflor, E. B., & Masson, P. H. (2003). Plant Gravitropism. Unraveling the Ups and Downs of a Complex Process. In Plant Physiology (Vol. 133, Issue 4, pp. 1677–1690). Oxford University Press (OUP). https://doi.org/10.1104/pp.103.032169

Blakeslee, J. J., Peer, W. A., & Murphy, A. S. (2005). Auxin transport. In Current Opinion in Plant Biology (Vol. 8, Issue 5, pp. 494–500). Elsevier BV. https://doi.org/10.1016/j.pbi.2005.07.014

Blilou, I., Xu, J., Wildwater, M., Willemsen, V., Paponov, I., Friml, J., Heidstra, R., Aida, M., Palme, K., & Scheres, B. (2005). The PIN auxin efflux facilitator network controls growth and patterning in Arabidopsis roots. In Nature (Vol. 433, Issue 7021, pp. 39–44). Springer Science and Business Media LLC. https://doi.org/10.1038/nature03184

Breiman, L. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655 (2001).

Brown, D.E., Rashotte, A.M., Murphy, A.S. et al. (2001) Flavonoids act as negative regulators of auxin transport in vivo in Arabidopsis. Plant Physiology, 126, 524–535.

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T.-H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Jr., Peshkin, L., Tabin, C. J., … Whited, J. L. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. In Cell Reports (Vol. 18, Issue 3, pp. 762–776). Elsevier BV. https://doi.org/10.1016/j.celrep.2016.12.063

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

Borin, G. P., Carazzolle, M. F., Dos Santos, R., Riaño-Pachón, D. M., & Oliveira, J.

(2018). Gene Co-expression Network Reveals Potential New Genes Related to Sugarcane Bagasse Degradation in *Trichoderma reesei* RUT-30. *Frontiers in bioengineering and biotechnology*, *6*, 151. https://doi.org/10.3389/fbioe.2018.00151

Botté, C. Y., Yamaryo-Botté, Y., Janouškovec, J., Rupasinghe, T., Keeling, P. J., Crellin, P., ... & McFadden, G. I. (2011). Identification of plant-like galactolipids in Chromera velia, a photosynthetic relative of malaria parasites. Journal of Biological Chemistry, 286(34), 29893-29903.

Bout, S., & Vermerris, W. (2003). A candidate-gene approach to clone the sorghum Brown midrib gene encoding caffeic acid O-methyltransferase. In Molecular Genetics and Genomics (Vol. 269, Issue 2, pp. 205–214). Springer Science and Business Media LLC. https://doi.org/10.1007/s00438-003-0824-4

Buell, C. R. (2008). Poaceae Genomes: Going from Unattainable to Becoming a Model Clade for Comparative Plant Genomics. In Plant Physiology (Vol. 149, Issue 1, pp. 111–116). Oxford University Press (OUP). https://doi.org/10.1104/pp.108.128926

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., & Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PloS one*, *8*(3), e57438. https://doi.org/10.1371/journal.pone.0057438

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., & Araus, J. L. (2012). High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. Journal of integrative plant biology, 54(5), 312–320.

Cai, J., Luo, J., Wang, S. & Yang, S. Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79 (2018).

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. Cell Syst. 4, 46–59. doi: 10.1016/j.cels.2016.10.014

Cardoso-Silva, C. B., Aono, A. H., Mancini, M. C., Sforça, D. A., da Silva, C. C., Pinto, L. R., Adams, K. L., & de Souza, A. P. (2022). Taxonomically Restricted Genes Are Associated With Responses to Biotic and Abiotic Stresses in Sugarcane (*Saccharum* spp.). *Frontiers in plant science*, *13*, 923069. https://doi.org/10.3389/fpls.2022.923069

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., Asp, T., Jensen, J., & Janss, L. (2018). Optimized Use of Low-Depth Genotyping-by-Sequencing for Genomic Prediction Among Multi-Parental Family Pools and Single Plants in Perennial

Ryegrass (Lolium perenne L.). Frontiers in plant science, 9, 369.

Chen, T., & Guestrin, C. Xgboost: A scalable tree boosting system. In KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 785–794 (ACM, New York, 2016).

Chen, K., Li, G., Bressan, R. A., Song, C., Zhu, J., & Zhao, Y. (2020). Abscisic acid dynamics, signaling, and functions in plants. In Journal of Integrative Plant Biology (Vol. 62, Issue 1, pp. 25–54). Wiley. https://doi.org/10.1111/jipb.12899

Childs, K. L., Davidson, R. M., and Buell, C. R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One 6:e22196. doi: 10.1371/journal.pone.0022196

Chung, CW., Hsiao, TH., Huang, CJ. et al. Machine learning approaches for the genomic prediction of rheumatoid arthritis and systemic lupus erythematosus. BioData Mining 14, 52 (2021). https://doi.org/10.1186/s13040-021-00284-5

Costa, C., Schurr, U., Loreto, F., Menesatti, P., & Carpentier, S. (2019). Plant Phenotyping Research Trends, a Science Mapping Approach. Frontiers in plant science, 9, 1933.

Creux, N., & Harmer, S. (2019). Circadian Rhythms in Plants. In Cold Spring Harbor Perspectives in Biology (Vol. 11, Issue 9, p. a034611). Cold Spring Harbor Laboratory. https://doi.org/10.1101/cshperspect.a034611

Cristianini, N. & Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge University Press, 2000).

Crossa, J., Martini, J. W. R., Gianola, D., Pérez-Rodríguez, P., Jarquin, D., Juliana, P., Montesinos-López, O., & Cuevas, J. (2019). Deep Kernel and Deep Learning for Genome-Based Prediction of Single Traits in Multienvironment Breeding Trials. In Frontiers in Genetics (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fgene.2019.01168

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22, 961–975. doi: 10.1016/j.tplants. 2017.08.011

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*(2), 347–365. https://doi.org/10.1534/genetics.112.147983

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011).

The variant call format and VCFtools. In Bioinformatics (Vol. 27, Issue 15, pp. 2156–2158). Oxford University Press (OUP). https://doi.org/10.1093/bioinformatics/btr330

de Bem Oliveira I, Amadeu RR, Ferrão LFV, Muñoz PR. Optimizing whole-genomic prediction for autotetraploid blueberry breeding. Heredity (Edinb). 2020 Dec;125(6):437-448. doi: 10.1038/s41437-020-00357-x. Epub 2020 Oct 19. PMID: 33077896; PMCID: PMC7784927.

de C. Lara, L. A., Santos, M. F., Jank, L., Chiari, L., Vilela, M. de M., Amadeu, R. R., dos Santos, J. P. R., Pereira, G. da S., Zeng, Z.-B., & Garcia, A. A. F. (2019). Genomic Selection with Allele Dosage in Panicum maximum Jacq. In G3 Genes|Genomes|Genetics (Vol. 9, Issue 8, pp. 2463–2475). Oxford University Press (OUP). https://doi.org/10.1534/g3.118.200986

De Mendiburu, F., & De Mendiburu, M. F. Package 'agricolae'. R package version, 1–2 (2020).

Devos, K. M. (2010). Grass genome organization and evolution. In Current Opinion in Plant Biology (Vol. 13, Issue 2, pp. 139–145). Elsevier BV. https://doi.org/10.1016/j.pbi.2009.12.005

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707–726. doi: 10.1093/bioinformatics/16.8.707

Dhingani RM, Umrania VV, Tomar RS, et al. Introduction to QTL mapping in plants. Ann Plant Sci. 2015;4(04):1072–1079.

Doyle, M.R.; Davis, S.J.; Bastow, R.M.; McWatters, H.G.; Kozma-Bognar, L.; Nagy, F.; Millar, A.J.; Amasino, R.M. The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. Nature 2002, 419, 74–77.

Du, H., Zhu, J., Su, H., Huang, M., Wang, H., Ding, S., Zhang, B., Luo, A., Wei, S., Tian, X., & Xu, Y. (2017). Bulked Segregant RNA-seq Reveals Differential Expression and SNPs of Candidate Genes Associated with Waterlogging Tolerance in Maize. Frontiers in plant science, 8, 1022.

Edae, E. A., & Rouse, M. N. (2019). Bulked segregant analysis RNA-seq (BSR-Seq) validated a stem resistance locus in Aegilops umbellulata, a wild relative of wheat. PloS one, 14(9), e0215492.

Edwards, S.M., Buntjer, J.B., Jackson, R. et al. The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet 132, 1943–1952 (2019).

https://doi.org/10.1007/s00122-019-03327-y

Elkind, Y., Edwards, R., Mavandad, M., Hedrick, S. A., Ribak, O., Dixon, R. A., & Lamb, C. J. (1990). Abnormal plant development and down-regulation of phenylpropanoid biosynthesis in transgenic tobacco containing a heterologous phenylalanine ammonia-lyase gene. In Proceedings of the National Academy of Sciences (Vol. 87, Issue 22, pp. 9057–9061). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.87.22.9057

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Eves-van den Akker, S. (2021). Plant–nematode interactions. In Current Opinion in Plant Biology (Vol. 62, p. 102035). Elsevier BV. https://doi.org/10.1016/j.pbi.2021.102035

Fè, D., Cericola, F., Byrne, S. et al. Genomic dissection and prediction of heading date in perennial ryegrass. BMC Genomics 16, 921 (2015). https://doi.org/10.1186/s12864-015-2163-3

Fernández-Marcos, M., Sanz, L., Lewis, D.R. et al. (2013) Control of auxin transport by reactive oxygen and nitrogen species, in Polar Auxin Transport, Signaling and Communication in Plants, vol. 17 (eds R. Chen and F. Baluska), Springer-Verlag, Berlin, pp. 103–117

Ferrão, L., Amadeu, R. R., Benevenuto, J., de Bem Oliveira, I., & Munoz, P. R. (2021). Genomic Selection in an Outcrossing Autotetraploid Fruit Crop: Lessons From Blueberry Breeding. *Frontiers in plant science*, *12*, 676326. https://doi.org/10.3389/fpls.2021.676326

Ferreira, R. C. U., Cançado, L. J., Do Valle, C. B., Chiari, L., and de Souza, A. P. (2016). Microsatellite loci for Urochloa decumbens (Stapf) R.D. Webster and cross-amplification in other Urochloa species. BMC. Res. Notes 9:152. doi: 10.1186/s13104-016-1967-9

Ferreira R. C. U, da Costa Lima Moraes A, Chiari L, Simeão RM, Vigna BBZ, de Souza AP. An Overview of the Genetics and Genomics of the *Urochloa* Species Most Commonly Used in Pastures. Front Plant Sci. 2021 Dec 13;12:770461. doi: 10.3389/fpls.2021.770461. PMID: 34966402; PMCID: PMC8710810.

Figueiredo, U. J. de, Nunes, J. A. R., & Valle, C. B. do. (2012). Estimation of genetic parameters and selection of Brachiaria humidicola progenies using a selection index. In Crop

Breeding and Applied Biotechnology (Vol. 12, Issue 4, pp. 237–244). FapUNIFESP (SciELO). https://doi.org/10.1590/s1984-70332012000400002

Francisco, F. R., Aono, A. H., da Silva, C. C., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., Fritsche-Neto, R., Souza, L. M., & de Souza, A. P. (2021). Unravelling Rubber Tree Growth by Integrating GWAS and Biological Network-Based Approaches. *Frontiers in plant science*, *12*, 768589. https://doi.org/10.3389/fpls.2021.768589

Franke, R., Hemm, M. R., Denault, J. W., Ruegger, M. O., Humphreys, J. M., & Chapple, C. (2002). Changes in secondary metabolism and deposition of an unusual lignin in the ref8 mutant of Arabidopsis. In The Plant Journal (Vol. 30, Issue 1, pp. 47–59). Wiley. https://doi.org/10.1046/j.1365-313x.2002.01267.x

Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139. https://doi.org/10.1006/jcss.1997.1504 (1997).

Friml J, Vieten A, Sauer M, Weijers D, Schwarz H, et al. 2003. Efflux-dependent auxin gradients establish the apical-basal axis of Arabidopsis. Nature 426:147–53

Gaut, B. S. (2002). Evolutionary dynamics of grass genomes. In New Phytologist (Vol. 154, Issue 1, pp. 15–28). Wiley. https://doi.org/10.1046/j.1469-8137.2002.00352.x

Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1 (2006).

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS One 9, e90346. doi: 10.1371/journal.pone.0090346

Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., & Hayes, B. J. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proceedings. Biological sciences*, *283*(1835), 20160569. https://doi.org/10.1098/rspb.2016.0569

Gorelova, V.; Ambach, L.; Rébeillé, F.; Stove, C.; Van Der Straeten, D. Folates in plants: Research advances and progress in crop biofortification. Front. Chem. 2017, 5, 21.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883

Granato, I., Cuevas, J., Luna-Vázquez, F., Crossa, J., Montesinos-López, O., Burgueño, J., & Fritsche-Neto, R. (2018). BGGE: A New Package for Genomic-Enabled Prediction Incorporating Genotype × Environment Interaction Models. In G3 Genes|Genomes|Genetics (Vol. 8, Issue 9, pp. 3039–3047). Oxford University Press (OUP). https://doi.org/10.1534/g3.118.200435

Grieneisen VA, Xu J, Maree AF, Hogeweg P, Scheres B. 2007. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. Nature 449:1008–13

Grinberg, N. F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K. P., Kelly, R., Blackmore, T., Thorogood, D., King, R. D., Armstead, I., Powell, W., & Skøt, L. (2016). Implementation of Genomic Prediction in Lolium perenne (L.) Breeding Populations. In Frontiers in Plant Science (Vol. 7). Frontiers Media SA. https://doi.org/10.3389/fpls.2016.00133

Grunewald, W., De Smet, I., Lewis, D.R. et al. (2012) Transcription factor WRKY23 assists auxin distribution patterns during Arabidopsis root development through local control on flavonol biosynthesis. Proceedings of the National Academy of Sciences of the United States of America, 109, 1554–1559.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. Bioinformatics. 2014 Oct;30(19):2811-2. doi: 10.1093/bioinformatics/btu393. Epub 2014 Jun 14. PMID: 24930139.

Guo X, Cericola F, Fè D, Pedersen MG, Lenk I, et al. 2018. Genomic prediction in tetraploid ryegrass using allele frequencies based on genotyping by sequencing. Front Plant Sci. 9:1165.

Guo, Z., Zhou, S., Wang, S., Li, W. X., Du, H., & Xu, Y. (2021). Identification of major QTL for waterlogging tolerance in maize using genome-wide association study and bulked sample analysis. Journal of applied genetics, 62(3), 405–418.

Haile TA, Walkowiak S, N'Diaye A, Clarke JM, Hucl PJ, Cuthbert RD, Knox RE, Pozniak CJ. Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. Theor Appl Genet. 2021 Jan;134(1):381-398. doi: 10.1007/s00122-020-03703-z. Epub 2020 Nov 1. PMID: 33135095.

Hanley, S. J., Pellny, T. K., de Vega, J. J., Castiblanco, V., Arango, J., Eastmond, P. J., Heslop-Harrison, J. S. P., & Mitchell, R. A. C. (2021). Allele mining in diverse accessions of tropical grasses to improve forage quality and reduce environmental impact. Annals of Botany, 128(5), 627–637. https://doi.org/10.1093/aob/mcab101

Heer, K., Behringer, D., Piermattei, A., Bässler, C., Brandl, R., Fady, B., Jehl, H.,

Liepelt, S., Lorch, S., Piotti, A., Vendramin, G. G., Weller, M., Ziegenhagen, B., Büntgen, U., & Opgenoorth, L. (2018). Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (Abies albaMill.). In Molecular Ecology (Vol. 27, Issue 6, pp. 1428–1438). Wiley. https://doi.org/10.1111/mec.14538

Islam, M. S., McCord, P. H., Olatoye, M. O., Qin, L., Sood, S., Lipka, A. E., & Todd, J. R. (2021). Experimental evaluation of genomic selection prediction for rust resistance in sugarcane. In The Plant Genome (Vol. 14, Issue 3). Wiley. https://doi.org/10.1002/tpg2.20148

Jacobs, M. and Rubery, P.H. (1988) Naturally-occurring auxin transport regulators. Science, 241, 346–349

Jank L., Barrios S. C., do Valle C. B., Simeão R. M., Alves G. F. (2014). The value of improved pastures to Brazilian beef production. *Crop Pasture Sci.* 65, 1132–1137. doi: 10.1071/CP13319

Jia, C., Zhao, F., Wang, X., Han, J., Zhao, H., Liu, G., & Wang, Z. (2018). Genomic Prediction for 25 Agronomic and Quality Traits in Alfalfa (Medicago sativa). Frontiers in plant science, 9, 1220.

Jeong, S., Kim, J. Y., & Kim, N. (2020). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Scientific reports*, *10*(1), 19653. https://doi.org/10.1038/s41598-020-76759-y

Juliana P, He X, Marza F, Islam R, Anwar B, Poland J, Shrestha S, Singh GP, Chawade A, Joshi AK, Singh RP and Singh PK (2022) Genomic Selection for Wheat Blast in a Diversity Panel, Breeding Panel and Full-Sibs Panel. *Front. Plant Sci.* 12:745379. doi: 10.3389/fpls.2021.745379

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., Wang, B., Liu, Z., Chen, J., Li, W., Zhang, M., Xie, S., & Lai, J. (2012). Genome-wide genetic changes during modern breeding of maize. Nature genetics, 44(7), 812–815.

Jones, C., De Vega, J., Worthington, M., Thomas, A., Gasior, D., Harper, J., et al. (2021). A comparison of differential gene expression in response to the onset of water stress between three hybrid Brachiaria genotypes. Front. Plant Sci. 12:637956. doi: 10.3389/fpls.2021.637956

Jones, L., Ennos, A. R., & Turner, S. R. (2001). Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of Arabidopsis. In The Plant Journal (Vol. 26, Issue 2, pp. 205–216). Wiley.

https://doi.org/10.1046/j.1365-313x.2001.01021.x

Jouhet, J., Maréchal, E., & Block, M. A. (2007). Glycerolipid transfer for the building of membranes in plant cells. Progress in lipid research, 46(1), 37-55.

Karthika, V., Babitha, K.C., Kiranmai, K. et al. Involvement of DNA mismatch repair systems to create genetic diversity in plants for speed breeding programs. Plant Physiol. Rep. 25, 185–199 (2020). https://doi.org/10.1007/s40502-020-00521-9

Kim, J., Kim, H.-S., Choi, S.-H., Jang, J.-Y., Jeong, M.-J., & Lee, S. (2017). The Importance of the Circadian Clock in Regulating Plant Metabolism. In International Journal of Molecular Sciences (Vol. 18, Issue 12, p. 2680). MDPI AG. https://doi.org/10.3390/ijms18122680

Kim, M. K., & Kim, W. T. (2018). Telomere Structure, Function, and Maintenance in Plants. In Journal of Plant Biology (Vol. 61, Issue 3, pp. 131–136). Springer Science and Business Media LLC. https://doi.org/10.1007/s12374-018-0082-y

Kobayashi, K., Kondo, M., Fukuda, H., Nishimura, M., & Ohta, H. (2007). Galactolipid synthesis in chloroplast inner envelope is essential for proper thylakoid biogenesis, photosynthesis, and embryogenesis. In Proceedings of the National Academy of Sciences (Vol. 104, Issue 43, pp. 17216–17221). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0704680104

Kombrink, A., Sánchez-Vallet, A., & Thomma, B. P. H. J. (2011). The role of chitin detection in plant–pathogen interactions. In Microbes and Infection (Vol. 13, Issues 14–15, pp. 1168–1176). Elsevier BV. https://doi.org/10.1016/j.micinf.2011.07.010

Korte, A., Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9, 29 (2013). https://doi.org/10.1186/1746-4811-9-29

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., & Li, Y. (2018). Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. In Frontiers in Genetics (Vol. 9). Frontiers Media SA. https://doi.org/10.3389/fgene.2018.00237

Li, X., Yang, Y., Yao, J. et al. FLEXIBLE CULM 1 encoding a cinnamyl-alcohol dehydrogenase controls culm mechanical strength in rice. Plant Mol Biol 69, 685–697 (2009). https://doi.org/10.1007/s11103-008-9448-8

Liang, M., Miao, J., Wang, X., Chang, T., An, B., Duan, X., Xu, L., Gao, X., Zhang, L., Li, J., & Gao, H. (2020). Application of ensemble learning to genomic selection in chinese simmental beef cattle. In Journal of Animal Breeding and Genetics (Vol. 138, Issue 3,

pp. 291–299). Wiley. https://doi.org/10.1111/jbg.12514

Liang, M., Chang, T., An, B., Duan, X., Du, L., Wang, X., Miao, J., Xu, L., Gao, X., Zhang, L., Li, J., & Gao, H. (2021). A Stacking Ensemble Learning Framework for Genomic Prediction. In Frontiers in Genetics (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fgene.2021.600040

Lipka, A. E., Lu, F., Cherney, J. H., Buckler, E. S., Casler, M. D., & Costich, D. E. (2014). Accelerating the Switchgrass (Panicum virgatum L.) Breeding Cycle Using Genomic Selection Approaches. In D. D. Fang (Ed.), PLoS ONE (Vol. 9, Issue 11, p. e112227). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0112227

Liu, S., Feuerstein, U., Luesink, W., Schulze, S., Asp, T., Studer, B., Becker, H. C., & Dehmer, K. J. (2018). DArT, SNP, and SSR analyses of genetic diversity in Lolium perenne L. using bulk sampling. BMC genetics, 19(1), 10.

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y. C., Cheng, F., et al. (2020). Computational network biology: data, models, and applications. Phys. Rep. 846, 1–66. doi: 10.1016/j.physrep.2019.12.004

Lorenz, A. J., & Smith, K. P. (2015). Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. In Crop Science (Vol. 55, Issue 6, pp. 2657–2667). Wiley. https://doi.org/10.2135/cropsci2014.12.0827

Louis Lello, Steven G Avery, Laurent Tellier, Ana I Vazquez, Gustavo de los Campos, Stephen D H Hsu, Accurate Genomic Prediction of Human Height, Genetics, Volume 210, Issue 2, 1 October 2018, Pages 477–497, https://doi.org/10.1534/genetics.118.301267

Luo, Z., Yu, Y., Xiang, J. & Li, F. Genomic selection using a subset of snps identified by genome-wide association analysis for disease resistance traits in aquaculture species. Aquaculture 539, 736620 (2021).

Ma, W., Qiu, Z., Song, J. et al. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta 248, 1307–1318 (2018). https://doi.org/10.1007/s00425-018-2976-9

Mateescu, R. G., Garrick, D. J., & Reecy, J. M. (2017). Network Analysis Reveals Putative Genes Affecting Meat Quality in Angus Cattle. *Frontiers in genetics*, *8*, 171. https://doi.org/10.3389/fgene.2017.00171

Martins, F. B., Moraes, A. C. L., Aono, A. H., Ferreira, R. C. U., Chiari, L., Simeão, R. M., Barrios, S. C. L., Santos, M. F., Jank, L., do Valle, C. B., Vigna, B. B. Z., & de Souza, A. P. (2021). A Semi-Automated SNP-Based Approach for Contaminant Identification in

Biparental Polyploid Populations of Tropical Forage Grasses. In Frontiers in Plant Science (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fpls.2021.737919

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Endelman, J. B., & Fritsche-Neto, R. (2019a). On the accuracy of genomic prediction models considering multi-trait and allele dosage in Urochloa spp. interspecific tetraploid hybrids. Molecular Breeding, 39(7), 1-16.

Matias, F. I., Vidotti, M. S., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Carley, C. A. S., et al. (2019b). Association mapping considering allele dosage: an example of forage traits in an interspecific segmental allotetraploid Urochloa spp. panel. Crop Sci. 59, 2062–2076. doi: 10.2135/cropsci2019.03.0185

Medina, C. A., Kaur, H., Ray, I., & Yu, L.-X. (2021). Strategies to Increase Prediction Accuracy in Genomic Selection of Complex Traits in Alfalfa (Medicago sativa L.). In Cells (Vol. 10, Issue 12, p. 3372). MDPI AG. https://doi.org/10.3390/cells10123372

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157:1819–1829.

Miao, J. & Niu, L. A survey on feature selection. Procedia Comput. Sci. 91, 919–926 (2016).

Millar AJ. 2016. The intracellular dynamics of circadian clocks reach for the light of ecology and evolution. Annu Rev Plant Biol 67: 595–618.doi:10.1146/annurev-arplant-043014-115619

Montesinos-López, O.A., Montesinos-López, A., Pérez-Rodríguez, P. et al. A review of deep learning applications for genomic selection. BMC Genomics 22, 19 (2021). https://doi.org/10.1186/s12864-020-07319-x

Mohan M, Nair S, Bhagwat A, et al. Genome mapping, molecular markers and marker-assisted selection in crop plants. Mol Breed. 1997;3(2):87–103.

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., & Crossa, J. (2019). Multi-Trait, Multi-Environment Genomic Prediction of Durum Wheat With Genomic Best Linear Unbiased Predictor and Deep Learning Methods. In Frontiers in Plant Science (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fpls.2019.01311

Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, Evgeny M Zdobnov, BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes,

Molecular Biology and Evolution, Volume 38, Issue 10, October 2021, Pages 4647–4654

Murad Leite Andrade, M. H., Acharya, J. P., Benevenuto, J., de Bem Oliveira, I., Lopez, Y., Munoz, P., Resende, M. F. R., Jr., & Rios, E. F. (2022). Genomic prediction for canopy height and dry matter yield in alfalfa using family bulks. In The Plant Genome. Wiley. https://doi.org/10.1002/tpg2.20235

Mutwil, M., Usadel, B., Schuؾtte, M., Loraine, A., Ebenhoؾh, O., & Persson, S. (2009). Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm . In Plant Physiology (Vol. 152, Issue 1, pp. 29–43). Oxford University Press (OUP). https://doi.org/10.1104/pp.109.145318

Oliver, S. (2000) Guilt-by-association goes global. Nature, 403, 601–602. https://doi.org/10.1038/35001165

Parker Gaddis, K. L., Null, D. J., & Cole, J. B. (2016). Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. Journal of dairy science, 99(8), 6420–6435. https://doi.org/10.3168/jds.2015-10444

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14(4), 417-419.

Pedregosa, F. et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

Peer, W.A. and Murphy, A.S. (2007) Flavonoids and auxin transport: modulators or regulators? Trends in Plant Science, 12, 556–563.

Peer, W.A., Bandyopadhyay, A., Blakeslee, J.J. et al. (2004) Variation in expression and protein localization of the PIN family of auxin efflux facilitator proteins in flavonoid mutants with altered auxin transport in Arabidopsis thaliana. The Plant Cell, 16, 1898–1911.

Pellny TK, Lovegrove A, Freeman J, et al. 2012. Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-seq transcriptome. Plant Physiology 158: 612–627.

Pereira, G. S., Garcia, A. A. F., Margarido, G. R. A. (2018a). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinform. 19, 398. doi: 10.1186/s12859-018-2433-6

Pereira J. F., Azevedo A. L. S., Pessoa-Filho M., Romanel E. A. D. C., Pereira A. V., Vigna B. B. Z., et al. (2018b). Research priorities for next-generation breeding of tropical forages in Brazil. *Crop Breed. Appl. Biotechnol.* 18 314–319.

10.1590/1984-70332018v18n3n46

Perez, P., and de los Campos, G., 2014 Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198 (2): 483-495.

Petrasch S, Mesquida-Pesci SD, Pincot DDA, Feldmann MJ, López CM, Famula R, Hardigan MA, Cole GS, Knapp SJ, Blanco-Ulate B. Genomic prediction of strawberry resistance to postharvest fruit decay caused by the fungal pathogen Botrytis cinerea. G3 (Bethesda). 2022 Jan 4;12(1):jkab378. doi: 10.1093/g3journal/jkab378. PMID: 34791166; PMCID: PMC8728004.

Pessoa-Filho, M., Sobrinho, F. S., Fragoso, R. R., Silva Junior, O. B., and Ferreira, M. E. (2019). “A Phased Diploid Genome Assembly for the Forage Grass Urochloa Ruziziensis Based on Single-Molecule Real-Time Sequencing.” in International Plant and Animal Genome Conference XXVII, 2019, San Diego. Available at: https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1107378/a-phased-diploid-genome-assembly-for-the-forage-grass-urochloa-ruziziensis-based-on-single-molecule-real-time-sequencing.

Piles, M., Bergsma, R., Gianola, D., Gilbert, H., & Tusell, L. (2021). Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. In Frontiers in Genetics (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fgene.2021.611506

Pimenta, R.J.G., Aono, A.H., Burbano, R.C.V. *et al.* Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. *Sci Rep* 11, 15730 (2021). https://doi.org/10.1038/s41598-021-95116-1

Pimenta, R. J. G., Aono, A. H., Burbano, R. C. V., da Silva, M. F., dos Anjos, I. A., de Andrade Landell, M. G., Gonçalves, M. C., Pinto, L. R., & de Souza, A. P. (2022). Multiomic investigation of sugarcane mosaic virus resistance in sugarcane. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2022.08.18.504288

Pincot, D. D. A., Hardigan, M. A., Cole, G. S., Famula, R. A., Henry, P. M., Gordon, T. R., et al. (2020). Accuracy of genomic selection and long−term genetic gain for resistance to *Verticillium* wilt in strawberry. *Plant Genome* 13:e20054. doi: 10.1002/tpg2.20054

Piquemal, J., Lapierre, C., Myton, K., O’connell, A., Schuch, W., Grima-pettenati, J., & Boudet, A.-M. (2002). Down-regulation of Cinnamoyl-CoA Reductase induces significant changes of lignin profiles in transgenic tobacco plants. In The Plant Journal (Vol. 13, Issue 1, pp. 71–83). Wiley. https://doi.org/10.1046/j.1365-313x.1998.00014.x

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PloS one, 7(2), e32253.

Popescu, M. C., Balas, V., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. WSEAS Trans. Circuits Syst. 8, 579–588 (2009).

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, X., & Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. *Acta biochimica et biophysica Sinica*, *51*(10), 981–988. https://doi.org/10.1093/abbs/gmz080

Resende, M. D. V. de. (2016). Software Selegen-REML/BLUP: a useful tool for plant breeding. In Crop Breeding and Applied Biotechnology (Vol. 16, Issue 4, pp. 330–339). FapUNIFESP (SciELO). https://doi.org/10.1590/1984-70332016v16n4a49

Rios, E. F., Andrade, M. H. M. L., Resende, M. F. R., Jr, Kirst, M., de Resende, M. D. V., de Almeida Filho, J. E., Gezan, S. A., & Munoz, P. (2021). Genomic prediction in family bulks using different traits and cross-validations in pine. In A. E. Lipka (Ed.), G3 Genes|Genomes|Genetics (Vol. 11, Issue 9). Oxford University Press (OUP). https://doi.org/10.1093/g3journal/jkab249

Rosolen, R. R., Aono, A. H., Almeida, D. A., Ferreira Filho, J. A., Horta, M., & De Souza, A. P. (2022). Network Analysis Reveals Different Cellulose Degradation Strategies Across *Trichoderma harzianum* Strains Associated With XYR1 and CRE1. *Frontiers in genetics*, *13*, 807243. https://doi.org/10.3389/fgene.2022.807243

Saballos, A., Sattler, S. E., Sanchez, E., Foster, T. P., Xin, Z., Kang, C., Pedersen, J. F., & Vermerris, W. (2012). Brown midrib2 (Bmr2) encodes the major 4-coumarate:coenzyme A ligase involved in lignin biosynthesis in sorghum (Sorghum bicolor (L.) Moench). In The Plant Journal (Vol. 70, Issue 5, pp. 818–830). Wiley. https://doi.org/10.1111/j.1365-313x.2012.04933.x

Sánchez-Vallet, A., Mesters, J. R., & Thomma, B. P. H. J. (2015). The battle for chitin recognition in plant-microbe interactions. In FEMS Microbiology Reviews (Vol. 39, Issue 2, pp. 171–183). Oxford University Press (OUP). https://doi.org/10.1093/femsre/fuu003

Sandhu, K., Aoun, M., Morris, C., & Carter, A. (2021). Genomic Selection for End-Use Quality and Processing Traits in Soft White Winter Wheat Breeding Program with Machine and Deep Learning Models. In Biology (Vol. 10, Issue 7, p. 689). MDPI AG.

https://doi.org/10.3390/biology10070689

Santelia, D., Henrichs, S., Vincenzetti, V. et al. (2008) Flavonoids redirect PIN-mediated polar auxin fluxes during root gravitropic responses. Journal of Biological Chemistry, 283, 31218–31226.

Schaefer, R. J., Michno, J. M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., et al. (2018). Integrating coexpression networks with GWAS to prioritize causal genes in maize. Plant Cell 30, 2922–2942. doi: 10.1105/tpc.18.00299

Schilmiller, A. L., Stout, J., Weng, J.-K., Humphreys, J., Ruegger, M. O., & Chapple, C. (2009). Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in Arabidopsis. In The Plant Journal (Vol. 60, Issue 5, pp. 771–782). Wiley. https://doi.org/10.1111/j.1365-313x.2009.03996.x

Schneider, M., Shrestha, A., Ballvora, A., & Léon, J. (2022). High-throughput estimation of allele frequencies using combined pooled-population sequencing and haplotype-based data processing. *Plant methods*, *18*(1), 34. https://doi.org/10.1186/s13007-022-00852-8

Scossa, F., Alseekh, S., & Fernie, A. R. (2021). Integrating multi-omics data for crop improvement. *Journal of plant physiology*, *257*, 153352. https://doi.org/10.1016/j.jplph.2020.153352

Seo, M., Koshiba, T. Transport of ABA from the site of biosynthesis to the site of action. J Plant Res 124, 501–507 (2011). https://doi.org/10.1007/s10265-011-0411-4

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. In Genome Research (Vol. 13, Issue 11, pp. 2498–2504). Cold Spring Harbor Laboratory. https://doi.org/10.1101/gr.1239303

Simeão RM, Valle CB do, Alves GF, Moreira DAL, Silva DR da, Araújo D de F, Ferreira RCU, Barrios SCL, Jank L, Caramalac GR, Naka IM, Calixto S and Carvalho J. de (2012) Melhoramento de Brachiaria ruziziensis tetraploide sexual na Embrapa: métodos e avanços. Embrapa, Campo Grande. Documentos 194: 1-32.

Simeão, R, Silva, A., Valle, C., Resende, M. D., & Medeiros, S. (2016). Genetic evaluation and selection index in tetraploid Brachiaria ruziziensis. In H.-P. Piepho (Ed.), Plant Breeding (Vol. 135, Issue 2, pp. 246–253). Wiley. https://doi.org/10.1111/pbr.12353

Simeão, R. M., Valle, C. B., & Resende, M. D. V. (2016). Unravelling the inheritance,QSTand reproductive phenology attributes of the tetraploid tropical grass

Brachiaria ruziziensis(Germain et Evrard). In O. A. Rognli (Ed.), Plant Breeding (Vol. 136, Issue 1, pp. 101–110). Wiley. https://doi.org/10.1111/pbr.12429

Simeão RM, Resende MDV, Alves RS, Pessoa-Filho M, Azevedo ALS, Jones CS, Pereira JF and Machado JC (2021) Genomic Selection in Tropical Forage Grasses: Current Status and Future Applications. Front. Plant Sci. 12:665195. doi: 10.3389/fpls.2021.665195

Simeão-Resende, R. M., Casler, M. D., and Resende, M. D. V. (2014). Genomic selection in forage breeding: accuracy and methods. Crop Sci. 54, 143–156. doi: 10.2135/cropsci2013.05.0353

Soneson C, Love MI, Robinson MD (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." F1000Research, 4. doi: 10.12688/f1000research.7563.1.

Song, J., Wang, Z. RNAi-mediated suppression of the phenylalanine ammonia-lyase gene in Salvia miltiorrhiza causes abnormal phenotypes and a reduction in rosmarinic acid biosynthesis. J Plant Res 124, 183–192 (2011). https://doi.org/10.1007/s10265-010-0350-5

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23, 1164–1167. doi: 10.1093/bioinformatics/btm069

Stakhova, L.; Stakhov, L.; Ladygin, V. Effects of exogenous folic acid on the yield and amino acid content of the seed of Pisum sativum L. and Hordeum vulgare L. Appl. Biochem. Microbiol. 2000, 36, 85–89.

Steinfath, M., Gärtner, T., Lisec, J. et al. Prediction of hybrid biomass in Arabidopsis thaliana by selected parental SNP and metabolic markers. Theor Appl Genet 120, 239–247 (2010). https://doi.org/10.1007/s00122-009-1191-2

Stevens P.F. (2001) Onwards. Angiosperm phylogeny website. Version 12, July 2012 [and more or less continuously updated since]. Available at: http://www.mobot.org/MOBOT/research/APweb/ (Accessed April, 2022).

Thaikua, S., Ebina, M., Yamanaka, N., Shimoda, K., Suenaga, K., and Kawamoto, Y. (2016). Tightly clustered markers linked to an apospory-related gene region and quantitative trait loci mapping for agronomic traits in Brachiaria hybrids. Grassl. Sci. 62, 69–80. doi: 10.1111/grs.12115

Thakral, V., Yadav, H., Padalkar, G., Kumawat, S., Raturi, G., Kumar, V., ... & Singh, M. (2022). Recent Advances and Applicability of GBS, GWAS, and GS in Polyploid Crops. *Genotyping by Sequencing for Crop Improvement*, 328-354.

Tong, H., & Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. In Journal of Plant Physiology (Vol. 257, p. 153354). Elsevier BV. https://doi.org/10.1016/j.jplph.2020.153354

Tuteja, N. (2003). Plant DNA helicases: the long unwinding road. In Journal of Experimental Botany (Vol. 54, Issue 391, pp. 2201–2214). Oxford University Press (OUP). https://doi.org/10.1093/jxb/erg246

Varshney, R. K. (2021). The Plant Genome special issue: Advances in genomic selection and application of machine learning in genomic prediction for crop improvement. In The Plant Genome (Vol. 14, Issue 3). Wiley. https://doi.org/10.1002/tpg2.20178

Vigna, B. B. Z., de Oliveira, F. A., de Toledo-Silva, G., da Silva, C. C., do Valle, C. B., and de Souza, A. P. (2016a). Leaf transcriptome of two highly divergent genotypes of Urochloa humidicola (Poaceae), a tropical polyploid forage grass adapted to acidic soils and temporary flooding areas. BMC Genomics 17:910. doi: 10.1186/s12864-016-3270-5

Vigna, B. B. Z., Santos, J. C. S., Jungmann, L., do Valle, C. B., Mollinari, M., Pastina, M. M., et al. (2016b). Evidence of allopolyploidy in Urochloa humidicola based on cytological analysis and genetic linkage mapping. PLoS One 11:e0153764. doi: 10.1371/journal.pone.0153764

Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered. 2002;93(1):77–8.

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 132(3), 669–686.

Wagner, A., Donaldson, L., Kim, H., Phillips, L., Flint, H., Steward, D., Torr, K., Koch, G., Schmitt, U., & Ralph, J. (2008). Suppression of 4-Coumarate-CoA Ligase in the Coniferous Gymnosperm Pinus radiata . In Plant Physiology (Vol. 149, Issue 1, pp. 370–383). Oxford University Press (OUP). https://doi.org/10.1104/pp.108.125765

Waldmann, P., Pfeiffer, C., & Mészáros, G. (2020). Sparse Convolutional Neural Networks for Genome-Wide Prediction. In Frontiers in Genetics (Vol. 11). Frontiers Media SA. https://doi.org/10.3389/fgene.2020.00025

Walter, A., Liebisch, F., & Hund, A. (2015). Plant phenotyping: from bean weighing to image analysis. Plant methods, 11, 14.

Wang, X., Xu, Y., Hu, Z., & Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. In The Crop Journal (Vol. 6, Issue 4, pp.

330–340). Elsevier BV. https://doi.org/10.1016/j.cj.2018.03.001

Wang, X., Shi, S., Wang, G. et al. Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. J Animal Sci Biotechnol 13, 60 (2022). https://doi.org/10.1186/s40104-022-00708-0

Wang, Y., Sun, G., Zeng, Q. et al. Predicting Growth Traits with Genomic Selection Methods in Zhikong Scallop (Chlamys farreri). Mar Biotechnol 20, 769–779 (2018).

Wang, Z., Chapman, D., Morota, G., & Cheng, H. (2020). A Multiple-Trait Bayesian Variable Selection Regression Method for Integrating Phenotypic Causal Networks in Genome-Wide Association Studies. *G3 (Bethesda, Md.)*, *10*(12), 4439–4448. https://doi.org/10.1534/g3.120.401618

Wickham, H., and Chang, W. (2016). Package 'ggplot2'. Vienna: R Foundation for Statistical Computing. doi: 10.1007/978-3-319-24277-4

Winkel-Shirley B (2001) It takes a garden. How work on diverse plant species has contributed to an understanding of flavonoid metabolism. Plant Physiol 127:1399–1404

Wolc, A., & Dekkers, J. (2022). Application of Bayesian genomic prediction methods to genome-wide association analyses. *Genetics, selection, evolution : GSE*, *54*(1), 31. https://doi.org/10.1186/s12711-022-00724-8

Wolfe, C.J., Kohane, I.S. and Butte, A.J. (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinform., 6, 227–227. https://doi.org/10.1186/1471-2105-6-227

Worthington M., Perez J. G., Mussurova S., Silva-Cordoba A., Castiblanco V., Jones C., et al. . (2021). A new genome allows the identification of genes associated with natural variation in aluminium tolerance in *Brachiaria* grasses. *J. Exp. Bot.* 72, 302–319. doi: 10.1093/jxb/eraa469

Woodward, A. W. (2005). Auxin: Regulation, Action, and Interaction. In Annals of Botany (Vol. 95, Issue 5, pp. 707–735). Oxford University Press (OUP). https://doi.org/10.1093/aob/mci083

Xu, B., Escamilla-Treviño, L. L., Sathitsuksanoh, N., Shen, Z., Shen, H., Percival Zhang, Y.-H., Dixon, R. A., & Zhao, B. (2011). Silencing of 4-coumarate:coenzyme A ligase in switchgrass leads to reduced lignin content and improved fermentable sugar yields for biofuel production. In New Phytologist (Vol. 192, Issue 3, pp. 611–625). Wiley. https://doi.org/10.1111/j.1469-8137.2011.03830.x

Xu, Y., Wang, X., Ding, X. et al. Genomic selection of agronomic traits in hybrid rice

using an NCII population. Rice 11, 32 (2018). https://doi.org/10.1186/s12284-018-0223-4

Yan, Z., Huang, H., Freebern, E., Santos, D. J., Dai, D., Si, J., et al. (2020). Integrating RNA-Seq with GWAS reveals novel insights into the molecular mechanism underpinning ketosis in cattle. BMC Genomics 21:489. doi: 10.1186/s12864-020-06909-z

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, *5*, 3231. https://doi.org/10.1038/ncomms4231

Yang, J., Jiang, H., Yeh, C. T., Yu, J., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2015). Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. The Plant journal : for cell and molecular biology, 84(3), 587–596.

Yoon, J., Choi, H., & An, G. (2015). Roles of lignin biosynthesis and regulatory genes in plant development. In Journal of Integrative Plant Biology (Vol. 57, Issue 11, pp. 902–912). Wiley. https://doi.org/10.1111/jipb.12422

Yu, H., Zhang, F., Wang, G., Liu, Y., & Liu, D. (2012). Partial deficiency of isoleucine impairs root development and alters transcript levels of the genes involved in branched-chain amino acid and glucosinolate metabolism in Arabidopsis. In Journal of Experimental Botany (Vol. 64, Issue 2, pp. 599–612). Oxford University Press (OUP). https://doi.org/10.1093/jxb/ers352

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., Li, J., & Simianer, H. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PloS one, 9(3), e93017. https://doi.org/10.1371/journal.pone.0093017

Zhang, H., Wang, X., Pan, Q., Li, P., Liu, Y., Lu, X., Zhong, W., Li, M., Han, L., Li, J., Wang, P., Li, D., Liu, Y., Li, Q., Yang, F., Zhang, Y. M., Wang, G., & Li, L. (2019). QTG-Seq Accelerates QTL Fine Mapping through QTL Partitioning and Whole-Genome Sequencing of Bulked Segregant Samples. Molecular plant, 12(3), 426–437.

Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, Whitaker VM, Pérez-Enciso M. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. Front Plant Sci. 2020 Feb 6;11:25. doi: 10.3389/fpls.2020.00025. PMID: 32117371; PMCID: PMC7015897.

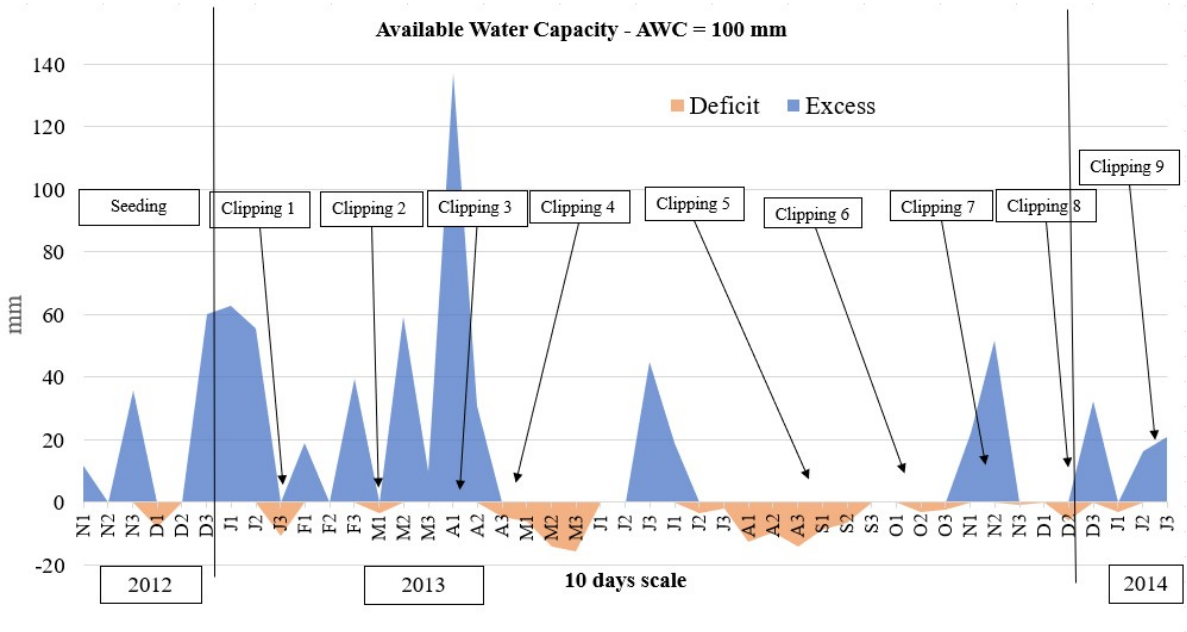
Zhou, W., Bellis, E.S., Stubblefield, J., Causey, J., Qualls, J., Walker, K. and Huang, X. (2019) Minor QTLs mining through the combination of GWAS and machine learning

feature selection. BioRxiv, 712190. https://doi.org/10.1101/712190

Zou, C., Wang, P., & Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. Plant biotechnology journal, 14(10), 1941–1955.

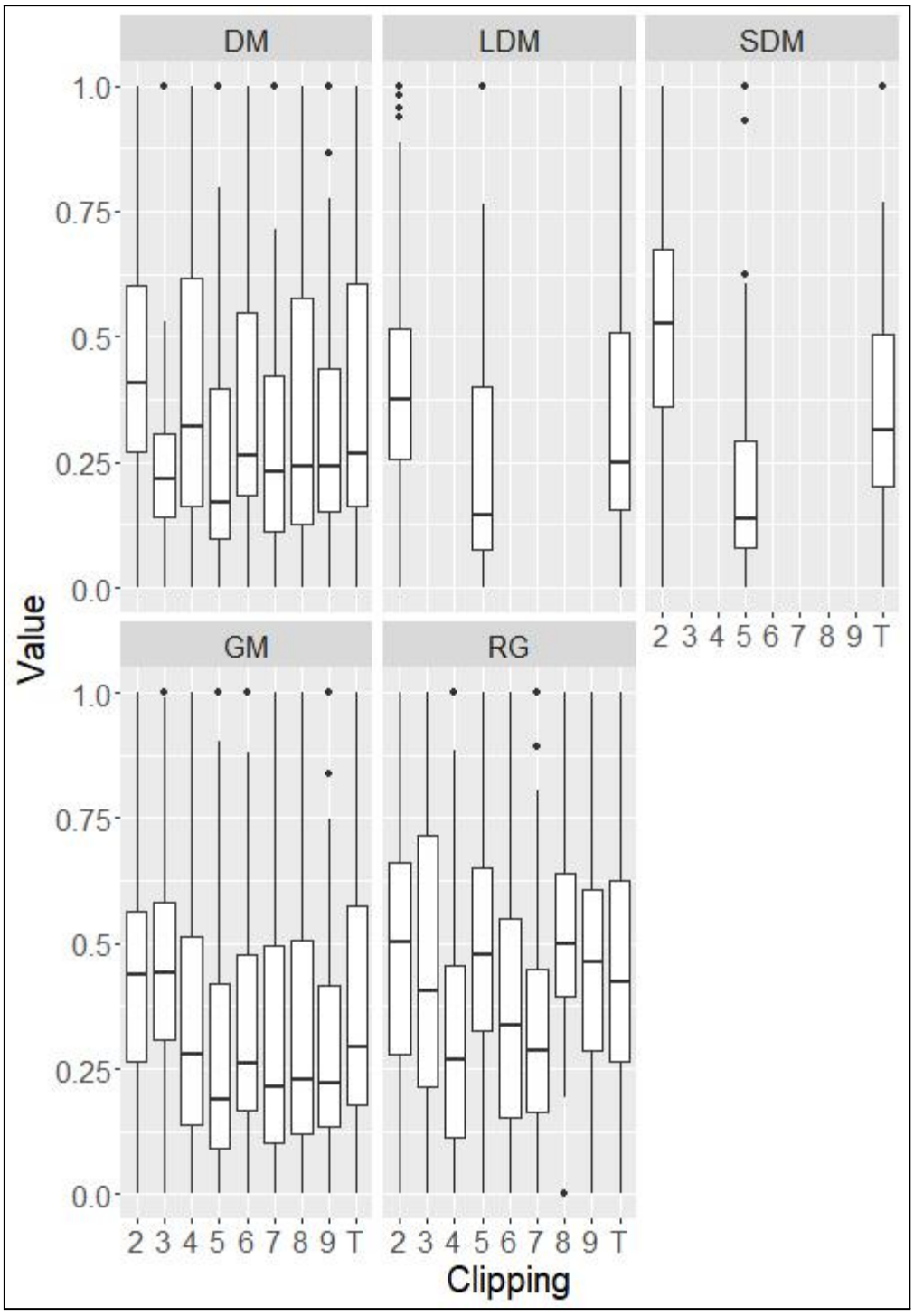
Supplementary Information

1- Supplementary Materials and Methods

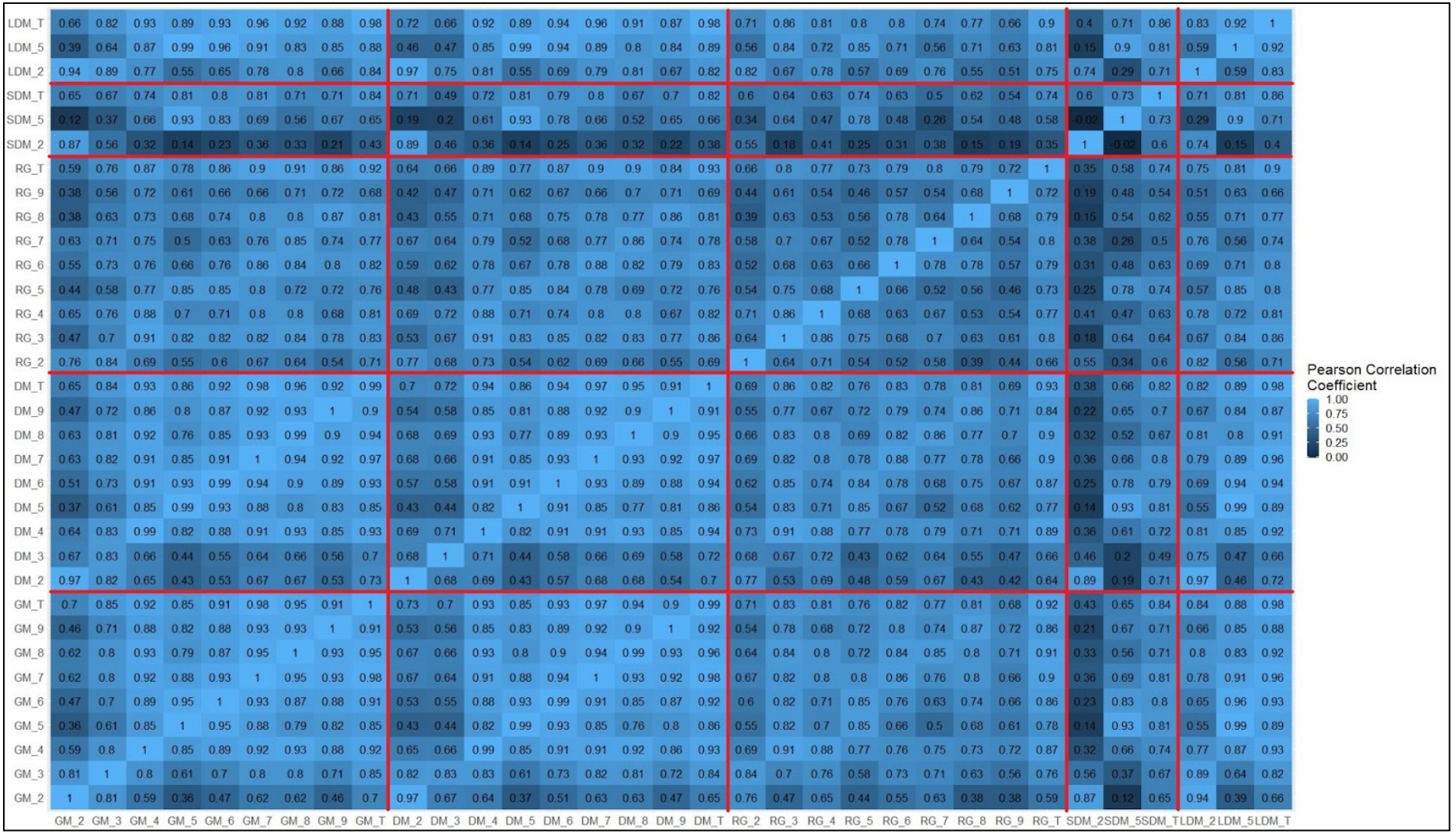


Supplementary Figure S1. Climatological water balances through available water capacity (AWC = 100mm) in a 10 days scale during the period of 2012 November to 2014 January in Embrapa Gado de Corte, Campo Grande, MS.

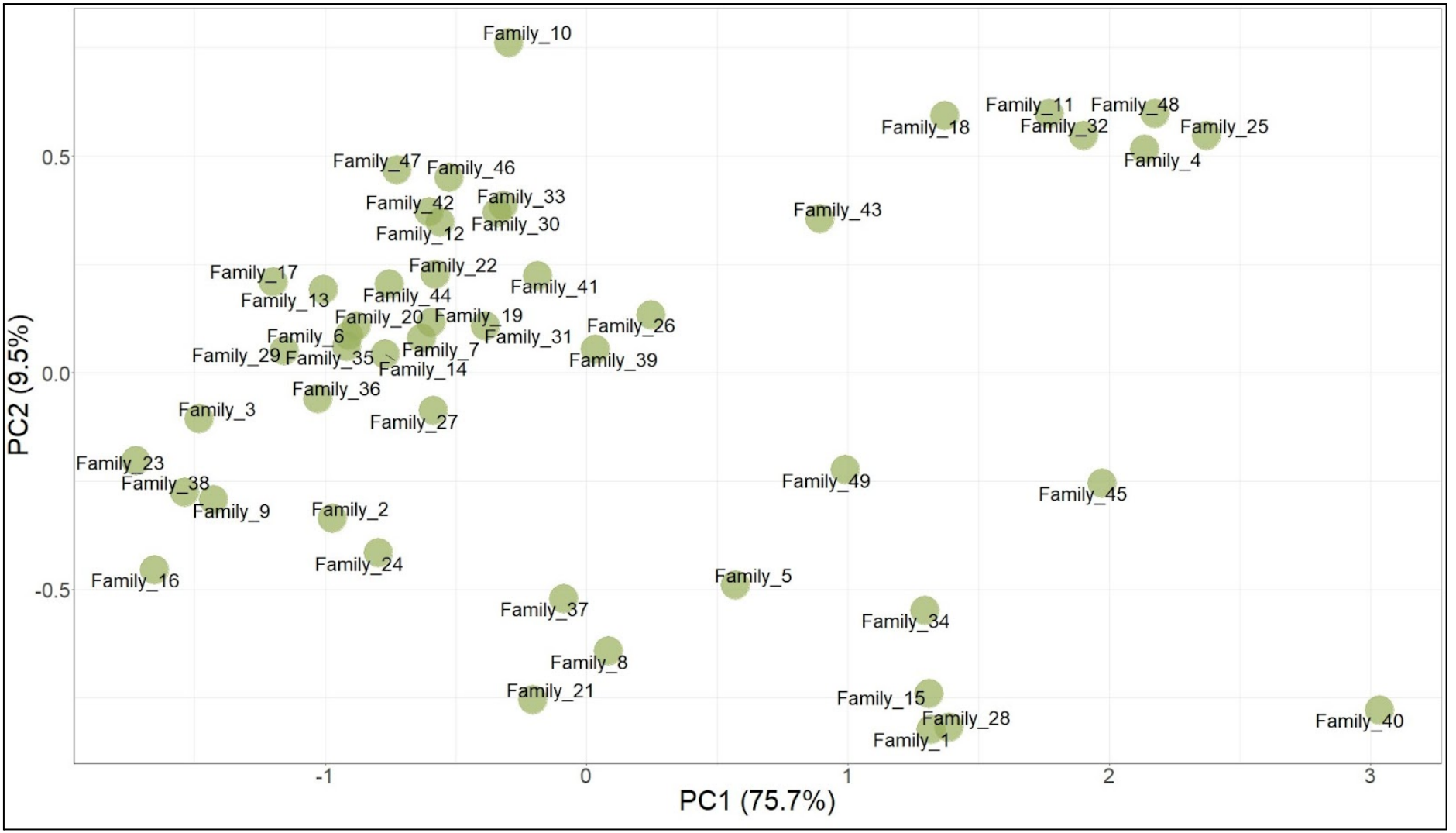
2 - Supplementary Results



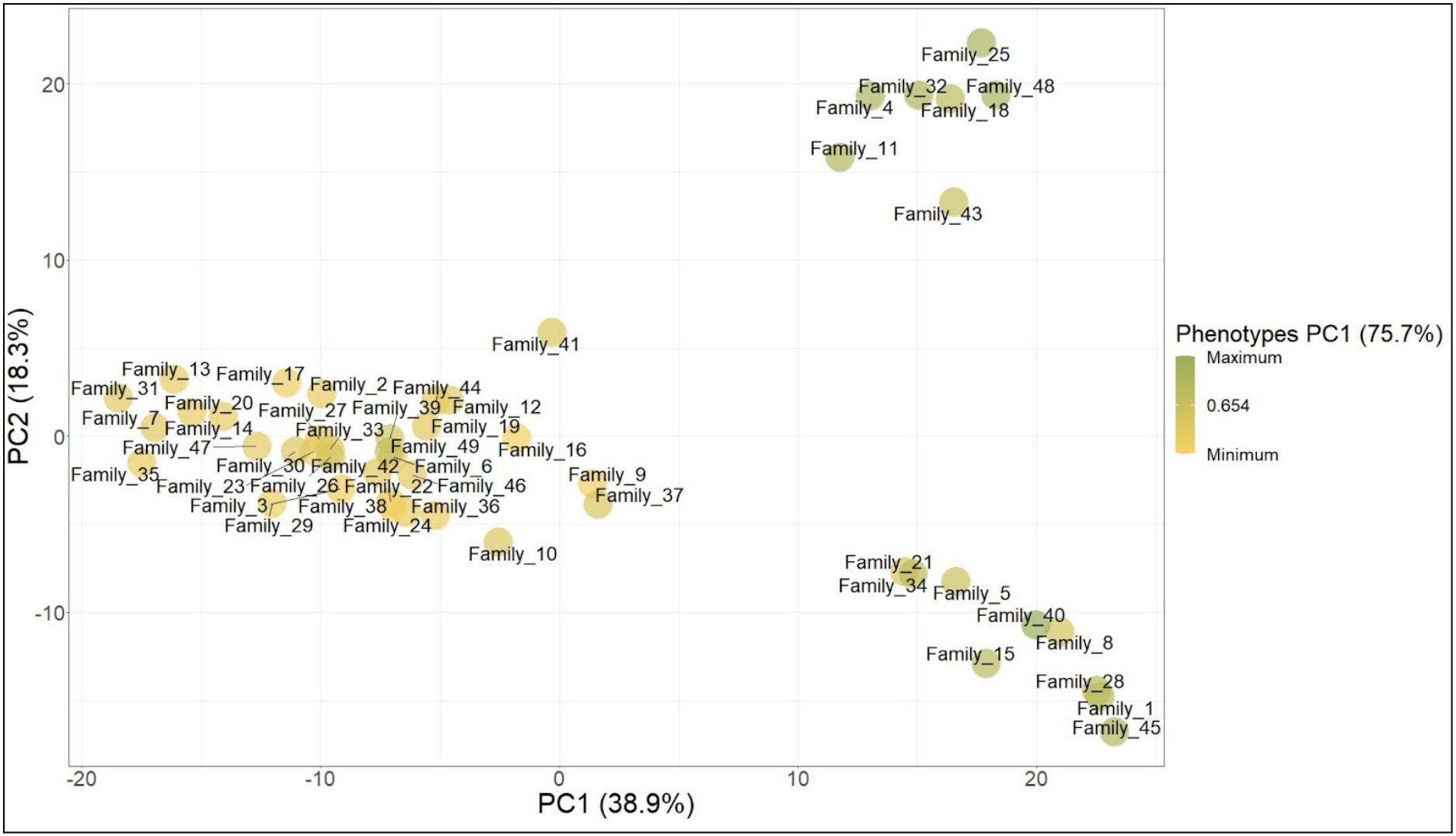
Supplementary Figure S2. Box plots of the scaled phenotype clippings. DM, LDM, SDM, GM and RG stands for dry matter, leaf dry matter, stem dry matter, green matter and regrowth.



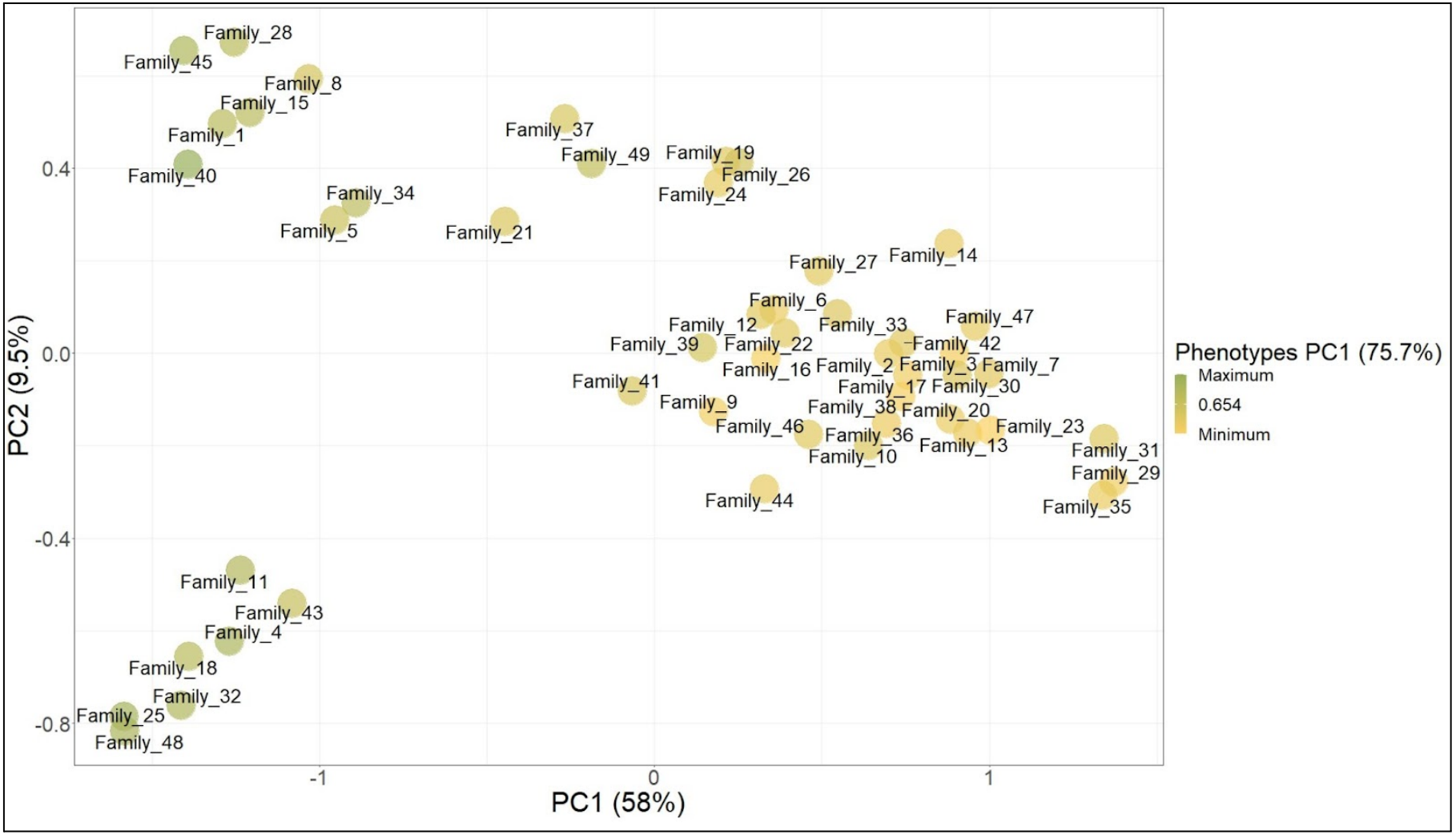
Supplementary Figure S3. Heat map showing the correlation between all trait clippings. Red lines separate different phenotypes. DM, LDM, SDM, GM and RG stands for dry matter, leaf dry matter, stem dry matter, green matter and regrowth.



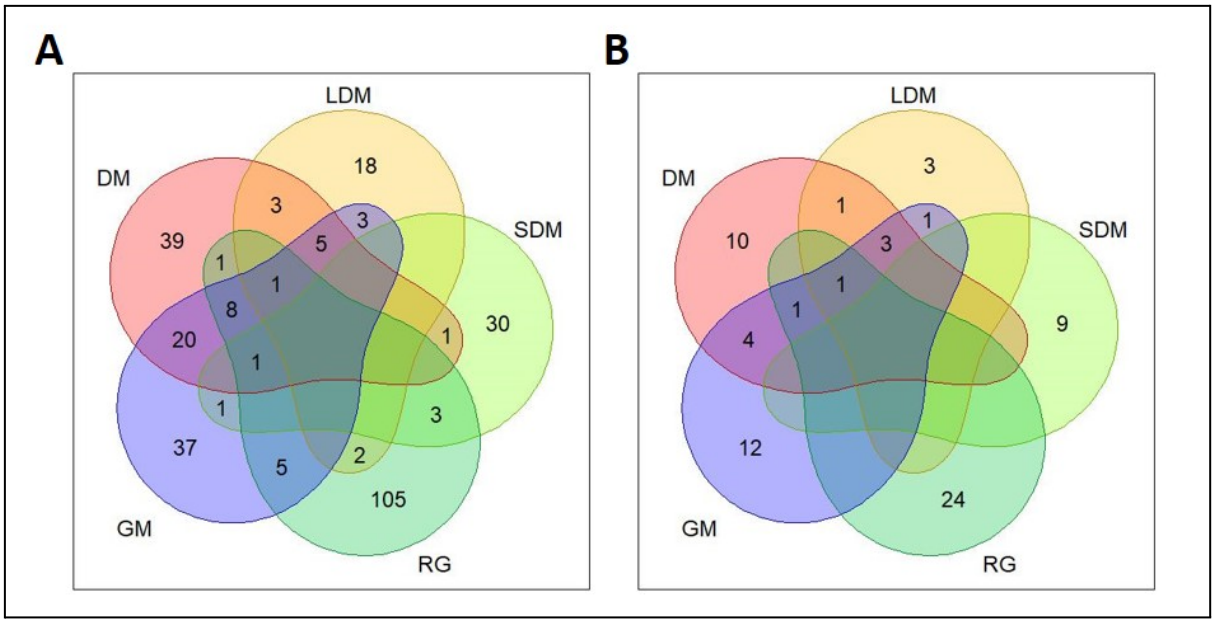
Supplementary Figure S4. Principal component analysis scatter plot of the families phenotypes. The axes represent the first and second principal components, which explain 75.7% and 9.5% of the variance, respectively.



Supplementary Figure S5. Principal component analysis scatter plot of the families genotyping, total of 28,106 markers. The axes represent the first and second principal components, which explain 38.9% and 18.3% of the variance, respectively.



Supplementary Figure S6. Principal component analysis scatter plot of the families genotyping, considering only the Major Importance markers (69 markers). The axes represent the first and second principal components, which explain 58% and 9.5% of the variance, respectively.

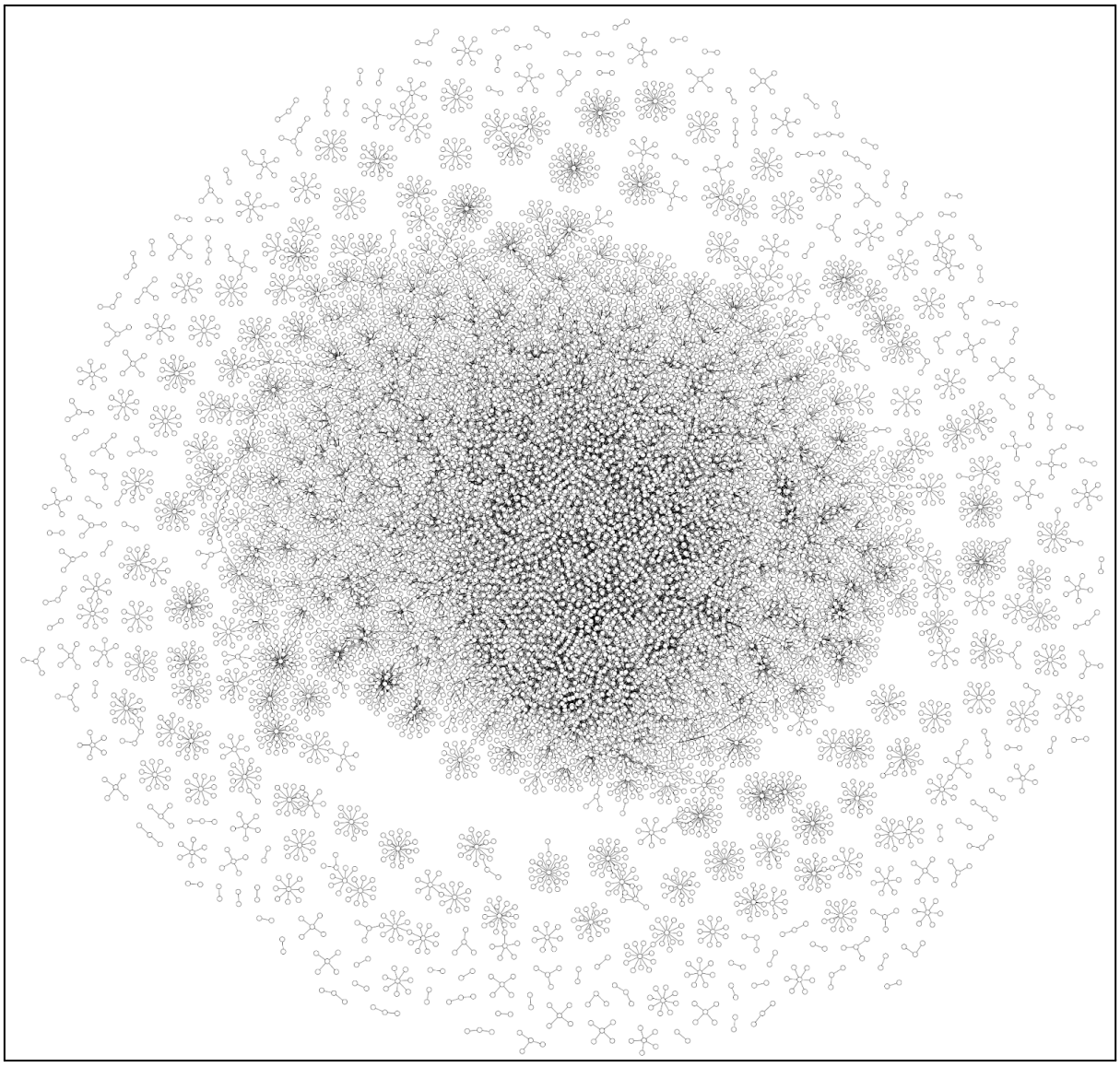


Supplementary Figure S7. Venn's diagrams showing the logic relation between the sets of markers identified for each phenotype. (A) Total 283 markers from the FI-2 data set and (B) 69 markers selected by the Gini importance condition. DM, LDM, SDM, GM and RG stands for dry matter, leaf dry matter, stem dry matter, green matter and regrowth.

2.1 - Transcriptome assembly, evaluation and annotation

To obtain a set of genes expressed by the species and later evaluate their co-expression, we used a previously published transcriptome of 11 *U. ruziziensis* genotypes. The sequencing of the libraries produced a total of 1.7 billion reads, of which 95.5% (Supplementary Table S9) were retained and used in the *de novo* assembly. The transcriptome comprised 575,524 transcripts, of which 223,593 were unigenes with a transcript N50 length of 1,227 bp. After the expression level filtration, the set was reduced to 288,487 transcripts, representing 49,445 unigenes. The evaluation of the assembly completeness was done comparing the 49,445 unigenes to the Viridiplantae database. From the 425 total BUSCO groups searched, we found 297 Complete sequences (69.8%), 48.2% as a single-copy and 21.6% as duplicated copies, in addition to 74 (17.4%) and 54 (12.8%) fragmented and missing sequences, respectively.

In the functional annotation, we aligned the transcripts to the UniProt database and obtained 197,045 GO terms associated using the Trinotate software; from those, 6,156 were unique GO terms. This set of genes and GO terms were used for the biological process GO terms enrichment of the genes linked to the Major importance markers and the sets obtained in the co-expression networks.



Supplementary Figure S8. *Urochloa ruziziensis* 11 genotypes transcriptome gene co-expression network (GCN).

Supplementary Tables

Supplementary Table S1. Progeny means narrow sense heritability (h^2) for all phenotype clippings.

Phenotype	Clipping	Heritability (h^2)
Green Matter	2	0.70
	3	0.71
	4	0.73
	5	0.89
	6	0.87
	7	0.87
	8	0.90
	9	0.87
	T	0.87
Dry Matter	2	0.70
	3	0.54
	4	0.83
	5	0.88
	6	0.87
	7	0.87
	8	0.89
	9	0.86
	T	0.89
Regrowth	2	0.73
	3	0.86
	4	0.84
	5	0.78
	6	0.67
	7	0.72
	8	0.63
	9	0.63
	T	0.85
Leaf Dry Matter	2	0.81
	5	0.92

	T	0.88
Stem Dry Matter	2	0.44
	5	0.84
	T	0.74

Supplementary Table S2. Modeling results for all models and phenotype clippings organized to show de Tukey's test results comparing the models for the same phenotype clipping.

Trait	Clipping	Model	Predictive ability CD	Tukey's test comparing models	Predictive ability FI-1	Tukey's test comparing models	Predictive ability FI-2	Tukey's test comparing models	Tukey's test comparing data sets correlation (CD/FI-1/FI-2)	MSE CD	Tukey's test comparing models	MSE FI-1	Tukey's test comparing models	MSE FI-2	Tukey's test comparing models	Tukey's test comparing data sets MSE (CD/FI-1/FI-2)
Green Matter	2	RKHS	0.653	ab	0.848	a	0.857	a	b/a/a	0.029	de	0.017	e	0.018	cd	a/c/b
		BRR	0.625	bc	0.822	b	0.837	bc	b/a/a	0.030	d	0.018	d	0.018	cd	a/b/c
		SVM	0.668	a	0.840	a	0.829	c	b/a/a	0.029	e	0.019	cd	0.022	a	a/b/c
		RF	0.606	c	0.841	a	0.850	ab	b/a/a	0.033	c	0.019	c	0.018	bc	a/b/c
		AB	0.551	d	0.805	c	0.844	abc	c/b/a	0.036	b	0.020	b	0.017	d	a/b/c
		MLP	0.541	d	0.738	d	0.845	ab	c/b/a	0.041	a	0.028	a	0.019	b	a/b/c
	3	RKHS	0.743	a	0.837	b	0.870	a	c/b/a	0.030	e	0.021	c	0.016	d	a/b/c
		BRR	0.714	b	0.812	c	0.788	d	c/a/b	0.033	d	0.023	b	0.027	b	a/c/b
		SVM	0.749	a	0.835	b	0.850	b	c/b/a	0.030	e	0.022	c	0.019	c	a/c/b
		RF	0.695	b	0.858	a	0.869	a	b/a/a	0.036	c	0.019	d	0.017	d	a/b/c
		AB	0.656	c	0.865	a	0.877	a	b/a/a	0.039	b	0.018	e	0.016	d	a/b/c

		MLP	0.630	d	0.794	d	0.809	c	b/a/a	0.044	a	0.026	a	0.029	a	a/b/c
	4	**RKHS**	0.847	a	0.906	b	0.894	b	c/a/b	0.020	d	0.014	c	0.015	c	a/c/b
		BRR	0.828	bc	0.874	c	0.855	c	c/a/b	0.022	c	0.017	b	0.020	b	a/c/b
		SVM	0.839	ab	0.901	b	0.895	b	b/a/a	0.022	c	0.017	b	0.015	c	a/c/b
		RF	0.829	abc	0.923	a	0.912	a	c/a/b	0.023	bc	0.011	d	0.013	d	a/c/b
		AB	0.819	c	0.930	a	0.902	ab	c/a/b	0.023	b	0.010	e	0.014	cd	a/c/b
		MLP	0.739	d	0.863	d	0.829	d	c/a/b	0.032	a	0.018	a	0.026	a	a/c/b
	5	**RKHS**	0.843	a	0.909	b	0.866	c	c/a/b	0.018	d	0.012	c	0.016	bc	a/c/b
		BRR	0.823	b	0.877	c	0.854	cd	c/a/b	0.019	c	0.015	b	0.019	b	a/b/a
		SVM	0.833	ab	0.902	b	0.843	d	b/a/b	0.020	b	0.015	b	0.018	bc	a/b/a
		RF	0.839	ab	0.931	a	0.917	a	c/a/b	0.019	cd	0.010	d	0.012	c	a/c/b
		AB	0.827	ab	0.927	a	0.892	b	c/a/b	0.019	c	0.011	d	0.015	bc	a/c/b
		MLP	0.735	c	0.854	d	0.734	e	b/a/b	0.028	a	0.016	a	0.069	a	a/c/b
	6	**RKHS**	0.826	a	0.885	b	0.863	b	c/a/b	0.020	d	0.015	c	0.018	c	a/c/b
		BRR	0.798	b	0.858	c	0.845	c	c/a/b	0.023	b	0.017	b	0.019	c	a/c/b
		SVM	0.824	a	0.878	b	0.843	c	c/a/b	0.022	c	0.017	b	0.022	b	a/c/b
		RF	0.816	ab	0.926	a	0.903	a	c/a/b	0.022	c	0.010	d	0.012	e	a/c/b
		AB	0.807	ab	0.922	a	0.888	a	c/a/b	0.022	bc	0.010	d	0.013	d	a/c/b
		MLP	0.711	c	0.796	d	0.813	d	b/a/a	0.032	a	0.025	a	0.024	a	a/c/b
	7	**RKHS**	0.795	a	0.889	b	0.844	b	c/a/b	0.023	cd	0.014	c	0.019	c	a/c/b

		BRR	0.783	a	0.853	c	0.809	c	c/a/b	0.024	bc	0.017	b	0.022	b	a/c/b
		SVM	0.793	a	0.879	b	0.822	c	c/a/b	0.024	b	0.017	b	0.021	b	a/c/b
		RF	0.795	a	0.921	a	0.923	a	b/a/a	0.023	cd	0.011	d	0.010	e	a/b/b
		AB	0.793	a	0.911	a	0.909	a	b/a/a	0.022	d	0.011	d	0.012	d	a/c/b
		MLP	0.670	b	0.841	c	0.732	d	c/a/b	0.035	a	0.018	a	0.034	a	a/c/b
	8	**RKHS**	0.858	a	0.927	ab	0.881	b	c/a/b	0.025	e	0.014	c	0.020	c	a/c/b
		BRR	0.840	a	0.902	d	0.818	d	b/a/c	0.027	cd	0.017	a	0.030	a	b/c/a
		SVM	0.857	a	0.921	bc	0.873	bc	c/a/b	0.026	d	0.017	a	0.021	c	b/c/a
		RF	0.848	a	0.932	a	0.913	a	c/a/b	0.027	bc	0.013	d	0.015	d	a/c/b
		AB	0.839	a	0.932	a	0.917	a	c/a/b	0.028	b	0.012	d	0.014	d	a/c/b
		MLP	0.778	b	0.911	cd	0.858	c	c/a/b	0.036	a	0.016	b	0.024	b	a/c/b
	9	**RKHS**	0.799	a	0.877	b	0.897	c	c/b/a	0.023	cd	0.016	c	0.012	d	a/b/c
		BRR	0.792	a	0.862	c	0.866	d	b/a/a	0.024	bcd	0.016	b	0.016	b	a/b/b
		SVM	0.799	a	0.868	bc	0.888	c	c/b/a	0.024	b	0.018	a	0.014	c	a/b/b
		RF	0.802	a	0.920	a	0.944	a	c/b/a	0.023	d	0.011	d	0.008	f	a/b/c
		AB	0.790	a	0.912	a	0.931	b	c/b/a	0.024	bc	0.011	d	0.009	e	a/b/c
		MLP	0.682	b	0.848	d	0.836	e	b/a/a	0.035	a	0.017	ab	0.023	a	a/b/c
	Total	**RKHS**	0.807	a	0.879	b	0.867	c	b/a/a	0.025	d	0.017	d	0.018	d	a/c/b
		BRR	0.793	a	0.840	c	0.852	d	c/b/a	0.026	bc	0.021	b	0.019	c	a/b/c
		SVM	0.806	a	0.872	b	0.843	d	c/a/b	0.026	bcd	0.019	c	0.021	b	a/b/c

		RF	0.795	a	0.919	a	0.937	a	c/b/a	0.027	b	0.012	e	0.009	f	a/b/c
		AB	0.796	a	0.915	a	0.920	b	b/a/a	0.026	cd	0.012	e	0.012	e	a/b/b
		MLP	0.704	b	0.809	d	0.769	e	c/a/b	0.037	a	0.024	a	0.036	a	a/b/b
Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing models**	**Predictive ability FI-1**	**Tukey's test comparing models**	**Predictive ability FI-2**	**Tukey's test comparing models**	**Tukey's test comparing data sets correlation (CD/FI-1/FI-2)**	**MSE CD**	**Tukey's test comparing models**	**MSE FI-1**	**Tukey's test comparing models**	**MSE FI-2**	**Tukey's test comparing models**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**
Dry Matter	**2**	**RKHS**	0.601	a	0.851	ab	0.537	c	b/a/c	0.033	d	0.018	d	0.044	c	b/c/a
		BRR	0.580	ab	0.866	a	0.606	b	c/a/b	0.034	d	0.015	e	0.039	d	b/c/a
		SVM	0.604	a	0.843	b	0.447	d	b/a/c	0.033	d	0.020	c	0.049	b	b/c/a
		RF	0.547	bc	0.795	cd	0.638	b	c/a/b	0.037	c	0.023	a	0.035	d	a/c/b
		AB	0.518	cd	0.789	d	0.698	a	c/a/b	0.039	b	0.022	b	0.029	e	a/c/b
		MLP	0.505	d	0.810	c	0.380	e	b/a/c	0.045	a	0.019	c	0.091	a	a/c/b
	3	**RKHS**	0.754	a	0.851	a	0.802	b	c/a/b	0.015	d	0.011	d	0.014	b	a/c/b
		BRR	0.732	ab	0.801	b	0.799	b	b/a/a	0.015	cd	0.014	c	0.014	b	a/b/b

		SVM	0.750	a	0.814	b	0.808	b	b/a/a	0.017	b	0.016	b	0.015	b	a/b/b
		RF	0.731	ab	0.856	a	0.840	a	c/a/b	0.016	c	0.012	d	0.013	b	a/c/b
		AB	0.715	b	0.863	a	0.808	b	c/a/b	0.015	d	0.011	e	0.019	b	b/c/a
		MLP	0.566	c	0.702	c	0.692	c	b/a/a	0.026	a	0.020	a	0.048	a	b/c/a
	4	**RKHS**	0.869	a	0.930	a	0.935	ab	b/a/a	0.019	d	0.012	c	0.010	d	a/b/c
		BRR	0.845	c	0.898	c	0.892	c	b/a/a	0.021	c	0.014	b	0.016	b	a/c/b
		SVM	0.865	ab	0.915	b	0.927	b	c/b/a	0.020	c	0.014	b	0.011	c	a/c/b
		RF	0.847	bc	0.932	a	0.941	a	c/b/a	0.023	b	0.010	d	0.009	e	a/b/c
		AB	0.833	c	0.932	a	0.935	ab	b/a/a	0.023	b	0.010	d	0.009	e	a/b/b
		MLP	0.773	d	0.884	d	0.894	c	b/a/a	0.030	a	0.017	a	0.017	a	a/b/b
	5	**RKHS**	0.841	a	0.913	b	0.882	ab	c/a/b	0.017	c	0.011	d	0.014	c	a/c/b
		BRR	0.816	b	0.881	c	0.878	ab	b/a/a	0.019	b	0.014	b	0.015	bc	a/c/b
		SVM	0.831	ab	0.905	b	0.847	c	c/a/b	0.019	b	0.013	c	0.018	bc	a/c/b
		RF	0.836	a	0.933	a	0.895	a	c/a/b	0.018	b	0.009	e	0.020	b	a/c/b
		AB	0.826	ab	0.934	a	0.868	b	c/a/b	0.018	b	0.009	e	0.015	c	b/c/a
		MLP	0.722	c	0.842	d	0.686	d	b/a/c	0.028	a	0.017	a	0.049	a	b/c/a
	6	**RKHS**	0.829	a	0.900	b	0.884	b	c/a/b	0.021	c	0.014	c	0.016	c	a/c/b
		BRR	0.812	a	0.881	c	0.856	d	c/a/b	0.023	b	0.016	b	0.019	b	a/c/b
		SVM	0.828	a	0.897	b	0.871	c	c/a/b	0.023	b	0.015	b	0.019	b	a/c/b
		RF	0.819	a	0.928	a	0.917	a	b/a/a	0.023	b	0.010	d	0.011	d	a/b/b

		AB	0.813	a	0.925	a	0.916	a	b/a/a	0.022	b	0.010	d	0.011	d	a/c/b
		MLP	0.732	b	0.871	c	0.846	d	c/a/b	0.032	a	0.017	a	0.022	a	a/c/b
	7	**RKHS**	0.794	a	0.875	b	0.849	b	c/a/b	0.021	b	0.015	d	0.017	b	a/c/b
		BRR	0.782	a	0.821	c	0.838	b	c/b/a	0.022	b	0.019	b	0.018	b	a/b/c
		SVM	0.794	a	0.864	b	0.854	b	b/a/a	0.022	b	0.017	c	0.017	b	a/b/c
		RF	0.789	a	0.915	a	0.913	a	b/a/a	0.022	b	0.011	e	0.011	c	a/b/b
		AB	0.788	a	0.920	a	0.900	a	c/a/b	0.022	b	0.010	f	0.012	c	a/c/b
		MLP	0.661	b	0.796	d	0.712	c	c/a/b	0.034	a	0.021	a	0.047	a	a/c/b
	8	**RKHS**	0.875	a	0.937	bc	0.911	a	c/a/b	0.021	d	0.012	d	0.016	c	a/c/b
		BRR	0.855	a	0.908	d	0.907	ab	b/a/a	0.023	c	0.015	b	0.016	c	a/c/b
		SVM	0.871	a	0.927	c	0.897	b	c/a/b	0.022	c	0.014	c	0.020	b	a/c/b
		RF	0.860	a	0.950	a	0.915	a	c/a/b	0.025	b	0.009	e	0.016	c	a/c/b
		AB	0.865	a	0.946	ab	0.895	b	c/a/b	0.023	c	0.009	e	0.021	b	a/c/b
		MLP	0.790	b	0.904	d	0.799	c	b/a/b	0.032	a	0.016	a	0.051	a	a/c/b
	9	**RKHS**	0.792	a	0.885	bc	0.787	b	b/a/b	0.023	c	0.015	c	0.022	c	a/c/b
		BRR	0.782	a	0.861	d	0.752	c	b/a/c	0.024	c	0.016	b	0.027	b	b/c/a
		SVM	0.788	a	0.875	c	0.776	b	b/a/b	0.025	b	0.018	a	0.023	bc	b/c/a
		RF	0.791	a	0.913	a	0.925	a	c/b/a	0.024	c	0.012	e	0.011	d	a/b/c
		AB	0.782	a	0.896	b	0.935	a	c/b/a	0.023	c	0.013	d	0.009	d	a/b/c
		MLP	0.671	b	0.835	e	0.582	d	b/a/c	0.035	a	0.018	a	0.063	a	a/b/c

	Total	RKHS	0.835	a	0.916	b	0.912	c	b/a/a	0.021	c	0.013	c	0.013	d	a/b/b
		BRR	0.820	a	0.885	d	0.895	d	b/a/a	0.022	b	0.016	b	0.015	c	a/b/c
		SVM	0.833	a	0.904	c	0.896	d	b/a/a	0.023	b	0.015	b	0.016	b	a/b/c
		RF	0.830	a	0.934	a	0.943	a	b/a/a	0.022	b	0.010	d	0.010	e	a/b/c
		AB	0.825	a	0.934	a	0.931	b	b/a/a	0.021	c	0.009	d	0.009	f	a/c/b
		MLP	0.740	b	0.876	d	0.879	e	b/a/a	0.033	a	0.016	a	0.026	a	a/c/b

Trait	Clipping	Model	Predictive ability CD	Tukey's test comparing models	Predictive ability FI-1	Tukey's test comparing models	Predictive ability FI-2	Tukey's test comparing models	Tukey's test comparing data sets correlation (CD/FI-1/FI-2)	MSE CD	Tukey's test comparing models	MSE FI-1	Tukey's test comparing models	MSE FI-2	Tukey's test comparing models	Tukey's test comparing data sets MSE (CD/FI-1/FI-2)
Regrowth	2	RKHS	0.639	a	0.851	b	0.868	b	c/b/a	0.037	d	0.019	c	0.015	c	a/b/c
		BRR	0.594	b	0.843	b	0.821	c	c/a/b	0.040	c	0.018	c	0.021	b	a/c/b
		SVM	0.648	a	0.843	b	0.828	c	b/a/a	0.036	d	0.021	a	0.021	b	a/c/b
		RF	0.591	b	0.857	b	0.896	a	c/b/a	0.041	c	0.020	b	0.014	d	a/b/c
		AB	0.556	c	0.808	c	0.896	a	b/a/a	0.043	b	0.015	d	0.013	e	a/b/c

		MLP	0.493	d	0.808	c	0.819	c	b/a/a	0.050	a	0.022	a	0.025	a	a/b/c
	3	**RKHS**	0.873	ab	0.946	a	0.940	b	c/a/b	0.023	d	0.011	c	0.011	d	a/b/b
		BRR	0.878	a	0.932	b	0.905	c	c/a/b	0.022	e	0.013	b	0.017	b	a/c/b
		SVM	0.870	ab	0.930	b	0.934	b	b/a/a	0.025	c	0.016	a	0.015	c	a/c/b
		RF	0.861	b	0.946	a	0.956	a	c/b/a	0.026	b	0.010	d	0.008	e	a/b/c
		AB	0.845	c	0.950	a	0.957	a	b/a/a	0.026	b	0.009	e	0.008	e	a/b/c
		MLP	0.820	d	0.923	c	0.867	d	c/a/b	0.032	a	0.015	a	0.027	a	a/b/c
	4	**RKHS**	0.796	a	0.907	b	0.859	b	c/a/b	0.023	c	0.012	b	0.017	c	a/c/b
		BRR	0.800	a	0.899	b	0.842	c	c/a/b	0.022	d	0.011	c	0.019	b	a/c/b
		SVM	0.794	a	0.899	b	0.850	bc	c/a/b	0.023	c	0.014	a	0.018	bc	a/c/b
		RF	0.782	ab	0.918	a	0.911	a	b/a/a	0.026	b	0.011	c	0.012	d	a/c/b
		AB	0.771	b	0.925	a	0.900	a	c/a/b	0.026	b	0.010	d	0.014	d	a/c/b
		MLP	0.693	c	0.872	c	0.752	d	c/a/b	0.034	a	0.015	a	0.040	a	a/c/b
	5	**RKHS**	0.706	a	0.875	a	0.826	b	c/a/b	0.026	c	0.013	c	0.016	c	a/c/b
		BRR	0.679	b	0.859	c	0.787	c	c/a/b	0.027	b	0.014	b	0.020	b	a/c/b
		SVM	0.697	ab	0.859	bc	0.827	b	c/a/b	0.026	bc	0.015	a	0.016	c	a/c/b
		RF	0.691	ab	0.870	ab	0.869	a	b/a/a	0.026	bc	0.013	bc	0.013	d	a/b/b
		AB	0.677	b	0.876	a	0.870	a	b/a/a	0.027	b	0.012	d	0.012	d	a/b/b
		MLP	0.578	c	0.835	d	0.657	d	c/a/b	0.036	a	0.015	a	0.040	a	a/b/b
	6	**RKHS**	0.696	a	0.834	c	0.833	b	b/a/a	0.039	cd	0.023	c	0.024	b	a/b/b

		BRR	0.699	a	0.832	c	0.831	b	b/a/a	0.038	d	0.024	bc	0.023	b	a/b/b
		SVM	0.699	a	0.822	c	0.825	bc	b/a/a	0.037	d	0.024	b	0.024	b	a/b/b
		RF	0.680	a	0.862	b	0.884	a	c/b/a	0.040	c	0.019	d	0.016	c	a/b/c
		AB	0.647	b	0.889	a	0.889	a	b/a/a	0.045	b	0.016	e	0.015	c	a/b/b
		MLP	0.592	c	0.788	d	0.812	c	c/b/a	0.051	a	0.029	a	0.028	a	a/b/b
	7	**RKHS**	0.753	a	0.889	ab	0.862	b	c/a/b	0.023	d	0.012	b	0.015	c	a/c/b
		BRR	0.753	a	0.888	ab	0.848	b	c/a/b	0.022	d	0.011	d	0.016	b	a/c/b
		SVM	0.750	a	0.876	b	0.857	b	c/a/b	0.023	cd	0.014	a	0.015	c	a/c/b
		RF	0.741	a	0.893	a	0.896	a	b/a/a	0.025	b	0.012	c	0.011	d	a/b/c
		AB	0.745	a	0.896	a	0.905	a	b/a/a	0.024	bc	0.011	d	0.010	d	a/b/c
		MLP	0.645	b	0.852	c	0.783	c	c/a/b	0.033	a	0.014	a	0.024	a	a/b/c
	8	**RKHS**	0.673	ab	0.834	c	0.860	bc	c/b/a	0.028	b	0.016	c	0.014	cd	a/b/c
		BRR	0.683	ab	0.834	c	0.821	d	c/a/b	0.028	b	0.016	c	0.017	b	a/c/b
		SVM	0.672	b	0.817	d	0.847	c	c/b/a	0.027	b	0.018	b	0.015	c	a/c/b
		RF	0.696	a	0.858	b	0.872	ab	c/b/a	0.026	c	0.014	d	0.013	de	a/b/c
		AB	0.666	b	0.879	a	0.882	a	b/a/a	0.027	b	0.012	e	0.012	e	a/b/b
		MLP	0.533	c	0.788	e	0.735	e	c/a/b	0.039	a	0.019	a	0.032	a	a/b/b
	9	**RKHS**	0.654	a	0.811	c	0.813	b	b/a/a	0.028	e	0.018	c	0.017	d	a/b/c
		BRR	0.628	b	0.823	bc	0.764	d	c/a/b	0.030	c	0.016	de	0.021	b	a/c/b
		SVM	0.649	ab	0.790	d	0.785	c	b/a/a	0.029	de	0.020	a	0.019	c	a/c/b

		RF	0.631	ab	0.830	ab	0.843	a	b/a/a	0.029	cd	0.016	d	0.014	e	a/b/c
		AB	0.577	c	0.838	a	0.839	a	b/a/a	0.033	b	0.015	e	0.015	e	a/b/b
		MLP	0.534	d	0.787	d	0.732	e	c/a/b	0.039	a	0.019	b	0.032	a	a/b/b
	Total	**RKHS**	0.789	a	0.887	b	0.886	b	b/a/a	0.022	d	0.014	c	0.013	d	a/b/b
		BRR	0.769	ab	0.873	c	0.857	d	c/a/b	0.024	c	0.014	c	0.017	b	a/c/b
		SVM	0.786	a	0.878	bc	0.869	c	b/a/a	0.023	d	0.015	b	0.015	c	a/c/b
		RF	0.772	ab	0.927	a	0.932	a	b/a/a	0.025	c	0.010	d	0.008	e	a/b/c
		AB	0.752	b	0.935	a	0.939	a	b/a/a	0.026	b	0.008	e	0.007	f	a/b/c
		MLP	0.677	c	0.851	d	0.855	d	b/a/a	0.034	a	0.016	a	0.018	a	a/b/c

Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing models**	**Predictive ability FI-1**	**Tukey's test comparing models**	**Predictive ability FI-2**	**Tukey's test comparing models**	**Tukey's test comparing data sets correlation (CD/FI-1/FI-2)**	**MSE CD**	**Tukey's test comparing models**	**MSE FI-1**	**Tukey's test comparing models**	**MSE FI-2**	**Tukey's test comparing models**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**
Leaf Dry Matter	**2**	**RKHS**	0.717	a	0.864	a	0.835	b	c/a/b	0.027	d	0.016	d	0.019	c	a/c/b
		BRR	0.694	ab	0.870	a	0.743	c	c/a/b	0.029	c	0.015	e	0.029	b	a/b/a

		SVM	0.717	a	0.838	b	0.823	b	b/a/a	0.027	cd	0.019	a	0.021	c	a/b/a
		RF	0.670	bc	0.863	a	0.875	a	b/a/a	0.032	b	0.017	c	0.014	d	a/b/c
		AB	0.649	c	0.876	a	0.879	a	b/a/a	0.033	b	0.014	e	0.014	d	a/b/b
		MLP	0.600	d	0.839	b	0.750	c	c/a/b	0.039	a	0.018	b	0.040	a	a/b/b
	5	**RKHS**	0.863	a	0.896	b	0.897	b	b/a/a	0.015	d	0.013	d	0.012	d	a/b/c
		BRR	0.844	b	0.866	c	0.840	d	b/a/b	0.016	c	0.015	c	0.017	b	b/c/a
		SVM	0.849	ab	0.900	b	0.882	c	c/a/b	0.018	b	0.016	b	0.015	c	b/c/a
		RF	0.858	ab	0.937	a	0.924	a	c/a/b	0.016	c	0.009	e	0.011	d	a/c/b
		AB	0.846	ab	0.931	a	0.909	b	c/a/b	0.016	cd	0.010	e	0.015	c	a/b/a
		MLP	0.749	c	0.777	d	0.823	e	c/b/a	0.026	a	0.023	a	0.022	a	a/b/a
	Total	**RKHS**	0.818	a	0.895	b	0.900	c	b/a/a	0.020	c	0.013	d	0.011	c	a/b/c
		BRR	0.805	a	0.874	c	0.842	e	c/a/b	0.021	bc	0.014	c	0.018	a	a/c/b
		SVM	0.813	a	0.874	c	0.861	d	b/a/a	0.022	b	0.015	b	0.016	b	a/c/b
		RF	0.812	a	0.917	a	0.952	a	c/b/a	0.020	c	0.011	e	0.007	e	a/b/c
		AB	0.804	a	0.916	a	0.929	b	c/b/a	0.021	bc	0.010	e	0.008	d	a/b/c
		MLP	0.695	b	0.839	d	0.848	de	b/a/a	0.032	a	0.018	a	0.018	a	a/b/c

Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test**	**Predictive**	**Tukey's test compari**	**Predictive**	**Tukey's test compari**	**Tukey's test compari**	**MSE CD**	**Tukey's test compar**	**MSE FI-1**	**Tukey's test**	**MSE FI-2**	**Tukey's test**	**Tukey's test comparing**

				comparing models	ability FI-1	ng models	ability FI-2	ng models	ng data sets correlation (CD/FI-1/FI-2)		ing models		comparing models		comparing models	data sets MSE (CD/FI-1/FI-2)
Stem Dry Matter	2	RKHS	0.490	a	0.814	a	0.664	b	c/a/b	0.045	c	0.023	b	0.033	c	a/c/b
		BRR	0.477	a	0.795	b	0.652	b	c/a/b	0.046	c	0.023	b	0.036	b	a/c/b
		SVM	0.491	a	0.794	b	0.643	b	c/a/b	0.045	c	0.025	a	0.036	bc	a/c/b
		RF	0.472	a	0.773	cd	0.818	a	c/b/a	0.046	c	0.026	a	0.021	d	a/b/c
		AB	0.429	b	0.760	d	0.795	a	c/b/a	0.050	b	0.026	a	0.022	d	a/b/c
		MLP	0.386	c	0.782	bc	0.536	c	c/a/b	0.058	a	0.023	b	0.062	a	a/b/c
	5	RKHS	0.809	a	0.920	ab	0.904	b	c/a/b	0.020	c	0.013	c	0.011	c	a/b/c
		BRR	0.772	b	0.881	c	0.848	d	c/a/b	0.021	b	0.015	b	0.016	a	a/c/b
		SVM	0.798	ab	0.909	b	0.912	ab	b/a/a	0.022	b	0.017	a	0.013	b	a/c/b
		RF	0.791	ab	0.922	a	0.921	a	b/a/a	0.021	b	0.012	c	0.012	b	a/b/b
		AB	0.778	b	0.910	b	0.908	ab	b/a/a	0.022	b	0.015	b	0.015	a	a/b/b
		MLP	0.688	c	0.854	d	0.864	c	b/a/a	0.030	a	0.017	a	0.016	a	a/b/b
	Total	RKHS	0.498	ab	0.716	b	0.694	b	c/a/b	0.036	b	0.025	bc	0.025	d	a/b/b
		BRR	0.491	ab	0.712	b	0.612	d	c/a/b	0.037	b	0.024	c	0.030	b	a/c/b
		SVM	0.492	ab	0.707	b	0.667	c	c/a/b	0.036	b	0.026	b	0.027	c	a/c/b

		RF	0.507	a	0.773	a	0.835	a	c/b/a	0.035	b	0.020	d	0.015	e	a/b/c
		AB	0.473	b	0.772	a	0.837	a	c/b/a	0.037	b	0.019	d	0.015	e	a/b/c
		MLP	0.341	c	0.642	c	0.667	c	b/a/a	0.049	a	0.030	a	0.032	a	a/b/c

Supplementary Table S3. Modeling results for all models and phenotype clippings organized to show de Tukey's test results comparing the clippings of a phenotype for each model.

Trait	Clipping	Model	Predictive ability CD	Tukey's test comparing clippings	Predictive ability FI-1	Tukey's test comparing clippings	Predictive ability FI-2	Tukey's test comparing clippings	Tukey's test comparing data sets correlation (CD/FI-1/FI-2)	MSE CD	Tukey's test comparing clippings	MSE FI-1	Tukey's test comparing clippings	MSE FI-2	Tukey's test comparing clippings	Tukey's test comparing data sets MSE (CD/FI-1/FI-2)
Green Matter	2	RKHS	0.653	f	0.837	e	0.857	de	b/a/a	0.029	b	0.017	b	0.018	bc	a/c/b
	3		0.743	e	0.848	e	0.870	cd	c/b/a	0.030	a	0.021	a	0.016	d	a/b/c
	4		0.847	ab	0.906	b	0.894	ab	c/a/b	0.020	e	0.014	d	0.015	e	a/c/b
	5		0.843	ab	0.909	b	0.866	d	c/a/b	0.018	f	0.012	e	0.016	d	a/c/b
	6		0.826	bc	0.885	cd	0.863	d	c/a/b	0.020	e	0.015	c	0.018	bc	a/c/b
	7		0.795	d	0.889	c	0.844	e	c/a/b	0.023	d	0.014	d	0.019	b	a/c/b
	8		0.858	a	0.927	a	0.881	bc	c/a/b	0.025	c	0.014	d	0.020	a	a/c/b
	9		0.799	d	0.877	d	0.897	a	c/b/a	0.023	d	0.016	c	0.012	f	a/b/c
	Total		0.807	cd	0.879	cd	0.867	cd	b/a/a	0.025	c	0.017	b	0.018	c	a/c/b
Dry Matter	2		0.601	e	0.851	e	0.537	f	b/a/c	0.033	a	0.018	a	0.044	a	b/c/a
	3		0.754	d	0.851	e	0.802	e	c/a/b	0.015	f	0.011	e	0.014	ef	a/c/b

	4		0.869	a	0.930	a	0.935	a	b/a/a	0.019	d	0.012	e	0.010	g	a/b/c
	5		0.841	b	0.913	b	0.882	c	c/a/b	0.017	e	0.011	e	0.014	e	a/c/b
	6		0.829	b	0.900	c	0.884	c	c/a/b	0.021	c	0.014	c	0.016	d	a/c/b
	7		0.794	c	0.875	d	0.849	d	c/a/b	0.021	c	0.015	b	0.017	c	a/c/b
	8		0.875	a	0.937	a	0.911	b	c/a/b	0.021	c	0.012	e	0.016	d	a/c/b
	9		0.792	c	0.885	d	0.787	e	b/a/b	0.023	b	0.015	b	0.022	b	a/c/b
	Total		0.835	b	0.916	b	0.912	b	b/a/a	0.021	c	0.013	d	0.013	f	a/b/b
Regrowth	**2**		0.639	f	0.851	e	0.868	c	c/b/a	0.037	b	0.019	b	0.015	cd	a/b/c
	3		0.873	a	0.946	a	0.940	a	c/a/b	0.023	e	0.011	g	0.011	f	a/b/b
	4		0.796	b	0.907	b	0.859	c	c/a/b	0.023	ef	0.012	f	0.017	b	a/c/b
	5		0.706	d	0.875	d	0.826	de	c/a/b	0.026	d	0.013	f	0.016	c	a/c/b
	6		0.696	d	0.834	f	0.833	d	b/a/a	0.039	a	0.024	a	0.024	a	a/b/b
	7		0.753	c	0.889	c	0.862	c	c/a/b	0.023	e	0.012	f	0.015	d	a/c/b
	8		0.673	e	0.834	f	0.860	c	c/b/a	0.028	c	0.016	d	0.014	e	a/b/c
	9		0.654	ef	0.811	g	0.813	e	b/a/a	0.028	c	0.018	c	0.017	b	a/b/c
	Total		0.789	b	0.887	c	0.886	b	b/a/a	0.022	f	0.014	e	0.013	e	a/b/b
Leaf Dry Matter	**2**		0.717	c	0.864	b	0.835	b	c/a/b	0.027	a	0.016	a	0.019	a	a/c/b
	5		0.863	a	0.896	a	0.897	a	b/a/a	0.015	c	0.013	b	0.012	b	a/b/c
	Total		0.818	b	0.895	a	0.900	a	b/a/a	0.020	b	0.013	b	0.011	c	a/b/c

Stem Dry Matter	**2**		0.490	b	0.814	b	0.664	c	c/a/b	0.045	a	0.025	a	0.033	a	a/c/b
	5		0.809	a	0.920	a	0.904	a	c/a/b	0.020	c	0.023	b	0.011	c	a/b/c
	Total		0.498	b	0.716	c	0.694	b	c/a/b	0.036	b	0.013	c	0.025	b	a/b/b
Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing clippings**	**Predictive ability FI-1**	**Tukey's test comparing clippings**	**Predictive ability FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets correlation (CD/FI-1/FI-2)**	**MSE CD**	**Tukey's test comparing clippings**	**MSE FI-1**	**Tukey's test comparing clippings**	**MSE FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**
Green Matter	**2**	**BRR**	0.625	d	0.822	f	0.837	c	b/a/a	0.030	b	0.018	c	0.018	e	a/b/c
	3		0.714	c	0.812	f	0.788	e	c/a/b	0.033	a	0.023	a	0.027	b	a/c/b
	4		0.828	a	0.874	bc	0.855	ab	c/a/b	0.022	f	0.015	g	0.020	d	a/c/b
	5		0.823	a	0.877	b	0.854	ab	c/a/b	0.019	g	0.017	ef	0.019	d	a/b/a
	6		0.798	b	0.858	d	0.845	bc	c/a/b	0.023	e	0.017	d	0.019	d	a/c/b
	7		0.783	b	0.853	de	0.809	d	c/a/b	0.024	d	0.017	de	0.022	c	a/c/b
	8		0.840	a	0.902	a	0.818	d	b/a/c	0.027	c	0.017	de	0.030	a	b/c/a
	9		0.792	b	0.862	cd	0.866	a	b/a/a	0.024	d	0.016	f	0.016	f	a/b/b

	Total		0.793	b	0.840	e	0.852	ab	c/b/a	0.026	c	0.021	b	0.019	d	a/b/c
Dry Matter	**2**		0.580	e	0.866	c	0.606	g	c/a/b	0.034	a	0.015	d	0.039	a	b/c/a
	3		0.732	d	0.801	e	0.799	e	b/a/a	0.015	h	0.014	f	0.014	f	a/b/b
	4		0.845	a	0.898	a	0.892	ab	b/a/a	0.021	f	0.014	e	0.016	d	a/c/b
	5		0.816	b	0.881	b	0.878	b	b/a/a	0.019	g	0.014	f	0.015	e	a/c/b
	6		0.812	b	0.881	b	0.856	c	c/a/b	0.023	cd	0.016	bc	0.019	c	a/c/b
	7		0.782	c	0.821	d	0.838	d	c/b/a	0.022	e	0.019	a	0.018	c	a/b/c
	8		0.855	a	0.908	a	0.907	a	b/a/a	0.023	bc	0.015	cd	0.016	d	a/c/b
	9		0.782	c	0.861	c	0.752	f	b/a/c	0.024	b	0.016	b	0.027	b	b/c/a
	Total		0.820	b	0.885	b	0.895	a	b/a/a	0.022	de	0.016	bc	0.015	e	a/b/c
Regrowth	**2**		0.594	f	0.843	e	0.821	e	c/a/b	0.040	a	0.018	b	0.021	b	a/c/b
	3		0.878	a	0.932	a	0.905	a	c/a/b	0.022	g	0.013	e	0.017	e	a/c/b
	4		0.800	b	0.899	b	0.842	cd	c/a/b	0.022	fg	0.011	f	0.019	d	a/c/b
	5		0.679	d	0.859	d	0.787	f	c/a/b	0.027	d	0.014	d	0.020	cd	a/c/b
	6		0.699	d	0.832	ef	0.831	de	b/a/a	0.038	b	0.023	a	0.023	a	a/b/b
	7		0.753	c	0.888	b	0.848	bc	c/a/b	0.022	f	0.011	f	0.016	g	a/c/b
	8		0.683	d	0.834	ef	0.821	e	c/a/b	0.027	d	0.016	c	0.017	f	a/c/b
	9		0.628	e	0.823	f	0.764	g	c/a/b	0.030	c	0.016	c	0.021	bc	a/c/b
	Total		0.769	c	0.873	c	0.857	b	c/a/b	0.024	e	0.014	d	0.017	fg	a/c/b

Leaf Dry Matter	**2**		0.694	c	0.870	a	0.743	b	c/a/b	0.029	a	0.015	ab	0.029	a	a/b/a
	5		0.844	a	0.866	a	0.840	a	b/a/b	0.016	c	0.015	a	0.017	b	b/c/a
	Total		0.805	b	0.874	a	0.842	a	c/a/b	0.021	b	0.014	b	0.018	b	a/c/b
Stem Dry Matter	**2**		0.477	b	0.795	b	0.652	b	c/a/b	0.046	a	0.023	b	0.036	a	a/c/b
	5		0.772	a	0.881	a	0.848	a	c/a/b	0.021	c	0.015	c	0.016	c	a/c/b
	Total		0.491	b	0.712	c	0.612	c	c/a/b	0.037	b	0.024	a	0.030	b	a/c/b
Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing clippings**	**Predictive ability FI-1**	**Tukey's test comparing clippings**	**Predictive ability FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets correlation (CD/FI-1/FI-2)**	**MSE CD**	**Tukey's test comparing clippings**	**MSE FI-1**	**Tukey's test comparing clippings**	**MSE FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**
Green Matter	**2**	**SVM**	0.668	f	0.840	d	0.829	de	b/a/a	0.029	b	0.019	c	0.022	a	a/b/c
	3		0.749	e	0.835	d	0.850	c	c/b/a	0.030	a	0.022	a	0.019	c	a/c/b
	4		0.839	ab	0.901	b	0.895	a	b/a/a	0.022	e	0.017	e	0.015	e	a/c/b
	5		0.833	b	0.902	b	0.843	cd	b/a/b	0.020	f	0.015	f	0.018	d	a/b/a
	6		0.824	bc	0.878	c	0.843	cd	c/a/b	0.022	e	0.017	d	0.022	ab	a/c/b

	7		0.793	d	0.879	c	0.822	e	c/a/b	0.024	d	0.017	e	0.021	ab	a/c/b
	8		0.857	a	0.921	a	0.873	b	c/a/b	0.026	c	0.017	e	0.021	b	b/c/a
	9		0.799	d	0.868	c	0.888	ab	c/b/a	0.024	d	0.018	d	0.014	e	a/b/b
	Total		0.806	cd	0.872	c	0.843	cd	c/a/b	0.026	c	0.019	b	0.021	b	a/b/c
Dry Matter	**2**		0.604	e	0.843	e	0.447	g	b/a/c	0.033	a	0.020	a	0.049	a	b/c/a
	3		0.750	d	0.814	f	0.808	e	b/a/a	0.017	g	0.016	d	0.015	f	a/b/b
	4		0.865	a	0.915	ab	0.927	a	c/b/a	0.020	e	0.014	e	0.011	g	a/c/b
	5		0.831	b	0.905	bc	0.847	d	c/a/b	0.019	f	0.013	f	0.018	d	a/c/b
	6		0.828	b	0.897	c	0.871	c	c/a/b	0.023	cd	0.015	d	0.019	cd	a/c/b
	7		0.795	c	0.864	d	0.854	cd	b/a/a	0.022	d	0.017	c	0.017	e	a/b/c
	8		0.871	a	0.927	a	0.897	b	c/a/b	0.022	cd	0.014	e	0.020	c	a/c/b
	9		0.788	c	0.875	d	0.776	f	b/a/b	0.025	b	0.018	b	0.023	b	b/c/a
	Total		0.833	b	0.904	bc	0.896	b	b/a/a	0.023	c	0.015	d	0.016	e	a/b/c
Regrowth	**2**		0.648	f	0.843	e	0.828	d	b/a/a	0.036	b	0.021	b	0.021	b	a/c/b
	3		0.870	a	0.930	a	0.934	a	b/a/a	0.025	e	0.016	e	0.015	ef	a/c/b
	4		0.794	b	0.899	b	0.850	c	c/a/b	0.023	f	0.014	f	0.018	d	a/c/b
	5		0.697	d	0.859	d	0.827	d	c/a/b	0.026	d	0.015	e	0.016	e	a/c/b
	6		0.699	d	0.822	f	0.825	d	b/a/a	0.037	a	0.024	a	0.024	a	a/b/b
	7		0.750	c	0.876	c	0.857	bc	c/a/b	0.023	f	0.014	f	0.015	f	a/c/b
	8		0.672	e	0.817	f	0.847	c	c/b/a	0.028	c	0.018	d	0.015	ef	a/c/b

	9		0.649	f	0.790	g	0.785	e	b/a/a	0.029	c	0.020	c	0.019	c	a/c/b
	Total		0.786	b	0.878	c	0.869	b	b/a/a	0.023	f	0.015	e	0.015	ef	a/c/b
Leaf Dry Matter	**2**		0.717	c	0.838	c	0.823	c	b/a/a	0.027	a	0.019	a	0.021	a	a/b/a
	5		0.849	a	0.900	a	0.882	a	c/a/b	0.018	c	0.016	b	0.015	c	b/c/a
	Total		0.813	b	0.873	b	0.861	b	b/a/a	0.022	b	0.015	b	0.016	b	a/c/b
Stem Dry Matter	**2**		0.491	b	0.794	b	0.643	c	c/a/b	0.045	a	0.025	a	0.036	a	a/c/b
	5		0.798	a	0.909	a	0.912	a	b/a/a	0.022	c	0.017	b	0.013	c	a/c/b
	Total		0.492	b	0.707	c	0.667	b	c/a/b	0.036	b	0.026	a	0.027	b	a/c/b

Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing clippings**	**Predictive ability FI-1**	**Tukey's test comparing clippings**	**Predictive ability FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets correlation (CD/FI-1/FI-2)**	**MSE CD**	**Tukey's test comparing clippings**	**MSE FI-1**	**Tukey's test comparing clippings**	**MSE FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**
Green Matter	**2**	**RF**	0.606	e	0.841	d	0.850	e	b/a/a	0.033	b	0.019	a	0.018	a	a/b/c
	3		0.695	d	0.858	c	0.869	d	b/a/a	0.036	a	0.019	a	0.017	b	a/b/c
	4		0.829	ab	0.923	ab	0.912	bc	c/a/b	0.023	d	0.011	d	0.013	d	a/c/b

	5		0.839	ab	0.931	a	0.917	b	c/a/b	0.019	f	0.010	e	0.012	e	a/c/b
	6		0.816	bc	0.926	ab	0.903	c	c/a/b	0.022	e	0.010	e	0.012	e	a/c/b
	7		0.795	c	0.921	b	0.923	b	b/a/a	0.023	d	0.011	d	0.010	f	a/b/b
	8		0.848	a	0.932	a	0.913	bc	c/a/b	0.027	c	0.013	b	0.015	c	a/c/b
	9		0.802	c	0.920	b	0.944	a	c/b/a	0.023	d	0.011	d	0.008	h	a/b/c
	Total		0.795	c	0.919	b	0.937	a	c/b/a	0.027	c	0.012	c	0.009	g	a/b/c
Dry Matter	**2**		0.547	f	0.795	e	0.638	e	c/a/b	0.037	a	0.023	a	0.035	a	a/c/b
	3		0.731	e	0.856	d	0.840	d	c/a/b	0.016	f	0.012	b	0.013	d	a/c/b
	4		0.847	ab	0.932	b	0.941	a	c/b/a	0.023	c	0.010	d	0.009	f	a/b/c
	5		0.836	abc	0.933	b	0.895	c	c/a/b	0.018	e	0.009	e	0.015	c	a/c/b
	6		0.819	c	0.928	b	0.917	b	b/a/a	0.023	c	0.010	d	0.011	e	a/b/b
	7		0.789	d	0.915	c	0.913	b	b/a/a	0.022	d	0.011	c	0.011	e	a/b/b
	8		0.860	a	0.950	a	0.915	b	c/a/b	0.025	b	0.009	e	0.016	b	a/c/b
	9		0.791	d	0.913	c	0.925	b	c/b/a	0.023	c	0.012	b	0.011	e	a/b/c
	Total		0.830	bc	0.934	b	0.943	a	b/a/a	0.022	cd	0.010	d	0.009	f	a/b/c
Regrowth	**2**		0.591	f	0.857	e	0.896	d	c/b/a	0.041	a	0.020	a	0.014	b	a/b/c
	3		0.861	a	0.946	a	0.956	a	c/b/a	0.026	c	0.010	f	0.008	e	a/b/c
	4		0.782	b	0.918	b	0.911	c	b/a/a	0.026	c	0.011	e	0.012	c	a/c/b
	5		0.691	d	0.870	d	0.869	f	b/a/a	0.026	c	0.013	d	0.013	c	a/b/b
	6		0.680	d	0.862	de	0.884	e	c/b/a	0.040	a	0.019	a	0.016	a	a/b/c

	7		0.741	c	0.893	c	0.896	d	b/a/a	0.025	d	0.012	e	0.011	d	a/b/c
	8		0.696	d	0.858	de	0.872	f	c/b/a	0.026	c	0.014	c	0.013	c	a/b/c
	9		0.631	e	0.830	f	0.843	g	b/a/a	0.029	b	0.016	b	0.014	b	a/b/c
	Total		0.772	b	0.927	b	0.932	b	b/a/a	0.025	d	0.010	f	0.008	e	a/b/c
Leaf Dry Matter	**2**		0.670	c	0.863	c	0.875	c	b/a/a	0.032	a	0.017	a	0.014	a	a/b/c
	5		0.858	a	0.937	a	0.924	b	c/a/b	0.016	c	0.009	c	0.011	b	a/c/b
	Total		0.812	b	0.917	b	0.952	a	c/b/a	0.020	b	0.011	b	0.007	c	a/b/c
Stem Dry Matter	**2**		0.472	c	0.773	b	0.818	c	c/b/a	0.046	a	0.026	a	0.021	a	a/b/c
	5		0.791	a	0.922	a	0.921	a	b/a/a	0.021	c	0.012	c	0.012	c	a/b/b
	Total		0.507	b	0.773	b	0.835	b	c/b/a	0.035	b	0.020	b	0.015	b	a/b/c

Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing clippings**	**Predictive ability FI-1**	**Tukey's test comparing clippings**	**Predictive ability FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets correlation (CD/FI-1/FI-2)**	**MSE CD**	**Tukey's test comparing clippings**	**MSE FI-1**	**Tukey's test comparing clippings**	**MSE FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**
Green Matter	**2**	**AB**	0.551	f	0.805	e	0.844	f	c/b/a	0.036	b	0.020	a	0.017	a	a/b/c

	3		0.656	e	0.865	d	0.877	e	b/a/a	0.039	a	0.018	b	0.016	a	a/b/c
	4		0.819	abc	0.930	a	0.902	cd	c/a/b	0.023	ef	0.010	f	0.014	bc	a/c/b
	5		0.827	ab	0.927	ab	0.892	de	c/a/b	0.019	g	0.011	ef	0.015	b	a/c/b
	6		0.807	bcd	0.922	abc	0.888	de	c/a/b	0.022	f	0.010	f	0.013	cd	a/c/b
	7		0.793	d	0.911	c	0.909	bc	b/a/a	0.022	f	0.011	de	0.012	de	a/c/b
	8		0.839	a	0.932	a	0.917	ab	c/a/b	0.028	c	0.012	c	0.014	bc	a/c/b
	9		0.790	d	0.912	c	0.931	a	c/b/a	0.024	e	0.011	de	0.009	f	a/b/c
	Total		0.796	cd	0.915	bc	0.920	ab	b/a/a	0.026	d	0.012	cd	0.012	e	a/b/b
Dry Matter	**2**		0.518	f	0.789	f	0.698	g	c/a/b	0.039	a	0.022	a	0.029	a	a/c/b
	3		0.715	e	0.863	e	0.808	f	c/a/b	0.015	g	0.011	c	0.019	c	b/c/a
	4		0.833	b	0.932	bc	0.935	a	b/a/a	0.023	bc	0.010	d	0.009	e	a/b/b
	5		0.826	b	0.934	ab	0.868	e	c/a/b	0.018	f	0.009	e	0.020	bc	b/c/a
	6		0.813	bc	0.925	bc	0.916	bc	b/a/a	0.022	cd	0.010	cd	0.011	de	a/c/b
	7		0.788	cd	0.920	c	0.900	cd	c/a/b	0.022	de	0.010	cd	0.012	d	a/c/b
	8		0.865	a	0.946	a	0.895	d	c/a/b	0.023	bc	0.009	e	0.021	b	a/c/b
	9		0.782	d	0.896	d	0.935	a	c/b/a	0.024	b	0.013	b	0.009	e	a/b/c
	Total		0.825	b	0.934	ab	0.931	ab	b/a/a	0.021	e	0.009	de	0.010	e	a/c/b
Regrowth	**2**		0.556	f	0.885	cde	0.896	cd	b/a/a	0.043	b	0.015	b	0.013	c	a/b/c
	3		0.845	a	0.950	a	0.957	a	b/a/a	0.026	de	0.009	ef	0.008	e	a/b/c
	4		0.771	b	0.925	b	0.900	cd	c/a/b	0.026	e	0.010	e	0.014	b	a/c/b

	5		0.677	d	0.876	e	0.870	f	b/a/a	0.027	de	0.012	c	0.012	c	a/b/b
	6		0.647	e	0.889	cd	0.889	de	b/a/a	0.045	a	0.016	a	0.015	a	a/b/b
	7		0.745	c	0.896	c	0.905	c	b/a/a	0.024	f	0.011	d	0.010	d	a/b/c
	8		0.666	de	0.879	de	0.882	ef	b/a/a	0.027	d	0.012	c	0.012	c	a/b/b
	9		0.577	f	0.838	f	0.839	g	b/a/a	0.033	c	0.015	ab	0.015	ab	a/b/b
	Total		0.752	bc	0.935	b	0.939	b	b/a/a	0.026	de	0.008	f	0.007	e	a/b/c
Leaf Dry Matter	**2**		0.649	c	0.876	c	0.879	c	b/a/a	0.033	a	0.014	a	0.014	b	a/b/b
	5		0.846	a	0.931	a	0.909	b	c/a/b	0.016	c	0.010	b	0.015	a	a/b/a
	Total		0.804	b	0.916	b	0.929	a	c/b/a	0.021	b	0.010	b	0.008	c	a/b/c
Stem Dry Matter	**2**		0.429	c	0.760	b	0.795	c	c/b/a	0.050	a	0.026	a	0.022	a	a/b/c
	5		0.778	a	0.910	a	0.908	a	b/a/a	0.022	c	0.015	c	0.015	b	a/b/b
	Total		0.473	b	0.772	b	0.837	b	c/b/a	0.037	b	0.019	b	0.015	b	a/b/c

Trait	**Clipping**	**Model**	**Predictive ability CD**	**Tukey's test comparing clippings**	**Predictive ability FI-1**	**Tukey's test comparing clippings**	**Predictive ability FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets correlation**	**MSE CD**	**Tukey's test comparing clippings**	**MSE FI-1**	**Tukey's test comparing clippings**	**MSE FI-2**	**Tukey's test comparing clippings**	**Tukey's test comparing data sets MSE (CD/FI-1/FI-2)**

									(CD/FI-1/ FI-2)							
Green Matter	2		0.541	g	0.738	e	0.845	ab	c/b/a	0.041	b	0.019	c	0.019	e	a/b/c
	3		0.630	f	0.794	d	0.809	d	b/a/a	0.044	a	0.022	a	0.029	cd	a/b/c
	4		0.739	b	0.863	b	0.829	bcd	c/a/b	0.032	de	0.017	e	0.026	d	a/c/b
	5		0.735	bc	0.854	bc	0.734	f	b/a/b	0.028	f	0.015	f	0.069	a	a/c/b
	6		0.711	bcd	0.796	d	0.813	cd	b/a/a	0.032	e	0.017	d	0.024	de	a/c/b
	7		0.670	e	0.841	c	0.732	f	c/a/b	0.035	c	0.017	e	0.034	bc	a/c/b
	8		0.778	a	0.911	a	0.858	a	c/a/b	0.036	c	0.017	e	0.024	de	a/c/b
	9		0.682	de	0.848	bc	0.836	abc	b/a/a	0.035	cd	0.018	d	0.023	de	a/b/c
	Total		0.704	cd	0.809	d	0.769	e	c/a/b	0.037	c	0.019	b	0.036	b	a/b/b
Dry Matter	2	MLP	0.505	f	0.810	d	0.380	f	b/a/c	0.045	a	0.020	a	0.091	a	a/c/b
	3		0.566	e	0.702	e	0.692	d	b/a/a	0.026	d	0.016	d	0.048	c	b/c/a
	4		0.773	ab	0.884	b	0.894	a	b/a/a	0.030	c	0.014	e	0.017	d	a/b/b
	5		0.722	c	0.842	c	0.686	d	b/a/c	0.028	d	0.013	f	0.049	c	b/c/a
	6		0.732	c	0.871	b	0.846	b	c/a/b	0.032	bc	0.015	d	0.022	d	a/c/b
	7		0.661	d	0.796	d	0.712	d	c/a/b	0.034	b	0.017	c	0.047	c	a/c/b
	8		0.790	a	0.904	a	0.799	c	b/a/b	0.032	bc	0.014	e	0.051	c	a/c/b
	9		0.671	d	0.835	c	0.582	e	b/a/c	0.035	b	0.018	b	0.063	b	a/b/c
	Total		0.740	bc	0.876	b	0.879	ab	b/a/a	0.033	bc	0.015	d	0.026	d	a/c/b

Regrowth	**2**		0.493	f	0.808	d	0.819	b	b/a/a	0.050	a	0.021	b	0.025	de	a/b/c
	3		0.820	a	0.923	a	0.867	a	c/a/b	0.032	d	0.016	e	0.027	cd	a/b/c
	4		0.693	b	0.872	b	0.752	d	c/a/b	0.034	cd	0.014	f	0.040	a	a/c/b
	5		0.578	d	0.835	c	0.657	e	c/a/b	0.036	c	0.015	e	0.040	a	a/b/b
	6		0.592	d	0.788	e	0.812	b	c/b/a	0.051	a	0.024	a	0.028	c	a/b/b
	7		0.645	c	0.852	c	0.783	c	c/a/b	0.033	cd	0.014	f	0.024	e	a/b/c
	8		0.533	e	0.788	e	0.735	d	c/a/b	0.039	b	0.018	d	0.032	b	a/b/b
	9		0.534	e	0.787	e	0.732	d	c/a/b	0.039	b	0.020	c	0.032	b	a/b/b
	Total		0.677	bc	0.851	c	0.855	a	b/a/a	0.034	cd	0.015	e	0.018	f	a/b/c
Leaf Dry Matter	**2**		0.600	c	0.839	a	0.750	c	c/a/b	0.039	a	0.019	a	0.040	a	a/b/b
	5		0.749	a	0.777	b	0.823	b	c/b/a	0.026	c	0.016	b	0.022	b	a/b/a
	Total		0.695	b	0.839	a	0.848	a	b/a/a	0.032	b	0.015	b	0.018	c	a/b/c
Stem Dry Matter	**2**		0.386	b	0.782	b	0.536	c	c/a/b	0.058	a	0.025	a	0.062	a	a/b/c
	5		0.688	a	0.854	a	0.864	a	b/a/a	0.030	c	0.017	b	0.016	c	a/b/b
	Total		0.341	c	0.642	c	0.667	b	b/a/a	0.049	b	0.026	a	0.032	b	a/b/c

Supplementary Table S4. Number of scenarios that the model was considered as the best scenario by the Tukey's test. The total number of scenarios is the number of phenotype clippings (33).

	Predictive ability			MSE		
	CD	FI-1	FI-2	CD	FI-1	FI-2
RKHS	33	11	5	30	3	8
BRR	21	3	1	12	6	4
SVM	32	1	1	13	0	3
RF	25	28	32	10	16	26
AB	15	27	24	8	28	26
MLP	0	0	1	0	1	0

Supplementary Table S5. Number of markers (Feature Quantity) selected by the three Feature Selection (FS) methods and the Feature Intersections (FI) for all phenotype clippings.

Trait	Clipping	FS	Feature Quantity
Green Matter	2	FS-1	161
		FS-2	1,002
		FS-3	798
		FI-1	105
		FI-2	9
	3	FS-1	132
		FS-2	1,014
		FS-3	709
		FI-1	104
		FI-2	11
	4	FS-1	154
		FS-2	830
		FS-3	711
		FI-1	102
		FI-2	7
	5	FS-1	141
Dry Matter	2	FS-1	167
		FS-2	1,109
		FS-3	807
		FI-1	99
		FI-2	6
	3	FS-1	169
		FS-2	870
		FS-3	676
		FI-1	107
		FI-2	5
	4	FS-1	146
		FS-2	794
		FS-3	694
		FI-1	107
		FI-2	18
	5	FS-1	146
Regrowth	2	FS-1	156
		FS-2	1,126
		FS-3	837
		FI-1	122
		FI-2	16
	3	FS-1	146
		FS-2	571
		FS-3	563
		FI-1	107
		FI-2	18
	4	FS-1	142
		FS-2	822
		FS-3	658
		FI-1	107
		FI-2	6
	5	FS-1	156

		FS-2	656				FS-2	780				FS-2	1,112		
		FS-3	594				FS-3	617				FS-3	782		
		FI-1	104				FI-1	81				FI-1	96		
		FI-2	6				FI-2	7				FI-2	11		
	6	FS-1	148			6	FS-1	144			6	FS-1	150		
		FS-2	756				FS-2	754				FS-2	922		
		FS-3	711				FS-3	685				FS-3	686		
		FI-1	92				FI-1	94				FI-1	107		
		FI-2	12				FI-2	17				FI-2	13		
	7	FS-1	144			7	FS-1	154			7	FS-1	175		
		FS-2	746				FS-2	808				FS-2	922		
		FS-3	673				FS-3	713				FS-3	751		
		FI-1	91				FI-1	99				FI-1	111		
		FI-2	10				FI-2	9				FI-2	23		
	8	FS-1	136			8	FS-1	143			8	FS-1	156		
		FS-2	484				FS-2	636				FS-2	982		
		FS-3	564				FS-3	628				FS-3	721		
		FI-1	92				FI-1	88				FI-1	99		
		FI-2	5				FI-2	9				FI-2	15		
	9	FS-1	129			9	FS-1	151			9	FS-1	171		

		FS-2	792				**FS-2**	745				**FS-2**	1,039
		FS-3	712				**FS-3**	704				**FS-3**	853
		FI-1	102				**FI-1**	94				**FI-1**	114
		FI-2	14				**FI-2**	7				**FI-2**	11
	Total	**FS-1**	130			**Total**	**FS-1**	138			**Total**	**FS-1**	145
		FS-2	795				**FS-2**	723				**FS-2**	838
		FS-3	636				**FS-3**	601				**FS-3**	727
		FI-1	99				**FI-1**	99				**FI-1**	104
		FI-2	15				**FI-2**	14				**FI-2**	16

Trait	Clipping	FS	Feature Quantity
Leaf Dry Matter	2	FS-1	167
		FS-2	993
		FS-3	774
		FI-1	108
		FI-2	7
	5	FS-1	146
		FS-2	717
		FS-3	679
		FI-1	101
		FI-2	14
	Total	FS-1	137
		FS-2	766
		FS-3	573
		FI-1	106
		FI-2	12

Trait	Clipping	FS	Feature Quantity
Stem Dry Matter	2	FS-1	146
		FS-2	1,098
		FS-3	822
		FI-1	115
		FI-2	12
	5	FS-1	153
		FS-2	602
		FS-3	578
		FI-1	103
		FI-2	13
	Total	FS-1	163
		FS-2	1,170
		FS-3	826
		FI-1	118
		FI-2	12

Supplementary Table S6. Number of scenarios that the datasets Complete Data (CD), Feature Intersection 1 (FI-1) and Feature Intersection 2 (FI-2) were considered as the best scenario by Tukey's test. Considering 6 models and 33 phenotype clippings, there were a total of 198 scenarios.

	Predictive ability	MSE
CD	0	0
FI-1	168	136
FI-2	100	89

Supplementary Table S7. Clippings that were considered as the best scenarios in regard to model, phenotype and dataset. Models highlighted in red were the best models for the dataset.

	Predictive ability					MSE				
	CD									
Models	Green Matter	Dry Matter	Regrowth	Leaf Dry Matter	Stem Dry Matter	Green Matter	Dry Matter	Regrowth	Leaf Dry Matter	Stem Dry Matter
RKHS	4,5,8	4,8	3	5	5	5	3	4,T	5	5
BRR	4,5,8	4,8	3	5	5	5	5	3,4	5	5
SVM	4,8	4,8	3	5	5	5	2	4	5	5
RF	4,5,8	4,5,8	3	5	5	5	3	7,T	5	5
AB	4,5,8	8	3	5	5	5	3	7	5	5
MLP	8	8	3	5	5	5	3,5	3,4,7,T	5	5
	FI-1									
RKHS	8	4,8	3	5,T	5	5	3,4,5,8	3	5,T	T
BRR	8	4,8	3	2,5,T	5	4	3,5	4	2,T	5
SVM	8	8	3	5	5	5	5	4	5,T	5
RF	4,5,6,8	8	3	5	5	5,6	5,8	3	5	5
AB	4,5,6,8	5,8,T	3	5	5	4,5,6	5,8,T	3,T	5,T	5
MLP	8	8	3	2,T	5	5	5	4,7	5,t	5
	FI-2									
RKHS	4,8	4	3	5,T	5	9	4	3	T	5
BRR	4,5,9,T	4,8,T	3	5,T	5	9	3	7,T	5,T	5
SVM	4,9	4	3	5	5	4,9	4	3,7,8,T	5	5
RF	9,T	4,T	3	T	5	9	4,T	3,T	T	5

AB	8,9,T	4,9,T	3	T	5	9	4,6,9,T	3,T	T	5,T
MLP	2,8,9	4,T	3,T	T	5	2,6,8,9	4,6,T	T	T	5

Supplementary Table S8. Functional annotation of the transcripts physically linked to the Major Importance markers.

Chromosome	**Marker Position**	**Phenotype Clipping**	**Gini Importance**	**Transcript Start**	**Transcript End**	**Transcript ID**	**Enzime**	GO	Kegg
1	3602030	DM_2	0.248	3598955	3599293	DN66813_c0_g1	Transcription factor-like protein DPB	GO:0005737,GO:0005634,GO:0005667, GO:0003677,GO:0000981,GO:0046982, GO:0042023,GO:0000082,GO:0051726	KEGG:ath:AT5G03415-transmembrane protein
1	3602030	DM_2	0.248	3600810	3600408	DN72459_c0_g1	Transcription factor-like protein DPB	GO:0005737,GO:0005634,GO:0005667, GO:0003677,GO:0000981,GO:0046982, GO:0042023,GO:0000082,GO:0051726	KEGG:ath:AT5G03415
1	3602030	DM_2	0.248	3606241	3606958	DN85127_c1_g1	4-coumarate--CoA ligase-like 1	GO:0005777,GO:0005524,GO:0016405, GO:0004321	KEGG:osa:4331650
1	5918888	GM_2	0.087	5915196	5915769	DN89084_c0_g1	3-ketoacyl-CoA synthase 10 {ECO:0000303|PubMed:18465198}	GO:0005783,GO:0005789,GO:0016021, GO:0005777,GO:0102756,GO:0006633, GO:0009409	KEGG:ath:AT2G26250

1	5918888	GM_2	0.087	5915633	5916152	DN74840_c1_g1	3-ketoacyl-CoA synthase 10 {ECO:0000303|PubMed:18465198}	GO:0005783,GO:0005789,GO:0016021,GO:0005777,GO:0102756,GO:0006633,GO:0009409	KEGG:ath:AT2G26250
1	5918888	GM_2	0.087	5915730	5914800	DN89084_c1_g1	3-ketoacyl-CoA synthase 1 {ECO:0000303|PubMed:10074711};3-ketoacyl-CoA synthase 4 {ECO:0000303|PubMed:18465198};3-ketoacyl-CoA synthase 10 {ECO:0000303|PubMed:18465198}	GO:0022626,GO:0005783,GO:0005789,GO:0016021,GO:0016020,GO:0009922,GO:0102756,GO:0070417,GO:0009409,GO:0009416,GO:0042761,GO:0010025,GO:0006633	KEGG:ath:AT1G19440;KEGG:ath:AT1G01120;KEGG:ath:AT2G26250
1	5918888	GM_2	0.087	5916004	5915528	DN84880_c0_g1	3-ketoacyl-CoA synthase 9 {ECO:0000303|PubMed:18465198}	GO:0005789,GO:0016021,GO:0102756,GO:0006633,GO:0009409	KEGG:ath:AT2G16280
1	5918888	GM_2	0.087	5921068	5921587	DN87697_c1_g1	Protein WHAT'S THIS FACTOR 1 homolog, chloroplastic {ECO:0000305}	GO:0009507,GO:0003729,GO:0000373,GO:0006397,GO:0015979	KEGG:osa:4339664;KEGG:ath:AT4G01037
1	5918888	GM_2	0.087	5921878	5921618	DN92024_c3_g2			
1	5927892	LDM_2	0.120	5927725	5928559	DN65451_c0_g2	Protein TIFY 11b {ECO:0000305}	GO:0005634,GO:0031347,GO:2000022	KEGG:osa:4331834
1	5927892	LDM_2	0.120	5928559	5927998	DN65451_c0_g3	Protein TIFY 11b {ECO:0000305}	GO:0005634,GO:0031347,GO:2000022	KEGG:osa:4331834

1	6758756	RG_2	0.083	6757183	6756739	DN83885_c1_g2	Membrane-associated progesterone-binding protein 4 {ECO:0000303|Ref.7}	GO:0012505,GO:0016021,GO:0016020	KEGG:ath:AT4G14965
1	6758756	RG_2	0.083	6761535	6761812	DN92221_c2_g1			
1	6758756	RG_2	0.083	6761796	6761532	DN91118_c5_g1	Glutathione S-transferase	GO:0005737,GO:0004364,GO:0016034,GO:0006749	
1	6758756	RG_2	0.083	6763726	6764490	DN87101_c0_g1	BTB/POZ domain-containing protein NPY2;BTB/POZ domain-containing protein NPY5	GO:0071944,GO:0009958,GO:0016567	KEGG:ath:AT2G14820;KEGG:ath:AT4G37590
1	8916524	RG_9	0.095	8915411	8914400	DN73443_c2_g1	Subtilisin-like protease SBT1.6 {ECO:0000303|PubMed:16193095}	GO:0005886,GO:0004252,GO:0008236	KEGG:ath:AT4G34980
1	8916524	RG_9	0.095	8916073	8916336	DN61834_c4_g2			
1	8916524	RG_9	0.095	8916337	8916028	DN59818_c2_g4			
1	8916524	RG_9	0.095	8916342	8916037	DN92072_c5_g1	Altered inheritance rate of mitochondria protein 25	GO:0005739,GO:0005886,GO:0017128,GO:0034605,GO:0034599	KEGG:sce:YJR100C
1	8916524	RG_9	0.095	8918403	8919467	DN71029_c0_g1	Actin-depolymerizing factor 5	GO:0015629,GO:0005737,GO:0051015,GO:0051017	KEGG:osa:4332220
1	8916524	RG_9	0.095	8918681	8918117	DN87804_c1_g2	Actin-depolymerizing factor 5	GO:0015629,GO:0005737,GO:0051015,GO:0051017	KEGG:osa:4332220

1	11178365	DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	11174855	11175173	DN91018_c1_g3	Trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM3 {ECO:0000305|PubMed:17190829}	GO:0005829,GO:0009506,GO:0008460,GO:0016853,GO:0016491,GO:0050377,GO:0010280,GO:0010315,GO:0051555	KEGG:ath:AT3G14790
1	11178365	DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	11174865	11175260	DN68526_c0_g1	Trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM3 {ECO:0000305|PubMed:17190829};Bifunctional dTDP-4-dehydrorhamnose 3,5-epimerase/dTDP-4-dehydrorhamnose reductase {ECO:0000305};Trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM1 {ECO:0000305|PubMed:17190829}	GO:0005829,GO:0009506,GO:0008460,GO:0016853,GO:0016491,GO:0050377,GO:0010280,GO:0010315,GO:0051555,GO:0010253,GO:0005886,GO:0008830,GO:0008831,GO:0016616,GO:0010489,GO:0010490,GO:0071555,GO:0019305,GO:0005739,GO:0030154	KEGG:ath:AT3G14790;KEGG:ath:AT1G63000;KEGG:ath:AT1G78570

1	11178365	DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	11175252	11176197	DN68526_c0_g4	Trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM1 {ECO:0000305|PubMed:17190829}	GO:0005829,GO:0005739,GO:0009506,GO:0008460,GO:0016853,GO:0016491,GO:0050377,GO:0010280,GO:0010315,GO:0030154,GO:0071555,GO:0051555,GO:0042127	KEGG:ath:AT1G78570
1	11178365	DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	11175837	11174804	DN77992_c1_g1	Trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM3 {ECO:0000305|PubMed:17190829}	GO:0005829,GO:0009506,GO:0008460,GO:0016853,GO:0016491,GO:0050377,GO:0010280,GO:0010315,GO:0051555	KEGG:ath:AT3G14790
1	11178365	DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	11177248	11176664	DN68526_c0_g2	Bifunctional dTDP-4-dehydrorhamnose 3,5-epimerase/dTDP-4-dehydrorhamnose reductase {ECO:0000305}	GO:0005886,GO:0009506,GO:0008830,GO:0008831,GO:0016616,GO:0010489,GO:0010490,GO:0071555,GO:0019305	KEGG:ath:AT1G63000
1	12048129	GM_4	0.082	12045182	12045469	DN85328_c3_g1			
1	12048129	GM_4	0.082	12046838	12047210	DN90895_c1_g2			
1	12048129	GM_4	0.082	12050316	12050604	DN92055_c5_g1			
1	12048129	GM_4	0.082	12050566	12050337	DN92407_c3_g1			

1	12048129	GM_4	0.082	12050606	12050353	DN70501_c2_g1			
1	14847909	DM_2	0.237	14851807	14851183	DN74650_c1_g3			
1	14847909	DM_2	0.237	14851833	14852628	DN74650_c1_g1	Sperm-associated antigen 1A	GO:0005737,GO:0005829	
1	14847909	DM_2	0.237	14852724	14851743	DN72444_c0_g1	Sperm-associated antigen 1A	GO:0005737,GO:0005829	
1	22557144	RG_4	0.091	22552665	22552287	DN89753_c2_g1			
1	22557144	RG_4	0.091	22558990	22559356	DN72187_c4_g2			
1	22557144	RG_4	0.091	22559802	22559512	DN77388_c0_g1			
1	22557144	RG_4	0.091	22560516	22560246	DN78712_c0_g2			
1	51382809	GM_8	0.108	51379525	51379886	DN78004_c0_g1	Malate dehydrogenase, cytoplasmic {ECO:0000303|Ref.1}	GO:0005886,GO:0030060,GO:0005975, GO:0006108,GO:0006734,GO:0006107	KEGG:zma:542598
1	51382809	GM_8	0.108	51386990	51386678	DN69968_c0_g3	Myb-related protein 306;Transcription factor MYB4 {ECO:0000305};Transcription factor MYB30 {ECO:0000305}	GO:0005634,GO:0003677,GO:0045893	KEGG:osa:4336407;KEGG:osa:4330027
1	59423187	SDM_2	0.270	59419823	59419185	DN64929_c2_g2			
1	59423187	SDM_2	0.270	59427372	59428373	DN77232_c2_g1			
1	62623601	SDM_T	0.148	62621154	62621601	DN71302_c0_g2	ARF guanine-nucleotide exchange factor GNOM	GO:0005829,GO:0005768,GO:0010008, GO:0005886,GO:0005085,GO:0042802, GO:0042803,GO:0010540,GO:0007155,	KEGG:ath:AT1G13980

								GO:0071555,GO:0070417,GO:0000911, GO:0009793	
1	62623601	SDM_T	0.148	62622551	62619731	DN71302_c0_g3	ARF guanine-nucleotide exchange factor GNL1;ARF guanine-nucleotide exchange factor GNOM	GO:0005829,GO:0005794,GO:0000139, GO:0005085,GO:0006897,GO:0080119, GO:0015031,GO:0032012,GO:0006890, GO:0005768,GO:0010008,GO:0005886, GO:0042802,GO:0042803,GO:0010540, GO:0007155,GO:0071555,GO:0070417, GO:0000911,GO:0009793	KEGG:ath:AT5G39500;KEGG:ath:AT1G13980
1	70240593	DM_T;GM_8	0.129; 0.439	70236251	70238578	DN91367_c1_g2			
1	70240593	DM_T;GM_8	0.129; 0.439	70236369	70235645	DN73962_c1_g1	Renalase {ECO:0000250|UniProtKB:Q48MT7}	GO:0050660,GO:0051287,GO:0050661	KEGG:pst:PSPTO_1126
1	70240593	DM_T;GM_8	0.129; 0.439	70241197	70241684	DN65798_c1_g5	Patatin-like protein 7	GO:0005886,GO:0047617,GO:0004620, GO:0019374,GO:0016042,GO:0006644, GO:0040008	KEGG:ath:AT3G54950
1	70240593	DM_T;GM_8	0.129; 0.439	70241561	70240676	DN65798_c1_g1	Patatin-like protein 7	GO:0005886,GO:0047617,GO:0004620, GO:0019374,GO:0016042,GO:0006644, GO:0040008	KEGG:ath:AT3G54950

1	70240593	DM_T;GM_8	0.129;0.439	70243020	70242555	DN80775_c1_g2	Patatin-like protein 3;Patatin-like protein 8	GO:0016787,GO:0006952	KEGG:osa:4342389;KEGG:ath:AT4G29800
2	2268441	RG_9	0.369	2265414	2264627	DN92198_c4_g1			
2	2268441	RG_9	0.369	2272425	2272732	DN74648_c2_g1			
2	5498836	RG_4	0.085	5496862	5496586	DN88838_c2_g2			
2	5498836	RG_4	0.085	5497560	5497258	DN74008_c0_g1			
2	41678485	RG_6	0.119	41677728	41678023	DN88838_c2_g2			
2	41678485	RG_6	0.119	41681686	41682343	DN59956_c2_g1			
2	41678485	RG_6	0.119	41682454	41684045	DN85683_c0_g1			
2	41678485	RG_6	0.119	41682735	41682432	DN91174_c2_g4			
2	49786056	SDM_2	0.105	49784917	49785178	DN70501_c2_g1			
2	49786056	SDM_2	0.105	49784964	49785194	DN92407_c3_g1			
2	49786056	SDM_2	0.105	49785215	49784925	DN92055_c5_g1			
2	49786056	SDM_2	0.105	49786005	49786595	DN58994_c4_g1	Myb-related protein 306;Transcription factor MYB60 {ECO:0000303|PubMed:9839469};Transcription factor MYB60 {ECO:0000303|PubMed:18647406}	GO:0005634,GO:0003677,GO:0003700,GO:0000976,GO:0080148,GO:0009737,GO:0009646,GO:0009733,GO:0009637,GO:0009409,GO:0009416,GO:1902074,GO:0009744,GO:0009411,GO:0009414,GO:0009651	KEGG:ath:AT1G08810;KEGG:vvi:100233072

2	49786056	SDM_2	0.105	49786741	49787164	DN76319_c2_g1	Transcription factor MYB94 {ECO:0000303|PubMed:9839469};Myb-related protein 306	GO:0005634,GO:0003700,GO:0043565, GO:0000976,GO:0045893,GO:1904278, GO:0006355,GO:0009733,GO:0009414	KEGG:ath:AT3G47600
2	49786056	SDM_2	0.105	49787052	49787738	DN88364_c1_g1	Myb-related protein 306	GO:0005634	
2	49786056	SDM_2	0.105	49788044	49786983	DN76319_c4_g2			
2	52975994	RG_7	0.210	52972075	52972424	DN74990_c0_g2			
2	64677456	DM_9	0.087	64674447	64675358	DN77139_c1_g1	Indole-3-acetic acid-amido synthetase GH3.8 {ECO:0000305}	GO:0005737,GO:0016881,GO:0016208, GO:0010279,GO:0010252,GO:0006952, GO:0009555	KEGG:osa:4343785
2	64677456	DM_9	0.087	64675018	64674410	DN63283_c0_g1	Indole-3-acetic acid-amido synthetase GH3.8 {ECO:0000305};Probable indole-3-acetic acid-amido synthetase GH3.1	GO:0005737,GO:0016881,GO:0016208, GO:0010279,GO:0010252,GO:0006952, GO:0009555,GO:1900424,GO:0009733	KEGG:osa:4343785;KEGG:osa:4327043
2	64677456	DM_9	0.087	64682359	64682668	DN83067_c0_g1	FRIGIDA-like protein 3	GO:0009536,GO:0030154	KEGG:ath:AT5G48385
2	64677456	DM_9	0.087	64682428	64681984	DN80966_c0_g1	FRIGIDA-like protein 3	GO:0009536,GO:0030154	KEGG:ath:AT5G48385
2	67098278	DM_4	0.163	67094490	67093879	DN72905_c0_g1	Dihydroneopterin aldolase 1 {ECO:0000305};Dihydroneopterin aldolase 2 {ECO:0000305}	GO:0005737,GO:0004150,GO:0046656, GO:0006760	KEGG:ath:AT3G11750

2	67098278	DM_4	0.163	67101745	67102220	DN72984_c1_g2			
2	71727123	GM_4	0.077	71723998	71724817	DN88440_c4_g1	Sugar transport protein MST6 {ECO:0000305}	GO:0016021,GO:0005886,GO:0015145	KEGG:osa:4343590
2	71727123	GM_4	0.077	71724470	71724823	DN62517_c2_g5	Sugar transport protein MST3 {ECO:0000305}	GO:0016021,GO:0015145	KEGG:osa:4342198
2	71727123	GM_4	0.077	71724673	71725134	DN69828_c1_g4	Sugar transport protein MST3 {ECO:0000305}	GO:0016021,GO:0015145	KEGG:osa:4342198
2	71727123	GM_4	0.077	71724778	71724459	DN89543_c1_g3	Sugar transport protein MST3 {ECO:0000305};Sugar transport protein MST6 {ECO:0000305}	GO:0016021,GO:0015145,GO:0015293	KEGG:osa:4342198;KEGG:osa:4343590
2	71727123	GM_4	0.077	71724906	71725223	DN88440_c5_g1	Sugar transport protein MST6 {ECO:0000305};Sugar transport protein MST3 {ECO:0000305}	GO:0016021,GO:0005886,GO:0015145	KEGG:osa:4343590;KEGG:osa:4342198
2	71727123	GM_4	0.077	71725037	71725680	DN69828_c0_g1	Sugar transport protein MST6 {ECO:0000305}	GO:0016021,GO:0005886,GO:0015145	KEGG:osa:4343590
2	71727123	GM_4	0.077	71725122	71724675	DN69828_c1_g2	Sugar transport protein MST3 {ECO:0000305}	GO:0016021,GO:0015145	KEGG:osa:4342198
2	71727123	GM_4	0.077	71725183	71724819	DN69828_c1_g3	Sugar transport protein MST6 {ECO:0000305}	GO:0016021,GO:0005886,GO:0015145	KEGG:osa:4343590
2	71727123	GM_4	0.077	71725455	71724470	DN65192_c3_g3	Sugar transport protein MST4 {ECO:0000305}	GO:0016021,GO:0005886,GO:0005358,GO:0009737,GO:0009651	KEGG:osa:4332079

2	71727123	GM_4	0.077	71725785	71724985	DN88440_c5_g2	Sugar transport protein MST6 {ECO:0000305}	GO:0016021,GO:0005886,GO:0015145	KEGG:osa:4343590
2	71727123	GM_4	0.077	71727517	71727813	DN90485_c1_g2			
2	71727123	GM_4	0.077	71730677	71730978	DN74008_c0_g1			
2	72127644	GM_6	0.104	72130649	72129278	DN62703_c0_g2			
3	780392	GM_3	0.085	775480	774712	DN78485_c1_g1	FCS-Like Zinc finger 1 {ECO:0000303|PubMed:24901469};FCS-Like Zinc finger 2 {ECO:0000303|PubMed:24901469}	GO:0005737,GO:0005634,GO:0019900,GO:0046872,GO:0019902,GO:0071456,GO:0009749,GO:1905582,GO:0042594	KEGG:ath:AT5G47060;KEGG:ath:AT4G17670
3	780392	GM_3	0.085	777709	778287	DN92370_c2_g1	Shewanella-like protein phosphatase 1 {ECO:0000303|PubMed:21976480}	GO:0009507,GO:0030145,GO:0016791	KEGG:ath:AT1G07010
3	780392	GM_3	0.085	780624	780885	DN70501_c2_g1			
3	780392	GM_3	0.085	780922	780632	DN92055_c5_g1			
3	780392	GM_3	0.085	781602	782459	DN71855_c0_g2			
3	780392	GM_3	0.085	781981	781576	DN63968_c0_g2	FCS-Like Zinc finger 1 {ECO:0000303|PubMed:24901469}	GO:0005737,GO:0005634,GO:0019900,GO:0046872,GO:0019902,GO:0071456,GO:0009749,GO:1905582,GO:0042594	KEGG:ath:AT5G47060

3	780392	GM_3	0.085	782786	781912	DN71855_c0_g1	MYG1 exonuclease {ECO:0000305};MYG1 protein C694.04c	GO:0005737,GO:0005759,GO:0005739, GO:0005730,GO:0005654,GO:0005634, GO:0004518,GO:0005829	KEGG:hsa:60314;KEGG:spo:SPAC694.04c
3	780392	GM_3	0.085	783088	783363	DN88498_c2_g1	MYG1 exonuclease {ECO:0000305};MYG1 protein C694.04c	GO:0005737,GO:0005759,GO:0005739, GO:0005730,GO:0005654,GO:0005634, GO:0004518,GO:0005829	KEGG:hsa:60314;KEGG:spo:SPAC694.04c
3	1134135	GM_5	0.103	1131615	1131867	DN90011_c5_g1			
3	1134135	GM_5	0.103	1131867	1131615	DN84073_c1_g1			
3	1134135	GM_5	0.103	1133660	1134379	DN58624_c1_g1			
3	1134135	GM_5	0.103	1134467	1133789	DN58624_c1_g2			
3	9317234	RG_7	0.137	9314767	9315028	DN92411_c3_g1			
3	15115892	GM_3	0.280	15116186	15116531	DN68616_c1_g2			
3	15115892	GM_3	0.280	15116915	15116593	DN64750_c2_g2			
3	15115892	GM_3	0.280	15118371	15118627	DN92413_c6_g1			
3	15115892	GM_3	0.280	15118576	15118998	DN92413_c6_g2			
3	15115892	GM_3	0.280	15118615	15118948	DN63580_c0_g1			
3	15115892	GM_3	0.280	15118623	15118140	DN63607_c1_g1	Uncharacterized mitochondrial protein AtMg00310;LINE-1 retrotransposable element ORF2 protein;Transposon TX1	GO:0005739,GO:0046872,GO:0003964, GO:0006310,GO:0090305,GO:0032197	KEGG:ath:ArthMp029

							uncharacterized 149 kDa protein		
3	15115892	GM_3	0.280	15118638	15119023	DN59037_c0_g1			
3	15115892	GM_3	0.280	15118638	15119036	DN63580_c0_g2			
3	15115892	GM_3	0.280	15118672	15118345	DN64020_c2_g3			
3	15115892	GM_3	0.280	15118936	15118531	DN59037_c1_g1			
3	15115892	GM_3	0.280	15118948	15118613	DN59037_c1_g2			
3	15115892	GM_3	0.280	15119072	15119451	DN62166_c1_g1			
3	15115892	GM_3	0.280	15119197	15119608	DN92413_c6_g3			
3	15115892	GM_3	0.280	15119295	15118983	DN92415_c9_g1			
3	15115892	GM_3	0.280	15119447	15119104	DN61033_c1_g2			
3	15115892	GM_3	0.280	15119518	15119830	DN86079_c2_g3			
3	15115892	GM_3	0.280	15119524	15119952	DN59798_c2_g1			
3	15115892	GM_3	0.280	15119534	15119806	DN92413_c5_g1			
3	15115892	GM_3	0.280	15119612	15119317	DN86325_c1_g1			
3	15115892	GM_3	0.280	15119644	15119964	DN78301_c2_g3			
3	15115892	GM_3	0.280	15119661	15119354	DN71673_c1_g2			
3	15115892	GM_3	0.280	15119663	15119415	DN64835_c0_g1			
3	15115892	GM_3	0.280	15119678	15119917	DN60624_c1_g1			
3	15115892	GM_3	0.280	15119687	15119964	DN61222_c3_g3			

3	15115892	GM_3	0.280	15119836	15119389	DN72295_c1_g2			
3	15115892	GM_3	0.280	15119964	15119687	DN78301_c2_g4			
3	15115892	GM_3	0.280	15119964	15120739	DN79343_c0_g5			
3	46155579	RG_2	0.372	46150902	46151275	DN60337_c0_g2			
3	46155579	RG_2	0.372	46150922	46151294	DN60337_c0_g1			
3	46155579	RG_2	0.372	46151310	46150996	DN60337_c0_g3			
3	46155579	RG_2	0.372	46154651	46155035	DN89780_c1_g2			
3	46155579	RG_2	0.372	46154884	46155271	DN87303_c1_g1			
3	46155579	RG_2	0.372	46155372	46155086	DN89780_c1_g1			
3	46155579	RG_2	0.372	46155762	46156033	DN61834_c4_g2			
3	46155579	RG_2	0.372	46156057	46155735	DN92072_c5_g1	Altered inheritance rate of mitochondria protein 25	GO:0005739,GO:0005886,GO:0017128, GO:0034605,GO:0034599	KEGG:sce:YJR100C
3	46155579	RG_2	0.372	46156058	46155731	DN59818_c2_g4			
3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156; 0.294; 0.133; 0.099	52373263	52373581	DN82633_c1_g2	Phosphatidylinositol/phosphatidylcholine transfer protein SFH4;Phosphatidylinositol/phosphatidylcholine transfer protein SFH6;Phosphatidylinositol/phosphatidylcholine transfer protein SFH8	GO:0000139,GO:0005886	KEGG:ath:AT1G19650;KEGG:ath:AT4G39170;KEGG:ath:AT2G21520

3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156;0.294;0.133;0.099	52373578	52373263	DN64082_c0_g1	Phosphatidylinositol/phosphatidylcholine transfer protein SFH8;Phosphatidylinositol/phosphatidylcholine transfer protein SFH4	GO:0000139,GO:0005886	KEGG:ath:AT2G21520;KEGG:ath:AT1G19650
3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156;0.294;0.133;0.099	52378743	52379111	DN78984_c0_g2	Proteasome subunit alpha type-1-B	GO:0005829,GO:0005634,GO:0009506,GO:0000502,GO:0005839,GO:0019773,GO:0010498	KEGG:ath:AT1G47250
3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156;0.294;0.133;0.099	52378914	52379279	DN78984_c0_g3	Proteasome subunit alpha type-1	GO:0005737,GO:0005634,GO:0005839,GO:0019773,GO:0010498	KEGG:osa:4328215
3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156;0.294;0.133;0.099	52381354	52381659	DN78984_c0_g1	Proteasome subunit alpha type-1	GO:0005737,GO:0005634,GO:0005839,GO:0019773,GO:0010498	KEGG:osa:4328215
3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156;0.294;0.133;0.099	52382175	52381665	DN71801_c0_g3	Proteasome subunit alpha type-1	GO:0005737,GO:0005634,GO:0005839,GO:0019773,GO:0010498	KEGG:osa:4328215
3	52378212	DM_5;GM_5;LDM_5;LDM_T	0.156;0.294;	52382857	52383542	DN88548_c1_g2	Aspartic proteinase 36 {ECO:0000303|PubMed:27872247};Aspartic proteinase 39	GO:0031225,GO:0005737,GO:0005829,GO:0005886,GO:0090406,GO:0004190,	KEGG:ath:AT5G36260;KEGG:ath:AT1G65240

			0.133; 0.099				{ECO:0000303|PubMed:27872247}	GO:0009555,GO:0009846,GO:0009860, GO:0010183,GO:0030163,GO:0006508	
3	58121164	RG_3	0.393	58124003	58124436	DN83581_c0_g1			
3	58174943	DM_5;DM_7;GM_7;LDM_T	0.369; 0.088; 0.095; 0.145	58170573	58170836	DN92024_c3_g2			
4	3959984	RG_7	0.219	3964656	3963915	DN62556_c1_g1	Mini zinc finger protein 1	GO:0005737,GO:0005634,GO:0003700, GO:0046872,GO:0000976	KEGG:osa:4349655
4	9291205	RG_8	0.090	9286742	9285019	DN85079_c0_g1			
4	9291205	RG_8	0.090	9294013	9294324	DN59818_c2_g4			
4	9291205	RG_8	0.090	9294016	9294333	DN92072_c5_g1	Altered inheritance rate of mitochondria protein 25	GO:0005739,GO:0005886,GO:0017128, GO:0034605,GO:0034599	KEGG:sce:YJR100C
4	9291205	RG_8	0.090	9294285	9294014	DN59818_c2_g5			
4	9291205	RG_8	0.090	9294311	9294045	DN61834_c4_g2			
4	9291205	RG_8	0.090	9295494	9296449	DN84649_c1_g2			
4	9291205	RG_8	0.090	9295909	9295363	DN67033_c0_g1			
4	14151549	GM_2;LDM_2	0.536; 0.339	14147826	14148107	DN70501_c2_g1			
4	14151549	GM_2;LDM_2	0.536; 0.339	14148112	14147853	DN92055_c5_g1			

4	14151549	GM_2;LDM_2	0.536;0.339	14148398	14148662	DN61834_c4_g2			
4	14151549	GM_2;LDM_2	0.536;0.339	14148422	14148692	DN59818_c2_g5			
4	14151549	GM_2;LDM_2	0.536;0.339	14148692	14148367	DN59818_c2_g4			
4	14151549	GM_2;LDM_2	0.536;0.339	14148694	14148371	DN92072_c5_g1	Altered inheritance rate of mitochondria protein 25	GO:0005739,GO:0005886,GO:0017128,GO:0034605,GO:0034599	KEGG:sce:YJR100C
4	25045056	RG_5	0.179	25044102	25045355	DN88735_c0_g1	RNA demethylase ALKBH10B {ECO:0000305}	GO:0032451,GO:0046872,GO:0003729,GO:1990931,GO:0006402	KEGG:ath:AT4G02940
4	25045056	RG_5	0.179	25045115	25045441	DN83624_c1_g3			
4	25045056	RG_5	0.179	25045355	25044921	DN90901_c1_g1			
4	25045056	RG_5	0.179	25046764	25046259	DN83624_c1_g1	RNA demethylase ALKBH10B {ECO:0000305}	GO:0032451,GO:0046872,GO:0003729,GO:1990931,GO:0006402	KEGG:ath:AT4G02940
4	25045056	RG_5	0.179	25047378	25046926	DN71468_c0_g2			
4	25045056	RG_5	0.179	25048105	25047339	DN71468_c0_g1			
4	25045056	RG_5	0.179	25048425	25049998	DN77535_c0_g1			
5	10809190	SDM_2	0.152	10806421	10805844	DN62812_c1_g1	Glycerophosphodiester phosphodiesterase GDPD4 {ECO:0000305}	GO:0016021,GO:0008889,GO:0006071	KEGG:ath:AT1G71340

5	10809190	SDM_2	0.152	10809358	10809589	DN77996_c0_g1	Probable solanesyl-diphosphate synthase 3, chloroplastic {ECO:0000305};Solanesyl-diphosphate synthase 2, chloroplastic {ECO:0000303|PubMed:20421194}	GO:0009507,GO:0046872,GO:0008299,GO:0009536,GO:0004337,GO:0004659	KEGG:osa:4351957;KEGG:osa:107276142
5	10809190	SDM_2	0.152	10813994	10810596	DN90414_c2_g1			
5	58853938	RG_4	0.614	58857380	58857898	DN91398_c0_g1			
6	27936260	DM_3	0.111	27940640	27940246	DN71614_c2_g1			
6	44238660	RG_9	0.144	44242237	44242527	DN92055_c5_g1			
6	44238660	RG_9	0.144	44242527	44242274	DN70501_c2_g1			
7	19129582	RG_6	0.255	19129069	19129440	DN68681_c0_g1	Auxin-responsive protein IAA4	GO:0005634,GO:0009734	KEGG:osa:4327919
7	19129582	RG_6	0.255	19130506	19129459	DN81791_c0_g1	Auxin-responsive protein IAA4	GO:0005634,GO:0009734	KEGG:osa:4327919
7	19129582	RG_6	0.255	19131391	19131082	DN68681_c0_g4			
7	42826434	DM_4;DM_7;DM_8;DM_T;GM_4;GM_T;LDM_T;RG_8;RG_T	0.258;0.305;0.218;0.274;0.646;0.228;	42822392	42822686	DN92326_c7_g2			

			0.410; 0.426; 0.229						
7	42826434	DM_4;DM_7 ;DM_8;DM_ T;GM_4;GM _T;LDM_T;R G_8;RG_T	0.258; 0.305; 0.218; 0.274; 0.646; 0.228; 0.410; 0.426; 0.229	42822526	42822845	DN91682_c3_g2	Phenylacetaldehyde reductase {ECO:0000303|PubMed:20650544};Cinnamoyl-CoA reductase 1 {ECO:0000303|PubMed:24985707, ECO:0000303|PubMed:25217505};Cinnamoyl-CoA reductase 1 {ECO:0000305};Tetraketide alpha-pyrone reductase 1	GO:0016491,GO:0005737,GO:0016621, GO:0000166,GO:0007623,GO:0010597, GO:0009809,GO:0009699,GO:0016616, GO:0006952,GO:0005783,GO:0005634, GO:0009555,GO:0010584,GO:0048316	KEGG:osa:4331085;KEGG:ath:AT4G35420
7	42826434	DM_4;DM_7 ;DM_8;DM_ T;GM_4;GM _T;LDM_T;R G_8;RG_T	0.258; 0.305; 0.218; 0.274; 0.646; 0.228; 0.410; 0.426; 0.229	42822771	42822341	DN92411_c5_g2			

7	42826434	DM_4;DM_7;DM_8;DM_T;GM_4;GM_T;LDM_T;RG_8;RG_T	0.258;0.305;0.218;0.274;0.646;0.228;0.410;0.426;0.229	42823900	42823573	DN91031_c2_g1	Polyadenylate-binding protein RBP47C';RNA-binding protein L {ECO:0000303|PubMed:20217123}	GO:0010494,GO:0005829,GO:0005634,GO:0003729,GO:0008143,GO:0034605,GO:0006397,GO:0005737	KEGG:ath:AT1G47500;KEGG:osa:4337064
7	42826434	DM_4;DM_7;DM_8;DM_T;GM_4;GM_T;LDM_T;RG_8;RG_T	0.258;0.305;0.218;0.274;0.646;0.228;0.410;0.426;0.229	42827903	42827488	DN78529_c0_g1	DEAD-box ATP-dependent RNA helicase 25	GO:0005524,GO:0016887,GO:0003723	KEGG:osa:4326614
7	42826434	DM_4;DM_7;DM_8;DM_T;GM_4;GM_T;LDM_T;RG_8;RG_T	0.258;0.305;0.218;0.274;0.646;0.228;0.410;	42830407	42830773	DN78268_c1_g2			

			0.426; 0.229						
7	42826434	DM_4;DM_7;DM_8;DM_T;GM_4;GM_T;LDM_T;RG_8;RG_T	0.258; 0.305; 0.218; 0.274; 0.646; 0.228; 0.410; 0.426; 0.229	42830590	42831767	DN85036_c0_g1			
7	43749191	DM_3	0.220	43745383	43745105	DN87407_c0_g1			
7	43749191	DM_3	0.220	43748265	43747985	DN64763_c1_g1	E3 ubiquitin-protein ligase SINAT5 {ECO:0000305}	GO:0005737,GO:0005634,GO:0061630, GO:0008270,GO:0016567	KEGG:ath:AT5G53360
7	43749191	DM_3	0.220	43748265	43747988	DN65961_c3_g1			
7	43749191	DM_3	0.220	43748933	43749538	DN65203_c0_g1	Cinnamoyl-CoA reductase-like SNL6 {ECO:0000305}	GO:0003854,GO:0016616,GO:0042742, GO:0009809	KEGG:osa:4326581
7	43749191	DM_3	0.220	43749446	43749812	DN82034_c1_g1	Cinnamoyl-CoA reductase-like SNL6 {ECO:0000305}	GO:0003854,GO:0016616,GO:0042742, GO:0009809	KEGG:osa:4326581
7	43749191	DM_3	0.220	43749788	43749480	DN65203_c0_g3	Cinnamoyl-CoA reductase-like SNL6 {ECO:0000305}	GO:0003854,GO:0016616,GO:0042742, GO:0009809	KEGG:osa:4326581
7	43749191	DM_3	0.220	43753312	43753568	DN82223_c4_g4			

7	43749191	DM_3	0.220	43753415	43754060	DN73479_c1_g1			
7	53131598	RG_5	0.296	53127527	53127175	DN77132_c1_g3	G-type lectin S-receptor-like serine/threonine-protein kinase At1g11330;G-type lectin S-receptor-like serine/threonine-protein kinase SD1-1;Receptor-like serine/threonine-protein kinase SD1-8	GO:0005576,GO:0016021,GO:0005886, GO:0005524,GO:0005516,GO:0030246, GO:0004672,GO:0106310,GO:0004674, GO:0004712,GO:0006955,GO:0006468, GO:0048544,GO:0009506,GO:0031625, GO:0005773	KEGG:ath:AT1G11330;KEGG:ath:AT4G21380;KEGG:ath:AT4G27300
7	53131598	RG_5	0.296	53129629	53129375	DN62551_c0_g2	Receptor-like serine/threonine-protein kinase SD1-7;G-type lectin S-receptor-like serine/threonine-protein kinase At1g11330	GO:0016021,GO:0005634,GO:0005886, GO:0005524,GO:0030246,GO:0004672, GO:0106310,GO:0004674,GO:0004712, GO:0031625,GO:0009738,GO:0071215, GO:0046777,GO:0006468,GO:0005576, GO:0005516,GO:0006955	KEGG:ath:AT1G65790;KEGG:ath:AT1G11330
7	53131598	RG_5	0.296	53130559	53130817	DN70501_c2_g1			
7	53131598	RG_5	0.296	53130603	53130833	DN92407_c3_g1			
7	53131598	RG_5	0.296	53130854	53130568	DN92055_c5_g1			
7	54377956	GM_9;DM_3;GM_3	0.123;0.494;0.316	54377699	54378060	DN77103_c1_g1	F-box protein SKIP24	GO:0005634,GO:0005886	KEGG:ath:AT1G08710

7	54377956	GM_9;DM_3;GM_3	0.123;0.494;0.316	54378425	54377831	DN81386_c1_g1	F-box protein SKIP24	GO:0005634,GO:0005886	KEGG:ath:AT1G08710
7	54377956	GM_9;DM_3;GM_3	0.123;0.494;0.316	54380492	54380798	DN69861_c2_g2			
7	54377956	GM_9;DM_3;GM_3	0.123;0.494;0.316	54381259	54382150	DN70964_c3_g1			
7	54380453	GM_9;DM_3;GM_3	0.123;0.494;0.316	54383819	54384110	DN90485_c1_g2			
7	54380453	GM_9;DM_3;GM_3	0.123;0.494;0.316	54384128	54383825	DN92407_c3_g1			
7	54380453	GM_9;DM_3;GM_3	0.494;0.316	54385397	54385017	DN75652_c0_g2	NAC transcription factor 47 {ECO:0000305};NAC domain-containing protein 2;NAC transcription factor 56 {ECO:0000303|PubMed:15029955};NAC transcription factor NAM-B2;NAC transcription factor ONAC010;NAC transcription factor NAM-2	GO:0005634,GO:0003700,GO:0000976,GO:0009793,GO:0010365,GO:0009413,GO:0019900,GO:0071456,GO:0009788,GO:0009611,GO:0080060,GO:0045995,GO:0009753,GO:0048317,GO:0048731,GO:0003677,GO:0006355,GO:0048653	KEGG:ath:AT1G01720;KEGG:ath:AT3G04070;KEGG:ath:AT3G15510

7	56928375	DM_2	0.180	56927853	56929373	DN81954_c1_g1	Laccase-7	GO:0048046,GO:0005507,GO:0052716, GO:0016491	KEGG:osa:4324802
7	57318671	SDM_5	0.231	57316152	57317025	DN86224_c1_g1	Cytochrome P450 94C1 {ECO:0000305}	GO:0005789,GO:0016021,GO:0043231, GO:0018685,GO:0020037,GO:0005506	KEGG:ath:AT2G27690
7	57318671	SDM_5	0.231	57316582	57315812	DN69510_c1_g1	Cytochrome P450 94C1 {ECO:0000305};Cytochrome P450 {ECO:0000250|UniProtKB:O81117}	GO:0005789,GO:0016021,GO:0043231, GO:0018685,GO:0020037,GO:0005506, GO:0009611,GO:0004497	KEGG:ath:AT2G27690
7	57318671	SDM_5	0.231	57316806	57315679	DN74642_c1_g1	Cytochrome P450 94C1 {ECO:0000305}	GO:0005789,GO:0016021,GO:0043231, GO:0018685,GO:0020037,GO:0005506	KEGG:ath:AT2G27690
7	57318671	SDM_5	0.231	57317141	57316457	DN71916_c0_g1	Cytochrome P450 94C1 {ECO:0000305}	GO:0005789,GO:0016021,GO:0043231, GO:0018685,GO:0020037,GO:0005506	KEGG:ath:AT2G27690
7	57318671	SDM_5	0.231	57317804	57318129	DN92072_c5_g1	Altered inheritance rate of mitochondria protein 25	GO:0005739,GO:0005886,GO:0017128, GO:0034605,GO:0034599	KEGG:sce:YJR100C
7	57318671	SDM_5	0.231	57317806	57318135	DN59818_c2_g4			
7	57318671	SDM_5	0.231	57318077	57317809	DN59818_c2_g5			
7	57318671	SDM_5	0.231	57318101	57317840	DN61834_c4_g2			
8	15859892	DM_6;GM_6	0.212; 0.303	15857413	15857121	DN88838_c2_g2			

8	15859892	DM_6;GM_6	0.212; 0.303	15858125	15857411	DN85683_c0_g1			
8	15859892	DM_6;GM_6	0.212; 0.303	15859106	15858422	DN59956_c2_g1			
8	37896003	RG_5	0.160	37900805	37899730	DN77486_c1_g1			
8	39023791	DM_6;GM_6	0.215; 0.253	39022061	39022357	DN65405_c2_g1			
8	39023791	DM_6;GM_6	0.215; 0.253	39022439	39022108	DN89584_c2_g4			
8	39023791	DM_6;GM_6	0.215; 0.253	39024960	39025299	DN91658_c2_g1			
8	39023791	DM_6;GM_6	0.215; 0.253	39025462	39024925	DN59875_c0_g2			
8	39023791	DM_6;GM_6	0.215; 0.253	39026234	39025903	DN79105_c0_g1			
8	43907416	GM_7	0.062	43904185	43904414	DN92413_c6_g1			
8	43907416	GM_7	0.062	43904640	43904324	DN90841_c2_g1			
8	43907416	GM_7	0.062	43905103	43904857	DN75855_c3_g1			
8	43907416	GM_7	0.062	43905306	43905673	DN86079_c2_g3			
8	43907416	GM_7	0.062	43905383	43905649	DN92413_c5_g1			
8	43907416	GM_7	0.062	43905469	43905788	DN59798_c2_g1			

8	43907416	GM_7	0.062	43905469	43905753	DN60624_c1_g1			
8	51586400	LDM_2	0.302	51583525	51582885	DN71539_c0_g2			
8	51586400	LDM_2	0.302	51587113	51587418	DN83969_c1_g2	Probable protein S-acyltransferase 19;Probable protein S-acyltransferase 20	GO:0005783,GO:0005794,GO:0016021, GO:0005886,GO:0019706,GO:0018230, GO:0006612	KEGG:ath:AT4G15080;KEGG:ath:AT3G22180
8	51586400	LDM_2	0.302	51589562	51590151	DN79785_c2_g1	Dihydroneopterin aldolase 1 {ECO:0000305}	GO:0005737,GO:0004150,GO:0046656, GO:0006760	KEGG:ath:AT3G11750
8	51586400	LDM_2	0.302	51590571	51590035	DN79785_c2_g2	Dihydroneopterin aldolase 1 {ECO:0000305}	GO:0005737,GO:0004150,GO:0046656, GO:0006760	KEGG:ath:AT3G11750
8	51586400	LDM_2	0.302	51591296	51590982	DN66068_c0_g2	Dihydroneopterin aldolase 1 {ECO:0000305};Dihydroneopterin aldolase 2 {ECO:0000305}	GO:0005737,GO:0004150,GO:0046656, GO:0006760	KEGG:ath:AT3G11750;KEGG:ath:AT5G62980
8	53956999	RG_3	0.136	53954409	53954959	DN71644_c0_g5			
8	53956999	RG_3	0.136	53958430	53959303	DN80257_c0_g1	Mitochondrial import receptor subunit TOM9-2	GO:0016021,GO:0005742,GO:0005739, GO:0000325,GO:0009536,GO:0015450	KEGG:ath:AT5G43970
8	53956999	RG_3	0.136	53960775	53961141	DN89705_c0_g1	Histone-lysine N-methyltransferase SUVR4;Probable inactive histone-lysine N-methyltransferase SUVR1;Probable inactive	GO:0005694,GO:0005730,GO:0009506, GO:0018024,GO:0008270,GO:0006325, GO:0034968,GO:0042802,GO:0031047, GO:0005634	KEGG:ath:AT3G04380;KEGG:ath:AT1G04050;KEGG:ath:AT5G43990

							histone-lysine N-methyltransferase SUVR2		
9	19780116	GM_2	0.069	19779176	19779551	DN92411_c5_g2			
9	19780116	GM_2	0.069	19779418	19779104	DN91682_c3_g2	Phenylacetaldehyde reductase {ECO:0000303|PubMed:20650544};Cinnamoyl-CoA reductase 1 {ECO:0000303|PubMed:24985707, ECO:0000303|PubMed:25217505};Cinnamoyl-CoA reductase 1 {ECO:0000305};Tetraketide alpha-pyrone reductase 1	GO:0016491,GO:0005737,GO:0016621, GO:0000166,GO:0007623,GO:0010597, GO:0009809,GO:0009699,GO:0016616, GO:0006952,GO:0005783,GO:0005634, GO:0009555,GO:0010584,GO:0048316	KEGG:osa:4331085;KEGG:ath:AT4G35420
9	19780116	GM_2	0.069	19779551	19779239	DN92326_c7_g2			
9	19780116	GM_2	0.069	19780842	19780322	DN62073_c2_g1	DNA-binding protein RHL1	GO:0005730,GO:0003677,GO:0042023	KEGG:ath:AT1G48380
9	19780116	GM_2	0.069	19782323	19783017	DN62450_c0_g2			
9	19780116	GM_2	0.069	19782496	19782011	DN89834_c1_g1			
9	19780116	GM_2	0.069	19783051	19782773	DN62450_c0_g1			
9	21047849	SDM_T	0.198	21043035	21043308	DN61834_c4_g2			
9	21047849	SDM_T	0.198	21043061	21043327	DN59818_c2_g5			

9	21047849	SDM_T	0.198	21043315	21043010	DN92072_c5_g1	Altered inheritance rate of mitochondria protein 25	GO:0005739,GO:0005886,GO:0017128, GO:0034605,GO:0034599	KEGG:sce:YJR 100C
9	21047849	SDM_T	0.198	21043327	21043006	DN59818_c2_g4			
9	21047849	SDM_T	0.198	21046974	21045821	DN88525_c0_g3	Pre-mRNA-splicing factor syf2	GO:0071013,GO:0005634,GO:0071014, GO:0000974,GO:0071007	KEGG:xtr:49670 6;KEGG:gja:107 110818
9	21047849	SDM_T	0.198	21047992	21047514	DN88525_c0_g1			
9	21047849	SDM_T	0.198	21049096	21049797	DN91275_c1_g2			
9	21047849	SDM_T	0.198	21049666	21050979	DN74723_c0_g1			
9	21047849	SDM_T	0.198	21049925	21049500	DN74723_c0_g3			
9	21047849	SDM_T	0.198	21052383	21052910	DN59424_c0_g1			
9	41402643	RG_T	0.161	41398940	41399653	DN89022_c1_g4			
9	41402643	RG_T	0.161	41402256	41401712	DN88921_c1_g3			
9	41402643	RG_T	0.161	41405269	41404845	DN91410_c2_g1			
9	50511582	GM_7;GM_8 ;GM_T	0.553; 0.317; 0.156	50510305	50509992	DN89722_c3_g3			
9	50511582	GM_7;GM_8 ;GM_T	0.554; 0.317; 0.156	50514195	50513877	DN74841_c0_g1	G-type lectin S-receptor-like serine/threonine-protein kinase SD2-5	GO:0016021,GO:0005886,GO:0005524, GO:0005516,GO:0030246,GO:0004672, GO:0106310,GO:0004674,GO:0004712, GO:0031625	KEGG:ath:AT4 G32300

9	50511582	GM_7;GM_8;GM_T	0.554;0.317;0.156	50514347	50514699	DN90710_c2_g2	G-type lectin S-receptor-like serine/threonine-protein kinase SD2-5	GO:0016021,GO:0005886,GO:0005524,GO:0005516,GO:0030246,GO:0004672,GO:0106310,GO:0004674,GO:0004712,GO:0031625	KEGG:ath:AT4G32300
Scaffold89	16485	RG_8	0.081	16215	15985	DN64763_c1_g1	E3 ubiquitin-protein ligase SINAT5 {ECO:0000305}	GO:0005737,GO:0005634,GO:0061630,GO:0008270,GO:0016567	KEGG:ath:AT5G53360
Scaffold89	16485	RG_8	0.081	20419	20746	DN85080_c2_g2	Probable galactinol--sucrose galactosyltransferase 2	GO:0009506,GO:0016757,GO:0052692,GO:0034484	KEGG:ath:AT3G57520
Scaffold89	16485	RG_8	0.081	20515	20276	DN82730_c1_g1	Probable galactinol--sucrose galactosyltransferase 2	GO:0009506,GO:0016757,GO:0052692,GO:0034484	KEGG:ath:AT3G57520
Scaffold89	16485	RG_8	0.081	21454	21117	DN77042_c0_g6	Probable galactinol--sucrose galactosyltransferase 2	GO:0009506,GO:0016757,GO:0052692,GO:0034484,GO:0006979	KEGG:ath:AT3G57520
Scaffold365	37575	SDM_5	0.110	34241	35502	DN63127_c0_g1			
Scaffold365	37575	SDM_5	0.110	36899	37289	DN71748_c0_g2			
Scaffold365	37575	SDM_5	0.110	40753	41026	DN89448_c0_g2			
Scaffold365	37575	SDM_5	0.110	41041	40752	DN88340_c2_g3			

Supplementary Table S9. Number of sequencing reads before and after processing for each sample.

Sample	Library	Number of raw reads	Number of reads after processing	Remaining reads (%)
1	SRR8417281	148292760	135972832	91.7
2	SRR8417289	142004516	129537976	91.2
3	SRR8417290	144161184	131574900	91.3
4	SRR8417280	166628484	150736860	90.5
5	SRR8417287	153180772	140252060	91.6
6	SRR8417279	167171908	158219380	94.6
7	SRR8417288	154749616	142042352	91.8
8	SRR8417286	154998656	141191016	91.1
9	SRR8417283	162881616	147582888	90.6
10	SRR8417282	154189316	140215916	90.9
11	SRR8417291	143835128	131937544	91.7

Supplementary Table S10. Biological processes GO terms enrichment of the genes physically linked to the Major Importance markers associated with the agronomic traits.

Phenotype	Importance	GO-ID	Term	Annotated	Significant	Expected	p-value
DM_4;LDM_2	0.163; 0.302	GO:0046656	folic acid biosynthetic process	16	4	0.06	2.8e-07
DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	GO:0051555	flavonol biosynthetic process	19	4	0.07	5.8e-07
DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	GO:0010315	auxin efflux	40	4	0.14	1.3e-05
RG_8	0.081	GO:0034484	raffinose catabolic process	21	3	0.08	5.8e-05
STD_T	0.148	GO:0010274	hydrotropism	7	2	0.03	0.00027
STD_T	0.148	GO:0001736	establishment of planar polarity	8	2	0.03	0.00036
DM_5;GM_5;GM_9;LDM_5	0.133;0.394;0.119;0.129	GO:0019305	dTDP-rhamnose biosynthetic process	10	2	0.04	0.00057
SDM_T	0.148	GO:0032012	regulation of ARF protein signal transdu...	13	2	0.05	0.00098
DM_2; GM_2	0.248;0.069	GO:0042023	DNA endoreduplication	56	3	0.2	0.00110
DM_3;DM_4;DM_7;DM_8;DM_T;GM_2;GM_4;GM_T;LDM_T;	0.220;0.258;0.305;0.218;0.274;0.069;0.646;0.228;0.410;	GO:0009809	lignin biosynthetic process	145	4	0.53	0.00189

RG_8;RG_T	0.426;0.229						
STD_T	0.148	GO:0048209	regulation of vesicle targeting, to, from or within Golgi	19	2	0.07	0.00212
RG_9	0.095	GO:0051017	actin filament bundle assembly	20	2	0.07	0.00235
STD_T	0.148	GO:0009942	longitudinal axis specification	21	2	0.08	0.00260
GM_2;SDM_T	0.087;0.148	GO:0070417	cellular response to cold	81	3	0.29	0.00317
DM_2	0.248	GO:0000082	G1/S transition of mitotic cell cycle	25	2	0.09	0.00367
GM_3;	0.085	GO:1905582	response to mannose	33	2	0.12	0.00635
SDM_T	0.148	GO:0007155	cell adhesion	34	2	0.12	0.00673
DM_T;GM_8	0.129;0.439	GO:0019374	galactolipid metabolic process	37	2	0.13	0.00793

Supplementary Table S11. General co-expression network GO terms enrichment considering first neighbors and second neighbors.

GO-ID	Neighbor	Term	Annotated	Significant	Expected	p-value
GO:0051555	first	flavonol biosynthetic process	19	3	0.15	0.00044
GO:0010332	first	response to gamma radiation	9	2	0.07	0.00221
GO:0010315	first	auxin efflux	40	3	0.32	0.00399
GO:0060862	first	negative regulation of floral organ absc...	14	2	0.11	0.00543
GO:0080168	first	abscisic acid transport	16	2	0.13	0.00709
GO:0046656	first	folic acid biosynthetic process	16	2	0.13	0.00709
GO:0072710	first	response to hydroxyurea	1	1	0.01	0.008
GO:0072718	first	response to cisplatin	1	1	0.01	0.008
GO:0010086	first	embryonic root morphogenesis	1	1	0.01	0.008
GO:0009830	first	cell wall modification involved in absci...	1	1	0.01	0.008

GO:0042823	first	pyridoxal phosphate biosynthetic process	19	2	0.15	0.00994
GO:0048314	second	embryo sac morphogenesis	4	3	0.21	0.00052
GO:0010200	second	response to chitin	43	8	2.21	0.00213
GO:0048235	second	pollen sperm cell differentiation	12	4	0.62	0.00248
GO:0070716	second	mismatch repair involved in maintenance ...	2	2	0.1	0.00265
GO:1904430	second	negative regulation of t-circle formatio...	2	2	0.1	0.00265
GO:0051351	second	positive regulation of ligase activity	2	2	0.1	0.00265
GO:0010431	second	seed maturation	58	9	2.99	0.00288
GO:0042752	second	regulation of circadian rhythm	85	8	4.38	0.00449
GO:0070192	second	chromosome organization involved in meio...	64	8	3.3	0.0046
GO:0000373	second	Group II intron splicing	63	9	3.25	0.00476
GO:0032508	second	DNA duplex unwinding	48	6	2.47	0.00604
GO:0009249	second	protein lipoylation	3	2	0.15	0.00768
GO:1990426	second	mitotic recombination-dependent replicat...	3	2	0.15	0.00768
GO:0016233	second	telomere capping	9	3	0.46	0.00906
GO:0016560	second	protein import into peroxisome matrix, d...	9	3	0.46	0.00906
GO:0042138	second	meiotic DNA double-strand break formatio...	17	4	0.88	0.00972

Supplementary Table S12. General co-expression network top degree genes.

Gene	Protein	Degree
TRINITY_DN58237_c0_g2	Cytadherence high molecular weight protein 3	59
TRINITY_DN53084_c1_g1		52
TRINITY_DN57655_c0_g3	40S ribosomal protein S6	51
TRINITY_DN57071_c0_g1		50
TRINITY_DN57279_c0_g1		48
TRINITY_DN57900_c0_g1		48
TRINITY_DN57910_c2_g3	60S ribosomal protein L9	48
TRINITY_DN58785_c0_g1	Protein ELF4-LIKE 4	47
TRINITY_DN57269_c0_g1	14-3-3 protein zeta	45
TRINITY_DN57379_c0_g1		45
TRINITY_DN57626_c0_g1	Fatty acid-binding protein	45
TRINITY_DN57181_c0_g1	Protein SENSITIVE TO UV 2 {ECO:0000303|PubMed:19619159; ECO:0000303|PubMed:28556304}	43
TRINITY_DN58673_c1_g1	3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ {ECO:0000255|HAMAP-Rule:MF_00406}	43
TRINITY_DN57660_c0_g1	Putative lipid-transfer protein DIR1	42
TRINITY_DN57273_c0_g2		38
TRINITY_DN59010_c0_g1		38
TRINITY_DN57941_c1_g1	Probable 2-oxoglutarate/Fe(II)-dependent dioxygenase	37
TRINITY_DN57941_c1_g2		37
TRINITY_DN59015_c0_g1		37
TRINITY_DN59043_c0_g1	Probable glutamate carboxypeptidase LAMP1 {ECO:0000305}	36
TRINITY_DN57010_c0_g1		35
TRINITY_DN57379_c0_g2		35
TRINITY_DN59461_c0_g1		35
TRINITY_DN51913_c0_g1	Fumarylacetoacetase {ECO:0000305}	34

TRINITY_DN54379_c0_g1	Probable pyridoxal 5'-phosphate synthase subunit PDX1.1	34
TRINITY_DN57315_c0_g1		34
TRINITY_DN57453_c0_g2	LINE-1 reverse transcriptase homolog	34
TRINITY_DN58234_c2_g1		34
TRINITY_DN56725_c0_g1	54S ribosomal protein L12; mitochondrial	33
TRINITY_DN57089_c0_g2		33
TRINITY_DN57654_c1_g2	Amino-acid permease BAT1 homolog	33
TRINITY_DN57990_c0_g1		33
TRINITY_DN57452_c0_g1		32
TRINITY_DN57543_c0_g1	Zinc finger protein ZAT7	32
TRINITY_DN58974_c0_g1		32
TRINITY_DN59466_c0_g2		32
TRINITY_DN56773_c0_g1	Phosphoglycerate kinase; cytosolic	31
TRINITY_DN58187_c0_g1	Protein argonaute MEL1	31
TRINITY_DN56338_c0_g1		30
TRINITY_DN57255_c0_g1	3-ketoacyl-CoA synthase 11 {ECO:0000303|PubMed:18465198}	29
TRINITY_DN53411_c0_g1	AT-rich interactive domain-containing protein 4	28
TRINITY_DN56677_c0_g1	Costars family protein WS02710_H03	28
TRINITY_DN56991_c0_g2		28
TRINITY_DN57570_c0_g2		28
TRINITY_DN58418_c1_g1		28
TRINITY_DN58466_c0_g2		28
TRINITY_DN58616_c0_g1	Vacuolar protein sorting-associated protein 32 homolog 1	28
TRINITY_DN53294_c0_g1		26
TRINITY_DN54735_c0_g1	CRM-domain containing factor CFM3; chloroplastic/mitochondrial {ECO:0000305}	26
TRINITY_DN56221_c0_g1		26
TRINITY_DN56783_c2_g1	Eukaryotic translation elongation factor 2 {ECO:0000312|FlyBase:FBgn0000559}	26

TRINITY_DN57176_c0_g2	P-(S)-hydroxymandelonitrile lyase	26
TRINITY_DN57571_c0_g1	Uracil phosphoribosyltransferase	26
TRINITY_DN57856_c0_g1		26
TRINITY_DN55210_c0_g1	Cell division cycle 5-like protein	25
TRINITY_DN57003_c0_g1	GATA transcription factor 4	24
TRINITY_DN59271_c0_g2	Probable WRKY transcription factor 19	24
TRINITY_DN54950_c1_g1	Aspartyl protease family protein At5g10770	23
TRINITY_DN55298_c0_g1		23
TRINITY_DN55329_c0_g1		23
TRINITY_DN58379_c0_g11		23
TRINITY_DN58974_c0_g2	PLASTID TRANSCRIPTIONALLY ACTIVE protein 6; chloroplastic {ECO:0000303|PubMed:16326926}	23
TRINITY_DN55496_c0_g1	Repetitive proline-rich cell wall protein 1	22
TRINITY_DN56838_c0_g2	Stress-associated endoplasmic reticulum protein 2	22
TRINITY_DN58065_c0_g1		22
TRINITY_DN58176_c0_g1	BTB/POZ and MATH domain-containing protein 1	22
TRINITY_DN55766_c0_g1		21
TRINITY_DN59061_c0_g2	Cyclase-like protein 4 {ECO:0000303|PubMed:25974367}	21
TRINITY_DN59168_c0_g1		21
TRINITY_DN59289_c0_g2	Glutathione S-transferase	21
TRINITY_DN52325_c0_g1		20
TRINITY_DN53975_c0_g1		20
TRINITY_DN57462_c0_g1		20
TRINITY_DN53002_c0_g1	Type II inositol polyphosphate 5-phosphatase 15 {ECO:0000305}	19
TRINITY_DN57971_c0_g1	indole-2-monooxygenase	19
TRINITY_DN58383_c0_g1		19
TRINITY_DN59678_c0_g3	Protein VASCULATURE COMPLEXITY AND CONNECTIVITY {ECO:0000303|PubMed:25149602}	19
TRINITY_DN46718_c0_g1		18

TRINITY_DN57074_c0_g1		18
TRINITY_DN57631_c0_g2	Disease resistance protein RGA5 {ECO:0000305}	18
TRINITY_DN58979_c0_g1	ERI1 exoribonuclease 2	18
TRINITY_DN59718_c0_g1	Pre-rRNA-processing protein TSR2 homolog	18
TRINITY_DN53274_c0_g1		17
TRINITY_DN57608_c0_g1	Putative DNA glycosylase At3g47830 {ECO:0000305}	17
TRINITY_DN57824_c0_g1	Obtusifoliol 14-alpha demethylase	17
TRINITY_DN58492_c0_g1		17
TRINITY_DN56756_c1_g1	Tubulin beta-8 chain	16
TRINITY_DN58383_c0_g2		16
TRINITY_DN58430_c0_g2		16
TRINITY_DN58798_c2_g1	Disease resistance protein RPM1 {ECO:0000303|PubMed:7638602}	16
TRINITY_DN53127_c0_g1		15
TRINITY_DN53713_c0_g1		15
TRINITY_DN55298_c1_g1	Probable terpene synthase 3; chloroplastic {ECO:0000305}	15
TRINITY_DN55832_c0_g1		15
TRINITY_DN56769_c0_g2	Alpha/beta hydrolase domain-containing protein 17C {ECO:0000305}	15
TRINITY_DN58512_c0_g1		15
TRINITY_DN50098_c0_g1		14
TRINITY_DN51747_c0_g1		14
TRINITY_DN53643_c0_g1		14
TRINITY_DN57710_c0_g1	Protein LAZ1 {ECO:0000303|PubMed:20830211}	14
TRINITY_DN57931_c0_g1	Thiol protease SEN102	14
TRINITY_DN47482_c1_g1		13
TRINITY_DN50751_c0_g1		13
TRINITY_DN54867_c0_g2		13
TRINITY_DN55334_c1_g5	Glycerol kinase	13
TRINITY_DN56583_c0_g2		13

TRINITY_DN56604_c1_g2	Disease resistance protein RGA5 {ECO:0000305}	13
TRINITY_DN57813_c0_g1	Uncharacterized lipoprotein syc1174_c	13
TRINITY_DN59033_c4_g1	Tabersonine/lochnericine 19-hydroxylase {ECO:0000303|PubMed:31009114}	13
TRINITY_DN59075_c3_g1	Putative wall-associated receptor kinase-like 16	13
TRINITY_DN57985_c0_g1	F-box protein At5g07610	12
TRINITY_DN58624_c1_g2		12
TRINITY_DN59076_c1_g1	Serine carboxypeptidase-like 20	12
TRINITY_DN53858_c0_g1		11
TRINITY_DN56749_c0_g1		11
TRINITY_DN57792_c3_g1	Pyridoxal 5'-phosphate synthase subunit PDX1	11
TRINITY_DN57936_c1_g1	Beta-galactosidase 6	11
TRINITY_DN58623_c3_g2	Autonomous transposable element EN-1 mosaic protein	11
TRINITY_DN52042_c0_g1		10
TRINITY_DN52814_c0_g1		10
TRINITY_DN53611_c0_g1	Terpene synthase 2; chloroplastic {ECO:0000303|PubMed:27662898; ECO:0000312|EMBL:AAX99149.1}	10
TRINITY_DN54955_c0_g1	Indole-3-acetic acid-induced protein ARG7	10
TRINITY_DN55413_c0_g1		10
TRINITY_DN55564_c0_g1		10
TRINITY_DN55597_c0_g1	Indole-3-acetaldehyde oxidase	10
TRINITY_DN57942_c0_g1	Peptidyl-prolyl cis-trans isomerase FKBP13; chloroplastic {ECO:0000303|PubMed:12424338}	10
TRINITY_DN59272_c1_g2		10
TRINITY_DN52101_c0_g1		9
TRINITY_DN52981_c0_g1		9
TRINITY_DN56774_c0_g1		9
TRINITY_DN58614_c0_g2	UDP-glucuronate:xylan alpha-glucuronosyltransferase 1	9

TRINITY_DN58753_c1_g1	4-hydroxy-tetrahydrodipicolinate synthase; chloroplastic	9
TRINITY_DN59020_c0_g1	PHD finger protein ALFIN-LIKE 8	9
TRINITY_DN60001_c1_g1	Pentatricopeptide repeat-containing protein MRL1; chloroplastic	9
TRINITY_DN52518_c0_g1		8
TRINITY_DN53223_c0_g1		8
TRINITY_DN57285_c0_g2		8
TRINITY_DN57345_c0_g2	ER membrane protein complex subunit 6	8
TRINITY_DN57631_c0_g1		8
TRINITY_DN57874_c0_g1	Phosphatidylinositol N-acetylglucosaminyltransferase subunit P	8
TRINITY_DN59245_c0_g1	Zinc finger SWIM domain-containing protein 7	8
TRINITY_DN51484_c0_g1	Pyruvate; phosphate dikinase 2 {ECO:0000303|PubMed:1668653}	7
TRINITY_DN54046_c1_g1	Probable carboxylesterase 15	7
TRINITY_DN55450_c0_g1	Cytochrome P450 72A15	7
TRINITY_DN57136_c0_g1	Peroxisomal catalase	7
TRINITY_DN59013_c3_g1	Zinc finger BED domain-containing protein RICESLEEPER 1	7
TRINITY_DN59056_c0_g5		7
TRINITY_DN59165_c0_g2		7
TRINITY_DN59169_c1_g1	Ferredoxin-1	7
TRINITY_DN40950_c0_g1		6
TRINITY_DN54038_c1_g1	Calmodulin-interacting protein 111	6
TRINITY_DN54316_c0_g1	Inner membrane protein ALBINO3; chloroplastic	6
TRINITY_DN57035_c0_g1		6
TRINITY_DN57847_c0_g1		6
TRINITY_DN58081_c3_g2	Ubiquitin	6
TRINITY_DN58624_c1_g1		6
TRINITY_DN58972_c0_g1	Cysteine-rich receptor-like protein kinase 28	6
TRINITY_DN59034_c1_g1	Proteasome subunit beta type-7-B	6

TRINITY_DN59037_c1_g1		6
TRINITY_DN59191_c1_g1		6
TRINITY_DN59295_c0_g1		6
TRINITY_DN59704_c1_g1	Transcription initiation factor IIF subunit beta	6
TRINITY_DN53550_c0_g1	Heat shock 70 kDa protein 15	5
TRINITY_DN55158_c0_g1	Pentatricopeptide repeat-containing protein At5g03800	5
TRINITY_DN55332_c1_g2		5
TRINITY_DN57043_c0_g3	Glutamine synthetase cytosolic isozyme 1-1 {ECO:0000305}	5
TRINITY_DN82367_c0_g3		5
TRINITY_DN57989_c0_g1	Polyubiquitin	5
TRINITY_DN58088_c0_g1	DNA mismatch repair protein Msh2	5
TRINITY_DN58221_c0_g1		5
TRINITY_DN58317_c0_g1	SEC14 cytosolic factor	5
TRINITY_DN58627_c0_g1	Cell division control protein 48 homolog C	5
TRINITY_DN58646_c1_g3		5
TRINITY_DN58685_c0_g1		5
TRINITY_DN58997_c2_g2		5
TRINITY_DN59013_c3_g2		5
TRINITY_DN59056_c0_g2		5
TRINITY_DN59150_c0_g1		5
TRINITY_DN72507_c0_g1	Purine nucleoside phosphorylase {ECO:0000250|UniProtKB:P00491}	5
TRINITY_DN84865_c4_g1	Putative linoleate 9S-lipoxygenase 3	5
TRINITY_DN91462_c0_g1	Purple acid phosphatase 2	4
TRINITY_DN77043_c0_g1	ESCRT-related protein CHMP1 {ECO:0000305}	4
TRINITY_DN80008_c0_g3	Protein CONSERVED IN THE GREEN LINEAGE AND DIATOMS 27; chloroplastic {ECO:0000303|PubMed:23043051}	4
TRINITY_DN68918_c1_g1	Pentatricopeptide repeat-containing protein At5g11310; mitochondrial	4

TRINITY_DN88704_c2_g2	Probable anion transporter 5; chloroplastic	4
TRINITY_DN56407_c0_g2	Elongation factor P {ECO:0000255|HAMAP-Rule:MF_00141}	4
TRINITY_DN57054_c0_g2	40S ribosomal protein S15a-2	4
TRINITY_DN57179_c0_g1		4
TRINITY_DN70226_c1_g3		4
TRINITY_DN77132_c1_g3	Receptor-like serine/threonine-protein kinase SD1-8	4
TRINITY_DN80131_c0_g2	HMG-Y-related protein A	4
TRINITY_DN81488_c0_g3	Trigger factor-like protein TIG; Chloroplastic	4
TRINITY_DN84330_c0_g1	Protection of telomeres protein 1a {ECO:0000303|PubMed:17627276}	4
TRINITY_DN57391_c2_g2		4
TRINITY_DN57428_c0_g1	Sterol 14-demethylase	4
TRINITY_DN57429_c3_g3		4
TRINITY_DN89418_c1_g2		4
TRINITY_DN89727_c1_g1	BAG family molecular chaperone regulator 8; chloroplastic	4
TRINITY_DN90739_c2_g1	Chloroplast envelope membrane protein	4
TRINITY_DN57550_c2_g1	60S ribosomal protein L14-2	4
TRINITY_DN65039_c0_g1	Serine carboxypeptidase-like 50	4
TRINITY_DN72749_c1_g1	Alpha-mannosidase I MNS5	4
TRINITY_DN81059_c0_g2	GATA transcription factor 8	4
TRINITY_DN57809_c0_g2		4
TRINITY_DN83148_c2_g3	F-box protein PP2-B10	4
TRINITY_DN78432_c1_g2		4
TRINITY_DN58317_c0_g2	Protein real-time	4
TRINITY_DN66027_c2_g1		4
TRINITY_DN58751_c6_g6	Ubiquitin-40S ribosomal protein S27a-1	4
TRINITY_DN59251_c0_g1		4
TRINITY_DN65997_c0_g3	UPF0481 protein At3g47200	4
TRINITY_DN66393_c0_g1	Glutathione S-transferase U18	4

TRINITY_DN66640_c2_g1		4
TRINITY_DN67579_c0_g2	Putative vesicle-associated membrane protein 726	4
TRINITY_DN70447_c1_g3		4
TRINITY_DN73671_c0_g1	Indole-3-glycerol phosphate lyase; chloroplastic	4
TRINITY_DN76743_c0_g1		4
TRINITY_DN82976_c0_g2		4
TRINITY_DN83055_c0_g3	Monodehydroascorbate reductase	4
TRINITY_DN85601_c0_g1	Amino-acid permease BAT1	4
TRINITY_DN86021_c0_g1		4
TRINITY_DN87062_c0_g2	Cyclin-dependent kinase G-1	4
TRINITY_DN90421_c2_g4		4
TRINITY_DN59889_c0_g1		4
TRINITY_DN65456_c0_g1	Elongation of fatty acids protein 3-like	4
TRINITY_DN68412_c0_g2	Exocyst complex component EXO70B2 {ECO:0000303|PubMed:16942608}	4
TRINITY_DN69503_c1_g1		4
TRINITY_DN69886_c0_g2	Zinc finger protein 1	4
TRINITY_DN87104_c0_g1	Organic cation/carnitine transporter 3	4
TRINITY_DN87710_c2_g3		4
TRINITY_DN91275_c0_g1	Putative MYST-like histone acetyltransferase 1	4
TRINITY_DN70554_c0_g7	Fanconi anemia group D2 protein homolog	4
TRINITY_DN80936_c2_g1	1-aminocyclopropane-1-carboxylate oxidase homolog 11	4
TRINITY_DN75008_c2_g2	Putative glucan endo-1;3-beta-glucosidase GVI	4
TRINITY_DN69278_c1_g3		4

Supplementary Table S13. Wet season co-expression network GO terms enrichment considering first neighbors and second neighbors**.**

GO-ID	Neighbor	Term	Annotated	Significant	Expected	p-value
GO:0051555	first	flavonol biosynthetic process	19	3	0.06	3.20E-05
GO:0010315	first	auxin efflux	40	3	0.13	0.00031
GO:0080168	first	abscisic acid transport	16	2	0.05	0.00125
GO:0009097	first	isoleucine biosynthetic process	23	2	0.08	0.00259
GO:0072710	first	response to hydroxyurea	1	1	0	0.0033
GO:0072718	first	response to cisplatin	1	1	0	0.0033
GO:0010086	first	embryonic root morphogenesis	1	1	0	0.0033
GO:0009830	first	cell wall modification involved in absci...	1	1	0	0.0033
GO:0009624	first	response to nematode	101	3	0.33	0.00452
GO:1903175	first	fatty alcohol biosynthetic process	3	1	0.01	0.00986
GO:1990426	first	mitotic recombination-dependent replicat...	3	1	0.01	0.00986
GO:0006297	first	nucleotide-excision repair, DNA gap fill...	3	1	0.01	0.00986
GO:0045004	first	DNA replication proofreading	3	1	0.01	0.00986
GO:0006287	first	base-excision repair, gap-filling	3	1	0.01	0.00986
GO:0016233	second	telomere capping	9	3	0.26	0.0018
GO:0016584	second	nucleosome positioning	19	4	0.55	0.0019
GO:1990426	second	mitotic recombination-dependent replicat...	3	2	0.09	0.0024
GO:0031365	second	N-terminal protein amino acid modificati...	22	4	0.63	0.0033
GO:0061077	second	chaperone-mediated protein folding	145	10	4.19	0.004
GO:0048235	second	pollen sperm cell differentiation	12	3	0.35	0.0043
GO:2000300	second	regulation of synaptic vesicle exocytosi...	4	2	0.12	0.0048
GO:0048314	second	embryo sac morphogenesis	4	2	0.12	0.0048

GO:0000722	second	telomere maintenance via recombination	4	2	0.12	0.0048
GO:0006148	second	inosine catabolic process	4	2	0.12	0.0048
GO:1900030	second	regulation of pectin biosynthetic proces...	5	2	0.14	0.0078
GO:0000730	second	DNA recombinase assembly	5	2	0.14	0.0078
GO:2000028	second	regulation of photoperiodism, flowering	120	8	3.46	0.0084

Supplementary Table S14. Dry season co-expression network GO terms enrichment considering first neighbors and second neighbors.

GO-ID	Neighbor	Term	Annotated	Significant	Expected	p-value
GO:0051555	first	flavonol biosynthetic process	19	3	0.1	0.00012
GO:0010315	first	auxin efflux	40	3	0.21	0.00113
GO:0060862	first	negative regulation of floral organ absc...	14	2	0.07	0.00228
GO:0042823	first	pyridoxal phosphate biosynthetic process	19	2	0.1	0.00421
GO:0048825	first	cotyledon development	46	3	0.24	0.00505
GO:0036297	first	interstrand cross-link repair	27	2	0.14	0.00842
GO:0048314	second	embryo sac morphogenesis	4	3	0.13	0.00012
GO:1904430	second	negative regulation of t-circle formatio...	2	2	0.06	0.00101
GO:0051351	second	positive regulation of ligase activity	2	2	0.06	0.00101
GO:0070716	second	mismatch repair involved in maintenance ...	2	2	0.06	0.00101
GO:0042138	second	meiotic DNA double-strand break formatio...	17	4	0.54	0.00172
GO:0006303	second	double-strand break repair via nonhomolo...	18	4	0.57	0.00215
GO:0051555	second	flavonol biosynthetic process	19	4	0.6	0.00266
GO:0016584	second	nucleosome positioning	19	4	0.6	0.00266
GO:1990426	second	mitotic recombination-dependent replicat...	3	2	0.1	0.00295
GO:0015031	second	protein transport	1097	48	34.81	0.00298

GO:0000373	second	Group II intron splicing	63	7	2	0.00371
GO:0009414	second	response to water deprivation	507	30	16.09	0.00509
GO:0048235	second	pollen sperm cell differentiation	12	3	0.38	0.00564
GO:0010315	second	auxin efflux	40	5	1.27	0.00831
GO:0051103	second	DNA ligation involved in DNA repair	5	2	0.16	0.00943
GO:0006627	second	protein processing involved in protein t...	5	2	0.16	0.00943
GO:0000730	second	DNA recombinase assembly	5	2	0.16	0.00943
GO:0043007	second	maintenance of rDNA	5	2	0.16	0.00943
GO:0010200	second	response to chitin	43	6	1.36	0.00977
GO:0036297	second	interstrand cross-link repair	27	4	0.86	0.00986

RESUMO DOS RESULTADOS

CAPÍTULO I

(1) Desenvolvemos uma metodologia para identificação de contaminantes oriundos de apomixia, autofecundação ou cruzamento indesejado, em populações biparentais de espécies tetra/hexaploides. A metodologia foi capaz de identificar 100% dos contaminantes em populações simuladas com 200 indivíduos e genotipados com pelo menos 689 marcadores moleculares;

(2) Aplicamos a metodologia para a identificação de contaminantes em três populações reais de forrageiras tropicais. Nas populações tetraplóides de *U. decumbens* e *M. maximus*, foram identificados 2 e 52 amostras como clones apomíticos. Já na população hexaploide de *U. humidicola*, foram identificadas 65 amostras como clones apomíticos;

(3) Implementamos a metodologia em um R Shiny web app com visualização gráfica chamado PolyCID.

CAPÍTULO II

(1) Comparamos modelos convencionais e de aprendizado de máquina para predição genômica em famílias de meios irmãos de características relacionadas a crescimento e produção de biomassa. O melhor modelo (RKHS) apresentou uma média aproximada de 0.76 de habilidade preditiva utilizando o conjunto completo de marcadores;

(2) A seleção de features foi capaz de reduzir o conjunto de marcadores necessários para fazer as predições, e aumentou a habilidade preditiva dos modelos. O melhor modelo (RF) obteve uma média aproximada de 0.89 de habilidade preditiva utilizando conjuntos reduzidos de aproximadamente 11 marcadores;

(3) Utilizando a métrica Importância Gini do modelo RF. selecionamos 69 marcadores de maior importância como associados as características agronômicas preditas;

(4) Identificamos 217 genes (obtidos no transcriptoma) fisicamente ligados aos 69 marcadores selecionados, e construímos um mapa físico relacionando os marcadores, genes, características fenotípicas e estações. Do total de genes identificados, 100 tiveram anotação funcional;

(5) O enriquecimento dos termos GO dos genes identificados mostrou que eles estão relacionados principalmente aos processos biossintéticos de lignina, flavonol e ácido fólico, no transporte de auxina e no processo metabólico de galactolipídeos, vias metabólicas que atuam no crescimento e produção de biomassa;

(6) Utilizando o transcriptoma, construímos uma rede global de co-expressão, e isolamos os genes previamente identificados junto aqueles co-expressos como primeiros e segundos vizinhos. A rede dos genes associados às características de interesse possui 2704 genes (nós) e 3453 arestas;

(7) O enriquecimento dos termos GO dos genes presentes na rede de co-expressão isolada mostrou que, além das vias metabólicas já citadas, vários genes estão relacionados ao metabolismo e manutenção de DNA, resposta a estresses bióticos e regulação do ciclo circadiano;

(8) Identificamos genes na rede com alto grau de conexões indicando sua atuação em conjunto com muitos outros genes, como proteínas ribossomais que influenciam na expressão de outros genes, proteína ELF4-LIKE 4 associada ao ciclo circadiano e proteína 14-3-3 zeta associada a crescimento, desenvolvimento e resposta a estresses;

(9) Separando a rede em temporada de seca e chuva, identificamos termos enriquecidos compartilhados relacionados à biossíntese de flavonol, transporte de auxina e replicação de DNA durante a mitose. Foram identificados termos específicos para a temporada de chuva relacionados a transporte de ácido abscísico, biossíntese de isoleucina e resposta a nematóides, enquanto para temporada de seca, termos relacionados ao transporte de proteínas, biossíntese de fosfato piridoxal e a resposta a quitina e à privação de água.

CONCLUSÃO

Apesar do surgimento das tecnologias de sequenciamento de nova geração e a consequente redução drástica nos custos de sequenciamento de DNA e RNA que aconteceu no fim da década de 2000, programas de melhoramento de espécies órfãs que possuem genomas de alta complexidade, como as forrageiras tropicais, ainda passam por dificuldades para implementar o uso desses dados no melhoramento e desenvolver novas variedades com maior ganho genético. Nesse sentido, esta tese teve como objetivo investigar o uso desses dados em forrageiras tropicais, especialmente espécies do gênero *Urochloa*, para produzir métodos, ferramentas e conhecimentos que auxiliem suas aplicações nos programas de melhoramento.

Quanto à metodologia e a ferramenta (polyCID) desenvolvidas no **Capítulo I** para a identificação de contaminantes em populações biparentais tetra/hexaplóide, concluímos que marcadores do tipo SNP, desde que genotipados em dosagens alélicas e em quantidade suficiente, são eficientes para a identificação de contaminantes. Além disso, a ferramenta proporciona uma análise de fácil execução e entendimento que a depender do tamanho do conjunto de dados pode ser realizada em poucos minutos.

No **Capítulo II**, realizamos uma análise multi ômica com dados de *U. ruziziensis*, integrando métodos como predição genômica e redes de co-expressão. Concluímos que GWFP é um método aplicável a espécie, e que técnicas de aprendizado de máquina integradas com seleção de features são capazes de melhorar as métricas de predição e reduzir o número de marcadores necessários para a predição, evidenciando que a abordagem representa uma grande redução de custos para o programa de melhoramento. Além disso, os marcadores selecionados podem ser considerados como possíveis QTL e utilizados para a identificação de genes putativos com influência no fenótipo. Com a integração dos genes identificados e a rede de co-expressão modelada com dados de transcriptoma, é possível através dos módulos de co-expressão estender a investigação para vias metabólicas mais amplas e inferir funções para genes não anotados.

Em relação a tese como um todo, a principal conclusão do trabalho é que muitos dos métodos e ferramentas do melhoramento molecular, desenvolvidos e amplamente aplicados em plantas modelo, com as devidas adaptações podem ser utilizados no melhoramento de espécies órfãs com genomas de alta complexidade como as forrageiras tropicais. Os resultados apresentados nesta tese e em outros trabalhos fornecem um arcabouço de conhecimento capaz de ampliar significativamente o uso do melhoramento molecular nos programas de melhoramento dessas espécies.

PERSPECTIVAS

Com o desenvolvimento da ferramenta polyCID para a identificação de contaminantes, é esperado que inicialmente ela auxilie programas de melhoramento que estejam trabalhando na construção de mapas genéticos de ligação. A metodologia foi criada para lidar com dados do tipo SNP com informação de dosagem alélica, que são os dados necessários para a construção de mapas genéticos em poliploides. Com a contínua redução dos custos para a genotipagem, espera-se que no futuro os programas consigam fazer a análise de forma rotineira em seus cruzamentos, reduzindo as perdas no ganho genético devido a presença de contaminantes.

Em relação a investigação da aplicabilidade de GWFP em *U. ruziziensis* para as características agronômicas de crescimento e produção de biomassa, os resultados foram promissores e, consequentemente, a abordagem poderia ser testada em outras características, especialmente as de valor nutricional, as quais essa mesma população já foi fenotipada. A espécie é conhecida por ter alta qualidade nessas características, o que a torna ainda mais importante, servindo para cruzamentos interespecíficos com espécies mais produtivas. Espera-se que com resultados como esses, os programas se sintam mais confiantes em de fato implementar a metodologia em seus programas.

REFERÊNCIAS BIBLIOGRÁFICAS

ABIEC 2022 - Associação brasileira das indústrias exportadoras de carne. Beef report: overview of livestock in Brazil.

Almeida, M. C. D. C., Chiari, L., Jank, L., and Valle, C. B. D. (2011). Diversidade genética molecular entre cultivares e híbridos de Brachiaria spp. e Panicum maximum. Ciênc. Rural 41, 1998–2003. doi: 10.1590/S0103-84782011001100024

Amadeu, R. R., Lara, L. A. C., Munoz, P., and Garcia, A. A. F. (2020). Estimation of molecular pairwise relatedness in autopolyploid crops. G3 Genes Genomes Genet. 10, 4579–4589. doi: 10.1534/g3.120.401669

Anderson, E. C. (2012). Large-scale parentage inference with SNPs: an efficient algorithm for statistical confidence of parent pair allocations. Stat. Appl. Genet. Mol. Biol. 11:296–302. doi: 10.1515/1544-6115.1833

Aono, A.H., Costa, E.A., Rody, H.V.S. et al. Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. Sci Rep 10, 20057 (2020). https://doi.org/10.1038/s41598-020-77063-5

Aono, A. H., Ferreira, R., Moraes, A., Lara, L., Pimenta, R., Costa, E. A., Pinto, L. R., Landell, M., Santos, M. F., Jank, L., Barrios, S., do Valle, C. B., Chiari, L., Garcia, A., Kuroshu, R. M., Lorena, A. C., Gorjanc, G., & de Souza, A. P. (2022). A joint learning approach for genomic prediction in polyploid grasses. Scientific reports, 12(1), 12499. https://doi.org/10.1038/s41598-022-16417-7

Azevedo, A. L. S., Costa, P. P., Machado, M. A., de Paula, C. M. P., & Sobrinho, F. S. (2011). High degree of genetic diversity among genotypes of the forage grass Brachiaria ruziziensis (Poaceae) detected with ISSR markers. In Genetics and Molecular Research (Vol. 10, Issue 4, pp. 3530–3538). Genetics and Molecular Research. https://doi.org/10.4238/2011.november.17.5

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., & Shiu, S.-H. (2019-A). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. In G3 Genes|Genomes|Genetics (Vol. 9, Issue 11, pp. 3691–3702). Oxford University Press (OUP). https://doi.org/10.1534/g3.119.400498

Balbino, L. C.; Barcellos, A. O.; Stone, L. F. Marco referencial: integração lavoura-pecuária-floresta. 1. ed. Brasília, DF: Embrapa, 2011. 130 p.

Balbinot Jr., A. A. et al. Integração lavoura-pecuária: intensificação de uso de áreas agrícolas. Ciência Rural, 39, p. 1925-1933, 2009.

Bateman, A. J. (1947). Contamination of seed crops. J. Genet. 48, 257–275. doi: 10.1007/BF02989385

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Science, 34(1), 20-25

Bernardo, R. (2016). Bandwagons I, too, have known. Theoretical and applied genetics, 129(12), 2323-2332.

Bernini, C., and Marin-Morales, M. A. (2001). Karyotype analysis in Brachiaria (Poaceae) species. Cytobios 104, 157–171.

Bicknell, R. A. (2004). Understanding apomixis: recent advances and remaining conundrums. Plant Cell Online 16, S228–S245. doi: 10.1105/tpc.017921

Braga, I., Yamamoto, C. J. T., Custódio, C. C., and Machado-Neto, N. B. (2017). Differentiation of Urochloa brizantha cultivars by inter-simple sequence repeat (ISSR) markers in seed samples. Afr. J. Biotechnol. 16, 607–614. doi: 10.5897/AJB2016.15638

Borin, G. P., Carazzolle, M. F., Dos Santos, R., Riaño-Pachón, D. M., & Oliveira, J. (2018). Gene Co-expression Network Reveals Potential New Genes Related to Sugarcane Bagasse Degradation in Trichoderma reesei RUT-30. Frontiers in bioengineering and biotechnology, 6, 151. https://doi.org/10.3389/fbioe.2018.00151

Cai, J., Luo, J., Wang, S. & Yang, S. Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79 (2018).

Campante, P. A glance at the Brazilian seed market. Agropages: 2018 seed special. p. 10-12. 2018. Disponível em: http://news.agropages.com/News/NewsDetail---26900.htm; Acesso em: 30 out.2018.

Cao, P., Zhao, Y., Wu, F., Xin, D., Liu, C., Wu, X., Lv, J., Chen, Q., & Qi, Z. (2022). Multi-Omics Techniques for Soybean Molecular Breeding. In International Journal of Molecular Sciences (Vol. 23, Issue 9, p. 4994). MDPI AG. https://doi.org/10.3390/ijms23094994

Cardoso-Silva, C. B., Aono, A. H., Mancini, M. C., Sforça, D. A., da Silva, C. C., Pinto, L. R., Adams, K. L., & de Souza, A. P. (2022). Taxonomically Restricted Genes Are Associated With Responses to Biotic and Abiotic Stresses in Sugarcane (Saccharum spp.). Frontiers in plant science, 13, 923069. https://doi.org/10.3389/fpls.2022.923069

Chiari, L., da Rocha, M., do Valle, C. B., and Salgado, L. R. (2008). Variabilidade genética em acessos e cultivares de quatro espécies de Brachiaria estimada por marcadores RAPD. Boletim de Pesquisa e Desenvolvimento n°. 24. Embrapa Beef Cattle, Campo Grande-MS.

Childs, K. L., Davidson, R. M., and Buell, C. R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One 6:e22196. doi: 10.1371/journal.pone.0022196

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22, 961–975. doi: 10.1016/j.tplants. 2017.08.011

Deo, T. G., Ferreira, R., Lara, L., Moraes, A., Alves-Pereira, A., de Oliveira, F. A., Garcia, A., Santos, M. F., Jank, L., and de Souza, A. P. (2020). High-resolution linkage map with allele dosage allows the identification of regions governing complex traits and apospory in Guinea grass (Megathyrsus maximus). Front. Plant Sci. 11, 15. doi: 10.3389/fpls.2020.00015

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science, 19(9), 592–601. https://doi.org/10.1016/j.tplants.2014.05.006

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707–726. doi: 10.1093/bioinformatics/16.8.707

Dhingani RM, Umrania VV, Tomar RS, et al. Introduction to QTL mapping in plants. Ann Plant Sci. 2015;4(04):1072–1079.

Dias-Filho, M.B. Uso de pastagens para a produção de bovinos de corte no Brasil: passado, presente e futuro. Documentos no. 418. Embrapa Amazônia Oriental, Belém, 2016.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. doi: 10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. In The Plant Genome (Vol. 4, Issue 3, pp. 250–255). Wiley. https://doi.org/10.3835/plantgenome2011.08.0024

Euclides, V. P. B.; Valle, C. B.; Macedo, M. C. M.; Almeida, R. G.; Montagner, D. B.; Barbosa, R. A. (2010). Brazilian scientific progress in pasture research during the first decade of XXI century. Revista Brasileira de Zootecnia 39: 151-168.

Ferreira, R. C. U., Cançado, L. J., Do Valle, C. B., Chiari, L., and de Souza, A. P. (2016). Microsatellite loci for Urochloa decumbens (Stapf) R.D. Webster and cross-amplification in other Urochloa species. BMC. Res. Notes 9:152. doi: 10.1186/s13104-016-1967-9

Ferreira, R. C. U., Costa Lima Moraes, A. da, Chiari, L., Simeão, R. M., Vigna, B. B. Z., & de Souza, A. P. (2021). An Overview of the Genetics and Genomics of the Urochloa Species Most Commonly Used in Pastures. In Frontiers in Plant Science (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fpls.2021.770461

Ferreira, R. C. U., Lara, L. A. D. C., Chiari, L., Barrios, S. C. L., do Valle, C. B., Valério, J. R., et al. (2019). Corrigendum: genetic mapping with allele dosage information in tetraploid Urochloa decumbens (Stapf) R. D. Webster reveals insights into spittlebug (Notozulia entreriana Berg) resistance. Front. Plant Sci. 10:92. doi: 10.3389/fpls.2019.00855

Figueiredo U.J. (2011). Estimação de parâmetros genéticos e fenotípicos emprogênies de Brachiaria humidicola. UFLA. Genética e Melhoramento de Plantas. Lavras. Dissertação. 75p.

Figueiredo, U. J. de, Nunes, J. A. R., & Valle, C. B. do. (2012). Estimation of genetic parameters and selection of Brachiaria humidicola progenies using a selection index. Crop Breeding and Applied Biotechnology, 12(4), 237–244. https://doi.org/10.1590/s1984-70332012000400002

Francisco, F. R., Aono, A. H., da Silva, C. C., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., Fritsche-Neto, R., Souza, L. M., & de Souza, A. P. (2021). Unravelling Rubber Tree Growth by Integrating GWAS and Biological Network-Based Approaches. Frontiers in plant science, 12, 768589. https://doi.org/10.3389/fpls.2021.768589

Garcia, M., Vigna, B. B. Z., Sousa, A. C. B., Jungmann, L., Cidade, F. W., Toledo-Silva, G., et al. (2013). Molecular genetic variability, population structure and mating system in tropical forages. Trop. Grass – Forr Trop 1, 25–30. doi: 10.17138/TGFT(1)25-30

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. Journal of Animal Breeding and Genetics, 124(6), 323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x

Grashei, K. E., Ødegård, J., and Meuwissen, T. H. E. (2018). Using genomic relationship likelihood for parentage assignment. Genet. Sel. Evol. 50:26. doi: 10.1186/s12711-018-0397-7

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., et al. (2011). Current trends in microsatellite genotyping. Mol. Ecol. Resour. 11, 591–611. doi: 10.1111/j.1755-0998.2011.03014.x

Hanley, S. J., Pellny, T. K., de Vega, J. J., Castiblanco, V., Arango, J., Eastmond, P. J., Heslop-Harrison, J. S. P., & Mitchell, R. A. C. (2021). Allele mining in diverse accessions of tropical grasses to improve forage quality and reduce environmental impact. Annals of

Botany, 128(5), 627–637. https://doi.org/10.1093/aob/mcab101 Zootecnia, v. 39, p. 151-168. 2010.

Hayes, B. J. (2011). Technical note: efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J. Dairy Sci. 94, 2114–2117. doi: 10.3168/jds.2010-3896

Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., et al. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. PLoS ONE 9:e94851. doi: 10.1371/journal.pone.0094851

Heer, K., Behringer, D., Piermattei, A., Bässler, C., Brandl, R., Fady, B., Jehl, H., Liepelt, S., Lorch, S., Piotti, A., Vendramin, G. G., Weller, M., Ziegenhagen, B., Büntgen, U., & Opgenoorth, L. (2018). Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (Abies albaMill.). In Molecular Ecology (Vol. 27, Issue 6, pp. 1428–1438). Wiley. https://doi.org/10.1111/mec.14538

Hill, M.C., Kadow, Z.A., Long, H. et al. Integrated multi-omic characterization of congenital heart disease. Nature 608, 181–191 (2022). https://doi.org/10.1038/s41586-022-04989-3

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., et al. (2016). The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. Appl. Plant Sci. 4:1600025. doi: 10.3732/apps.1600025

Huang, K., Guo, S. T., Shattuck, M. R., Chen, S. T., Qi, X. G., Zhang, P., et al. (2015). A maximum-likelihood estimation of pairwise relatedness for autopolyploids. Heredity 114, 133–142. doi: 10.1038/hdy.2014.88

Huisman, J. (2017). Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. Mol. Ecol. Resour. 17, 1009–1024. doi: 10.1111/1755-0998.12665

Jank, L.; Barrios, S. C.; Valle, C. B.; Simeão, R. M.; Alves, G. F. The value of improved pastures to Brazilian beef production. Crop and Pasture Science, v. 65, p. 1132-1137. 2014.

Jank, L.; Carvalho P. F.; Valle, C. B. New grasses and legumes: advances and perspectives for the tropical zones of Latin America. In: Reynolds SG & Frame J (Org.) Grasslands: developments, opportunities, perspectives. Roma, FAO, India, Science Publishers. p. 55-79. 2005.

Jha, N. K., Jacob, S. R., Nepolean, T., Jain, S. K., and Kumar, M. B. A. (2016). SSR markers based DNA fingerprinting and it's utility in testing purity of eggplant hybrid seeds. Qual. Assur. Saf. Crops Foods 8, 333–338. doi: 10.3920/QAS2015.0689

Jones, C., De Vega, J., Worthington, M., Thomas, A., Gasior, D., Harper, J., et al. (2021). A comparison of differential gene expression in response to the onset of water stress between three hybrid Brachiaria genotypes. Front. Plant Sci. 12:637956. doi: 10.3389/fpls.2021.637956

Jones, O. R., and Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol. Ecol. Resour. 10, 551–555. doi: 10.1111/j.1755-0998.2009.02787.x

Jungmann, L., Sousa, A. C. B., Paiva, J., Francisco, P. M., Vigna, B. B. Z., and do Valle, C. B. (2009b). Isolation and characterization of microsatellite markers for Brachiaria brizantha (Hochst. Ex A. Rich.). Stapf. Conserv. Genet. 10, 1873–1876. doi: 10.1007/s10592-009-9839-7

Jungmann, L., Vigna, B. B. Z., Boldrini, K. R., Sousa, A. C. B., do Valle, C. B., Resende, R. M. S., et al. (2010). Genetic diversity and population structure analysis of the tropical pasture grass Brachiaria humidicola based on microsatellites, cytogenetics, morphological traits, and geographical origin. Genome 53, 698–709. doi: 10.1139/G10-055

Jungmann, L., Vigna, B. B. Z., Paiva, J., Sousa, A. C. B., do Valle, C. B., Laborda, P. R., et al. (2009a). Development of microsatellite markers for Brachiaria humidicola (Rendle) Schweick. Conserv. Genet. Resour. 1, 475–479. doi: 10.1007/s12686-009-9111-y

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16, 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Keller- Grein, G.; Maass, B.L.; Hanson, J. Natural variation in Brachiaria and existing germplasm collections. In J.W. Miles et al. (ed.) Brachiaria: Biology, agronomy, and improvement. CIAT, Cali, Colombia, and CNPGC/EMBRAPA, Campo Grande, MS, Brazil, p. 16-42, 1996.

Kemble, H., Nghe, P., and Tenaillon, O. (2019). Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. Evol. Appl. 12, 1721–1742. doi: 10.1111/eva.12846

Knoch, D., Werner, C.R., Meyer, R.C. et al. Multi-omics-based prediction of hybrid performance in canola. Theor Appl Genet 134, 1147–1165 (2021). https://doi.org/10.1007/s00122-020-03759-x

Kuwi, S. O., Kyalo, M., Mutai, C. K., Mwilawa, A., Hanson, J., Djikeng, A., et al. (2018). Genetic diversity and population structure of Urochloa grass accessions from Tanzania using simple sequence repeat (SSR) markers. Braz. J. Bot. 41, 699–709. doi: 10.1007/s40415-018-0482-8

Li, X., Yang, Y., Yao, J. et al. FLEXIBLE CULM 1 encoding a cinnamyl-alcohol dehydrogenase controls culm mechanical strength in rice. Plant Mol Biol 69, 685–697 (2009). https://doi.org/10.1007/s11103-008-9448-8

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y. C., Cheng, F., et al. (2020). Computational network biology: data, models, and applications. Phys. Rep. 846, 1–66. doi: 10.1016/j.physrep.2019.12.004

Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569, 655–662 (2019). https://doi.org/10.1038/s41586-019-1237-9

Lopes FCF Paciullo DSC, Mota EF, Pereira JC, Azambuja AA, Motta ACS, Rodrigues GS and Duque ACA (2010) Composição química e digestibilidade ruminal in situ da forragem de quatro espécies do gênero Brachiaria. Arq. Bras. Med. Vet. Zootec. 62(4): 883-888.

Luo, Z., Yu, Y., Xiang, J. & Li, F. Genomic selection using a subset of snps identified by genome-wide association analysis for disease resistance traits in aquaculture species. Aquaculture 539, 736620 (2021).

Macedo, M.C.M. Aspectos edáficos relacionados com a produção de Brachiaria brizantha cultivar Marandu. In: Barbosa, R.A (ed.) Morte de pastos de braquiárias. Campo Grande/MS: Embrapa Gado de Corte, pp.35-65

Mateescu, R. G., Garrick, D. J., & Reecy, J. M. (2017). Network Analysis Reveals Putative Genes Affecting Meat Quality in Angus Cattle. Frontiers in genetics, 8, 171. https://doi.org/10.3389/fgene.2017.00171

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Endelman, J. B., et al. (2019B). On the accuracy of genomic prediction models considering multi-trait and allele dosage in Urochloa spp. interspecific tetraploid hybrids. Mol. Breed. 39:100. doi: 10.1007/s11032-019-1002-7

Matias, F. I., Vidotti, M. S., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Carley, C. A. S., et al. (2019a). Association mapping considering allele dosage: an example of forage traits in an interspecific segmental allotetraploid Urochloa spp. panel. Crop Sci. 59, 2062–2076. doi: 10.2135/cropsci2019.03.0185

Mendonça S.A. (2012). Avaliação agronômica e modo de reprodução de híbridos intraespecíficos de Brachiaria decumbens. UEP. Zootecnia. Botucatu. Dissertação. 51p.

Meuwissen T H, Hayes B J, Goddard M E, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Miao, J. & Niu, L. A survey on feature selection. Procedia Comput. Sci. 91, 919–926 (2016).

Miles, J. W. Apomixis for cultivar development in tropical forage grasses. Crop Science, 47:S238-S249. 2007.

Miles, J.W.; Valle, C.B. Manipulation of apomixis in Brachiaria breeding. In: Miles JW, Mass BL, Valle CB (eds.). Brachiaria: biology, agronomy and improvement. Colombia: CIAT. p. 164-177, 1996.

Mohan M, Nair S, Bhagwat A, et al. Genome mapping, molecular markers and marker-assisted selection in crop plants. Mol Breed. 1997;3(2):87–103.

Mollinari, M., Olukolu, B. A., Pereira, G. D. S., Khan, A., Gemenet, D., Yencho, G. C., et al. (2020). Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. G3 Genes Genomes Genet. 10, 281–292. doi: 10.1534/g3.119.400620

Montesinos-López, O.A., Montesinos-López, A., Pérez-Rodríguez, P. et al. A review of deep learning applications for genomic selection. BMC Genomics 22, 19 (2021). https://doi.org/10.1186/s12864-020-07319-x006.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., & Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell, 21(8), 2194-2202.

Nakaya, A., & Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? Annals of Botany, 110(6), 1303–1316. https://doi.org/10.1093/aob/mcs109

Namazzi, C., Sserumaga, J. P., Mugerwa, S., Kyalo, M., Mutai, C., Mwesigwa, R., et al. (2020). Genetic diversity and population structure of Brachiaria (syn Urochloa) ecotypes from Uganda. Agronomy 10:1193. doi: 10.3390/agronomy10081193

Nunes, S.G.; Boock, A.; Penteado, M.I.O.; Gomes, D.T. Brachiaria brizantha cv. Marandu. CNPGC/EMBRAPA, Campo Grande, MS, Brasil. 31p, 1984.

Oliver, S. (2000) Guilt-by-association goes global. Nature, 403, 601–602. https://doi.org/10.1038/35001165

Ould Estaghvirou, S., Ogutu, J. O., Schulz-Streeck, T., Knaak, C., Ouzunova, M., Gordillo, A., & Piepho, H.-P. (2013). Evaluation of approaches for estimating the accuracy of

genomic prediction in plant breeding. BMC Genomics, 14(1), 860. https://doi.org/10.1186/1471-2164-14-860

Parker Gaddis, K. L., Null, D. J., & Cole, J. B. (2016). Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. Journal of dairy science, 99(8), 6420–6435. https://doi.org/10.3168/jds.2015-10444

Patella, A., Palumbo, F., Galla, G., and Barcaccia, G. (2019). The molecular determination of hybridity and homozygosity estimates in breeding populations of lettuce (Lactuca sativa L.). Genes 10:916. doi: 10.3390/genes10110916

Penteado, M.I.; Santos, A.C.; Rodrigues, I.F.; Valle, C.B.; Seixas, M.A; Esteves A. (2000) Determinação de ploidia e avaliação da quantidade de DNA total em diferentes espécies do gênero Brachiaria. Boletim de Pesquisa no. 11. Embrapa Gado de Corte, Campo Grande.

Pérez, P., & de los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. In Genetics (Vol. 198, Issue 2, pp. 483–495). Oxford University Press (OUP). https://doi.org/10.1534/genetics.114.164442

Pessoa-Filho, M., Azevedo, A. L. S., Sobrinho, F. S., Gouvea, E. G., Martins, A. M., and Ferreira, M. E. (2015). Genetic diversity and structure of Ruzigrass germplasm collected in Africa and Brazil. Crop Sci. 55, 2736–2745. doi: 10.2135/cropsci2015.02.0096 (Accessed 30, 2021).

Pessoa-Filho, M., Martins, A.M. & Ferreira, M.E. Molecular dating of phylogenetic divergence between Urochloa species based on complete chloroplast genomes. BMC Genomics 18, 516 (2017). https://doi.org/10.1186/s12864-017-3904-2

Pessoa-Filho, M., Sobrinho, F. S., Fragoso, R. R., Silva Junior, O. B., and Ferreira, M. E. (2019). “A Phased Diploid Genome Assembly for the Forage Grass Urochloa Ruziziensis Based on Single-Molecule Real-Time Sequencing.” in International Plant and Animal Genome Conference XXVII, 2019, San Diego. Available at: https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1107378/a-phased-diploid-genome-assembly-for-the-forage-grass-urochloa-ruziziensis-based-on-single-molecule-real-time-sequencing.

Piles, M., Bergsma, R., Gianola, D., Gilbert, H., & Tusell, L. (2021). Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. In Frontiers in Genetics (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fgene.2021.611506

Pimenta, R.J.G., Aono, A.H., Burbano, R.C.V. et al. Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. Sci Rep 11, 15730 (2021). https://doi.org/10.1038/s41598-021-95116-1

Pimenta, R. J. G., Aono, A. H., Burbano, R. C. V., da Silva, M. F., dos Anjos, I. A., de Andrade Landell, M. G., Gonçalves, M. C., Pinto, L. R., & de Souza, A. P. (2022). Multiomic investigation of sugarcane mosaic virus resistance in sugarcane. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2022.08.18.504288

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253. doi: 10.1371/journal.pone,0.0032253

Rao, X., & Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. Acta biochimica et biophysica Sinica, 51(10), 981–988. https://doi.org/10.1093/abbs/gmz080

Renvoize, S.A.; Clayton, W.D.; Kabuye, C.H.S. Morphology,taxonomy and natural dis tribution 01 Brachiaria (Tr in .) Griseb. In: Miles, J. W.; Maass, B.L.; Valle, C.B. do, ed. 8rachiaria: Biology, Agronomy and Improvement. Cali:CIAT/Brasília: EMBRAPA-CNPGC, 1996. p .1-15.

Richardson, S., Tseng, G. C. & Sun, W. Statistical methods in integrative genomics. Annu. Rev. Stat. Appl. 3, 181–209 (2016).

Rios, E. F., Andrade, M. H. M. L., Resende, M. F. R., Jr, Kirst, M., de Resende, M. D. V., de Almeida Filho, J. E., Gezan, S. A., & Munoz, P. (2021). Genomic prediction in family bulks using different traits and cross-validations in pine. In A. E. Lipka (Ed.), G3 Genes|Genomes|Genetics (Vol. 11, Issue 9). Oxford University Press (OUP). https://doi.org/10.1093/g3journal/jkab249

Risso-Pascotto, C., Pagliarini, M. S., and do Valle, C. B. (2005). Meiotic behavior in interspecific hybrids between Brachiaria ruziziensis and Brachiaria brizantha (Poaceae). In Euphytica Vol. 145 (Springer Science and Business Media LLC.), 155–159.

Risso-Pascotto, C., Pagliarini, M. S., and Valle, C. B. (2003). A mutation in the spindle checkpoint arresting meiosis II in Brachiaria ruziziensis. Genome 46, 724–728. doi: 10.1139/g03-037

Rosolen, R. R., Aono, A. H., Almeida, D. A., Ferreira Filho, J. A., Horta, M., & De Souza, A. P. (2022). Network Analysis Reveals Different Cellulose Degradation Strategies

Across Trichoderma harzianum Strains Associated With XYR1 and CRE1. Frontiers in genetics, 13, 807243. https://doi.org/10.3389/fgene.2022.807243

Ryschawy, J. et al. Mixed crop-livestock systems: An economic and environmental-friendly way of farming? Animal, v. 6, p. 1722-1730, 2012.

Salgado, L. R., Lima, R., Santos, B. F. D., Shirakawa, K. T., Vilela, M. D. A., Almeida, N. F., et al. (2017). De novo RNA sequencing and analysis of the transcriptome of signalgrass (Urochloa decumbens) roots exposed to aluminum. Plant Growth Regul. 83, 157–170. doi: 10.1007/s10725-017-0291-2

Salton, J. C. et al. Integrated crop-livestock system in tropical Brazil: Toward a sustainable production system. Agriculture, Ecosystems & Environment, v. 190, p. 70-79, 2014.

Sanderson, M. A. et al. Diversification and ecosystem services for conservation agriculture: Outcomes from pastures and integrated crop–livestock systems. Renewable Agriculture and Food Systems, v. 28, p. 129-144, 2013.

Santos, J. C. S., Barreto, M. A., Oliveira, F. A., Vigna, B. B. Z., and Souza, A. P. (2015). Microsatellite markers for Urochloa humidicola (Poaceae) and their transferability to other Urochloa species. BMC. Res. Notes 8:83. doi: 10.1186/s13104-015-1044-9

Santos, J. M. D., Barbosa, G. V. D. S., Neto, C. E. R., and Almeida, C. (2014). Efficiency of biparental crossing in sugarcane analyzed by SSR markers. Crop Breed. Appl. Biotechnol. 14, 102–107. doi: 10.1590/1984-70332014v14n2a18

Schulz-Streeck, T., Ogutu, J. O., Karaman, Z., Knaak, C., & Piepho, H. P. (2012). Genomic Selection using Multiple Populations. In Crop Science (Vol. 52, Issue 6, pp. 2453–2461). Wiley. https://doi.org/10.2135/cropsci2012.03.0160

Scossa, F., Alseekh, S., & Fernie, A. R. (2021). Integrating multi-omics data for crop improvement. In Journal of Plant Physiology (Vol. 257, p. 153352). Elsevier BV. https://doi.org/10.1016/j.jplph.2020.153352

Seiffert NF (1980) Gramíneas Forrageiras do gênero Brachiaria. Circular Técnica n° 1, Embrapa-CNPGC, jan. 1980, ed. 1984, Campo Grande - MS http://www.cnpgc.embrapa.br/publicacoes/ct/ct01/index.html acesso 02 abril 2015.

Silva, P.I., Martins, A.M., Gouvea, E.G. et al. Development and validation of microsatellite markers for Brachiaria ruziziensisobtained by partial genome assembly of Illumina single-end reads. BMC Genomics 14, 17 (2013). https://doi.org/10.1186/1471-2164-14-17

Simeão, R., Silva, A., Valle, C., Resende, M. D., and Medeiros, S. (2016). Genetic evaluation and selection index in tetraploid Brachiaria ruziziensis. Plant Breed. 135, 246–253. doi: 10.1111/pbr.12353

Simeão-Resende, R. M., Casler, M. D., and Resende, M. D. V. (2014). Genomic selection in forage breeding: accuracy and methods. Crop Sci. 54, 143–156. doi: 10.2135/cropsci2013.05.0353

Simioni, C.; Valle, C.B. Chromosome duplication in Brachiaria (A.Rich.) Stapf allows intraspecific crosses. Crop Breeding and Applied Biotechnology. v.9, p.328 - 334, 2009.

Smith, R. L. (1972). Sexual reproduction in Panicum maximum Jacq. Crop Sci. 12, 624–627. doi: 10.2135/cropsci1972.0011183X001200050021x

Souza, J. S., Chiari, L., Simeão, R. M., de Mendonça Vilela, M., & Salgado, L. R. (2018). Development, Validation and Characterization of Genic Microsatellite Markers. In American Journal of Plant Sciences (Vol. 09, Issue 02, pp. 281–295). Scientific Research Publishing, Inc. https://doi.org/10.4236/ajps.2018.92023

Souza Sobrinho F, Auad AM and Lédo FJS (2010) Genetic variability in Brachiaria ruziziensis for resistance to spittlebugs. Crop Breeding and Applied Biotechnology 10: 83-88.

Spielmann, A., Harris, S. A., Boshier, D. H., and Vinson, C. C. (2015). Orchard: paternity program for autotetraploid species. Mol. Ecol. Resour. 15, 915–920. doi: 10.1111/1755-0998.12370

Steinfath, M., Gärtner, T., Lisec, J. et al. Prediction of hybrid biomass in Arabidopsis thaliana by selected parental SNP and metabolic markers. Theor Appl Genet 120, 239–247 (2010). https://doi.org/10.1007/s00122-009-1191-2

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. In Bioinformatics and Biology Insights (Vol. 14, p. 117793221989905). SAGE Publications. https://doi.org/10.1177/1177932219899051

Swenne, A.; Louant, B-P.; Dujardin, M. Induction par la colchicine de formes autotétraploïdes chez Brachiaria ruziziensis Germain et Evrard (Graminée). Agronomie Tropicale, v. 36, p. 134-141, 1981.

Tegegn, A., Kyalo, M., Mutai, C., Hanson, J., Asefa, G., Djikeng, A., et al. (2019). Genetic diversity and population structure of Brachiaria brizantha (A.Rich.) Stapf accessions from Ethiopia. Afr. J. Range Forage Sci. 36, 129–133. doi: 10.2989/10220119.2019.1573760

Thaikua, S., Ebina, M., Yamanaka, N., Shimoda, K., Suenaga, K., and Kawamoto, Y. (2016). Tightly clustered markers linked to an apospory-related gene region and quantitative

trait loci mapping for agronomic traits in Brachiaria hybrids. Grassl. Sci. 62, 69–80. doi: 10.1111/grs.12115

Timbó, A. L. de O., Souza, P. N. da C., Pereira, R. C., Nunes, J. D., Pinto, J. E. B. P., Souza Sobrinho, F. de, & Davide, L. C. (2014). Obtaining tetraploid plants of ruzigrass (Brachiaria ruziziensis). In Revista Brasileira de Zootecnia (Vol. 43, Issue 3, pp. 127–131). FapUNIFESP (SciELO). https://doi.org/10.1590/s1516-35982014000300004

Valle, C.B. Coleção de germoplasma de espécies de Brachiaria no CIAT: estudos básicos visando ao melhoramento genético. Documentos no. 46. Embrapa Gado de Corte, Campo Grande, 1990.

Valle, C.B.; Simioni, C.; Resende, R.M.S.; Jank, L. (2008) Melhoramento genético de Brachiaria. In: Resende, R.M.S.; Valle C.B.; Jank, L. (eds.) Melhoramento de Forrageiras Tropicais. Embrapa Gado de Corte, Campo Grande, MS. pp 13-54.

Valle, C.B., Jank, L., and Resende, R.M.S. (2009). O melhoramento de forrageiras tropicais no Brasil. Ceres, 460-472.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. In Journal of Dairy Science (Vol. 91, Issue 11, pp. 4414–4423). American Dairy Science Association. https://doi.org/10.3168/jds.2007-0980

Vasaikar, S. V., Straub, P., Wang, J., & Zhang, B. (2017). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. In Nucleic Acids Research (Vol. 46, Issue D1, pp. D956–D963). Oxford University Press (OUP). https://doi.org/10.1093/nar/gkx1090

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. Genet. Mol. Biol. 39, 312–328. doi: 10.1590/1678-4685-GMB-2016-0027

Vigna, B. B. Z., de Oliveira, F. A., de Toledo-Silva, G., da Silva, C. C., do Valle, C. B., and de Souza, A. P. (2016b). Leaf transcriptome of two highly divergent genotypes of Urochloa humidicola (Poaceae), a tropical polyploid forage grass adapted to acidic soils and temporary flooding areas. BMC Genomics 17:910. doi: 10.1186/s12864-016-3270-5

Vigna, B. B. Z., Jungmann, L., Francisco, P. M., Zucchi, M. I., do Valle, C. B., and de Souza, A. P. (2011a). Genetic diversity and population structure of the Brachiaria brizantha germplasm. Trop. Plant Biol. 4, 157–169. doi: 10.1007/s12042-011-9078-1

Vigna, B. B. Z., Santos, J. C. S., Jungmann, L., do Valle, C. B., Mollinari, M., Pastina, M. M., et al. (2016a). Evidence of allopolyploidy in Urochloa humidicola based on

cytological analysis and genetic linkage mapping. PLoS One 11:e0153764. doi: 10.1371/journal.pone.0153764

Wang, X., Xu, Y., Hu, Z., & Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. In The Crop Journal (Vol. 6, Issue 4, pp. 330–340). Elsevier BV. https://doi.org/10.1016/j.cj.2018.03.001

Whalen, A., Gorjanc, G., and Hickey, J. M. (2019). Parentage assignment with genotyping-by-sequencing data. J. Anim. Breed. Genet = Zeits. Tierz. Zucht. 136, 102–112. doi: 10.1111/jbg.12370

Wolfe, C.J., Kohane, I.S. and Butte, A.J. (2005) Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. BMC Bioinform., 6, 227–227. https://doi.org/10.1186/1471-2105-6-227

Worthington, M., Ebina, M., Yamanaka, N., Heffelfinger, C., Quintero, C., Zapata, Y. P., et al. (2019). Translocation of a parthenogenesis gene candidate to an alternate carrier chromosome in apomictic Brachiaria humidicola. BMC Genomics 20:41. doi: 10.1186/s12864-018-5392-4

Worthington, M., Heffelfinger, C., Bernal, D., Quintero, C., Zapata, Y. P., Perez, J. G., et al. (2016). A parthenogenesis gene candidate and evidence for segmental allopolyploidy in apomictic Brachiaria decumbens. Genetics 203, 1117–1132. doi: 10.1534/genetics.116.190314

Worthington M., Perez J. G., Mussurova S., Silva-Cordoba A., Castiblanco V., Jones C., et al. . (2021). A new genome allows the identification of genes associated with natural variation in aluminium tolerance in Brachiaria grasses. J. Exp. Bot. 72, 302–319. doi: 10.1093/jxb/eraa469

Wright, I. A. et al. Integrating crops and livestock in subtropical agricultural systems. Journal of the Science of Food and Agriculture, v. 92, p. 1010-1015, 2012.

Zargar, S. M., Raatz, B., Sonah, H., Bhat, J. A., Dar, Z. A., Agrawal, G. K., & Rakwal, R. (2015). Recent advances in molecular marker techniques: insight into QTL mapping, GWAS and genomic selection in plants. Journal of crop science and biotechnology, 18(5), 293-308.

Zhao, X., Zhang, J., Zhang, Z., Wang, Y., and Xie, W. (2017). Hybrid identification and genetic variation of Elymus sibiricus hybrid populations using EST-SSR markers. Hereditas 154:15. doi: 10.1186/s41065-017-0053-1

Zhou, W., Bellis, E.S., Stubblefield, J., Causey, J., Qualls, J., Walker, K. and Huang, X. (2019) Minor QTLs mining through the combination of GWAS and machine learning feature selection. BioRxiv, 712190. https://doi.org/10.1101/712190

Zwart, A. B., Elliott, C., Hopley, T., Lovell, D., and Young, A. (2016). polypatex: anrpackage for paternity exclusion in autopolyploids. Mol. Ecol. Resour. 16, 694–700. doi: 10.1111/1755-0998.12496

Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform. 2018;19:1370-1381.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications, 5, 3231. https://doi.org/10.1038/ncomms4231

ANEXOS

Declaração de Bioética

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "***Identificação de contaminantes e análises multiômicas aplicadas no melhoramento molecular de forrageiras tropicais poliplóides***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____
Nome do(a) aluno(a): Felipe Bitencourt Martins

Assinatura: _____
Nome do(a) orientador(a): Anete Pereira de Souza

Data: 7 de fevereiro de 2023

Declaração de Direitos Autorais

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Identificação de contaminantes e análises multiômicas aplicadas no melhoramento molecular de forrageiras tropicais poliplóides**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 7 de fevereiro de 2023

Assinatura : _____
Nome do(a) autor(a): **Felipe Bitencourt Martins**
RG n.° 48.898.020-3

Assinatura : _____
Nome do(a) orientador(a): **Anete Pereira de Souza**
RG n.° 8.680.325-6